Bundeskriminalamt

Nathalie Guzy
Christoph Birkel
Robert Mischkowitz (Hrsg.)

Viktimisierungsbefragungen in Deutschland

Band 2

Methodik und Methodologie

Viktimisierungsbefragungen in Deutschland

Band 2

Methodik und Methodologie

Polizei + Forschung

Band 47.2

Herausgegeben vom
Bundeskriminalamt
Kriminalistisches Institut

Bundeskriminalamt

Nathalie Guzy,
Christoph Birkel,
Robert Mischkowitz (Hrsg.)

Viktimisierungsbefragungen in Deutschland

Band 2

Methodik und Methodologie

BKA

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de** abrufbar.

Alle Publikationen der BKA-Reihe Polizei + Forschung (ausgenommen VS-NfD-eingestufte Bände) sind im Internet im PDF-Format unter www.bka.de (Kriminalwissenschaften/Kriminalistisches Institut) eingestellt.

Korrektorat und Redaktion:

Lars Wiedemann

Bundeskriminalamt
Kriminalistisches Institut

Lektorat: Wissenschaftslektorat Zimmermann, Magdeburg
Herstellung: Griebsch und Rochol Druck GmbH, Hamm

Vorwort

Die Polizeiliche Kriminalstatistik (PKS) dokumentiert seit nunmehr 62 Jahren die Kriminalitätslage in Deutschland. Doch können in der PKS nur *die* Straftaten ausgewiesen werden, die der Polizei – sei es durch die Anzeige von Bürgerinnen und Bürgern, sei es durch eigene Ermittlungen – zur Kenntnis gelangen. Diesem so genannten Hellfeld der Kriminalität steht ein – je nach Deliktsart unterschiedlich großes – „Dunkelfeld“ gegenüber.

Seit mehr als vier Jahrzehnten versuchen Wissenschaftlerinnen und Wissenschaftler mit persönlichen, postalischen, telefonischen und in den letzten Jahren zunehmend auch Online-Opferbefragungen Zugang zum Dunkelfeld zu bekommen. Je nach Forschungsfrage und -intention wurden entweder einzelne Opfergruppen wie Jugendliche, Seniorinnen und Senioren oder Frauen untersucht oder Kriminalitätslagen in begrenzten lokalen oder regionalen Bereichen beleuchtet (sog. „Kriminologische Regionalanalysen“).

Mit dem vorliegenden Sammelband setzt das BKA sein Engagement auf dem Gebiet der Opferbefragungen fort mit dem Ziel, die vorhandenen einzelnen Forschungsergebnisse zu bündeln, um eine bislang noch fehlende systematische Darstellung des aktuellen Forschungsstandes vorzulegen. Neben der zentralen Dokumentation der Forschungsergebnisse lassen sich im vorliegenden Werk auch Anregungen für künftige wissenschaftliche Arbeiten und Projekte finden. Mit dieser Wissensbasis sollen nicht nur Wissenschaftlerinnen und Wissenschaftler, sondern auch Angehörige von Politik und Polizeipraxis angesprochen werden, indem kriminalpolitische Anregungen und Umsetzungsmöglichkeiten in der Polizeipraxis besonders herausgestellt sowie Hinweise und Interpretationshilfen zum besseren Verständnis der Ergebnisse, Möglichkeiten und Grenzen von Opferbefragungen dargelegt werden.

Die Realisierung eines solchen Werks ist ohne die Unterstützung vieler Mitwirkender nicht möglich. Zunächst einmal sind an dieser Stelle die Autorinnen und Autoren zu nennen, die sich pro bono dazu bereit erklärt haben, einen Teil ihrer oft geringen zeitlichen Ressourcen zur Verfügung zu stellen, um die einzelnen Beiträge und damit das Herzstück des vorliegenden Sammelbands zu verfassen. Vor allem Prof. Dr. Helmut Kury, Dr. Joachim Obergfell-Fuchs, Privatdozent Dr. Dietrich Oberwittler und Prof. Dr. Peter Wetzels danke ich für ihren Einsatz, den sie – neben ihren Beiträgen – bereits während der Konzeption der Sammelbandstruktur gezeigt haben. Mit Hilfe aller Beteiligten ist es gelungen, ein Werk zu schaffen, dass zukünftigen Opferbefragungs-Projekten eine reichhaltige und hilfreiche Wissensbasis bieten wird.

Holger Münch

Präsident des Bundeskriminalamts

Inhaltsverzeichnis

2 Amtliche Daten der Kriminalstatistik versus Daten aus Opferbefragungen – Vergleichsschwierigkeiten und Kombinationsmöglichkeiten

3 Analyse der Ergebnisse von Viktimisierungsbefragungen

4 Grenzen von Opferbefragungen

5 Zusammenfassung und Implikationen für die Praxis

Einleitung

Nathalie Guzy, Christoph Birkel und Robert Mischkowitz

Dass die methodische Vorgehensweise in Umfragen erhebliche Auswirkungen auf die Ergebnisse derselben hat, darf heutzutage sowohl unter Methodikern als auch unter Laien als unbestritten gelten. Die Frage allerdings, in welche Richtung(en) und mit welchem Umfang Effekte einzelner Methoden auf die jeweiligen Ergebnisse wirken bzw. welcher optimale methodische Umgang mit derartigen Effekten für verschiedene (inhaltliche) Fragestellungen angezeigt ist, kann dagegen nur noch ein kleiner Teil meist spezialisierter Methodikerinnen oder Methodiker beantworten.

Vergleichbar mit dem Forschungsstand zu den (deliktspezifischen) Ergebnissen aus Dunkelfeld-Opferbefragungen zeichnet sich auch der Forschungsstand zur Methodik und Methodologie durch eine Vielzahl einzelner, teils veralteter, größtenteils englischsprachiger und auf ausländischen Datensätzen beruhender Forschungspapiere aus. Der Großteil der Veröffentlichungen basiert auf den Daten des amerikanischen *National Crime Victimization Surveys* (NCVS), auf dessen Basis insbesondere in den 80er und 90er Jahren diverse methodische Untersuchungen durchgeführt wurden (z. B. Skogan 1981, 1986). Auch wenn ein bedeutender Teil der dort generierten Erkenntnisse für deutsche bzw. europäische Opferbefragungen genutzt werden kann (so z. B. zum Problem des Telescoping, der Ehrlichkeit von Antworten oder der Fragereihenfolge), muss bei vielen Erkenntnissen – nicht zuletzt aufgrund der speziellen Methodik des NCVS – bezweifelt werden, dass der dort generierte Forschungsstand problemlos auf deutsche Dunkelfeld-Opferbefragungen übertragbar ist (so z. B. zur Bedeutung und Systematik von Ausfällen durch Umzüge der Befragungspersonen oder zur Operationalisierung und Zählung von Opfererlebnissen).

Zwar finden sich auch in Deutschland vereinzelte Untersuchungen und Veröffentlichungen, die sich den Methoden von Opferbefragungen widmen, diese sind aber selten bzw. beziehen sich nur auf ausgewählte (Teil-)Fragestellungen (z. B. zu Antworttendenzen bei der Erhebung von Strafeinstellungen oder sozialer Erwünschtheit bei der Abfrage sensibler Opfererlebnisse; siehe Reuband 2007, für einen Überblick Wetzels 1996). Hinzu kommt der Umstand, dass der Großteil dieser Untersuchungen bereits älteren Datums ist und somit nicht nur hinsichtlich der jeweiligen Fragestellung, sondern auch bezüglich der methodischen Vorgehensweise – zumindest partiell – als veraltet gelten kann bzw. nicht mehr dem methodischen State of the Art entspricht.

Zusammenfassend kann also festgestellt werden, dass Veröffentlichungen über den aktuellen methodologischen und erhebungspraktischen Forschungsstand in Deutschland auch zu deliktspezifischen Viktimisierungsbefragungen ebenso wenig existieren wie Arbeiten, aus denen praktische Anwendungshilfen für die Interpretation, Analyse und Bewertung von Ergebnissen aus Viktimisierungsbefragungen hervorgehen.

Der vorliegende Sammelband zielt daher neben der Darstellung des aktuellen Forschungsstands im Bereich deliktspezifischer Dunkelfeld-Opferbefragungen (Band I) auf die Aufbereitung und kritische Diskussion der methodologischen Grundlagen und Probleme bei deren Durchführung und Bewertung ab. Insbesondere Letzteres spielt in der bisher vorliegenden Forschungsliteratur nur eine untergeordnete Rolle – obwohl diese Informationen für eine adäquate Einordnung und Bewertung von Daten aus Opferbefragungen zentral sind.[1]

Mit dem vorliegenden zweiten Band werden somit zwei Ziele verfolgt: *Erstens* soll den Lesern und Leserinnen eine möglichst umfassende Grundlage zur Durchführung eigener Opferbefragungen gegeben werden – und zwar derart, dass diese dem aktuellen nationalen und (sofern nicht vorhanden) ggf. internationalen Forschungsstand entspricht. *Zweitens* soll der Sammelband das notwendig erscheinende Methodenwissen vermitteln, um Ergebnisse aus Opferbefragungen adäquat, d. h. vor dem Hintergrund ihrer Methodik bewerten bzw. interpretieren zu können.

Der Aufbau dieses zweiten Sammelbands orientiert sich eng an dem „Lebenszyklus von Umfragen“ sowie dem darauf basierenden Fehlermodell von Groves und Kollegen (2009). Als besonders relevant für die Durchführung und Bewertung von Opferbefragungen wurden insbesondere die Bereiche Stichprobenbildung, Effekte des Erhebungsmodus, der Fragebogengestaltung sowie der statistischen Auswertung, insbesondere unter (kausaltheoretischer) Berücksichtigung des Erhebungsdesigns identifiziert. Ebenfalls Berücksichtigung finden sollte der opferbefragungsspezifische Bereich der Gegenüberstel-

[1] Mit dem Abschlussbericht der Arbeitsgruppe „Regelmäßige Durchführung von Opferbefragungen (BUKS)“ wurden zwar erstmalig zentrale methodologische und inhaltliche Gesichtspunkte nationaler Dunkelfeld-Opferbefragungen zusammengefasst, der Bericht konnte jedoch bisher weder einer breiteren Öffentlichkeit zugänglich gemacht werden, noch liefert er einen Gesamtüberblick über die aktuelle Forschungslage in Deutschland. Auch das von der UN erstellte „Manual on Victimization Surveys“ liefert keinen adäquaten Ersatz für eine Zusammenstellung des aktuellen Forschungstands, da dessen Fokus auf forschungspraktischen Empfehlungen liegt mit dem Ziel international vergleichbare Ergebnisse zu generieren (UNODC/UNECE 2010).

lung von Hell- und Dunkelfelddaten.[2] Erfreulicherweise konnten für all diese Themen renommierte Autoren und Autorinnen gewonnen werden.

Den ersten Teil zu erhebungsmethodischen Grundlagen von Opferbefragungen beginnen *Schnell/Noack* mit einer Einführung in die Grundlagen und Besonderheiten der Stichprobenziehung bei Opferbefragungen, die Wechselbeziehung zwischen Erhebungsmodus und Auswahlgrundlage sowie die sich dabei ergebenden Probleme durch Nonresponse. Es folgt der Beitrag von *Kury/Guzy/Leitgöb* zu den verschiedenen Effekten sowie Vor- und Nachteilen einzelner Erhebungsverfahren, insbesondere mit Blick auf Erkenntnisse aus dem Bereich von Opferbefragungen. *Killias* vervollständigt den Bereich der Erhebungsmethoden durch einen Beitrag zu neueren internetbasierten Erhebungsmethoden sowie Mixed-Mode-Surveys, die aufgrund der aktuellen technischen und methodischen Entwicklungen gesondert betrachtet werden. In dem Beitrag von *Faulbaum* werden die Grundlagen der Fragebogen- und Item-Entwicklung mit besonderem Fokus auf Opferbefragungen ausgeführt. Dieser Themenkomplex wird durch den Beitrag von *Waubert de Puiseau, Hoffmann und Musch* ergänzt, und zwar mittels Darstellung der Problematik sozial erwünschter Antworten und der Abfrage sensibler Viktimisierungserfahrungen. Der erste Teil des zweiten Bands wird durch den Beitrag von *Hatt* abgeschlossen, der sich mit datenschutzrechtlichen Besonderheiten und Problemen im Zusammenhang mit Opferbefragungen beschäftigt.

Der zweite Teil des vorliegenden Sammelbands ist der Gegenüberstellung von amtlichen Daten der Kriminalstatistik und Daten aus Opferbefragungen gewidmet. Da in Deutschland mangels regelmäßiger, statistikbegleitender Opferbefragungen diesbezüglich nur limitierte Erfahrungen vorliegen, stellt *Norris* zunächst Erfahrungen und Möglichkeiten auf Basis des *British Crime Surveys* (mittlerweile *Crime Survey for England and Wales*) dar. Im Anschluss daran werden von *Heinz* die methodischen Schwierigkeiten bei der Gegenüberstellung von Daten der Polizeilichen Kriminalstatistik und Opferbefragungen in Deutschland herausgearbeitet.

Der dritte Teil des Sammelbands konzentriert sich auf die Analyse der Ergebnisse von Opferbefragungen. In dem Beitrag von *Hanslmaier/Baier* werden anhand von Beispielen die gängigen statistischen Analyseverfahren sowie deren Probleme bei einer adäquaten Auswertung von Opferbefragungen vorgestellt. *Leitgöb/Seddig* widmen sich dann den gängigen Forschungsdesigns

---

[2] Der Vollständigkeit halber soll an dieser Stelle dennoch erwähnt werden, dass in der ursprünglichen Planung des Sammelbands zwei weitere Beiträge, und zwar zu den Themen „geografische Kriminalitätsanalyse“ sowie „forschungsethische Grundlagen“ geplant waren. Der erste Beitrag konnte kurzfristig krankheitsbedingt nicht umgesetzt werden, für den zweiten Beitrag konnte – trotz intensiver Bemühungen – kein Autor bzw. keine Autorin gewonnen werden.

von Opferbefragungen und den damit zusammenhängenden Möglichkeiten zur Aufdeckung kausaler Effekte (als eines der zentralen Ziele von Opferbefragungen).

Der Sammelband endet mit dem Beitrag von *Kury* zu den Grenzen von Opferbefragungen und einer Zusammenfassung zentraler Erkenntnisse sowie Skizzierung ihrer praktischen Implikationen durch die Herausgeberschaft.

Zu guter Letzt sei noch auf eine konzeptionelle Besonderheit dieses Sammelbands hingewiesen: Wie bereits im ersten Band dieses Kompendiums ausgeführt wurden alle Beiträge für eine breite Leserschaft konzipiert und formuliert. Dabei wurde stets versucht, spezifische Fachbegriffe allgemeinverständlich zu erklären, um die Beiträge auch für Nichtstatistikerinnen und Nichtstatistiker bzw. Nichtmethodikerinnen und Nichtmethodiker nachvollziehbar zu machen. Es dürfte allerdings verständlich sein, dass ein Methodenband, der das Ziel verfolgt, den aktuellen Forschungsstand im Lichte aktueller und innovativer Methoden darzustellen, nicht in allen Teilen und in der gesamten Detailtiefe auch von Nichtmethodikern und Nichtmethodikerinnen verstanden werden kann. Es haben jedoch alle Autorinnen und Autoren versucht, die Kernaussagen und zentralen Punkte eines jeden Beitrags allgemein verständlich darzulegen, so dass diese auch ohne spezielle Fachkenntnisse verstanden werden können (auch wenn möglicherweise einzelne Passagen nicht detailliert nachvollzogen werden können).

Für diese Bemühungen – sicherlich eine der größten Herausforderungen dieses Sammelbands – sei an dieser Stelle nochmals allen Autorinnen und Autoren ausdrücklich gedankt.

Literatur

Groves, Robert M.; Fowler, Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor und Tourangeau, Roger (2009): Survey Methodology, 2. Aufl. Hoboken, NJ: Wiley.

Reuband, Karl-Heinz (2007): Reihenfolgeeffekte bei Fragen zum Sanktionsverlangen: Macht es in Umfragen einen Unterschied, welche Strafe an welche Stelle der Antwortvorgaben genannt wird? In: Monatsschrift für Kriminologie und Strafrechtsreform, 90, 5, S. 409–417.

Skogan, Wesley (1981): Issues in the Measurement of Victimization. U.S. Department of Justice. Washington DC. URL: https://www.ncjrs.gov/pdffiles1/Digitization/74682NCJRS.pdf.

Skogan, Wesley (1986): Methodological Issues in the Study of Victimization. In: Fattah, Ezzat A. (Hg.): From Crime Policy to Victim Policy. Basingstoke: Macmillan, S. 80–116.

UNODC/UNECE (2010): UNODC-UNECE Manual on Victimization Surveys. URL: http://www.unodc.org/unodc/en/data-and-analysis/Manual-on-victim-surveys.html.

Wetzels, Peter (1996): Kriminalität und Opfererleben: Immer öfter immer das Gleiche? Defizite und Perspektiven repräsentativer Opferbefragungen als Methode empirisch-viktimologischer Forschung in der Kriminologie. In: Monatsschrift für Kriminologie und Strafrechtsreform, 79, 1, S. 1–24.

1 Erhebungsmethodische Grundlagen

Stichproben, Nonresponse und Gewichtung für Viktimisierungsstudien

Rainer Schnell und Marcel Noack

1 Vollerhebungen und Stichproben[1]

Nur bei wenigen Projekten stehen genug Ressourcen zur Verfügung, um alle interessierenden Personen (die Population oder Grundgesamtheit) untersuchen oder befragen zu können (Vollerhebung). Daher werden zumeist nur Teilmengen einer Population untersucht. Dies wird als „Stichprobenuntersuchung“ bezeichnet. Der Sinn einer Stichprobenuntersuchung besteht darin, von der Stichprobe auf die Grundgesamtheit schließen zu können. Für diese Verallgemeinerbarkeit einer Stichprobenuntersuchung ist die Art der Stichprobenziehung von zentraler Bedeutung. Wird diese Ziehung nicht über einen berechenbaren Zufallsprozess durchgeführt, dann ist die resultierende Stichprobe nachträglich nur sehr schwierig als verallgemeinerbar zu rechtfertigen. Mathematisch korrekt ist für nahezu alle praktischen Anwendungen lediglich die Verwendung echter Zufallsstichproben (Abschnitt 5). In diesem Beitrag werden wir nur solche Verfahren vorstellen, die für die Ziehung echter Zufallsstichproben prinzipiell geeignet sind. Vor allem für die tatsächliche Konstruktion, Ziehung und Gewichtung bundesweiter Stichproben der allgemeinen Bevölkerung sind umfangreichere praktische Kenntnisse erforderlich, als sie in einer Einführung vermittelt werden können. Der vorliegende Beitrag stellt nur diejenigen Details dar, die uns für die Beurteilung und Planung der Stichproben von Viktimisierungsstudien unverzichtbar erscheinen.[2]

---

[1] Einzelne Teile des Beitrags basieren auf früheren Arbeiten der Autoren, vor allem auf Schnell 2012 und Schnell u. a. 2013. Für kritische Anmerkungen danken wir Sabrina Torregroza und Christian Borgs.

[2] Dementsprechend setzen wir nur Grundkenntnisse der Inferenzstatistik, wie sie jedes einführende Lehrbuch der Statistik jenseits der deskriptiven Statistik vermittelt, voraus.

2 Konsequenzen des Studiendesigns für die Art der Stichprobenziehung

Beim Design einer Untersuchung muss man sich darüber klar sein, ob Aussagen über den Zustand einer Untersuchungsgruppe zu einem bestimmten Zeitpunkt erforderlich sind (Querschnittserhebungen) oder Aussagen über Veränderungen derselben Personen (Panel- oder Kohortenstudien) gefordert werden. Erhebungen zu einem Zeitpunkt erlauben nur sehr begrenzt Aussagen über Veränderungen, da die Verwendung von Fragen über vergangene Zustände (Retrospektivfragen) immer unter den vielfältigen Problemen autobiografischer Erinnerungen leidet (Schwarz 2007). Aufgrund der Dauer und der erheblichen Kosten von Panelstudien wird trotz der prinzipiell unüberwindbaren Probleme von Retrospektivfragen zumeist eine Entscheidung für eine einmalige Querschnittserhebung gefällt.[3] Wir gehen im Folgenden davon aus, dass es sich bei dem Projekt, für das eine Stichprobenziehung erforderlich ist, um eine einmalige Erhebung handelt.[4]

3 Angestrebte Grundgesamtheit, Auswahlgesamtheit und Inferenzpopulation

Zu Beginn eines Forschungsprojekts muss zunächst die Grundgesamtheit festgelegt werden, für die Aussagen beabsichtigt sind. Die Menge dieser Personen wird als angestrebte Grundgesamtheit bezeichnet. So könnte man z. B.

---

[3] Übersichten über das Design von Viktimisierungsstudien finden sich bei Lynch 2006, Groves/Cork 2008 und Aebi/Linde 2014. Hinweise für das Design von Viktimisierungsstudien in Deutschland finden sich bei Schnell/Hoffmeyer-Zlotnik 2002.

[4] Wird eine wiederholte Erhebung geplant, bleiben alle prinzipiellen Erwägungen gleich; allerdings wird je nach Art der Wiederholung ein höherer Aufwand erforderlich. Handelt es sich um eine unabhängige Wiederholungsstudie (wiederholte Querschnitte) wird nur ein höherer Dokumentationsaufwand notwendig, da das Stichprobenverfahren bei der Wiederholung exakt repliziert werden muss. Ansonsten können Veränderungen zwischen den Erhebungszeitpunkten nicht mehr auf Veränderungen der Personen zurückgeführt werden. Ist eine wiederholte Befragung derselben Personen beabsichtigt (Panel- oder Kohortenstudie), dann wird neben dem erhöhten Dokumentationsaufwand eine in der Regel größere Ausgangsstichprobe erforderlich, da allein schon durch Tod und Wanderung bei den Teilnehmern der ersten Welle mit Verlusten in der zweiten Welle gerechnet werden muss (mindestens 10 % pro Jahr bei einer Zufallsstichprobe aus der Bevölkerung). Dazu kommen weitere Ausfälle durch Verweigerung. Schließlich muss mit erheblichem Aufwand zur erneuten Kontaktierung und Sicherstellung der Befragung derselben Person (Identitätsmanagement) gerechnet werden. Die resultierenden unvermeidlichen Probleme haben Konsequenzen für die Planung und Durchführung schon bei der Stichprobenziehung. Die Details sprengen den Rahmen dieser Übersicht; es muss auf die Literatur zur Planung von Längsschnittstudien verwiesen werden. Einzelheiten finden sich bei Schnell 2012.

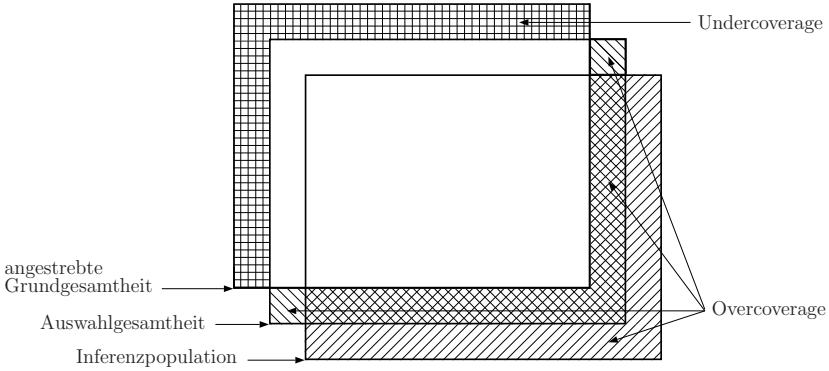
die zu einem Stichtag in Deutschland lebenden Menschen als angestrebte Grundgesamtheit betrachten. Eine solche Definition stößt auf zahlreiche Probleme. Abgesehen von den Ungenauigkeiten der verwendeten Begriffe liegt das Problem vor allem darin, dass es keine vollständigen Listen oder Datenbanken dieser Grundgesamtheit gibt. Es gibt in Deutschland lediglich Annäherungen an solche Listen, z. B. die Gesamtheit aller Einwohnermeldedateien der mehr als 5.600 Meldeämter oder die Datei der mehr als 90 Millionen vergebenen Steueridentifikationsnummern. Man könnte die Menge der Personen, die in diesen Listen enthalten sind (die sogenannte Auswahlgesamtheit oder *frame population*) als Annäherung an die angestrebte Grundgesamtheit betrachten. Zwar ist faktisch weder eine vollständige Einwohnermeldeliste noch die Datei der Steueridentifikationsnummern für Stichprobenziehungen verfügbar, neben den unüberwindlichen Problemen des Zugangs bestehen aber noch prinzipielle Probleme: Derartige Listen sind unvollständig, enthalten Duplikate und überzählige Personen.

Im Zusammenhang mit Stichproben werden solche Probleme als Coverage-Probleme bezeichnet. Es wird in der Regel unterschieden zwischen Undercoverage und Overcoverage (*Abbildung 1*). In den Einwohnermeldedateien fehlen z. B. illegale Migranten, diese werden zum Undercoverage gezählt. Es finden sich aber in der Regel auch mehrfach erfasste Personen (z. B. durch Haupt- und Nebenwohnsitze) sowie Personen, die nicht zur angestrebten Grundgesamtheit gehören (z. B. Verstorbene oder ins Ausland Verzogene), die als Overcoverage gerechnet werden. Ziel der Konstruktionen einer Auswahlgesamtheit (eines *frames*) ist die möglichst große Übereinstimmung zwischen angestrebter Grundgesamtheit und Auswahlgesamtheit, also ein möglichst geringes Ausmaß an Overcoverage und Undercoverage.[5]

[5] Über die fehlerhaften Listen hinaus können auch Fehler, die durch Interviewer oder die befragten Personen entstehen, zu Overcoverage führen. Als Beispiele können kaserniertes militärisches Personal auf Heimaturlaub, das durch den Interviewer fälschlich als Teil des zu befragenden Haushalts angesehen und interviewt wird, oder aufgrund fehlerhafter Angaben befragte Personen (minderjährige Person gibt an, volljährig zu sein) genannt werden.

Abbildung 1:

Verhältnis von angestrebter Grundgesamtheit, Auswahlgesamtheit und Inferenzpopulation (Schnell u. a. 2013, 262)



Durch die tatsächliche Durchführung fallen weitere Personen aus der Auswahlgesamtheit heraus (z. B. durch mangelnde Sprachkenntnisse). Gelegentlich werden auch Personen berücksichtigt, die nicht zur angestrebten Grundgesamtheit gehören, z. B. wenn bei telefonischen Befragungen minderjährige Personen befragt werden, aber nur Erwachsene befragt werden sollen. Die Menge der Personen, aus denen die tatsächlich Befragten eine Zufallsstichprobe darstellen würden, wird als „Inferenzpopulation“ bezeichnet, weil nur über diese Aussagen gemacht werden können. Die Inferenzpopulation sollte also der angestrebten Grundgesamtheit möglichst entsprechen.

4 Aus Bevölkerungserhebungen ausgeschlossene Populationen

In den meisten Projekten erfolgt die Definition der Grundgesamtheit bis zur Ziehung der Stichprobe kaum explizit. Selbst für die Vorbereitung der Stichprobenziehung wird die Grundgesamtheit selten exakt definiert. Üblich sind ungenaue Kurzdefinitionen der Grundgesamtheit wie z. B. „erwachsene Wohnbevölkerung“ oder „in Privathaushalten lebende deutsche Staatsangehörige ab 18 Jahren“. Solche Definitionen sind für eine praktische Umsetzung nicht exakt genug. Die Ungenauigkeiten sind dabei keinesfalls folgenlos, da nicht klar ist, welche Populationen ausgeschlossen sind und welche nicht (Unterkapitel 4.6).

Im Folgenden werden einige der ausgeschlossenen Populationen erwähnt und ihre Größenordnung abgeschätzt.[6] Am bedeutsamsten ist in diesem Zusammenhang die Bevölkerung in Institutionen.

4.1 Bevölkerung in Institutionen

In der älteren deutschen Literatur wurde diese Gruppe als „Anstaltsbevölkerung“ bezeichnet, im Zensus 2011 als „Sonderbereiche“. Darunter fallen laut § 2 Abs. 5 S. 1–3 des Zensusgesetzes 2011 Gemeinschafts-, Anstalts- und Notunterkünfte, Wohnheime und ähnliche Unterkünfte.[7]

Im Zensus 2011 wurde zwischen sensiblen und nicht sensiblen Sonderbereichen unterschieden. Zu Ersteren gehören Behindertenwohnheime, spezielle Krankenhäuser (wie z. B. Palliativstationen, Hospize, psychiatrische Kliniken), Justizvollzugsanstalten und andere Einrichtungen des Maßregelvollzugs sowie Flüchtlingsunterkünfte und Unterkünfte für Wohnungslose. Sowohl zu sensiblen als auch nicht sensiblen Sonderbereichen können Kinder- und Jugendheime sowie Mutter-Kind-Heime zählen. Als nicht sensible Sonderbereiche galten Studentenwohnheime, Arbeiterheime und sonstige Wohnheime, Alten- und Pflegeheime, Internate, Schulen des Gesundheitswesens sowie Klöster.

Legt man den Anteil der Personen zugrunde, die in Niedersachsen in sensiblen Sonderbereichen wohnen (Mayer 2013), dann wären insgesamt in Deutschland ca. 271.000 Personen in sensiblen Sonderbereichen, für die nicht sensiblen Sonderbereiche entsprechend 1.375.000 Personen zu erwarten. Damit wären insgesamt 1,646 Millionen Personen oder etwas mehr als 2 % der Bevölkerung in Sonderbereichen zu finden. Verglichen mit den USA (2,5 %)[8], erscheint dies als eine plausible Größenordnung.

---

6 Eine ausführliche Übersicht über die Größe faktisch aus der Befragung der „allgemeinen Bevölkerung“ ausgeschlossener Populationen in Deutschland findet sich bei Schnell 1991. Trotz des Alters der Studie ist dies bislang die einzige Publikation zu diesem Problem in Deutschland.

7 Die Dokumentation des Zensus 2011 ist auch in diesem Bereich spärlich; einige wenige Einzelheiten zur Erhebung in Sonderbereichen finden sich bei Geiger/Styhler 2012 und Mayer 2013.

8 Für die USA werden in „Group Quarters“ insgesamt nahezu 8 Millionen Personen geschätzt, die Hälfte davon in Institutionen (National Research Council 2012, 25).

4.2 Personen mit besonderen Wohnungsbedingungen

In allen Industriegesellschaften lebt ein Teil der Bevölkerung nicht in Wohngebäuden, sondern in anderen Gebäuden oder Unterkünften. Nach den ersten Ergebnissen (Statistische Ämter des Bundes und der Länder 2014) der Gebäude- und Wohnungszählung gab es 2011 3,4 % der Wohnungen in „sonstigen Gebäuden mit Wohnraum“ (Schulen, Gewerbeobjekte etc.). Falls die Stichprobenziehung durch eine Begehung vor Ort erfolgt (z. B. in den sogenannten Random-Route-Verfahren), werden Personen in Nichtwohngebäuden allgemein häufig nicht erfasst, so z. B. Hausmeister, Mönche, Nonnen oder Personen in Bereitschaftsunterkünften der Polizei oder der Bundeswehr (Schnell 1991).

Ein anderes Problem sind „bewohnte Unterkünfte“ wie Wohn- oder Bauwagen, Baracken, Gartenlauben, fest verankerte Wohnschiffe, Schrebergartengebäude und Weinberghütten sowie Dauercampende. Obwohl der größte Teil dieser Population an irgendeiner Stelle administrativ erfasst ist, kann der tatsächliche Wohnort von der erfassten Anschrift abweichen. Je nach Art der verwendeten Auswahlgrundlage, z. B. Listen von Telefonnummern oder Einwohnermelderegister, können solche Populationen enthalten sein oder nicht. Im Einzelfall hängt die Auswahl dieser Personen jedoch zumeist von willkürlichen Entscheidungen ab. Für den Zensus 2011 werden ca. 10.000 bewohnte Unterkünfte mit ca. 15.000 Wohnungen nachgewiesen (Statistische Ämter des Bundes und der Länder 2014).[9] Sollte sich die Haushaltszusammensetzung gegenüber der Volkszählung von 1987 (VZ87) nicht verändert haben, entspräche dies 27.000 Personen. Schließlich sollen noch die Menschen erwähnt werden, die ohne jede Unterkunft auf der Straße leben: Die Bundesarbeitsgemeinschaft Wohnungslosenhilfe gibt diese Zahl mit 24.000 für 2012 an (Bundesarbeitsgemeinschaft Wohnungslosenhilfe 2013).

4.3 Klandestine Populationen

Mit dem Begriff ‚klandestine Populationen‘ werden seit einigen Jahren Subgruppen bezeichnet, die sich willentlich einer amtlichen Erfassung entziehen. Hierzu gehören zunächst einmal Personen ohne Aufenthaltsgenehmigung oder -duldung (ohne Scheinlegale oder registrierte Ausreisepflichtige) (Vogel/Äßner 2011). Dies dürfte den größten Teil dieser Populationen darstellen. Für Deutschland existieren nur höchst unvollkommene Schätzungen in der Grö-

[9] Für die Volkszählung von 1987 wurden 25.400 Haushalte mit 45.600 Personen unter solchen Bedingungen gezählt (Schnell 1991). Es erscheint erstaunlich, dass sich diese Zahl trotz Wiedervereinigung gegenüber der VZ87 mehr als halbiert haben soll.

ßenordnung von 100.000 bis 675.000 Personen (Vogel/Äßner 2011), vereinzelt finden sich auch höhere Angaben, z. B. bei CLANDESTINO Project (2009). In anderen Ländern werden durchaus erhebliche Größenordnungen (für die USA z. B. 8 Millionen Personen) erwartet. Aufgrund der besonderen rechtlichen Situation dieser Personen muss von einer besonders hohen Viktimisierungswahrscheinlichkeit ausgegangen werden. Entsprechend werden Opferbefragungen, die diesen Personenkreis ausschließen, systematisch zu niedrige Viktimisierungsraten ermitteln. Ein ähnliches Argument gilt für andere klandestine Populationen wie z. B. Straftäterinnen und Straftäter. Bei anderen Populationen als illegalen Migranten werden aber in der Regel nur kleine Anteile an der Gesamtpopulation vermutet, sodass kaum ein Effekt auf Viktimisierungsraten oder andere Populationsparameter der Gesamtbevölkerung erwartet werden kann.

4.4 Populationen mit gesundheitlichen Problemen

Personen mit gesundheitlichen Problemen, beispielsweise Patienten in psychiatrischen Kliniken, pflegebedürftige oder demente Personen, stellen ein weiteres Problem dar (Schnell 1991). Wie bei fast allen Gesundheitsstatistiken ist die Datenlage in Deutschland auch für die Prävalenz von Demenz unbefriedigend. Auf der Basis der vorliegenden Analysen (Ziegler/Doblhammer 2009; Doblhammer u. a. 2012) erscheint die Schätzung eines Anteils von Dementen bei der über 65-jährigen Bevölkerung von 6 bis 7 % als realistisch. Dies entspräche einer Gesamtzahl von 1,17 Millionen Erkrankten. Das Robert-Koch-Institut geht davon aus, dass 60 % der Menschen mit Demenz in Privathaushalten gepflegt werden (Weyerer 2005). Das entspräche 709.000 Personen. Diese Personengruppe dürfte aus nahezu allen Befragungen ausfallen.[10]

4.5 Populationen mit Sprachbarrieren

Wenn nicht besondere Maßnahmen ergriffen werden, dann wird faktisch unabhängig von der Definition der Grundgesamtheit immer nur die deutschsprachige Bevölkerung befragt. Man muss sich in Erinnerung rufen, dass bei der Befragung einer „deutschsprachigen“ Bevölkerung letztlich immer das Datenerhebungspersonal über die Zugehörigkeit zur Grundgesamtheit entscheidet; entsprechende Interviewereffekte sind trivialerweise erwartbar.

---

[10] Ein nicht unbeträchtlicher Teil davon dürfte körperlicher Gewalt ausgesetzt sein. Zu den Problemen der Schätzung des Anteils demenziell Erkrankter, die Opfer körperlicher Gewalt werden, siehe Weissenberger-Leduc/Weiberg (2011, 35–38).

Dieser unpräzise Ausschluss der „nicht deutschsprachigen Bevölkerung“ ist für Viktimisierungsstudien kaum zu ignorieren. Legt man die Daten der RAM-Studie 2006/2007 der fünf größten Ausländerpopulationen (Personen aus der Türkei, Ex-Jugoslawien, Italien, Polen, Griechenland; zusammen ca. 57 % der ausländischen Personen in Deutschland) zugrunde, dann ergibt sich ein Anteil von ca. 10 % dieser Personen, die nicht gut genug Deutsch sprechen können, um sich problemlos im Alltag verständigen zu können (Haug 2008). Nimmt man an, dass dieser Anteil auch für andere Ausländergruppen in Deutschland gilt, dann ergäben sich ca. 670.000 Personen, mit denen eine Verständigung auf Deutsch Probleme bereiten würde. Berücksichtigt man nur die über 14-Jährigen, dann handelt es sich grob um mehr als eine halbe Million Personen (ca. 0,7 % der Bevölkerung über 14 Jahren), die allein aufgrund ihrer Deutschkenntnisse ausgeschlossen würden.

4.6 Konsequenzen des Ausschlusses spezieller Populationen

Der Ausschluss zahlreicher und sehr spezieller Populationen bleibt nicht ohne Folgen für die Schätzung von Viktimisierungsraten. In nahezu allen ausgeschlossenen Subgruppen kann von erhöhten Viktimisierungsraten ausgegangen werden, entsprechend führt der Ausschluss dieser Subgruppen zu niedrigeren Viktimisierungsraten (hierzu insbesondere Lynn 1997).[11] Wie fast immer bei Befragungen werden die Ergebnisse sozialpolitisch umso erfreulicher, je schlechter die Erhebungen durchgeführt werden. Daher ist es nicht verwunderlich, dass z. B. der Ausschluss der Bevölkerung in Institutionen aus Erhebungen zumeist undiskutiert bleibt.[12]

Man kann argumentieren, dass Viktimisierungsstudien nicht versuchen, eine unverzerrte Schätzung der Population zu erreichen, sondern lediglich einen Indikator für den Zustand einer Population liefern sollen. Dann muss man aber konstante Verzerrungen unterstellen, wenn Vergleiche über die Zeit oder verschiedene Studien oder mehrere Länder erfolgen sollen. Dies ist ohne exakte Kontrolle der Art und der Größe der ausgeschlossenen Population nicht möglich. Daher muss der Dokumentation der Art und Größe der ausgeschlossenen Populationen gerade bei Viktimisierungsstudien erhöhte Aufmerksamkeit gewidmet werden. Dies bedeutet, dass den Eigenschaften der

---

11 Dies wurde schon für den British Crime Survey (BCS) kritisiert (Smith 2006, 10).

12 Eine der wenigen Ausnahmen findet sich im Hinblick auf Viktimisierungsstudien in den USA in einer neueren Veröffentlichung (National Research Council 2014, 124): „The frame for the ancillary listing of group quarters, which is an important part of the secondary sample for the National Crime Victimization Survey because their residents may be at higher risk for sexual violence, is seriously flawed in terms of both the building and enumeration of this secondary frame.“

für die jeweilige Studie verwendeten Erhebungsinstrumente, den Interviewern und den Auswahlgrundlagen mehr Aufmerksamkeit in Hinsicht auf ausgeschlossene Populationen gewidmet werden muss als z. B. bei einer Wahlabsichtsbefragung.

5 Formen von Auswahlverfahren

Schlüsse von einer Stichprobe auf eine Grundgesamtheit lassen sich ohne schwerlich zu rechtfertigende sonstige Annahmen nur dann mathematisch begründen, wenn die Stichprobe durch einen Zufallsprozess gezogen wird. Dabei ist es von entscheidender Bedeutung, dass die Wahrscheinlichkeit für die Auswahl jedes einzelnen Elements der Grundgesamtheit berechnet werden kann. Es ist dabei nicht wesentlich, ob die Wahrscheinlichkeiten gleich sind oder nicht.[13] Die Auswahlwahrscheinlichkeiten müssen aber berechenbar und größer als Null sein.

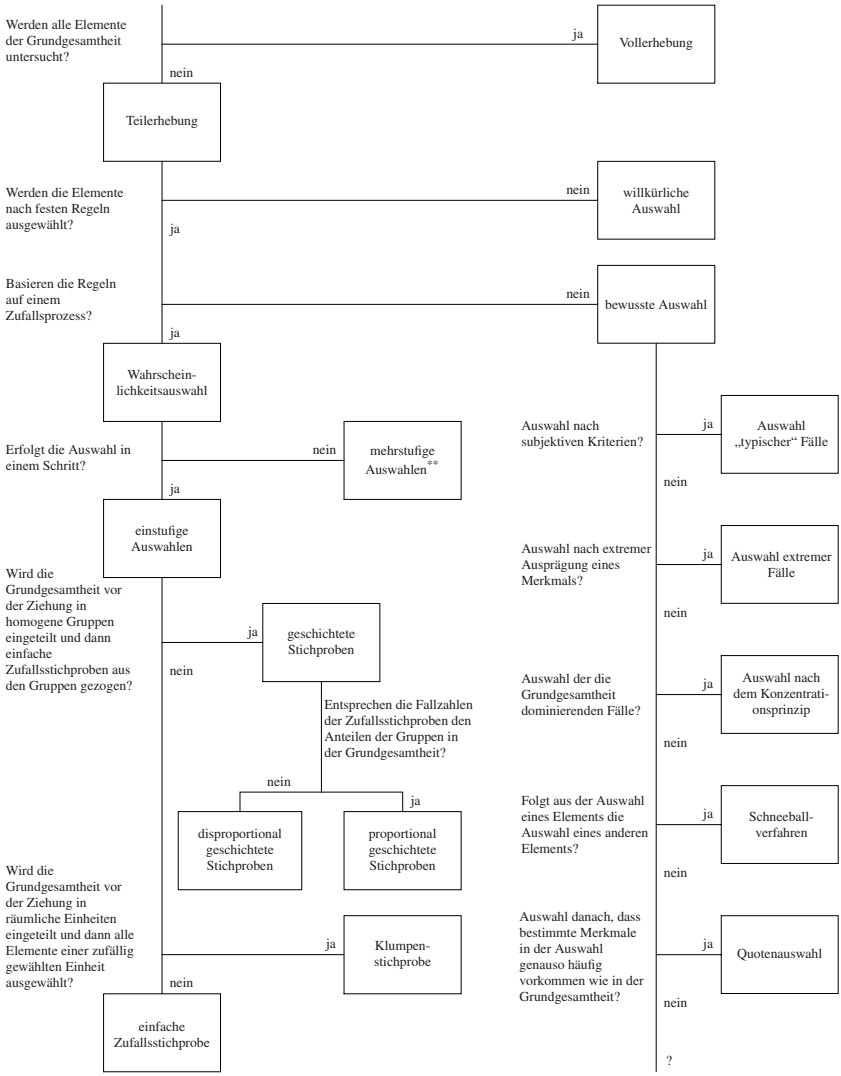
5.1 Willkürliche und bewusste Auswahlverfahren

Bei allen sogenannten willkürlichen oder bewussten Auswahlen sind die Auswahlwahrscheinlichkeiten nicht berechenbar und/oder nicht größer als null (*Abbildung 2*, rechte Seite der Abbildung).

[13] Falls ungleiche Auswahlwahrscheinlichkeiten vorliegen, kann dies durch eine Gewichtung berücksichtigt werden.

Abbildung 2:

Übersicht über Auswahlverfahren (Schnell u. a. 2013, 260)



** Mehrstufige Auswahlen bestehen aus Kombinationen einstufiger Verfahren mit unterschiedlichen Auswahleinheiten.

Willkürliche Auswahlen sind nicht regelgeleitet, sondern man greift auf etwas Verfügbares zurück (sogenannte *convenience samples*). Solche Stichproben sind prinzipiell nicht verallgemeinerbar. Im Gegensatz dazu liegt bewussten Auswahlen zwar eine Regel zugrunde, die Auswahlwahrscheinlichkeiten sind aber trotzdem nicht berechenbar. Daher sind auch bewusste Auswahlen nicht verallgemeinerbar. Dazu gehören z. B. die Auswahl extremer Fälle oder „typischer Fälle" (häufig irreführend von Laien als *theoretical sampling* bezeichnet). Dazu gehört auch die Auswahl nach dem Konzentrationsprinzip, also z. B. die Auswahl der zehn größten Schulen aus einem Bezirk. Ebenso nicht verallgemeinerbar ist das sogenannte „Schneeballverfahren": Hat man wie auch immer ein Mitglied einer seltenen Subgruppe gefunden, ist es plausibel, dass dieses Mitglied andere Mitglieder der Subgruppe kennt: Man befragt dann ausgehend von der Indexperson weitere benannte Personen.[14]

Schließlich gehört zu dieser mathematisch nicht zu rechtfertigenden Klasse von Verfahren auch das Quotenverfahren. Dabei werden Interviewern Quoten bestimmter Personenmerkmale oder Kombinationen dieser Merkmale vorgegeben, z. B. fünf Männer im Alter von 30 bis 39 Jahren und fünf Frauen im Alter von 30 bis 39 Jahren, jeweils katholisch und berufstätig. Innerhalb dieser Kombination kann aber der Interviewer die Befragten beliebig auswählen (also: willkürlich). Damit sind die Auswahlwahrscheinlichkeiten nicht berechenbar und das Verfahren mathematisch nicht zu rechtfertigen.[15]

5.2 Auswahlverfahren mit berechenbaren Auswahlwahrscheinlichkeiten

Verfahren mit berechenbaren Auswahlwahrscheinlichkeiten lassen sich mit einem einfachen Modell erläutern: Jedes Element der Grundgesamtheit bekommt eine Losnummer zugeteilt, die Lose werden in einer Trommel gemischt und dann nacheinander gezogen. In diesem Fall handelt es sich um ein einstufiges Auswahlverfahren einer einfachen Zufallsstichprobe ohne Zurücklegen. Heute werden die Personen auf einer Liste einfach nummeriert und

---

14 Eine Variante dieses Verfahrens wird zurzeit als *Respondent Driven Sampling* auch außerhalb der Statistik bekannter. Für das Verfahren müssen einige schwierig zu prüfende Annahmen getroffen werden (z. B. darf die Wahrscheinlichkeit einer Selektion von Merkmalen des letzten Selektionsschritts abhängen, aber nicht mehr von vorherigen Selektionsschritten), daher ist es prinzipiell unklar, wann das Verfahren angewandt werden kann oder nicht (einführend: Schonlau u. a. 2014). Im Zweifelsfall sind auch hier die Auswahlwahrscheinlichkeiten unbestimmbar.

15 In der Politik und in einigen Bereichen der Marktforschung ist das Quotenverfahren aufgrund seiner schnellen und wenig aufwendigen Durchführung immer noch verbreitet, in wissenschaftlichen Anwendungen spielt es kaum noch eine Rolle. Einzelheiten finden sich bei Schnell u. a. 2013.

dann durch Zufallszahlen aus einem Zufallszahlengenerator gezogen. Dies ist zum Beispiel das häufigste Verfahren, wenn für eine Gemeinde aus dem Einwohnermelderegister gezogen wird.[16]

Manchmal verfügt man nicht über eine Liste aller Personen, sondern nur über eine Liste von Ansammlungen („Klumpen“ oder „Cluster“) von Personen. Das klassische Beispiel sind Schulen: Man hat keine Liste der Schülerschaft, aber eine Liste von Schulen. Jede Schülerin bzw. jeder Schüler gehört zu genau einer Schule. In diesem Fall würde man die Klumpen zufällig ziehen und – im einfachsten Fall – die gesamte Schülerschaft in einem Klumpen auswählen. Dies wäre eine Klumpenstichprobe. Ein anderes Beispiel für dieses Auswahlverfahren sind Flächenstichproben: Kann jede Person genau einer Fläche zugeordnet werden, dann kann man Flächen (wie z. B. Häuser) als Klumpen einer Klumpenstichprobe verwenden. Das Problem von Klumpenstichproben besteht darin, dass Personen innerhalb eines Klumpens einander ähnlicher sind als zufällig ausgewählte Personen. Dieses Problem wird als „Klumpeneffekt“ bezeichnet, der letztlich die Präzision der Schätzungen verringert und bei den Analysen berücksichtigt werden muss (Abschnitt 8). Die praktische Konsequenz bei Klumpenstichproben besteht darin, möglichst viele Klumpen und möglichst wenig Personen pro Klumpen zu ziehen.

Häufig gibt es Einteilungen der Grundgesamtheit, deren Vergleich von besonderem Interesse ist. Diese Einteilungen werden als Schichten bezeichnet. Ist die Größe der Schichten in der Grundgesamtheit bekannt, kann die Präzision von Stichprobenschätzungen durch Berücksichtigung dieser Schichtung verbessert werden.[17] Möchte man eine Stichprobe aus Deutschland ziehen und Vergleiche zwischen Bundesländern anstellen, dann wird häufig nach Bundesländern geschichtet. Das bedeutet nichts anderes, als dass pro Bundesland eine unabhängige Stichprobe gezogen wird. Die gesamte Stichprobe ist dann eine geschichtete Stichprobe aus der gesamten Bundesrepublik.

Bei geschichteten Stichproben wird danach unterschieden, ob die Größe der Schichten ihrem Anteil in der Grundgesamtheit entspricht oder nicht. Sind die Anteile der Schichten in der Stichprobe genauso groß wie in der Grund-

---

16 In der Praxis findet man häufig eine Annäherung an dieses Verfahren, das als systematische Stichprobe bezeichnet wird. Dabei wird z. B. jede zehnte Person auf der Liste ausgewählt. Das Problem solcher Verfahren besteht darin, dass die Liste selbst systematisch geordnet sein kann und daher systematische Fehler entstehen. Ähnliches gilt für Buchstaben oder Geburtstagsverfahren, bei denen Personen mit bestimmtem Geburtsdatum oder Anfangsbuchstaben von Namen ausgewählt werden. Da echte Zufallsverfahren heute keine zusätzlichen Kosten verursachen, sollte im Regelfall immer eine echte Zufallsstichprobe gezogen werden, um jede unerwünschte Systematik der Auswahl zu verhindern.

17 Diese Verbesserung ist dann möglich, wenn die Schichten unterschiedliche Kennwerte und Varianzen aufweisen. Dies ist bei den üblichen Schichtungen nach Bundesländern und Ortsgrößen meistens der Fall.

gesamtheit, dann handelt es sich um eine proportionale Schichtung, andernfalls um eine disproportionale Schichtung.[18] Seit der Wende wird bei Bevölkerungsstichproben häufig disproportional nach alten und neuen Bundesländern geschichtet: Neue Bundesländer werden in höherem Ausmaß berücksichtigt als es ihrem Bevölkerungsanteil entspricht.

Die meisten bundesweiten Bevölkerungsstichproben (oder *national samples* in anderen Ländern) verwenden Kombinationen der bisher erläuterten Auswahlverfahren, in der Regel also geschichtete Stichproben von Klumpen, aus denen dann einfache Zufallsstichproben gezogen werden. Solche Kombinationen sind technisch mehrstufige Auswahlen, die häufig als „komplexe Stichproben" bezeichnet werden. Wichtig dabei ist, dass die Auswahlwahrscheinlichkeit auf jeder einzelnen Stufe berechenbar ist: Wird z. B. auf der letzten Stufe nicht berechenbar ausgewählt, dann ist die Stichprobe nicht „komplex", sondern willkürlich und damit unbrauchbar. Einzelheiten komplexer Stichproben hängen vom Erhebungsmodus der Befragung ab und werden in Kapitel 9 diskutiert.

6 Das Total-Survey-Error-Modell

Einführende oder rein mathematische Lehrbücher konzentrieren sich bei der Diskussion um Auswahlverfahren häufig ausschließlich auf Stichprobenschwankungen, also die Varianz der Schätzungen bei wiederholten Ziehungen aus einer stabilen Grundgesamtheit. Die Wurzel aus dieser Varianz ist der Standardfehler, das üblicherweise verwendete Maß für die Präzision einer Schätzung. In einführenden Lehrbüchern wird dies in der Regel weiter vereinfacht auf die ausschließliche Schätzung des Standardfehlers einfacher Zufallsstichproben.

Das Resultat solcher Vereinfachungen sind dann z. B. irreführende Aussagen wie:

> Von 1.250 Befragten entscheiden sich auf die Frage, wen sie am nächsten Sonntag wählen wollen, 40 Prozent für eine Partei. Die Fehlertoleranz liegt hier bei

[18] Eine disproportionale Schichtung kann zum Beispiel verwendet werden, um die Schätzungen in jeder Schicht *h* mit einer ausreichenden Präzision durchführen zu können. Falls die resultierenden Stichprobengrößen n_h innerhalb kleiner Schichten einer proportional geschichteten Stichprobe nicht ausreichen, um die interessierenden Parameter mit der gewünschten Präzision zu schätzen, dann könnte eine disproportional geschichtete Stichprobe zur Lösung dieses Problems verwendet werden. In solch einer disproportionalen Stichprobe sind dann die kleinen Schichten überproportional vertreten, die großen Schichten hingegen unterproportional, sodass die gewünschte Präzision in allen Schichten erreicht werden kann (Kalton 1983, 24 f.). Für spezielle Allokationskriterien für disproportionale Schichtungen wie die varianzoptimale Allokation nach Neyman siehe Lohr 2010.

> rund +/– 3 Prozentpunkten. Das heißt, der Anteil dieser Partei bei allen Wahlberechtigten liegt zwischen 37 und 43 Prozent.[19]

Das ist natürlich in mehrfacher Hinsicht falsch; die tatsächlichen Fehlertoleranzen sind sehr viel größer (Schnell und Noack 2014). Das Problem besteht nicht darin, dass die verwendeten Formeln falsch sind oder falsch gerechnet wurde: Die Voraussetzungen für die Anwendungen der einfachen Modelle sind aber nicht gegeben. Im Beispiel des ZDF-Politbarometers dürften die tatsächlichen „Fehlertoleranzen" nicht bei +/– 3 Prozent, sondern bei +/– 9 Prozent liegen.[20]

Um eines deutlich zu machen: Die Tatsache, dass die tatsächlichen „Fehlertoleranzen" komplexer Stichproben in der Forschungspraxis sehr viel größer sind, als die naiven Berechnungen aus Einführungslehrbüchern glauben lassen, ist in der Statistik vollkommen unumstritten. Die Probleme der korrekten Berechnung basieren zum einem darauf, dass man für eine korrekte Berechnung sehr viel mehr über die Erhebung wissen muss, als man üblicherweise einer Pressemitteilung entnehmen kann (z. B. Intraklassenkorrelationen, Klumpengrößen, Schichtung, Gewichtungsfaktoren etc.)[21]. Wir werden dies in Kapitel 8 genauer darstellen. Zum anderen sind die resultierenden Intervalle bei den in der Praxis üblichen kleinen Stichproben so groß, dass man den Nutzen einer solchen Erhebung kaum begründen kann: Wer würde das ZDF-Politbarometer schauen, wenn die Fehlergrenzen korrekt angegeben würden? Dann müsste das Ergebnis lauten: „Die CDU würde zwischen 31 und 49 % der Stimmen erhalten". Aus diesen beiden Gründen ist die Berechnung korrekter Intervalle in der Praxis – vor allem in der Meinungsforschung – nicht weit verbreitet.

Der Unterschied zwischen der naiven Schätzung und den korrekten empirischen Ergebnissen lässt sich am folgenden Beispiel zweier Viktimisierungsstudien zeigen. Schnell und Kreuter (2000) verglichen die Ergebnisse zweier im Auftrag des Bundesministeriums für Justiz 1997 durchgeführter Studien (eine Mehrthemenbefragung, eine Befragung im Rahmen des Sozialwissenschaften-Bus III/97), die nahezu zeitgleich vom selben Institut mit den gleichen Fragen in Deutschland erhoben wurden. Das Ergebnis illustriert *Abbildung 3*.

---

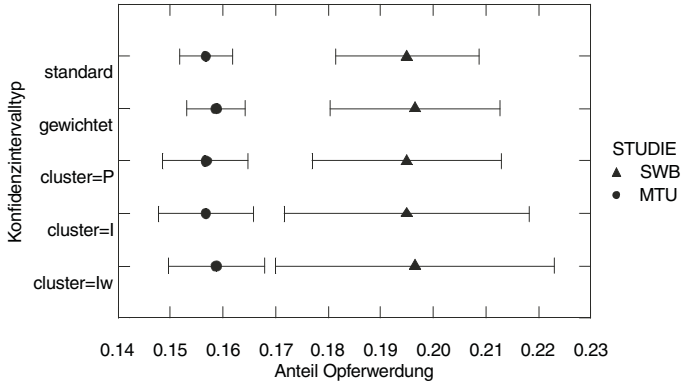
19 Homepage des ZDF-Politbarometers am 04. 11. 2013.

20 Berechnet aus allen Wahlprognosen für die CDU im Datensatz bei Schnell und Noack (2014). Die CDU erzielte bei den Prognosen im Mittel 40,3 %, die mittlere Stichprobengröße liegt bei 1.540 Fällen. Für diese etwas günstigeren Ausgangsdaten wurden die Konfidenzintervalle so vergrößert, dass 95 % der Konfidenzintervalle die Ergebnisse für die CDU bei der höchstens einen Monat nach der Prognose stattfindenden Bundestagswahl enthalten. Dies ist erst dann der Fall, wenn die Konfidenzintervalle +/– 9 % umfassen.

21 Der Intraklassenkorrelationskoeffizient stellt ein Maß für die Homogenität der Cluster dar.

Abbildung 3:

Konfidenzintervalle des Anteils der Opferwerdung zweier nahezu zeitgleicher deutschlandweiter Erhebungen desselben Instituts mit identischen Fragen (Schnell/Kreuter 2000, 102)



Offensichtlich sind die Ergebnisse unvereinbar: Die Konfidenzintervalle überlappen sich nicht, unabhängig davon ob man ein naives Standardkonfidenzintervall („standard“) berechnet, zusätzlich die Gewichtung in das Modell („gewichtet“) aufnimmt, den Erhebungsort („cluster = P“) oder den Interviewer („cluster = I“) als Klumpeneffekt einschließt oder alle Probleme simultan berücksichtigt (Gewichtung und Interviewer als Klumpen, „cluster = Iw“). Solche signifikanten Unterschiede bei einer unabhängigen Replikation bei einer unveränderten Grundgesamtheit sollten selten (in weniger als 0,6 % der Studien) auftreten. Die vermutliche Ursache für solche Ergebnisse liegt darin, das bei einem Schätzergebnis einer Befragung nicht nur die Unsicherheit durch die Stichprobe (*Sampling Error*) berücksichtigt werden muss, sondern auch alle anderen Fehlerquellen.[22]

In der Statistik werden diese anderen Fehlerquellen zusammenfassend als *non-sampling errors* bezeichnet.[23] Die Größe dieser Fehler kann die der Standardfehler deutlich übersteigen.

---

22 Ein Vergleich der pro Land und Erhebungsjahr geschätzten Anteile von Personen mit „Furcht vor Kriminalität“ für die europaweiten Surveys „ESS“, „ICVS“ und „Eurobarometer“ ergab ebenfalls teilweise unvereinbare Ergebnisse. Analog werden auch hier *non-sampling errors* als vermutliche Ursache dafür angesehen (Noack 2015, 94–123).

23 Eine Übersicht über die Literatur zu nahezu allen dieser Fehlerquellen gibt Weisberg 2005.

In der Diskussion um die Qualität eines Surveys spielt daher in der wissenschaftlichen Literatur zunehmend ein erweitertes statistisches Fehlermodell eine zentrale Rolle. Dieses Fehlermodell wird als Total-Survey-Error-Modell bezeichnet.[24] Definiert man den Fehler der Schätzung einer Statistik $\hat{\mu}$ eines Parameters μ für einen Survey als

$$Fehler = \hat{\mu} - \mu, \tag{1}$$

dann ist das in der Surveystatistik übliche Gütemaß für die Schätzung der sogenannte *mean-squared error* (MSE).[25] Der MSE ist eine Kombination des Ausmaßes der Abweichung der Schätzungen vom Populationswert (Bias) und des Ausmaßes der Streuung der Schätzungen vom Populationswert (Varianz der Schätzungen):

$$MSE(\hat{\mu}) = B^2 + Var(\hat{\mu}), \tag{2}$$

wobei $B = E(\hat{\mu} - \mu)$ den Bias, E den Erwartungswert und $Var(\mu)$ die Varianz der Schätzungen darstellt. Beim Design und der Durchführung eines Surveys versucht man den MSE für die interessierenden Schätzungen zu minimieren. Üblicherweise führt man in der Gleichung für die Schätzung des MSE die Quellen des Bias eines Surveys einzeln auf:

$$MSE = \left(B_{spez} + B_{nr} + B_{cover} + B_{mess} + B_{da}\right)^2 + Var_{sampling} + Var_{mess} + Var_{da}, \tag{3}$$

wobei

B_{spez} = Spezifikationsfehler
B_{nr} = Nonresponsebias
B_{cover} = Coveragebias
B_{mess} = Messfehler
B_{da} = Datenaufbereitungsbias
$Var_{sampling}$ = Varianz der Kennwerteverteilung
Var_{mess} = Messfehlervarianz
Var_{da} = Datenaufbereitungsvarianz
ist (Biemer/Lyberg 2003, 59).

24 Dieser Abschnitt wurde Schnell 2012 entnommen.
25 Die Darstellung folgt hier Biemer 2010.

Mit Ausnahme von Spezifikationsfehler, Messfehler und Datenaufbereitungsfehler werden alle Fehlerquellen in diesem Kapitel behandelt. Die hier nicht behandelten Fehlerquellen betreffen Operationalisierungs- und Messfehler (durch Interviewer, Befragte, Erhebungsinstrument und Erhebungsmodus) sowie Datenaufbereitungs- und Datenanalysefehler – diesbezüglich muss auf die jeweilige Spezialliteratur verwiesen werden.[26]

Im Prinzip ist die Schätzung aller einzelnen Bestandteile des MSE zumindest mit vereinfachenden Annahmen möglich, wenngleich auch außerordentlich aufwendig. Das Modell des Total-Survey-Errors wird daher fast immer nur als regulative Idee verwendet. Bisher wurde das Modell für Erhebungen in Deutschland kaum thematisiert. Einen empirischen Versuch am Beispiel der Ungenauigkeiten der Wahlprognosen in Deutschland findet man bei Schnell und Noack (2014).

7 Standardfehler und Konfidenzintervalle: Ermittlung der benötigten Stichprobengröße bei einfachen Zufallsstichproben und vereinfachten Annahmen

Es gibt keine absolute Mindestgröße einer Zufallsstichprobe. Die notwendige Größe einer Stichprobe hängt nahezu ausschließlich davon ab, mit welcher Genauigkeit man eine Aussage treffen möchte. Die Genauigkeit einer Schätzung auf der Basis einer Stichprobe wird durch die Größe der Konfidenzintervalle ausgedrückt.

Die Breite eines Konfidenzintervalls (selten auch Vertrauensintervall genannt) wird zunächst durch den Standardfehler bestimmt. Der Standardfehler ist dabei nicht zu verwechseln mit der Standardabweichung, also der Streuung der Messungen um ihren Mittelwert. Im Gegensatz dazu ist der Standardfehler definiert als die Standardabweichung der Stichprobenkennwertverteilung, also der Verteilung der geschätzten Stichprobenstatistiken (z. B.

[26] Spezifikationsfehler sind Unterschiede zwischen den tatsächlich gemessenen Variablen und dem eigentlichen Messziel, wobei es sich nicht um Messfehler, sondern um Probleme einer für das Ziel des Surveys unangemessenen Operationalisierung handelt. Zu den Datenaufbereitungsfehlern gehören Fehler durch die Dateneingabe, die Codierung der Antworten, in der Gewichtung und der Datenanalyse. Für Fehler in diesen Stufen einer Erhebung muss auf die entsprechende Literatur verwiesen werden (z. B. Schnell u. a. 2013, 420–429).

Anteilswerte oder Mittelwerte), die für alle möglichen Stichproben der Größe n aus einer Population der Größe N berechnet werden.[27]

Für den Anteilswert ergibt sich der Standardfehler aus dem Anteilswert und der Stichprobengröße:

$$se(p) = \sqrt{\frac{p(1-p)}{n}} \tag{4}$$

Für die Präzision der Schätzung spielt die Größe der Population (N) keine Rolle, sondern lediglich die Größe der Stichprobe (n) und die Größe des Anteilswerts (p). Es ist demnach also für die Präzision der Schätzung unerheblich, ob beispielsweise die Viktimisierungsrate für ein bestimmtes Delikt für die Bundesrepublik Deutschland oder eine einzelne Großstadt geschätzt werden soll.[28] Diese Tatsache scheint für Laien schwer akzeptabel zu sein: Eine Stichprobe für eine Großstadt darf nicht kleiner sein als eine Stichprobe für ein gesamtes Land. Diese unangenehme Konsequenz ist mathematisch ebenso unzweifelhaft wie der Politik nur schwierig zu vermitteln.

Die Breite eines Konfidenzintervalls wird weiterhin durch die Festlegung der Irrtumswahrscheinlichkeit bestimmt. Wird diese zu groß gewählt (z. B. 50 %), ist das Konfidenzintervall zwar sehr schmal, wird aber den Populationsparameter für die Hälfte der realisierten Konfidenzintervalle nicht enthalten. Wird die Irrtumswahrscheinlichkeit hingegen zu klein gewählt (0,001 %), so wird zwar nahezu jedes realisierte Konfidenzintervall den Populationsparameter enthalten, die Konfidenzintervalle werden aber wegen ihrer großen Breite faktisch unbrauchbar. Aus diesem Grund wird die Irrtumswahrscheinlichkeit üblicherweise auf 5 % festgelegt. Dieser Wert gibt die Wahrscheinlichkeit an, einen Alpha-Fehler (auch „Fehler erster Art“) zu begehen (siehe hierzu Fahrmeir u. a. 2007, 415–417). In der Regel werden also 95-%-Konfidenzintervalle verwendet.[29]

---

27 In Veröffentlichungen außerhalb der Statistik ist die Verwendung von Groß- und Kleinbuchstaben häufig irreführend (dies wird durch die Verwendung von Textprogrammen wie Word, welche die Großschreibung von Einzelbuchstaben häufig fälschlich erzwingen, befördert). Eindeutig und korrekt hingegen ist die Verwendung von Großbuchstaben für Kennzahlen der Grundgesamtheit und von Kleinbuchstaben für Kennzahlen einer Stichprobe.

28 Dies gilt mit einer unbedeutenden Einschränkung: Ist die Population sehr klein oder die Stichprobe sehr groß ($n/N \geq 0.05$), werden die Ergebnisse *präziser* als in Formel (4) angegeben. In diesen Fällen verkleinert sich das Konfidenzintervall (für den Mittelwert und bei einfachen Zufallsstichproben) dann um $\sqrt{1 - n/N}$. Einzelheiten zu dieser „finiten Populationskorrektur“ (fpc) können in mathematischen Lehrbüchern zur Stichprobentheorie (z. B. bei Lohr 2010) nachgelesen werden.

29 Da die Summe aus Konfidenzniveau (Irrtumswahrscheinlichkeit) und Signifikanzniveau 100 % beträgt, ergibt ein Signifikanzniveau von 5 % somit ein Konfidenzniveau von 100 % – 5 % = 95 %.

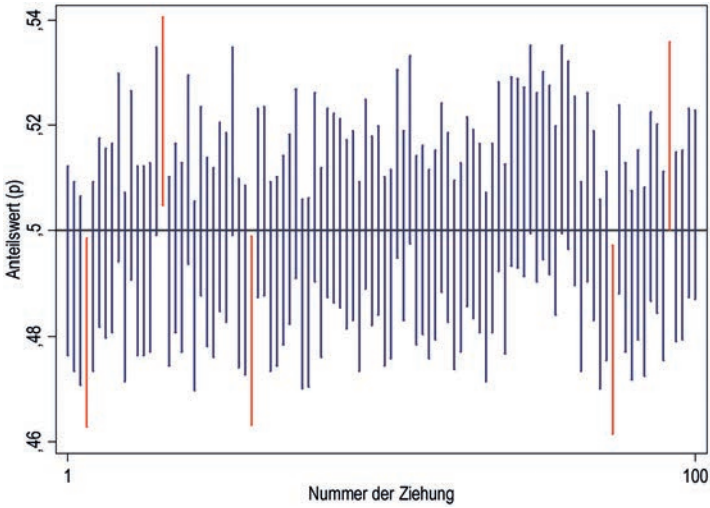
Zur Verdeutlichung: Die Formel zur Schätzung eines 95-%-Konfidenzintervalls des Anteilswerts ist über

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}\right] \qquad (5)$$

gegeben. Werden nun 100 Stichproben der Größe n = 3.000 aus einer Population der Größe N = 100.000 gezogen, werden die meisten der 100 berechneten Konfidenzintervalle (ca. 95) den Populationsparameter enthalten, einige wenige (ca. 5) aber auch nicht (*Abbildung 4*).[30]

Abbildung 4:

95-%-Konfidenzintervalle für Anteilswerte aus 100 verschiedenen einfachen Zufallsstichproben (n = 3.000) aus der gleichen Grundgesamtheit mit π = 0,05. Fünf Konfidenzintervalle enthalten (zufällig) den Populationsmittelwert π = 0,05 nicht. In der Abbildung sind dies die zwei Intervalle mit den höchsten und die drei Intervalle mit den niedrigsten Intervallgrenzen.

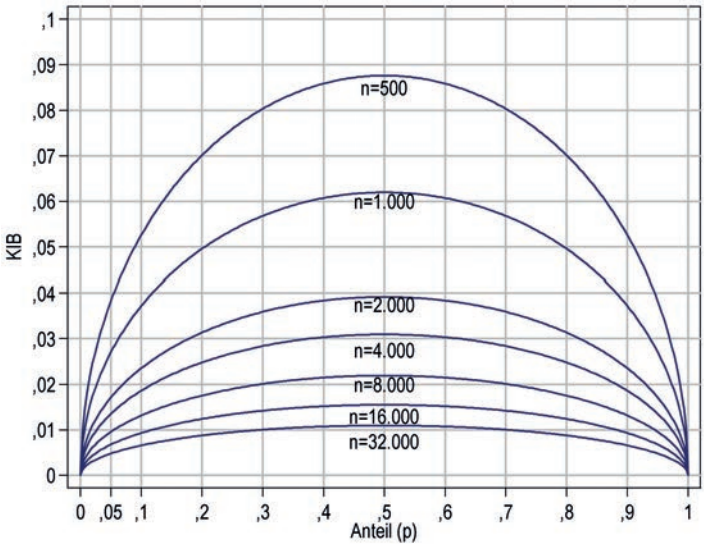


[30] In diesem Zusammenhang ist es sinnlos, davon zu sprechen, dass ein realisiertes Konfidenzintervall den Populationsparameter mit 95 % Wahrscheinlichkeit enthält. Der Populationsparameter liegt entweder innerhalb der Grenzen des Konfidenzintervalls oder nicht.

Die Breite der Konfidenzintervalle ist dabei davon unabhängig, ob die Population 10.000, 100.000 oder 80.000.000 Elemente umfasst, sie ist nicht von *N*, sondern von der Stichprobengröße *n* abhängig. Der Zusammenhang zwischen Konfidenzintervallbreite und Größe der Stichprobe kann in einem Nomogramm dargestellt werden (*Abbildung 5*).

Abbildung 5:

Breite des Konfidenzintervalls (KIB) für gegebene Anteilswerte (p) und verschiedene Stichprobengrößen *n* (in Anlehnung an Schnell/Hoffmeyer-Zlotnik 2002)



Die Breite eines Konfidenzintervalls für Anteilswerte ist maximal für $p = 0{,}50$. Für eine Stichprobe mit $n = 1.000$ Fällen ergibt sich demnach eine Breite des 95-%-Konfidenzintervalls für Anteilswerte von über 6 %. Um die Breite auf 1 % zu reduzieren, wäre bereits eine Stichprobengröße von über 38.000 Fällen notwendig.[31]

Um einen Anteilswert mit einer bestimmten Genauigkeit schätzen zu können, werden also zwei Dinge benötigt: erstens eine möglichst genaue Vorstellung über die Größe des Anteilswerts sowie zweitens die gewünschte Irrtumswahr-

[31] Generell gilt die Faustregel, dass eine Halbierung der Breite eines Konfidenzintervalls eine Vervierfachung der Stichprobengröße erfordert.

scheinlichkeit. Wenn hierfür Zahlen festgelegt wurden, kann die Breite des Konfidenzintervalls über

$$KIB = 2 * z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{6}$$

geschätzt werden, wobei $z_{\alpha/2}$ den Z-Wert der inversen Standardnormalverteilung für die gegebene Irrtumswahrscheinlichkeit α (für gewöhnlich 5 %) darstellt (Bortz 2005, 104, Formel 3.24). Stellt man diese Formel um, lässt sich der benötigte Stichprobenumfang mit

$$n = \frac{4 * z_{\alpha/2}^2 p(1-p)}{KIB^2} \tag{7}$$

schätzen, wobei *KIB* hier die gewünschte Breite des Konfidenzintervalls (z. B. 0,01 für eine Breite von 1 % oder 0,05 für eine Breite von 5 %) bezeichnet (Bortz 2005, 104, Formel 3.26).

8 Designeffekte

Den bisherigen Überlegungen liegt die Annahme zugrunde, dass es sich um einfache Zufallsstichproben handelt. Dies ist jedoch faktisch für keinen bundesweiten Survey der Fall, mit dem Aussagen über die allgemeine Bevölkerung getroffen werden sollen. Das Design solcher Surveys umfasst üblicherweise die Schichtung, Klumpung oder Auswahl der Populationselemente in mehreren Stufen.

Werden beispielsweise natürlich vorkommende räumliche Einheiten als Klumpen für die Stichprobenziehung verwendet, resultieren aufgrund des sogenannten Klumpeneffekts im Normalfall weniger präzise Ergebnisse, als bei Verwendung einer einfachen Stichprobe gleicher Größe (Kish 1965, 164). Die Ursache liegt darin, dass Personen mit einem ähnlichen soziodemografischen Hintergrund dazu tendieren, in der gleichen Nachbarschaft zu leben. Dies führt zu einer größeren klumpeninternen Homogenität als bei rein zufälligem Siedlungsverhalten.

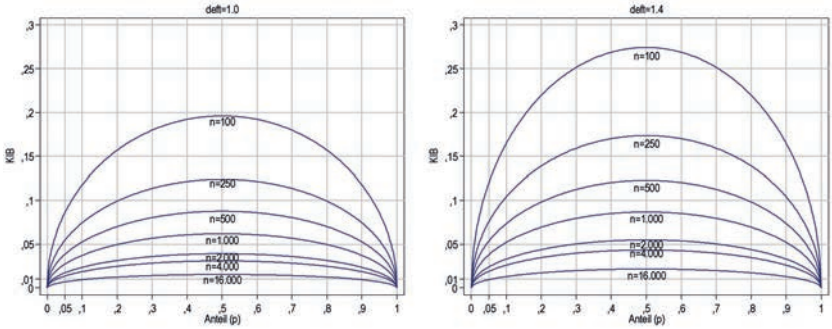
Als Beispiele können das Familieneinkommen (Converse/Traugott 1986, 1095) oder die Fragen nach vorhandenen „Incivilities“ in der jeweiligen Nachbarschaft (Schnell/Kreuter 2005, 401) angeführt werden. Generell bezeichnet man die Vergrößerung der Konfidenzintervalle bei komplexen Stichproben durch Klumpung, Schichtung und Gewichtung sowie einige andere Faktoren als Designeffekt.

Liegen solche Designeffekte vor, dann ist die Berechnung von Konfidenzintervallen über Formel (5) sowie die naive Berechnung statistischer Tests nicht mehr korrekt. In nahezu allen Fällen führen Designeffekte zu größeren Standardfehlern und damit auch zu konservativeren statistischen Tests, also weniger fälschlich signifikanten Ergebnissen.[32]

Die Auswirkungen solcher Designeffekte werden in den Abbildungen 6 und 7 dargestellt.[33] *Abbildung 6* gibt die Breite der 95-%-Konfidenzintervalle für Anteilswerte in Abhängigkeit von der Stichprobengröße sowie des Anteilswerts p an. Hier zeigt sich, dass die Konfidenzintervalle umso breiter werden, je kleiner die Stichprobe ausfällt und je näher der Anteilswert an $p = 0{,}5$ liegt. Ergänzend ist deutlich zu erkennen, dass sich die Breite der jeweiligen Konfidenzintervalle deutlich erhöht, wenn nicht von einer einfachen Zufallsstichprobe ausgegangen wird (linke Abbildung), sondern ein komplexes Stichprobendesign mit einem Designeffekt von $deft = \sqrt{deff} = 1{,}4$ angenommen wird (rechte Abbildung).[34]

Abbildung 6:

Breite des Konfidenzintervalls (KIB) für gegebene Anteilswerte (p) und verschiedene Stichprobengrößen n bei unterschiedlichen Designeffekten (deft=1.0 und deft=1.4)



32 Aus diesem Grund sind die naiven Varianzschätzungen wie z. B. in Ahlborn u. a. 1993 nicht vertretbar. Dort wird argumentiert, dass der Klumpeneffekt durch Schichtung kompensiert werden könne. Dies ist mathematisch zwar denkbar, dürfte aber bei kaum einer Anwendung möglich sein. Bei kriminologischen Fragestellungen kann eine solche Kompensation für die allgemeine Bevölkerung nahezu ausgeschlossen werden. Die Formeln und Ergebnisse bei Ahlborn u. a. 1993 sollten daher nicht für die Planung von Viktimisierungsstudien verwendet werden.

33 Die Beispiele sind an Schnell/Hoffmeyer-Zlotnik 2002 angelehnt.

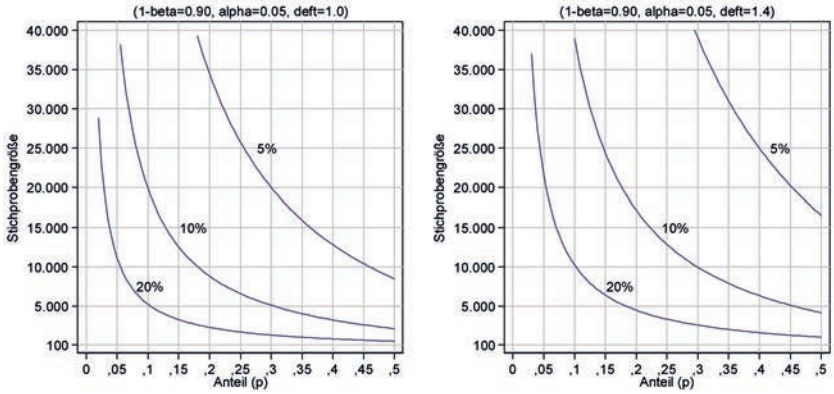
34 Der Wert $deft = 1{,}4$ stellt in der Survey-Literatur die gängige „rule of thumb“ dar (Schnell/Kreuter 2005, 390). Zur Berechnung von $deff$ siehe die Formeln (8) und (10).

Wird demnach für ein Merkmal mit einer Prävalenzrate von 5 % in einem Survey mit einem Designeffekt $deft$ = 1,4 ein 95-%-Konfidenzintervall berechnet, so liegt die tatsächliche Breite dieses Konfidenzintervalls bei einer Stichprobengröße von n = 2.000 nicht bei 1,91 % (linke Abbildung), sondern bei 2,67 % (rechte Abbildung). Um unter diesen Bedingungen ein 1-%-Punkt breites Konfidenzintervall zu erhalten, ist eine Stichprobengröße von n = 14.307 erforderlich. Bei einer einfachen Zufallsstichprobe ohne Designeffekt wären hingegen bereits n = 7.300 Fälle ausreichend.

Ist nicht die einmalige Schätzung einer Prävalenzrate, sondern deren Veränderung von Interesse, kann die zu einer Entdeckung dieser Differenz notwendige Stichprobengröße in *Abbildung 7* abgelesen werden. *Abbildung 7* zeigt die benötigte Fallzahl in Abhängigkeit vom Anteilswert, um eine relative Veränderung des Anteilswerts um x % mit einer Wahrscheinlichkeit von $1 - \beta = 0{,}9$ auch zu entdecken.[35]

Abbildung 7:

Benötigte Fallzahl in Abhängigkeit des Anteilswerts und der relativen Veränderung in Prozent (alpha = 0,05; 1-beta = 0,9; deft = 1,0/1,4; Schnell/ Hoffmeyer-Zlotnik 2002)



35 Die Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt in einer Stichprobe auch zu entdecken, wird in der Statistik als *Power* bezeichnet und ist als $1 - \beta$ definiert, wobei β die Wahrscheinlichkeit für einen Fehler der zweiten Art ist. Ein Fehler der zweiten Art ist die Beibehaltung der Nullhypothese, obwohl sie falsch ist. Die Berechnung der *Power* ist relativ aufwendig, da Stichprobengröße, Effektstärke und Irrtumswahrscheinlichkeit bekannt sein müssen. In der Praxis wird in der Regel eine *Power* von über 0,9 angestrebt. Wie man vor allem anhand der *Abbildung 7* sieht, erfordert eine *Power* von 0,9 in der Regel deutlich größere Stichproben als gemeinhin unter Laien vermutet.

Um die Veränderung eines Anteilswerts von ρ = 5 % um 20 % (also von 5 % auf 5 % × 1,20 auf 6 %) mit einer Wahrscheinlichkeit von 90 % (1-beta) entdecken zu können, ist bei einem Designeffekt von deft = 1,0 eine Stichprobengröße von n= 11.120 erforderlich. Liegt ein Designeffekt von 1,4 vor, so erhöht sich die notwendige Stichprobengröße auf 21.684.[36] Sollen kleinere Veränderungen entdeckt werden, so ist dies nur durch eine deutliche Vergrößerung der Stichprobe zu erreichen.

Bislang haben wir keine Möglichkeit vorgestellt, die Größe des Designeffekts zu berechnen. Exakt ist ein Designeffekt ($deff$) definiert als Quotient des Standardfehlers $\widehat{\sigma^2}_{\theta,SRS}$ einer einfachen Zufallsstichprobe (SRS) und des Standardfehlers der gegebenen komplexen Stichprobe $\widehat{\sigma^2}_{\theta,Komplex}$, wobei θ für einen beliebigen Parameter (z. B. μ oder π) steht. Der Designeffekt $deff$ ist also gleich

$$deff = \frac{\widehat{\sigma^2}_{\theta,Komplex}}{\widehat{\sigma^2}_{\theta,SRS}} \tag{8}$$

Werte größer als 1 zeigen eine geringere Präzision des komplexen Designs im Vergleich zu einer einfachen Zufallsstichprobe gleicher Größe. Häufig ist es einfacher, mit der Wurzel aus dem Designeffekt $deff$ zu rechnen, die als $deft$ bezeichnet wird.

Dies kann am Beispiel der angesprochenen Konfidenzintervalle verdeutlicht werden. Die korrekt berechneten Konfidenzintervalle für komplexe Designs verbreitern sich im Vergleich zu den naiv berechneten Konfidenzintervallen (siehe Formel (5)) um den Faktor $\sqrt{deff}$.

$$\left[\bar{x} - z_{1-\alpha/2}\sqrt{deff}\,\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\sqrt{deff}\,\frac{s}{\sqrt{n}}\right] \tag{9}$$

Die Schätzung des Standardfehlers einer gegebenen komplexen Stichprobe $\widehat{\sigma^2}_{\theta,Komplex}$ kann auf verschiedene Arten erfolgen.[37] Die einfachste Art der Schätzung basiert auf dem sogenannten Intraklassenkorrelationskoeffizienten (ρ, auch ICC). ρ gibt die Homogenität des untersuchten Merkmals innerhalb der verwendeten Klumpen an. Je ähnlicher sich die Elemente innerhalb der

36 Siehe hierzu Schnell/Hoffmeyer-Zlotnik 2002, 10-13.

37 Die korrekte Schätzung über die verbreitetsten Ansätze wie Resampling-Verfahren (Jackknife oder Bootstrap) oder Taylor-Linearisation ist beispielsweise in Stata über die „svy"-Kommandos oder in R über das Paket „survey" möglich. Für Details zu diesen und weiteren Verfahren siehe Wolter 2007.

Klumpen sind, desto größer fällt ρ aus.[38] Neben ρ wird für die Berechnung des Designeffekts ebenfalls die durchschnittlichen Anzahl der Interviews innerhalb der gezogenen Klumpen $\bar{b}$ benötigt. Sind beide Größen bekannt, kann $deff$ über

$$deff = 1 + \rho(\bar{b} - 1) \tag{10}$$

berechnet werden (Kish 1965, 162; Lohr 2010, 174). Diese Größe des Designeffekts hängt also nicht nur von der Homogenität der Klumpen, sondern auch ihrer Größe ab. Die Verwendung großer Klumpen kann also ebenfalls zu einem deutlichen Präzisionsverlust führen.

Da nicht nur räumliche Einheiten als Klumpen angesehen werden können, sondern auch die in einer Studie beteiligten Interviewer, ist auch die Zahl der zu bearbeitenden Fälle pro Interviewer (*Workload*) für den Designeffekt von Interesse. Dies betrifft insbesondere CATI-Studien (*Computer Assisted Telephone Interview*), bei denen hohe Interviewer-Workloads in der Praxis häufig vorkommen.

Daher kann sich auch dann ein großer Wert für $deff$ ergeben, wenn die Klumpen intern zwar heterogen sind ($\rho \approx 0$), die durchschnittliche Anzahl der Interviews pro Interviewer aber ausreichend groß ist, um die geringen Werte für ρ auszugleichen (Schnell/Kreuter 2005, 394). Für CATI-Studien stellen $\rho = 0{,}001$ und $\bar{b} = 70$ übliche Werte dar (Tucker 1983; Groves/Magilavy 1986; Groves 1989). Für diese Werte ergibt sich ein Designeffekt von $deff = 1 + 0{,}01 * (70 - 1) = 1{,}69$ (Schnell/Kreuter 2005, 394). Die effektive Stichprobengröße

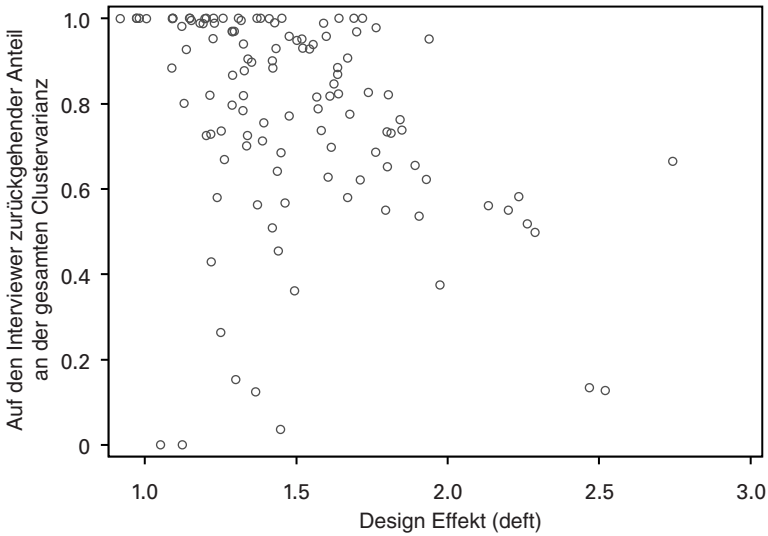
$$n' = \frac{n}{deff} \tag{11}$$

für diese Werte ergibt, dass eine komplexe Stichprobe der Größe $n = 1.000$ mit einem Designeffekt von 1,69 zu einer effektiven Stichprobengröße von $n' = \frac{1.000}{1{,}69} = 591{,}7$ führt. Das bedeutet, dass die Stichprobe, obwohl 1.000 Personen befragt wurden, effektiv so ungenau ist, als wären lediglich 591 Personen befragt worden.

[38] Der Intraklassenkorrelationskoeffizient kann beispielsweise über Varianzanalysen oder Mehrebenenmodelle berechnet werden. Weitere Details finden sich bei Lohr 2010, 174.

Abbildung 8:

Anteil des Interviewers am Designeffekt und die Größe des Designeffekts für 118 Fragen aus dem Defect-Projekt (Schnell/Kreuter 2005, 402)



Überdies muss beachtet werden, dass Designeffekte in von Interviewern durchgeführten Studien mit räumlicher Klumpung sowohl auf die räumlichen als auch die interviewerbedingten homogenisierenden Effekte zurückgehen. In diesem Zusammenhang konnten Schnell und Kreuter (2005) unter Nutzung eines zur Trennung dieser beiden Effekte notwendigen speziellen Stichprobendesigns (interpenetrierende Stichproben) zeigen, dass der Designeffekt für viele Merkmale stärker auf die Interviewer zurückgeht als auf die räumliche Klumpung (*Abbildung 8*). Das Verhalten der Interviewer wirkt sich also auf die tatsächliche Größe der Konfidenzintervalle aus. Diese Tatsache wird in der Forschungspraxis und in der mathematischen Statistik häufig ignoriert. Die Konsequenz ist eine Überschätzung der Genauigkeit statistischer Schätzungen auf der Basis realisierter Stichproben.

9 Mehrstufige Auswahlverfahren für verschiedene Erhebungsmodi

In den meisten Projekten der empirischen Sozialforschung erfolgt die Wahl des Erhebungsmodus nahezu ausschließlich anhand der erwarteten Kosten,

sodass telefonische oder postalische Erhebungen die Regel sind. Stehen die Mittel für die erheblich teureren persönlichen (Face-to-Face-)Befragungen zur Verfügung, dann kann dieser Erhebungsmodus dann eingesetzt werden, wenn telefonische oder postalische Erhebung aus methodischer Sicht ausscheiden (z. B. falls Dokumente herangezogen werden sollen oder die Befragten weder telefonisch noch schriftlich erreicht oder befragt werden können). Für diese Erhebungsmodi stehen geeignete Auswahlverfahren zur Verfügung, die wir im Folgenden detaillierter darstellen werden. Für Web-Befragungen gilt das nicht. Weder existieren geeignete Listen, aus denen ausgewählt werden kann, noch können Auswahlwahrscheinlichkeiten berechnet werden. Entsprechend besteht keine Möglichkeit, auf der Basis eines Web-Surveys auf eine „allgemeine Bevölkerung" zu schließen. Daher raten wir auf absehbare Zeit von der Verwendung von Web-Befragungen für Bevölkerungserhebungen ab.[39]

9.1 Telefonische Befragungen

Die Stichprobenziehung für telefonische Befragungen (CATI) ist nicht trivial. Üblicherweise muss die Ziehung in mehreren Stufen erfolgen, da im Allgemeinen keine vollständige Liste aller Telefonanschlüsse vorliegt. Aus diesem Grund muss mit einer Technik gearbeitet, die keine vollständigen Listen voraussetzt.

9.1.1 Random Digit Dialing

Die Grundidee der Random-Digit-Dialing-Technik (RDD) lässt sich am einfachsten am Beispiel der USA erläutern: US-amerikanische Telefonnummern sind zehnstellig, wobei die ersten drei Ziffern einer Region und die nächsten drei Ziffern einer Vermittlungsstelle entsprechen. Die letzten vier Ziffern bilden zusammen mit den ersten sechs Nummern die Teilnehmernummer. Da zwar keine vollständigen Listen von Telefonnummern, aber vollständige Listen „aktiver" Region-Vermittlungsstelle-Kombinationen (die ersten sechs Stellen) vorliegen, können diese zur Stichprobenziehung verwendet werden. Die einfachste Variante von RDD sieht dann vor, eine Zufallsstichprobe aus der Liste der Region-Vermittlungsstelle-Kombinationen zu ziehen und die Teilnehmernummern durch das Anhängen einer vierstelligen Zufallszahl zu

[39] Einzelheiten finden sich bei Schnell 2012.

generieren. Dieser Ansatz ist allerdings recht ineffizient, da hier viele Nummern generiert werden, die keinem Privathaushalt zugeordnet sind. Diese nicht zielführenden Nummern machen ca. 80 % der generierten Nummern aus (Schnell u. a. 2013, 280–281).

Eine effizientere, zweistufige Methode besteht in der 1970 von Mitofsky vorgeschlagenen und 1978 von Waksberg weiterentwickelten sogenannten Mitofsky-Waksberg-Methode. Diese besteht in einem ersten Schritt aus der Einteilung der letzten vier Ziffern in 100er-Blöcke (andere Einteilungen sind ebenfalls möglich), zum Beispiel 678-560-0000 bis 678-560-0099 (Link/Fahimi 2008). Handelt es sich bei der ersten antelefonierten Nummer aus einem der zufällig ausgewählten 100er-Blöcke (z. B. 678-560-0054) um einen Privatanschluss, so werden in einem zweiten Schritt weitere zufällig gewählte Nummern innerhalb des Blocks antelefoniert, ansonsten scheidet der gesamte Block aus (Schnell u. a. 2013, 281).

Aufgrund der technischen Details bei der Vergabe der Telefonnummern durch die Telekom ist eine Stichprobenziehung über das RDD-Verfahren in Deutschland in dieser Weise nicht möglich. Alternativ finden deshalb Telefonbücher oder Telefon-CDs als Auswahlgrundlage Verwendung. Als erste Auswahlstufe werden hier die Ortsnetze der Telekom verwendet, als zweite Auswahlstufe dann eine Zufallsstichprobe aus den in den Telefon-CDs eingetragenen Nummern gezogen. In einem dritten Schritt wird die zu befragende Person ausgewählt. Dies kann über eine vereinfachte Version einer als „Schwedenschlüssel“ bekannten Zufallsauswahl oder über die Frage, welche Person als letzte Geburtstag hatte (*last birthday method*), geschehen (Schnell u. a. 2013, 281–282).

Ein überaus gewichtiger Punkt ist bei dieser Vorgehensweise die Auswahl der Telefonnummern im zweiten Auswahlschritt. Beschränkt sich die Auswahl auf die in den Telefon-CDs gelisteten Nummern, so führt dies zum Verlust aller nicht eingetragenen Telefonnummern. Neben Personen ohne Festnetzanschluss (kein Telefon oder nur Mobiltelefon, hierzu Abschnitt 9.1.2) betrifft dies insbesondere Personen ohne eingetragene Nummer. Dieses Vorgehen ist nicht vertretbar. Somit empfiehlt sich ein Verfahren, in dem auch die nicht in den Telefonbüchern oder Telefon-CDs vorhandenen Nummern berücksichtigt werden. Dies geschieht durch über bestimmte Verfahren generierte zusätzliche Telefonnummern.

Eine Möglichkeit besteht in der Addition einer Zufallszahl zu einer existierenden, zufällig aus der Telefon-CD ausgewählten Telefonnummer (*randomized last digit*, RLD), eine weitere darin, die letzten beiden Ziffern einer exis-

tierenden Nummer durch zufällig generierte Nummern zu ersetzen (Schnell u. a. 2013, 281–282).[40]

Sowohl der sinkende Anteil in den Telefonbüchern erfasster Festnetznummern als auch die steigende Anzahl von Personen, die nur noch über ein Mobiltelefon verfügen (*Cell-Phone-Only*, CPO),[41] stellen ein Problem für telefonische Befragungen dar. Aus diesem Grund werden in der Bundesrepublik vermehrt andere Auswahlverfahren für telefonische Befragungen (Dual-Frame-Verfahren) als die bisher beschriebenen eingesetzt (Schnell u. a. 2013, 282).

9.1.2 Dual-Frame-Stichproben zur Berücksichtigung von Mobiltelefonen

Die steigenden Teilnehmerzahlen im Bereich des Mobilfunks sowie die Auflösung des Telekom-Monopols hatten erhebliche Auswirkungen auf die Stichprobenziehung für telefonische Befragungen. Um die daraus resultierenden Probleme zu lösen, werden in Deutschland seit 2007 die erst seit wenigen Jahren verfügbaren Daten der Bundesnetzagentur über an die Telekommunikationsanbieter vergebene Nummern zur Stichprobenziehung verwendet.[42] Der aus diesen Daten gebildete Nummernraum umfasst alle überhaupt nutzbaren Telefonnummern, nicht nur die tatsächlich verwendeten Nummern sowohl für Festnetz als auch Mobilfunk.

---

40 Dieses auf die letzten vier Ziffern angewendete Verfahren stellte von 1999 bis 2007 die Basis der vom „Arbeitskreis deutscher Marktforschungsinstitute“ (ADM) verwendeten Telefonstichproben dar. Bei dieser Variante werden allerdings viele Nummern generiert, die entweder nicht vergeben („kein Anschluss unter dieser Nummer“) oder keine Privatanschlüsse sind. Obwohl solche Varianten des Mitofsky-Waksberg-Designs bereits 1988 von Lepkowski ausführlich diskutiert wurden, wird dieses Verfahren im deutschsprachigen Raum häufig irreführend nach den ersten Anwendern dieses Verfahrens in der BRD als Gabler-Häder-Design bezeichnet (Schnell u. a. 2013: 282). Seit 2007 basieren die ADM-Telefonstichproben auf den von der Bundesnetzagentur vergebenen Rufnummernblöcken (Heckel u. a. 2014, 142–144) und sind damit nicht mehr als modifiziertes Mitofsky-Waksberg-Design zu verstehen.

41 Nach Berechnungen von Hunsicker/Schroth (2014, 9) mit Daten der Forschungsgruppe Wahlen und des Politbarometers hat sich der Anteil Wahlberechtigter, die nur noch über ein Mobiltelefon verfügen, von 8 % im Jahr 2006 auf 14 % in den Jahren 2012/2013 nahezu verdoppelt. Weitere dort angeführte vergleichbare Dual-Frame-Studien weisen für unterschiedliche Grundgesamtheiten Werte zwischen 11 % und 12 % aus (INFAS 12 %, ADM 12,4 %, Cella 2 11 %). Weitere Hinweise auf eine Zunahme der CPO-Problematik sowie deutliche Unterschiede zwischen den Ländern in Europa für 2006 und 2010 auf Basis verschiedener „Special Eurobarometer: E-Communications Household Surveys“ geben Heckel/Wiese 2012, 111. Aufgrund der spärlichen Informationen hinsichtlich der Datenerhebung und der somit unklaren methodischen Güte dieser Surveys sind diese Aussagen allerdings nicht als endgültig gesichert anzusehen.

42 Für Details hinsichtlich des Vorgehens des ADM siehe Glemser u. a. (2014), zu Details der Erstellung der Auswahlgrundlage siehe Heckel u. a. (2014).

Damit liegt es nahe, jeweils eine Auswahlgrundlage für das Festnetz und eine Auswahlgrundlage für die Mobilfunknetze zu konstruieren und diese dann zu kombinieren. Solche Kombinationen zweier Auswahlgrundlagen werden in der Stichprobentheorie allgemein als *Dual-Frames* bezeichnet. Da eine Person aber sowohl mehrere Festnetz- als auch Mobilnummern besitzen kann, müssen Dual-Frame-Stichproben erst derart gewichtet werden, dass jede Person die gleiche Auswahlwahrscheinlichkeit besitzt (Schnell u. a. 2013, 282–283). In der Praxis in Deutschland basieren diese Gewichtungen auf einer Reihe plausibler Annahmen, so z. B. dass ein Mobiltelefon exakt einer Person sowie ein Festnetzanschluss jeder Person im Haushalt zugeordnet ist. Weiter wird angenommen, dass die Wahrscheinlichkeit, parallel über den Festnetz- und den Mobiltelefonframe gleichzeitig ausgewählt zu werden, gleich null ist. Unter diesen Annahmen lässt sich die kombinierte Auswahlwahrscheinlichkeit als

$$\pi_i = k_i^F \frac{n^F}{N^F} \frac{1}{z_i} + k_i^C \frac{n^C}{N^C} \tag{12}$$

berechnen (Häder u. a. 2009, 29). Dabei stellt k_i^F die Zahl der Nummern dar, unter der ein Haushalt über das Festnetz erreicht werden kann, k_i^C bezeichnet die Zahl der Nummern, unter der eine Person über Mobiltelefone erreicht werden kann, n ist die Zahl der Telefonnummern in der Stichprobe (F oder C), N die Zahl der Telefonnummern in der Grundgesamtheit (F oder C) und z_i die Zahl der Personen im Haushalt.

Entscheidend für die Validität der derart berechneten Auswahlwahrscheinlichkeit ist die Gültigkeit der Angaben k_i^F, k_i^C und z_i, die von den subjektiven Wahrnehmungen der Befragten abhängen und somit Raum für Fehler lassen. Damit sind diese Werte prinzipiell schon aufgrund der potenziellen Unkenntnis der Befragten als fehleranfällig anzusehen (Schnell 2012, 273). Die daraus resultierenden Probleme sind nicht abschließend geklärt.

9.2 Schriftliche und postalische Befragung

Gelegentlich werden schriftliche und postalische Befragungen miteinander verwechselt, obwohl es sich um klar abgrenzbare Erhebungsformen handelt.

Schriftliche Befragungen erfolgen innerhalb von Organisationen, bei denen vollständige Listen der zu befragenden Personen vorliegen.[43] Dabei handelt

[43] Schriftliche Befragungen ohne solche Listen entsprechen einer Stichprobe ohne bekannte Auswahlregel, sind also wie die entsprechenden Web-Surveys willkürliche oder bewusste Stichproben. Damit sind sie keine geeignete Basis für verallgemeinerbare Aussagen.

es sich um Organisationen mit festen Mitgliedern (Universitäten, Verwaltungen, Schulen). Werden die Personen innerhalb der Organisation befragt (z. B. im Klassenraum, bei Vorlesungen oder gemeinsamen Veranstaltungen), dann handelt es sich um eine individuelle schriftliche Befragung in einer Gruppe. Das Standardbeispiel wäre eine Befragung im Klassenverband oder während einer Vorlesung oder Personalversammlung. In solchen Fällen handelt es sich um Klumpenstichproben, d. h., die Klumpen werden zufällig ausgewählt und innerhalb der Klumpen werden alle schriftlich befragt. Das Auswahlverfahren ist vergleichsweise trivial durchzuführen. Die zu beachtende Regel wurde schon erwähnt: Möglichst viele Klumpen bei gleicher Fallzahl sind besser als wenige große Klumpen. Allerdings müssen Nonresponse-Probleme bei der Auswahl der Klumpen (Beispiel: Alle Mitglieder einer Vorlesung fallen aus, falls der Dozent nicht kooperiert) und der einzelnen Mitglieder bedacht werden. Weiterhin müssen die Klumpeneffekte bei der Analyse beachtet werden: Eine Schülerbefragung von 2.000 einzelnen Personen hat einen kleineren Standardfehler als eine Befragung von 50 Klassen mit je 40 Personen. Schließlich muss beachtet werden, dass das Antwortverhalten in Gruppen besonderen Dynamiken unterliegen kann: In vielen Fällen ist es kaum möglich, eine Zusammenarbeit mehrerer Personen in einer Gruppe zu unterbinden. Bei Studierenden ist dies z. B. in Vorlesungen nur unter Klausurbedingungen möglich, andere Gruppen sind nicht unbedingt disziplinierter. Sollte als Erhebungsform die Befragung in Gruppen gewählt werden, sollte dies daher entsprechend dokumentiert und die Zugehörigkeit jeder Person zu einer Gruppe Bestandteil des Datensatzes werden, da ansonsten keine korrekte Berechnung der Standardfehler möglich ist.

Bei postalischen Befragungen liegt fast immer eine Liste der zu befragenden Personen vor.[44] Bei einer bundesweiten Bevölkerungsbefragung würde man in der Regel zunächst eine Schichtung nach Bundesländern vornehmen. Innerhalb der Schichten sollte man eine möglichst hohe Zahl von Gemeinden (in der Regel z. B. 160, 210 oder 240)[45] proportional zu ihrer Einwohnerzahl ziehen (*Probability Proportional to Size*, PPS). In jeder gezogenen Gemeinde würde man eine konstante (kleine) Zahl (in der Regel weniger als 20) Personen aus der Einwohnermeldedatei ziehen. Praktisch steht eine solche bundesweite Ziehung vor dem Problem, dass die Gemeinden nicht kooperieren müssen und es zu entsprechenden Ausfällen auf Gemeindeebene kommt.

---

44 Versuche, bundesweite postalische Befragungen auf Haushaltsebene ohne Einwohnermeldedateien durchzuführen, sind selten. Ein entsprechendes Experiment war Bestandteil des Defect-Projekts (Schnell/Kreuter 2000).

45 Diese Zahlen gehen auf die Entwicklung der Stichproben für die Musterstichprobenpläne des Arbeitskreises Deutscher Markt- und Sozialforschungsunternehmen (ADM) zurück. Zur Begründung der Zahlen und einer Geschichte der Entwicklung diese Zahlen siehe Schnell 1997, 58–59.

Weiterhin variieren die Gebühren für solche Ziehungen in so erheblichem Umfang, dass man gelegentlich schon aus finanziellen Gründen auf einzelne Gemeinden verzichtet. Schließlich ist zu beachten, dass die Dauer der Ziehung in mehr als 160 Gemeinden mehr als ein halbes Jahr dauern kann: Dann sind im Mittel mehr als 5 % der Befragten bereits wieder verzogen. Postalische Befragungen mit Einwohnermeldedateien sind bundesweit also keineswegs unproblematisch.

9.3 Face-to-Face

Da in der Bundesrepublik kein vollständiges Zentralregister für die allgemeine Bevölkerung existiert, ist auch keine Zufallsstichprobe realisierbar, die solch eine Liste als *Sampling-Frame* benötigt. Demnach müssen andere Ansätze zur Konstruktion von Stichproben verwendet werden (Schnell 2012, 204–205).

9.3.1 Random Walks

Für bundesweite Erhebungen hat sich in Deutschland seit Ende der 70er Jahre ein Stichprobenplan als Standard etabliert, der auf die Musterstichprobenpläne des „Arbeitskreises deutscher Markt- und Sozialforschungsinstitute" (ADM) zurückgeht (für die Historie der ADM-Stichproben siehe Löffler u. a. 2014). Im klassischen ADM-Design wurden zwischen 160 und 240 *Sampling-Points* aus einer Datei von ca. 80.000 Bundestagsstimmbezirken proportional zur Zahl der Stimmberechtigten gezogen. In den ausgewählten Stimmbezirken wurde dann ein *Random Walk* durchgeführt, bei dem eine Person ausgehend von einem zufällig gewählten Startpunkt einen Zufallsweg beschreitet. Die Grundregel für einen solchen Zufallsweg könnte z. B. lauten: Auf der linken Straßenseite gehen, rechts abbiegen wann immer es möglich ist. Im Detail werden die Regelwerke aufwendig (was ist eine Straße, was passiert in Sackgassen etc.), aber im Prinzip könnte ein Zufallsweg entstehen.[46] Die begehende Person listet auf ihrem Zufallsweg dann z. B. jeden dritten Haushalt (in der Regel: Klingeln). Die Liste dieser Haushalte (Name und Adresse) bildet dann die Auswahlgrundlage einer Haushaltsstichprobe.[47] In der Praxis wurden die Begehung und die Befragung häufig derselben Person übertragen (*Standard-Random*), was zu vielen Implementierungsproblemen

---

46 Details finden sich bei Schnell (2012, 206–207).

47 In den Haushalten wurde dann noch eine Person zumindest näherungsweise zufällig ausgewählt, in der Regel mit einer speziellen Zufallszahlentabelle („Schwedenschlüssel") (siehe hierzu Schnell u. a. 2013, 276).

geführt hat. Durch die Verfügbarkeit digitaler Karten und elektronischer Gebäudedateien ist in modernen Gesellschaften ein *Random Walk* weitgehend obsolet geworden. *Random Walks* sollten für Befragungen nur noch dann verwendet werden, wenn keine digitalen Karten und Gebäudedateien verfügbar sind, z. B. in Entwicklungsländern, Katastrophen- oder Kriegsgebieten.

9.3.2 Einwohnermeldeamtsstichproben

Als Goldstandard gilt für Deutschland seit Mitte der 90er Jahre die Durchführung von Einwohnermeldestichproben.

Zu diesem Zweck ist die Kooperation der Einwohnermeldeämter unverzichtbar. Erschwert wird die Stichprobeziehung aus den Einwohnermeldeämtern dadurch, dass weder der Zugang zu den Daten bundeseinheitlich geregelt ist, noch die Gebührensätze über alle Gemeinden einheitlich sind.[48] Eher im Gegenteil schwanken die Gebührensätze in einem erstaunlichen Ausmaß. Die notwendige Kooperation mit mehreren Hundert Gemeinden, die sowohl frei über die Kooperation als auch die Gebührensätze entscheiden können, führt bei bundesweiten Stichproben neben hohen Kosten auch zu einer langen Dauer für die Stichprobenziehung (Schnell 2012, 194).

Doch obwohl die Daten der Einwohnermeldeämter in der Bundesrepublik als bestmögliche Auswahlgrundlage für Stichproben der allgemeinen Bevölkerung gelten, sind auch diese Daten mit Problemen behaftet.[49] So leiden auch die Daten der Einwohnermeldeämter unter Overcoverage und Undercoverage. Ein vom statistischen Bundesamt im Rahmen der Vorbereitung des Zensus 2011 durchgeführter Registertest (Stichtag 5. Dezember 2001) erbrachte bundesweit 1,7 % Personen, die zwar angetroffen, aber nicht in den Registern gefunden wurden (Undercoverage). Der gegenteilige Fall („Karteileichen“, Overcoverage) belief sich bundesweit auf 4,1 % der Einträge. Unter Berücksichtigung der zeitlichen Verzögerung zwischen Bevölkerungsbewegung und der Aktualisierung der Registereinträge („temporäre Karteileichen“) sinkt dieser Wert auf 2,9 %. Allerdings liegen deutliche Unterschiede hinsichtlich der Raten für Undercoverage (zwischen 1,0 % und 3,1 %) und Overcoverage (zwischen 2,6 % und 8,1 %) zwischen den Bundesländern vor. Ebenso deutliche Unterschiede zeigen sich je nach Größe für die einzelnen Gemeinden, wobei größere Gemeinden auch größere Fehlerraten aufweisen (Schnell 2012,

---

48 Einen detaillierteren Überblick über Einwohnermeldestichproben findet sich in von der Heyde 2014.

49 Für die vielen praktischen Probleme von Zufallsstichproben aus Einwohnermelderegistern siehe Albers 1997.

194; siehe auch Tabelle 1 in Statistische Ämter des Bundes und der Länder 2004, 814). Demnach liegen Coverage-Probleme insbesondere in Großstädten vor. Sollten sich die Coverageraten zwischen soziodemografischen Gruppen unterscheiden, werden diese Anteile verzerrt geschätzt. Da die Daten der Zensus-Testerhebung für wissenschaftliche Analysen nicht zur Verfügung stehen, kann dies nicht geprüft werden. Entsprechende Untersuchungen in den USA und dem Vereinigten Königreich zeigen aber, dass es sich bei unterrepräsentierten Personen eher um mobile und sozial randständige Bevölkerungsgruppen handelt (Schnell 2012, 195). Liegen hier systematisch höhere Viktimisierungsraten vor, so würde die tatsächlich vorliegende Kriminalitätsbelastung über eine Einwohnermeldestichprobe unterschätzt.

9.3.3 Gebäudestichproben

Seit 2006 existiert eine neue amtliche Datenbasis, in der alle Gebäude in Deutschland enthalten sind. Da durch diese Liste jedem Gebäude eine eindeutige Auswahlwahrscheinlichkeit zugeordnet werden kann, ist die Ziehung einer Stichprobe aus der allgemeinen Bevölkerung mit dieser Liste als Auswahlgrundlage möglich.[50] Da Einwohnerzahlen für Aggregate unterhalb der Ebene „Stadt“ in der Praxis aus verschiedenen Gründen schwierig zu erhalten sind, wird dieser Schritt in der vorgeschlagenen Stichprobenziehung umgangen und stattdessen zur Auswahl Wohngebäude verwendet. Die Auswahl von Gebäuden ist sowohl als einfache Zufallsstichprobe als auch als geschichtete und/oder geklumpte Stichprobe möglich. Beispielsweise ist es für eine Face-to-Face-Befragung sinnvoll, Städte als natürlich vorkommende Klumpen im Sampling-Prozess zu verwenden, um die Interviewer-Reisekosten im Vergleich zu einer einfachen Zufallsstichprobe geringzuhalten. Im Hinblick auf die unterschiedliche Anzahl der Wohnungen pro Gebäude in Klein- und Großstädten scheint eine Schichtung bezüglich der Größe der Städte sinnvoll zu sein. Innerhalb dieser Schichten wird die Auswahl einer für alle Städte gleichen Zahl von Gebäuden (*Secondary Sampling Unit*, SSU) aus den Städten (*Primary Sampling Unit*, PSU) per PPS-Sampling empfohlen (Schnell 2008, 7–8).[51]

Für Wohngebäude mit mehreren Wohnungen wird die Auswahl jeweils einer Wohnung empfohlen. Die Zahl der Wohngebäude (ohne Wohnheime) wird in

---

50 Das Design einer bundesweiten Bevölkerungsstichprobe auf Basis der Gebäudedatei geht auf einen Antrag des Erstautors bei der Deutschen Forschungsgemeinschaft (DFG) aus dem Jahr 2007 zurück. Details dieses „G-Plans“ finden sich bei Schnell 2008, 7.

51 Dies entspricht einer PPS-Stichprobe auf Gebäudeebene und erscheint sinnvoller als eine PPS-Stichprobe auf Personenebene, da sich die Gebäudestatistik langsamer ändert, als die Bevölkerung.

Deutschland für das Jahr 2011 auf 19.050.663 Gebäude mit 40.857.381 Wohnungen geschätzt (Statistische Ämter des Bundes und der Länder 2014, 5). Da 82,3 % der Wohngebäude in Deutschland aus höchstens zwei Wohnungen bestehen, ist eine Auswahl der Wohnung innerhalb eines Wohngebäudes mit nur einer Wohnung (65,1 %) nicht notwendig oder in Wohngebäuden mit zwei Wohnungen (17,2 %) durch Münzwurf zu realisieren (diese Zahlen basieren auf den Angaben des Statistischen Bundesamts (Statistische Ämter des Bundes und der Länder 2014, 8)). Für Gebäude mit mehr als zwei Wohnungen (17,7 %) müssen andere Auswahlmechanismen genutzt werden. Schnell (2008, 8) schlägt hierfür vor, die Klingeltafeln per Mobiltelefon zu fotografieren und per MMS an den Supervisor zu senden, der dann die zu kontaktierende Wohnung per einfacher Zufallsziehung ohne Zurücklegen festlegt und diese Information per SMS an den Interviewer übergibt.

Für das Jahr 2012 wird die Anzahl an Privathaushalten vom Statistischen Bundesamt auf 40.656.000 geschätzt. Davon sind 41 % Ein- und 35 % Zweipersonenhaushalte. Damit stellen Haushalte mit drei oder mehr Mitgliedern 24 % aller Haushalte dar. Innerhalb der Haushalte wird die Anzahl aller erwachsenen Haushaltsmitglieder durch die Interviewer erfragt und in ein CAPI-System eingegeben. In Ein-Personen-Haushalten oder Mehrpersonenhaushalten mit nur einer erwachsenen Person ist eine Auswahl nicht notwendig. Für Haushalte mit zwei erwachsenen Personen erfolgt die Auswahl über das CAPI-System mit einer Wahrscheinlichkeit von jeweils 50 % für die Kontaktperson oder die andere erwachsene Person im Haushalt. In Mehrpersonenhaushalten erfolgt die zufällige Auswahl aus einer Liste aller erwachsenen Haushaltsmitglieder durch das CAPI-System (Schnell 2008, 9).[52]

9.4 Web-Surveys

Das zentrale Problem aller Web-Surveys ist die Tatsache, dass es keine geeigneten Auswahlgrundlagen für allgemeine Bevölkerungsstichproben gibt. Für wenige hochspezialisierte Populationen sind vollständige, aktuelle und in aktiver Nutzung befindliche E-Mail-Listen verfügbar, wenngleich sehr seltene Ausnahmen. Damit verbleiben nur wenige Optionen für die Herstellung solcher Listen (Einzelheiten finden sich bei Schnell 2012).

Weitverbreitet ist das Ziehen einer Zufallsstichprobe aus einer administrativen Liste oder einer Telefonstichprobe, die dann online befragt wird, soweit

---

52 Erfolgt die Auswahl nicht zufällig über eine bestimmbare Regel, sondern willkürlich über die Interviewer, so ist eine Berechnung der Inklusionswahrscheinlichkeiten nicht möglich. Somit handelt es sich bei einer solchen Stichprobe um keine Zufallsstichprobe (siehe Unterkapitel 5.1).

dies möglich ist. Manchmal wird versucht, den ausgewählten Personen einen Internetzugang zu ermöglichen, falls dies vor der Ziehung nicht der Fall war. In der Regel sind die Ausfälle von Offline-Rekrutierungen zur Online-Befragung erheblich[53] und immer systematisch: Sensible Populationen (z. B. illegale, alte und/oder kranke Personen) fallen spätestens bei der Online-Befragung aus.

Bei Studien zur Art der Nutzung sozialer Medien mag dies irrelevant sein, nicht aber für Gesundheitssurveys oder Viktimisierungsbefragungen. Solche systematischen Ausfälle führen immer zu einer Unterschätzung der Viktimisierung und lassen sich kaum durch Gewichtungen kompensieren.[54]

Der Nachweis, dass eine Internet-basierte Befragung einer allgemeinen Bevölkerung zu vergleichbaren Resultaten wie eine dem Stand der Surveymethodologie entsprechende Zufallsstichprobe führt, steht weltweit noch aus und ist auf lange Zeit angesichts der selektiven Internetnutzung (Schnell 2012) in der allgemeinen Bevölkerung nicht zu erwarten. So lautet eine klare Empfehlung der *American Association for Public Opinion Research* (AAPOR) im Hinblick auf Online-Befragungen: „Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values“ (AAPOR 2010, 758).

Wir raten daher dringend von der Verwendung von Internetsurveys bei Viktimisierungsstudien ab.

10 Zum Begriff der Repräsentativität

Außerhalb der fachlichen Diskussion wird Repräsentativität als ein wichtiges Merkmal für die Beschreibung von Stichproben begriffen. Hiermit soll ausgedrückt werden, dass die Verteilung der in einer Stichprobe vorliegenden Merkmale deren Verteilung in der Grundgesamtheit entspricht. Doch im Gegensatz zur landläufigen Meinung stellt ‚Repräsentativität‘ keinen in der Stichprobentheorie verwendeten Begriff dar.[55]

---

53 Für Deutschland berichten Bandilla u. a. 2009 von 11 % der Befragten des Allbus 2006, die tatsächlich online an einer Befragung teilnahmen.

54 An dieser Stelle muss darauf hingewiesen werden, dass die Beweislast für die Gültigkeit einer Methode bei denen liegt, die eine neue Methode vorstellen, nicht umgekehrt. Ein solcher Nachweis kann auch nicht durch das einmalige Präsentieren eines günstigen Ergebnisses erbracht werden, da bei vielen Versuchen ein Ergebnis immer zufällig korrekt sein könnte.

55 Einzelheiten zur nahezu ausschließlich missbräuchlichen Verwendung des Begriffs ‚Repräsentativität‘ finden sich in einer Artikelserie bei Kruskall (1979a; 1979b; 1979c; 1980).

Die einzige Möglichkeit, die Übereinstimmung der Merkmalsverteilung zwischen Stichprobe und Grundgesamtheit innerhalb berechenbarer Fehlergrenzen sicherstellen zu können, liegt in der Verwendung von Zufallsstichproben. Nur auf diese Weise sind die beiden Begriffe ‚repräsentativ' und ‚Zufallsauswahl' synonym (Schnell u. a. 2013, 296). Die Bezeichnung einer nicht zufällig gezogenen Stichprobe wie z. B. einer Quotenstichprobe als repräsentativ ist demnach bedeutungslos. Auch der Nachweis, dass die Verteilung einiger Merkmale einer Quotenstichprobe oder eines Websurveys der Verteilung der Merkmale in der Grundgesamtheit entspricht, sagt nichts über die Verteilung der restlichen in der Stichprobe vorliegenden Merkmale aus. Beispielsweise sagt die Unverzerrtheit demografischer Merkmale – wie Alter, Geschlecht, Bildungsstand – nichts über die Unverzerrtheit anderer inhaltlich relevanter Merkmale wie Viktimisierungserfahrungen oder Kriminalitätsfurcht.[56]

11 Nonresponse

Ein zentrales Problem der empirischen Sozialforschung liegt im Ausfall für die Befragung vorgesehener Personen. Diese Ausfälle werden als *Nonresponse* bezeichnet.[57] Je nachdem, ob es sich um einen totalen Ausfall der zur Befragung vorgesehenen Person handelt oder die befragte Person nur einige Fragen nicht beantwortet, wird nach Unit- und Item-Nonresponse unterschieden.

Um die statistischen Konsequenzen solcher Ausfälle abschätzen zu können, sind die möglichen Ausfallmechanismen von besonderer Bedeutung. Je nach den Eigenschaften des vermuteten Ausfallmechanismus ist mit verzerrten Schätzungen zu rechnen. Demnach ergeben sich die weiteren Schritte der Datenanalyse daraus, welcher Ausfallmechanismus angenommen wird. In der neueren Literatur werden drei Prozesse unterschieden:

– MCAR: missing completely at random,

– MAR: missing at random,

– MNAR: missing not at random.

---

56 Theoretisch begründet und mit Simulationsbeispielen belegt findet sich dieses Ergebnis erstmals bei Schnell 1993. Ein neuerer empirischer Hinweis unter Verwendung medizinischer Registerdaten mit fast 20.000 Fällen findet sich bei Vercambre/Gilbert 2012.

57 Weder Populationssurveys im Allgemeinen noch Nonresponse im Besonderen scheinen im Rahmen kriminologischer Forschung bisher die notwendige Aufmerksamkeit erfahren zu haben: Die 5.662 Seiten starke „Encyclopedia of Criminology and Criminal Justice" von Bruinsma/Weisburd (2014) enthält neben den Beiträgen von Johnson (2014) zu „Sample Selection Models" und Aebi/Linde (2014) zu „National Victimization Surveys" keine weiteren Beiträge zu diesen Themen.

Im einfachsten Fall liegt MCAR vor. Die Befragten fehlen damit völlig zufällig, das Fehlen ist also durch keine Variable vorhersagbar. Sowohl die Nonrespondenten als auch die Respondenten stellen somit eine Zufallsstichprobe aus allen zur Befragung vorgesehenen Personen dar.[58] Im Vergleich zu einer Stichprobe ohne Nonresponse werden die Schätzungen durch die ausfallbedingte Reduktion der Stichprobengröße lediglich etwas unpräziser ausfallen.[59] MCAR würde also dann vorliegen, wenn Personen rein zufällig ausfallen und dies nicht beispielsweise von Bildungsstand, Geschlecht, Alter oder beruflicher Stellung abhängt.

Falls hingegen der Ausfall durch Merkmale wie Bildungsstand, Geschlecht, Alter oder berufliche Stellung erklärt werden kann, dann liegt MAR vor. Die Schätzungen können dann immer noch unverzerrt durchgeführt werden, wenngleich spezielle Analysemethoden notwendig sind (für Unit-Nonresponse siehe beispielsweise Särndal/Lundström 2005; Bethlehem u. a. 2011 oder Valliant u. a. 2013; für Item-Nonresponse beispielsweise Schnell 1986; Schafer 1997; Little/Rubin 2002 oder Enders 2010).

Sollte aber der Ausfall direkt mit dem Thema der Befragung derart in Zusammenhang stehen, dass das Fehlen einer Beobachtung nur durch die fehlende Information selbst erklärt werden kann, dann liegt MNAR vor. Bei Viktimisierungssurveys wäre der Ausfall der Person z. B. vom Viktimisierungsstatus selbst abhängig, wobei dieser nicht durch andere Variablen vorhergesagt werden kann. In diesem Fall ist eine Korrektur nur über komplexe Modelle möglich (Sample-Selection-Modelle), für die die explizite Modellierung des Ausfallmechanismus erforderlich ist. Sollte die korrekte Modellierung des

---

58 Und damit ebenfalls eine Zufallsstichprobe aus der Grundgesamtheit, sofern es sich bei der Stichprobe um eine Zufallsstichprobe gehandelt hat.

59 Liegt als Ausfallmechanismus nicht MCAR vor, so kann der Ausfall von Untersuchungseinheiten nicht einfach ignoriert werden (siehe auch Unterkapitel 11.4). Aber auch das „Nachziehen“ in einer vor der Durchführung der Untersuchung nicht exakt definierten Stichprobe, um trotzdem eine gewünschte Stichprobengröße und damit Präzision zu erreichen, ist in der Realität nicht geeignet, dieses Problem zu lösen. Ausfälle durch nicht erreichte oder verweigernde Personen können nicht einfach durch leicht erreichbare oder kooperationsbereite Personen ersetzt werden, wenn sich diese Gruppen systematisch von den teilnehmenden Personen unterscheiden. Hieraus resultieren verzerrte Schätzungen (siehe Unterkapitel 11.3). Somit würde das Nonresponse-Problem nicht gelöst, sondern durch das Nachziehen lediglich verdeckt, wodurch das Ausmaß des Nonresponse nicht mehr angegeben werden kann. Die Stichprobe würde damit faktisch zu einer Quotenstichprobe und wäre damit als willkürliche Stichprobe nicht mehr verallgemeinerbar (Schnell u. a. 2013, 307).

unbekannten Ausfallmechanismus jedoch nicht möglich sein, ist auch mit diesen Modellen keine korrekte Analyse erreichbar.[60]

Insgesamt kann festgehalten werden, dass das Nonresponse-Problem praktisch nicht ignoriert werden kann, da auch das Nichtbeachten der Ausfälle faktisch unterstellt, das der unproblematische Ausfallmechanismus MCAR vorliegt. Dies ist in der Praxis kaum gegeben: Im Regelfall dürften die meisten Ausfallmechanismen in der empirischen Sozialforschung MAR sein und bedürfen entsprechender Berücksichtigung im Design der Studie und der Analyse. Sollten sich hingegen Hinweise auf MNAR ergeben, dann ist eine aufwendige statistische Modellbildung erforderlich. Dabei muss aber beachtet werden, dass Annahmen der resultierenden Modelle prinzipiell nicht mit den zur Verfügung stehenden Daten getestet werden können. Für die Forschungspraxis würden wir bei Hinweisen auf MNAR zu zusätzlichen empirischen Studien mit anderen Methoden raten (z. B. Nutzung administrativer Daten, verdeckte Beobachtung usw.), nicht hingegen zur statistischen Modellierung.

Natürlich ist das Ausmaß eines Nonresponse-Problems vom Mechanismus und der Größe des Nonresponse abhängig. Die Diskussion der Größe des Nonresponse ist einfacher als die Diskussion seiner Mechanismen, daher beschränken sich sogar die meisten methodischen Arbeiten allein auf die Größe des Nonresponse. Die Feststellung der Größe des Nonresponse erfolgt meistens mit der Berechnung einer Ausschöpfungsquote.

11.1 Ausschöpfungsquote

Das quantitative Ausmaß des Nonresponse wird meistens über die Ausschöpfungsquote als Gegenteil der Nonresponse-Quote angegeben.[61] Für die Berechnung der Ausschöpfungsquote liefert die *American Association for Pub-*

---

60 Sample-Selection-Modelle werden nach ihrem Erfinder auch als Heckman-Modelle bezeichnet. Sie bestehen aus einer Modell- und einer Selektionsgleichung. Die Modellgleichung lautet $y_{1i}^* = x_{1i}\beta_1 + u_{1i}$, wobei y_{1i}^* das latente Gegenstück zu der beobachtbaren Variable y_{1i} darstellt. Die Selektionsgleichung ist über $y_{2i}^* = x_{2i}\beta_1 + u_{2i}$ gegeben. Wenn $y_{2i}^* > 0$ gilt, ist y_{1i} mit $y_{1i} = y_{1i}^*$ beobachtet, im Fall $y_{2i}^* \leq 0$ liegt keine Beobachtung für y_{1i} vor, also $y_{1i} = 0$ (Puhani 2000, 54). Das Hauptproblem dieser Modelle liegt in der Tatsache, dass sich ihre statistischen Annahmen mit den gegebenen Daten prinzipiell nicht prüfen lassen (Schnell 2012, 178). Simulationsstudien zu diesem Problem lassen die routinemäßige Anwendung dieser Modelle höchst fragwürdig erscheinen (Stolzenberg 1997; Vella 1998).

61 Die Angaben der Nonresponse-Quote sind nicht immer einfach zu bewerten. So werden in Quotenstichproben „Ausfälle“ durch andere Personen mit passenden Quotenmerkmalen ersetzt. Dieses Vorgehen verdeckt das Nonresponse-Problem lediglich, ohne es zu lösen.

lic Opinion Research insgesamt sechs Definitionen (AAPOR 2011, 44–45), deren eindeutigste als *minimum response rate* (RR1) bekannt ist. Sie ist über

$$RR1 = \frac{I}{(I+P)+(R+NC+O)+(UH+UO)} \tag{13}$$

gegeben (AAPOR 2011, 44). Die einzelnen Größen bezeichnen vollständige Interviews (*I*), partielle Interviews (*P*), Verweigerungen und Abbrüche (*R*), nicht Erreichte (*NC*), andere Ausfallgründe (*O*), unbekannt, ob es sich um einen Haushalt handelt oder nicht (*UH*) und andere unbekannte Ursachen (*UO*).

Um den Unterschied zwischen *O* einerseits sowie *UH* und *UO* andererseits zu verdeutlichen, soll hierauf weiter eingegangen werden. *O* liegt vor, wenn eine Zielperson zur Grundgesamtheit gehört, nicht verweigert, es aber trotzdem durch Krankheit oder Sprachprobleme nicht zu einem Interview kommt. Im Gegensatz ist für *UH* bzw. *UO* beispielsweise durch nicht aufgefundene oder nicht bearbeitete Adressen unbekannt, ob in dem Haus überhaupt eine Zielperson existiert (Schnell 2012, 163).

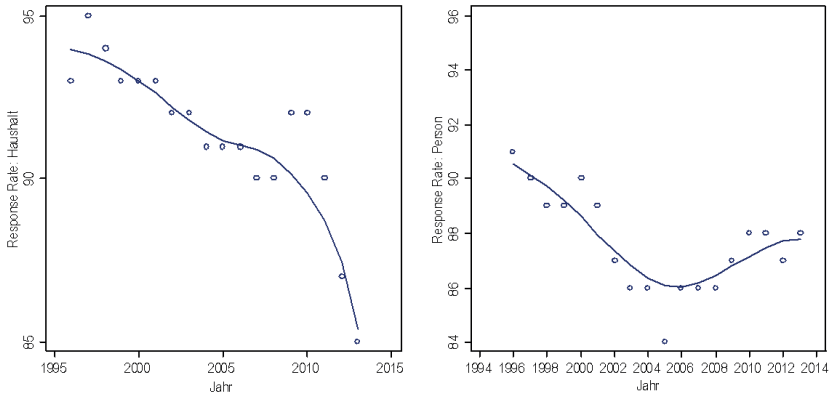
Trotz der starken Bemühungen sind die Ausschöpfungsquoten auch in den aufwendigsten Viktimisierungsstudien in den letzten Jahren zurückgegangen (*Abbildung 9*).[62] Der Effekt der verstärkten Bemühungen gegeben einen Kontakt ist in der rechten Abbildung deutlich zu sehen. Man beachte, dass trotz des Rückgangs beider Ausschöpfungsquoten die Höhe deutlich über den entsprechenden Zahlen für Deutschland liegt. Dies mag zum Teil sicherlich an den erheblich höheren Kosten pro Fall liegen, die für den amerikanischen *National Crime Victim Survey* (NCVS) akzeptiert werden.

---

62 Die der Abbildung zugrunde liegenden Daten wurden den Methodenberichten der jeweiligen Erhebung („National Crime Victimization Survey Technical Documentation") des NCVS auf der Homepage des amerikanischen Justizministeriums (www.bjs.gov/content/pub/pdf/ncvstd13.pdf) entnommen.

Abbildung 9:

Ausschöpfungsrate (*Response Rate*) für Haushalte und Personen im NCVS 1996–2013. Eingezeichnete Linien: Lowess, Glättungsparameter f = 0,8



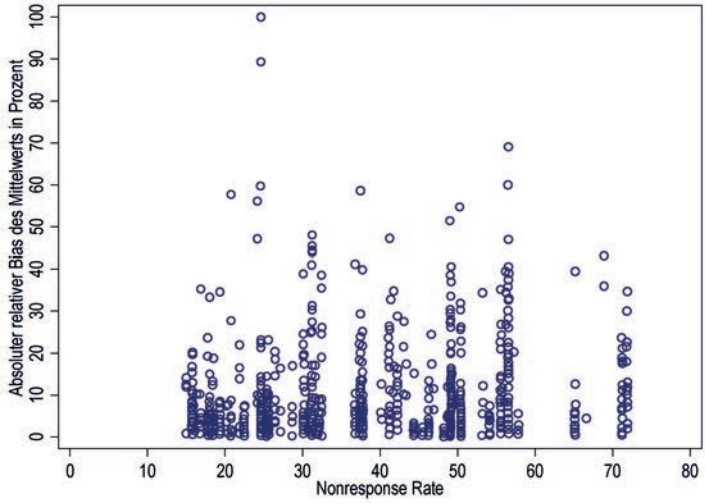
Die bloße Angabe der Ausschöpfungsquote ist für eine angemessene Abschätzung des Nonresponse-Problems allerdings nicht ausreichend. So konnten Groves und Peytcheva (2008) in der von ihnen durchgeführten Meta-Analyse anhand von 959 Nonresponse-Bias-Schätzungen aus 59 Studien zeigen, dass die Nonresponse-Rate nahezu keine Erklärungskraft für den Nonresponse-Bias besitzt (*Abbildung 10*).[63] Über eine einfache lineare Regression werden lediglich 4 % der Varianz des Nonresponse-Bias durch die Nonresponse-Rate erklärt (R^2 = 0,04; Groves und Peytcheva 2008, 174).[64] Für das Auftreten von Nonresponse-Bias müssen also weitere Größen von Bedeutung sein.

63 Zu diesem Ergebnis kommen auch Klausch u. a. 2015.

64 Die Abbildung entspricht der zweiten Abbildung bei Groves/Peytcheva (2008, 172). In Ermangelung des Datensatzes wurden die Daten aus dieser Abbildung für Abbildung 10 rekonstruiert. Die Grundlage der Abbildung bilden 959 Datenpunkte. Da durch Overplotting aber nur 527 Datenpunkte erkennbar sind, stimmt zwar die Darstellung optisch mit dem Original überein, Berechnungen mit der rekonstruierten Datenmatrix wären aber irreführend.

Abbildung 10:

Absoluter relativer Nonresponse-Bias für 959 Schätzungen in Abhängigkeit der Nonresponse-Rate in insgesamt 59 Surveys (Datenquelle: Groves/Peytcheva 2008, 172)



11.2 Nonresponse-Bias

Um zu verstehen, warum Nonresponse zu verzerrten Schätzungen führen kann, hilft die Betrachtung der Formel des Nonresponse-Bias. Dieser Bias ist in seiner einfachsten Form über

$$\bar{y}_{res} - \bar{y}_{all} = \frac{n_{non}}{n}(\bar{y}_{res} - \bar{y}_{non}) \qquad (14)$$

gegeben mit den entsprechenden Werten für alle Befragten (*all*), Respondenten (*res*) und Nonrespondenten (*non*) (z. B. Groves 1989, 134). An dieser Formel kann abgelesen werden, dass das primäre Problem – wie bereits gesehen – nicht im Ausmaß des Nonresponse, also dem Anteil von Nonrespondenten an allen zur Befragung vorgesehenen Personen $\frac{n_{non}}{n}$, sondern vielmehr in der Differenz zwischen Respondenten und Nonrespondenten $\bar{y}_{res} - \bar{y}_{non}$ liegt. Ist diese Differenz klein, unterscheiden sich Respondenten und Nonrespondenten also nicht (oder kaum), so kann auch bei einem großen Anteil an Nonrespondenten trotzdem von unverzerrten Schätzungen ausgegangen werden. Unterscheiden sich Respondenten und Nonrespondenten

allerdings systematisch, so sind die Schätzungen auch bei einem kleinen Anteil an Nonrespondenten verzerrt.[65] Dies ist besonders dann der Fall, wenn ein Zusammenhang zwischen Ausfallursache und dem Thema der Befragung besteht.

Im Hinblick auf die Ausfallursache ist an dieser Stelle darauf hinzuweisen, dass es sich bei Nonrespondenten um keine homogene Population handelt, auch wenn dies implizit in Formel 14 unterstellt wird. Für einen angemessenen Umgang mit Nonresponse ist somit eine weitere Unterteilung des Nonresponse in verschiedene Ausfallursachen notwendig.

11.3 Ausfallursachen

In der Literatur werden im Allgemeinen mindestens drei Kategorien von Ursachen für den Ausfall einer zur Befragung vorgesehenen Person genannt:

1. Verweigerung
2. Teilnahmeunfähigkeit
3. Nichterreichbarkeit

Interessanterweise ist es möglich, dass sich die Effekte des Nonresponse in den jeweiligen Gruppen unterscheiden, also verschiedene Gruppen von Nonrespondenten im Vergleich zu den Respondenten systematisch niedrigere

---

65 Eine Möglichkeit zur Schätzung des maximalen Bias beruht auf der Annahme, dass sich die Wahrscheinlichkeit für eine Teilnahme einer Person als *Responsepropensity* ρ schätzen lässt. Der maximal mögliche Bias ist dann durch

$$B_{max}(y,\rho) = \frac{\big(1 - R(\rho)\big)S(y)}{2\bar{\rho}} \geq \left|\frac{\mathrm{Cov}(y,\rho)}{\bar{\rho}}\right|$$

mit

$$R(\rho) = 1 - 2S(\rho)$$

gegeben, wobei $S(\rho)$ die Standardabweichung der Responsepropensities, $S(y)$ die Populationsvarianz der abhängigen Variablen, $\bar{\rho}$ den Mittelwert der Responsepropensities und $\mathrm{Cov}(y, \rho)$ die Kovarianz zwischen den Responsepropensities und der abhängigen Variablen darstellt. Diese Wahrscheinlichkeit ρ, dass eine für die Stichprobe ausgewählte Person auch antwortet, wird dabei durch eine Reihe von Hilfsvariablen x_j, beispielsweise über eine logistische Regression, geschätzt (Schouten u. a. 2009, 105). Der Bias ist dabei umso größer, je stärker die Korrelation der Responsepropensities ρ mit der untersuchten Variablen y ausfällt (Schouten u. a. 2009, 107). Die zentrale Schwäche dieses sogenannten R-Indikatoren-Ansatzes besteht in der Auswahl der Hilfsvariablen x_j. Wenn der Nonresponse-Mechanismus nicht mit den zur Schätzung der Responseprobabilities verwendeten Hilfsvariablen korreliert, bleibt der Bias unbemerkt (Schnell 2012, 174). Somit würde eine unverzerrte Schätzung angenommen, obwohl dies nicht der Fall ist.

oder höhere Mittelwerte aufweisen oder auch gar keine Differenz zu beobachten ist. Daher ist es prinzipiell möglich (wenn auch in der Praxis eher unwahrscheinlich), dass eine Erhöhung der gesamten Ausschöpfung zu einer Vergrößerung der Verzerrung führen kann, und zwar wenn Subgruppen mit kleiner Differenz zu den Respondenten stärker ausgeschöpft werden als Subgruppen mit großer Differenz. Eine Erweiterung von Formel (14) soll dies verdeutlichen. Diese erweiterte Formel ist über

$$\bar{y}_{res} - \bar{y}_{all} = \frac{n_{nc}}{n}(\bar{y}_{res} - \bar{y}_{nc}) + \frac{n_{rf}}{n}(\bar{y}_{res} - \bar{y}_{rf}) + \frac{n_{na}}{n}(\bar{y}_{res} - \bar{y}_{na}) + \frac{n_{ot}}{n}(\bar{y}_{res} - \bar{y}_{ot}) \quad (15)$$

gegeben, wobei *nc* für nicht erreichte Personen (*noncontacts*), *rf* für Verweigerer (*refusals*), *na* für teilnahmeunfähige Personen (*not able*) und *ot* für andere Gründe (*other*) steht (siehe Schnell 2012, 171). Im Folgenden soll auf die einzelnen Ausfallursachen näher eingegangen werden.

11.3.1 Verweigerung

In der Literatur werden viele Gründe als Ursachen für die Verweigerung der Teilnahme an einem Survey diskutiert, beispielsweise die Belastung durch Länge oder Häufigkeit der Befragung, politisches Desinteresse, altersbedingter Rückzug aus öffentlichen Angelegenheiten, Kriminalitätsfurcht oder unklare Konsequenzenbefürchtungen (Schnell 2012, 159). All diesen Hypothesen ist gemein, dass sie sich als Spezialfälle der Rational-Choice-Theorie interpretieren lassen (Schnell 1997, 157–216). Demnach ist eine Teilnahme dann zu erwarten, wenn der erwartete Nutzen die erwarteten Kosten übersteigt. Die Kosten und der Nutzen werden von den einzelnen Befragten individuell bewertet, sodass deren Einschätzung von den Besonderheiten und jeweiligen situativen Bedürfnissen und Erwartungen der Befragten abhängig ist.

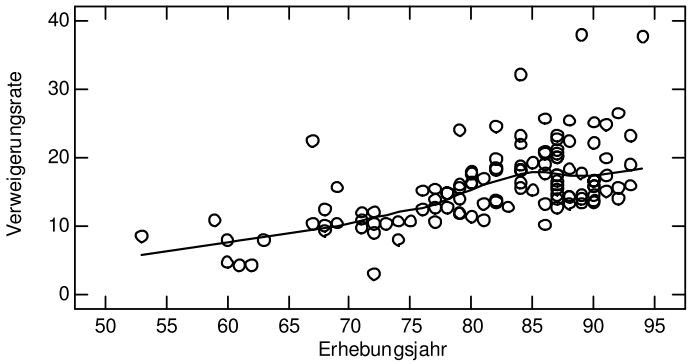
Durch die starke Routinisierung großer Teile des alltäglichen Lebens und Handelns reichen oftmals schon kleine Reize aus, um entsprechende Handlungsskripte bei den Befragten auszulösen (Schnell 2012, 159). Somit wird erklärbar, warum gelegentlich schon minimale Veränderungen in der Erhebungssituation (kleine Geschenke, Sprache oder Aussehen der Interviewer) zu größeren Veränderungen im Anteil der Verweigerungen führen können. Die Entscheidung zur Verweigerung scheint also stark von der Situation abzuhängen und ist somit nicht als über die Situation hinaus stabile Eigenschaft der Befragten zu sehen. Sowohl die hohen Konvertierungsraten von Verweigerern (bis zu maximal 30 %) als auch der Zeitpunkt der Verweigerung (gewöhnlich in den ersten Sekunden des Interviews, noch bevor das Thema der

Befragung erläutert wurde) können als Hinweis darauf interpretiert werden.[66] Dazu passend zeigen sich in der Regel nur schwache Korrelationen zwischen Verweigerungsverhalten und Hintergrundvariablen. Entsprechend existieren keinerlei empirische Hinweise auf einen „harten Kern“ von Verweigerern (Schnell 1997, 151, 186; Schnell 2012, 159–160).

Trotzdem muss bei jeder Studie geklärt werden, ob Verweigerungsgründe systematisch mit dem Thema der Untersuchung in Zusammenhang stehen. In einem solchen Fall sind verzerrte Schätzungen zu erwarten. So könnten beispielsweise ältere Personen Befragungen eher verweigern, weil sie die Kontaktaufnahme als Teil eines potenziellen Trickbetrugs wahrnehmen. Von weiblichen Personen könnte die Kontaktaufnahme in persönlichen Befragungen als Versuch wahrgenommen werden, Zutritt zur Wohnung in Absicht eines sexuellen Übergriffs zu erlangen. In diesen Fällen würde der Ausfallgrund mit den inhaltlichen Merkmalen einer Viktimisierungsstudie zusammenhängen, sodass beispielsweise das Kriminalitätsfurchtniveau älterer Personen vor Trickbetrug oder die Angst weiblicher Personen vor sexuellen Übergriffen unterschätzt werden würden.

Abbildung 11:

Entwicklung der Verweigerungsraten in akademischen Surveys in der BRD 1953–1994 (Schnell u. a. 2013, 301)



[66] Die Höhe der Konvertierungsraten hängt maßgeblich von den Merkmalen und dem Verhalten der Interviewer ab. Ein Beispiel für ein vollständiges Verweigerungs-Reduktions-Training in deutscher Sprache findet sich bei Schnell (2012, 223–225).

Wie *Abbildung 11* zu entnehmen sind die Verweigerungsraten spätestens seit den 70er Jahren deutlich gestiegen. Man muss bedenken, dass dieser Trend trotz der gegenteiligen Bemühungen der Institute zu beobachten ist. Generell dürfte diese ansteigende Verweigerungstendenz unumkehrbar sein. Allerdings zeigen die großen Streuungen der Verweigerungsraten zwischen verschiedenen Studien gleichsam, dass ein großer Teil des Verweigerungsverhaltens offensichtlich von den Details der Feldarbeit abhängt. Es macht für eine endgültige Verweigerung einen deutlichen Unterschied, ob alle Regeln für ein erfolgreiches Interview (Ankündigung, Belohnung, Interviewerwechsel, Konvertierungsversuche, Wechsel des Erhebungsmodus etc., zu den Details Schnell 2012, 181–183, 223–225) beachtet wurden oder nicht. Entscheidend ist dabei, dass nicht eine einzelne Maßnahme zu einer deutlichen Verbesserung der Ausschöpfung führt, sondern nur die konsequente Anwendung aller Maßnahmen. Entsprechend kostenintensiv sind korrekt durchgeführte Erhebungen.

11.3.2 Erkrankung/Teilnahmeunfähigkeit

Es existieren verschiedene Gründe, warum eine Person nicht an einer Befragung teilnehmen kann, beispielsweise nicht ausreichende Sprachkenntnisse, Analphabetismus in einer postalischen Befragung, psychische Probleme, chronischer Alkohol- und Drogenmissbrauch oder schwere Erkrankungen (z. B. Demenz). Sollte der Grund der Nichtbefragbarkeit mit dem Thema des Surveys in Zusammenhang stehen, so ist mit verzerrten Schätzungen zu rechnen (Schnell u. a. 2013, 302).

Dies wäre z. B. dann der Fall, wenn eine zur Befragung vorgesehene Person aufgrund einer Viktimisierung entweder physisch oder psychisch nicht in der Lage ist, an der Befragung teilzunehmen. Ebenso würden die Viktimisierungsraten unterschätzt, wenn ein systematischer Zusammenhang zwischen einer vorgefallenen Viktimisierung und Deutschkenntnissen bestünde.

11.3.3 Nichterreichbarkeit

Für die meisten Studien stellen weder Verweigerungen noch Befragungsunfähigkeit prinzipiell das größte Problem dar, sondern schwer- und nicht erreichbare Personen. Damit sind Personen gemeint, die trotz mehrfacher Kontaktversuche an ihrem Wohnsitz nicht angetroffen werden. Neben Personen mit besonders vielen Sekundärkontakten (z. B. politisch Aktive, Vereinsmitglieder usw.) trifft dies auch auf längere Zeit Verreiste, Personen, deren tatsächlicher Aufenthaltsort nicht mit ihrem Wohnsitz übereinstimmt (z. B. Montagearbeiter), und Personen mit ungewöhnlichen Arbeitszeiten (z. B.

Krankenpflegepersonal) zu (Schnell u. a. 2013, 303). Diese Ausfälle erfolgen nicht zufällig, sondern hängen offensichtlich mit bestimmten Merkmalen der Personen zusammen. Übereinstimmend damit zeigt sich, dass die Erreichbarkeit der zur Befragung vorgesehenen Personen mit zahlreichen sozialwissenschaftlich relevanten Variablen korreliert. Aus diesem Grund können schwierig erreichbare Personen nicht einfach durch leicht erreichbare Personen ersetzt werden (Schnell 1998). Sind Personen schwieriger oder nicht erreichbar, weil sie vielen Aktivitäten außerhalb der eigenen Wohnung nachgehen, und weisen diese Personen eine erhöhte Wahrscheinlichkeit auf, Opfer eines Verbrechens zu werden, so wird die entsprechende Viktimisierungsrate unterschätzt. Dieser Zusammenhang wird zum Beispiel von Hindelang u. a. (1978, 250) oder Cohen/Felson (1979, 589) beschrieben.

Daher müssen auch schwierig Erreichbare in die Stichprobe aufgenommen werden. Zumeist wird über mehrere Kontaktversuche (*Callbacks*) zu verschiedenen Tageszeiten versucht, die zur Befragung vorgesehene Person zu erreichen. Erfolgsversprechende Zeiten für Kontaktversuche liegen in den frühen Abendstunden oder am Wochenende, wobei bei persönlichen Befragungen auch ein Wechsel der Interviewer wünschenswert ist, da sich Interviewer in ihren Kontaktstrategien unterscheiden. Interessanterweise lässt sich auch die Nichterreichbarkeit durch Incentives beeinflussen. Insgesamt zeigt sich, dass sich die Zahl der Nichterreichten durch eine flexible Kontaktstrategie und eine hohe Zahl von Callbacks deutlich reduzieren lässt. Allerdings steigen mit zunehmender Callback-Zahl auch die Kosten pro Interview (Schnell u. a. 2013, 303).

11.4 Vermeidung und Kontrolle von Nonresponse statt Korrektur

Wie bereits ausgeführt ist die Responserate allein nicht geeignet, um Aussagen über mögliche auf Nonresponse zurückgehende Verzerrungen zu treffen. Man benötigt für jede neue Studie erneut eine Analyse der Ursachen des Nonresponse.[67] In der Regel bedeutet dies, für jede Gruppe der Ausfallmechanismen (Erkrankung, Verweigerung, Nichterreichbarkeit) die möglichen Effekte auf die jeweilige Studie zu analysieren. Dies muss bereits vor der ersten

[67] Hierfür sind Informationen über die Erhebung, sogenannte Paradaten (z. B. Datum, Uhrzeit, Nummer und Ergebnis des Kontaktversuchs) notwendig. Diese Daten sind für die Erhebungsinstitute prinzipiell leicht verfügbar und nahezu kostenneutral zu erhalten, trotzdem sind nicht alle Erhebungsinstitute bereit, diese Paradaten auch zur Verfügung zu stellen. Demnach sollte im Vorfeld durch den Auftraggeber vertraglich festgehalten werden, welche (Para-)Daten als Teil des Datensatzes vom Erhebungsinstitut zu liefern sind (siehe hierzu Anhang F in Schnell u. a. 2013).

Feldphase erfolgen. Im Anschluss an diese Analyse muss dann das Design der Studie (z. B. Erhebungsmodus, Klumpung und Schichtung, Interviewerkontrolle, Interviewerallokation, Incentives, Tracking-Maßnahmen, Interviewerschulung, Verweigerungstraining etc.) angepasst werden.

An dieser Stelle sei explizit darauf hingewiesen, dass die einzig erfolgsversprechende Strategie im Umgang mit dem Nonresponse-Problem darin besteht, bereits in der Feldphase alle möglichen Schritte zu unternehmen, um Unit-Nonresponse so weit wie möglich zu verhindern. Die beste Lösung des Nonresponse-Problems besteht darin, kein Nonresponse-Problem zu haben.[68] Methodologen sind sich einig, dass Nonresponse ein bereits vor einer Erhebung zu berücksichtigendes und minimierendes Problem ist. In der Praxis zeigen sich hingegen naive Anwender durch ein hohes Ausmaß an Nonresponse überrascht und betrachten es fälschlich als unabwendbares Naturereignis. Weder Verweise auf zahlreiche andere Studien mit Nonresponse-Problemen noch heroische Annahmen über die vermeintliche Neutralität der Ausfälle sind akzeptable Handlungsstrategien. Dies gilt auch für alle Versuche, Nonresponse nachträglich allein durch spezielle Gewichtungsverfahren zu „korrigieren“ (siehe Unterkapitel 11.6).

11.5 Nonresponse in neueren deutschen Viktimisierungssurveys

Da für die Bundesrepublik im Gegensatz zu den Vereinigten Staaten und zum Vereinigten Königreich kein amtlicher Viktimisierungssurvey wie der NCVS bzw. der *Britisch Crime Survey* (BCS) existiert, ist die kriminologische Forschung in Deutschland auf eigene Erhebungen angewiesen. Interessanterweise existiert aber bislang keine Übersicht über Nonresponse in neueren Erhebungen, in denen kriminologisch relevante Fragen Teil des Frageprogramms waren. Diese Lücke wurde durch die von den Autoren betreute Abschlussarbeit von Klingwort (2014) geschlossen. Gesucht wurden Surveys anhand folgender Kriterien:

– Feldzeit ab dem 01. 01. 2001 und

– Stichprobengröße ≥ 1.500.

[68] Anderson u. a. (1983) schreiben in Hinsicht auf fehlende Werte allgemein dieses Motto ohne Quellenangabe bereits Snedecor zu.

Weiterhin sollten Fragen nach

– Viktimisierung in Form von körperlicher Gewalt und/oder

– Viktimisierung in Form von Wohnungseinbruch und/oder

– zur Kriminalitätsfurcht

erhoben worden sein. In die resultierende Auflistung wurden sowohl spezielle Viktimisierungssurveys als auch Mehrthemenbefragungen mit einem kriminologischen Teil aufgenommen.

Besonderes Augenmerk lag dabei auf der Definition der Grundgesamtheit als „bundesweite allgemeine Bevölkerung". Trotzdem wurden auch die Erhebungen in die Analyse einbezogen, die zwar den ersten Kriterienkatalog erfüllten, sich aber nur auf einzelne Regionen oder spezielle Subpopulationen bezogen.[69]

Von den 34 identifizierten Projekten mit insgesamt 105 Erhebungen mit kriminologischem Bezug seit 2001, die alle Kriterien erfüllten, beziehen sich lediglich 13 Projekte auf die allgemeine Bevölkerung (Klingwort 2014, 10–32).[70]

---

69 Hierbei handelt es sich um die Projekte „Lebenssituation, Sicherheit und Gesundheit von Frauen in Deutschland", „Kriminalität und Gewalt im Leben alter Menschen", „Jugendliche in Deutschland als Opfer und Täter von Gewalt", „Lebenssituation und Belastung von Frauen mit Behinderungen und Beeinträchtigungen in Deutschland", „Gender-based Violence, Stalking and Fear of Crime", „Repräsentativbefragung zu Viktimisierungserfahrungen in Deutschland", „Violence against Women: an EU-wide Survey", „Muslime in Deutschland" (Erwachsene, Studenten, Schüler), „Jugendgewalt und Jugenddelinquenz in Hannover", „Second International Self-Reported Study of Delinquency", „Sicherheit und Kriminalität in Stade" (Schüler, Erwachsene), „European Union Minorities and Discrimination Survey", „Jugendliche als Opfer und Täter von Gewalt im Bundesland Sachsen-Anhalt", „Kriminalitäts- und Terrorismusfurcht in Hessen", „Jugendliche als Opfer und Täter von Gewalt in Wolfsburg", „Kinder- und Jugenddelinquenz im Bundesland Saarland", „Jugendliche als Opfer und Täter von Gewalt in Berlin", „Jugendliche als Opfer und Täter von Gewalt im Landkreis Emsland", „Gewalt im Strafvollzug", „Youth Deviance and Youth Violence" und „Jugendkriminalität in der modernen Stadt" (2001–2009, 2011, 2013) (Klingwort 2014, 33–55).

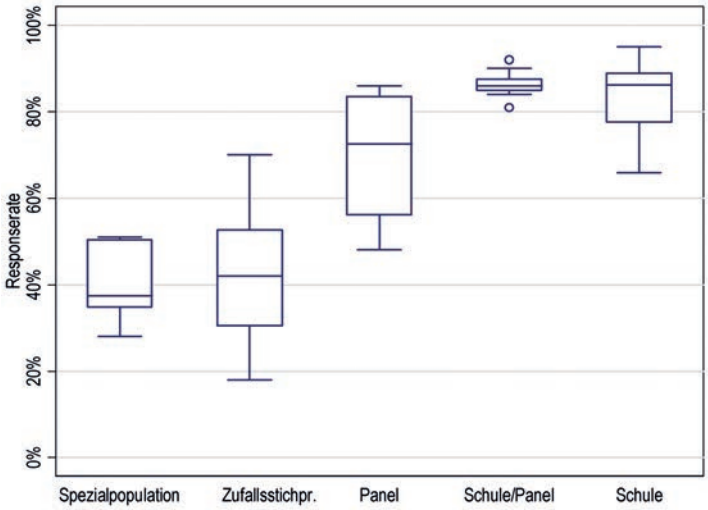
70 Hierbei handelt es sich um folgende Projekte: „Die Ängste der Deutschen" (jährlich 2001–2014), „Kriminalitätsfurcht, Strafbedürfnisse und wahrgenommene Kriminalitätsentwicklung" (2004, 22006, 2010), „European Survey of Crime and Safety 2005 (EU ICS)", „Studie zur Gesundheit Erwachsener in Deutschland", „Kriminalität und Sicherheitsempfinden", „International Crime Victims Survey-Pilotstudie 2010" (CATI und Online-Panel), „Deutscher Viktimisierungssurvey 2012 (Barometer Sicherheit in Deutschland)", „Sicherheitsreports" (2011, 2012, 2013), „Das Sozio-oekonomische Panel (SOEP, jährlich 2001–2012)", „Eurobarometer" (Standard Eurobarometer zwei Mal jährlich 2001–2013, Nr. 56–59, 61–79), „Special Eurobarometer 2010", „European Social Survey" (2002, 2004, 2006, 2008, 2010, 2012), „Allgemeine Bevölkerungsumfrage der Sozialwissenschaften 2008" und „Allgemeine Bürgerbefragungen der Polizei in Nordrhein-Westfalen" (Klingwort 2014, 10–32). Die letztgenannte Studie würde die Aufnahmekriterien eigentlich nicht erfüllen, da sie auf NRW beschränkt ist. Da NRW aber fast 22 % der Bevölkerung der Bundesrepublik umfasst, wird diese Studie trotzdem aufgeführt.

In diesen Projekten wurden insgesamt 71 Erhebungen durchgeführt. Davon sind 17 Erhebungen Quotenstichproben. Da sich für Quotenstichproben keine Nonresponse-Quoten angeben lassen (obwohl Quotenstichproben auch ein Nonresponse-Problem besitzen), sind diese Studien für Nonresponse-Analysen irrelevant und wurden aus der Betrachtung ausgeschlossen. Die folgenden Analysen basieren auf 18 Erhebungen aus dieser Gruppe.[71]

Die verbleibenden 21 Projekte bestehen aus 34 Erhebungen, wobei für drei dieser Erhebungen keine Responseraten angegeben werden können.[72] Somit basiert *Abbildung 12* auf 18 + 31 = 49 Erhebungen.

Abbildung 12:

Responseraten nach Auswahlverfahren und Art der Zielpopulation (Datengrundlage: Klingwort 2014)



71 Für die Eurobarometererhebungen (24 Erhebungen) sind in den Dokumentationen keine Informationen über Nonresponse enthalten. Für das SOEP (12 Wellen) wurden keine Responseraten recherchiert, da dies hier für ein seit vielen Jahren laufendes Panel nicht sinnvoll erscheint. Informationen zu den Ausschöpfungsquoten der einzelnen SOEP-Wellen können Kroh 2014 entnommen werden.

72 Dies sind die Erhebungen „Gender-based Violence, Stalking and Fear of Crime“ (nicht dokumentiert), „Repräsentativbefragung zu Viktimisierungserfahrungen in Deutschland“ (Quotenstichprobe) und „Jugendkriminalität in der modernen Stadt 2013“ (Information liegt nicht vor, aktuellster Methodenbericht von Bentrup und Verneuer 2014 bezieht sich auf 2011).

Die Responseraten dieser Erhebungen sind nach Erhebungstyp zusammengefasst. Hier zeigt sich für die Random-Erhebungen eine Responserate von durchschnittlich ca. 41 %, für die Erhebungen von Spezialpopulationen von ungefähr ca. 40 %. *Abbildung 12* enthält eine Reihe interessanter Details.

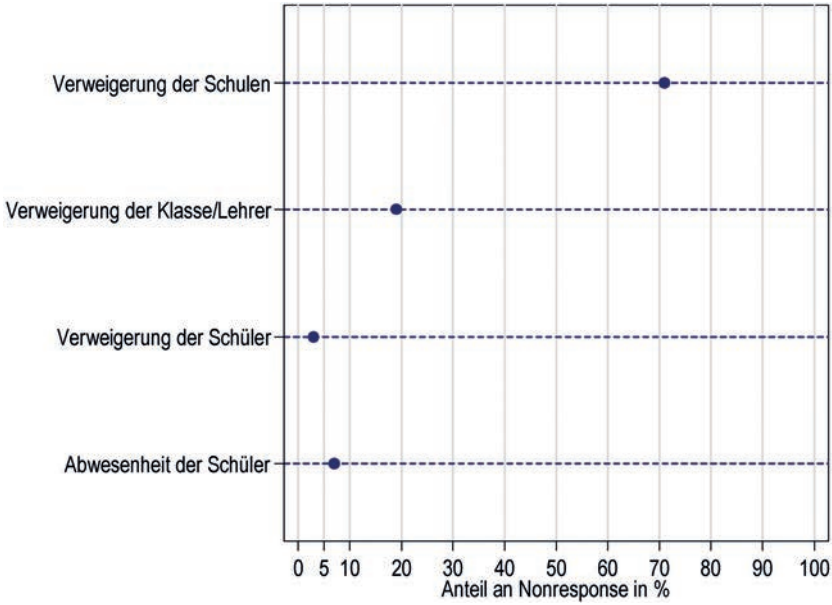
Erstens kann eine große Streuung innerhalb einiger der Gruppen beobachtet werden. So liegen beispielsweise die Responseraten für Erhebungen auf der Basis von Zufallsstichproben zwischen 20 % und 70 %. Schwankungen dieser Größe innerhalb von wenigen Jahren sind nicht zufällig. Die oben beschriebenen kumulativen Effekte scheinbar trivialer Details wie Ankündigung, Feldzeit, Incentives, Konvertierungsversuche etc. resultieren in eindrucksvollen Differenzen von mehr als 50 % Unterschied in der Ausschöpfung einer Zufallsstichprobe aus der bundesweiten „allgemeinen" Bevölkerung. Dieser große Unterschied in der Ausschöpfungsquote ist das wichtigste Ergebnis der Arbeit von Klingwort (2014).

Nicht so bedeutsam für die Methodenforschung im Allgemeinen ist das erwartbare zweite Ergebnis der Studie: die Beobachtung hoher Responseraten für Panelstudien und Schulbefragungen sowie besonders für Panelstudien mit Schulen. Allgemein sinkt die Kooperationsbereitschaft der Befragten in Panelstudien – falls die erste Befragung nicht traumatisch war – mit der Anzahl der teilgenommenen Wellen nur langsam. Der Ausfall aus einem Panel (*Dropout*, *Attrition*) lässt sich durch Belohnungen für die Teilnehmer (Untersuchungsberichte sind keine geeigneten Belohnungen) und hohen Aufwand für das Verfolgen kooperationsbereiter, aber verzogener Personen („Tracking") weiter reduzieren.

Schülerbefragungen sind aus verschiedenen Gründen als problematisch anzusehen. Zwar ist bei einer Befragung der Schülerschaft einer Schule innerhalb der Klassen eine hohe Kooperationsbereitschaft erwartbar, doch setzt dies voraus, dass die jeweilige Schule kooperiert. Dies ist keineswegs selbstverständlich. Der Großteil des Nonresponse in der Studie „Second International Self-Reported Study of Delinquency" (2006) ging beispielsweise auf die komplette Verweigerung der Teilnahme einiger Schulen zurück (Enzmann 2010, 51, siehe *Abbildung 13*). Entsprechend sollte bei der Beurteilung von Schülerbefragungen beachtet werden, ob ein solcher Komplettausfall einer Schule (korrekt) als Nonresponse codiert oder als vermeintlich unsystematischer Ausfall betrachtet wurde.

Abbildung 13:

Nonresponse nach Ursache in der Schülerbefragung „Second International Self-Reported Study of Delinquency“ von 2006 (Enzmann 2010, 51)



Weiterhin ist in Deutschland eine Einwilligung der Eltern zur Befragung Minderjähriger notwendig (Schnell 2012, 166). Auch diese Einwilligung kann durchaus systematisch mit kriminologischen Variablen zusammenhängen.

Das schwerwiegendere Problem bei Schülerbefragungen besteht aber darin, dass Absentismus der Schulpflichtigen an die Delinquenz eben dieser gekoppelt zu sein scheint (Vaughn u. a. 2013, 773). Wenn also gerade delinquente Jugendliche nicht erscheinen oder in einer Panelstudie die Schule eher abbrechen als andere Schülerinnen und Schüler, so stellen die verbleibenden Panelteilnehmer keine zufällige Auswahl aus der gesamten Schülerschaft dar.[73]

[73] Ein deutsches Beispiel für diesen Effekt findet sich in der Duisburger kriminologischen Schuluntersuchung bei Pöge 2007.

Daher ist auch die Dauer der Feldzeit innerhalb der Schulen für Schülerbefragungen von Bedeutung. Wird die Befragung nur an einem Stichtag durchgeführt und nicht über einen längeren Zeitraum ist mit deutlich niedrigeren Responseraten und eher mit systematischen Ausfällen zu rechnen. Bei Studien mit solchen durch Absentismus systematisch erfolgenden Ausfällen ist die geschätzte Delinquenz demnach lediglich als Untergrenze der tatsächlichen Delinquenz anzusehen. Da auch bei Jugendlichen aufgrund des sogenannten *Victim-Offender Overlap* Täter, insbesondere bei Gewaltdelikten überproportional häufig Opfer dieser Delikte werden (Shaffer/Ruback 2002), wird demnach durch Absentismus nicht nur die Anzahl der Täter, sondern auch die die Anzahl der Opfer unterschätzt.

Abschließend möchten wir nochmals betonen, dass die Ausschöpfungsquote allein keinen Hinweis auf mögliche Nonresponse-Effekte liefert. Allerdings sollten mittlere Ausschöpfungen von 40 % ein deutlicher Hinweis darauf sein, das Nonresponse bereits bei der Planung einer Erhebung berücksichtigt werden muss – und nicht erst bei der Analyse.

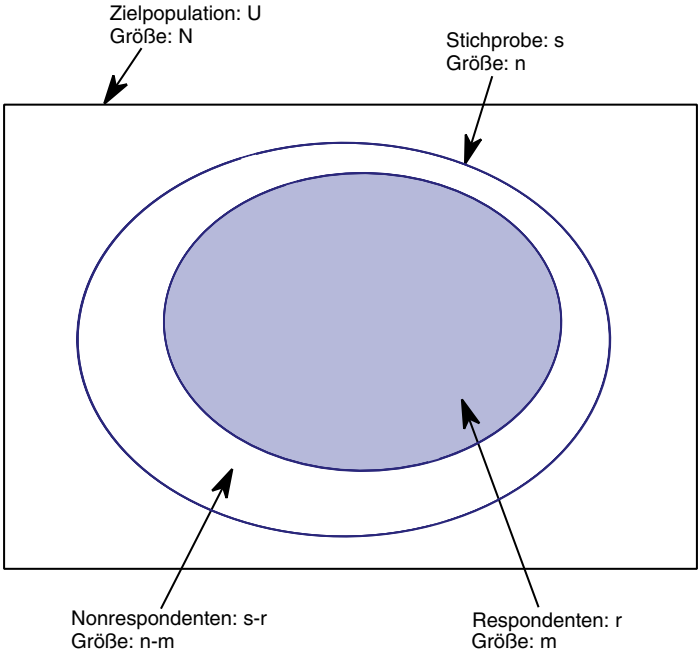
11.6 Korrekturverfahren für Nonresponse-Bias

Trotz aller Anstrengungen während der Feldphase ist Nonresponse unvermeidlich. Üblicherweise wird versucht, zumindest offensichtliche Verzerrungen der Stichprobe nachträglich durch Gewichtungsverfahren zu kompensieren.[74] Dies bedeutet, dass jedem Fall im Datensatz ein bei Analysen zu berücksichtigendes Gewicht zugeordnet wird. *Abbildung 14* illustriert die im Folgenden verwendete Notation.

[74] In der Statistik wird zwischen Design- und Korrekturgewichten unterschieden. Designgewichte korrigieren für unterschiedliche Auswahlwahrscheinlichkeiten durch das Design der Stichprobe, z. B. werden bei disproportionalen Stichproben Personen aus Bundesländern, die überproportional gezogen wurden, entsprechend geringer gewichtet. Designgewichte sind in der Statistik unstrittig. Ob und wie Korrekturgewichte verwendet werden sollten, ist nicht abschließend geklärt. Bei Längsschnittstudien kommen noch Längsschnittgewichte hinzu, die häufig wiederum Design- und Korrekturgewichte enthalten. Das Gesamtgewicht eines Falls ist immer das Produkt der Einzelgewichte.

Abbildung 14:

Gezogene und realisierte Stichprobe als Teilmengen der Zielpopulation (als identisch mit der Auswahlgesamtheit angenommen; in Anlehnung an Särndal/Lundström 2005, 44)



Für den Fall einer Zufallsstichprobe ohne Zurücklegen der Größe n mit identischen Inklusionswahrscheinlichkeiten $\pi_k = n/N$ und Gewichten $d_k = 1/\pi_k$ für alle Elemente der Stichprobe und ohne Nonresponse stellt das über die Werte y_k berechnete Stichprobentotal $\hat{Y}$ über den Horvitz-Thompson-Schätzer

$$\hat{Y} = \sum_r d_k y_k \tag{16}$$

eine unverzerrte Schätzung für das Total der Population Y mit

$$Y = \sum_U y_k \tag{17}$$

dar.[75] Fallen allerdings Personen durch Nonresponse aus, wird Y nicht mehr unverzerrt über $\hat{Y}$ geschätzt (Särndal und Lundström 2005, 57). Die Schätzung muss dann durch Gewichtung korrigiert werden und erfolgt für alle befragbaren Personen ($k \in r$) über

$$\hat{Y}_W = \sum_r w_k y_k, \tag{18}$$

wobei w_k die Korrekturgewichte der Respondenten sind. Es existieren mehrere Methoden, um diese Korrekturgewichte zu konstruieren, alle diese Methoden haben jedoch gemeinsam, dass der gewichtete Schätzer bessere Eigenschaften aufweisen soll, als die ungewichtete Schätzung.

Hierzu werden zusätzliche Informationen in Form von Hilfsvariablen (*auxiliary variables*) benötigt. Dies können zusätzliche Variablen sein, die für alle zur Befragung vorgesehenen Personen (also Respondenten und Nonrespondenten) oder die gesamte Population vorliegen, aber auch Informationen über die Verteilung der Hilfsvariablen in der Population. So können beispielsweise die zur Stichprobenziehung verwendeten Auswahlgrundlagen zusätzliche Variablen enthalten (z. B. bei Einwohnermeldedateien das Alter). Es können aber auch Daten aus anderen administrativen Registern, Aggregatstatistiken (z. B. Zahl der Krankenhausentlassungen pro Jahr) oder Daten über den Prozess der Datenerhebung selbst (Paradaten) zur Konstruktion der Korrekturgewichte verwendet werden.[76]

Notwendige Voraussetzung ist allerdings, dass die zur Gewichtung herangezogenen Hilfsvariablen mit dem Nonresponse-Mechanismus in Zusammenhang stehen. Sind sie vollkommen unabhängig von der Ausfallursache, wird die Gewichtung keine bessere Schätzung erbringen als die ungewichtete Schätzung.

Aktueller Standard in der Statistik für die Korrektur von Nonresponse ist der sogenannte Calibration-Ansatz (Deville/Särndal 1992), der zahlreiche ältere Verfahren als Spezialfall enthält. Der erste Schritt einer solchen Gewichtung besteht in der Auswahl geeigneter Hilfsvariablen x_k. Anschließend werden im zweiten Schritt Gewichte w_k gesucht, die die sogenannte Kalibrierungsgleichung

$$\sum_r w_k \mathrm{x}_k = \mathrm{X} \tag{19}$$

---

75 Diese Inklusionswahrscheinlichkeiten können sich für komplexere Stichprobendesigns zwischen den Elementen der Stichprobe (bspw. in disproportional geschichteten Stichproben) unterscheiden. Ist π_i konstant, spricht man von einer EPSEM-Stichprobe (*Equal Probability of Selection Method*, Kish 1965, 21).

76 Einzelheiten finden sich vor allem bei Särndal/Lundström 2005 und Valliant u. a. 2013, zu Paradaten allgemein siehe die Beiträge in dem Sammelband von Kreuter 2013.

erfüllen. Gewichte, die dieser Gleichung genügen, werden im Hinblick auf X als kalibriert bezeichnet (Särndal/Lundström 2005, 58).

Wie bereits erwähnt müssen im Falle von Stichproben mit ungleichen Inklusionswahrscheinlichkeiten oder der Schätzung des Population Totals über den Horvitz-Thompson-Schätzer (Formel (16)) zusätzlich die durch das Design erforderlichen Gewichte d_k berücksichtigt werden, die als Kehrwert der Inklusionswahrscheinlichkeit über $d_k = 1/\pi_k$ definiert sind. Designgewichte sind notwendig, aber nicht in der Lage, Nonresponse zu korrigieren. Die Designgewichte werden daher mit einem zusätzlichen Faktor υ_k korrigiert. Diese neuen korrigierten Gewichte sind dann als

$$w_k = d_k \upsilon_k \tag{20}$$

definiert.

Wird ein linearer Zusammenhang zwischen den Hilfsvariablen x_k und dem Korrekturfaktor v_k angenommen, ergibt sich der Korrekturfaktor über

$$\upsilon_k = 1 + \lambda' \mathrm{x}_k, \tag{21}$$

wobei in einem nächsten Schritt der Vektor λ aus dieser Gleichung bestimmt werden muss. Wird $w_k = d_k\,(1 + \lambda' x_k)$ in Formel (19) eingesetzt und nach λ aufgelöst, so ergibt sich

$$\lambda' = \left(\mathrm{X} - \textstyle\sum_r d_k \mathrm{x}_k\right)' \left(\textstyle\sum_r d_k \mathrm{x}_k \mathrm{x}'_k\right)^{-1} \tag{22}$$

als Lösung. Die Korrekturgewichte w_k ergeben sich dann über

$$w_k = d_k + d_k \lambda' \mathrm{x}_k \tag{23}$$

und führen damit zur kalibrierten Schätzung

$$\hat{Y}_W = \textstyle\sum_r w_k y_k \tag{24}$$

als Produkt der konkreten Messung y_k und des Korrekturgewichts w_k (Särndal/Lundström 2005, 57–59).[77]

[77] Für Kalibrierungen stehen für Statistikprogramme wie STATA oder R kostenlose zusätzliche Makros zur Berechnung zur Verfügung, daneben gibt es auch teilkommerzielle Lösungen wie z. B. BASCULA als Teil von BLAISE, dem CATI-System des niederländischen CBS.

Es muss nochmals betont werden, dass die Übereinstimmung einiger Verteilungen der Stichprobe mit einigen Verteilungen in der Grundgesamtheit nicht beweist, dass die Stichprobe eine „repräsentative“ Stichprobe, frei von Verzerrungen oder eine Zufallsstichprobe ist. Solche Übereinstimmungen zeigen nur, dass der Selektionsmechanismus nicht mit den überprüften Variablen zusammenhängt (Schnell 1993). Trotz übereinstimmender Randverteilungen (gleichgültig ob vor oder nach einer Gewichtung oder Kalibrierung) zwischen Stichprobe und Grundgesamtheit sind irreführende Verallgemeinerung oder verzerrte Schätzungen keineswegs ausgeschlossen oder weniger wahrscheinlich. Nehmen wir z. B. an, ältere Menschen würden aus gesundheitlichen Gründen weniger an Befragungen teilnehmen. Trotzdem wird es einige ältere Menschen in der Stichprobe geben, die vermutlich etwas gesünder als die älteren Nichtteilnehmer sind. Werden diese gesunden Älteren nun höher gewichtet, dann wird der Anteil der Gesunden in der Grundgesamtheit überschätzt, obwohl der Anteil der Älteren, eventuell auch Geschlecht und Bildung mit der Verteilung in der Grundgesamtheit übereinstimmen. Gewichten verringert Verzerrungen nur dann, wenn die Gewichtungsvariablen mit dem Selektionsmechanismus stark zusammenhängen. Das lässt sich empirisch anhand einer gegebenen Stichprobe allein nicht überprüfen, sondern nur mittels einer Stichprobe ohne Nonresponse oder der Grundgesamtheit. Stehen diese Daten nicht zur Verfügung, müssen heroische Annahmen getroffen werden. Das ist in keiner Weise problematisch, muss aber klar in einer Studiendokumentation thematisiert werden.

Wir haben hierzu ein fiktives Beispiel mit den Daten des *British Crime Survey* (BCS) 2010–2011 berechnet. Für dieses Beispiel wurde die Stichprobe des BCS (etwa n = 47.000) als Population verwendet, aus der eine Zufallsstichprobe der Größe n = 3.000 gezogen wurde. Für diese Stichprobe wurde dann der Ausfall von ca. 25 % der Befragten in Abhängigkeit von der Anzahl der erwachsenen Haushaltsmitglieder, des durch die Interviewer wahrgenommenen Ausmaßes an Incivilities sowie der Lage des Haushalts (Innenstadt: ja/nein) modelliert.[78] Als Hilfsinformationen für die Kalibrierung konnten zwei Quellen verwendet werden: einerseits die Informationen aus der Population (BCS-Gesamtstichprobe). Hier werden „Populationstotale“ für fünf Altersklassen und Geschlecht verwendet, die beispielsweise der amtlichen Statistik entnommen werden könnten. Weiterhin können die für alle Personen (also Respondenten und Nonrespondenten) vorliegenden Variablen „Zahl der er-

[78] Die Gewichte, mit denen diese Variablen in die Ausfallmodellierung eingingen, wurden über eine Regression des Kriminalitätsfurcht-Standardindikators auf diese Variablen ermittelt. Zu den resultierenden vorhergesagten Werten wurde eine Zufallskomponente addiert ($N(0; 0.25)$) und die 25 % größten Werte dann als fehlend markiert. Ohne diese Zufallskomponente wäre durch die Kalibrierung eine unverzerrte Schätzung erreicht worden.

wachsenen Haushaltsmitglieder“, „durch Interviewer wahrgenommene Incivilities“ sowie „Innenstadtlage des Haushalts“ verwendet werden. Diese Art von Informationen kann entweder im Sampling-Frame vorliegen oder durch die Interviewer während der Erhebung gesammelt werden (Paradaten). Mit diesen Informationen wurden dann die kalibrierten Korrekturgewichte in einem zweistufigen Verfahren konstruiert (Typ B, siehe Särndal/Lundström 2005, 81–83). Für dieses Beispiel ergibt die ungewichtete Berechnung des Anteils „ängstlicher Personen“ einen Wert von 24,8 %.[79] Wird die Schätzung mit den durch das Calibrate-Verfahren konstruierten Korrekturgewichten durchgeführt, ergibt sich ein Wert von 26,8 %, was einer Steigerung des Anteils ängstlicher Personen von ungewichteter zu gewichteter Schätzung von 8,1 % entspricht. Der Populationsparameter μ beträgt 26,2 %. Damit ist die kalibrierte Schätzung $\bar{y}_w$ mit einer Abweichung von 0,6 % deutlich genauer als die unkalibrierte Schätzung $\bar{y}_u$ mit einer Abweichung von –1,4 %.

Häufig wird übersehen, dass sich durch Gewichtung der Gesamtfehler gemessen als MSE (siehe Gleichung 2) erhöhen kann. Dies ist dann der Fall, wenn der gewichtungsbedingte Präzisionsgewinn durch Biasreduktion durch eine ebenfalls gewichtungsbedingte Varianzinflation des Schätzers übertroffen wird (Kish 1965, 424–433; Elliot/Little 2000, 192). Daher können ungewichtete, verzerrte Schätzer bessere Ergebnisse liefern als gewichtete, unverzerrte Schätzer.[80] Damit ist insbesondere dann zu rechnen, wenn die Gewichte selbst eine große Streuung aufweisen oder es sich um kleine Stichproben handelt (Elliot/Little 2000, 192).[81]

Für ungewichtete Daten kann der MSE durch

$$MSE(\bar{y}_u) = S^2\left(1 + \frac{B^2}{S^2}\right) = S^2\left(1 + \left(\frac{\bar{y}_u - \bar{y}_w}{\sigma_{y_u}}\right)^2\right) \qquad (25)$$

79 Als Variable wurde der Standardindikator zur Kriminalitätsfurchtmessung verwendet und vor der Analyse dichotomisiert.

80 Dieser als *Bias-Variance-Tradeoff* bekannte Effekt ist dadurch erklärbar, dass sich die Varianz eines Schätzers, zum Beispiel des Mittelwerts, durch die Gewichtung von S^2/n auf $S^2\left(1 + s_k^2/\bar{k}^2\right)/n$ erhöht (Kish 1992). Hierbei stellt s_k^2 die Varianz der Gewichte und $\bar{k}$ den Mittelwert der Gewichte dar.

81 Aus diesem Grund werden in der Praxis Gewichte häufig numerisch begrenzt („getrimmt“), sodass kein Fall z. B. ein Gewicht über 3 oder 10 erhält. Der Vollständigkeit halber soll erwähnt werden, dass vor allem bei Kalibrierungen auch negative Gewichte entstehen können. Da dies in der Regel schwer zu vermitteln ist, werden Gewichte daher häufig auch nach unten begrenzt.

berechnet werden. Für die gewichteten Daten kann die Varianz über

$$Var(\bar{y}_w) = S^2(1 + L) = S^2\left(1 + \frac{s_k^2}{\bar{k}^2}\right) \tag{26}$$

mit dem quadrierten Variationskoeffizienten (L) der Gewichte (k) berechnet werden (Kish 1992, 191).

Für das vorherige fiktive Beispiel mit den Daten des BCS ergibt sich ein Wert für $Var(\bar{y}_w)$ von $1 + L = 1.11$ und entsprechend ein Faktor für den $MSE(\bar{y}_u)$ von $1 + (B/S)^2 = 5{,}89$. Da bessere Schätzungen kleinere MSE-Werte aufweisen, sollten die Daten hier gewichtet werden.[82]

12 Empfehlungen

Die Gesamtkosten, die Probleme und die Dauer von Befragungen werden von Laien zumeist unterschätzt (Schnell 2012, 24–25). Das gilt insbesondere für die notwendigen Stichprobengrößen für Viktimisierungsstudien und die Maßnahmen zur Reduktion von Nonresponse.

Die tatsächlichen Konfidenzintervalle sind zumeist sehr viel größer, als es naive Analysen erwarten lassen. Aus diesem Grund sind vorbildliche Studien wie der NCVS und der BCS von beeindruckender Größe. Der British Crime Survey (nun *Crime Survey for England and Wales*, CSEW) umfasste 2012/2013 insgesamt 37.759 Personen; der NCVS lag 2013 bei 90.630 Haushalten mit 160.040 Personen. Entsprechend den hohen methodischen Anforderungen an Viktimisierungsstudien sind solche Surveys kostspielig: Die jährlichen Kosten des NCVS wurden 2012 auf 27 Millionen Dollar geschätzt.[83] Diese Kosten ergeben sich vor allem durch die zusätzlichen Maßnahmen zur Verhinderung von Nonresponse, vor allem durch aufwendige Mehrfachkontakte in verschiedenen Erhebungsmodi, Verweigerungskonvertierungen und den Einsatz sicherer Incentives.

---

82 Natürlich ist auch hier zusätzlich der in Kapitel 8 auf komplexe Stichprobendesigns zurückgehende Präzisionsverlust relevant.

83 Bereinigt man dies um die Einwohnerzahl (Faktor 0,25) und das Bruttosozialprodukt pro Kopf (Faktor 0,85) sowie den Währungskurs (Faktor 0,80), dann entspräche dies 4,6 Millionen Euro; in dieser Größenordnung dürfte auch der BCS liegen. Deutschland leistet sich mit dem Mikrozensus eine Erhebung von 830.000 Personen in ca. 370.000 Haushalten mit geschätzten Kosten von 21,6 Millionen Euro (Bundestagsdrucksache 17/10041 vom 19. 06. 2012). Eine Erhebung der Größe des NCVS oder des BCS/CSEW ist in Deutschland unter den gegebenen politischen Bedingungen kaum durchsetzbar.

Es gibt keine Möglichkeit, mit der solche Kosten vermieden werden können, falls man belastbare Aussagen für politisch interessante Subgruppen (wie Bundesländer oder Personen mit Migrationshintergrund) treffen möchte. Um das nochmals deutlich zum Ausdruck zu bringen: Es gibt weder statistische Zaubertricks noch „moderne Erhebungsmethoden", die mit kleineren Kosten vergleichbar präzise Resultate wie sehr große, langwierige und aufwendige Surveys produzieren können. Dies wird sich keineswegs im Laufe der Zeit bessern – eher im Gegenteil.

Wir raten daher im Zweifelsfall eher von kleineren Erhebungen ab. Belastbare Aussagen lassen sich mit kommunalen Studien (gar durch schriftliche Befragungen oder Lehrforschungsprojekte) mit geringem Aufwand nicht erzielen. Die Zukunft wissenschaftlicher Viktimisierungsstudien sind wenige, sehr große und methodisch sehr aufwendige Erhebungen, die ein einzelnes Institut nicht leisten kann. Die für solche Erhebungen notwendigen finanziellen Mittel werden die Zahl solcher Studien erheblich reduzieren; vermutlich wird sich Deutschland eine solche Studie nur in größeren Intervallen leisten wollen.

Trotz dieser Überlegungen gibt es Fragestellungen, für die quantitative Vorstudien sinnvoll sein können. Hierzu gehören vor allem Erhebungen in Institutionen und die Erhebung von Spezialpopulationen.[84] Empfehlungen für die Durchführung solcher Erhebungen haben wir im Folgenden zusammengefasst.

– Erhebung in einer Institution: Verwendung einer Liste der Insassen, der dort arbeitenden oder wohnenden Personen. Uneingeschränkte Zufallsauswahl, bei sehr kleinen Institutionen Vollerhebung. Bei besonderem Interesse an kleinen Teilgruppen geschichtete Auswahl. Schriftliche Befragung mit mehrfacher Mahnung.

– Spezialpopulationen: Liegt eine Liste der Mitglieder der Population vor, sollte eine uneingeschränkte Zufallsstichprobe gezogen werden, wobei der Erhebungsmodus den Kosten entsprechend gewählt werden kann. Liegt keine Liste vor, müssen spezielle Verfahren eingesetzt werden; die Durchführung solcher Verfahren sollte delegiert werden. Von der Verwendung kumulierter Screening-Interviews in Telefonbussen oder Schneeballverfahren sollte abgesehen werden.

---

[84] Die Ziehung von Zufallsstichproben seltener Populationen (konventionell weniger als 5 % oder 1 % der „allgemeinen" Bevölkerung) ist ein eigenes Gebiet innerhalb der Stichprobenverfahren. Eine Übersicht findet sich bei Schnell u. a. 2013.

– Kommunale Erhebung: Einwohnermelderegister als Auswahlgrundlage, Stichprobe größer als 2.000, uneingeschränkte Zufallsauswahl. Schriftliche Befragung mit mehrfacher Mahnung. Nonresponse-Stichprobe mit Face-to-Face-Erhebung zur Kontrolle des Nonresponse-Bias. Aufwand mindestens zwei Personenjahre.

– Bundesweite Erhebung: Stichprobenziehung und Datenerhebung an ein großes sozialwissenschaftliches Institut delegieren. Keine Websurveys, schriftliche Befragungen kaum empfehlenswert. CATI als Erstmodus, Dual-Frame-Stichprobe größer als 2.000, Nonresponse-Stichprobe (Face-to-Face) dringend empfohlen. Reine Erhebungskosten oberhalb von 150.000 Euro.

13 Literatur

AAPOR (2011): Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. Lenexa, KS: American Association for Public Opinion Research.

AAPOR (2010): AAPOR Report on Online Panels. Prepared for the AAPOR Executive Council by a Task Force operating under the auspices of the AAPOR Standards Committee, with members including: Reg Baker, Stephen Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, Mike Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, Sunghee Lee, Paul J. Lavrakas, Michael Link, Linda Piekarski,Kumar Rao, Douglas Rivers, Randall K. Thomas, Dan Zahs. In: Public Opinion Quarterly, 74, S. 711–781.

Aebi, Marcelo F.; Linde, Antonia (2014): National Victimization Surveys. In: Bruinsma, Gerben; Weisburd, David (Hg.): Encyclopedia of Criminology and Criminal Justice. New York: Springer, S. 3228–3242.

Ahlborn, Wilfried; Böker, Fred und Lehnick, Dirk (1993): Stichprobengrößen bei Opferbefragungen in der Dunkelfeldforschung, Wiesbaden: Bundeskriminalamt.

Albers, Ines (1997): Einwohnermelderegister-Stichproben in der Praxis. In: Gabler, Siegfried; Hoffmeyer-Zlotnik, Jürgen H. P. (Hg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 117–126.

Anderson, Andy B.; Basilevsky, Alexander und Hum, Derek P. I. (1983): Missing Data. A Review of the Literature. In: Rossi, Peter H.; Wright, James D. und Anderson, Andy B. (Hg.): Handbook of Survey Research. New York: Academic Press, S. 415–494.

Bandilla, Wolfgang; Kaczmirek, Lars; Blohm, Michael; Neubarth, Wolfgang; Jackob, Nikolaus; Schoen, Harald und Zerback, Thomas (2009): Coverage- and Nonresponse-Effekte bei Online-Bevölkerungsumfragen. In: Jackob, Nikolaus; Schoen, Harald und Zerback, Thomas (Hg.): Sozialforschung im Internet. Methodologie und Praxis der Online-Befragung. Wiesbaden: VS-Verlag, S. 129–144.

Bentrup, Christina; Verneuer, Lena (2014): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2011. In: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden, Nr. 20/2014.

Bethlehem, Jelke; Cobben, Fannie und Schouten, Barry (2011): Handbook of Nonresponse in Household Surveys. Hoboken: Wiley.

Biemer, Paul P. (2010): Total Survey Error. Design, Implementation, and Evaluation. In: Public Opinion Quarterly, 74, S. 817–848.

Biemer, Paul P.; Lyberg, Lars E. (2003): Introduction to Survey Quality, Hoboken: Wiley.

Bortz, Jürgen (2005): Statistik für Human- und Sozialwissenschaftler, 6. Aufl. Heidelberg: Springer.
Bruinsma, Gerben; Weisburd, David (Hg.) (2014): Encyclopedia of Criminology and Criminal Justice. New York: Springer.
Bundesarbeitsgemeinschaft Wohnungslosenhilfe (2013): Zahl der Wohnungslosen in Deutschland weiter gestiegen. Pressemitteilung. URL: http://www.bagw.de/media/doc/PRM_2013_08_01_Zahl_der_Wohnungslosen.pdf. – Download vom 09. 04. 2015.
CLANDESTINO Project (2009): Undocumented Migration: Counting the Uncountable. Data and Trends Across Europe. Final Report. Athens.
Cohen, Lawrence E.; Felson, Marcus (1979): Social Change and Crime Rate Trends. A Routine Activity Approach. In: American Sociological Review, 44, S. 588–608.
Converse, Philip E.; Traugott, Michael W. (1986): Assessing the Accuracy of Polls and Surveys. In: Science, 234, S. 1094–1098.
Deville, Jean-Claude; Särndal, Carl-Erik (1992): Calibration Estimators in Survey Sampling. In: Journal of the American Statistical Association, 87, S. 376–382.
Doblhammer, Gabriele; Schulz, Anne; Steinberg, Juliane und Ziegler, Uta (2012): Demografie der Demenz. Bern: Verlag Hans Huber.
Elliot, Michael R.; Little, Roderick J. A. (2000): Model-based Alternatives to Trimming Survey Weights. In: Journal of Official Statistics, 16, S. 191–209.
Enders, Craig K. (2010): Applied Missing Data Analysis. New York: Guilford Press.
Enzmann, Dirk (2010): Germany. In: Junger-Tas, Josine; Marshall, Ineke H.; Enzmann, Dirk; Killias, Martin; Steketee, Majone und Gruszczynska, Beate (Hg.): Juvenile Delinquency in Europe and Beyond. Results of the Second International Self-reported Delinquency Study. Dordrecht: Springer, S. 47–64.
Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris und Tutz, Gerhard (2007): Statistik, 6. Aufl. Berlin/Heidelberg: Springer.
Geiger, Marion; Styhler, Doris (2012): ZENSUS 2011: Erhebungsteil Sonderbereiche. In: Bayern in Zahlen, 5, S. 280–285.
Glemser, Axel; Meier, Gerd und Heckel, Christiane (2014): Dual-Frame: Stichprobendesign für CATI-Befragungen im mobilen Zeitalter. In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 167–190.
Groves, Robert M. (1989): Survey Errors and Survey Costs. New York: Wiley.

Groves, Robert M.; Cork, Daniel L. (Hg.) (2008): Surveying Victims. Options for Conducting the National Crime Victimization Survey, Washington: The National Academies Press.

Groves, Robert M.; Magilavy, Lou J. (1986): Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. In: Public Opinion Quarterly, 50, S. 251–266.

Groves, Robert M.; Peytcheva, Emilia (2008): The Impact of Nonresponse Rates on Nonresponse Bias. A Meta-analysis. In: Public Opinion Quarterly, 72, S. 167–189.

Häder, Sabine; Gabler, Siegfried und Heckel, Christiane (2009): Stichprobenziehung für die CELLA-Studie. In: Häder, Michael; Häder, Sabine (Hg.): Telefonbefragungen über das Mobilfunknetz. Konzept, Design und Umsetzung einer Strategie zur Datenerhebung. Wiesbaden: VS-Verlag, S. 21–49.

Haug, Sonja (2008): Sprachliche Integration von Migranten in Deutschland. Bundesamt für Migration und Flüchtlinge, Workingpaper 14.

Heckel, Christiane; Glemser, Alex und Meier, Gerd (2014): Das ADM-Telefonstichproben-System. In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 137–166.

Heckel, Christiane; Wiese, Kathrin (2012): Sampling Frames for Telephone Surveys in Europe. In: Häder, Sabine; Häder, Michael und Kühne, Mike (Hg.): Telephone Surveys in Europe. Research and Practice. Berlin/Heidelberg: Springer, S. 103–119.

von der Heyde, Christian (2014): Einwohnermeldeamts-Stichproben (EWA-Stichproben). In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 191–195.

Hindelang, Michael J.; Gottfredson, Michael R. und Garofalo, James (1978): Victims of Personal Crime. An Empirical Foundation for a Theory of Personal Victimization. Cambridge: Ballinger.

Hunsicker, Stefan; Schroth, Yvonne (2014): Dual-Frame-Ansatz in politischen Umfragen. Arbeitspapiere der Forschungsgruppe Wahlen e. V., Mannheim, Nr. 2 – April 2014.

Johnson, Brian (2014): Sample Selection Models. In: Bruinsma, Gerben; Weisburd, David (Hg.): Encyclopedia of Criminology and Criminal Justice. New York: Springer, S. 4561–4580.

Kalton, Graham (1983): Introduction to Survey Sampling, Newbury Park: Sage.

Kish, Leslie (1965): Survey Sampling, New York: Wiley.Kish, Leslie (1992): Weighting for Unequal P_i. In: Journal of Official Statistics, 8, S. 183–200.

Klausch, Thomas; Hox, Joop und Schouten, Barry (2015): Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population. In: Journal of the Royal Statistical Society, Series A (Statistics in Society), Early View (Online Version of Record published before inclusion in an issue). URL: http://onlinelibrary.wiley.com/doi/10.1111/rssa.12102/full – Download vom 09. 04. 2015.

Klingwort, Jonas (2014): Nonresponse in aktuellen deutschen Viktimisierungssurveys, Bachelorarbeit. Universität Duisburg-Essen: Institut für Soziologie.

Kreuter, Frauke (Hg.) (2013): Improving Surveys with Paradata – Analytic Uses of Process Information. Hoboken: Wiley.

Kroh, Martin (2014): Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2012). SOEP Survey Papers 177: Series D. Berlin: DIW/SOEP.

Kruskal, William; Mosteller, Frederick (1979a): Representative Sampling, I: Non-scientific Literature. In: International Statistical Review, 47, S. 13–24.

Kruskal, William; Mosteller, Frederick (1979b): Representative Sampling, II: Scientific Literature, Excluding Statistics. In: International Statistical Review, 47, S. 111–127.

Kruskal, William; Mosteller, Frederick (1979c): Representative Sampling, III: the Current Statistical Literature. In: International Statistical Review, 47, S. 245–265.

Kruskal, William; Mosteller, Frederick (1980): Representative Sampling, IV: the History of the Concept in Statistics 1895–1939. In: International Statistical Review, 48, S. 169–195.

Lepkowski, James M. (1988): Telephone Sampling Methods in the United States. In: Groves, Robert M.; Biemer, Paul P.; Lyberg, Lars E.; Massey, James T.; Nicholls, William L. und Waksberg, Joseph (Hg.): Telephone Survey Methodology. New York: Wiley, S. 73–98.

Link, Michael W.; Fahimi, Mansour (2008): Telephone Survey Sampling. In: Levy, Paul S.; Lemeshow, Stanley (Hg.): Sampling of Populations. Methods and Applications, 4. Aufl. Hoboken: Wiley, S. 455–487.

Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical Analysis with Missing Data, 2. Aufl. Hoboken: Wiley.

Löffler, Ute; Behrens, Kurt und von der Heyde, Christian: (2014): Die Historie der ADM-Stichproben. In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 67–83.

Lohr, Sharon L. (2010): Sampling: Design and Analysis, 2. Aufl. Boston: Brooks/Cole.
Lynch, James P. (2006): Problems and Promise of Victimization Surveys for Cross-national Research. In: Crime and Justice, 34, S. 229–287.
Lynn, Peter (1997): Sampling Frame Effects on the British Crime Survey. In: Journal of the Royal Statistical Society, Series A (Statistics in Society), 160, S. 253–269.
Mayer, Raimund (2013): Zensus 2011: Erhebung an Anschriften mit Sonderbereichen. In: Statistische Monatshefte Niedersachsen, 12, S. 672–679.
Mitofsky, Warren (1970): Sampling of Telephone Households, unveröffentlichtes CBS-News-Memorandum.
National Research Council (2012): Small Populations, Large Effects: Improving the Measurement of the Group Quarters Population in the American Community Survey, Washington DC: The National Academies Press.
National Research Council (2014): Estimating the Incidence of Rape and Sexual Assault, Washington DC: The National Academies Press.
Noack, Marcel (2015): Methodische Probleme bei der Messung von Kriminalitätsfurcht und Viktimisierungserfahrungen, Wiesbaden: Springer VS.
Pöge, Andreas (2007): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2005 (Vier-Wellen-Panel). In: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden, Nr. 13/2007.
Puhani, Patrick A. (2000): The Heckman Correction for Sample Selection and Its Critique. In: Journal of Economic Surveys, 14, S. 53–68.
Särndal, Carl-Erik; Lundström, Sixten (2005): Estimation in Surveys with Nonresponse. Chichester: Wiley.
Schafer, Joseph L. (1997): Analysis of Incomplete Multivariate Data. Boca Raton: Chapman & Hall/CRC.
Schnell, Rainer (1986): Missing-Data-Probleme in der empirischen Sozialforschung. Dissertation. Ruhr-Universität Bochum.
Schnell, Rainer (1991): Wer ist das Volk? Zur faktischen Grundgesamtheit bei „allgemeinen Bevölkerungsumfragen“: Undercoverage, Schwererreichbare und Nichtbefragbare. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43, S. 106–137.
Schnell, Rainer (1993): Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität“ und Gewichtungsverfahren. In: Zeitschrift für Soziologie, 22, S. 16–32.
Schnell, Rainer (1997): Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen, Opladen: Leske+Budrich.

Schnell, Rainer (1998): Besuchs- und Berichtsverhalten der Interviewer. In: Statistisches Bundesamt (Hg.): Interviewereinsatz und -qualifikation, Nummer 11 in Spektrum Bundesstatistik. Stuttgart: Metzler-Poeschel, S. 156–170.

Schnell, Rainer (2008): Avoiding Problems of Traditional Sampling Strategies for Household Surveys in Germany: Some New Suggestions, DIW Data-Documentation 33, Berlin.

Schnell, Rainer (2012): Survey-Interviews. Methoden standardisierter Befragungen, Wiesbaden: VS-Verlag.

Schnell, Rainer; Hill, Paul. B. und Esser, Elke (2013): Methoden der empirischen Sozialforschung, 10. Aufl. München: Oldenbourg.

Schnell, Rainer; Hoffmeyer-Zlotnik, Jürgen H. P. (2002): Methodik für eine regelmäßige Opferbefragung, Gutachten im Auftrag des BMI/BMJ, Universität Konstanz.

Schnell, Rainer; Kreuter, Frauke (2000): Untersuchungen zur Ursache unterschiedlicher Ergebnisse sehr ähnlicher Viktimisierungssurveys. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 52, S. 96–117.

Schnell, Rainer; Kreuter, Frauke (2005): Separating Interviewer and Sampling-point Effects. In: Journal of Official Statistics, 21, S. 389–410.

Schnell, Rainer; Noack, Marcel (2014): The Accuracy of Pre-Election Polling of German General Elections. In: MDA – Methods, Data, Analyses, 8, S. 5–24.

Schonlau, Matthias; Weidmer, Beverly und Kapteyn, Arie (2014): Recruiting an Internet Panel Using Respondent-driven Sampling. In: Journal of Official Statistics, 30, S. 291–310.

Schouten, Barry; Cobben, Fannie und Bethlehem, Jelke (2009): Indicators for the Representativeness of Survey Response. In: Survey Methodology, 35, S. 101–113.

Schwarz, Norbert (2007): Retrospective and Concurrent Self-reports. The Rationale for Real-time Data Capture. In: Stone, Arthur A.; Shiffman, Saul; Atienza, Audie A. und Nebeling, Linda (Hg.): The Science of Real-time Data Capture. Self-reports in Health Research. New York: Oxford University Press, S. 11–26.

Shaffer, Jennifer N.; Ruback, R. Barry (2002): Violent Victimization as a Risk Factor for Violent Offending Among Juveniles, Washington DC: Office of Juvenile Justice and Delinquency Prevention.

Smith, Adrian (2006): Crime Statistics. An Independent Review. Carried out for the Secretary of State for the Home Department, London: Home Office.

Statistische Ämter des Bundes und der Länder (2004): Ergebnisse des Zensustests. In: Wirtschaft und Statistik, 8, S. 813–833.

Statistische Ämter des Bundes und der Länder (Hg.) (2014): Gebäude- und Wohnungsbestand in Deutschland. Erste Ergebnisse der Gebäude- und Wohnungszählung 2011, Hannover: Landesamt für Statistik Niedersachsen.

Stolzenberg, Ross M.; Relles, Daniel A. (1997): Tools for Intuition about Sample Selection Bias and Its Correction. In: American Sociological Review, 62, S. 494–507.

Tucker, Clyde (1983): Interviewer Effects in Telephone Surveys. In: Public Opinion Quarterly, 47, S. 84–95.

Valliant, Richard; Dever, Jill A. und Kreuter, Frauke (2013): Practical Tools for Designing and Weighting Survey Samples. New York: Springer.

Vaughn, Michael G.; Maynard, Brandy; Salas-Wright, Christopher; Perron, Brian E. und Abdon, Arnelyn (2013): Prevalence and Correlates of Truancy in the US. Results from a National Sample. In: Journal of Adolescence, 36, S. 767–776.

Vella, Francis (1998): Estimating Models with Sample Selection Bias. A Survey. In: Journal of Human Resources, 33, S. 127–169.

Vercambre, Marie-Noël; Gilbert, Fabien (2012): Respondents in an Epidemiologic Survey Had Fewer Psychotropic Prescriptions than Nonrespondents. An Insight into Health-related Selection Bias Using Routine Health Insurance Data. In: Journal of Clinical Epidemiology, 65, S. 1181–1189.

Vogel, Dita; Aßner, Manuel (2011): Umfang, Entwicklung und Struktur der irregulären Bevölkerung in Deutschland. Expertise im Auftrag der deutschen nationalen Kontaktstelle für das Europäische Migrationsnetzwerk (EMN) beim Bundesamt für Migration und Flüchtlinge. URL: http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/EMN/Expertisen/emn-wp-41-expertise-de.pdf – Download vom 09. 04. 2015.

Waksberg, Joseph (1978): Sampling Methods for Random Digit Dialing. In: Journal of the American Statistical Association, 19, S. 103–113.

Weisberg, Herbert F. (2005): The Total Survey Error Approach. A Guide to the New Science of Survey Research, Chicago: The University of Chicago Press.

Weissenberger-Leduc, Monique; Weiberg, Anja (2011): Gewalt und Demenz. Ursachen und Lösungsansätze für ein Tabuthema in der Pflege. Wien: Springer.

Weyerer, Siegfried (2005): Altersdemenz. In: Gesundheitsberichterstattung des Bundes, 28, S. 1–33.

Wolter, Kirk M. (2007): Introduction to Variance Estimation, 2. Aufl. New York: Springer.

Ziegler, Uta; Doblhammer, Gabriele (2009): Prävalenz und Inzidenz von Demenz in Deutschland. In: Das Gesundheitswesen, 71, S. 281–290.

Effekte des Erhebungsmodus

Helmut Kury, Nathalie Guzy und Heinz Leitgöb

1 Einleitung

Ein wesentlicher Bestandteil der empirischen Sozialforschung beruht auf Meinungsumfragen bzw. allgemein auf Befragungen (Scheuch 1973, 66), die seit den frühen 1940er Jahren zur dominierenden Methode der Datensammlung in diesem Forschungsbereich wurden (Hippler/Schwarz 1987, 102) und die bald eine geradezu vorherrschende Stellung einnahmen. In diesem Kontext wurden auch in der Kriminologie und Viktimologie Opferbefragungen längst zu Standardmethoden der Erkenntnisgewinnung (Kaiser u. a. 1991). Seit den 1960er Jahren werden weltweit, vorwiegend in den westlichen Industrieländern, vermehrt Opferbefragungen durchgeführt. Zu deren Verbreitung trugen nach einigen Vorläufern – auch zur Methodik (vgl. etwa Ennis 1967; zusammenfassend Kury 1992, 142ff.) – vor allem die von der US-President's Commission on Law Enforcement and Administration of Justice (1967) in den USA seit 1972 ständig durchgeführten groß angelegten Umfragen zur Kriminalitätssituation bei. Sparks (1981, 4) bezeichnet die vorbereitenden Umfragen (auch Biderman 1967, Reiss 1967), gerade auch wegen ihrer methodischen Qualität, zu Recht als „landmarks in the study of crime", die „a substantial impact on academic criminology" hatten (5). Bald begannen weitere Länder nach dem Vorbild der USA ebenfalls regelmäßige Opfer-/Dunkelfeld-Befragungen durchzuführen. „Few innovations in the social sciences rival the importance of the modern survey" (Presser 1984, 93). Deutschland schloss sich diesem Trend – bis heute – zwar nicht an, was immer wieder bedauert wurde, begann jedoch in den 1970er Jahren ebenfalls große, allerdings zunächst regional begrenzte, Opferstudien durchzuführen, organisiert etwa in einer fruchtbaren Zusammenarbeit zwischen dem Bundeskriminalamt und dem Max-Planck-Institut für ausländisches und internationales Strafrecht (Stephan 1976; Villmow/Stephan 1983). In den 1980er Jahren gab es einen weiteren großen Fortschritt mit dem Aufkommen international vergleichender Studien (Hofer 2009). 1988 wurde der erste „International Crime and Victimization Survey – ICVS" durchgeführt, an dem sich auch Deutschland mit der ersten in ganz Westdeutschland durchgeführten Opferbefragung beteiligte (Kury 1991; van Dijk 2009). Ein Jahr später folgte, durchgeführt von denselben beiden Forschungsgruppen, die erste deutsch-deutsche Opferstudie (Kury u. a. 1992).

Mit dem raschen Aufkommen von Umfragen („Victim Surveys“) wurde jedoch bald und nicht zu Unrecht auch mehr und mehr Kritik an der Methodik der Befragungen laut. Wesentlich zur Verbreitung der Umfrageforschung trug deren – allerdings nur scheinbar – leichte Anwendbarkeit und Durchführbarkeit bei. Bereits Sparks (1981, 7) bemängelte an den von ihm hochgelobten US-Crime Surveys, dass sie mit ungebührlicher Hast eingeführt worden seien. Schon in den 1930er und 40er Jahren begann die Diskussion zur Verweigererproblematik und den Möglichkeiten, damit umzugehen, insbesondere die Antwortquote zu erhöhen. Yates (1933) hat als einer der ersten Techniken zum Umgang mit fehlenden Werten aufgrund von Verweigerungen entwickelt (Cochran 1983). Madow u. a. (1983) gaben einen ersten umfassenden Überblick zu Problemen von „incomplete data in sample surveys“.

Die Methodenkritik an Umfragen kam – zu Recht – nicht zum Erliegen, da man es sich vielfach „zu einfach“ machte (vgl. bereits McNemar 1946). Selg u. a. (1992, 71) betonen in diesem Zusammenhang, dass Befragungen „selten mit der Sorgfalt durchgeführt (werden), die eine ernstzunehmende wissenschaftliche Arbeit verlangt“. Rosenthal beschäftigte sich bereits in den 1960er Jahren, als das persönliche Interview als der „Königsweg“ der praktischen Sozialforschung angesehen wurde (König 1962, 328), intensiv mit dem Einfluss der Interviewerinnen und Interviewer auf Umfrageergebnisse (Rosenthal/Rosnow 1969). Vor dem Hintergrund dieser Kritik wurden in den USA bereits früh differenzierte methodische Standards für die Umfrageforschung veröffentlicht, etwa von Dodd (1947), der 41 verschiedene Bereiche anspricht, wie Stichprobengewinnung, Interviewerauswahl und -überwachung oder Fragebogenentwicklung. Um eine größere Vergleichbarkeit zu erreichen setzte er sich etwa für eine Standardisierung von Surveys ein – Bemühungen, die bis heute anhalten (etwa Kury/Obergfell-Fuchs 2003), jedoch kaum zu Erfolg führten. Kritikpunkte wurden unterschiedlich aufgegriffen, die Erhebungsinstrumente etwa veränderten sich mehr, als dass sie einheitlich wurden. So wird etwa im „European Sourcebook of Crime and Criminal Justice Statistics 2014“ (Aebi u. a. 2014, 341) hinsichtlich der berichteten internationalen Ergebnisse aus Opferstudien betont: „Readers must keep in mind that the results of national victimization surveys conducted in different countries cannot be compared because their methodology differs“ (vgl. etwa auch Ahlf 2007, 528).

Wie Kreuter (2002, 55) hinsichtlich der Erfassung von Verbrechensfurcht mit den vielfach kritisierten „Standardindikatoren“ betont, scheint es heute zum „‘guten Ton‘ zu gehören, die mangelnde Qualität der Indikatoren zu beklagen und einen der beiden dann trotzdem in die Studien aufzunehmen“. Es interessieren die vielfach in den Medien verbreiteten Ergebnisse, Methodenfragen treten dann in der Regel in den Hintergrund.

Der folgende Beitrag beschäftigt sich mit einem der bekanntesten und vergleichsweise gut erforschten Methodeneffekten in Umfragen: dem Einfluss des Erhebungsverfahrens („Modeeffekt"). So liegt mittlerweile eine Vielzahl an Untersuchungen vor, die belegen, dass das Erhebungsverfahren einen erheblichen Einfluss auf die Ergebnisse von Umfragen ausüben kann. Vergleichsweise geringe Kenntnisse sind in diesem Kontext allerdings über entsprechende Effekte in Opferbefragungen bekannt. Es finden sich nur vereinzelt Beiträge, die verschiedene, durch den Erhebungsmodus induzierte Einflüsse auf die Ergebnisse in Opferbefragungen untersuchen. Dieses Wissen ist jedoch nicht nur für Forscherinnen und Forscher und die adäquate Auswahl eines geeigneten Erhebungsverfahrens wichtig, sondern auch für eine angemessene Bewertung der Ergebnisse von Opferbefragungen.

Im Folgenden wird zunächst auf die Definition und Ätiologie von Modeeffekten eingegangen, wobei die notwendige Differenzierung zwischen einer „engen" und „weiten" Definition von Mode-Effekten vorgestellt wird (Abschnitt 2). Anschließend werden Einflüsse des Erhebungsmodus auf den Erhebungs- und Antwortprozess, wie Auswahlgrundlage und Stichprobendesign, Nonresponse und Messfehler, dargestellt (Abschnitt 3). In weiterer Folge werden die wesentlichen Erhebungsmodi, telefonische, postalische und persönliche Befragung, im Kontext möglicher Einflussfaktoren dargestellt (Abschnitt 4). Dabei wird ein besonderer Fokus auf relevante Erkenntnisse aus dem Bereich von Opferbefragungen gelegt und insbesondere der Forschungsstand zum Einfluss des Erhebungsmodus auf die Messung von Prävalenzen vorgestellt.[1]

Der Beitrag endet schließlich mit Empfehlungen für ein Vorgehen (Abschnitt 5) sowie zusammenfassenden Bullet Points (Abschnitt 6).

2 Definition und Ätiologie von Modeeffekten

2.1 Definition

Während sich der Begriff des *Modus der Datenerhebung* (auch *Erhebungsmodus* bzw. *-verfahren* genannt; *mode of data collection*, *survey administration mode*) auf die konkrete Art und Weise bezieht, wie Daten im Zuge einer Umfrage generiert werden bzw. der Fragebogen administriert wird (z. B. persönlich-mündlich bzw. face-to-face, schriftlich-postalisch, telefonisch, on-

[1] Neuere Erhebungsverfahren wie Online und/oder Websurveys bleiben bewusst unberücksichtigt, da diese in einem separaten Beitrag behandelt werden (Killias in diesem Band).

line- bzw. webbasiert), umfasst der Terminus *Modeeffekte* die Gesamtheit aller Einflüsse, die ein bestimmter Erhebungsmodus auf das *Antwortverhalten* von Befragten ausübt (Groves u. a. 2009; Jans 2008). In diesem Sinne können Modeeffekte als die Reaktivität des Antwortverhaltens auf einen Erhebungsmodus und aus methodologischer Perspektive als Messfehler verstanden werden, der zumindest für einen Teil der Divergenz zwischen dem „wahren Wert" und der tatsächlich gegebenen Antwort verantwortlich ist. Diese ausschließlich auf den Antwortprozess abstellende Definition von Modeeffekten soll in der Folge als „*enge Definition*" bezeichnet werden. Sie ist insofern als restriktiv zu charakterisieren, als sie den Umstand ignoriert, dass der Erhebungsmodus auch Einfluss auf andere Phasen des Umfrageprozesses nimmt, so z. B. auf die Konstruktion des Fragebogens, die Auswahlgrundlage („Sampling Frame"), das Stichprobendesign oder die Kontaktierung und Rekrutierung der Befragten. In der Konsequenz muss somit davon ausgegangen werden, dass die verschiedenen Erhebungsverfahren neben variierenden Messfehlern auch mit unterschiedlich ausgeprägten Coverage-, Stichproben- sowie Nonresponsefehlern behaftet sind.[2] Dieses holistische Verständnis von Modeeffekten sei als „*weite Definition*" eingeführt.

2.2 Ätiologie

Die Ursachen von Modeeffekten nach der engen Definition (d. h. auf das Antwortverhalten selbst) sind auf Unterschiede entlang einer Reihe von Dimensionen zurückzuführen. Während de Leeuw (1992, 2008) drei Klassen von zentralen Faktoren (*interviewer effects*, *media related factors*, *factors influencing information transmission*) herausarbeitet hat, etablierten sich – darauf aufbauend – in den letzten Jahren folgende fünf Ursachen von Modeeffekten (Groves u. a. 2009; Jans 2008):

(i) *Ausmaß der Beteiligung von Interviewerinnen und Interviewern (degree of interviewer involvement*)

Datenerhebungsverfahren unterscheiden sich hinsichtlich des Grades der Einbindung von Interviewerinnen und Interviewern voneinander. Während postalische bzw. onlinebasierte Befragungen in aller Regel selbstadministriert – d. h. ohne Interviewende – ablaufen, steigt der Grad der Interviewerbetei-

[2] Unter „Coverage-Error" wird der Fehler bezeichnet, der durch eine nicht-adäquate Deckung der eigentlich interessieren Grundgesamtheit und der zur Verfügung stehenden Auswahlgrundlage zur Ziehung der Stichprobe entsteht. Als „Nonresponse Error" wird der Fehler bezeichnet, der durch Antwortverweigerungen entsteht (Groves u. a. 2009)

ligung von selbstadministrierten Befragungen unter Leitung von Interviewerinnen und Interviewern (für Rückfragen) über telefonbasierte Interviews bis hin zu persönlichen (face-to-face) Interviews (zur empirischen Befundlage siehe etwa Schaeffer u. a. 2010, 450f.).

Eine Interviewerbeteiligung kann die Qualität von Umfrageergebnissen sowohl erhöhen als auch reduzieren (Groves u. a. 2009, 154). So können Interviewerinnen und Interviewer einerseits wichtige Aufgaben wie die Motivierung von Befragten, die Bereinigung von Unklarheiten, die Kompensation von limitierten Lese- und schriftsprachlichen Ausdrucksfähigkeiten übernehmen (siehe etwa de Leeuw 2008; Weisburd 2005), andererseits kann durch die Anwesenheit einer Interviewerin bzw. eines Interviewers das Antwortverhalten aber auch in negativer Weise beeinflusst werden und – insbesondere bei sensiblen bzw. emotionalen Fragen (siehe etwa Lensvelt-Mulders 2008; Tourangeau u. a. 2000; Tourangeau/Yan 2007) – verstärkt zu sozial erwünschten Antworten (social desirability) bzw. bewussten Falschangaben (misreporting) führen (siehe dazu auch den Beitrag von Waubert de Puiseau u. a. in diesem Band). Darüber hinaus können Interviewende bei Befragten einen Zwang zur Bereitstellung möglichst konsistenter Antwortmuster (consistency motif; siehe etwa Podsakoff u. a. 2003) hervorrufen. Für detailreiche allgemeine Ausführungen zu Interviewereffekten in Surveys sowie deren Begegnung durch ein gezieltes Interviewertraining sei auf Groves (2004), Kreuter (2008a), Lessler u. a. (2008) sowie Schaeffer u. a. (2010) verwiesen.

(ii) *Grad der Interaktion mit den Befragten (degree of interaction with the respondent)*

Der Grad der Interviewer-Befragten-Interaktion zielt auf das Ausmaß des Einflusses der Forscherinnen und Forscher auf den Befragungsprozess ab und ist in hohem Maße mit der Intensität der Einbindung von Interviewerinnen und Interviewern in diesen Prozess verwoben. Eine ausgeprägte Interviewer-Befragten-Interaktion, z. B. im Zuge eines persönlich-mündlichen Interviews, erlaubt den Forschenden – vermittelt über die geschulten und instruierten Interviewenden – eine stärkere Kontrolle der Erhebungssituation als bei einem selbstadministrierten Verfahren ohne Interviewende.

(iii) *Grad der Privatsphäre (degree of privacy)*

Datenerhebungsverfahren variieren auch hinsichtlich des Ausmaßes an Privatsphäre, die sie den Befragten bei der Beantwortung der Fragen gewähren. Der von den Befragten wahrgenommene Grad an Privatheit kann sowohl die Entscheidung für oder gegen die Teilnahme an einer Befragung erheblich beeinflussen, als auch, insbesondere bei sehr sensiblen Fragen, die Bereitschaft zur Bekanntgabe der korrekten Antwort entscheidend mitbestimmen.

(iv) *Kommunikationskanäle (channels of communication)*

Als vierte Dimension, entlang der sich Datenerhebungsverfahren grundlegend voneinander unterscheiden, müssen die zur Verfügung stehenden Kommunikationskanäle genannt werden. Diese beziehen sich auf die verschiedenen sinnesorganischen Möglichkeiten zur Wahrnehmung von Informationen sowie auf deren Vermittlung an die Umwelt. Während interviewerbasierte Erhebungsverfahren in der Regel auf der verbalen Kommunikation zwischen der bzw. dem Interviewenden und der bzw. dem Befragten beruhen, allerdings auch nichtverbale Kommunikation und Wahrnehmung mit einschließen können (wie spontanes oder zögerliches Antwortverhalten oder Kommentare zu einem vorliegenden Fragebogen) liegt selbstadministrierten Befragungen eine ausschließlich visuelle Informationsvermittlung zugrunde (die bzw. der Befragte liest eine Frage im postalischen bzw. webbasierten Fragebogen). Je höher die Beteiligung der Interviewerinnen und Interviewer, desto vielfältiger ist der Einsatz von unterschiedlichen Kommunikationskanälen.

(v) *Einsatz von Technologie (technology use)*

Letztlich lassen sich Verfahren der Datenerhebung noch nach der Art und dem Ausmaß des Einsatzes von (Erhebungs-)Technologien differenzieren (siehe exemplarisch für face-to-face Interviews Fuchs u. a. 2000). Allgemein gilt es zunächst festzuhalten, dass die letzten beiden Jahrzehnte durch eine rapide fortschreitende Entwicklung der computergestützten Erhebung von Umfragedaten gekennzeichnet sind. So hat sich bei face-to-face Befragungen das *computer assisted personal interviewing* (*CAPI)* und bei telefonischen Befragungen das *computer assisted telephone interviewing, (CATI)* weitgehend durchgesetzt. Während der Einsatz computergestützer Verfahren einerseits eine größere Gestaltungsfreiheit im Design der Erhebungsinstrumente erlaubt (z. B. durch die Integration von komplexen Filterführungen, die eine höhere Kontrolle über den Befragungsprozess gestattet oder aber die zufällige Anordnung von Items zur Vermeidung von Primacy oder Recency Effekten[3]), erfordern technologieintensive Erhebungsverfahren andererseits auch eine Reihe von Kenntnissen, da der Erfolg von computergestützen Datenerhebungsverfahren in erheblichem Maße von dem Zugang, der Akzeptanz und der Vertrautheit der Befragten mit der zum Einsatz gebrachten Technologie abhängt.

---

[3] Dabei handelt es sich um systematische Verzerrungen durch Antworttendenzen, demnach bei längeren Antwortskalen entweder die erste Antwortkategorie (Primacy Effekt) bzw. die letzte Antwortkategorie (Recency Effekt) bevorzugt wird.

3 Einflüsse des Erhebungsmodus auf den Erhebungs- und Antwortprozess

Ein hilfreiches Modell um die, gemäß der weiten Definition relevanten Modeeffekte innerhalb des Umfrageprozesses systematisch zu erfassen, stellt das „Survey Lifecycle"-Modell von Groves u. a. (2009) dar.

Das Modell unterscheidet zwei Gruppen von Fehlerquellen, einerseits die Messung selbst (*measurement*), andererseits die Repräsentativität der Daten (*representation).* Während auf der Ebene der Messung Fehler durch die Operationalisierung der Konstrukte (*Validity*), die Fragebeantwortung (*Measurement Error*) und die Dateneingabe (*Processing Error*) entstehen können, sind auf Repräsentativitätsebene Verzerrungen insbesondere durch die Auswahlgrundlage (*Coverage Error*), das Stichprobendesign (*Sampling Error*), fehlende Teilnahme (*Nonresponse Error*) und Gewichtungsprozeduren (*Adjustment Error*) zu berücksichtigen (Groves u. a. 2009).

Zwar besitzen alle Fehlerquellen grundsätzlich das Potenzial unterschiedliche Ergebnisse in Abhängigkeit des Erhebungsmodus zu produzieren, allerdings haben sich insbesondere die mit der Stichprobenziehung in Zusammenhang stehenden Effekte (*Coverage und Sampling Errors*), Fehler durch systematische Ausfallprozesse (*Nonresponse Error*) sowie die modebedingten Effekte auf die Messung bzw. den Antwortprozess als besonders relevant erwiesen (Groves u. a. 2009):

3.1 Effekte durch die Auswahlgrundlage und das Stichprobendesign (*Coverage und Sampling Error*)

Das Erhebungsverfahren ist in der Regel mit einer bestimmten Auswahlgrundlage verknüpft – sei es weil die Auswahlgrundlage das Erhebungsverfahren bestimmt oder umgekehrt. So werden telefonische Bevölkerungsbefragungen meist über die Auswahl (z. B. in Telefonbüchern) gelisteter oder zufällig generierter Telefonnummern (Random Digit Dialing, vgl. Häder/Gabler 1998) durchgeführt, während schriftliche und/oder persönlich-mündliche Befragungen zur Stichprobenziehung in der Regel auf offizielle Adressdaten (z. B. das Einwohnermelderegister) zurückgreifen. Jede Auswahlgrundlage zeichnet sich allerdings durch eine bestimmte Auswahlgesamtheit (d. h. Personen, die über die jeweilige Auswahlgrundlage überhaupt erreicht werden können) und in Folge auch durch einen speziellen „Coverage Error" aus. Da jedoch verfügbare Auswahlgrundlagen häufig Personen mit tendenziell erhöhten Viktimisierungsrisiken ausschließen, wie z. B. wohnsitzlose oder nicht gemeldete Personen, Personen in Anstalten (Pflegeheime, Gefängnisse) oder Personen ohne festen Telefonanschluss, sind – je nach Erhebungsverfahren –

unterschiedlich große Verzerrungen zu erwarten (vgl. auch Cantor/Lynch 2000; Griffin-Saphire 1984; Stangelang 1996; UNODC/UNECE 2011; Woltman u. a. 1980).

Darüber hinaus muss bedacht werden, dass die Auswahlgrundlage häufig an ein spezifisches Stichprobendesign gekoppelt ist, das wiederum mit bestimmten Stichprobenfehlern verbunden ist. So erlauben bspw. telefonische Befragungen meist keine (adäquate) Schichtung der Stichprobe (eine Ausnahme stellt die Schichtung über Vorwahlnummern dar), während gleichzeitig eine Auswahl von Befragungspersonen auf Ebene des kontaktierten Haushalts erforderlich ist. Dadurch handelt es sich bei Telefonbefragung regelmäßig um zweistufige Zufallsstichproben, die aufgrund der notwendigen Designgewichtung deutlich größere Stichprobenfehler aufweisen als einfache oder geschichtete Stichproben auf Basis von Adressdaten (vgl. dazu ausführlich Schnell/Noack in diesem Band).

3.2 Effekte durch Nonresponse

Die Auswahl des Erhebungsmodus kann auch gravierende Einflüsse auf die Höhe der Ausschöpfung, die damit zusammenhängenden Verzerrungen durch Nonresponse (Nonresponse-Error bzw. Nonresponse-Bias) sowie den möglichen Umgang damit ausüben. So erreichen persönlich-mündliche Befragungen in der Regel deutlich höhere Ausschöpfungsquoten als etwa telefonische oder schriftliche Befragungen (de Leeuw 1992; Hox/de Leeuw 1994; Groves u. a. 1988; Kury 1991, 287f.). In den letzten Jahren werden zunehmend niedrigere Ausschöpfungsquoten erreicht bzw. es müssen zu deren Steigerung größere Anstrengungen unternommen werden, was vor allem auf eine steigende „Befragungsmüdigkeit“ aufgrund der deutlich gestiegenen Zahl von Umfragen zurückgeführt wird. So erreichte etwa die vom BKA und MPI 1990 durchgeführte viktimisierungsbezogene Face-to-Face-Befragung bei 10.860 Personen in Ost- und Westdeutschland noch eine Antwortquote von auswertbaren Interviews von 76,6 % in Ost- und von 70,1 % in Westdeutschland (Kury u. a. 1996, 27). Die neueste, 2012 von beiden Institutionen in Deutschland gemeinsam geplante und differenziert umgesetzte Telefonbefragung (CATI) mit rund 35.500 Personen erreichte – durchschnittlich bis zu sieben Kontaktversuchen – eine Antwortquote von vollständig realisierten Interviews von 25,2 % (Schiel u. a. 2013, 30f.). Bei einer Zusatzstichprobe von Personen mit türkischem Migrationshintergrund lag die Antwortquote bei 21,1 % (ebd., 36). Wollinger u. a. (2014, 76) betonen, dass sie bei ihrer schriftlichen Befragung von Opfern von Wohnungseinbruch wohl nur deshalb eine Rücklaufquote von 68,7 % erreichen konnte, weil dem Fragebogen fünf Euro als „Aufwandsentschädigung“ beigelegt wurden (vgl. auch Kunz 2013, 10f.).

Auch wenn niedrige Ausschöpfungsquoten nicht zwingend mit Verzerrungseffekten der Zielvariablen einhergehen (Groves 2006; Groves/Peytcheva 2008), so ist die Repräsentativität der Ergebnisse jedoch zumindest in Frage zu stellen – vor allem wenn erwartet werden muss, dass sich Nichtbefragte (Nonrespondenten) hinsichtlich der interessierenden Merkmale systematisch von den befragten Personen unterscheiden. Dies scheint durchaus wahrscheinlich – zumindest was die Verbreitung von Opfererlebnissen angeht (z. B. Stangeland 1996, van Dijk u. a. 1990; Müller/Schröttle 2004; für eine Zusammenfassung des Forschungsstandes vgl. Guzy 2015). Da zudem die Gründe für Nonresponse in Umfragen in Abhängigkeit des Erhebungsverfahrens erheblich variieren, diese jedoch in unterschiedlichem Zusammenhang mit Opfererlebnissen stehen können, sind weitere Abweichungen allein durch die unterschiedliche Zusammensetzung der Nonrespondenten zwischen Modes denkbar (Groves u. a. 2009).

3.3 Effekte durch Messfehler

Als besonders relevante Einflussgröße und gemäß der engen Definition von Modeeffekten zentraler Faktor muss der Effekt des Erhebungsverfahrens auf die Messung selbst, d. h. den Antwortprozess, bewertet werden.

In der Literatur wird eine Vielzahl an Effekten auf das Antwortverhalten diskutiert, die nachweislich auch durch das Erhebungsverfahren selbst beeinflusst werden. Zu nennen sind bspw. die Nichtbeantwortung einzelner Fragen (Item Nonresponse), Effekte durch die Formulierung der Fragen, durch die Anordnung der Fragen (Reihenfolgeeffekte) bzw. der vorgegebenen Antwortmöglichkeiten bei geschlossenen Fragen, Erinnerungsprobleme beim Abruf relevanter Informationen, soziale Erwünschtheit oder sonstige Antworttendenzen wie z. B. die Tendenz zu extremen Antworten (zur Beeinflussung dieser Effekte durch das Erhebungsverfahren vgl. zusammenfassend Madow u. a. 1983; Groves u. a. 2009; Tourangeau u. a. 2000). Gerade bei sensiblen Themen werden solche Effekte besonders relevant (Hindelang u. a. 1979; Kerschke-Risch 1993, 21).

Um die modespezifischen Entstehungsbedingungen und Prozesse, die zu Messfehlern führen, besser verstehen zu können, erscheint es hilfreich auf das kognitive Modell des Antwortprozesses von Tourangeau u. a. (2000) zurückzugreifen. In diesem Ansatz wird der Antwortprozess in vier Aufgaben untergliedert, und zwar i) das Verständnis der Frage, ii) den Abruf der für die Beantwortung der Frage notwendigen Informationen, iii) die Meinungsbildung und iv) die Antwortgabe. Tourangeau u. a. (2000) konnten zeigen, dass jede dieser Aufgaben durch das Erhebungsverfahren beeinflusst werden kann. Darauf aufbauend haben Guzy/Leitgöb (2015) dargelegt, auf welchen Stufen

des Antwortprozesses Modeeffekte im Rahmen der Abfrage von Opfererlebnissen theoretisch erwartbar sind. Folgende Prozesse scheinen dabei von grundsätzlicher Bedeutung:

(i) *Verständnis der Frage:* Es scheint plausibel, dass bei interviewerad-ministrierten Erhebungsverfahren (durch das Vorlesen der Frage) von den Befragten häufiger alle relevanten Erklärungen und Frageteile wahrgenommen werden als in selbstadministrativen Erhebungsverfahren. So konnten Studien nachweisen, dass in schriftlichen Befragungen regelmäßig Teile von – insbesondere längeren – Fragen übersprungen werden (Galesic u. a. 2008). Dies spielt insbesondere in Opferbefragungen eine bedeutende Rolle, da in den Fragen zu Opfererlebnissen relevante Deliktbeschreibungen und Referenzzeiträume erläutert werden, die für die korrekte Beantwortung und damit die Qualität der Ergebnisse entscheidend sind. Darüber hinaus stehen in interviewerad-ministrierten Umfragen geschulte Personen zur Verfügung, um Unklarheiten auszuräumen. Dies führt in der Regel nicht nur zu vollständigeren Antworten (Bowling 2005), sondern auch zu einer höheren Validität und damit Datenqualität der Ergebnisse (Schober/Conrad 1997). Hinzu kommt, dass auch Personen ohne bzw. mit eingeschränkten Lesekompetenzen an der Befragung teilnehmen können (Tourangeau u. a. 2000). Andererseits muss berücksichtigt werden, dass sich interviewer-administrierte Erhebungsverfahren grundsätzlich durch eine höhere kognitive Belastung auszeichnen, da relevante Fragen akustisch (zeitnah) verarbeitet werden müssen, und bei Bedarf nicht mehrfach nachgelesen und überdacht werden könnten (Krosnick 1991).

(ii) *Informationsabruf*: Für den Bereich des Informationsabrufs sind ebenfalls Effekte durch das Erhebungsverfahren wahrscheinlich – insbesondere durch den mit einem Erhebungsverfahren verbundenen Zeitdruck zur Beantwortung von Fragen. Auch dies spielt in Opferbefragungen eine besondere Rolle, da nicht nur relevante Erlebnisse erinnert und zeitlich korrekt zugeordnet werden müssen, sondern kriminalitätsbezogene Einstellungen (die nicht immer präsent sind) teilweise erst generiert werden müssen (so z. B. die angemessene Strafe für bestimmte Deliktbeschreibungen oder das Sicherheitsgefühl nachts in der eigenen Wohngegend).

(iii) *Antwortauswahl und -abgabe*: Im Zuge der Antwortgabe muss die generierte Antwort einer der vorgegebenen Antwortkategorien zugeordnet und z. B. hinsichtlich Konsistenz geprüft werden (Tourangeau u. a. 2000). In der Umfrageforschung haben in diesem Zusammenhang verschiedene Antwortverzerrungen Aufmerksamkeit erfahren, so z. B. so-

ziale Erwünschtheit, bewusste Falschangaben, fehlende Angaben, Reihenfolgeeffekte, Kontexteffekte oder sonstige Antworttendenzen wie inhaltsunspezifische Zustimmungen (Aquieszenz) oder die Neigung mittlere oder extreme Antwortkategorien auszuwählen (für einen Überblick vorhandener Antworttendenzen inkl. deren empirischer Evidenz vgl. Biemer u. a. 2004; Lyberg u. a. 1997). Der aktuelle Forschungsstand bietet zahlreiche Belege, dass das jeweilige Ausmaß dieser Effekte, die freilich – je nach Frage – auch in Opferbefragungen relevant sind, vom Erhebungsmodus abhängt. Da Antworttendenzen jedoch grundsätzlich als Vereinfachung kognitiver Anstrengungen betrachtet werden können (Krosnick 1991), sind in der Regel stärkere Antworttendenzen in den (kognitiv stärker beanspruchenden) Telefonbefragungen zu beobachten (Bishop u. a. 1988; Dillman u. a. 1996). Effekte durch soziale Erwünschtheit oder Falschangaben sind ebenfalls häufiger bei interviewer-administrierten Befragungsformen festzustellen (wobei persönlich-mündliche Verfahren aufgrund der starken Interviewereinbindung die höchste Gefahr von sozial erwünschten Angaben aufweisen, vgl. Tourangeau/Smith 1996; de Leeuw 1992; Kreuter u. a. 2008c). Fehlende Angaben werden dagegen etwas häufiger bei selbst-administrierten Erhebungsformen beobachtet (Kreuter u. a. 2008c; Tourangeau/Smith 1996; für einen Überblick des Forschungsstands vgl. Roberts 2007; Tourangeau u. a. 2000) – vmtl. gerade wegen des reduzierten sozialen Drucks einer Antwortgabe.

Dass etwa durch die Gestaltung der Erhebungsinstrumente verursachte Messfehler in ihrem möglichen Ausmaß bis heute unterschätzt werden, zeigen zahlreiche Untersuchungen. So konnte Kury (1994 in einer experimentellen Methodenstudie zu einem Item aus der Untersuchung von Sessar (1992) zu Sanktionseinstellungen zeigen, dass allein die Umkehrung der fünf Antwortalternativen auf die Frage zu deutlich anderen Ergebnissen führte. Wurden drei weitere Antwortalternativen zu milderen bzw. härteren Sanktionen hinzugefügt, waren die Ergebnisse mit den Antworten auf das Originalitem nicht mehr vergleichbar (Kury 1994, 90f.).

Hinweise auf die Bedeutung des Fragekontextes liefert auch die einzige deutsche repräsentative Längsschnittbefragung zu den „Ängsten der Deutschen“, die seit 1991 mit weitgehend demselben Design jährlich durchgeführt wird und in einem „neutralen“ Kontext nach erlebten Ängsten, unter anderem auch die „Angst vor Straftaten“, fragt: Hier werden durchgehend niedrigere Werte zur Verbrechensangst festgestellt als in spezifischen Victim Surveys, in denen der Befragte deutlich stärker auf das Thema Kriminalität hingewiesen wird (R+V-Versicherung 2014; Kreuter 2002, 75ff.). Allerdings kann nicht mit Sicherheit ausgeschlossen werden, dass auch hier Unterschiede in der Fragenformulierung einen (zusätzlichen) Einfluss auf die Ergebnisse haben.

4 Modeeffekte nach Erhebungsmodus

Nachdem in den beiden vorausgehenden Abschnitten der Rahmen skizziert wurde, in dem die Ursachen und Prozesse für die Entstehung von Modeeffekten grundsätzlich zu verorten bzw. zu interpretieren sind, wird im Folgenden der Forschungsstand zu den Effekten der einzelnen Erhebungsverfahren auf den Erhebungsprozess bzw. die Ergebnisse von Opferbefragungen dargestellt (vgl. auch Guzy 2014).

4.1 Telefonische Befragungen

Telefonbefragungen stellen national wie international (aktuell noch) die am häufigsten verwendete Erhebungsmethode dar (ADM 2011, für die internationale Forschung Kalkgraff-Skjak/Harkness 2003). Die Vorteile liegen einerseits in der kostengünstigen und schnellen Befragungsdurchführung, andererseits in der hohen Kontrollmöglichkeit der Interviewerinnen und Interviewer (bspw. durch Mithören von Interviews) und der hohen Befragungsstandardisierung – insbesondere bei der Verwendung computerunterstützter Verfahren. Darüber hinaus kann angenommen werden, dass die Deliktdefinitionen durch die Anwesenheit einer Interviewerin bzw. eines Interviewers präziser kommuniziert werden, da diese für Rückfragen zur Verfügung stehen und Frageteile nicht oder zumindest seltener überlesen bzw. übersehen werden können (vgl. auch van Dijk u. a. 2010).

Als nachteilig sind die häufiger anzutreffenden Antworttendenzen wie Aquieszenz- und Recencyeffekte (Schwarz u. a. 1985, Bishop u. a. 1988) sowie die geringere Flexibilität hinsichtlich Länge und Gestaltung des Fragebogens zu bewerten (Steeh 2008). Erstere spielen allerdings für die Abfrage von Viktimisierungserlebnissen keine herausragende Rolle, da diese in der Regel mit dichotomen Antwortformaten abgefragt werden, die im Vergleich zu längeren Antwortlisten seltener durch Antworttendenzen beeinflusst werden (Tourangeau u. a. 2000). Als deutlich gravierender sind dagegen Effekte durch die Einbindung einer Interviewerin bzw. eines Interviewers zu bewerten. Nahezu alle mode-vergleichenden Studien kommen zu dem Ergebnis, dass sensible Delikte in interviewer-administrierten Befragungen, insbesondere etwa in selbstadministrierten Erhebungsteilen, häufiger berichtet werden (Cantor/Lynch 2000; Müller/Schröttle 2004; van Dijk u. a. 2010). Dies könnte aber nicht nur durch die höhere Anonymität und die damit zusammenhängende höhere Bereitschaft Opfererlebnisse zuzugeben erklärt werden, sondern auch auf den Umstand zurückgeführt werden, dass Opfer in interviewer-administrierten Umfragen die Frage nach einer Opfererfahrung aus Scham eher

verneinen anstatt sie zu verweigern (Kreuter 2008c; Skogan 1981; Guzy/Leitgöb 2015). Dieser Einfluss der Scham dürfte verständlicherweise in selbstadministrierten Erhebungen reduziert sein.

Bemerkenswert sind in diesem Zusammenhang auch die Ergebnisse von Cantor/Lynch (2000), die feststellten, dass nach dem Wechsel des amerikanischen National Crime Victim Surveys (NCVS) von einer klassischen Telefonbefragung zu einer computerunterstützen Telefonbefragung (CATI) die Opferraten signifikant anstiegen.

Seit einigen Jahren verlieren Telefonbefragungen sowohl in Umfragen insgesamt, als auch in Opferbefragungen zunehmend an Bedeutung, da einerseits die Ausschöpfungsquoten erheblich abgenommen haben und andererseits immer weniger Personen überhaupt noch über Festnetz zu erreichen sind (ITU World Telecommunication/ICT Indicators Database 2010). Damit einhergehend stellt die zunehmende Verbreitung von Mobiltelefonen und so genannten Mobile Onlys (= Personen die ausschließlich über das Mobiltelefon erreichbar sind) eine Herausforderung dar (Eurostat 2011; zur methodologischen Problematik Häder/Häder 2009). So konnte bspw. eine finnische Untersuchung auf Basis der EU-ICS Daten von 2005 feststellen, dass Mobile Onlys sich nicht nur hinsichtlich ihrer soziodemographischen Verteilung von über das Festnetz befragten Personen unterscheiden, sondern dass diese Personengruppe auch signifikant erhöhte Viktimisierungsraten aufweist (Hideg/Manchin 2005), was insbesondere mit der Stichprobenselektion zu tun haben dürfte, da vor allem jüngere Personen über Mobiltelefone erreichbar sind und diese gleichzeitig im Durchschnitt eine erhöhte Kriminalitäts- und Opferrate zeigen. Mit der inzwischen erheblichen Verbreitung von Mobiltelefonen über alle Altersklassen dürfte dieser Effekt allmählich verschwinden. Zu beachten ist auch, dass die Situation, in welcher Befragte über Mobilfunk erreicht werden, kaum zu kontrollieren ist und somit „Umwelteinflüsse“ eine erhebliche Rolle auf das Antwortverhalten spielen können.

4.2 Postalische Befragungen

Postalische Befragungen stellen in aller Regel Erhebungsverfahren mit einer guten Auswahlgrundlage (meist einem amtlichen Adressregister) und folglich einer guten Coverage (= Übereinstimmung zwischen Auswahlgrundlage und interessierenden Grundgesamtheit) dar. Ferner können Fragen mit vielen Antwortkategorien und/oder hoher Sensibilität abgefragt werden. Insbesondere Letzteres führt prinzipiell zu ehrlicheren und weniger sozial erwünschten Antworten (de Leeuw/Hox 2008; Tourangeau/Smith 1996), was gerade in Viktimisierungsbefragungen bedeutende Vorteile mit sich bringt (Kury 1994). So erhoben etwa Müller und Schröttle (2004, 10) bei der groß angelegten und

methodisch differenziert durchgeführten Face-to-Face-Befragung von Frauen in Deutschland Ereignisse zu den hochsensiblen Bereichen Gewalt in der Partnerschaft und der eigenen Herkunftsfamilie, und zwar mittels eines schriftlichen Selbstausfüllers (drop-off) im Anschluss an den mündlichen Teil, wobei die Befragten die Ergebnisse in einem Umschlag verschließen konnten. Dabei konnten substantiell mehr Gewaltereignisse als bei der nur mündlichen Befragung erfasst werden, obwohl die Interviewerinnen besonders geschult waren und auch im mündlichen Teil die Themen sensibel ansprachen.

Ferner stehen Befragte in schriftlichen Befragungen unter einem geringeren Zeitdruck, gerade etwa im Vergleich zu telefonischen Umfragen, so dass Ihnen mehr Zeit zur Erbringung der Erinnerungsleistung zur Verfügung steht. Dies führt in der Regel zu genaueren Ergebnissen (Chang/Krosnick 2009; Tourangeau u. a 2000). Für den Bereich der Viktimisierungsbefragungen konnte Kury (1994) bspw. feststellen, dass gerade leichte Nichtkontaktdelikte signifikant häufiger in schriftlichen Befragungen angegeben werden als in mündlichen Befragungen. Bei schweren und deutlich leichter zu erinnernden Delikten konnten diese Effekte nicht nachgewiesen werden. Guzy und Leitgöb (2015) konnten diese Ergebnisse allerdings auf Basis des International Crime Victim Survey (ICVS) nicht bestätigen, was die Komplexität der Zusammenhänge unterstreicht.

Unglücklicherweise weisen schriftliche Erhebungsverfahren im Vergleich zu anderen Vorgehensweisen häufig ausgesprochen niedrige Ausschöpfungsquoten und hohe Itemnonresponseraten auf (zu Ausnahmen, bspw. bei Verwendung von Incentives, siehe Becker/Mehlkop 2007; Kunz 2013), was regelmäßig zu einem ausgeprägten Bildungsbias führt (de Leeuw 1992). Für beide Effekte muss ein nicht unerheblicher Einfluss auf die Ergebnisse von Viktimisierungsbefragungen erwartet werden. Schneekloth und Leven (2003, 19) betonen allerdings zu Recht: „Tatsächlich misst die Ausschöpfung jedoch nur [. . .], wie groß der Spielraum für Selektivität durch Nonresponse ist. Sie besagt nichts über die tatsächliche Selektivität“ (vgl. auch Deutsche Forschungsgemeinschaft 1999, 104). Nonresponse führt vor allem dann zu einer Verschlechterung der Aussagekraft von Umfragen, wenn damit systematisch eine Unterausschöpfung von bestimmten Bevölkerungsgruppen verbunden ist. Kunz (2013) konnte in ihrer Methodenstudie etwa zeigen, dass sich eine unterschiedliche Antwortquote bei einzelnen Bevölkerungsgruppen beispielsweise bereits durch die Zusicherung von Anonymität ergeben kann. Bei anonymer Befragung lag die Teilnahmequote ca. 5 % höher, vor allem zeigte sich aber eine Altersverzerrung in der (nicht-anonym durchgeführten) Stichprobe. Zu Recht betont die Autorin, dass anonyme im Vergleich zu nicht-anonymen Befragungen zu weniger verzerrten realisierten Stichproben führen, was wiederum validere Ergebnisse erzeugt (Kunz 2013, 16). Dennoch muss auch be-

dacht werden, dass mit verstärkten Bemühungen zur Erhöhung der Stichprobenausschöpfung auch eine stärkere Verzerrung der Umfrageergebnisse einhergehen kann – bspw. wenn dadurch nur bestimmte Bevölkerungsgruppen erreicht werden (Groves 1989; Groves/Couper 1998).

Ferner muss bedacht werden, dass sich Viktimisierungsbefragungen durch teilweise komplexe Filterführungen auszeichnen, welche nicht selten zu Filterfehlern und zu falsch oder nicht beantworteten Items führen (vgl. auch Cantor/Lynch 2000). Dies bestätigt auch die deutsche Pilotstudie der Testerhebung „Translating and Testing a Victimisation Survey Module“ (Statistisches Bundesamt/Bundeskriminalamt 2010). Gerade in Viktimisierungsbefragungen, die auf die Messung seltener Ereignisse abzielen, können diesbezügliche Fehler gravierend sein. Darüber hinaus steht Befragten in schriftlichen Befragungen keine Interviewerin bzw. kein Interviewer für Rück- oder Verständnisfragen zur Verfügung. Auch wenn von Interviewenden grundsätzlich die Gefahr verzerrender Effekte auf die Datenqualität ausgeht (Loosveldt 2008), können sie bei Verständnisschwierigkeiten (z. B. wenn Befragte ihr Opfererlebnis nicht eindeutig einer Deliktform zuordnen können) durchaus hilfreich sein und genauere Ergebnisse produzieren (vgl. bspw. Schober/Conrad 1997).

Ein weiterer Nachteil schriftlicher Befragungen muss außerdem in der langen Erhebungs- und Datenaufbereitungsphase gesehen werden, da in der Regel mehrere Erinnerungswellen mit entsprechenden Rücklaufzeiten sowie die manuelle Eingabe der Daten mit anschließender Plausiblisierung notwendig ist sowie die Tatsache, dass in der Regel keine Kontrolle darüber vorliegt wer den Fragebogen tatsächlich ausgefüllt hat und ob es sich tatsächlich um eine Person der Stichprobe oder Grundgesamtheit handelt.

4.3 Persönlich-mündliche Erhebungsverfahren

Persönlich-mündliche Befragungen weisen gegenüber vielen anderen Erhebungsverfahren bedeutende Vorteile auf: Es werden in der Regel hohe Ausschöpfungsquoten erreicht (so auch ein Vergleich der EU-Testerhebungen, van Dijk u. a. 2010), es existieren gute Auswahlgrundlagen in Form von Adresslisten oder geographischen Sample-Units (auch wenn diese in weniger dicht besiedelten Regionen wenig praktikabel sein können) und die maximal mögliche Befragungsdauer liegt deutlich über jener anderer Interviewformen (Lohr 2008; Loosveldt 2008).

Doch obwohl persönlich-mündliche Befragungen in der Umfrageforschung regelmäßig als besonders geeignete und anderen Erhebungsverfahren in vielen Punkten überlegene Methode betrachtet werden (König 1962, 328; Scheuch 1973, 66), lassen sich mittlerweile empirisch, z. B. gegenüber telefonischen Befragungen, nur wenige qualitative Vorteile feststellen: „[. . .] when

face to face and telephone surveys are compared only small effects are discovered. Face to face interviews [...] result in data with slightly less item nonresponse and slightly more statements to open question. No differences were found concerning response validity (record checks) and social desirability“ (de Leeuw 1992, 34). Und auch für den Bereich der Viktimisierungsbefragung ließen sich bisher nur teilweise Unterschiede zwischen telefonischen und persönlich-mündlichen Befragungen feststellen. So wurden in der Bochumer Dunkelfeldstudie von 1998 weder bei Diebstahl noch bei Körperverletzung signifikant voneinander abweichende Prävalenzen gemessen (Schwind u. a. 2001). Deutlichere Differenzen finden sich dagegen zwischen persönlich-mündlichen und schriftlichen Erhebungsverfahren, da hier größere Unterschiede hinsichtlich der wahrgenommenen Anonymität vorliegen (Tourangeau/Smith 1996). Dies bestätigt sich auch in schriftlichen Viktimisierungssurveys, in denen sensible Delikte regelmäßig häufiger angegeben werden als in mündlichen Befragungen (Kury 1994; Cantor/Lynch 2000; Wetzels u. a. 1994; Müller/Schröttle 2004).

Hinsichtlich weiterer Einflüsse, die durch den Einsatz von Interviewerinnen und Interviewern entstehen, berichten Bailar u. a. (1977) sowie Bailey u. a. (1978) auf Basis von Daten des face-to-face administrierten *National Crime Survey (NCS)* stärkere *interviewerbezogene Clustereffekte*[4] für Viktimisierungsitems als für die weniger sensitiven Fragen zur Soziodemographie. Weiterhin konnten Schnell und Kreuter (2000; 2005) im Rahmen der *DEFECT-Studie* feststellen, dass Interviewereffekte sogar die Stichprobengenauigkeit (*Designeffekt*)[5] von Items zu Kriminalitätsfurcht, subjektiven Viktimisierungswahrscheinlichkeiten, tatsächlichen Viktimierungserfahrungen, Sicherheitsmaßnahmen usw. beeinflussen, und zwar in deutlich stärkerem Maße als jene Effekte, die auf das clusterbasierte Samplingdesign (Personen in Haushalten in Sampling Points; zu den Details siehe Schnell/Kreuter 2000, 91f.) zurückzuführen sind.

---

[4] Ein interviewerbezogener Clustereffekt liegt vor, wenn sich die Befragten innerhalb eines Clusters – als Cluster wird im vorliegenden Fall eine Gruppe von Befragten definiert, die alle von der bzw. vom selben Interviewerin bzw. Interviewer befragt wurden – in den interessierenden Merkmalen ähnlicher sind, als wenn jeder Befragte von einem anderen Interviewer befragt worden wäre (für eine allgemeine Definition von Clustereffekten siehe etwa Campbell/Berbaum 2010). Die Ursache dieses Effekts liegt in dem über die Interviewenden variierenden Ausmaß an (ungewollter) Einflussnahme auf das Antwortverhalten der Befragten (siehe dazu ausführlich Schaeffer u. a. 2010). Die Varianz zwischen den interviewerbezogenen Clustern (Interviewervarianz; siehe Kreuter 2008b) repräsentiert einen Teil der *Methodenvarianz* (method effects produced by measurement context; Podsakoff u. a. 2003).

[5] Als Designeffekt wird allgemein das Verhältnis der Varianz eines interessierenden Schätzers auf Basis eines arbiträren Stichprobendesigns zur Varianz basierend auf einer einfachen Zufallsauswahl (ohne Interviewereffekte) bezeichnet.

5 (Praktische) Empfehlungen für die Auswahl geeigneter Erhebungsverfahren

Der Überblick macht deutlich, dass es „das“ geeignete Erhebungsverfahren nicht gibt und dass die einzelnen Erhebungstechniken jeweils spezifische Vor- und Nachteile haben. Seit dem Aufkommen der Umfrageforschung in der ersten Hälfte des letzten Jahrhunderts hat die Zahl der Umfragen, vor allem auch telefonischer und internetbasierter Befragungen, dramatisch zugenommen, was auch zu einer „Befragungsmüdigkeit“ in der Bevölkerung und damit einer Reduzierung der Antwortbereitschaft geführt hat. Die in den letzten Jahren zunehmend gestiegene Diskussion um Datensicherheit hat ein Übriges zu einer größeren Zurückhaltung hinsichtlich der Offenlegung persönlicher Angaben geführt, insbesondere wenn es um die Preisgabe sehr persönlicher Ereignisse geht, wie das bei Viktimisierungen nicht selten der Fall ist. Grundsätzlich sollte die Wahl eines geeigneten Erhebungsverfahren unter Berücksichtigung der mit einem Mode verbundenen Messfehler (z. B. durch die Gefahr sozial erwünschter Antworten und Falschangaben, vgl. Abschnitt 3), sowie der mode-spezifischen Kontaktierungsmöglichkeit getroffen werden (verfügbare Auswahlgrundlage, Antwortbereitschaft etc., vgl. Abschnitt 4).

Opferbefragungen zeichnen zweifellos ein deutlich genaueres Bild der Kriminalitätswirklichkeit, was insofern nicht erstaunlich ist, als das Dunkelfeld der Straftaten enorm hoch ist (Kury 2001). Wie genau dieses Bild letztlich allerdings ausfallen kann, hängt nicht unerheblich vom Design der Umfragen ab. Der Trend, Umfragen möglichst schnell und billig durchzuführen, kann kaum zu einer höheren Qualität der Daten beitragen. Erhebungsinstrumente müssen sorgfältig konstruiert werden, da von ihnen, wie inzwischen zahlreiche vorliegende Studien zeigen, erhebliche Einflüsse auf die gewonnenen Ergebnisse ausgehen können. Werden Einstellungen, etwa zu Sanktionen oder kriminalpolitischen Entscheidungen abgefragt, sollte berücksichtigt werden, dass teilweise ein erheblicher Teil der Befragten keine klaren Vorstellungen, etwa zum Sanktionsverhalten von Gerichten bzw. der „Notwendigkeit“ einer Strafverschärfung zu kriminalpräventiven Zwecken oder der Kriminalitätsfurcht, hat („Nonattitudes“). Die zu messenden Konstrukte sollten klar definiert sein, die Forscher sollten wissen, was sie messen wollen, wenn sie etwa „Kriminalitätsfurcht“ oder „Punitivität“ messen (Kury u. a. 2004; vgl. auch Faulbaum in diesem Band). Werden die Befragten aufgrund der Konstruktion des Erhebungsinstruments „gezwungen“ (z. B. durch eine fehlende „weiß nicht“-Kategorie) sich zu entscheiden, obwohl sie keine Vorstellungen haben, kann das zu falschen Einschätzungen führen, der Übernahme dessen, was man den Medien entnommen hat und worüber man sich eigentlich noch keine Gedanken gemacht hat.

Die Aussagekraft von Umfragen kann dadurch besser beurteilt werden, dass Methodenvariationen eingeführt werden, etwa durch eine Kombination mündlicher Interviews mit schriftlichen Befragungen (vgl. etwa Müller/ Schröttle 2004) bzw. der Kontext, in welchen einzelne Items eingebaut sind oder die Formulierung von Items, verändert wird. Bei persönlichen Interviews kann etwa die Datenqualität der gemachten Angaben von den Interviewerinnen und Interviewern zumindest ansatzweise eingeschätzt werden (Kury u. a. 1992). Freilich müssen bei der simultanen Verwendung verschiedener Erhebungsmodi (Mixed-mode-surveys) auch verschiedene, mit den einzelnen Erhebungsmodi verbundenen, Modeeffekte berücksichtigt werden. Für bestimmte Fragestellungen und Items kann die Entscheidung für Mixed-mode-surveys aber durchaus mit erheblichen Vorteilen verbunden sein (vgl. auch Killias in diesem Band; für detaillierte Ausführungen siehe Dillman u. a. 2014).

6 Zusammenfassung

– Die in Umfragen ausgewählte Datenerhebungsmethode hat, wie inzwischen zahlreich vorliegende nationale und internationale Studien zeigen, einen erheblichen Einfluss auf die letztlich gefundenen Ergebnisse. Aufgrund dessen sind Resultate aus Umfragen stets mit Vorsicht und vor dem Hintergrund des jeweiligen Erhebungsverfahrens zu interpretieren.

– Bei der Bewertung von Modeeffekten sollte zwischen einem „*engen*“ und einem „*weiten*“ Verständnis unterschieden werden. Während die enge Definition auf Effekte des Erhebungsmodus auf das Antwortverhalten selbst (z. B. durch die stärkere Neigung zu sozial erwünschten Antworten oder bestimmten Antworttendenzen) abzielt, verweist die weite Definition auf die Tatsache, dass die einzelnen Erhebungsverfahren mit spezifischen Auswahlgrundlagen, Stichprobenverfahren und Ausschöpfungsraten verbunden sind. Auch diese üben „Modeeffekte“ aus.

– Die Datenerhebung hängt von zahlreichen Faktoren ab, neben der Gestaltung des Erhebungsinstruments etwa der Beteiligung von Interviewerinnen und Interviewern und deren Engagement, dem Ausmaß der gewährten Privatsphäre oder dem Einsatz unterschiedlicher Technologien. All diese Merkmale haben unterschiedliche Effekte auf die Ergebnisse in Umfragen und sollten sowohl bei der Wahl eines geeigneten Erhebungsinstruments als auch der Bewertung von Ergebnissen aus Opferbefragungen berücksichtigt werden.

- Befragte müssen zur Beantwortung von Umfragen verschiedenen kognitiven Leistungen erfüllen: Verständnis der Frage, Informationsabruf sowie die Antwortauswahl und -abgabe. Für all diese Aufgaben sind Einflüsse durch die verschiedenen Erhebungsverfahren wahrscheinlich und/oder empirisch belegt.

- Selbst wenn eine repräsentative Stichprobe generiert werden konnte, können Ausfälle durch Nonresponse, die Verallgemeinerbarkeit der Resultate erheblich einschränken. Auch hier sind mode-spezifische Unterschiede wahrscheinlich.

- Die Ausgestaltung und Qualitätsprüfung eines Erhebungsinstrumentes, das die Gewinnung von möglichst aussagekräftigen und verallgemeinerbaren Ergebnissen ermöglicht, ist für alle Erhebungsverfahren ausgesprochen wichtig und schwierig – oft wird dafür zu wenig Zeit investiert.

- Nach gegenwärtigem Forschungsstand haben persönliche mündliche Interviews nach wie vor die wohl größte Chance, eine hohe Ausschöpfungsquote und valide Informationen zu liefern, unter der Voraussetzung, dass die Interviewerinnen und Interviewer gut geschult sind. Fehlereinflüsse bestehen vor allem bei der Erhebung wenig klar ausgeprägter Einstellungen („Nonattitudes“) sowie sensiblen Themen.

- Trotz aller Bedenken kann kein Zweifel daran bestehen, dass die Umfrageforschung, etwa in Form von Opferstudien, die empirische Kriminologie enorm bereichert und eine Fülle wesentlicher Informationen, etwa zum Dunkelfeld der Kriminalität oder zu Verbrechensfurcht, Punitivität oder Einstellungen zu Strafverfolgungsorganen wie der Polizei oder Justiz gebracht hat.

7 Weiterführende Literatur

Häder, Sabine (2000): Telefonstichproben. ZUMA How-to-Reihe, Nr. 6. URL: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/how-to6sh.pdf. – Download vom 12. 03. 2015.

Dillman, Don; Smyth, Jolene und Christian, Leah (2014): Internet, Phone, Mail, and Mixed Mode Surveys: The Tailored Design Method, 4. Aufl. New York: Wiley.

Lepkowski, James M.; Tucker, Clyde; J. Michael Brick; de Leeuw, Editz D.; Japec, Lilli; Lavrakas, Paul J.; Link, Michael W. und Sangster, Roberta L. (Hg.): Advances in telephone survey methodology. Hoboken, NJ: Wiley.

Baur, Nina; Blasius, Jörg (Hg.) (2014): Handbuch der Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS.

8 Literaturverzeichnis

ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (2011): Quantitative Interviews der Mitgliedsinstitute des ADM nach Befragungsart. https://www.adm-ev.de/g. – Download vom 12. 03. 2015

Aebi, Marcelo F.; Akdeniz, Galma; Barclay, Gordon; Campistol, Claudia; Caneppeke, Stefano; Gruszczynska, Beata; Harrendorf, Stefan; Heiskanen, Markku; Hysi, Vasilika; Jehle, Jörg-Martin; Jokinen, Anniina M; Kensey, Annie; Killias, Martin; Lewis, Chris G.; Savona, Ernesto; Smit, Paul und Pórisdóttir, Rannveig (2014): European Sourcebook of Crime and Criminal Justice Statistics 2014. 5. Aufl. Helsinki: European Institute for Crime Prevention and Control (HEUNI).

Ahlf, Ernst-Heinrich (2007): Seniorenkriminalität und -viktimität: alte Menschen als Täter und Opfer. In: Schneider, Hans Jörg (Hg.): Internationales Handbuch der Kriminologie. Band 1: Grundlagen der Kriminologie. Berlin: De Gruyter, S. 509–550.

Bailar, Barbara A., Bailey, Leroy, und Stevens, Joyce (1977). Measures of Interviewer Bias and Variance. In: Journal of Marketing Research, 14, S. 337–343.

Bailey, Leroy, Moore, Thomas F., und Bailar, Barbara A. (1978): An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample. In: Journal of the American Statistical Association, 73, S. 16–23

Becker, Rolf; Mehlkop, Guido (2011): Effects of Prepaid Monetary Incentives on Mail Survey Response Rates and on Self-reporting about Delinquency – Empirical Findings. In: Bulletin of Sociological Methodology, 111, S. 5–25.

Biderman, Albert D. (1967): Report on a pilot study in the district of Columbia on victimization and attitudes toward law enforcement. Field Surveys 1. Washington D.C.

Biemer, Paul P.; Groves Robert M.; Lyberg Lars E.; Mathiowetz Nancy A. und Sudman, Seyman (2004): Measurement Errors in Surveys. New York: Wiley.

Birkel, Christoph; Guzy, Nathalie; Hummelsheim, Dina; Oberwittler, Dietrich und Pritsch, Julian (2014): Der Deutsche Viktimisierungssurvey 2012. Erste Ergebnisse zu Opfererfahrungen, Einstellungen gegenüber der Polizei und Kriminalitätsfurcht. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht.

Bishop, George; Hippler, Hans-Jürgen; Schwarz, Norbert und Strack Fritz (1988): A Comparison of Response Effects in Self-Administered and Telephone Surveys. In: Groves Robert; Lyberg Lars; Massey, James; Nicholls, William und Waksberg, Joseph (Hg): Telephone Survey Methodology. New York: Wiley, S. 321–340.

Bowling, Ann (2005): Mode of Questionnaire Administration Can Have Serious Effects on Data Quality. In: Journal of Public Health 27, 3, S. 281–291.

Cantor, David; Lynch, James P. (2000): Self-Report Surveys as Measures of Crime and Criminal Victimization. In: Criminal Justice, 04/2000. URL: https://www.ncjrs.gov/criminal_justice2000/vol_4/04c.pdf. – Download vom 12. 08. 2014.

Chang, Linchiat C.; Krosnick, Jon A. (2009): National Surveys via RDD Telephone interviewing versus the Internet: Comparing Sample Representativeness and Response Quality. In: Public Opinion Quarterly, 73, S. 641–678.

Cochran, William G. (1983): Historical perspective. In: Madow, William G.; Olkin, Ilgram, Rubin, Donald B. (Hg.): Incomplete data in sample surveys. Vol. 2: Theory and bibliographies. New York: Academic Press, S. 11–25.

de Leeuw, Edith D. (1992): Data Quality in Mail, Telephone and Face to Face Surveys. URL: http://files.eric.ed.gov/fulltext/ED374136.pdf – Download vom 12. 08. 2014.

de Leeuw, Edith D. (2008): Choosing the Method of Data Collection. In: de Leeuw, Edith D.; Hox, Joop, J. und Dillman, Don A. (Hg.): International Handbook of Survey Methodology. New York: Psychology Press, Taylor & Francis Group, S. 113–135.

de Leeuw, Edith D.; Hox, Joop (2008): Self-Administered Questionnaires. In: De Leew, Edith; Hox, Joop und Dillman, Don (2008): International Handbook of Survey Methodology. New York: Psychology Press, Taylor & Francis Group, S. 239–263.

Deutsche Forschungsgemeinschaft (1999): Qualitätskriterien der Umfrageforschung. Berlin.

Dillman Don A.; Sangster, Roberta; Tarnai, John und Rockwood, Tood (1996): Understanding Differences in People's Answers to Telephone and Mail. In: New Directions for Program Evaluation, 70, S. 45–61.

Dillman, Don; Smyth, Jolene und Christian, Leah (2014): Internet, Phone, Mail, and Mixed Mode Surveys: The Tailored Design Method, 4. Aufl., New York: Wiley.

Dodd, Stuart C. (1947): Standards for surveying agencies. Public Opinion Quarterly, 11, S. 115–130.

Ennis, Phillip H. (1967): Criminal Victimization in the United States: A report of a national survey. Washington D.C.

Eurostat (2011): Mobilfunkteilnehmer (je 100 Einwohner). URL: http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&language=de&pcode=tin00059&plugin=0. – Download vom 12. 03. 2015.

Fuchs, Marek (2009): Item-Nonresponse in einer Befragung von Alten und Hochbetagten. Der Einfluss von Lebensalter und kognitiven Fähigkeiten. In: Weichbold, Martin; Bacher, Johann und Wolf, Christof (Hg.): Umfrageforschung. Herausforderungen und Grenzen. Sonderheft 9 der österreichischen Zeitschrift für Soziologie. Wiesbaden: VS Verlag, S. 333–349.

Fricker, Ronald D. Jr.; Schonlau, Matthias (2002): Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. In: Field Methods, 14, S. 347–367.

Galesic, Mirta; Tourangeau, Roger; Couper, Mick P. und Conrad, Frederic (2008): Eye-tracking data. New insights on response order effects and other cognitive shortcuts in survey responding. In: Public Opinion Quarterly, 72, S. 892–913.

Groves, Robert M. (2004): Survey Errors and Survey Costs. New York: Wiley & Sons.

Groves, Robert M.; Fowler, Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor und Tourangeau, Roger (2009): Survey Methodology. Hoboken: Wiley & Sons.

Groves, Robert M.; Biemer, Paul P.; Lyberg, Lars E.; Massey, James; Nicholls, William L. und Waksberg, Joseph (1988): Telephone Survey Methodology. New York: Wiley.

Groves, Robert M. (2006): Nonresponse rates and nonresponse bias in household surveys. In: Public Opinion Quarterly, 70, 5, S. 646–675.

Groves, Robert M.; Peytcheva, Emilia (2008): The Impact of nonresponse Rates on Nonresponse Bias. A Meta Analysis. In: Public Opinion Quarterly, 72, 2, S. 167–189.

Groves, Robert M. (1989): Survey Errors and Survey Cost. New York: John Wiley & Sons.

Groves, Robert M.; Couper, Mick P. (1998): Nonresponse in Household Surveys. New York: Wiley.

Griffin-Saphire, Diane (1984): Estimation of victimization prevalence using data from the National Crime Survey. New York: Springer-Verlag.

Guzy, Nathalie (2015): (Unit-)Nonresponse-bias in Dunkelfeld-Opferbefragungen. In: Wolf, Christoph; Schupp, Jürgen (Hg.): Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen (= Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute). Wiesbaden: Springer VS, S. 161–207.

Guzy, Nathalie (2014): International vergleichende Viktimisierungssurveys. Aktuelle Herausforderungen und Ergebnisse des Methodentests „ICVS-2“. In: Eifler, Stefanie; Pollich, Daniela: Empirische Forschung über Kriminalität. Perspektiven und Herausforderungen. Wiesbaden: Springer VS, S. 149–182.

Guzy, Nathalie; Leitgöb, Heinz (2015): Assessing Mode Effects in Online and Telephone Victimization Surveys. In: International Review of Victimology, Vol. 21, 1, S. 101–131.

Häder, Michael; Häder, Sabine (2009): Telefonbefragungen über das Mobilfunknetz. Konzept, Design und Umsetzung einer Strategie zur Datenerhebung. Wiesbaden: Springer VS Verlag.

Häder, Sabine; Gabler, Siegfried (1998): Ein neues Stichprobendesign für telefonische Umfragen in Deutschland. In: Gabler, Siegfried; Häder, Sabine und Hoffmeyer-Zlotnik, Jürgen H. P. (Hg.): Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag, S. 69–88.

Hideg, Gergely; Manchin, Robert (2005): The Inclusion of Mobile only Person in the Finnish ICS. EU ICS Working Paper Series. URL: http://www.europeansafetyobservatory.eu/doc/The%20Inclusion%20of%20Mobile-only%20Persons%20in%20the%20Finnish%20ICS.pdf. – Download vom 12. 03. 2015.

Hindelang, Michael J.; Hirschi, Travis und Weis, Joseph G. (1979): Correlates of delinquency. The illusion of discrepancy between self-report and official measures. American Sociological Review, 44, S. 995–1014.

Hippler, Hans-J.; Schwarz, Norbert (1987): Response effects in surveys. In: Hippler, Hans-J.; Schwarz, Norbert und Sudman, Seymour (Hg.): Social information processing and survey methodology. New York: Springer, S. 102–122.

Hofer, Hans von (2009): Der internationale Kriminalitätsvergleich mit Hilfe der Statistik. In: Schneider, Hans Joachim (Hg.): Internationales Handbuch der Kriminologie. Band 2: Besondere Probleme der Kriminologie. Berlin: De Gruyter, S. 121–144.

ITU World Telecommunication/ICT Indicators Database (2010): Fixed Telephone Line. URL: http://www.itu.int/ITU-D/ict/statistics/. – Download vom 12. 03. 2015.

Jans, Matthew (2008): Mode Effects. In: Lavrakas, Paul J. (Hg.): Encyclopedia of Survey Research Methods, Band 1. Thousand Oaks: Sage, S. 475–480.

Kaiser, Günther; Kury, Helmut und Albrecht, Hans-J. (Hg.) (1991): Victims and criminal justice. Band 4. Freiburg: MPI für ausländisches und internationales Strafrecht.

Kalkgraff-Skjak, Knut; Harkness, Janet (2003): Data Collection Methods. In: Harkness, Janet; van de Vijver, Fons J. R. und Mohler, Peter Ph. (2003): Cross-Cultural Survey Methods. Hoboken, NJ: Wiley, S. 79–194.

Kerschke-Risch, Pamela (1993): Gelegenheit macht Diebe – Doch Frauen klauen auch. Opladen: Westdeutscher Verlag.

König, René (1962) (Hg.): Handbuch der empirischen Sozialforschung. Stuttgart: Enke.

Kreuter, Frauke (2002): Kriminalitätsfurcht: Messung und methodische Probleme. Opladen: Leske + Budrich.
Kreuter, Frauke (2008a): Interviewer Effects. In: Lavrakas, Paul J. (Hg.): Encyclopedia of Survey Research Methods, Band 1. Thousand Oaks: Sage, S. 369–371.
Kreuter, Frauke (2008b): Interviewer Variance. In: Lavrakas, Paul J. (Hg.): Encyclopedia of Survey Research Methods, Band 1. Thousand Oaks: Sage, S. 384–385.
Kreuter, Frauke; Presser, Stanley und Tourangeau, Roger (2008c): Social Desirability Bias in CATI, IVR, and Web Surveys. In: Public Opinion Quarterly 72, 5, S. 847–865.
Krosnick Jon A. (1991): Response strategies for coping with the cognitive demands of attitudes measures in surveys. In: Applied Cognitive Psychology, 5, S. 213–236.
Kunz, Franziska (2013): Auswirkungen der Erhebungsanonymität auf Teilnahmebereitschaft und Antwortverhalten in postalischen Befragungen zu selbstberichteter Kriminalität. Ein Methodenexperiment. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht.
Kury, Helmut (1991): Victims of crime – results of a representative telephone survey of 5.000 citizens of the former Federal Republik of Germany. In: Kaiser, Günther; Kury, Helmut und Albrecht, Hans-J. (Hg.), Victims and criminal justice. Band 50. Freiburg: Max-Planck-Institut für ausländisches und internationales. Strafrecht, S. 265–304.
Kury, Helmut (1992): Kriminalität und Viktimisierung in Ost- und Westdeutschland. Ergebnisse der ersten vergleichenden Victim Survey in der ehemaligen DDR und BRD. In: Kury, Helmut (Hg.): Gesellschaftliche Umwälzung. Kriminalitätserfahrungen, Straffälligkeit und soziale Kontrolle. Das Erste deutsch-deutsche kriminologische Kolloquium. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht, S. 141–228.
Kury, Helmut; Dörmann, Uwe; Richter, Harald und Würger, Michael (1992): Opfererfahrungen und Meinungen zur Inneren Sicherheit in Deutschland. Ein empirischer Vergleich von Viktimisierungen, Anzeigeverhalten und Sicherheitseinschätzung in Ost und West vor der Vereinigung. Wiesbaden: Bundeskriminalamt (2. Aufl. 1996).
Kury, Helmut (1994): Zum Einfluss der Datenerhebung auf die Ergebnisse von Umfragen. In: Monatsschrift für Kriminologie und Strafrechtsreform, 77, S. 22–33.
Kury, Helmut (2001): Das Dunkelfeld der Kriminalität. Oder: Selektionsmechanismen und andere Verfälschungsstrukturen. In: Kriminalistik, 55, S. 74–84.

Kury, Helmut; Obergfell-Fuchs, Joachim (2003). Standardinventar für Bevölkerungsbefragungen zu Kriminalität und Kriminalitätsfurcht – Ergebnisse von Pretests. In: Dölling, Dieter; Feltes, Thomas; Heinz, Wolfgang und Kury, Helmut (Hg.): Kommunale Kriminalprävention – Analysen und Perspektiven – Ergebnisse der Begleitforschung zu den Pilotprojekten in Baden-Württemberg. Holzkirchen/Obb.: Felix Verlag, S. 233–249.

Kury, Helmut; Lichtblau, Andrea und Neumaier, Andre (2004): Was messen wir, wenn wir Kriminalitätsfurcht messen? In: Kriminalistik, 58, S. 457–465.

Lensvelt-Mulders, Gerty (2008): Surveying Sensitive Topics. In: de Leeuw, Edith D.; Hox, Joop, J. und Dillman, Don A. (Hg.): International Handbook of Survey Methodology. New York: Psychology Press, Taylor & Francis Group, S. 461–460.

Lessler, Judith T.; Eyerman, Joe und Wang, Kevin (2008): Interviewer Training. In: de Leeuw, Edith D.; Hox, Joop, J. und Dillman, Don A. (Hg.): International Handbook of Survey Methodology. New York: Taylor & Francis Group, S. 442–478.

Loosveldt, Geert (2008): Face-to-Face Interviews. In: de Leeuw, Edith D.; Hox, Joop und Tillman, Don A. (2008): International Handbook of Survey Methodology. New York: Taylor & Francis Group, S. 201–220.

Lohr, Sharon (2008): Coverage and Sampling. In: de Leeuw, Edith D.; Hox, Joop und Dillmann, Don A. (2008): International Handbook of Survey Methodology. New York: Taylor & Francis Group, S. 97–112.

Lyberg, Lars; Biemer, Paul; Collins, Martin; de Leeuw, Edith D.; Dippo, Cathryn, Schwartz, Nobert und Trewin, Dennis (1997): Survey Measurement and Process Quality. New York: Wiley.

Madow, William G.; Olkin, Ingram, und Rubin, Donald B. (Hg.) (1983): Incomplete data in sample surveys. New York: Academic Press.

McNemar, Quinn (1946): Opinion-attitude methodology. In: Psychological Bulletin, 43, S. 289–374.

Müller, Ursula; Schröttle, Monika (2004): Lebenssituation, Sicherheit und Gesundheit von Frauen in Deutschland. Eine repräsentative Untersuchung zu Gewalt gegen Frauen in Deutschland. Im Auftrag des Bundesministeriums für Familie, Senioren, Frauen und Jugend. Bielefeld: Interdisziplinäres Frauenforschungs-Zentrum der Universität Bielefeld. URL: http://fra.europa.eu/DVS/DVT/vaw.php. – Download vom 12. 03. 2015.

Podsakoff, Philip M.; MacKenzie, Scott B.; Podsakoff, Nathan P. und Lee, Jeong-Yeon (2003): Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. In: Journal of Applied Psychology, 88, 5, S. 879–903.

Presser, Stanley (1984): The use of survey data in basic research in the social sciences. In: Turner, Charles F.; Martin, Elizabeth (Hg.), Surveying subjective phenomena. Band 2. New York: Russell Sage Foundation, S. 93-116.

R+V-Versicherung (2014): Die Ängste der Deutschen 2014. URL: https://www.ruv.de/de/presse/download/pdf/aengste-der-deutschen-2014/grafiken-bundesweit.pdf. – Download vom 12. 03. 2015.

Reiss, Albert J. (1967): Studies in crime and law enforcement in major metropolitan areas. President's Commision on Law Enforcement and Administration of Justice, Field Surveys III, Band 1. Washington, D.C.: U.S. Government Printing Office.

Roberts, Caroline (2007): Mixing modes of data collection in surveys: A methodological review. URL: http://eprints.ncrm.ac.uk/418/1/Methods-ReviewPaperNCRM-008.pdf. – Download vom 12. 03. 2015.

Rosenthal, Robert; Rosnow, Ralph L. (1969): Artefact in behavioural research. New York: Academic Press.

Schaeffer, Nora C.; Dykema, Jennifer und Maynard, Douglas W. (2010): Interviews and Interviewing. In: Marsden, Peter V.; Wright, James D. (Hg.): Handbook of Survey Research. Howard House: Emerald Group Publishing Limited, S. 437–470.

Scheuch, Erwin K. (1973): Das Interview in der Sozialforschung. In: König, R. (Hg.): Handbuch der empirischen Sozialforschung, Band 2. Stuttgart: Enke, S. 66–190.

Schiel, Stefan; Dickmann, Christian; Gilberg, Reiner und Malina, Aneta (2013): Repräsentative Bevölkerungsbefragung im Rahmen des BaSiD-Teilvorhabens „Sicherheitsgefährdungen durch Kriminalität". Methodenbericht. Bonn: infas Institut für angewandte Sozialwissenschaften.

Schneekloth, Ulrich; Leven, Ingo (2003): Woran bemisst sich eine „gute" allgemeine Bevölkerungsumfrage? Analysen zu Ausmaß, Bedeutung und zu den Hintergründen von Nonresponse in zufallsbasierten Stichprobenerhebungen am Beispiel des ALLBUS. In: ZUMA-Nachrichten, 53, S. 16–57.

Schnell, Rainer; Kreuter, Frauke (2000): Das DEFECT-Projekt: Sampling-Errors und Nonsampling-Errors in komplexen Bevölkerungsstichproben. In: ZUMA Nachrichten, 24, 47, S. 89–102.

Schnell, Rainer; Kreuter, Frauke (2005): Separating Interviewer and Sampling-Point Effects. In: Journal of Official Statistics, 21, 3, S. 389–410.

Schober, Michael; Conrad, Frederick (1997): Does Conversational Interviewing Reduce Survey Measurement Error? In: Public Opinion Quarterly, 61, S. 287–308.

Schwarz, Norbert; Hippler, Hans-Jürgen; Deutsch, Brigitte und Strack, Fritz (1985): Response Categories: Effects on Behavioural Reports and Comparative Judgments. In: Public Opinion Quarterly, 49, S. 388–395.

Schwind, Hans Dieter; Fetchenhauer, Detlef; Ahlborn, Wilfried und Weiß, Rüdiger (2001): Kriminalitätsphänomene im Langzeitvergleich am Beispiel einer deutschen Großstadt. Bochum 1975-1986-1998. Neuwied: Luchterhand

Selg, Herbert; Klapprott, Jürgen und Kamenz, Rudolf (1992): Forschungsmethoden der Psychologie. Stuttgart: Kohlhammer.

Sessar, Klaus (1992): Wiedergutmachen oder Strafen? Einstellungen in der Bevölkerung und der Justiz. Pfaffenweiler: Centaurus.

Skogan, Wesley G. (1986): Methodological Issues in the Study of Victimization. In: Fattah, Ezzat A. (Hg.): From Crime Policy to Victim Policy: Restoring the Justice System. Basingstoke: Palgrave Macmillan, S. 80–116.

Sparks, Richard F. (1981): Surveys of victimization – An optimistic assessment. In: Tonry, Michael; Morris, Norval (Hg.): Crime and justice: An annual review of research. Vol. 3. Chicago: University of Chicago Press, S. 1–60.

Statistisches Bundesamt; Bundeskriminalamt (2010): Abschlussbericht der deutschen Testerhebung. URL: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-127774. – Download vom 12. 03. 2015.

Stangeland, Per (1996): Effects on Victim Survey Crime Rates. Australian Institute of Criminology; Research Paper. URL: http://www.aic.gov.au/en/publications/previous%20series/proceedings/1-27/~/media/publications/proceedings/27/stangeland.pdf. – Download vom 12. 03. 2015.

Steeh, Charlotte (2008): Telephone Surveys. In: de Leeuw, Edith D.; Hox, Joop und Dillmann, Don A. (2008): International Handbook of Survey Methodology. New York: Taylor & Francis Group, S. 221–238.

Stephan, Egon (1976): Die Stuttgarter Opferbefragung. Wiesbaden: BKA-Forschungsreihe.

U.S. President's Commission on Law Enforcement and Administration of Justice (1967): The challenge of crime in a free society. Washington, D.C.: U.S. Government Printing Office.

Tourangeau, Roger; Rips, Lance J. und Rasinski, Kenneth (2000): The Psychology of Survey Response. Cambridge: Cambridge University Press.

Tourangeau, Roger; Yan, Ting (2007): Sensitive Questions in Surveys. In: Psychological Bulletin, 133, 5, S. 859–883.

Tourangeau, Roger; Smith Tom W. (1996): Asking sensitive questions. The impact of date collection mode, question format and question context. In: Public Opinion Quarterly, 60, S. 275–304.

van Dijk, Jan (2009). Criminological Research in the Framework of the United Nations. In: Schneider, Hans Joachim (Hg.): Internationales Handbuch der Kriminologie. Band 2: Besondere Probleme der Kriminologie. Berlin: De Gruyter, S. 227–253.

van Dijk, Jan; Mayhew, Pat und Killias, Martin (1990): Experiences of Crime around the World: Key findings from the 1989 international crime survey. Deventer: Kluwwer Law and Tacation Publisher.

van Dijk, Jan; Mayhew, Pat; van Kesteren, John; Aebi, Marcelo und Linde, Antonia (2010): Final Report on the Study on Crime Victimisation. URL: http://arno.uvt.nl/show.cgi?fid=113047. – Download vom 12. 03. 2015.

Villmow, Bernhard; Stephan, Egon (1983): Jugendkriminalität in einer Gemeinde. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht.

Wetzels, Peter; Ohlemacher, Thomas; Pfeiffer, Chrstian und Strobl, Rainer (1994): Victimization Surveys: Recent Developments and Perspectives (KFN-Forschungsberichte Nr. 28). Hannover: Kriminologisches Forschungsinstitut Niedersachsen e. V.

Weisburd, Herbert F. (2005): The Total Survey Error Approach. A Guide to the New Science of Survey Research. Chicago: University of Chicago Press.

Wollinger, Gina R.; Dreißigacker, Arne; Blauert, Katharina; Bartsch, Tilman. und Baier, Dirk (2014): Wohnungseinbruch: Tat und Folgen. Ergebnisse einer Betroffenenbefragung in fünf Großstädten (=KFN-Forschungsbericht Nr. 124). Hannover: Kriminologisches Forschungsinstitut Niedersachsen e. V.

Woltman, Henry F.; Turner, Anthony und Bushery, John, M. (1980): A comparison of three mixed mode interviewing procedures in the National Crime Survey. In: Journal of the American Statistical Association, 75, 371, S. 534–543.

UNODC; UNECE (2011): UNODC-UNECE Manual on Victimization Surveys. URL: http://www.unodc.org/unodc/en/data-and-analysis/Manual-on-victim-surveys.html. – Download vom 12. 03. 2015.

Yates, Frank (1933): The analysis of replicated experiments when the field results are incomplete. In: Empire Journal of Experimental Agriculture 1, S. 129–142.

Plädoyer für einen Methoden-Mix: Wie man zu akzeptablen Kosten gute *Crime Surveys* macht

Martin Killias

1 Wo liegt das Problem?

Opferbefragungen oder *Crime (Victimisation) Surveys* kosten relativ viel Geld. Das führt zu vielerlei Sachzwängen. Forscher sehen sich oft gezwungen, die Stichproben ungebührlich zu verkleinern, was dazu führt, dass wichtige Fragen angesichts zu kleiner Fallzahlen nicht mehr beantwortet werden können. Falls die Befragung per Telefon stattfindet, ist die Interviewdauer ein erstrangiger Kostenfaktor, was dazu führt, dass oft wichtige Teile des Fragebogens – wie etwa solche zum Lebensstil – gestrichen werden. Die Stichprobe würde von ihrer Größe her zwar in die Tiefe gehende Analysen zulassen, doch fehlen dann die Angaben, welche die Verteilung von Opfererfahrungen verständlich werden ließen. Falls die für die Finanzierung Verantwortlichen dies erkennen, besteht die Gefahr, dass sie am Ende vorziehen, auf die Befragung überhaupt zu verzichten. Im vorliegenden Beitrag soll gezeigt werden, wie sich diese Stolpersteine erfolgreich umgehen und ein qualitativ befriedigendes Ergebnis erzielen lässt.

2 Die Interviewkosten als entscheidender Faktor

Weder die Verkleinerung der Stichproben noch die Verkürzung der Fragebogen ist zielführend. Damit aber hängt die Machbarkeit qualitativ hochwertiger Befragungen sehr eng mit den Kosten eines einzelnen Interviews zusammen. Als ein Team dreier Forschender – die Engländerin Pat Mayhews (Home Office), Jan van Dijk (Justizministerium der Niederlande, heute Universität Tilburg) und der Autor – sich 1988 zusammenfand, um die Möglichkeiten einer internationalen Befragung zum Thema Kriminalität zu diskutieren, erwies sich die damals neu verfügbare Methode der computer-gestützten Telefoninterviews (CATI) als so etwas wie das Ei des Kolumbus. Zuvor in zwei Surveys in der Schweiz anhand größerer Stichproben (von 3.000 und 3.500) erprobt (Killias 1989), versprach diese Methode erstmals die praktische Durchführbarkeit einer Befragung in einer Vielzahl von Ländern. Neben der Möglichkeit einer weitgehenden Standardisierung durch die CATI-Technologie war der offenkundige Hauptvorteil die finanzielle Tragbarkeit. Mit damals rund 20 heutigen Euro für ein Interview entstanden Kosten, die für die zustän-

digen Ministerien von vierzehn Ländern akzeptabel schienen. Am Ende erschien so innerhalb rund eines Jahres ein Band (van Dijk u. a. 1990) mit einer Vielzahl internationaler Vergleiche, die zuvor so nicht möglich gewesen wären – und auch einigen geläufigen Vorstellungen widersprachen.

Dieser äußerliche Erfolg einer internationalen Initiative blieb nicht ohne Nebengeräusche. Neben legitimer Kritik gab es auch ungewöhnlich heftige Reaktionen, dies vor allem in den Niederlanden, die plötzlich als ein Land mit hohen Kriminalitätsraten dastanden. Da nicht sein kann, was nicht sein darf, suchte man sofort nach Erklärungen bei der gewählten Methode der telefonischen Befragung (CATI) – was nicht viel kostet, so die ungeprüfte Prämisse, konnte ja auch nicht viel wert sein. Auch wurde die in einzelnen Ländern – vor allem in Deutschland – erreichte niedrige Ausschöpfungsrate vorschnell der Methode angelastet, obwohl der bei der Rekrutierung der Zielpersonen betriebene (oder unterlassene) Aufwand die entscheidende Variable darstellte. Die damalige Polemik führte im Laufe der nächsten Jahre zu verstärkter Methodenforschung, doch leider nur vereinzelt zu kontrollierten Experimenten, die allein solche Streitfragen zu entscheiden erlauben (Killias u. a. 2011, Rz. 1124). Wo Experimente durchgeführt wurden – so etwa von Scherpenzeel (1992, 2001) in den Niederlanden, von Kury (1994) in Erfurt und Schwind u. a. (2001) in Bochum –, zeigte sich konsistent, dass die Ergebnisse zwischen verschiedenen Befragungsmethoden weit weniger stark variieren als zuvor vorschnell behauptet worden war. Viel wichtiger sind andere Merkmale einer Befragung (Killias u. a. 2011, Rz. 246), so beispielsweise die Struktur des Fragebogens und insbesondere die Verortung der berichteten Ereignisse in Raum und (vor allem) Zeit. Fragt man beispielsweise direkt nach Ereignissen im Laufe der letzten zwölf Monate, erhält man deutlich (um mehrere Hundert Prozent) überzeichnete Raten (Scherpenzeel 1992, Lucia u. a. 2007), wogegen die Angaben sehr viel realistischer ausfallen, wenn man zunächst nach Erfahrungen in einem längeren Zeitraum (wie etwa den letzten fünf Jahren) fragt und anschließend die Befragten bittet, die berichteten Ereignisse genauer zu verorten: War das im Laufe des letzten Jahres oder schon länger her? War das in Deutschland (oder wo immer) oder in einem anderen Land?

Neue Befragungsmethoden wie CATI (und CAWI[1], dazu unten) garantieren im Vergleich zum persönlichen Interview eine größere Anonymität, was gerade bei sensibleren Themen – wie etwa häuslicher Gewalt – Vorteile verspricht. Kommen die befragte Person und ihre Interviewerin bzw. ihr Interviewer nicht in direkten Kontakt, steht ein Wiedererkennen bei späteren

[1] Für *computer-assisted web interview*, also elektronisch zu beantwortende Onlinefragebogen.

zufälligen Treffen nicht zu befürchten und entfällt das Eindringen in das Wohn- oder Arbeitsumfeld – mit der stets möglichen und oft störenden Anwesenheit weiterer Personen und Gefühlen von Peinlichkeit, die damit verbunden sein können. So betrachtet kann man nur staunen, welcher Aufwand beim neulich publizierten *European Survey on Violence against Women* betrieben wurde, im Rahmen dessen nicht weniger als 42.000 Frauen (d. h. 1.500 in jedem der 28 EU-Länder) persönlich (mit Papierfragebogen oder Fragebogen auf Laptops, CAPI) befragt wurden (Violence Against Women 2014, besonders 16–17). Die Unsummen an Forschungsgeld, die dafür benötigt wurden, werden später bei anderen und möglicherweise nicht weniger relevanten Forschungsvorhaben (gerade auch zu diesem Thema) fehlen. Zudem fehlen jegliche Angaben zu der Antwortrate insgesamt oder in den einzelnen EU-Staaten, weshalb die Ergebnisse auch diesbezüglich schwierig einzuschätzen sind.

3 Notwendige Anpassungen an neue Technologien

Seit dem Durchbruch von 1989 sind 25 Jahre vergangen. In dieser Zeit haben sich die technologischen Voraussetzungen erneut stark verändert. Das gilt beispielsweise auch für Befragungen an Schulen wie beim *International Self-Reported Delinquency Survey* (Junger-Tas u. a. 2010), die sich seit dem Durchbruch der Computer in der schulischen Alltagswelt die Voraussetzungen grundlegend verändert haben. Auch hier erlaubt der Rückgriff auf elektronische Interviews (CAWI) erhebliche Kosteneinsparungen im Vergleich zu klassischen schriftlichen Fragebogen, die in einer Schulstunde von der ganzen Klasse beantwortet werden müssen. Wie kontrollierte Experimente in der Schweiz, in Finnland und neuerdings in Deutschland gezeigt haben (Lucia u. a. 2007; Kivivuori 2007; Baier 2014), sind beide Methoden ungefähr gleichwertig, was die Ergebnisse über selbst berichtete Delinquenz anbelangt.[2] Was die berichteten Opfererfahrungen anbelangt, die vorliegend im Vordergrund des Interesses stehen, erbrachte das Lausanner Experiment (bei 588 in der P&P- bzw. 615 Schülerinnen und Schülern in der Online-Gruppe) nahezu identische Raten für jemals erlebte Viktimisierungen, nämlich 10 (P&P) vs. 9 Prozent (Online) bei Raub, je 5 Prozent bei Erpressung und ebenso viele bei sexuellen Übergriffen, ferner 11 vs. 13 Prozent bei Körperverletzung und schließlich 23 (P&P) vs. 24 Prozent (Online) für alle Opfererfahrun-

[2] Ein weiteres Methodenexperiment war im Rahmen des ISRD-3 in Österreich vorgesehen, konnte schlussendlich aber nicht realisiert werden (Mitteilung von Dr. Patrick Manzoni, Universität Zürich).

gen zusammen.[3] Die Raten sind somit fast identisch und variieren (nicht signifikant) nach beiden Richtungen. Leider werden im deutschen Methodenexperiment keine Ergebnisse hinsichtlich der Befragung der Schülerinnen und Schüler über erlittene Opferbefragungen (via P&P oder online) berichtet.

Jenseits der Vergleichbarkeit der Ergebnisse sollte auch ein forschungsethischer Aspekt nicht vergessen werden. Die Beantwortung eines Fragebogens am Computer garantiert deutlich mehr Anonymität als ein traditioneller schriftlicher Fragebogen. Angesichts der Platzverhältnisse in einem Klassenzimmer und der Erkennbarkeit der einzelnen Fragebogenseiten ist es schlicht unmöglich, diese vor neugierigen Blicken anderer Schülerinnen und Schüler völlig abzuschirmen, wogegen Bildschirme ab einem bestimmten seitlichen Winkel nicht ohne Weiteres einsehbar sind. Ein weiterer Vorteil elektronischer Fragebogen liegt, wie zwei weitere Experimente in der Schweiz (Walser/Killias 2012) und in Finnland (Kivivuori u. a. 2013) gezeigt haben, darin, dass die Anwesenheit eines externen Mitglieds des Forscherteams anstelle oder neben der Lehrperson entbehrlich ist, da die Ergebnisse bei beiden Vorgehensweisen sehr vergleichbar ausfallen. In größeren Ländern, wo Angehörige des Forscherteams unter Umständen große Distanzen zurücklegen müssen, erlaubt die Betreuung der Schülerinnen und Schüler durch eine Lehrperson erhebliche Kosteneinsparungen. Es mag gerade Leserinnen und Leser in Deutschland darum interessieren, dass in der Schweiz und in Finnland nationale Studien zu diesem Thema durchgeführt werden konnten, die trotz großer Stichproben nur einen Bruchteil der Kosten der bekannten Studie des Kriminologischen Forschungsinstituts Niedersachen zu Delinquenz unter Schülern verursacht haben.[4]

Die technologische Entwicklung hat selbstredend nicht nur die Klassenzimmer erfasst, sondern auch die tägliche Kommunikation nachhaltig verändert. So verfügen heute die meisten Menschen in westlichen Ländern über Mobiltelefone, wogegen Haushalte mit einem Festnetzanschluss im Laufe der letzten Jahre zurückgegangen und noch seltener als früher in einem Telefonbuch verzeichnet sind. Dies erschwert die erfolgreiche Anwendung der CATI-Technik vor allem bei jüngeren Altersgruppen massiv – zumindest soweit die Stichprobenziehung aufgrund der Telefonverzeichnisse erfolgt. Der Gedanke

---

3 Viel entscheidender ist, ob man im Fragebogen direkt nach Erfahrungen im Laufe der letzten zwölf Monate oder zunächst nach solchen „jemals“ oder in einem längeren Zeitraum fragt – und sich erst anschließend präziser nach der zeitlichen Einordnung (beispielsweise im Laufe des letzten Jahrs oder ähnlich) erkundigt. Im Lausanner Experiment (Lucia u. a. 2007) resultierten bei der direkten Frage nach den letzten zwölf Monaten um 50 bis über 100 Prozent höhere Raten.

4 Für rund 4.000 Interviews in der dritten Welle des *International Self-Reported Delinquency Survey* (ISRD-3) wurden für die Schweizer Stichprobe rund 100.000 Franken benötigt.

lag daher nahe, die zu Befragenden über einen schriftlichen Fragebogen (ausgewählt über ein Adressregister) zu kontaktieren und um die Beantwortung eines elektronischen Fragebogens zu bitten. Dies war die ursprüngliche Anlage anlässlich eines Pretests für eine EU-weite Befragung, die in Fortführung der früheren *International Crime Victim Survey* (ICVS) im Jahre 2010 in sechs Ländern (Deutschland, Niederlande, Dänemark, Schweden, England und Kanada) durchgeführt wurde (van Dijk 2013). Schlussendlich konnte dieses Design lediglich in zwei Ländern durchgeführt werden, wobei sich extrem niedrige Antwortraten (Response-Raten) von weniger als 10 Prozent zeigten. In den anderen Ländern, in denen nicht eine Bevölkerungsstichprobe, sondern ein Panel – beim Befragungsinstitut – „akkreditierter" Befragter kontaktiert wurde, lag die Ausschöpfung dagegen deutlich höher. Auch wenn die Befragungsinstitute bei der Bildung solcher Panels einen möglichst repräsentativen Bevölkerungsquerschnitt zu erreichen versuchen, darf man vermuten, dass Befragte, die längerfristig (gegen Entgelt oder andere Vorteile) rekrutiert werden und sich für wiederholte Interviews zu den verschiedensten Themen zur Verfügung stellen, eine in mehrerer Hinsicht nicht ganz zufällige Auswahl darstellen.

4 Ein pragmatisches „Experiment" in der Schweiz

Bei der Planung der letzten schweizerischen Opferbefragung von 2011 (Killias u. a. 2011) ging es einerseits darum, die technologische Entwicklung mitzumachen und Onlineinterviews anzustreben (CAWI), andererseits aber auch die unguten Erfahrungen des ICVS-Tests in Bezug auf die Response-Rate zu berücksichtigen. Ein weiteres Ziel bestand darin, an einer zufälligen (und damit repräsentativen) Bevölkerungsstichprobe festzuhalten. Zudem ging es darum, für Personen, die – wie häufig Senioren – per Internet nicht zu erreichen sind oder sonst nicht reagiert haben, eine zweite Kontaktmöglichkeit – via Telefon (CATI) – bereitzuhalten. So wurden Personen, die auf den Einladungsbrief nicht reagiert hatten, nach Ablauf von rund zwei Wochen telefonisch kontaktiert. Wenn sie sich bei dieser Erinnerung für das Ausfüllen des elektronischen Fragebogens entschieden haben, wurden ihnen neue Zugangsdaten mitgeteilt, andernfalls wurde das Interview telefonisch durchgeführt oder zu diesem Zweck ein Termin vereinbart. Diese Methoden-Kombination erwies sich als durchaus erfolgreich. Bei einer Stichprobe von insgesamt rund 15.000 – einschließlich diverser Zusatzstichproben für lokale Sicherheitsstudien über einzelne Kantone – ergab sich eine Response-Rate von 59,5 Prozent. Von den abgeschlossenen Interviews wurden 54 Prozent (oder 32 Prozent der Ausgangsstichprobe) via CAWI durchgeführt, die übrigen 46 Prozent (oder 27 Prozent der Ausgangsstichprobe) via CATI. Die hohe Antwortrate per Internet ist umso bemerkenswerter, als keine Mahnschreiben versandt

wurden, sondern die Zielpersonen nach rund drei Wochen direkt angerufen wurden, um ein CATI-Interview durchzuführen, wozu vorgängig die Telefonnummern herausgesucht werden mussten. Allerdings wurde ihnen beim ersten Anruf anheimgestellt, auf das Internet auszuweichen – und ihnen dazu das Passwort erneut mitgeteilt.

Neben der Möglichkeit, per Internet nicht antwortende Zielpersonen via CATI erneut zu kontaktieren, dürfte auch die Form der Kontaktaufnahme zum Erfolg beigetragen haben. Da die Stichprobe direkt aus den Bevölkerungsregistern gezogen wurde, konnten die Zielpersonen direkt ausgewählt werden. Anders als bei CATI, wo der Festnetzanschluss grundsätzlich zu einem Haushalt führt, entfiel damit das – immer etwas umständliche und nicht unproblematische – zufällige Auswählen einer Zielperson unter den Haushaltsmitgliedern. Theoretisch könnten zwar auf diese Weise zwei Personen innerhalb desselben Haushalts in die Stichprobe gelangen, was bei Delikten gegen den Haushalt (wie etwa Wohnungseinbruch und Fahrzeugdiebstähle) zu Doppelzählungen führen würde. Praktisch ist diese statistische Wahrscheinlichkeit jedoch gering und daher vernachlässigbar.[5] Wichtig war dagegen, dass der Kontaktbrief, in dem den Zielpersonen der Zweck der Befragung und das Passwort mitgeteilt wurden, vom Kommandanten der jeweiligen Kantonspolizei mitunterzeichnet wurde. In einem einzigen Kanton, wo dieses Vorgehen wegen interner Kommunikationspannen nicht zeitgerecht abgeschlossen werden konnte, lag die Antwortrate prompt ungefähr 10 Prozent tiefer. Bei einer breit angelegten Bevölkerungsbefragung zu häuslicher Gewalt im Kanton Genf – für Befragungen ein notorisch schwieriges Umfeld – ergab sich bei demselben Vorgehen wiederum eine Ausschöpfungsrate von 46 Prozent (Killias u. a. 2013). Dieses Vorgehen entsprach weitgehend dem international empfohlenen Prozedere bei einem Methoden-Mix (Dillman u. a. 2014).

5 Höhere Viktimisierungsraten wegen CAWI-Interviews?

In einem kontrollierten Experiment in den Niederlanden, im Rahmen dessen 8.000 Befragte zufällig auf vier verschiedene Arten (CAWI, CATI, postalische Befragung und oder CAPI) interviewt wurden (Buelens u. a. 2012), zeigten sich signifikant höhere Antwortraten bei CATI im Vergleich zu CAWI

[5] Wenn von rund 3,5 Millionen Privathaushalten rund 15.000 Personen (oder eine von 233) befragt wird, ist theoretisch damit zu rechnen, dass etwa 32 „Paare“ aus demselben Haushalt in der Stichprobe figurieren. In Wirklichkeit werden es allerdings eher 10 bis 15 sein, weil (1) ca. 35 % der Privathaushalte Einzelhaushalte sind und weil (2) vorliegend ungefähr 40 % aus irgendwelchen Gründen nicht befragt werden konnten. Dies führt zu vernachlässigbaren Schätzfehlern bei der Hochrechnung der Ergebnisse auf die Gesamtbevölkerung.

(von 61 gegenüber 29 Prozent). Demgegenüber zeigten sich bei der schweizerischen Befragung (wie oben berichtet) geringfügig höhere Antwortarten bei CAWI im Vergleich zu CATI. Dies könnte darauf zurückzuführen sein, dass via CATI in der schweizerischen Befragung nur kontaktiert wurde, wer zunächst einmal auf den Kontaktbrief und die Einladung zur Beantwortung eines Onlinefragebogens nicht reagiert hatte. Neben Personen ohne Internetanschluss betraf dies wohl vor allem Befragte, die von Anfang an weniger kontaktbereit waren und insofern eine „negative" Auswahl darstellten. Die viel schlechteren Antwortraten in den Niederlanden bei CAWI erklären sich wohl damit, dass bei dieser Methode die Befragten über E-Mail kontaktiert wurden, was das Verweigern im Vergleich zu CATI mit einem immerhin persönlichen Telefonkontakt deutlich erleichtert haben dürfte. Es entspricht darüber hinaus der Alltagserfahrung, dass Kontaktversuche via E-Mail deutlich weniger erfolgreich sind als solche über Telefon oder andere Mittel (Dillman u. a. 2009).

Wie nun hat sich die Befragungsmethode auf die Ergebnisse der Befragung ausgewirkt? Im erwähnten niederländischen Experiment (Buelens u. a. 2012) berichteten via CAWI Befragte höhere Viktimisierungsraten als telefonisch befragte Personen. Das könnte damit zusammenhängen, dass man anlässlich eines kürzeren Telefongesprächs leichter weniger schwere oder minder denkwürdige Ereignisse eher vergisst als beim Ausfüllen eines Formulars. Allerdings zeigte sich auch, dass die Onlineinterviews im Durchschnitt etwas weniger lange dauerten als die telefonischen, was nicht unbedingt dafür spricht, dass man am Telefon Ereignisse, die aus der Sicht der Befragten weniger wichtigen waren, weniger vollständig berichtet hätte.

Bei der schweizerischen Befragung zeigte sich eine ähnliche Tendenz, wie die in *Tabelle 1* berichteten Ereignisse zeigen.

Tabelle 1:

Einfluss (*Odds Ratios*, OR) der Befragungsmethode (CAWI vs. CATI) auf berichtete Opfererfahrungen unter Kontrolle demografischer Variablen (Killias 2012)

2-Jahre (2009-10)	Diebstahl persönlicher Gegenstände			Diebstahl von oder aus Fahrzeugen			Einbruch (mit Versuchen)			Gewaltdelikte		
	Coeff.	OR	*Sig.*	Coeff.	OR	*Sig.*	Coeff.	OR	*Sig.*	Coeff.	OR	*Sig.*
Geschlecht	-0,022	0,979	*n.s.*	-0,091	0,913	*n.s.*	0,107	1,113	*n.s.*	0,139	1,149	*n.s.*
Alter	-0,573	0,564	***	-0,519	0,595	***	0,174	1,190	*n.s.*	-0,633	0.531	***
Bildungsniveau	0,454	1,575	**	0,161	1,175	*n.s.*	0,519	1,680	*	0,232	1,261	*n.s.*
Einkommen	-0,012	0,988	*n.s.*	0,070	1,072	*n.s.*	0,023	1,024	*n.s.*	0,230	1,259	*n.s.*
CAWI vs. CATI	0,314	1,369	*n.s.*	0,221	1,248	*n.s.*	0,551	1,734	*	0,418	1,519	*
CAWI vs. CATI (bivariat)	0,507	1,660	**	0,386	1,471	**	0,595	1,812	**	0,580	1,786	***

* p<0.05, ** p<0.01, *** p<0.001

Wie man vermuten konnte, sind die Unterschiede zwischen CAWI und CATI nach Kontrolle demografischer Variablen (letzte Zeile) – hier Geschlecht, Alter und Bildung, die sowohl die Erreichbarkeit per Internet wie auch die Wahrscheinlichkeit von Opfererfahrungen beeinflussen – nur noch bei Einbruch und Gewaltdelikten (wie Raub und Sexualdelikten) signifikant. Dies überrascht, weil bei diesen Straftaten im Vergleich zu Diebstählen weniger plausibel erscheint, dass man sie während eines Interviews zu erwähnen vergisst, und ein Methodeneffekt daher weniger zu erwarten wäre (Guzy/Leitgöb 2015). Zudem stellt sich die Frage, wieso Scherpenzeel (1992; 2001), Kury (1994), Schwind u. a. (2001) sowie Lucia u. a. (2007) und Kivivuori (2007) in ihren Experimenten keinen analogen Methodeneffekt finden konnten.

Es drängt sich darum die Vermutung auf, dass weniger ein Methoden- als ein Selektionseffekt im Spiel gewesen sein könnte. In der schweizerischen Opferbefragung waren die telefonisch Befragten *a priori* wohl weniger motiviert, an der Befragung mitzuwirken, als die per Internet Antwortenden. Die Motivation, an einer Befragung mitzumachen, korreliert nun aber, wie seit Langem bekannt ist, mit der Betroffenheit der Zielperson (Farrington u. a. 1990). Wer sich vom Thema einer Befragung angesprochen fühlt, weil er oder sie allenfalls vor nicht allzu langer Zeit selbst Opfer einer Straftat geworden ist, wird wohl eher sofort oder, wie die Erfahrung bestätigte, in den nächsten Tagen den Onlinefragebogen beantworten. Dieser Selektionseffekt dürfte auch beim holländischen Experiment eine Rolle gespielt haben, wie die deutlich niedrigere Antwortrate unter den per Mail Kontaktierten vermuten lässt. Ein „korrekter“ Test der beiden Hypothesen – des Methoden- oder des Selektionseffekts – hätte derart vonstattengehen müssen, dass man die zu Befragenden zu einem Interview (gleich welcher Art) ins Befragungslabor aufgeboten und dort dann bei ihrer Ankunft zufällig auf die verschiedenen Befragungsmethoden verteilt hätte. Bei diesem Vorgehen wäre sichergestellt gewesen, dass alle Befragten in gleicher Weise zum Mitmachen motiviert sind – was immer dann an Unterschieden herauskäme, wäre eindeutig als Effekt der Befragungsmethode identifiziert.

6 Konsequenzen für die Durchführung weiterer Opferbefragungen

Im Hinblick auf die Durchführung weiterer Sicherheits- oder Opferbefragungen ergeben sich daraus einige praktische Folgerungen. Zunächst ist es nicht vertretbar, einem Methoden-Monismus zu huldigen – reine CATI- wie auch reine CAWI-Befragungen ergäben schlechte Response-Rates und, was wohl schwerer wiegt, möglicherweise verzerrte Ergebnisse, da die Antwortenden bei beiden Methoden nicht unbedingt miteinander vergleichbar sind. Für die in diesem Jahr (2015) angebahnte weitere schweizerische Befragung planen

wir daher erneut einen „Mix“ aus CAWI und CATI. Wegen des Kostendrucks – die Frage ist bekanntlich nicht, was wünschbar sein mag, sondern wie man mit bescheidenem Budget eine optimale Qualität der Studie sicherstellen kann – und der mehr als doppelt so hohen Kosten der CATI- im Vergleich zu den CAWI-Interviews planen wir deshalb, zunächst die aus Bevölkerungsregistern ausgewählten Zielpersonen anzuschreiben und für ein Onlineinterview zu motivieren. Dazu soll der Kontaktbrief wiederum von einer lokal bekannten Persönlichkeit (etwa dem lokalen Polizeichef oder einem regionalen Regierungsvertreter) mitunterzeichnet werden. Nach rund zwei Wochen werden Mahnschreiben an alle Adressaten versandt, die den Fragebogen noch nicht retourniert haben. Das System der Passwörter wird gestatten, völlig anonym die Personen, die geantwortet haben, aus der Kartei zu entfernen. Nach dem Mahnschreiben werden die noch verbleibenden Zielpersonen für ein kurzes CATI-Interview kontaktiert. Da die Kosten eines telefonischen Interviews einerseits direkt von der Gesprächsdauer abhängen, andererseits aber vor allem interessiert, um wie viel tiefer unter dieser Stichprobe „säumiger“ Zielpersonen die Opferraten liegen, genügt ein Kurzfragebogen mit den eigentlichen Kernfragen zur eigenen Viktimisierung. Da dafür im Allgemeinen nicht mehr als fünf Minuten benötigt werden, lassen sich mit dieser Verkürzung der Dauer die „Stückkosten“ von knapp 40 € für ein CATI-Interview um mehr als die Hälfte reduzieren. Wir hoffen, auf diese Weise zu akzeptablen finanziellen Bedingungen – rund 80.000 CHF (ca. 75.000 €) sind dafür vorgesehen – eine nationale Zufallsstichprobe von rund 2.000 Personen befragen und gleichzeitig eine relativ hohe Ausschöpfungsrate von 50 Prozent erreichen zu können. Die vorläufigen Ergebnisse der noch laufenden Befragung deuten darauf hin, dass dieses Ziel erreicht werden kann.

7 Zusammenfassung

– Befragungen zu seltenen Ereignissen – dazu gehören glücklicherweise schwerere Verbrechen – erfordern große Stichproben, wenn aussagekräftige Ergebnisse gefunden werden sollen. Da die Forschungsmittel regelmäßig begrenzt sind, ergibt sich die Stichprobengröße aus der Division des Gesamtbudgets durch die Kosten pro Interview. Gelingt es, diese zu senken, resultieren eine größere Stichprobe und damit eine aussagekräftigere, auch Differenzierungen ermöglichende Untersuchung.

– Mitte der Achtzigerjahre kamen computergestützte Telefoninterviews (CATI) auf den Markt. Ihre Kosten betrugen damals wie heute je nach Dauer zwischen 10 und 20 Prozent eines klassischen persönlichen Interviews. Für Befragungen zu Erfahrungen mit Kriminalität ergab sich aus dieser neuen Technologie ein entscheidender Vorteil. Größere nationale

Studien und die verschiedenen internationalen Befragungen wären ohne die CATI-Methode niemals durchführbar gewesen.

– Diverse Experimente, bei denen die Befragten zufällig per Telefon, persönlich oder brieflich befragt wurden, ergaben keine eindeutigen systematischen Unterschiede. Man kann somit davon ausgehen, dass die Methoden relativ gleichwertig sind. Allerdings haben persönliche Interviews gerade bei etwas sensiblen Themen den schwerwiegenden Nachteil, dass Interviewer und befragte Person sich gegenübersitzen und damit für einander nicht mehr anonym sind. In kleinstädtischen Verhältnissen kann dies sehr wohl eine Rolle spielen.

– Schülerinnen und Schüler werden – zu ihren Erfahrungen als Opfer wie auch als Täter(innen) von Straftaten – seit Langem schon mittels schriftlicher Fragebogen befragt. Mit dem Zwang zu Einsparungen zeigte sich auch hier, dass solche Interviews wesentlich kostengünstiger über den Computer (online) durchgeführt werden können. Ein größeres Methodenexperiment mit mehr als 1.000 Schülern und Schülerinnen in Lausanne (Schweiz) zeigte auch hier, dass sich die beiden Methoden hinsichtlich der Ergebnisse kaum unterscheiden. Hingegen garantiert das Ausfüllen eines elektronischen Fragebogens am Computer eine unvergleichlich größere Anonymität als die Beantwortung eines schriftlichen Fragebogens, dessen Layout auch aus Distanz oft deutlich zu erkennen ist.

– Neue Technologien stehen zunehmend auch für die Befragung Erwachsener zur Verfügung. In einer Welt, in der die meisten Privathaushalte über einen Internetanschluss verfügen, bietet es sich förmlich an, solche Befragungen über das Internet durchzuführen. Bei der schweizerischen Opferbefragung von 2011 haben 32% der Befragten bereits auf ein erstes Ankündigungsschreiben hin den Fragebogen online abgerufen und ausgefüllt. Weitere, oft ältere Personen wurden anschließend noch über Telefon (CATI) befragt (27 %), was zusammen eine Rücklaufquote von nahezu 60 % ergab. Dieser im internationalen Vergleich sehr gute Erfolg kam zustande, weil im Briefumschlag des Ankündigungsschreibens auch noch ein Brief des lokalen Polizeichefs figurierte, in dem dieser Sinn und Zweck der Befragung näher erläuterte.

– Obwohl bei experimentellen Methodentests kaum signifikante Unterschiede zwischen mehreren Methoden zutage traten, zeigte sich bei der schweizerischen Opferbefragung von 2011 – ähnlich wie in anderen Studien – dass die Opferraten unter Befragten, die spontan online geantwortet hatten, höher lagen als unter den telefonisch Befragten. Letztlich muss dies aber kein Methodeneffekt sein – viel plausibler scheint ein Selektions-

effekt insofern, als dieses Ergebnis darauf zurückzuführen ist, dass diejenigen, die sich sofort an den Computer setzten und den Fragebogen online beantworteten, möglicherweise motivierter waren als Befragte, die die Dinge zunächst liegen ließen und daraufhin persönlich für ein CATI-Interview kontaktiert wurden.

– Um solche Zweifel auszuräumen, würde sich ein Methodenexperiment anbieten, indem zunächst alle Befragten eingeladen würden, im Labor des Meinungsforschungsinstituts zu erscheinen, um dort interviewt zu werden. Nach Eintreffen der Versuchspersonen würden diese zufällig auf die zu vergleichenden Methoden verteilt. Da bei dieser Untersuchungsanlage die Trägheit in der Reaktion der Zielpersonen neutralisiert wäre, müssten alle auftretenden Unterschiede in den Ergebnissen dem Einfluss der verschiedenen Methoden zugeschrieben werden.

– Solange ein solches Methodenexperiment nicht durchgeführt ist, empfiehlt sich als Ausweg ein Methoden-Mix, bei dem die Zielpersonen zunächst brieflich angeschrieben und erst dann, wenn Sie sich darauf ein- oder zwei Wochen lang nicht gemeldet haben, telefonisch befragt würden. Wichtig erscheint, dass dem Briefumschlag ein (zweites) Schreiben beiliegt, in dem der lokale Polizeichef Sinn und Zweck der Befragung für die künftige Ausrichtung der Polizeiarbeit darlegt. Wie unsere Erfahrungen bei der Befragung von 2011 ergeben haben, erhöht sich mit diesem Vorgehen die Rücklaufquote um rund 10 %.

8 Literatur

Baier, Dirk (2014): Computergestützte vs. schriftliche Dunkelfeldbefragung: Ergebnisse eines Methodenexperiments. In: Eifler, Stefanie; Pollich, Daniela (Hg.): Empirische Forschung über Kriminalität. Wiesbaden: Springer Fachmedien, S. 123–148.

Buelens, Bart; van der Laan, Jan; Schouten, Bary; van den Brakel, Jan; Burger Joep and Klausch, Thomas (2012): Disentangling mode-specific selection and measurement bias in social surveys. The Hague: Statistics Netherland.

Dillman, Don A.; Phelps, Glenn; Tortora, Robert; Swift, Robert; Kohrell, Julie; Berck, Jodi und Messer, Benjamin L. (2009): Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. In: Social Science Research, 38, 1, S. 1–18.

Dillman, Don; Smyth, Jolene und Leah, Christian (2014): Internet, Phone, Mail, and Mixed Mode Surveys: The Tailored Design Method. 4. Aufl. New York: Wiley.

Farrington, David P.; Gallagher, Bernhard; Morely Linda; Ledger Raymond J. und West, Donald J. (1990): Minimizing attrition in longitudinal research: Method of Tracing and Securing cooperation in a 24-year follow-up study. In: Magnusson, David; Bergman, Lars R. (Hg.): Data Quality in Longitudinal Analysis. Cambridge University Press, S. 122–147.

Guzy, Nathalie; Leitgöb, Heinz (2015): Mode-effects in online and telephone victimization surveys. In: International Review of Victimology, 21, 1, S. 101–131

Junger-Tas, Josine; Marshall, Ineke H.; Enzmann, Dirk; Killias Martin; Gruszczynska, Beata und Steketee, Majone (2010): Juvenile Delinquency in Europe and Beyond: Results of the Second International Self-Report Delinquency Study. Berlin, New York: Springer.

Killias, Martin (1989): Les Suisses face au crime: Leurs expériences et attitudes à la lumière des sondages suisses de victimisation. Grüsch (Schweiz): Ruegger.

Killias, Martin (2012): Innovations in methodology and conservative reflexes among researchers: Some anecdotes from the First International Crime Victimisation Surveys (ICVS) and beyond. In: Groenhuijsen, Marc; Letschert, Rianne und Hazenbroek, Sylvia (Hg.): Liber amicorum prof. dr. mr. J. J. M. van Dijk. Nijmegen: Wolf Legal Publishers, S. 207–216.

Killias, Martin; Walser, Simone und Biberstein, Lorenz (2013): Étude cantonale de victimisation suite à des violences conjugales ou familiales. In: Bourgoz, David; Merenda, Florence; Delhumeau-Cartier, Cecile; Walser, Simone; Biberstein, Lorenz und Killias, Martin (Hg.): La violence domestique en chiffres – année 2012. Genf: OCSTAT, S. 11–23.

Killias, Martin; Kuhn, André und Aebi, Marcelo F. (2011): Grundriss der Kriminologie. Eine europäische Perspektive. 2. Aufl. Bern: Stämpfli.
Killias, Martin; Staubli, Silvia; Biberstein, Lorenz und Iadanza, Sandro (2011): Studie zur Kriminalität und Opfererfahrungen in der Schweiz. Universität Zürich: Kriminologisches Institut.
Kivivuori, Janne (2007): Delinquent Behaviour in Nordic Capital Cities. Helsinki: Scandinavian Research Council for Scandinavia/National Research Institute of Legal Policy, Finland.
Kivivuori, Janne; Salmi, Venla und Walser, Simone (2013): Supervision mode effects in computerized delinquency surveys at school: Finnish Replication of a Swiss Experiment. Journal of Experimental Criminology, 9, 1, S. 91–107.
Kury, Helmut (1994): Zum Einfluss der Art der Datenerhebung auf die Ergebnisse von Umfragen. Monatsschrift für Kriminologie und Strafrechtsreform, 77, 1, S. 22–33.
Lucia, Sonia; Herrmann, Leslie und Killias, Martin (2007): How important are interview methods and questionnaire designs in research on self-reported juvenile delinquency? An experimental comparison of Internet vs. paper-and-pencil questionnaires and different definitions of the reference period. Journal of Experimental Criminology, 3, 1, S. 39–64.
Scherpenzeel A. (1992): Response effecten in slachtoffer-enquêtes: Effecten van vraagformulering en dataverzamelingsmethode. Tijdschrift voor criminology, 34, 4, S. 296–305.
Scherpenzeel, Annette (2001): Mode effects in panel surveys: A comparison of CAPI and CATI. Neuenburg: Bundesamt für Statistik (Nr. 448-0100).
Schwind, Hans-D.; Fetchenhauer, Detlef; Ahlborn, Wilfried und Weiss, Rüdiger (2001): Kriminalitätsphänomene im Langzeitvergleich am Beispiel einer deutschen Großstadt (Bochum 1975–1986–1998). Neuwied: Luchterhand.
Van Dijk, Jan J. M. (2013): The International Crime Victims Survey 1988–2010: Latest results and prospects (ICVS Newsletter April 12, 2013). URL: http://www3.unil.ch/wpmu/icvs/2013/04/445 – Download vom 14. 01. 2015.
Van Dijk, Jan J. M.; Mayhew, Pat und Killias, Martin (1990): Experiences of Crime across the World. Key findings of the 1989 International Crime Survey. Deventer/Boston: Kluwer.
European Union Agency for Fundamental Rights (2014): Violence Against Women: An EU-wide Survey. Wien: European Union Agency for Fundamental Rights.
Walser, Simone; Killias, Martin (2012): Who should supervise students during self-report interviews? A controlled experiment on response behavior in online questionnaires. In: Journal of Experimental Criminology, 8, 1, S. 17–28.

Anzeigequoten als Indikator des Nichtwissens: Mess- und Konstruktionsprobleme

Dirk Enzmann

1 Einleitung

Von Thorsten Sellin stammt die berühmte Argumentationsfigur, dass „[...] der Wert einer Kriminalitätsrate als Indikator in dem Maße abnimmt, wie der Abstand des Verfahrens zum kriminellen Akt selbst zunimmt." (Sellin 1931, 346, Übers. d. Verf.) Demnach sind Verurteiltenziffern als Indikator der Kriminalitätslage weniger valide als polizeiliche Statistiken der Delikte und Tatverdächtigen. Sellin bezieht sich hier nur auf offizielle Statistiken von Polizei und Justiz. In Fortführung des Gedankens könnte man aber argumentieren, dass Polizeistatistiken wiederum weniger valide sind als die unmittelbaren Berichte von Opfern im Rahmen von Viktimisierungsstudien (die jedoch als systematisch erhobene Daten zu Sellins Zeiten noch nicht zur Verfügung standen).

Die Analyse des Umfangs des Anzeigeverhaltens (der Anzeigequote) hat für empirische Studien zur Kriminalitätslage eine besondere Bedeutung. Die wesentlichen Quellen unseres Wissens zur Kriminalitätslage sind Viktimisierungsbefragungen und Hellfeldstatistiken der Polizei. Allerdings finden sich regelmäßig substanzielle Unterschiede zwischen diesen beiden Datenquellen, sowohl was den geschätzten Umfang der Kriminalität betrifft als auch hinsichtlich der Veränderungen von Kriminalitätsraten im Zeitverlauf. Ohne wiederholte Viktimisierungsstudien, mit denen zugleich auch das Anzeigeverhalten erfasst wird, ließe sich nicht beurteilen, inwiefern eine Veränderung im Hellfeld der Kriminalität eine tatsächliche Veränderung der Kriminalitätslage oder eher ein verändertes Anzeigeverhalten widerspiegelt. Das Anzeigeverhalten hat deshalb eine so große Bedeutung für die Erklärung veränderter Kriminalitätsraten, weil die polizeilich registrierte Kriminalität zu einem wesentlichen Teil auf der Menge *angezeigter* Delikte beruht. So wird geschätzt, dass zwischen 77 % und 96 % der in den Polizeistatistiken registrierten Kriminalität auf Anzeigen nicht formeller Instanzen (z. B. durch Opfer, deren Angehörige oder Zeugen) beruht (Coleman/Moynihan 1996; Feltes 2009). Dies gilt jedoch deutlich weniger für sogenannte Kontrolldelikte wie Verstöße gegen das Betäubungsmittelgesetz, bei denen Kontrollaktivitäten der Polizei und insofern Anzeigen „von Amts wegen" eine große Rolle spielen.

Sich ändernde Diskrepanzen zwischen Hell- und Dunkelfelddaten können allerdings nicht nur durch eine Veränderung des Anzeigeverhaltens verursacht werden, sondern auch durch eine Veränderung der polizeilichen Registrierungspraxis. Auch für letztere lassen sich durch eine Betrachtung des Anzeigeverhaltens Hinweise gewinnen – nämlich wenn die Diskrepanzen von Hell- und Dunkelfelddaten nicht durch Anzeigeverhalten und die Kontrollaktivitäten der Polizei erklärt werden können und zugleich angenommen werden kann, dass das Volumen angezeigter Delikte valide erfasst worden ist. Da diese Annahme aber immer mit Unsicherheit behaftet ist, gilt dies auch für die aus dem Vergleich gewonnenen Hinweise auf Veränderungen der polizeilichen Registrierungspraxis.

Auf die Fragen, wie sich Anzeigequoten in Deutschland deliktspezifisch darstellen und im Zeitverlauf verändert haben, welche Rolle Anzeigequoten und polizeiliche Registrierungspraxis für die Diskrepanz von Hell- und Dunkelfelddaten spielen sowie auf Gründe von Anzeige und Nichtanzeige wird in Band 1 ausführlich eingegangen (Enzmann 2015). In diesem Kapitel werden speziell Probleme der Messung und Analyse von Anzeigequoten thematisiert. Dabei werden zunächst Varianten der Operationalisierung und Erfassung von Anzeigequoten dargestellt sowie die Abhängigkeit der Höhe der Anzeigequote von der Messmethode anhand eines Beispiels illustriert. Anschließend werden einige Empfehlungen zur Erfassung von Anzeigequoten gegeben und Techniken zum Umgang mit inzidenzbasierten Anzeigequoten vorgestellt.

2 Varianten der Erfassung von Anzeigequoten

2.1 Prävalenz- versus inzidenzbasierte Anzeigequoten

Eine Analyse der Anzeigequoten dient dazu, im Rahmen von Dunkelfeldstudien das Volumen der kriminellen *Ereignisse* abzuschätzen, von denen die Polizei erfährt und die Eingang in die PKS finden können (also üblicherweise nicht, den Prozentsatz der *Opfer*, die Anzeige erstatten, zu bestimmen). Zu diesem Zweck können in Viktimisierungsstudien prinzipiell zwei unterschiedliche Methoden benutzt werden. Im Anschluss an die Frage zur Anzahl der Delikte, die den Befragten im Referenzzeitraum (z. B. im letzten Jahr) widerfahren sind, kann man erfragen:

1. pro Delikt(skategorie) die Anzahl der Ereignisse (Viktimisierungen), von der die Polizei erfahren hat oder die der Polizei angezeigt wurden (Anzeigeinzidenz);

2. ob das letzte (jüngst geschehene) Delikt der Polizei mitgeteilt oder angezeigt wurde.[1]

Bei der ersten Methode kann pauschal nach der Anzahl der Viktimisierungen gefragt werden, die der Polizei angezeigt[2] wurden (bloße Häufigkeit), oder für jedes Ereignis einzeln die Tatsache einer Anzeige abgefragt werden. Bei der zweiten Methode gibt es einerseits die Variante, pro Delikt(skategorie) die Anzeige des letzten Delikts abzufragen oder nur eine Frage nach dem letzten Delikt über alle Delikte/Deliktskategorien hinweg zu stellen.[3] Bei Methode 1 wird die Menge aller Anzeigen erhoben (Anzeigeinzidenz), während bei Methode 2 (letztes Delikt) nur für eine Stichprobe der jüngsten Viktimisierung(en) das Anzeigeverhalten erfasst wird.

Würde die Psychologie des Antwortverhaltens keine Rolle spielen, würden beide Methoden – abgesehen von Zufallsschwankungen – zu gleichen Anzeigequoten führen, wobei allerdings der Stichprobenfehler (und damit das statistische Konfidenzintervall) bei Methode 2 größer ausfallen würde, da hier nur eine Teilmenge der Viktimisierungsereignisse berücksichtigt wird. Die Anzeigequote der Population würde also mit Methode 2 mit einer größeren Unsicherheit geschätzt werden.

Tatsächlich spielen aber kognitive Prozesse für das Antwortverhalten eine große Rolle. Hier ist insbesondere das Phänomen des Telescoping zu nennen (Sudman/Bradburn 1973; Skogan 1975; Averdijk/Elffers 2012). Darunter wird die Neigung von Befragten verstanden, ein Ereignis zeitlich nicht korrekt zu verorten: Die Verschiebung eines weiter zurückliegenden Ereignisses in den Referenzzeitraum wird als Vorverlagerung (*forward telescoping*) bezeichnet, der umgekehrte Fall als Rückverlagerung (*backward telescoping*). Zu beachten ist, dass Telescoping auch innerhalb des Referenzzeitraums stattfinden kann, z. B. wenn eine Person zum letzten Ereignis befragt wird, sie

---

1 Auch hier sind unterschiedliche Varianten möglich. So wird z. B. im CSEW (Crime Survey for England and Wales) pro Person zu insgesamt maximal sechs Vorfällen eine Mitteilung an die Polizei erfasst (siehe unten). Bei Seriendelikten werden (in jeweils sechs Deliktskategorien) nur maximal fünf Ereignisse registriert und nur Details zum jüngsten Delikt erfragt, dabei u. a. die Mitteilung an die Polizei (Office for National Statistics 2015, 15). Zur kritischen Bewertung der Kappungsgrenze bei fünf Ereignissen siehe Farrell und Pease (2007).

2 Im Folgenden wird nicht zwischen der Mitteilung an die Polizei und einer förmlichen Anzeigeerstattung unterschieden – entgegen einem verbreiteten Missverständnis kann eine Anzeigenannahme auch bei Antragsdelikten nicht verweigert werden und erfordert keine besondere Form (§ 158, Abs. 1 StPO).

3 Schließlich wäre auch denkbar, statt nach dem letzten Delikt nur nach dem schwersten Delikt zu fragen. Dies ist jedoch auf keinen Fall zu empfehlen, da die Schwere des Delikts eines der stärksten Motive für eine Anzeigeerstattung ist (siehe Enzmann 2015) und auf dieser Frage basierende Anzeigequoten systematisch zu hoch ausfallen würden.

sich bei der Antwort aber auf ein weiter zurückliegendes Ereignis im Referenzzeitraum bezieht.

Werden Ereignisse aufgrund ihres Schweregrades oder anderer damit verbundener besonderer Merkmale (z. B. die Tatsache einer Anzeige) besser erinnert, können sie Befragten als weniger weit zurückliegend erscheinen (Brown u. a. 1985). Zwar ist über die Größe der zu erwartenden Verzerrungen durch Telescoping wenig bekannt (Guzy/Leitgöb 2015, 105), es ist aber plausibel, anzunehmen, dass zeitliche Vorwärtsverlagerungen bei Delikten, die der Polizei angezeigt worden sind, stärker ausgeprägt sind als bei nicht angezeigten Delikten. Anzeigequoten sollten demnach bei der Frage zur Anzeige beim letzten Delikt (Methode 2) höher ausfallen als bei Schätzungen, die auf Inzidenzangaben basieren (Methode 1). Denkbar ist auch, dass Angaben zum letzten Delikt sich je nach Frage nicht auf ein und dasselbe Delikt beziehen. So können Auskünfte zu den Tatumständen auf das tatsächlich letzte Delikt bezogen sein, während bei der Antwort auf die Frage nach einer Anzeige bei der Polizei auf das letzte angezeigte Delikt „zurückgegriffen“ wird (hier als selektives Telescoping bezeichnet), was ebenfalls zu erhöhten Maßen der Anzeigequote führt.

Obwohl inzidenzbasierte Anzeigequoten vermutlich weniger durch Telescopingeffekte verzerrt sind, ergeben sich bei ihrer Konstruktion ebenfalls Probleme. So kann gegen die Methode eingewandt werden, dass Angaben zur Anzahl der Viktimisierungsereignisse fehlerbehaftet sind, insbesondere wenn es sich um weniger schwerwiegende und häufigere Delikte handelt oder wenn Personen befragt werden, die sich nur geringe Mühe bei der Genauigkeit der Angaben geben. Prinzipiell existiert dieses Problem auch bei Angaben zur Anzahl der *angezeigten* Ereignisse (und in erster Linie dann, wenn die Anzeige nicht für jedes Ereignis einzeln erfragt wird), allerdings ist hier das Problem wesentlich geringer, da diese Ereignisse besser erinnert werden können und zugleich seltener sind. Daneben gibt es das nicht unwesentliche praktische Problem, dass die Berechnung von inzidenzbasierten Anzeigequoten mit dem verbreiteten Statistikpaket SPSS nicht unmittelbar möglich ist bzw. fortgeschrittene Programmierkenntnisse erfordert,[4] wenn nicht manuelle Berechnungsschritte zwischengeschaltet werden, da hierbei von der Ebene der Befragten auf die Ebene der Ereignisse gewechselt werden muss. Das ist einer der Gründe, aus denen statt der inzidenzbasierten Anzeigequote, die das Verhältnis der Summe der angezeigten Delikte zur Summe der Viktimisierungsereignisse darstellt und als Prozent der Viktimisierungen angegeben wird, gelegentlich der Einfachheit halber eine prävalenzbasierte Anzeigequote be-

[4] Bei Statistikpaketen wie *R* oder Stata ist demgegenüber eine Weiterverarbeitung von Aggregatstatistiken ohne Umwege möglich.

rechnet wird, die angibt, wieviel Prozent der Opfer eines Referenzzeitraums mindestens eins der Delikte angezeigt haben. Da im Allgemeinen nicht alle Delikte angezeigt werden, im Referenzzeitraum aber durchaus Mehrfachviktimisierungen berichtet werden, muss die prävalenzbasierte Anzeigequote in der Regel höher ausfallen als die inzidenzbasierte. Die prävalenzbasierte Anzeigequote ist also nur ein schlechter Ersatz für eine inzidenzbasierte Quote, da sie bei Mehrfachviktimisierungen keine Auskunft über das Volumen der Ereignisse gibt, das der Polizei bekannt geworden ist.

Anzeigequoten, die auf Inzidenzen basieren, unterscheiden sich von Anzeigequoten des letzten Delikts auch in dem höheren statistischen Aufwand, der nötig, ist, um Konfidenzintervalle der Maße zu berechnen. Allerdings ist dies wohl nicht der Grund dafür, dass inzidenzbasierte Anzeigequoten seltener benutzt werden, da Konfidenzintervalle überhaupt nur viel zu selten berichtet werden (siehe Abschnitt 3.4).

2.2 Aktuelle Beispiele der Messung von Anzeigequoten

Im Folgenden werden einige aktuelle Beispiele der Erfassung von Anzeigequoten vorgestellt. Die dazu verwendeten Instrumente dienen nicht primär dazu, Anzeigeverhalten selbst zu erfassen, sondern werden im Zusammenhang mit Viktimisierungsstudien oder Studien selbstberichteter Delinquenz (letztere fokussiert auf das Verhalten Jugendlicher) eingesetzt, in denen auch Viktimisierungsereignisse erfragt werden. Dabei unterscheiden sich die Instrumente auch dahingehend, ob sie für persönliche Interviews oder als Selbstausfüll-Fragebogen (häufig in Gruppenbefragungen) verwendet werden. Zur ersten Gruppe gehören die Instrumente der National Crime Victimization Surveys (NCVS), der Crime Surveys for England and Wales (CSEW), der International Crime Victimization Surveys/EU International Crime Surveys (ICVS/EU-ICV) oder des Deutschen Viktimisierungssurveys (Birkel u. a. 2014). Die Instrumente zur selbstberichteten Delinquenz Jugendlicher, wie sie in den International Self-Report Delinquency (ISRD) Studien und in den Studien des Kriminologischen Forschungsinstituts Niedersachsen (KFN) verwendet wurden, sind durchweg Selbstausfüll-Fragebögen für Gruppenbefragungen.

Die Anzeigequoten, die sich auf Daten des NCVS[5] stützen, sind inzidenzbasiert. Im regelmäßig durchgeführten NCVS, der Viktimisierungsstudie mit

[5] Die Fragebögen (*Basic Screen Questionnaire* und *Crime Incidence Report*) finden sich unter http://www.bjs.gov/index.cfm?ty=dcdetail&iid=245, dort ist ebenfalls das NCVS Interview-Handbuch verfügbar, in dem in Abschnitt B-4 das Vorgehen bei Nachfragen zu den einzelnen Viktimisierungsereignissen beschrieben wird.

dem höchsten Qualitätsstandard, wird das Anzeigeverhalten für jedes Viktimisierungsereignis des halbjährlichen Referenzzeitraums erfragt, wobei allerdings Serientaten[6] wie ein einzelnes Delikt behandelt und bei der Berechnung von Jahresraten nicht berücksichtigt werden. Da Angaben zum Zeitpunkt (Monat) der Ereignisse vorliegen, kann auch eine auf das letzte Delikt bezogene Anzeigequote bestimmt werden.

Im NCVS sind die Fragen zur Anzeige in einem Fragebogenmodul enthalten, dem ein Screening-Instrument vorangestellt ist, in dem unter anderem für alle Deliktskategorien die Häufigkeit der Viktimisierungen im halbjährlichen Referenzzeitraum erfragt wird. Anschließend werden für *jede* Viktimisierung vom ersten bis zum letzten Ereignis des Referenzzeitraums (bei Serienstraftaten nur zum letzten Ereignis) Fragen zur jeweiligen Tat und den Umständen gestellt. Hierzu gehören auch die Fragen zur Anzeige, nämlich (a) ob die Polizei informiert wurde oder davon erfahren hat, (b) wie und durch wen die Polizei davon erfahren hat, (c) falls die Polizei nicht informiert wurde: eine ausführliche Abfrage der Gründe der Nicht-Benachrichtigung – oder falls die Polizei informiert wurde: eine Abfrage der Gründe, die Polizei zu benachrichtigen, (d) ob, wie schnell und wie die Polizei auf die Benachrichtigung reagiert hat, (e) ob es nach dem Ereignis noch Kontakt zur Polizei gab, von wem der Kontakt ausging und die Form des Kontaktes, (f) was die Polizei im Anschluss an das Ereignis getan hat und (g) ob eine Strafanzeige erstattet wurde. Da weitere ausführliche Fragen zur jeweiligen Viktimisierung selbst gestellt werden, kann eine Analyse des Anzeigeverhaltens prinzipiell auch diese Informationen berücksichtigen.

Etwas weniger präzise ist die Erfassung des Anzeigeverhaltens im regelmäßig durchgeführten CSEW.[7] Nach einer Reihe von Vorab-Fragen (Screening), unter anderem zur Anzahl der Viktimisierungen in den letzten zwölf Monaten bezüglich unterschiedlicher Delikte, werden anschließend zu insgesamt maximal sechs Ereignissen des einjährigen Referenzzeitraums weitere Fragen zur Tat und den Tatumständen inklusive des Anzeigeverhaltens gestellt (Office for National Statistics 2015, 15). Auch hiermit können also (eingeschränkt durch die Obergrenze) inzidenzbasierte Anzeigequoten berechnet werden. Prinzipiell lassen sich für jede Deliktskategorie auch auf das letzte Delikt be-

---

6 Im NCVS werden multiple Viktimisierungen als Serientaten behandelt, wenn pro Deliktskategorie sechs oder mehr Ereignisse im halbjährlichen Referenzzeitraum stattgefunden haben, die Ereignisse einander ähnlich sind *und* die befragte Person nicht in der Lage ist, sich an genügend Details zu erinnern, um die Ereignisse voneinander unterscheiden zu können.

7 Die Fragebögen finden sich unter http://www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/, Informationen zur Stichprobenziehung und Befragung (wie der Handhabung des Fragebogens) sind unter http://www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/user-guides/ verfügbar.

zogene Anzeigequoten bestimmen – sofern nicht bei einer Überschreitung der Obergrenze von sechs Ereignissen für die Viktimisierung in einer bestimmten Deliktskategorie keine Abfrage mehr erfolgt.[8]

Im CSEW werden die Fragen zum Anzeigeverhalten für die maximal sechs Ereignisse (Seriendelikte werden wie ein Ereignis behandelt) in einer speziellen Reihenfolge gestellt: Zuerst werden die Ereignisse zu Straftaten gegen die Person erfragt, danach zu Eigentums- und Vermögensdelikten und zuletzt zu KFZ-Diebstahl, wobei innerhalb dieser Kategorien erst das letzte Ereignis angesprochen wird und anschließend die weiteren in rückwärtiger Reihenfolge. Neben Auskünften zur Tat und den Umständen erfassen die Fragen, (a) ob, durch wen und über welche Kommunikationsmittel die Polizei von dem Ereignis erfahren hat und (b) ob die befragte Person eine Referenznummer zu dem Vorfall erhalten hat. Des Weiteren werden (c) Fragen zur Interaktion mit der Polizei (ob die Motivation des Täters oder der Täterin thematisiert wurde) sowie (d) zur Zufriedenheit mit der Polizei gestellt, und es wird (e) gefragt, was die Polizei gegen den Täter oder die Täterin unternommen hat und (f) woher die befragte Person das weiß. Da wie im NCVS weitere Fragen zu der Viktimisierung selbst gestellt werden, kann eine Analyse des Anzeigeverhaltens prinzipiell auch diese Informationen berücksichtigen (allerdings werden die Informationen zu den letzten drei Ereignissen weniger ausführlich erhoben).

In der zweiten und dritten ISRD Studie[9] zur selbstberichteten Delinquenz Jugendlicher der 7. bis 9. Jahrgangsstufe wird auch die Häufigkeit von Viktimisierungserfahrungen im vergangenen Jahr und die Anzahl der der Polizei mitgeteilten Viktimisierungen erfasst. Im Gegensatz zum NCVS und CSEW wird das Anzeigeverhalten aber nicht für jedes Ereignis einzeln erfragt. Unmittelbar nach der Frage, wie oft den Befragten eines der Delikte in den letzten zwölf Monaten widerfahren ist, wird bei jedem Delikt mit einer offenen Häufigkeitsfrage erfasst, von wie vielen Ereignissen die Polizei erfahren hat. Mit den ISRD-Daten können inzidenzbasierte Anzeigequoten der jeweiligen Delikte bestimmt werden, nicht jedoch Anzeigequoten zum letzten Delikt.

In den unregelmäßig durchgeführten Studien des ICVS/EU-ICS[10] wird das Anzeigeverhalten nur beim jeweils letzten Delikt erfragt. Die Häufigkeit der

---

8 Dass dies geschieht, ist allerdings unwahrscheinlich. So wurden in der Befragung 2013/14 von 34.902 Befragten nur 40 (0,11 %) zu sechs Viktimisierungen befragt (wie viele Befragte mehr als sechs Viktimisierungen erlebt haben, ist allerdings unklar) (Office for National Statistics 2014b, 18).

9 Siehe http://www.northeastern.edu/isrd/. Der Fragebogen des ISRD2-Projekts findet sich unter http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34658.

10 http://wp.unil.ch/icvs/key-publications/key-publications/

Viktimisierungserfahrungen wird auf einer fünfstufigen Skala (von ein, zwei, drei, vier sowie fünf und mehr Ereignissen) eines einjährigen Referenzzeitraums (volles Kalenderjahr vor der Befragung) erfasst. Dem gegenüber beträgt der Referenzzeitraum des Anzeigeverhaltens beim jeweils letzten Delikt fünf Jahre. Sollen die auf Angaben zum letzten Delikt basierenden Anzeigequoten dazu benutzt werden, das Volumen der im Referenzzeitraum geschehenen Viktimisierungsereignisse, das der Polizei mitgeteilt wurde, zu erfassen, ist es bei Mehrfachviktimisierungen möglich, dass das angezeigte Delikt auch jünger sein kann als die bezogen auf das Referenzintervall erfragten Viktimisierungsereignisse. Das Problem dabei ist, dass die Erfassungsmethode es nicht erlaubt, festzustellen, *ob* dies der Fall ist.[11] Die Methode des ICVS schränkt somit die Möglichkeit ein, anhand der Anzeigequote das Volumen der Delikte zu bestimmen, die der Polizei bekannt geworden sind. Außerdem ist denkbar, dass Telescopingeffekte die auf dem letzten Delikt basierenden Anzeigequoten erhöhen, was möglicherweise durch den großen Referenzzeitraum noch verstärkt wird.

In den Schülerbefragungen des KFN[12] wurden ebenfalls Viktimisierungserfahrungen sowie das Anzeigeverhalten erhoben. In den Befragungen der Jahre 1998 bis 2005 wurden die Häufigkeit der Viktimisierungen bezogen auf einzelne (Gewalt)Delikte im aktuellen Befragungsjahr und im davor liegenden Kalenderjahr sowie die darauf bezogenen Häufigkeiten der Anzeigen erfasst, womit sich deliktspezifisch inzidenzbasierte Anzeigequoten bestimmen lassen. Gleichzeitig wurde auch zum insgesamt letzten Delikt erfragt, ob dies der Polizei mitgeteilt wurde, womit Anzeigequoten des letzten Delikts bestimmbar sind. Zwar ermöglichen die Fragen zum letzten Delikt auch deliktspezifische Anzeigequoten, die Stichprobenbasis ist aber dadurch reduziert, dass die Angaben nur für ein Delikt erfragt wurden. In den späteren Schülerbefragungen der Jahre 2007 und 2008 wurde das Anzeigeverhalten nur noch bezogen auf das letzte Delikt erhoben (Baier u. a. 2009, 98), so dass ein Vergleich mit inzidenzbasierten Anzeigequoten der Vorjahre nicht mehr möglich ist.

In der jüngsten bundesweiten Viktimisierungsstudie des Jahres 2012 (Deutscher Viktimisierungssurvey 2012; Birkel u. a. 2014) wurde das Anzeigeverhalten ähnlich wie im NCVS und CSEW für jedes der abgefragten Delikte bezogen auf die letzten zwölf Monate inzidenzbasiert erfasst. Dazu wurde vom ersten bis zu maximal fünf Ereignissen dieses Zeitraums jeweils gefragt, (a) ob die Polizei über den Vorfall informiert wurde, (b) wer die Polizei infor-

---

11 Vgl. den Fragebogen und die Instruktion an die Interviewerinnen und Interviewer (Van Dijk u. a. 2007, 203 f.).

12 http://www.kfn.de/Forschungsbereiche_und_Projekte/Schuelerbefragungen.htm.

miert hat, (c) in welcher Form die Polizei informiert wurde, und zusätzlich (d) ob eine Anzeige erstattet wurde – hier mit den Antwortmöglichkeiten (1) ja, (2) versucht aber „abgewimmelt“ und (3) nein. Auf diese Weise ist es möglich, sowohl inzidenzbasierte Anzeigequoten zu bestimmen (siehe Birkel u. a. 2014, 40) als auch zwischen der Mitteilung einer Straftat an die Polizei und einer Anzeige, die bei Vorliegen der Voraussetzungen[13] mit ziemlicher Sicherheit auch polizeilich registriert wurde, zu differenzieren.

2.3 Illustration: Inzidenzbasierte und auf dem letzten Delikt basierende Anzeigequoten

Anhand der Schülerbefragungen des KFN der Jahre 1998 und 2000, in denen einerseits die Inzidenzen von Viktimisierungen und die darauf bezogenen Mitteilungen an die Polizei erfragt wurden, und andererseits erhoben wurde, ob die Polizei vom jüngsten dieser Delikte erfahren hat, kann demonstriert werden, inwiefern sich die Anzeigequoten bei unterschiedlichen Methoden zur Messung der Anzeigequoten (inzidenzbasiert, Frage nach dem letzten Delikt) unterscheiden.

Da die Berechnung inzidenzbasierter Anzeigequoten aufwändig ist, wird gelegentlich auf eine prävalenzbasierte Anzeigequote ausgewichen, bei der die Angaben zur Häufigkeit der Anzeige dichotomisiert werden (keine Anzeige versus Anzeige) und unter den Opfern der Anteil derjenigen berechnet wird, die mindestens ein Delikt des jeweiligen Referenzzeitraums angezeigt haben. Um zu demonstrieren, welchen Einfluss dieses Verfahren auf die Schätzung der Anzeigequote hat, werden im Folgenden auch derartige prävalenzbasierte Maße berechnet.

Mit Inzidenzangaben ist das Problem verbunden, dass Befragte gelegentlich nur grobe und überhöhte Schätzungen angeben und dass zuweilen sogar extrem unplausible Angaben wie zum Beispiel mehr als 100 Viktimisierungen durch Raub pro Kalenderjahr auftreten. Das Problem tritt in erster Linie bei Angaben zur Häufigkeit von Viktimisierungen auf, kann aber auch bei Angaben zur Häufigkeit von Anzeigen vorkommen. Methoden der Extremwert- und Ausreißeranalysen ermöglichen eine Korrektur derartiger Werte (siehe Abschnitt 3.2). Da dabei in erster Linie sehr hohe Viktimisierungsangaben reduziert werden, fällt bei adjustierten Inzidenzangaben die Anzeigequote tendenziell höher aus. Um den Einfluss der hier vorgeschlagenen Korrektur-

[13] Die polizeilichen Registrierungsrichtlinien schreiben vor, dass eine Strafanzeige immer zu erfassen ist, wenn überprüfte Anhaltspunkte zum Tatbestand, zum Tatort und zur Tatzeit bzw. zum Tatzeitraum vorliegen (Bundeskriminalamt 2014, 6).

methode, bei der quasi unmögliche Werte eliminiert und Ausreißerwerte auf niedrigere Werte adjustiert werden, zu demonstrieren, werden neben (a) prävalenzbasierten, (b) auf dem letzten Delikt basierenden und (c) auf den Rohwerten der Inzidenzen basierenden Anzeigequoten auch (d) Anzeigequoten anhand von adjustierten Inzidenzangaben berechnet und miteinander verglichen.

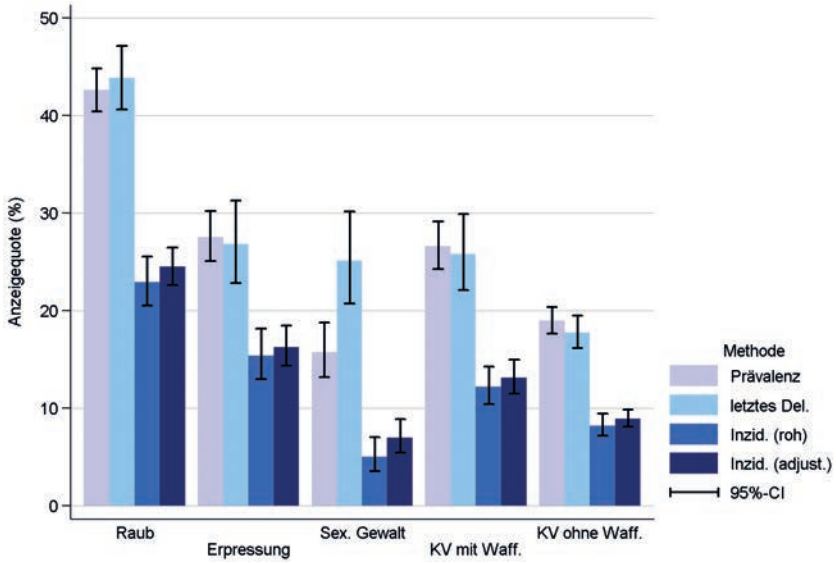
Die Analysen basieren auf einer Zusammenfassung der Datensätze von Schülerbefragungen des KFN, die in den Jahren 1998 und 2000 in Schulklassen der 9. Jahrgangsstufe mit überwiegend 14- bis 16-jährigen Jugendlichen in Hamburg, Hannover, Leipzig und München durchgeführt wurden. Die Inzidenzangaben der Befragung 1998 beziehen sich auf die Jahre 1996 bis zum Befragungszeitpunkt im Jahr 1998, die der 2000er Befragung auf das Jahr 1999 bis zum Befragungszeitpunkt im Jahr 2000. Für die Analysen der Anzeigequoten in diesem Abschnitt wurden zur Vergleichbarkeit der Maße nur diejenigen Fälle verwendet, die bei der Frage zum letzten Delikt angegeben haben, eines der fünf Gewaltdelikte erlebt zu haben, für die auch die Angaben zu den Inzidenzen erfasst wurden (Raub, Erpressung, sexuelle Gewalt, Körperverletzung (KV) mit Waffen und KV ohne Waffen). Die auf diese Weise verwendete Stichprobe umfasst 4.126 Viktimisierungsopfer mit gültigen Angaben sowohl zu den Inzidenzen als auch zur Anzeige beim letzten Delikt.

Ein Vergleich der in *Abbildung 1* dargestellten Anzeigequoten für die einzelnen Delikte zeigt, dass die inzidenzbasierten Anzeigequoten insgesamt deutlich niedriger sind als die auf Angaben zum letzten Delikt basierenden. Während über alle fünf Deliktsarten hinweg die inzidenzbasierte Anzeigequote zwischen 12,3 % [11,2 %–13,4 %][14] (Rohwerte) und 13,5 % [12,7 %–14,4 %] (adjustierte Werte) liegt, ist die Anzeigequote beim letzten Delikt mit 25,8 % [24,5 %–27,1 %] ungefähr doppelt so hoch. Eine derartige Diskrepanz findet sich bei allen Delikten mit Ausnahme der sexuellen Gewalt, bei der die auf Angaben zum letzten Delikt basierende Anzeigequote etwa viermal höher ausfällt als die inzidenzbasierten Anzeigequoten (25,2 % [20,7 %–30,2 %] versus 5,0 % [3,6 %–7,0 %] (Rohwerte) und 7,1 % [5,5 %–8,9 %] (adjustierte Werte)).

[14] 95 %-CI (Konfidenzintervalle) sind in eckigen Klammern angegeben; zur Interpretation siehe Fußnote 30.

Abbildung 1:

Anzeigequoten in Abhängigkeit von der Messmethode



Dass die auf Angaben zum letzten Delikt basierenden Anzeigequoten deutlich höher ausfallen, ist höchstwahrscheinlich Telescopingeffekten zuzuschreiben, die dazu führen, dass bei der Nennung des Delikts dasjenige ausgewählt wird (und subjektiv als jünger erscheint), das spontan besser erinnert werden kann. Das scheinen vor allem solche Delikte zu sein, bei denen die Polizei informiert wurde und die subjektiv als bedeutsamer oder als schwerwiegender erlebt wurden. Aber auch das Bedürfnis von Befragten, sozial erwünscht zu antworten und sich deshalb bei der Antwort eher auf ein angezeigtes Delikt zu beziehen, kann bei Fragen nach dem letzten Delikt zu einer erhöhten Anzeigequote führen.[15] Dabei müssen die Befragten das Delikt nicht deshalb ausgewählt haben, weil es angezeigt wurde, sondern weil sie glauben, dass der Interviewer oder die Interviewerin sich besonders für schwerwiegende Delikte (die generell eine höhere Anzeigewahrscheinlichkeit haben) interessiert.

---

15 Auch wenn gelegentlich angenommen wird (z. B. Schwind u. a. 2001, 116) und es plausibel ist, dass Befragte zumindest in der Befragungssituation eine Anzeigeerstattung für sozial erwünscht halten, gibt es dafür kaum empirisch belastbare Belege. Eine experimentelle Studie konnte dies jedenfalls nicht bestätigen (Greenberg/Ruback 1992, 88f.).

Dass bei der Frage nach dem letzten Delikt Ereignisse ausgewählt wurden, die sehr wahrscheinlich weiter zurückliegen als das tatsächlich zuletzt erlebte Delikt, lässt sich anhand der Daten nachweisen: Zunächst war für verschiedene Delikte jeweils erfragt worden, wie viele Viktimisierungen den Befragten im Befragungsjahr sowie im ersten und im zweiten Jahr davor zugestoßen waren. In der anschließenden Frage zum letzten Delikt haben dann 15,5 % der Befragten bei der zeitlichen Einordnung ein Datum angegeben, das vor dem Zeitraum liegt, in dem sie bei den vorangehenden Fragen zur Häufigkeit von Viktimisierungen schon mindestens ein gleichartiges Delikt genannt haben.[16] Da dieser Prozentsatz aber noch nicht einen derart großen Unterschied der inzidenzbasierten und der auf Angaben zum letzten Delikt basierenden Anzeigequoten erklärt, ist zu vermuten, dass Telescoping auch *innerhalb* des Jahres stattgefunden hat, das bei der zeitlichen Einordnung des letzten Delikts genannt wurde. Eine Vorverlagerung innerhalb des Referenzzeitraums ist vor allem bei Delikten mit einer hohen Inzidenzrate möglich, weil dabei die Wahrscheinlichkeit erhöht ist, dass sich innerhalb des Zeitraums *mehrere* Delikte ereignet haben, von denen nicht alle angezeigt wurden. Dementsprechend ist eine Verlagerung in dem Sinne, dass als letztes Delikt ein Ereignis ausgewählt wurde, das weiter zurückliegt als anhand der vorangegangenen Fragen zur Häufigkeit der Viktimisierungen in einzelnen Jahren zu erwarten gewesen wäre, bei den deutlich selteneren sexuellen Gewaltdelikten mit 20,8 % am höchsten.

Unter denen, die als letztes Delikt ein Ereignis nennen, das vor dem Zeitraum liegt, in dem sie bei den vorangehenden Fragen zur Häufigkeit von Viktimisierungen schon mindestens ein gleichartiges Delikt genannt haben (im Folgenden als „Verlagerer“ bezeichnet), sollte die auf den Angaben zum letzten Delikt basierende Anzeigequote höher sein als bei denen, die ein Ereignis ausgewählt haben, dass in dem letztmöglichen Zeitraum liegt. Ein Problem bei der Prüfung ist jedoch, dass sich unter diesen „Nicht-Verlagerern“ auch unerkannte Verlagerer befinden, nämlich solche, die (wie oben beschrieben) innerhalb des letztmöglichen Zeitraums verlagert haben. Wenn die Telescoping-These der selektiven Auswahl zutrifft, sollte eine höhere Anzeigequote jedenfalls dann feststellbar sein, wenn eine Verlagerung nicht auch innerhalb des letztmöglichen Zeitraums möglich ist. Da diese Möglichkeit bei seltenen Delikten weniger gegeben ist, sollte der Unterschied der auf Angaben zum

16 Aus Forscherperspektive stellt dies eine Vorverlagerung dar, da die Befragten ein Delikt, das weiter zurückliegt als anhand vorangegangener Fragen zur Häufigkeit der Viktimisierungen in einzelnen Jahren zu erwarten gewesen wäre, als das jüngste angeben. Wäre die Zusatzfrage, wann dieses „letzte Delikt“ geschehen ist, nicht gestellt worden, wäre diese Verlagerung nicht offensichtlich gewesen.

letzten Delikt basierenden Anzeigequoten zwischen Verlagerern und Nicht-Verlagerern bei seltenen Delikten am größten sein.

Ein Vergleich der entsprechenden Anzeigequoten zwischen Nicht-Verlagerern und Verlagerern je nach Gewaltdelikt zeigt, dass bei dem seltenen Delikt der sexuellen Gewalt die Anzeigequote der Verlagerer tatsächlich mit 35,8 % signifikant höher ist als die der Nicht-Verlagerer mit 22,4 % ($\chi^2_{(1)} = 5{,}11$, $p = ,024$), während sich insgesamt die auf dem letzten Delikt basierenden Anzeigequoten zwischen den beiden Gruppen nicht signifikant unterscheiden (26,4 % bei den Verlagerern vs. 25,7 % bei den Nicht-Verlagerern, $\chi^2_{(1)} = 0{,}14$, $p = ,705$). Vermutlich enthält die Gruppe der Nicht-Verlagerer ebenfalls einen substanziellen Anteil tatsächlich vorverlagernder Personen, bei denen eine Vorverlagerung aber nicht erkannt wurde, weil dies auch innerhalb des Referenzzeitraums geschehen kann. Ein Nachweis ist anhand der vorliegenden Daten nicht möglich.

Eine Betrachtung der prävalenzbasierten Anzeigequoten (*Abbildung 1*) zeigt, dass diese ähnlich hoch sind wie die auf den Angaben zum letzten Delikt beruhenden Anzeigequoten. Die Annahme, dass die auf dem letzten Delikt basierenden Anzeigequoten gegenüber den inzidenzbasierten deshalb erhöht sind, weil bei der Nennung des letzten Delikts vorzugsweise Delikte genannt werden, die auch angezeigt wurden, hilft, auch diesen Befund zu verstehen. Während bei den Angaben zum letzten Delikt die Befragten selbst die in Richtung Anzeige verzerrende Auswahl der Delikte treffen, ist dies bei der Berechnung der prävalenzbasierten Anzeigequote notwendig der Fall: Wenn mindestens eines der im Referenzzeitraum erlebten Delikte angezeigt wurde, wird das Opfer als anzeigend klassifiziert. Nur bei sehr seltenen Delikten, bei denen Mehrfachviktimisierungen im gleichen Referenzzeitraum unwahrscheinlich sind, sollten prävalenzbasierte Anzeigequoten niedriger sein und sich inzidenzbasierten Anzeigequoten annähern. Auch dies ist anhand der Ergebnisse in *Abbildung 1* zu beobachten (siehe die Anzeigequoten zur sexuellen Gewalt).

Schließlich zeigen die Ergebnisse in *Abbildung 1*, dass eine Korrektur der Inzidenzangaben zur Verringerung des Einflusses extrem hoher Häufigkeiten (dies wird ausführlich in Abschnitt 3.2 behandelt) zu einer etwas höheren inzidenzbasierten Anzeigequote führt, wobei die Unterschiede insgesamt aber nicht sehr ausgeprägt sind. Über alle Delikte hinweg erhöht eine Adjustierung der Inzidenzen die Anzeigequote um knapp 10 %, bei dem Delikt der sexuellen Gewalt allerdings um etwa 40 % von 5,0 % auf 7,1 % (siehe oben). Da ein Seriendelikt mit sehr hohen Inzidenzen bei sexueller Gewalt durchaus denkbar ist, dieses bei einer Anzeige aber wie ein Delikt behandelt würde, könnte die Erhöhung der Anzeigequote durch die hier erfolgte Adjustierung tatsächlich angemessen sein.

Die Ergebnisse in *Abbildung 1* demonstrieren eindrücklich, dass inzidenzbasierte und auf Angaben zum letzten Delikt basierte Anzeigequoten zu drastisch unterschiedlichen Ergebnissen führen. Deshalb ist die Frage wesentlich, welche Methode besser geeignet ist, anhand der Anzeigequote das Volumen der der Polizei mitgeteilten Ereignisse zu schätzen. Es spricht einiges dafür, dass die Angaben zum letzten Delikt aufgrund von Telescoping-Effekten überhöht sind. Insbesondere die Ähnlichkeit dieser Anzeigequoten mit prävalenzbasierten Anzeigequoten, die bei Mehrfachviktimisierungen und einer eher mäßig hohen Anzeigehäufigkeit deutlich erhöht sein müssen, demonstriert, dass auch auf Angaben zum letzten Delikt basierende Anzeigequoten sehr wahrscheinlich drastisch überhöht sind und damit zu einer Überschätzung der Menge der der Polizei mitgeteilten Ereignisse führen.

3 Methodische und praktische Empfehlungen

3.1 Erfassung inzidenzbasierter Anzeigequoten

Eine zentrale, aus diesen Ausführungen folgende Empfehlung ist, inzidenzbasierte Anzeigequoten zu benutzen, die (pro Delikt) auf Angaben zur Anzahl der Viktimisierungen sowie der Anzahl der angezeigten Delikte im jeweiligen Referenzzeitraum beruhen. Liegen die dafür nötigen Daten vor und sollte sich zeigen, dass die Inzidenzangaben unzuverlässig sind (einer der wichtigsten Einwände gegen inzidenzbasierte Anzeigequoten), können die Daten immer noch in Prävalenzangaben transformiert und zur Berechnung prävalenzbasierter Anzeigequoten benutzt werden. Man sollte sich dabei aber bewusst sein, dass eine derartige Reduktion der Information die Ungenauigkeit der Inzidenzschätzung nur übertüncht – geeigneter scheinen stattdessen Verfahren zur Eliminierung unplausibler Werte und zur Adjustierung oder Winsorisierung von Ausreißerwerten zu sein (siehe Abschnitte 3.2 und 3.3).

Wird die Anzahl der angezeigten Delikte nicht pauschal über eine Häufigkeitsfrage ermittelt (wie in den Befragungen der ISRD- oder KFN-Studien), sondern indem pro Viktimisierungsereignis erfragt wird, ob es angezeigt wurde oder nicht, werden auf diese Weise gleichzeitig auch die Inzidenzangaben zur Viktimisierung überprüft. Das reduziert die Wahrscheinlichkeit unzuverlässiger Angaben – insbesondere dann, wenn (wie im NCVS oder CSEW) auch weitere Angaben zur Tat und den Tatumständen erfragt werden. Wird dabei auch erfasst, wann (z. B. in welchem Monat) sich das jeweilige Delikt ereignet hat (wobei zu empfehlen ist, dies – wie im NCVS – für *alle* Vorfälle zu erheben, die keine Seriendelikte darstellen; das erlaubt es auch, dabei die

chronologische Reihenfolge der Ereignisse beizubehalten),[17] können anhand dieser Angaben auch auf dem letzten Delikt basierende Anzeigequoten berechnet werden. Das ermöglicht den Vergleich mit den Ergebnissen von anderen Studien, in denen ausschließlich Fragen zum letzten Delikt benutzt wurden.

Fragen zum Anzeigeverhalten bei jeder einzelnen Viktimisierung sind allerdings für Selbstausfüll-Fragebögen im Papier-und-Bleistift-Modus ungeeignet, da die dazu nötige Filterführung die meisten Befragten überfordern dürfte.[18] Sie können also nur in persönlichen Interviews oder in computergestützten Fragebögen eingesetzt werden, bei denen die Filterführung kontrolliert bzw. automatisiert werden kann.

Eine wichtige Frage ist, ob auch erfragt werden sollte, wie und auf welchem Weg die Polizei über die Straftat informiert worden ist (mündlich, schriftlich; online, telefonisch, persönlich) und wie die Anzeige aufgenommen wurde (entgegengenommen oder „abgewimmelt“), gegebenfalls sollte gefragt werden, ob bei der Anzeigeerstattung ein Schriftstück unterzeichnet wurde und ob ein Strafantrag gestellt wurde (um eine Strafanzeige von einem Strafantrag abzugrenzen). Eigentlich müsste in Deutschland die Aussage, dass die Polizei von einer Straftat erfahren hat, für die Bestimmung der Anzeigequote ausreichen, da die Polizei entsprechend dem Legalitätsprinzip verpflichtet ist, jede Mitteilung einer Straftat, bei der Hinweise auf Tatbestand, Tatort und Tatzeit vorliegen, zu registrieren (Bundeskriminalamt 2014, 6).[19] Genauere Fragen hierzu scheinen aber insbesondere in Deutschland nötig zu sein, da selbst bei Experten Unklarheiten über ein vermeintliches Formerfordernis bei einer Anzeigeerstattung zu bestehen scheinen (vgl. Fußnote 2 und Schwind u. a. 2001, 115). Bei der Auswertung der Daten ist es dann möglich, Diskrepanzen zwischen der Rate angezeigter und polizeilich registrierter Delikte genauer nach-

---

17 Sollen die Nachfragen zu den einzelnen Vorfällen auf ein Maximum beschränkt werden, sollte eine Maximalzahl pro Deliktskategorie gelten, damit sichergestellt ist, dass für jede Kategorie Angaben zum Anzeigeverhalten vorliegen. Überschreitet die Anzahl der Vorfälle das Maximum, sollte darauf geachtet werden, dass auch zum letzten Vorfall des Referenzzeitraums Informationen vorliegen.

18 Wenn bei Selbstausfüll-Fragebögen die Anzeige der Vorfälle über eine offene Häufigkeitsfrage erfasst wird, muss darauf geachtet werden, dass Nichtanzeigen von fehlenden Antworten unterschieden werden können – nicht alle Befragte sind bereit, in ein Antwortfeld eine Null einzutragen. Analoges gilt auch für die Frage der Viktimisierungsinzidenz, bei der sichergestellt werden muss, dass Null Vorfälle im Referenzzeitraum von einer fehlenden Angabe unterscheidbar sind.

19 Anders ist dies beispielsweise in den Niederlanden, wo die Polizei über größeren Entscheidungsspielraum verfügt, weshalb dort in den Befragungen zum Sicherheitsmonitor zusätzlich zur Frage der Mitteilung des Vorfalls an die Polizei erfasst wird, ob dabei ein Dokument unterzeichnet wurde (der jeweils aktuelle Fragebogen ist unter http://www.veiligheidsmonitor.nl/Werkwijze/Vragenlijst verfügbar).

zugehen, indem die bloße Mitteilung einer Straftat von einer förmlichen, mit Unterschrift versehenen Anzeige (die mit einer größeren Sicherheit auch registriert wurde) unterschieden wird.

Bei der Kodierung und Berechnung inzidenzbasierter Anzeigequoten müssen einige Besonderheiten beachtet werden: Da inzidenzbasierte Anzeigequoten auf Inzidenzangaben für Viktimisierungen und auf Inzidenzangaben für Anzeigen beruhen, muss sichergestellt werden, dass (a) die Häufigkeiten der Vorfälle mit Null (und nicht als „fehlend“) kodiert werden, wenn sich in dem jeweiligen Referenzzeitraum keine Viktimisierung ereignet hat, und (b) die Häufigkeiten der Anzeigen in den Fällen, in denen sich keine Viktimisierung im Referenzzeitraum ereignet hat, als „fehlend“ (nicht mit Null) kodiert werden, allerdings mit Null (und nicht als „fehlend“) für den Fall, dass sich im Referenzzeitraum mindestens ein Vorfall ereignet hat. Zur Berechnung der Anzeigequoten müssen (c) für alle Opfer pro Delikt(skategorie) jeweils die Summe der Anzeigen und die Summe der Viktimisierungen berechnet werden, so dass anschließend die Anzeigequote als das Verhältnis der Gesamtsumme der Anzeigen aller Opfer zur Gesamtsumme der Viktimisierungen aller Opfer berechnet werden kann.[20] Schließlich ist bei Signifikanztests inzidenzbasierter Anzeigequoten zu berücksichtigen, dass es sich hierbei innerhalb der Personen um abhängige Messungen handelt, für die spezielle statistische Verfahren wie Bootstrapping oder die Analyse sogenannter geclusterter Daten erforderlich sind (siehe Abschnitt 3.4).

3.2 Identifikation und Adjustierung von Ausreißern bei Inzidenzangaben

Zur Berechnung von inzidenzbasierten Anzeigequoten werden Häufigkeitsangaben sowohl zur Viktimisierung als auch zur Anzeige benutzt. Ein Problem derartiger Angaben ist, dass sie insbesondere bei häufigen Ereignissen fehlerbehaftet sein können – sei es dadurch, dass es für die Befragten schwierig ist, die Anzahl häufiger Ereignisse präzise zu schätzen, oder dass aufgrund mangelnder Motivation nur grobe oder sogar phantastisch hohe Angaben gemacht werden, was insbesondere bei der Befragung Jugendlicher vorkommen kann. Wenn die Häufigkeitsangaben wie bei den Schülerbefragungen des KFN oder in den ISRD Studien nicht – wie im NCVS oder CSEW – durch Nachfragen zu jedem Ereignis überprüft werden (was hilft, problematische

20 In SPSS müssen die Gesamtsummen über den Umweg von Aggregatdaten berechnet werden.

Angaben schon in der Befragungssituation zu identifizieren),[21] können die Daten nur mit Hilfe von Plausibilitätsregeln und statistischen Verfahren überprüft und gegebenenfalls korrigiert werden.

Da inzidenzbasierte Anzeigequoten vor allem durch die Angabe unplausibel häufiger Viktimisierungen verzerrt werden können, liegt es nahe, zu hohe Werte zu identifizieren und diese entweder zu eliminieren oder zu adjustieren. In einer Reanalyse von Inzidenzangaben im NCVS diskutieren Planty und Strom (2007) die Vor- und Nachteile verschiedener Varianten des Umgangs mit Seriendelikten[22] (in Berichten zu jährlichen Raten des NCVS werden Seriendelikte nicht berücksichtigt, da ihre zeitliche Einordnung unklar ist). Sie unterscheiden dabei fünf Möglichkeiten, die Anzahl von Ereignissen bei Seriendelikten zu berücksichtigen: Seriendelikte können (a) als ein Ereignis, (b) als sechs Ereignisse (was der Kappungsgrenze entspricht), (c) als der Median der Serienereignisse (etwa 10), (d) als Mittelwert der Serienereignisse (etwa 25) oder (e) entsprechend der berichteten Anzahl (6 bis zu 750 Viktimisierungen) gezählt werden. Seriendelikte nicht oder nur als ein Ereignis zu berücksichtigen, ist offensichtlich ähnlich unbefriedigend, wie die berichtete Inzidenz selbst zu benutzen: „Sich auf die berichteten Viktimisierungsinzidenzen zu stützen, ist wegen unbestimmbarer, durch Mehrfachopfer verursachte Verzerrungen ebenfalls problematisch.“ (Planty/Strom 2007, 197, Übers. d. Verf.) Auch die Kappung bei einem willkürlich festgelegten Maximum betrachten die Autoren als inadäquat, da dies nicht berücksichtigt, dass die Anzahl der Viktimisierungen von Delikts- und Opfermerkmalen abhängt. Offensichtlich sind für eine adäquate Behandlung von hohen Inzidenzangaben plausible Annahmen über die Verteilungsform der Viktimisierungen nötig, um Extremwerte und Ausreißer identifizieren zu können.

Ein häufig verwendetes, empirisch gestütztes Verfahren besteht darin, Werte, die mehr als zweieinhalb oder drei Standardabweichungen vom Mittelwert abweichen, als Ausreißer zu klassifizieren (Aguinis u. a. 2013, 283). Da dieses Kriterium aber unterstellt, dass die Daten normalverteilt sind, ist es insbesondere bei seltenen Ereignissen wie Viktimisierungen ungeeignet, bei denen häufig mehr als 80 % der Fälle den Wert Null (= „keine Viktimisierung“) aufweisen. Die Annahme, dass derartige Ereignisse stattdessen einer negati-

21 Siehe auch Abschnitt 3.1.

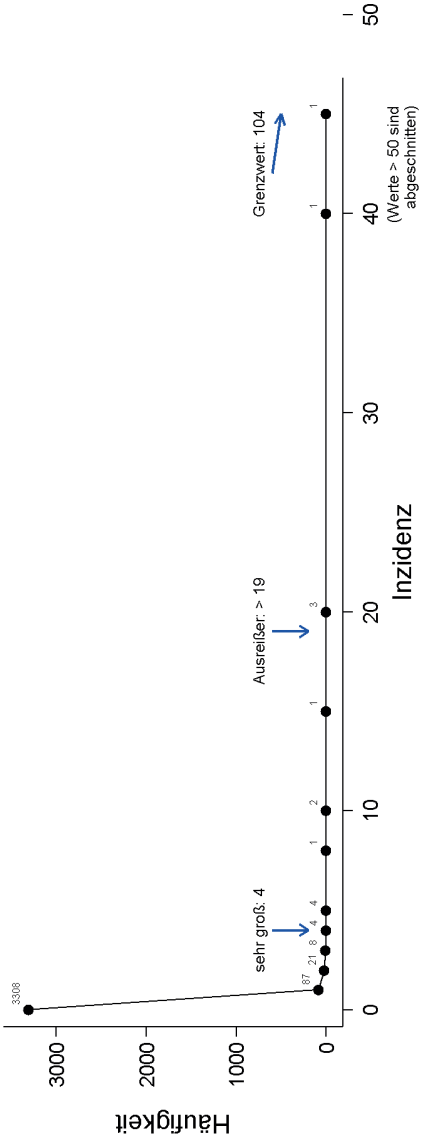
22 Zur Definition im NCVS vgl. Fußnote 6.

ven Binomialverteilung folgen, ist häufig angemessener[23] und wird deshalb dem hier vorgeschlagenen Verfahren zugrunde gelegt. Unter dieser Annahme lässt sich ein wesentlich plausibleres Kriterium definieren, ab dem Werte der Stichprobe als Ausreißer zu bezeichnen sind. Dazu können die Stichprobeninzidenzen der Ereignisse mit den theoretisch zu erwartenden Inzidenzen verglichen werden, die sich unter der Annahme einer spezifischen negativen Binomialverteilung mit dem Mittelwert und Überdispersionsparameter der Stichprobe ergeben. Als Ausreißer lassen sich dann Fälle definieren, deren Inzidenz in einer deutlich größeren Anzahl vorliegt, als aufgrund der theoretischen Verteilung der Inzidenzen zu erwarten wäre. Konkret werden im Folgenden alle Fälle als Ausreißer bezeichnet, deren Inzidenz so groß ist, dass sie theoretisch seltener als ein halbes Mal in der Stichprobe vorkommen sollten.

[23] Zählvariablen werden häufig als Poisson-verteilt angenommen, wobei der Mittelwert gleich der Varianz der Werte ist. Insbesondere bei voneinander abhängigen Ereignissen wie Mehrfachviktimisierungen ist die Annahme einer negativen Binomialverteilung aber wesentlich plausibler, bei der die Varianz der Werte größer als der Mittelwert ist. Diese Verteilungen sind deshalb zusätzlich zum Mittelwert noch durch einen sogenannten Überdispersions-Parameter gekennzeichnet (für eine ausführliche Darstellung siehe Hilbe 2011).

Abbildung 2:

Häufigkeitsverteilung von Inzidenzen mit normativem Grenzwert für Extreme und empirisch bestimmtem Ausreißerschwellenwert (Beispiel)



Wie groß der Schwellenwert für diese „Ausreißerinzidenz“ ist, hängt von den Kennwerten (Mittelwert und Überdispersionsparameter) der negativen Binomialverteilung ab, die anhand der Stichprobe gewonnen werden – der konkrete Schwellenwert (vgl. Wert 19 in *Abbildung 2*) ist also ein empirisch gewonnener Wert, der je nach Delikt und Stichprobe anders ausfallen kann. Da die zur Identifikation von Ausreißern benutzten Kennwerte für die negative Binomialverteilung allerdings selbst von Extremwerten der Stichprobe abhängen, sollten bei ihrer Bestimmung diejenigen Fälle nicht berücksichtigt werden, deren Inzidenz derart extrem ist, dass sie als quasi unmöglich bezeichnet werden können (wie z. B. die Angabe, im vergangenen Jahr mehr als zweimal pro Woche Opfer eines Raubdelikts geworden zu sein oder häufiger als wöchentlich ein bestimmtes Delikt bei der Polizei angezeigt zu haben). Der Grenzwert (z. B. 104) für nicht zu berücksichtigende Extremwerte kann nicht empirisch, sondern muss normativ festgelegt werden. Der empirische Schwellenwert[24] für die Ausreißerinzindenz sollte also erst nach Eliminierung der Extremwerte berechnet werden.

Es gibt verschiedene Möglichkeiten, mit den auf diese Weise identifizierten Ausreißern umzugehen. Angesichts dessen, dass abgesehen von quasi unmöglichen Extremwerten Ausreißer wirklich existieren, auch wenn die Höhe der jeweiligen Inzidenz fraglich ist, sollten Ausreißer nicht einfach aus der Stichprobe entfernt werden. Eine bessere Variante ist, die Ausreißer in der Stichprobe zu belassen, alle Ausreißerwerte aber auf den Ausreißerschwellenwert festzulegen. Die Rückstufung von Ausreißerwerten auf den Schwellenwert hat allerdings den Nachteil, dass damit die Verteilungsform der Inzidenzen am Verteilungsende einen zweiten Gipfel zu bekommen droht. Des Weiteren kann dann zwischen Personen, deren ursprüngliche Werte über diesem Schwellenwert lagen, nicht mehr differenziert werden.

Eine noch bessere Variante scheint es deshalb zu sein, für die als Ausreißer identifizierten Werte zufällige Ersatzwerte aus der theoretischen negativen Binomialverteilung zu ziehen, mit der die Ausreißer identifiziert wurden, dies aber unter der Restriktion, dass der Zufallswert mindestens so groß sein muss wie eine „sehr große“, aber noch realistische Inzidenz. Um diese festzulegen, ist ebenfalls eine normative Entscheidung nötig, die deliktspezifisch festlegt, welche Inzidenz als „sehr groß“ gelten soll. Bei schweren Gewaltdelikten könnten zum Beispiel Inzidenzen von vier als „sehr groß“ bezeichnet werden (vgl. *Abbildung 2*). So beträgt im CSEW die maximal registrierte einjährige

---

24 Der Schwellenwert wird anhand der theoretischen Dichtekurve einer negativen Binomialverteilung mit den empirischen Kennwerten der jeweiligen Inzidenzvariablen (Mittelwert und Überdispersonsparameter) und der aktuellen Stichprobengröße berechnet (siehe das nachfolgende Beispiel).

Inzidenz pro Delikt(skategorie) fünf. Wenn dann die unter dieser Restriktion erzeugten Zufallswerte, die mindestens so groß sind wie die „sehr große“ Inzidenz, sortiert werden und die Ausreißerwerte der Größe nach ersetzen, bleibt trotz der Adjustierung der Inzidenzen die Rangreihe der Personen weitgehend erhalten.[25]

3.3 Beispiel zur Identifikation und Adjustierung von Ausreißerinzidenzen

Das folgende Beispiel basiert auf einer Stichprobe von 3.443 Befragten, in der es 3.308 Nichtopfer und 135 (3,9 %) Opfer von Raubdelikten gibt (Jahresprävalenzrate).[26] Die Häufigkeitsverteilung der Inzidenzen ist in *Abbildung 2* dargestellt (2 Fälle mit den Inzidenzangaben 99 und 100 liegen außerhalb des sichtbaren Bereichs). Diese Werte liegen noch unterhalb des Grenzwerts für extrem unplausible Werte, der normativ auf 104 festgelegt wurde, was einer Viktimisierungsfrequenz von zweimal wöchentlich entspricht. Deshalb werden in diesem Beispiel *alle* Fälle für die Bestimmung des Ausreißerschwellenwerts benutzt. Die mittlere Inzidenzrate (durchschnittliche Anzahl der Viktimisierungen pro Person) beträgt 0,167, die Standardabweichung ist mit 2,718 sehr groß, und der Überdispersionsparameter der negativen Binomialverteilung dieser Daten beträgt 60,95.

Wird eine erwartete Häufigkeit von weniger als 0,5 als Kriterium für den Ausreißerschwellenwert der Inzidenzen angenommen, sind bei einer negativen Binomialverteilung mit diesem Überdispersionsparameter und Mittelwert bei dieser Stichprobengröße alle Werte größer 19 als Ausreißer zu bezeichnen. Insgesamt sind mit sieben der 3.443 Fälle (0,2 %) nur sehr wenige größer als dieser empirische Schwellenwert. *Tabelle 1* zeigt die Häufigkeitsverteilung der Inzidenzen für diese sieben Ausreißer sowie die Zufallswerte, mit denen sie ersetzt wurden. Die Zufallswerte wurden aus einer negativen Binomialverteilung mit einem Mittelwert von 0,167 und einem Überdispersionsparameter von 60,95 unter der Bedingung gezogen, dass der Zufallswert mindestens „sehr groß“ ist (in diesem Beispiel mindestens 4, vgl. *Abbildung 2*).

---

25 Ein Stata-Programm, mit dem sich die hier beschriebene Identifizierung und Adjustierung von Inzidenzwerten automatisieren lässt, kann in Stata mit “ssc install nb_adjust“ installiert werden, siehe http://econpapers.repec.org/software/bocbocode/s458051.htm.

26 Die Daten stammen aus Schülerbefragungen des KFN des Jahres 1998.

Tabelle 1:

Häufigkeiten ursprünglicher und adjustierter Inzidenzen der Ausreißer

	Inzidenz (ursprünglich)					
Inzidenz (adjustiert)	20	40	45	99	100	Summe
4	1					1
5	2					2
6		1				1
8			1			1
9				1		1
13					1	1
Summe	3	1	1	1	1	7

Tabelle 2 demonstriert, dass eine derartige Adjustierung der Inzidenzen von nur 0,2 % der Fälle die Kennwerte der Stichprobe deutlich verändert: Während die mittlere Inzidenzrate in etwa halbiert wurde, beträgt die Standardabweichung weniger als ein Viertel. *Tabelle 2* enthält auch die Kennwerte bei einer Reduktion aller Ausreißerwerte auf den Ausreißerschwellenwert (19; manchmal auch als „Zensierung“ oder „Winsorisierung“ bezeichnet) sowie nach dem Entfernen der Ausreißerwerte aus der Stichprobe. Hinsichtlich des Mittelwerts und der Streuung stellt die Adjustierung einen Kompromiss zwischen Reduktion und Eliminierung dar. Inhaltlich hätte die Eliminierung den wesentlichen Nachteil, dass die Stichprobe um Fälle „bereinigt“ würde, deren Werte möglicherweise tatsächlich existieren (wenn sie auch vermutlich nicht derartig hoch sind), was unter Umständen die Analyse von besonderen Risikopopulationen oder entsprechenden Teilgruppen ungültig werden ließe oder unzulässig einschränken würde.

Tabelle 2:

Kennwerte der Inzidenzvariablen je nach Ausreißerbehandlung

Ausreißer	*Mittelwert*	*SD*	*Minimum*	*Maximum*	*n*
original	0,167	2,718	0	100	3.443
reduziert	0,106	0,992	0	19	3.443
adjustiert	0,082	0,612	0	15	3.443
eliminiert	0,068	0,506	0	15	3.436

3.4 Die Notwendigkeit von Konfidenzintervallen

Unabhängig davon, ob inzidenzbasierte, auf Angaben zum letzten Delikt basierende oder prävalenzbasierte Anzeigequoten benutzt werden, wird in Analysen und bei der Interpretation von Anzeigequoten häufig ignoriert, dass es sich hierbei um *Stichprobendaten* handelt, die notwendig mit einem mehr oder weniger großen Stichprobenfehler behaftet sind. So werden weder im CSEW[27] noch in den Publikationen des KFN Konfidenzintervalle angegeben, anhand derer sich die Größe des Stichprobenfehlers ablesen ließe.[28]

Die Notwendigkeit, Konfidenzintervalle von Anzeigequoten bei der Interpretation zu berücksichtigen, wird deutlich, wenn man z. B. die in *Abbildung 3* dargestellten Daten der Schülerbefragungen des KFN betrachtet, die Anzeigequoten von Gewaltdelikten der Jahre 1997 und 1999 für vier Großstädte zeigen. In einer früheren Publikation haben wir darauf hingewiesen, dass insbesondere der Anstieg in Hannover besonders deutlich ist (Wilmers u. a. 2002, 38).[29] Während der Vergleich der Prozentangaben der Grafik links oben in *Abbildung 3* durchaus beeindruckende Zuwächse der Anzeigequoten um 7,4 % (Hamburg) bis zu 28,4 % (München) nahezulegen scheint, zeigen die Konfidenzintervalle[30] der übrigen drei Grafiken, dass tatsächlich in keiner der Städte die Veränderung zwischen 1997 und 1999 statistisch signifikant ist.

Die Grafiken in *Abbildung 3* demonstrieren darüber hinaus den Einfluss, den der Umgang mit Ausreißern bei inzidenzbasierten Anzeigequoten hat: Eine Eliminierung oder Adjustierung von Ausreißern führt im Allgemeinen zu höheren Anzeigequoten mit schmaleren Konfidenzintervallen.[31] Letzteres ist auf

---

27 Vgl. zum Beispiel Office for National Statistics (2014a, 9 ff. und die dazu gehörenden Tabellen).

28 Auch in früheren Publikationen zum NCVS fehlen Konfidenzintervalle, z. B. Hart und Rennison (2003); seit 2009 werden jedoch in der „*Criminal Victimization Series*“ regelmäßig Standardfehler angegeben, anhand derer sich Konfidenzintervalle berechnen lassen, siehe http://www.bjs.gov/index.cfm?ty=pbse&sid=6.

29 Da die Daten in *Abbildung 3* nicht gewichtet sind, weichen die Anzeigequoten von den in Wilmers u. a. (2002, 39) dargestellten Quoten geringfügig ab.

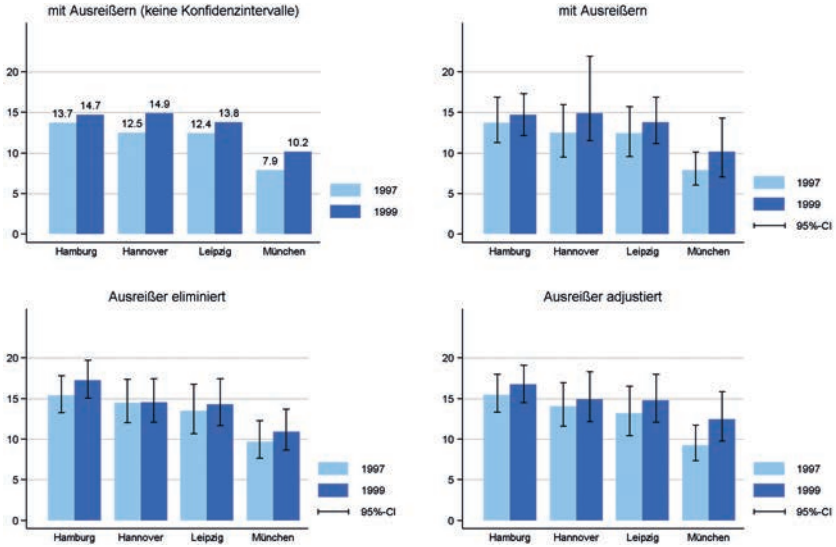
30 Werden Daten voneinander unabhängiger Stichproben miteinander verglichen, lässt sich am Grad der Überlappung der 95 %-Konfidenzintervalle unmittelbar erkennen, ob sich die Mittelwerte der beiden Stichproben signifikant unterscheiden: Als Faustregel gilt, dass ab einer Stichprobengröße von jeweils 10 der Unterschied mit mindestens $p < .05$ statistisch signifikant ist, wenn die Überlappung nicht mehr als 50 % der durchschnittlichen Armlänge (Distanz zwischen Mittelwert und Grenze des Konfidenzintervalls um diesen Mittelwert) beträgt (Cumming/Finch 2005). Dabei wird allerdings multiplen Vergleichen, die die Irrtumswahrscheinlichkeit erhöhen, nicht Rechnung getragen.

31 Man beachte, dass die Identifikation und Adjustierung oder Eliminierung von Ausreißern sowohl für die Inzidenzangaben der Viktimisierung als auch für die Inzidenzangaben der Anzeige durchgeführt wurden. Da dabei vor allem außergewöhnlich hohe Viktimisierungsinzidenzen korrigiert werden, erhöht sich dadurch die geschätzte Anzeigequote, die als das prozentuale Verhältnis der Anzeigeinzidenz zur Viktimisierungsinzidenz berechnet wird, s. o.

die deutlich geringere Varianz der Inzidenzen bei einer Reduktion, Adjustierung oder Eliminierung von Ausreißerwerten zurückzuführen. Besonders gut sichtbar wird dies für die Anzeigequote des Jahres 1999 in Hannover.

Abbildung 3:

Inzidenzbasierte Anzeigequoten von Gewaltdelikten (KFN-Schülerbefragungen)



Die Berechnung von Konfidenzintervallen für Anzeigequoten, die auf Angaben zum letzten Delikt basieren oder prävalenzbasiert sind, stellt statistisch gesehen keine größere Schwierigkeit dar und ist als Standardmethode in den meisten Statistikpaketen implementiert. Die statistische Theorie für die Berechnung *inzidenzbasierter* Anzeigequoten ist jedoch wesentlich komplexer, da die Basis hier nicht die Anzahl der Befragten, sondern die Anzahl der Viktimisierungsereignisse ist und darüber hinaus die Ereignisse bei Mehrfachviktimisierungen statistisch nicht voneinander unabhängig sind. Für derartige Probleme ist die Verwendung von sogenannten Bootstrap-Verfahren eine Alternative zum klassischen Signifikanztest (Hesterberg u. a. 2009). Dabei wird die Verteilung der Populationsparameter, die für den Signifikanztest und die Berechnung von Konfidenzintervallen benötigt wird, durch wiederholte Stichprobenziehungen aus den vorliegenden Daten empirisch bestimmt.[32]

[32] Bootstrap-Verfahren lassen sich sowohl mit SPSS unter Verwendung des Zusatzmoduls „Bootstrapping“ als auch mit Stata (seit Version 12.0) berechnen.

Alternativ zum Bootstrap-Verfahren ist es möglich, die Opfer als Cluster von Viktimisierungsereignissen zu betrachten[33] und die Daten dementsprechend so zu transformieren, dass die Anzahl der Fälle der Anzahl der Viktimisierungsereignisse entspricht. Dabei sind dann nicht mehr Personen, sondern Viktimisierungsereignisse die kleinste Einheit. Für Signifikanztests und zur Konstruktion von Konfidenzintervallen werden dann statistische Verfahren benutzt, die es erlauben, die Standardfehler von sogenannten geclusterten Daten korrekt zu berechnen.[34] Während das Bootstrap-Verfahren rechenaufwändig ist, erfordert das alternative Verfahren die Umstrukturierung des Datensatzes für jede zu analysierende Deliktskategorie. Der Vorteil ist, dass auf diese Weise problemlos mit gewichteten Daten gearbeitet werden kann und multiple logistische Regressions- und Mehrebenenmodelle zur Vorhersage der Anzeigewahrscheinlichkeit analysiert werden können.

Tabelle 3 zeigt die Zunahme der auf adjustierten Inzidenzen basierenden Anzeigequoten der vier Städte und der Gesamtstichprobe zusammen mit ihren Bootstrap-Konfidenzintervallen.[35] Alle Konfidenzintervalle schließen den Wert 0 ein, d. h. in keiner der Städte ist die Zunahme statistisch signifikant. Zwar findet sich in den Grafiken für *alle* Städte ein gemeinsamer Trend, jedoch ist auch dieser Gesamttrend statistisch nicht signifikant. Obwohl sich also in den Stichprobendaten zwischen 1997 und 1999 insgesamt ein Zuwachs der Anzeigequote von 13,4 % auf 15,0 % bzw. um 11,4 % findet, muss die Hypothese, in der Population habe die Anzeigebereitschaft zugenommen, nach gängiger statistischer Konvention zurückgewiesen werden – die Zunahme der in der PKS in diesem Zeitraum registrierten Häufigkeitsziffer für Gewaltdelikte trotz sinkender Viktimisierungsraten in den Befragungsdaten kann also entgegen dem ersten Anschein nicht durch eine Zunahme der Anzeigequote erklärt werden (zumindest reicht für einen statistisch signifikanten Nachweis dieser Größenordnung die Stichprobengröße nicht aus).

---

33 Dank an Christoph Birkel (BKA) für diesen Hinweis.

34 In SPSS ist hierfür das Zusatzmodul „Complex Samples“ notwendig, in Stata können hierzu die svy-Befehle verwendet werden.

35 Verwendet wurden sogenannte Bias-korrigierte und akzelerierte Konfidenzintervalle, die bessere asymptotische Eigenschaften haben als traditionelle Bootstrap-Konfidenzintervalle. Für weitere Details siehe Poi (2004).

Tabelle 3:

Differenz der Anzeigequoten 1999 – 1997 (Prozent)

	Differenz	*Bootstrap SE*	*95 %-CI von*	*bis*
Hamburg	1,30	1,68	–2,01	4,55
Hannover	0,84	2,08	–3,12	5,04
Leipzig	1,56	2,21	–2,79	5,84
München	3,19	1,87	–0,29	6,97
total	1,53	0,97	–0,38	3,43

Anmerkung: Adjustierte Inzidenzen; Bias-korrigierte und akzelerierte Bootstrap Konfidenzintervalle (10.000 Replikationen); $n = 2.348$ (1997) und 2.545 (1999)

Die Konfidenzintervalle der auf adjustierten Inzidenzen basierenden Anzeigequoten in *Abbildung 3* zeigen des Weiteren, dass auch die regionalen Unterschiede zwischen den Städten Hamburg, Hannover und Leipzig nicht signifikant sind, während der optische Vergleich der Konfidenzintervalle von Hamburg und München sowohl für 1997 und möglicherweise auch für 1999 eine signifikant niedrigere Anzeigequote Münchens zeigt, die dem bekannten Nord-Süd-Gefälle der Kriminalitätsraten im Hellfeld entspricht. Der statistische Test anhand von Bootstrap-Konfidenzintervallen bestätigt dies: Im Jahr 1997 beträgt die Differenz zwischen Hamburg und München 6,21 [9,34–3,05] Prozentpunkte (Bootstrap SE = 1,61), und auch die Differenz von 4,32 [7,79–0,23] Prozentpunkten im Jahr 1999 ist noch statistisch signifikant (Boostrap SE = 1,91).

4 Zusammenfassung

Insgesamt zeigt eine Auseinandersetzung mit den Mess- und Konstruktionsproblemen von Anzeigequoten, dass die Entscheidung über die Art und Weise der Erfassung und die jeweils benutzten Analysemethoden die erzielten Befunde und ihre Interpretation massiv beeinflussen.

– Nur mit wiederholt mit gleicher Methode durchgeführten Viktimisierungsstudien mit ausreichender Stichprobengröße ist es möglich, einigermaßen verlässliche Angaben zu Höhe und Veränderung von Anzeigequoten zu erhalten.

– Inzidenzbasierte Anzeigequoten erfassen das Volumen der kriminellen Ereignisse, von denen die Polizei erfahren hat, während prävalenzbasierte Anzeigequoten Auskunft über den Prozentsatz der Opfer geben, die Anzeige erstattet haben.

- Anzeigequoten, die auf Angaben zum letzten Delikt beruhen, führen zu ähnlich hohen Anzeigequoten wie prävalenzbasierte. Es ist zu erwarten, dass auf dem letzten Delikten basierende Anzeigequoten vermehrt die Anzeige schwererer Delikte erfassen und deshalb (sowie aufgrund von Telescoping-Effekten oder sozial erwünschtem Antwortverhalten) systematisch überhöht sind.

- Inzidenzbasierte Anzeigequoten fallen im Allgemeinen nur halb so groß aus wie prävalenzbasierte oder auf Angaben zum letzten Delikt basierende Anzeigequoten. Dies muss bei dem Vergleich von Anzeigequoten über Studien hinweg in Rechnung gestellt werden, um nicht methodisch bedingte Diskrepanzen fälschlicherweise als regionale oder zeitlich bedingte Unterschiede zu interpretieren.

- Sowohl aus inhaltlichen als auch aus methodischen Gründen sind inzidenzbasierte Anzeigequoten prävalenzbasierten oder auf Angaben zum letzten Delikt beruhenden vorzuziehen.

- Obwohl schon die Frage, ob eine Straftat der Polizei mitgeteilt wurde, für die Feststellung einer Strafanzeige ausreichen müsste, ist zu empfehlen, zusätzlich zu erfassen, ob dabei ein Dokument unterzeichnet wurde, da offenbar häufig Unklarheiten über die vermeintliche Formerfordernis einer Anzeige bestehen.

- Bei der Analyse inzidenzbasierter Anzeigequoten ist zu beachten, dass für Signifikanztests und die Konstruktion von Konfidenzintervallen spezielle statistische Verfahren wie Bootstrapping oder die Analyse geclusterter Daten erforderlich sind.

- Ganz allgemein ist zu fordern, dass Angaben zu Anzeigequoten mit Konfidenzintervallen versehen werden, um den Stichprobenfehler abschätzen zu können – dies gilt ebenso für prävalenzbasierte sowie für solche Anzeigequoten, die auf der Angabe zum letzten Delikt beruhen.

5 Literatur

Aguinis, Herman; Gottfredson, Ryan K. und Joo, Harry (2013): Best-practice recommendations for defining, identifying, and handling outliers. In: Organizational Research Methods, 16, S. 270–301.

Averdijk, Margit; Elffers, Henk (2012): The discrepancy between survey-based victim accounts and police records revisited. In: International Journal of Victimology, 18, S. 91–107.

Baier, Dirk; Pfeiffer, Christian; Simonson, Julia und Rabold, Susann (2009): Jugendliche in Deutschland als Opfer und Täter von Gewalt: Erster Forschungsbericht zum gemeinsamen Forschungsprojekt des Bundesministeriums des Innern und des KFN (= KFN Forschungsbericht Nr. 107). Hannover: Kriminologisches Forschungsinstitut Niedersachsen e. V. URL: http://www.kfn.de/versions/kfn/assets/fb107.pdf – Download vom 04. 06. 2015.

Birkel, Christoph; Guzy, Nathalie; Hummelsheim, Dina; Oberwittler, Dietrich und Pritsch, Julian (2014): Der Deutsche Viktimisierungssurvey 2012: Erste Ergebnisse zu Opfererfahrungen, Einstellungen gegenüber der Polizei und Kriminalitätsfurcht (= Schriftenreihe des Max-Planck-Instituts für ausländisches und internationales Strafrecht, Band A 7 10/2014). Freiburg i. Br.: Max-Planck-Institut für ausländisches und internationales Strafrecht. URL: https://www.mpicc.de/files/pdf3/a7_2014_Viktimisierungssurvey_2012.pdf – Download vom 04. 06. 2015.

Brown, Norman R.; Rips, Lance J. und Shevell, Steven K. (1985): The subjective dates of natural events in very-long-term memory. In: Cognitive Psychology, 17, S. 139–177.

Bundeskriminalamt (2014): Richtlinien für die Führung der Polizeilichen Kriminalstatistik i. d. F. vom 01. 01. 2014. Wiesbaden: BKA.

Coleman, Clive; Moynihan, Jenny (1996): Understanding Crime Data: Haunted by the Dark Figure. Buckingham: Open University Press.

Cumming, Geoff; Finch, Sue (2005): Inference by eye: Confidence intervals and how to read pictures of data. In: American Psychologist, 60, S. 170–180.

Enzmann, Dirk (2015): Anzeigeverhalten und polizeiliche Registrierung. In: Guzy, Nathalie; Birkel, Christoph und Mischkowitz, Robert (Hg.): Viktimisierungsbefragungen in Deutschland. Band 1: Ziele, Nutzen und Forschungsstand. Wiesbaden: Bundeskriminalamt, S. 509–540.

Farrell, Graham; Pease, Ken (2007): The sting in the tail of the British Crime Survey: Multiple victimizations. In: Hough, Mike; Maxfield, Mike (Hg.): Surveying Crime in the 21st Century (= Crime Prevention Studies, Vol. 22). Cullomptom: Willan Publ., S. 33–53.

Feltes, Thomas (2009): Aussagewert der polizeilichen Aufklärungsquote. In: Kriminalistik, 63, S. 36–41.

Greenberg, Martin S.; Ruback, R. Barry (1992): After the Crime: Victim Decision Making. New York: Plenum Press.

Guzy, Nathalie; Leitgöb, Heinz (2015): Assessing mode effects in online and telephone victimization surveys. In: International Review of Victimology, 21, S. 101–131.

Hart, Timothy C.; Rennison, Callie (2003): Reporting crime to the police, 1992-2000 (NCJ 195710). Washington, D.C.: U.S. Department of Justice, Bureau of Justice Statistics.

Hesterberg, Tim; Moore, David S.; Monaghan, Shaun; Clipson, Ashley und Epstein, Rachel (2009): Bootstrap methods and permutation tests. In: Moore, David S.; McCabe, George P. und Craig, Bruce (Hg.): Introduction to the Practice of Statistics. New York, NY: Freeman, S. 16–60.

Hilbe, Joseph M. (2011): Negative Binomial Regression. 2. Aufl. Cambridge: Cambridge University Press.

Office for National Statistics (2014a): Chapter 1: Overview of Violent Crime and Sexual Offenses 2012/13. Newport: ONS. URL: http://www.ons.gov.uk/ons/dcp171776_352364.pdf – Download vom 04. 06. 2015.

Office for National Statistics (2014b): The 2013/14 Crime Survey for England and Wales. Volume One: Technical Report. Newport: ONS. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/2013-14-crime-survey-for-england-and-wales-technical-report- - -volume-1.pdf – Download vom 04. 06. 2015.

Office for National Statistics (2015): User Guide to Crime Statistics for England and Wales. Newport: ONS. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/user-guides/user-guide-to-crime-statistics.pdf – Download vom 04. 06. 2015.

Planty, Michael; Strom, Kevin J. (2007): Unterstanding the role of repeat victims in the production of the annual US victimization rates. In: Journal of Quantitative Criminology, 23, S. 179–200.

Poi, Brian P. (2004): From the help desk: Some bootstrapping techniques. In: The Stata Journal, 4, S. 312–328.

Schwind, Hans-Dieter; Fechtenhauer, Detlef; Ahlborn, Wilfried und Weiß, Rüdiger (2001): Kriminalitätsprobleme im Langzeitvergleich am Beispiel einer deutschen Großstadt: Bochum 1975 – 1986 – 1998. Neuwied: Luchterhand.

Sellin, Thorsten (1931): The basis of a crime index. In: Journal of the American Institute of Criminal Law and Criminology, 22, S. 335–356.

Skogan, Wesley G. (1975): Measurement problems in official and survey crime rates. In: Journal of Criminal Justice, 3, S. 17–31.

Sudman, Seymour; Bradburn, Norman M. (1973): Effects of time and memory factors on response in surveys. In: Journal of the American Statistical Association, 68, S. 805–815.

van Dijk, Jan; van Kesteren, John und Smit, Paul (2007): Criminal Victimization in International Perspective: Key Findings from the 2004-2005 ICVS and EU ICS. Den Haag: WODC. URL: http://wp.unil.ch/icvs/files/2012/11/ICVS2004_051.pdf - Download vom 04. 06. 2015.

Wilmers, Nicola; Enzmann, Dirk; Schaefer, Dagmar; Herbers, Karin; Greve, Werner und Wetzels, Peter (2002): Jugendliche in Deutschland zur Jahrtausendwende: Gefährlich oder gefährdet? Ergebnisse wiederholter, repräsentativer Dunkelfelduntersuchungen zu Gewalt und Kriminalität im Leben junger Menschen 1998–2000. Baden-Baden: Nomos.

Fragebogenkonstruktion

Frank Faulbaum

1 Einleitende Bemerkungen

Umfragen stellen gegenwärtig noch die einzige Möglichkeit dar, große Haushalts- und Personenstichproben zu untersuchen. Sie unterscheiden sich von anderen Methoden der Datenerhebung vor allem dadurch, dass Daten unter Einsatz *systematischer* Methoden der Befragung mit dem Ziel der quantitativen Beschreibung einer Zielpopulation (auch: Grundgesamtheit) von Elementen hinsichtlich bestimmter Merkmale erhoben werden. Um verallgemeinerbare Erkenntnisse über Opfer unterschiedlicher Arten von Vergehen – von Opfern sexueller Belästigungen bis zu Opfern schwerer Verbrechen – zu gewinnen, erscheint es notwendig, größere Stichproben von Opfern zu untersuchen. Hier stellen Umfragen mit ihrem strukturierten Vorgehen bei allen von der qualitativen Sozialforschung vorgetragenen Nachteilen, wie z. B. die Reduktion der Kommunikation auf Frage-Antwort-Dyaden ohne vertiefendes Gespräch, die Erhebungsmethode der Wahl dar. Ihre Werkzeuge sind standardisierte Interviews, die unter Einsatz von Fragebögen entweder durch Interviewerinnen und Interviewer administriert werden (Interviewer-administrierte Fragebögen) oder ohne Interviewerinnen und Interviewer von den Befragten selbst administriert werden (selbst administrierte Fragebögen). Zahlreiche Befunde deuten darauf hin, dass sensitive Inhalte, wie sie im Fall von Opfererfahrungen vorliegen, eher im Rahmen selbst administrierter Interviews erhoben werden sollten, da in diesem Fall mit einer geringeren Zahl von Verweigerungen bei sensitiven Fragen zu rechnen ist (z. B. Jobe u. a. 1997; Tourangeau/Yan 2007).

Für das Erzielen von validen Ergebnissen, d. h. von Ergebnissen, die auf Antworten der Befragten beruhen, die sich auf die inhaltlich relevanten Zielvariablen des bzw. der Forschenden beziehen, und nicht etwa auf Merkmale wie soziale Erwünschtheit oder Selbstenthüllungstendenz, erscheint es geboten, bei der Formulierung von Fragen und Fragebögen auch die Merkmale des Befragten wie etwa die Affinität zur sozialen Bezugsgruppe oder die Selbstenthüllungstendenz in den Blick zu nehmen. Insgesamt sind Fehler und Ungenauigkeiten beim Entwurf von Fragen und ihre Folgen für die Interpretierbarkeit der Ergebnisse nach Abschluss der Erhebung nicht mehr zu korrigieren.

2 Bausteine von Fragebögen: Survey-Items

Der Fragebogen ist die zentrale Grundlage für die Steuerung der Interviewer-Befragten-Interaktion (bei Interviewer-administrierten Interviews) oder der bzw. des Befragen (bei selbst administrierten Interviews). Sein Entwurf muss einerseits die elementaren Bausteine des Instruments, andererseits aber auch die Sequenz dieser Bausteine und ihre Wirkungen auf die Befragten in den Blick nehmen. Wenn wir von einem Fragebogen sprechen, beziehen wir uns auf ihn als Steuerungsinstrument in standardisierten Interviews. Standardisierung bedeutet, dass für alle Befragten gleiche Befragungsbedingungen gelten:

– gleiche Einleitungstexte,

– gleiche Fragen und gleiche Antwortvorgaben,

– gleiche Reihenfolge der Fragen,

– gleiche Befragungshilfen (z. B. Listen, Kärtchen etc. bei Face-to-Face-Interviews),

– Anweisungen an den Administrator, die Bestandteile des Erhebungsinstruments in der vorgegebenen Form zu handhaben.

Die elementaren Bausteine eines Fragebogens bestehen in sprachlichen Ausdrücken, die neben weiteren Informationen u. a. eine Aufforderung an die Befragten enthalten, eine bestimmte Aufgabe zu erfüllen. Diese Aufgabe besteht für die bzw. den Befragten in der Regel darin, bestimmte Selbstauskünfte zu geben. Nicht immer wird dabei von der bzw. dem Befragten eine direkte Antwort auf eine Frage gefordert. Vielmehr können den Befragten durchaus auch komplexere Aufgaben wie Paarvergleiche oder Entscheidungsaufgaben, wie z. B. beim zur Vermeidung von Antwortverweigerungen bei sensitiven Fragen entwickelten RRT-Verfahren (Randomized-Response-Technik; Chauduri/Mukherjee 1988), gestellt werden. Im Fall der RRT wird der Befragte zur Wahl zwischen zwei alternativen Items aufgefordert, einem sensitiven und einem nicht sensitiven Item, mithilfe eines Zufallsmechanismus, z. B. eines Würfels oder einer Münze, ein Item auswählen und zu beantworten. Schließlich kann eine Frage auch in einer Bitte um Zustimmung zur Durchführung physischer Messungen (z. B. Bestimmung der Körpermasse) bestehen. In diesem Fall ist die Antwort das Ergebnis einer Messung, die von der bzw. dem Befragten nicht selbst durchgeführt wird.

Die elementaren Bausteine des Fragebogens beinhalten neben der eigentlichen Frage mit der Aufforderung, eine bestimmte Leistung zu erbringen, in der Regel mehrere zusätzliche Informationen. Als umfassende Bezeichnung

für diese Bausteine wird oft der Begriff ‚Survey-Item' verwendet (Saris/Gallhofer 2010).

Zusatzinformationen in einem Survey-Item können sein:

– *Szenarios oder Situationsbeschreibungen*, auf die sich die Antworten der bzw. des Befragten beziehen sollen, z. B. standardisierte Situations- oder Personenbeschreibungen, die auch als *Vignetten* bezeichnet und in *faktoriellen Surveys* verwendet werden, wie Beschreibungen von Gefährdungssituationen, Opfersituationen etc., bei denen reale oder vorgestellte alternative Verhaltensweisen, Einschätzungen etc. erhoben werden;

– *Orientierende Ausdrücke* wie „Wenn Sie einmal an Situation x denken". Sie dienen zur kognitiven und emotionalen Orientierung und Einstimmung der bzw. des Befragten und sollen einen bestimmten kognitiven und/oder emotionalen Zustand in der bzw. dem Befragten erzeugen;

– *Ausdrücke, die über eventuelle Hilfsmittel zur Beantwortung (Befragungshilfen) informieren* wie „Hier habe ich einige Kärtchen, auf denen Sätze stehen. Bitte …" oder „Ich lege Ihnen jetzt eine Liste vor, auf der verschiedene Berufe stehen. Bitte sagen Sie mir …". Davon wird vor allem bei Face-to-Face-Interviews ohne Computerunterstützung Gebrauch gemacht;

– *Definitionen, Erläuterungen und Klärungen durch die bzw. den Interviewer* bei Interviews, in denen die Bedeutung im Fragetext verwendeter sprachlicher Ausdrücke präzisiert wird; Beispiele sind Deliktbeschreibungen und -verdeutlichungen;

– Im Fall des *Dependent Interviewing* (DI; Jäckle 2009; Lynn u. a. 2006): Aufforderungen an die Befragten, sich an Antworten, die sie an früherer Stelle des Interviews oder früher in einem anderen Interview gegeben haben, zu erinnern, um die Validität der Antworten z. B. durch Vermeidung von Fehlklassifikationen, insbesondere in Panelerhebungen (z. B. Erinnerungen an Opferwerdungen, die am Anfang des Fragebogens erhoben wurden) zu erhöhen.

In der Aufgabenbeschreibung einer Frage werden oft Leistungen der bzw. des Befragten verlangt, die sich auf die Bewertung oder die Klassifikation von Aussagen beziehen. Wir wollen für diese Aussagen den Begriff ‚Item' verwenden, der vom Begriff ‚Survey-Item' zu trennen ist. Unter einem Item wird in diesem Sprachgebrauch ein sprachlicher Ausdruck verstanden, der *als Teil der in der Frage formulierten Aufgabe* auf einer Antwortdimension bewertet werden soll. Oft haben diese Ausdrücke den Charakter von Aussagen, die

Sachverhalte, Situationen etc. beschreiben. In diesem Sinne wird der Begriff traditionell in der psychometrischen Literatur verwendet (Guilford 1954).

In *Abbildung 1* ist ein Fall dargestellt, bei dem drei Aussagen, die verschiedene Ereignisse beschreiben, von den Befragten dahingehend beurteilt werden müssen, ob sie eingetreten sind oder nicht (Bewertungsdimension: „passiert – nicht passiert". Die Frage fordert von der bzw. dem Befragten aber nur eine Ja/Nein-Antwort in Abhängigkeit davon, ob mindestens eines der beschriebenen Ereignisse eingetreten ist.

In den Sozialwissenschaften findet sich auch das Verständnis eines Items als kleinster Einheit eines Fragebogens.

Abbildung 1:

Beispiel einer Itembatterie aus dem Deutschen Viktimisierungssurvey 2012

Es kommt auch vor, dass man ohne Waffen oder Gegenstände tätlich angegriffen wird, mit dem Ziel, jemanden absichtlich körperlichen Schaden oder Schmerzen zuzufügen.
Denken Sie bitte wieder an die letzten fünf Jahre, also die Zeit seit Anfang 2007: Bitte sagen Sie, ob Ihnen in dieser Zeit mindestens einmal einer der folgenden Vorfälle passiert ist.

- Jemand hat Sie seit Anfang 2007 mindestens einmal absichtlich geschlagen, getreten oder gewürgt, um Ihnen körperlichen Schaden oder Schmerzen zuzufügen.
- Jemand hat Ihnen seit Anfang 2007 mindestens einmal absichtlich Verbrennungen zugefügt.
- Jemand hat Sie seit Anfang 2007 mindestens einmal auf andere Weise absichtlich tätlich angegriffen, um Ihnen körperlichen Schaden oder Schmerzen zuzufügen.

Items

Items zum gleichen Thema und gleichen Bewertungsdimensionen können in Itemlisten bzw. Itembatterien zusammengefasst werden. Die inhaltlichen Dimensionen solcher Itembatterien lassen sich mithilfe von Techniken der explorativen Faktorenanalyse identifizieren. Die Items einer Itemliste können aber auch im Verlauf der Operationalisierung theoretischer (auch: latenter) Konstrukte wie „Kriminalitätsneigung" als beobachtete (auch: manifeste, empirische) Indikatoren für ein Konstrukt eingeführt werden. Ein statistisches Modell für die Beziehungen zwischen einem latenten Konstrukt und seinen manifesten Indikatoren wird auch als Messmodell bezeichnet. Die Einflussstärken des Konstrukts auf die Indikatoren und die Gesamtanpassung des Messmodells lassen sich im Rahmen der konfirmatorischen Faktorenanalyse bestimmen (Reinecke 2014).

Nachdem Fragen und Items entworfen sind, kann die Reihenfolge festgelegt werden. Vereinfacht ausgedrückt gilt:

Fragebogen = Fragen + Navigationsanweisungen.

Die Reihenfolge sollte nach Themen strukturiert und unter Berücksichtigung möglicher Kontexteffekte (Schuman/Presser 1981), d. h. vorangehender Fragen/Items auf nachfolgende vorgenommen werden.

3 Einteilungsprinzipien von Fragen

Je nach Antwortformat lassen sich Fragen grob einteilen in:

– Geschlossene Fragen (*closed-ended questions*): Alle Antwortmöglichkeiten sind durch Antwortvorgaben abgedeckt. *Voraussetzung*: Universum der Antwortalternativen ist bekannt.

– Offene Fragen (*open-ended questions*): Fragen ohne Antwortvorgaben.

– Hybridfragen (*auch halboffene Fragen*): feste Antwortvorgaben mit der Möglichkeit, zusätzliche, in den Antwortkategorien nicht vorgesehene Antworten zu geben (Beispiel: Sonstiges, und zwar . . .).

– *Voraussetzung*: Universum der Antwortalternativen ist nicht vollständig bekannt.

Andere Einteilungsprinzipien von Fragen basieren auf dem Inhalt der Frage bzw. auf der Art der in der Frage gewünschten Information. Eine populäre Einteilung unterscheidet zwischen den folgenden Fragetypen, von denen prinzipiell alle auch in Viktimisierungsbefragungen zum Einsatz kommen können:

– Faktfragen (*factual questions*): Fragen nach gegenwärtigen oder vergangenen Fakten, wobei sich diese auf Ereignisse oder das Verhalten des Befragten beziehen können (Tourangeau u. a. 2000), z. B.: „Haben Sie im letzten Monat persönlich einen Arzt aufgesucht? (Ja/Nein)“; „In welchem Jahr sind Sie geboren?“. Beziehen sich Faktfragen auf das Verhalten der bzw. des Befragten, wird gelegentlich von Verhaltensfragen gesprochen. In Viktimisierungsbefragungen können sich Faktfragen z. B. auf Viktimisierungserlebnisse und das Verhalten in Opfersituationen beziehen.

– Wissensfragen (*knowledge questions*) beziehen sich auf Kenntnisse der bzw. des Befragten, z. B. die Bekanntheit einer Produktmarke, einer Insti-

tution, einer gesetzlichen Regelung oder einer Person, etwa einer Politikerin oder eines Politikers.

– Einstellungs- und Meinungsfragen sind Beurteilungen bzw. Bewertungen bestimmter Aussagen (Items) auf verschiedenen Antwortdimensionen, z. B. eine Bewertung auf der Dimension „Wichtigkeit“: „Für wie wichtig halten Sie die folgenden Merkmale für Ihren Beruf und ihre berufliche Arbeit?“ (sehr wichtig/eher wichtig/eher unwichtig/sehr unwichtig); eine Qualitätsbewertung: „Was meinen Sie: Wie gut arbeitet die Polizei bei der Verbrechensbekämpfung“ (sehr gut, eher gut, eher schlecht, sehr schlecht).

– Fragen nach Überzeugungen, Einschätzungen gegenwärtiger, vergangener oder vermuteter zukünftiger Ereignisse und Zustände, z. B. „Was glauben Sie: Wird Person X eher wegen Mordes oder wegen fahrlässiger Tötung verurteilt?“ (Ja/Nein).

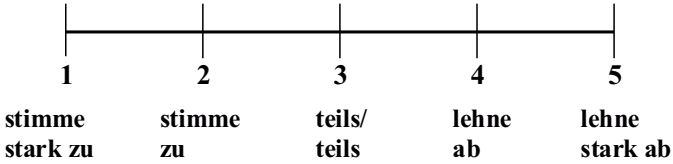
4 Antwortformate und Skalen

Mit einer Frage sind in standardisierten Interviews stets auch bestimmte Antwortvorgaben verbunden, in welche die Befragten ihre Antworten einpassen müssen. Im Grenzfall kann eine Frage in einem freien Format auch offen gestellt werden. Grundsätzlich erfordern bestimmte Fragen konventionelle, d. h. nach den in der Gesellschaft gelernten Konversationsregeln bestimmte Arten von Antworten. Einige Fragen, insbesondere solche nach Opfererlebnissen, können z. B. angemessen nur mit „Ja“ oder „Nein“ beantwortet werden. Es ist wichtig, dass bei der Konstruktion von Fragetexten und Antwortkategorien die Sprachkonventionen im Hinblick auf das Verhältnis von Frage und Antwort beachtet werden. Bei geschlossenen Fragen werden den Befragten verschiedene Antwortalternativen (auch: Antwortkategorien, Antwortvorgaben) präsentiert. Beziehen sich die Antwortalternativen auf eine bestimmte gemeinsame Dimension (Antwortdimension) oder eine Eigenschaft wie z. B. „Zufriedenheit“, so stellen sie Abstufungen auf einer Antwortskala (*response scale*) dar, denen Zahlen zugeordnet werden können. Vom Begriff ‚Anwortskala‘ ist der messtheoretische Begriff ‚Skala‘ zu unterscheiden (z. B. Orth 1974; Suppes/Zinnes 1963). Ob die Antworten auf einer Antwortskala eine Skala im messtheoretischen Sinn bilden, kann nur auf Basis messtheoretischer Annahmen entschieden werden.

Antwortskalen, auf denen *Urteile* abgestuft werden können, heißen auch Ratingskalen (*rating scales*). Ein bekanntes Beispiel für eine Ratingskala ist jener Typ einer fünfstufigen Antwortskala, der von Likert (1932) in seiner *Methode der summierten Ratings* verwendet wurde (*Abbildung 2*).

Abbildung 2:

Antwortskala vom Likert-Typ



In den meisten Fällen handelt es sich bei den Abstufungen auf Antwortskalen um Abstufungen in Form diskreter Kategorien. In diesem Fall spricht man auch von Kategorialskalen (*category scales*). Kategorialen Einstufungen können kontinuierliche Bewertungen der bzw. des Befragten auf einer Dimension zugrunde liegen, die die bzw. der Befragte in vorgegebene kategoriale Formate übertragen bzw. einfügen muss. In diesem Fall übersetzt die bzw. der Befragte seine subjektive Bewertung auf einer latenten Antwortskala in beobachtbare diskrete Kategorien. Im Grenzfall kann eine Antwortskala auch dichotom sein wie etwa eine Ja/Nein-Skala. Der statistische Zusammenhang zwischen kontinuierlichen latenten Antwortskalen und beobachteten diskreten Antwortkategorien ist Gegenstand der sog. Item-Response-Theorie (IRT; Hambelton u. a. 1991; Reeve/Mâsse 2004).

Werden die Abstufungen auf der Antwortskala numerisch dargestellt bzw. benannt und nur die Endpunkte verbalisiert, so spricht man auch von einer numerischen Skala (*numerival scale*). Sind alle Abstufungen verbalisiert, heißt die Skala Verbalskala oder verbalisierte Skala (*verbal scale*). Antwortskalen beziehen sich in der Regel auf eine bestimmte *Antwortdimension.* Beispiele für Antwortdimensionen sind:

– Grad der Zustimmung (Zustimmungsskalen),

– Wichtigkeit (Wichtigkeitsskalen),

– Zufriedenheit (Zufriedenheitsskalen),

– Häufigkeit (Häufigkeitsskalen),

– Intensität (Intensitätsskalen; Grad der Stärke),

– Ausmaß, in dem eine Aussage auf einen Sachverhalt zutrifft („Trifft zu“-Skalen),

– Wahrscheinlichkeit (Wahrscheinlichkeitsskalen),

– Sympathie (Sympathieskalen),

– Interesse (Interessenskalen),

– Sicherheit,

– Betroffenheit.

Antwortdimensionen stellen durch Adjektive bezeichnete Eigenschaften dar, die in ihrer Intensität durch Adverbien bzw. adverbiale Modifikatoren (*adverbial modifiers, intensifiers, qualifiers*) abgestuft werden können.

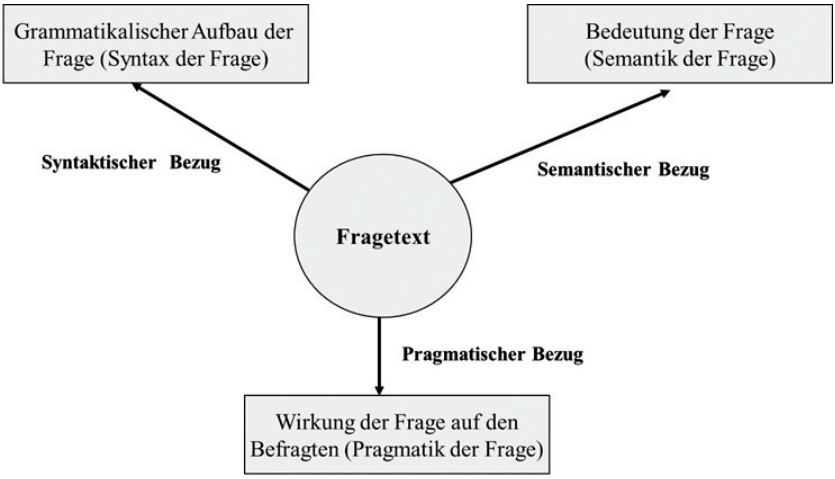
Verbale Skalen haben den Nachteil, dass die zur Verbalisierung der Skalenpunkte verwendeten Adverbien von unterschiedlichen Befragtengruppen unterschiedlich verstanden werden können (Wegener u. a. 1982; Schaeffer 1991). In Telefonumfragen müssen die Antwortkategorien vorgelesen und von der bzw. dem Befragten erinnert werden. Schon bei mehr als drei bis vier Quantifikatoren können dabei Primacy- oder Recency-Effekte auftreten, d. h., das erste oder das letzte Adverb wird erinnert. Zudem muss bei metrischer Interpretation verbaler Skalen die Gleichabständigkeit der verbalisierten Skalenpunkte gesichert sein (Rohrmann 1978). Die Bedeutung der Quantifikatoren hängt auch von den Antwortstrategien der Befragten ab (Hofmans u. a. 2007). Auf der anderen Seite entspricht die Verwendung verbaler Abstufungen eher dem Alltagssprachgebrauch und erscheint natürlicher. In der Regel wird man bei wenigen Abstufungen Verbalisierungen verwenden, bei vielen Abstufungen wie z. B. bei Fünfer- und Siebener-Skalen eher numerische Skalen.

5 Die Bedeutung von Fragetexten

Äußerungen im Interview, seien es mündliche oder schriftliche Fragen/Antworten, erhalten ihre kommunikative Funktion erst durch ihre Rolle als Zeichen. Unter semiotischer (zeichentheoretischer) Perspektive (Morris 1946) stellen Äußerungen im Interview Zeichen dar, die auf einem Zeichenträger wie Papier oder Bildschirm realisiert und durch drei Bezüge charakterisierbar sind: einen syntaktischen, einen semantischen und einen pragmatischen Bezug (*Abbildung 3*).

Abbildung 3:

Zeichentheoretische Bezüge von Fragetexten



Der syntaktische Bezug besteht darin, dass Zeichen nach bestimmten grammatikalischen Regeln erzeugt sind und insofern wohlgeformte sprachliche Ausdrücke darstellen. Einen semantischen Bezug haben Zeichen insofern, als sie etwas bedeuten, wobei zwischen der designativen, extensionalen Bedeutung (dem bezeichneten Gegenstand) und der detonativen, intensionalen Bedeutung (Sinn) unterschieden wird.

Der pragmatische Bezug thematisiert die Beziehung zwischen den Zeichen und ihren Nutzern. Er besteht darin, dass Zeichen in bestimmter Weise verwendet werden, etwa um bestimmte Ziele zu verfolgen bzw. bestimmte Wirkungen zu erzielen. Auch Fragen stellen Zeichen dar und stehen insofern ebenfalls in einem syntaktischen, semantischen und pragmatischen Bezug. Der syntaktische Aspekt bezieht sich also auf den grammatikalischen Aufbau der Frage, der semantische auf die Bedeutung der Frage und der pragmatische Aspekt auf die Frageverwendung und die Wirkung auf den Befragten.

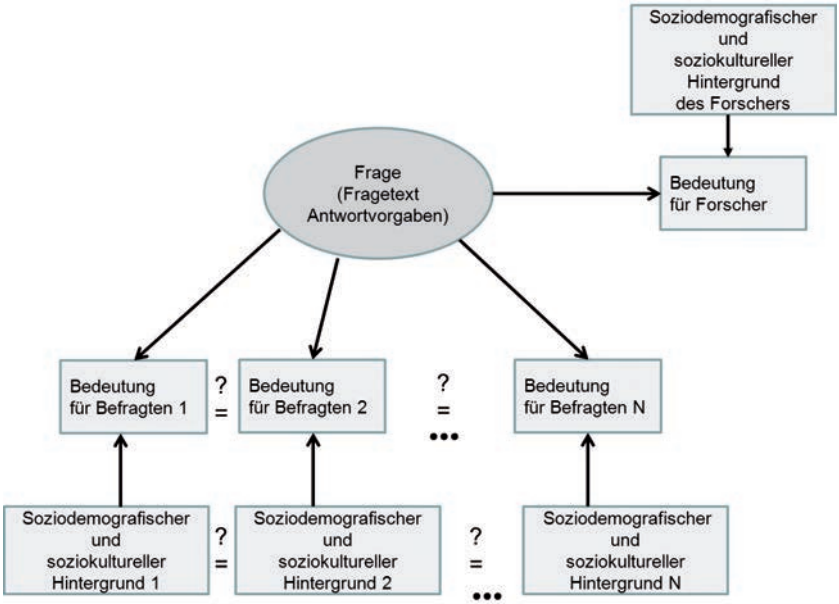
Von besonderer praktischer Bedeutung sind der semantische und der pragmatische Bezug von Fragen und Antwortvorgaben. Es ist bedeutsam, sich immer wieder klarzumachen, dass die Befragten nicht auf den Frage*text* reagieren, sondern auf die *Bedeutungen*, die sie dem Fragetext und den Bestandteilen der Antwortformate geben. Die Bedeutungen können bei Fragetexten mehr oder weniger komplex sein. Wenn z. B. in einem Item eine Episode beschrieben wird, besteht die designative Bedeutung in einer mehr oder weniger komplexen episodischen Struktur. Durch Worte bezeichnete Begriffe können in

eine mehr oder weniger komplexe Struktur von Begriffen eingebettet sein. Die Wirkung der Äußerungen geht dabei mitunter weit über die rein sprachliche Bedeutung hinaus, da diese in semantische Strukturen und Wissensstrukturen integriert wird und – über diese vermittelt – auch vergangene Erfahrungen, Ängste, Bilder etc. ansprechen kann. Dabei steht die Komplexität der semantischen Struktur nicht unbedingt in direktem Zusammenhang mit der Komplexität der syntaktischen Struktur. Schon einzelne Worte wie „Behörde“, „Straftat“, „Verbrechen“, „Opfer“, „Täter“, „Regierung“, „Familie“ sind in umfassendere Wissensstrukturen eingebettet.

Grundsätzlich sollten Forschende davon ausgehen, dass die Bedeutungszuordnungen von Fragetexten zwischen den Befragten variieren und sich auch von den Bedeutungen unterscheiden können, die sie selbst mit den sprachlichen Ausdrücken verbinden (*Abbildung 4*).

Abbildung 4:

Bedeutungsvariation von Fragen



In der Regel wird die Bedeutung von Fragen nicht hinterfragt. Stattdessen legt die bzw. der Forschende uneingestanden seine eigene Bedeutungswahrnehmung zugrunde und vergisst, dass die Befragten unterschiedliche Alltags-

interpretationen mit ihnen verbinden könnten, die von der von ihr bzw. ihm unterstellten Bedeutung abweichen. Verantwortlich für unterschiedliche Interpretationen von Fragetexten sind nicht zuletzt soziodemografische und soziokulturelle Unterschiede.

Interpretative Unterschiede zwischen Begriffsbezeichnungen wurden vor allem in den Arbeiten von Conrad und Schober (Conrad/Schober 2000; Conrad u. a. 2007; Schober/Conrad 1997; Schober u. a. 2004; Suessbrick u. a. 2000; Peytchev u. a. 2010; Redline 2013; Tourangeau u. a. 2006). untersucht. Im Mittelpunkt steht dabei der Begriff „Klärung" (clarification) von Bedeutungen; Klärungen von Bedeutungen können eingesetzt werden, um sicherzustellen, dass alle Befragten die Begriffe in der gleichen Weise interpretieren (z. B. Conrad/Schober 2000). Unter „Interpretation" wird dabei die Instantiierung bzw. Konkretisierung von Begriffen und Konstrukten verstanden. Instantiierung bezeichnet die Fixierung der semantischen Bedeutung. Dabei kann es sich um konkrete Ausprägungen des Begriffs oder um Beschreibungen handeln. So untersuchten Tourangeau u. a. (2006) Fehlzuordnungen (*malalignments*) von Begriffen und die mangelnde Übereinstimmung zwischen Begriff und Instanz des Begriffs in Bezug auf Alltagsbegriffe wie ‚Aufenthaltsort' (*residence*) und ‚Unfähigkeit' (*disability*) mithilfe von Vignetten als Träger der semantischen Definitionen. Ross und Murphy (1999) untersuchten die Instantiierungen von Nahrungsbegriffen (*food terms*). Schober und Conrad schlagen die Integration von Klärungen in das Interview vor und weichen damit bewusst von den strengen Regeln des standardisierten Interviews ab. Sie plädieren stattdessen für ein *conversational interviewing*.

Sollen Begriffe im Sinne der bzw. des Forschenden verstanden werden, empfiehlt es sich, Definitionen als Hilfen zur Verfügung zu stellen. Dies ist nicht immer einfach. Sowohl technische als auch juristische Definitionen sind für die bzw. den Befragten nicht immer verständlich. Ein Beispiel stellt die juristische Definition von Delikten dar, die bei Opfererfahrungen abgefragt werden. In diesem Fall muss versucht werden, die Begriffe so zu beschreiben, dass sie dem juristischen Verständnis entsprechen und gleichzeitig von der bzw. dem Befragten verstanden werden. Erleichtert werden kann die Transformation der Fachsprache in die Umgangssprache auch durch Beispiele (Tourangeau u. a. 2014). Eine geeignete Umsetzung fachsprachlicher Definitionen in die Umgangssprache ist gleichsam Voraussetzung für die Konstruktion von Screening-Instrumenten (*Screener*), mit deren Hilfe die Zielpersonen identifiziert werden können (Lynch 1993).

Angesichts der Globalisierung und zunehmender kultureller Heterogenität der befragten Teilgruppen der Bevölkerung in Umfragen ist damit zu rechnen, dass sich in allgemeinen Bevölkerungsumfragen kulturelle Unterschiede in den Bezeichnungen und den Konnotationen von Worten und sprachlichen Ausdrücken verstärkt niederschlagen. Worte mit quantitativen Ausprägungen

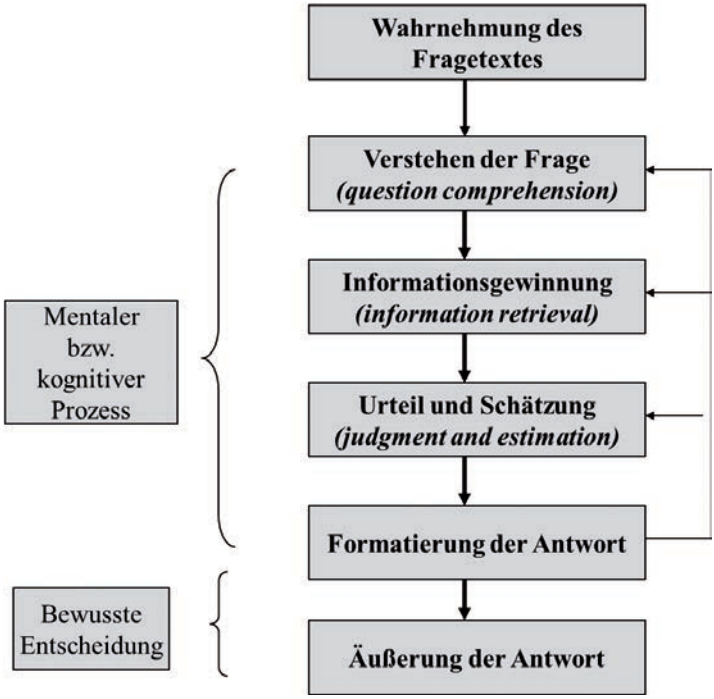
wie Häufigkeitsangaben können in Kombination mit Worten wie z. B. „Kriminalität“ durchaus mit unterschiedlichen subjektiven quantitativen Ausprägungen verbunden sein. Ausdrücke wie „zahlreiche Verbrechen“ können für jemanden, der in einer Region ohne nennenswerte Kriminalität aufgewachsen ist, etwas anderes bedeuten als für jemanden, der in einem sozialen Brennpunkt aufgewachsen ist. Weitere Beispiele sind Fragen nach der Schwere von Delikten, nach der Häufigkeit von Auseinandersetzungen in der Wohngegend oder nach der Wahrscheinlichkeit zukünftiger Opferwerdungen. Smith (2003, 2004) verglich länderübergreifend die Stärken adverbialer Modifikatoren. Schon eine frühe Studie von Kristof aus den 60er Jahren (Kristof 1966) über den Vergleich der Modifikatorstärken verschiedener adverbialer Modifikatoren zwischen Deutschland und den USA konnte starke Unterschiede deren quantitativer Bedeutungen nachweisen. So ist z. B. das angloamerikanische Adverb *slightly* mit einer stärkeren Intensität versehen als das deutsche Adverb *etwas*. Faulbaum, Wegener und Maag (1982) fanden Hinweise für Unterschiede in den Stärken zwischen Alters- und Geschlechtergruppen.

6 Der Antwortprozess

Die Erzeugung einer Antwort auf eine Frage erfordert von der bzw. dem Befragten das Erbringen einer Leistung. Diese Leistung beinhaltet u. a. bestimmte interne, d. h. in der bzw. dem Befragten ablaufende Handlungen und Prozesse, die in ihrem gesamten Ablauf als Antwortprozess (*response process*) bezeichnet werden (Tourangeau 1984, 1987; Tourangeau, u. a. 2000; *Abbildung 5*). Um eine Frage beantworten zu können, muss diese Frage zunächst wahrgenommen werden. Dabei stellt die akustische oder visuelle Wahrnehmung eines Fragetexts bereits eine Leistung der bzw. des Befragten dar, die nicht immer ohne geeignete Bewegungen des Körpers und des Wahrnehmungsorgans erbracht werden kann, das den Übertragungskanal der kommunizierten Frage kennzeichnet. So erfordert die visuelle Wahrnehmung eines Texts bei selbst administrierten Fragen die Fähigkeit zu lesen und damit bestimmte Blickbewegungen auszuführen (Jenkins/Dillman 1997). Das Hören einer Frage am Telefon erfordert, dass das Telefon an das Ohr gehalten und überhaupt in der durch die Klingeldauer vorgegebenen Zeit erreicht werden kann etc. Diese Beispiele zeigen, dass die Leistung eventuell nicht von allen Befragten erbracht werden kann, sofern nicht vorher eine Anpassung an die Leistungsfähigkeit der Befragten erfolgt ist. Diese Anpassung bezieht sich auch auf die Wahl einer geeigneten Kommunikationsform (z. B. selbst administriert versus Interviewer-administriert; Web versus postalisch; CAPI vs. CATI).

Abbildung 5:

Modell des Antwortprozesses



An der Erzeugung einer Antwort sind neben der Wahrnehmung alle zur Interpretation des Fragetexts notwendigen kognitiven Prozesse des Sprachverstehens beteiligt – Abruf syntaktischen (grammatikalischen), semantischen und pragmatischen Wissens, Aufbau semantischer Repräsentationen, Prozesse der Informationsgewinnung, Abrufs von Erfahrungen, Erinnerungen etc., Datierung von Ereignissen, Prozesse der Urteilsbildung einschließlich Auswahl von Entscheidungsalternativen und Prozesse der Informationsintegration sowie Schätzung einschließlich eventuell geforderter Berechnungen. Schließlich muss die Antwort formatiert und dann auch tatsächlich geäußert werden. Ob die gefundene Lösung (Antwort) tatsächlich geäußert wird oder nicht, muss als bewusste Entscheidung der oder des Befragten angesehen werden, die bzw. der diese Entscheidung noch einmal vor dem Hintergrund möglicher Nachteile für sich selbst überprüft, wozu insbesondere die Konsequenzen für das Selbstkonzept (Überblick über Selbst und Identität bei Leary 2007) gehö-

ren. Das dargestellte Modell kann in mehrerer Hinsicht weiter detailliert, modifiziert und/oder erweitert werden. So ist davon auszugehen, dass für die Lösung der im Fragetext erforderlichen Aufgabe weitere spezifische mentale Prozesse wie z. B. Schlussfolgerungsprozesse sowie beobachtbare Handlungen wie z. B. das Heraussuchen einer Rechnung, eines Vertrags etc. erforderlich sein können. Die Stufen des Antwortprozesses dienen auch als Hintergrundfolie für die Dimensionen, die in Fragebewertungssystemen angesprochen werden (siehe unten), sowie für den Vergleich verschiedener Administrationsformen (Guzy/Leitgöb 2014).

Von Bedeutung ist, dass die zur Erzeugung der Antwort notwendigen Prozesse tatsächlich stattfinden. Lassen aufgrund eines zu langen Fragebogens Konzentration und Motivation der Befragten nach, kann es zu einem Phänomen kommen, das als *Satisficing* bezeichnet wird und zur Herausbildung inhaltsunabhängiger Antwortstrategien führen kann (Krosnick 1991). Beim Satisficing werden die notwendigen mentalen Operationen entweder unvollständig (schwaches Satisficing) oder überhaupt nicht (starkes Satisficing) durchlaufen.

7 Determinanten des Antwortverhaltens und ihre Wirkungen

7.1 Überblick

Wichtige Determinanten des Antwortverhaltens sind:

- Nichterfüllung/Nichterfüllbarkeit der durch die Frage definierten Leistungsanforderungen und ihre Ursachen (z. B. physische Ausstattung, kognitive Kompetenz etc.),
- sensitive Bedeutung von Fragetexten und ihre Ursachen,
- Effekte der Kommunikationsform bzw. Befragungsart (Modeeffekte),
- Frageformulierungen,
- Gestaltung von Skalen und Antwortvorgaben,
- Gestaltung des Layouts,
- Anwesenheit von Interviewerinnen oder Interviewern,
- Stellung der Frage im Fragebogen (Kontext),

– Anwesenheit Dritter wie etwa der Täterin oder des Täters (z. B. Ehepartnerinnen oder -partner) beim Interview.

Die Wirkungen dieser Merkmale können in nicht adäquaten Antworten bestehen:

– Antwortverweigerungen,

– nicht vorgelesene bzw. nicht zu den zugelassenen Alternativen gehörende Antworten wie „weiß nicht“ (*don't know*; kurz: *DK*), „keine Meinung“ (*no opinion* bzw. *non attitude)* oder „trifft nicht zu“,

– ungenaue Antworten (z. B. „prima“ statt „sehr gut“, vage Antworten wie „ungefähr 10 Tage“),

– statt einer Antwort spontane Kommentare (z. B. „Das ist aber eine schwierige Frage“, „Die Frage verstehe ich nicht“, „Es ist unverschämt, mir eine solche Frage zu stellen“ etc.),

– voreilige Antworten, die sich nicht auf den vollständigen Fragetext beziehen können, da die bzw. der Befragte ihn gar nicht vollständig wahrgenommen hat.

Bei einigen nicht adäquaten Antworten wie z. B. ungenauen Angaben kann durch den Einsatz neutraler Nachfragetechniken versucht werden, die bzw. den Befragten schließlich zu einer adäquaten Antwort zu bewegen und somit eine Korrektur der Nicht-Adäquatheit zu erreichen (Prüfer/Stiegler 2002).

– Falschangaben: Es werden bewusst falsche Angaben gemacht. Diese treten etwa bei sensitiven Items bzw. Fragen zu sensiblen Delikten (z. B. bei sexueller Viktimisierung, häuslicher Gewalt) auf.

– Fehlerhafte Angaben: Angaben können aus unterschiedlichen Gründen fehlerhaft sein. Zu solchen Gründen zählen z. B. Gedächtniseffekte bei Erinnerungsaufgaben wie *Telescoping* (Neter/Waksberg 1964) oder *Vergessen.* Beim Telescoping können Ereignisse näher (*forward telescoping*) oder weiter entfernt vom Datum des Interviews (*backward telescoping*) erinnert werden. Ereignisse können irrtümlicherweise außerhalb (*external telescoping*) oder innerhalb einer Referenzperiode (*internal telescoping*) datiert werden. Diese Fehler werden umso größer, je weiter ein Ereignis zurückliegt. In jedem Fall werden Genauigkeit und Zuverlässigkeit der gelieferten Information beeinträchtigt. Um Telescopingeffekte zu vermeiden, werden in Viktimisierungsbefragungen immer Fünf- und Ein-Jahres-Prävalenzen abgefragt.

Weitere Gründe für fehlerhafte Angaben können Verständnisprobleme sein. Ein Beispiel sind Verständnisschwierigkeiten bei der Deliktdefinition mit der Folge entsprechender Zuordnungsschwierigkeiten insbesondere bei schwer abgrenzbaren Delikten wie z. B. Wohnungseinbruchsdiebstahl.

Messtheoretische Effekte betreffen die Zuverlässigkeit und die Validität von Fragen/Items (z. B. Bohrnstedt 2010; Lord/Novick 1968; Zeller/Carmines 1980). Die Zuverlässigkeit bzw. Reliabilität bezieht sich auf die Abweichung der beobachteten Messung vom wahren Wert einer Messung, d. h. den Messfehler. Je größer der Messfehler ausfällt, desto geringer ist die Reliabilität. Die Zuverlässigkeit einer Messung ist formal definiert als:

$$p_x = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_\varepsilon^2}{\sigma_x^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_x^2}.$$

In dieser Formel ist σ_τ^2 die Varianz der wahren Messungen (auch: *wahre Varianz*) und σ_ε^2 die Varianz der Fehlervariablen (auch: *Fehlervarianz*).

Validität bezeichnet den Grad bzw. das Ausmaß, mit dem ein Instrument (z. B. Test, Fragebogen, Item) das zu untersuchende Konstrukt misst. So ist etwa eine valide Messung der Geschlechterrollenidentität dann gegeben, wenn dieses Konstrukt und kein anderes durch die Messung erfasst wird. Kürzer ausgedrückt bezeichnet Validität das Ausmaß, in dem ein Messinstrument das misst, was es messen soll. Als theoretische Validität wird die Korrelation eines Indikators mit dem Konstrukt, das er messen soll, bezeichnet. Man kann zeigen, dass die theoretische Validität der Quadratwurzel aus der Reliabilität entspricht.

Die theoretische Validität lässt sich in Messmodellen mit multiplen Indikatoren bestimmen. Sie entspricht in diesem Fall den Faktorenladungen, wenn alle Variablen standardisiert, d. h. z-transformiert und die Fehler der Messungen nicht korreliert sind. Die Ladungen lassen sich mithilfe von Techniken der konfirmatorischen Faktorenanalyse schätzen (z. B. Reinecke 2014). Modelle für den Zusammenhang zwischen Konstrukten und ihren Indikatoren werden auch als Messmodelle bezeichnet (z. B. Bohrnstedt 2010; Reinecke 2014).

7.2 Sensitive Effekte und soziale Erwünschtheit

Da die Forscherin bzw. der Forscher erwartet, dass die bzw. der Befragte nur auf den Inhalt der Frage reagiert, stellen die beschriebenen Effekte in der Regel unerwünschte Nebeneffekte dar, die es zu reduzieren gilt. Die volle Breite der Befunde zu diesen Effekten kann an dieser Stelle nicht dargestellt werden.

Für Erhebungen im Bereich der Viktimisierung sind vor allem die *sensitiven Wirkungen* von Bedeutung, die durch Assoziationen des Fragetexts mit traumatischen Erlebnissen im Zusammenhang mit eigenen Opfererfahrungen, durch Verwicklungen in kriminelle, von der Gesellschaft sanktionierte Handlungen etc. entstehen können.

Fragen/Items, die sensitive Wirkungen erzeugen können, werden gemeinhin als *sensitive Fragen/Items* bezeichnet. Sie können nicht nur zu nicht adäquaten Antworten (siehe unten) führen, sondern auch sowohl den Messfehler als auch die inhaltliche Validität beeinflussen (siehe Überblick bei Tourangeau/Yan 2007). Obgleich eine verbindliche Definition einer sensitiven Frage schwer zu leisten ist, können folgende Arten von Fragen als sensitiv angesehen werden:

– Fragen, die zu sozial erwünschten (*socially desirable*) Antworten führen,

– Fragen, welche die Privatheit (*privacy*) der Befragten bedrohen,

– Fragen, die ein Risiko zur Enthüllung (*disclosure*) gegenüber Dritten beinhalten (Tourangeau u. a. 2000).

Eine Antwort auf eine Frage ist dann sozial erwünscht, wenn sie von der bzw. dem Befragten unter Berücksichtigung der Normen ihrer bzw. seiner Bezugsgruppe gegeben wird (zur Definition der sozialen Erwünschtheit DeMaio 1984; Edwards 1957; Krebs 1987). Eine im Zusammenhang mit der sozialen Erwünschtheit viel diskutierte Frage ist, ob die Tendenz, sozial erwünscht zu antworten, eine stabile Persönlichkeitseigenschaft, also eher einen *Trait* (Crowne/Marlowe 1964) oder eine itemspezifische, eher temporäre Reaktion bestimmter Respondenten auf bestimmte Fragen darstellt, also Ausdruck einer Strategie, mit den Inhalten der Frage umzugehen, ist. Paulhus (2002) unterscheidet zwischen Antwortstilen im Sinne eines über Fragebögen und Zeiten hinweg konsistenten Antwortverhaltens und einer temporären, aus der augenblicklichen Motivation entstandenen Antworttendenz (*response set*). Detailliertere Analysen ergaben Hinweise auf eine differenzierte Zusammensetzung des Konstrukts der sozialen Erwünschtheit. Beispiele für Bestandteile dieses Konstrukts sind die Neigung, eigene Fehler zuzugeben, und der „moralistische Bias“ im Sinne eines übertriebenen Gefühls für die eigenen moralischen Qualitäten (Paulhus 2002).

Die Privatheit bedrohende Fragen erkundigen sich z. B. nach dem Einkommen oder einer Wahlentscheidung (Partei, Kandidat; sogenannte Sonntagsfrage). Ein Risiko zur Enthüllung gegenüber Dritten ist gegeben, wenn Befragte die Gefahr sehen, dass ihre Antwort an Dritte weitergeben wird. Der Dritte

kann die Interviewerin oder der Interviewer sein, anwesende dritte Personen oder Organisationen (z. B. bei Mitarbeiterbefragungen).

Zahlreiche Studien belegen die Wirkungen sensitiver Fragen – insbesondere solcher nach Einkommen, Drogengebrauch oder Sexualität – auf das Ausmaß an *Item Nonresponse* (Tourangeau u. a. 1997). Auch bewusst falsche Angaben sind bei sensitiven Fragen durchaus zu erwarten (Jobe u. a. 1997). Dies belegen auch Studien zum *Overreporting* und *Underreporting* von Ereignissen, bei denen absichtlich falsche Tatsachenbehauptungen vorliegen. Das Phänomen des *Overreporting* betrifft den Sachverhalt, dass Personen Ereignisse und Verhaltensweisen berichten, die nicht stattgefunden haben. Man fand dieses Verhalten verstärkt bei Nichtwählerinnen und Nichtwählern in Bezug auf die Fragen nach der Teilnahme an politischen Wahlen. So bestand bei Nichtwählerinnen und Nichtwählern offensichtlich die Tendenz, statt ihrer tatsächlichen Nichtteilnahme eine Teilnahme zu berichten (Belli u. a. 1999; Bernstein 2001). Als Ursachen werden soziale Erwünschtheit und der Versuch, das Gesicht zu wahren, genannt. Das beschriebene Verhalten tritt offensichtlich vor allem bei gebildeteren und religiöseren Personen auf. Ein weiteres Beispiel ist das Anzeigeverhalten in Opferbefragungen. So wird regelmäßig angenommen, dass in Opferbefragungen mehr Personen angeben, eine Opferwerdung bei der Polizei angezeigt zu haben, als dies tatsächlich der Fall ist (dazu auch Heinz in Band I; Averdijk/Elffers 2012; Schwind u. a. 2001).

Neben dem Overreporting spielt bei sensitiven Fragen auch das *Underreporting* eine Rolle, z. B. von Abtreibungen (Peytchev u. a. 2010). Als Erklärung bietet sich in diesem Fall die Angst vor sozialer Stigmatisierung an. Mit ähnlichen Ergebnissen kann auch bei Erhebungen gerechnet werden, in denen Befragte gebeten werden, Angaben zu eigenem kriminellen Verhalten zu machen. Auch in Bezug auf Viktimisierungsbefragungen werden sexuelle Delikte regelmäßig nicht angegeben, worauf z. B. mode-vergleichende Studien hindeuten (Guzy/Leitgöb 2014 auch mit weiteren Belegen).

Die beschriebenen sensitiven Effekte stellen in Bezug auf die Forschungsfragestellung in der Regel unerwünschte und manchmal nicht intendierte Nebeneffekte dar. Sie führen zu einer Gefährdung der inhaltlichen Validität. Reagieren Befragte auf bestimmte Fragen sensitiv, besteht die Gefahr, dass mit der Frage nicht das von der bzw. dem Forschenden intendierte Konstrukt, sondern ausschließlich oder zusätzlich ein anderes Konstrukt wie z. B. soziale Erwünschtheit gemessen wird bzw. die ursprünglich für die Messung eines bestimmten Konstrukts vorgesehenen Indikatoren auch oder ausschließlich Indikatoren eines anderen Konstrukts sind. Nur in einem entsprechend gestalteten Umfragedesign lassen sich Effekte der beiden Konstrukte auf die beobachteten Indikatoren voneinander getrennt schätzen, sodass man das Ausmaß beurteilen kann, indem das Konstrukt der sozialen Erwünschtheit die Fragen/

Items beeinflusst. So kann etwa eine Skala der sozialen Erwünschtheit explizit in den Fragebogen aufgenommen werden, um sie als Kontrollvariable in statistische Modelle einzuführen (für weitere Ausführungen Waubert de Puiseau u. a. in diesem Band).

Bestimmte Effekte treten erst bei der Positionierung von Fragen im Fragebogen oder bei der Positionierung von Items in einer Itembatterie auf. Diese Effekte werden auch als Kontexteffekte oder Reihenfolgeeffekte bezeichnet und ihnen kommt gerade in Opferbefragungen eine besondere Bedeutung zu. Stellt man kriminalitätsbezogene Einstellungsfragen nach den Fragen zu Opfererlebnissen, besteht die Gefahr, dass diese durch die Erinnerung an das Opferwerden beeinflusst werden. Screeningfragen zu Opfererlebnissen sollten am Anfang des Fragebogens im Block abgefragt werden, damit die bzw. der Befragte nicht „lernt", dass ein „Ja" zu zahlreichen Nachfragen führt. Aufgrund von Überschneidungen bei Deliktdefinitionen sollte der Wohnungseinbruch vor weiteren speziellen Diebstahldelikten (d. h. Auto-, Fahrraddiebstahl) gefolgt von sonstigen Diebstählen abgefragt werden.

8 Verfahren zur Evaluation der Fragequalität

8.1 Überblick

Entwürfe von Fragen und Fragebögen bedürfen zur Optimierung ihrer Qualität und zur Abschätzung ihrer Wirkungen auf die Befragten stets einer Evaluation. Die Optimierung der Qualität erfordert zunächst die Identifikation möglicher Qualitätsbeeinträchtigungen bzw. Schwächen des Erhebungsinstruments. Auf der Basis dieser Diagnose können dann Verbesserungen des Entwurfs erfolgen, die wiederum in den Diagnoseprozess zurückgespielt werden müssen. Die Schwächenanalyse sollte sich auf sämtliche Aspekte eines Erhebungsinstruments beziehen, um möglichst alle negativen Wirkungen auf die Befragten im Interview auszuschließen. Zu diesen Aspekten gehören neben den Fragetexten das Layout der Fragen (bei selbst administrierten Interviews), die Interviewer-Instruktionen (bei Interviewer-administrierten Interviews) und die Navigation durch den Fragebogen. Einige dieser Aspekte wie z. B. das Verständnis von Fragetexten können von anderen Aspekten wie etwa dem Layout und dem Navigationsverhalten getrennt evaluiert werden.

Bei der Evaluation von Erhebungsinstrumenten kann ein Repertoire unterschiedlicher Verfahren zur Diagnose von Schwächen bzw. Qualitätsbeeinträchtigungen herangezogen werden. Dazu gehören:

– Fragebewertungssysteme (*question appraisal systems*),

– Expertenrunden, eventuell unter Einbeziehung von Fragebewertungssystemen,

– Gruppendiskussionen,

– empirische Pretestverfahren einschließlich der statistischen Analyse der Antwortverteilungen und der statistischen Überprüfung der Gütekriterien der Messung,

– Verfahren zur Überprüfung der Benutzerfreundlichkeit (*usability*) von Layout und Navigation bei CASI-Umfragen unter Einsatz spezifischer Vorrichtungen wie *Eye Tracking* zur Analyse von Blickbewegungen sowie spezifischer Hard- und Softwaretechnologien,

– Simulationsverfahren zur Funktionsprüfung programmierter Fragebögen,

– Verfahren zur Überprüfung der messtheoretischen Qualität, d. h. der Reliabilität und Validität.

Die praktischen Verfahren zur Überprüfung der messtheoretischen Qualität erfordern besondere Erhebungsdesigns und den Einsatz statistischer Verfahren. Die Darstellung dieser Verfahren und ihrer theoretischen Grundlagen ist hier leider nicht möglich. Stattdessen sei auf die weiterführende Literatur verwiesen (Bohrnstedt 2010; Saris/Gallhofer 2014).

Die Nutzung von Fragebewertungssystemen erlaubt bereits vor dem Einsatz empirischer Evaluationsverfahren eine Diagnose von Qualitätsbeeinträchtigungen bei Fragetexten. Grundlage solcher Systeme sind Klassifikationen von Problemen, die bei Fragen/Items auftreten können. Es ist anzuraten, bereits vor dem Einsatz empirischer Evaluationsverfahren Fragebewertungssysteme einzusetzen, damit sich die empirischen Verfahren auf die noch verbleibenden Probleme konzentrieren können. Mit dem Einsatz von Fragebewertungssystemen werden bereits im Vorfeld Reaktionen der Befragten ausgeschlossen, die sich auf offensichtliche Mängel der Fragen beziehen. Allerdings können Fragebewertungssysteme oft nur *mögliche* Gefährdungen der Qualität aufzeigen. Für einen konkreten Befragten muss eine über das Bewertungssystem als sensitiv bewertete Frage nicht unbedingt sensitiv sein.

Expertenrunden dienen in der Entwurfsphase dazu, einen Erstentwurf weiter durch den gemeinsamen Austausch von Argumenten zu verbessern und Qualitätsmängel aufzudecken. Im Kern geht es um die Verbesserung der Opera-

tionalisierung, indem z. B. geprüft wird, ob es angemessenere Inhalte für den Erstentwurf gibt, die das Konstrukt besser repräsentieren.

Gruppendiskussionen über Entwürfe von Fragen können dazu beitragen, Unterschiede im Frageverständnis zu entdecken. Teilnehmende könnten z. B. Personen mit Opfererfahrung und Personen ohne Opfererfahrung sein.

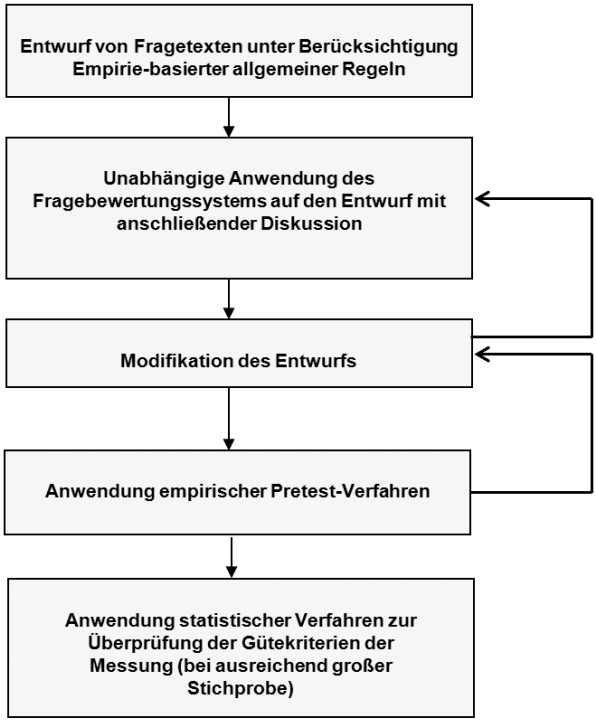
Empirische Pretestverfahren erlauben die Entdeckung von Problemen, die Befragte mit Fragen haben, und die Untersuchung des Frageverständnisses.

Verfahren zur Evaluation der Benutzerfreundlichkeit können zur Untersuchungen der Wirkung des Layouts und der Funktionalität programmierter Erhebungsinstrumente eingesetzt werden, wobei zahlreiche technische Hilfsmittel verwendet werden können, wie z. B. Eye-Tracking-Verfahren oder Videoaufzeichnungen. Insbesondere zur Entdeckung von Navigationsproblemen und Problemen der Filterführung, also zur Funktionsprüfung programmierten Fragebogen können auch Simulationen durchgeführt werden, indem der Fragebogen unter allen Besetzungen der Werte von Filtervariablen geprüft wird. Insbesondere Fehler in der Fragebogenprogrammierung lassen sich auf diese Weise identifizieren.

Die Evaluation von Fragebögen und Fragen ist in der Regel kein einfacher linearer Prozess. Nicht selten muss das evaluierte Instrument nach der Beseitigung von Mängeln erneut eine Evaluation durchlaufen, um die verbesserten Entwürfe noch einmal zu prüfen. *Abbildung 6* gibt einen Überblick über einen möglichen Ablauf der Qualitätsprüfung eines Fragebogens. In der Regel unterbleibt der letzte Schritt bezüglich des erforderlichen Stichprobenumfangs wegen zu hoher Kosten eines geeigneten Designs (z. B. bei Test-Retest-Verfahren, Fehlen multipler Indikatoren etc.).

Abbildung 6:

Beispiel für die Optimierung des Textentwurfs einer Frage



8.2 Die Evaluation *on desk*: Fragebewertungssysteme

Fragebewertungssysteme evaluieren jede Frage anhand eines Problemkatalogs. Beispiele für Fragebewertungssysteme sind das Bewertungssystem von Willis und Lessler (1999) und das darauf basierende Bewertungssystem von Faulbaum, Prüfer und Rexroth (2009). Der Problemkatalog des Systems von Willis und Lessler enthält folgende Problemarten:

Interviewer-bezogene Probleme (bei Interviewer-administrierten Interviews):

- Interviewer weiß nicht, welchen Teil der Frage er vorlesen soll,
- fehlende Informationen (Interviewer fehlen die Informationen, um die Frage angemessen zu stellen),
- Frage ist schwierig zu lesen (z. B. nicht voll ausgeschrieben);

Befragten-bezogene Probleme:

- Probleme mit Einleitungen, Anweisungen oder Erklärungen aus Sicht der Befragten (z. B. ungenaue oder widersprüchliche Anweisungen, komplizierte Anweisungen),
- unklare Bedeutung oder Absicht der Frage, wie
  - Probleme, die sich auf die Formulierung beziehen (z. B. Frage zu lang, zu komplizierte Wortwahl, grammatikalisch nicht korrekt),
  - Fachausdrücke,
  - Ungenauigkeit bzw. Mehrdeutigkeit,
  - Probleme, die sich auf die Definition der Zeiträume beziehen: Zeiträume sind ungenau, widersprüchlich oder gar nicht definiert,
- Probleme, die mit Annahmen über die Befragten zusammenhängen, wie
  - unangemessene Annahmen über den Befragten,
  - irrtümliche Annahmen einer Verhaltenskonstanz oder Konstanz von Erfahrungen, obwohl eine solche nicht existiert,
  - Ambivalenz: Frage beinhaltet mehr als nur eine Aussage,
- fehlendes Wissen/Erinnerungsvermögen, wie
  - fehlendes Wissen: Es ist unwahrscheinlich, dass der Befragte über das zur Beantwortung der Frage notwendige Wissen verfügt,
  - Erinnerung: Der Befragte ist nicht in der Lage, die Information aus dem Gedächtnis abzurufen,

  - Berechnung: Um die Frage zu beantworten, müssen aufwendigere Berechnungen angestellt werden,
  - Einstellung: Es ist unwahrscheinlich, dass der Befragte über die zu erhebende Einstellung verfügt,
- Sensibilität/Beeinflussung, wie
  - sensible Inhalte: Die Frage spricht peinliche oder private Themen an,
  - sensible Wortwahl,
  - soziale Akzeptanz: Der Befragte beantwortet die Frage entsprechend der sozialen Erwünschtheit,
- Antwortkategorien, wie
  - offene Fragen,
  - fehlende Übereinstimmung von Fragetext und Antwortkategorien: Die Antwortkategorien passen nicht zu den Fragen,
  - Fachausdrücke sind undefiniert, unklar oder zu komplex (z. B. bei Chemikalien und Medikamenten),
  - Ungenauigkeit: Ungenau formulierte Antwortkategorien lassen mehrere
  - Interpretationsmöglichkeiten zu,
  - Überschneidungen: Es existieren Antwortkategorien, die sich überschneiden,
  - fehlende Antwortkategorien: Es fehlen zu möglichen Antworten die Antwortkategorien,
- unlogische Anordnung.

Das umfassendere System von Faulbaum, Prüfer und Rexroth wird anhand von über 100 Beispielen aus bekannten nationalen und internationalen Umfragen erläutert.

8.3 Empirische Pretestverfahren

Abbildung 7 gibt einen Überblick zu Pretestverfahren, die vor allem bei Fragen und Fragebögen für Interviewer-administrierte Interviews Anwendung finden, von denen sich aber einige einzeln oder in Kombination auch bei Pretests für selbst administrierte Erhebungsinstrumente wiederfinden. Bei selbst administrierten Erhebungsinstrumenten kommen spezifische Verfahren zum Test der Handhabbarkeit (*usability*) hinzu.

Abbildung 7:

Pretestverfahren im Überblick (Faulbaum u. a. 2009, 96)

Von ganz besonderer Bedeutung ist bei Pretestverfahren die Zusammensetzung der Preteststichprobe. Probleme der Befragten mit Fragen lassen sich nur erkennen, wenn die Preteststichprobe Befragtenmerkmale abdeckt, die für das Verstehen von Fragen von Bedeutung sind. Fragen zum Thema „Arbeitslosigkeit“ erfordern z. B., dass auch Personen ohne Arbeit in die Preteststichprobe aufgenommen werden. Bei allgemeinen Bevölkerungsumfragen sollte in jedem Fall darauf geachtet werden, dass die Breite der soziodemografischen Merkmale Alter, Geschlecht und Bildung vertreten ist. Um dies zu garantieren, sollte unabhängig davon, ob es sich um ein kognitives Interview oder einen Feldpretest handelt, eine Quotierung nach zentralen soziodemografischen Variablen vorgenommen werden.

8.3.1 Kognitive Interviews

Kognitive Interviews (Willis 2005; Miller u. a. 2014; Prüfer/Rexroth 2005) sind ein Werkzeug zur Evaluation des Frageverständnisses und sollen einen Einblick in die kognitiven Prozesse während der Beantwortung von Fragen vermitteln. Eine Forscherin bzw. ein Forscher sollte grundsätzlich nicht davon ausgehen, dass sein eigenes Begriffsverständnis mit dem der Befragten übereinstimmt. Viele Forschende wären vermutlich überrascht, vielleicht sogar entsetzt, wenn sie zur Kenntnis nehmen müssten, wie weit das Verständnis der Befragten von ihrem eigenen entfernt ist. Zahlreiche statistische Modelle müssten vermutlich anders interpretiert werden. Es ist sogar damit zu rechnen, dass das Problem der mangelnden Vorhersagbarkeit und der Heterogenität des Frageverständnisses aufgrund der zunehmenden Heterogenität der Bevölkerung im Zusammenhang mit der Zunahme des Anteils von Personengruppen unterschiedlichen Migrationshintergrunds eher weiter zunehmen wird.

Konkret sollen mit den kognitiven Techniken eines kognitiven Interviews die folgenden Fragen beantwortet werden:

- Wie kommen die Antworten zustande?
- Was denken Befragte bei der Beantwortung einer Frage?
- Wie verstehen Befragte Fragen oder Begriffe?
- Verstehen Befragte Fragen so, wie von der bzw. dem Forschenden intendiert?

Um diese Fragen zu beantworten, können folgende Techniken eingesetzt werden:

- Nachfragetechniken (*Probing*),
- Paraphrasieren (*Paraphrasing*),
- Bewertung der Verlässlichkeit der Antwort (*Confidence Rating*),
- Sortiertechniken (*Card Sorting*),
- Technik des lauten Denkens (*Thinking Aloud*).

Alternative Einteilungsprinzipien rechnen das Paraphrasieren zu den Nachfragetechniken. Bei der Anwendung kognitiver Techniken kann man offen oder standardisiert vorgehen. Im Fall eines offenen Vorgehens sind die Techniken

und Nachfragen vor dem kognitiven Interview nicht festgelegt. Bei der standardisierten Vorgehensweise werden die Techniken bzw. Nachfragen vor dem Interview festgelegt und sind der Testleiterin bzw. dem Testleiter fest vorgegeben. Antworten der bzw. des Befragten sollten von der bzw. dem Testleitenden auch dann hinterfragt werden, wenn sie formal korrekt sind und das Verhalten der Testperson auf keine Probleme schließen lässt. Kognitive Interviews werden in der Regel persönlich-mündlich durchgeführt. Es gibt aber durchaus erfolgreiche Versuche, diese Interviews auch telefonisch zu führen (z. B. Noel 2013).

Mit Nachfragetechniken (*Probing*) werden mittels einer oder mehrerer Nachfragen Fragetexte, Begriffe oder gegebene Antworten hinterfragt – stets mit dem Ziel, über das Verständnis der Frage mehr Information zu erhalten. Dabei können folgende Arten der Nachfrage unterschieden werden:

- Nachfragen zum Verständnis (*comprehension probing*),
- Nachfragen zur Wahl der Antwortkategorie (*category selection probing*),
- Nachfragen zur Erinnerungsfähigkeit und zum vorhandenen Wissen (*information retrieval probing* bzw. *recall probing*).

Werden Nachfragen auf bestimmte Begriffe in der Antwort des Befragten bezogen, spricht man von bedingten Nachfragen (*conditional probing*). Beziehen sich Nachfragen auf keinen spezifischen Aspekt der Frage, spricht man von einer unspezifischen Nachfrage.

Sollen begriffliche Abgrenzungen bei den Befragten erhoben werden, empfiehlt sich die Integration von Sortiertechniken, mittels derer sich ermitteln lässt, welche empirischen Instanzen unter einem Begriff subsumiert werden. Beispielsweise kann erhoben werden, welche Verhaltensweisen zur Kleinkriminalität gerechnet werden, was nach Meinung der Befragten als Verbrechen gewertet wird oder einen Verkehrsunfall darstellt.

8.3.2 Feldpretest

Unter einem Feldpretest (auch: Standardpretest, konventioneller Pretest, klassischer Pretest, Beobachtungspretest) versteht man eine vom Stichprobenumfang her stark verkleinerte Testerhebung eines Fragebogens am Ende der Fragebogenentwicklung unter möglichst realistischen Bedingungen der Haupterhebung – streng genommen eine Simulation der Hauptstudie, d. h., dass er in derselben Befragungsart wie die Haupterhebung durchgeführt werden sollte. Der klassische Feldpretest wird in erster Linie bei Interviewer-administrierten Interviews eingesetzt. Dabei beobachtet die Interviewerin bzw.

der Interviewer, welche Probleme und Auffälligkeiten aufseiten der oder des Befragten auftreten, ohne diese aktiv zu hinterfragen (passive Vorgehensweise). Die beobachteten Probleme werden von der bzw. dem Interviewenden während des Interviews notiert und anschließend in sogenannten Erfahrungsberichten/Pretest-Reports fragenspezifisch dokumentiert. Die passive Vorgehensweise des Verfahrens liefert erfahrungsgemäß eher oberflächliche und begrenzte Informationen zum Frageverständnis. Das eigentliche Ziel des Feldpretests besteht also darin, neben der Überprüfung der durch passive Beobachtung feststellbaren Probleme des Frageverständnisses den gesamten Ablauf des Interviews und den gesamten Fragebogen – auch in technischer Hinsicht – zu testen.

Ein Feldpretest ist in der Regel relativ schnell und problemlos realisierbar. Der organisatorische Aufwand ist eher gering und die Kosten sind insbesondere bei Anwendung eines Quotenverfahrens moderat. Er liefert im Allgemeinen verlässliche Informationen über technische Mängel des Fragebogens und die Handhabbarkeit durch die Interviewenden. Neben Informationen über spontane, nicht adäquate Antworten und Kommentare der Befragten erlaubt der Feldpretest die Analyse von Antwortverteilungen und annähernd realistische Schätzungen der durchschnittlichen Interviewdauer sowie für jede Frage die durchschnittliche Dauer eines Frage-Antwort-Dialogs. Da ein Feldpretest einen Datensatz liefert, lassen sich durch dessen Analyse auch Filterfehler entdecken, die auf die Programmierung (bei programmierten Erhebungsinstrumenten) oder das Fragebogendesign zurückgehen.

Erweitert man den Begriff des Feldpretests auf selbst administrierte Interviews, könnte man auch Probeläufe postalischer Umfragen oder Webumfragen mit kombiniertem Selbstausfüllen eines Fragebogens in die Klasse der Feldpretests einordnen. Feldpretests können mit verschiedenen Dokumentationsformen verbunden sein.

So werden im Rahmen des Befragten- und/oder Interviewer-Debriefings die Befragten und/oder die Interviewer im Anschluss an das Interview noch einmal retrospektiv zu einzelnen Fragen und zum gesamten Verlauf des Interviews befragt. Dies können auch ausführliche Interviews zum Frageverständnis sein (sog. Intensivinterviews). Interviewer-Debriefings dienen der Erhebung von Informationen über:

– Dauer der Befragung (falls nicht durch die Befragungssoftware automatisch erhoben),

– Auftreten von Unterbrechungen und die Frage/das Item, bei der/dem die Unterbrechung auftritt,

– Interessantheit des Interviews für die Befragten,

– Interessantheit des Interviews für die Interviewenden,

– Schwierigkeit des Interviews für die Befragten,

– Schwierigkeit des Interviews für Interviewenden,

– Anwesenheit Dritter (bei Face-to-Face-Interviews),

– Motivation/Aufmerksamkeit der oder des Befragten,

– Einschätzung des Themas der Befragung generell (Interessantheit, Relevanz),

– Probleme einzelner Fragen.

Der Bericht des Interviewers bzw. der Interviewerin erfolgt entweder schriftlich in Form eines sogenannten Erfahrungsberichts bzw. Pretest-Reports (zumeist über jedes durchgeführte Interview) oder mündlich in Einzel- oder gemeinsamen Sitzungen, im Rahmen derer alle beteiligten Interviewerinnen und Interviewer über ihre Interviewerfahrungen berichten. Standardisierte Bewertungsverfahren des sogenannten *Behaviour Coding* erlauben eine qualitative Bewertung des Interviewer- und Befragtenverhaltens. Verfahren des Feldpretests lassen sich mit experimentellen Bedingungsvariationen (z. B. unterschiedliche Frageformen) verbinden (Split-Ballot-Verfahren; Fowler 2004) und im Rahmen mehr oder weniger komplexer experimenteller Designs realisieren (Tourangeau 2004).

9 Zusammenfassung

– Fragen und Fragebögen stellen die zentralen Instrumente zur Erhebung von Selbstauskünften dar. Diese Selbstauskünfte dienen der bzw. dem Forschenden als Indikatoren in Operationalisierungen von Konstrukten.

– Bei ihrer sprachlichen Umsetzung sollten die möglichen Wirkungen des Fragetexts und der Antwortvorgaben auf das Antwortverhalten der Befragten berücksichtigt, d. h. die möglichen Reaktionen der oder des Befragten beim Entwurf antizipiert werden. Antworten sind Ergebnisse mentaler und emotionaler Prozesse, die zum größten Teil von der bzw. dem Forschenden unbeobachtet im oder in der Befragten ablaufen und daher selbst nur den Status theoretischer Konstrukte haben können.

– Es sollte soweit wie möglich sichergestellt werden, dass alle Prozesskomponenten des Antwortverhaltens durchlaufen werden, die zur Beantwortung der Frage wichtig sind (Abwesenheit von Satisficing). Zu diesen Komponenten zählt maßgeblich der Prozess des Verstehens, der zu einem Frageverständnis führen sollte, das sich so weit wie möglich mit dem Verständnis des oder der Forschenden deckt.

– Es sollte so weit wie möglich sichergestellt werden, dass das Leistungsvermögen der Befragten durch die Fragenkonstruktion nicht überfordert wird (z. B. durch zu komplizierte und umständliche Formulierungen, durch zu komplizierte Berechnungen, die Abfrage von Ereignissen, die in der Vergangenheit zu weit zurückliegen, durch die Voraussetzung von Kenntnissen, die der oder die Befragte nicht haben kann, durch einen zu langen Fragebogen, bei dessen Bearbeitung die Konzentration leidet, etc.).

– Es sollte so weit wie möglich sichergestellt werden, dass die Motivation des Befragten zur Beantwortung der Frage gestärkt wird.

– Mängel von Fragen lassen sich mithilfe nicht empirischer wie empirischer Verfahren erkennen. Die Identifikation von Mängeln des Entwurfs von Fragen und Fragebögen ist die Voraussetzung für die weitere Optimierung der Fragequalität. Letztere ist von großer Bedeutung, weil Mängel im Entwurf von Fragen und Fragebögen einerseits den Messfehler und damit die Zuverlässigkeit von Fragen bzw. Items, andererseits auch deren Validität negativ beeinflussen können.

– Sowohl nicht empirische Verfahren der Evaluation von Fragen und Fragebögen, wie die Anwendung von Fragebewertungssystemen, als auch empirische Pretestverfahren können herangezogen werden, um Erhebungsinstrumente hinsichtlich ihrer Qualität zu optimieren.

10 Weiterführende Literatur

UNODC; UNECE (2011): UNODC-UNECE Manual on Victimization Surveys. URL: http://www.unodc.org/unodc/en/data-and-analysis/Manual-on-victim-surveys.html – Download vom 16. 12. 2014.

Skogan, Welsey G. (1986): Methodological Issues in the Study of Victimization. In: Fattah, Ezzat A. (Hg.): From Crime Policy to Victim Policy: Restoring the Justice System. Basingstoke: Palgrave Macmillan.

Cantor, David; Lynch, James P. (2000): Self-Report Surveys as Measures of Crime and Criminal Victimization. In: Duffee, David (Hg.): Criminal Justice 2000. Washington: National Institute of Justice, S. 85–138

Lynch, James P. (1993): The Effects of Survey Design on Reporting in Victimization Surveys – The United States Experience. In: Bilsky, Wolfgang; Pfeiffer, Christian und Wetzels, Peter (Hg.): Fear of Crime and Criminal Victimization. Stuttgart: Enke, S. 159–185.

11 Literatur

Averdijk, Margit; Elffers, Hank (2012): The Discrepancy Between Survey-Based Victim Accounts and Police Reports Revisited. In: International Journal of Victimology, 18, 2, S. 91–107.

Belli, Robert F.; Traugott, Michael W.; Young, Margaret und Mc Gonagle, Katherine A. (1999): Reducing vote overreporting in surveys. In: Public Opinion Quarterly, 63, S. 90–108.

Bernstein, Robert; Chadha, Anita und Montjoy, Robert (2001): Overreporting voting. In: Public Opinion Quarterly, 65, S. 22–44.

Bohrnstedt, George W. (2010): Measurement models for survey research. In: Marsden, Peter V.; Wright, James D. (Hg.): Handbook of survey research. Bingley: Emerald Book Publishers, S. 347–404

Chauduri, Arijit; Mukherjee, Rahul (1988): Randomized response: Theory and techniques. New York: Marcel Dekker.

Conrad, Frederick G.; Schober, Michael F. (2000): Clarifying question meaning in a household telephone survey. In: Public Opinion Quarterly, 64, S. 1–28.

Conrad, Frederick G.; Schober, Michael F. und Coiner, Tania (2007): Bringing features of human dialogue to web surveys. In: Applied Cognitive Psychology, 21, S. 165–187.

Crowne, Douglas; Marlowe, David (1964): The approval motive. New York: John Wiley.

DeMaio, Theresa J. (1984): Social desirability and survey measurement. In: Turner, Charles; Martin, Elizabeth (Hg.): Surveying subjective phenomena. New York: Russell Sage, S. 257–282.

Edwards, Allen L. (1957): The social desirability variable in personality assessment and research. New York: Dryden.

Faulbaum, Frank (2014): Total survey error. In: Blasius, Jörg; Baur, Nina (Hg.): Handbuch der Empirischen Sozialforschung. Wiesbaden: Springer VS, S. 439–453.

Faulbaum, Frank; Prüfer, Peter und Rexroth, Margrit (2009): Was ist eine gute Frage? Wiesbaden: Springer VS.

Fowler Jr., Floyd J. (2004): The case for more split-sample experiments in developing survey instruments. In: Presser, Stanley; Couper, Mick P.; Lessler, Judith T.; Martin, Elizabeth; Martin, Jean; Rothgeb, Jennifer M. und Singer, Eleanor (Hg.): Methods for testing and evaluating survey questionnaires. Hoboken, NJ: John Wiley, S. 173–188.

Guilford, Joy P. (1954): Psychometric methods. New York: McGraw-Hill.

Guzy, Nathalie; Leitgöb, Heinz (2014): Mode effects in online and telephone based victimisation surveys. In: International Review of Victimology, Published online before print September 9, 2014, DOI: 10.1177/0269758014547995.

Hofmans, Joeri; Theuns, Peter; Baekelandt, Sven; Mairesse, Olivier; Schillewaert, Niels und Cools, Walentina (2007): Bias and changes in perceived intensity of verbal qualifiers effected by scale orientation. In: Survey Research Methods, 1, S. 97–108.

Jäckle, Annette (2009): Dependent interviewing: A framework and application to current research. In: Lynn, Peter (Hg.): Methodology of longitudinal surveys. Chichester: John Wiley, S. 93–111.

Jenkins, Cleo R.; Dillman, Don A. (1997): Towards a theory of self-administered questionnaires. In: Lyberg, Lars E.; Biemer, Paul B.; Collins, Martin; de Leeuw, Edith D.; Dippo, Cathryn; Schwarz, Norbert und Trewin, Dennis (Hg.): Survey Measurement and Process Quality. New York: John Wiley, S. 165–196.

Jobe, Jared; Pratt, William F.; Tourangeau, Roger; Baldwin, Alison K. und Rasinski, Kenneth A. (1997): Effects of interview mode on sensitive questions in a fertility survey. In: Lyberg, Lars E.; Biemer, Paul B.; Collins, Martin; de Leeuw, Edith D.; Dippo, Cathryn; Schwarz, Norbert und Trewin, Dennis (Hg.): Survey Measurement and Process Quality. New York: John Wiley, S. 311–329.

Krebs, Dagmar (1987): Soziale Empfindungen. Frankfurt/M.: Campus.

Kristof, Walter (1966): Das Cliffsche Gesetz im Deutschen. In: Psychologische Forschung, 29, S. 22–31.

Krosnick, Jon A. (1991): Response strategies for coping with the cognitive demands of attitude measures in surveys. In: Applied Cognitive Psychology, 5, S. 213–236.

Leary, Mark R. (2007): Motivational and emotional aspects of the self. In: Annual Review of Psychology, 58, S. 317–344.

Likert, Rensis (1932): A technique for the measurement of attitudes. In: Archives for Psychology, 22, S. 1–55.

Lord, Frederic M.; Novick, Melvin R. (1968): Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lynch, James P. (1993): The effects of survey design on reporting in victimization surveys – The United States experience. In: Bilsky, Wolfgang; Pfeiffer, Christian und Wetzels, Peter (Hg.): Fear of Crime and Criminal Victimization. Stuttgart: Enke, S. 159–185.

Lynn, Peter; Jäckle, Annette; Jenkins, Stephen P. und Sala, Emanuela (2006): The effects of dependent interviewing on responses to questions on income sources. In: Journal of Official Statistics, 22, S. 357–384.

Miller, Kristen; Chepp, Valerie; Willson, Stephanie und Padilla, Jose Luis (2014): Cognitive interviewing methodology. Hoboken, NJ: John Wiley.

Morris, Charles W. (1946): Signs, language and behavior. New York: Prentice-Hall.

Neter, John; Waksberg, Joseph (1964): A study of response errors in expenditures data from household interviews. In: Journal of the American Statistical Association, 59, S. 17–55.

Noel, Harmoni (2013): Conducting cognitive interviews over the phone: Benefits and challenges. In: Proceedings of the 68th Annual Conference of the American Association of Public Opinion Research, Boston, S. 4466–4478.

Orth, Bernhard (1974): Einführung in die Theorie des Messens. Stuttgart: Kohlhammer.

Paulhus, Delroy L. (2002): Socially desirable responding: The evolution of a construct. In: Braun, Henry T.; Jackson, Douglas N. und Wiley, David E. (Hg.): The role of constructs in psychological and educational measurement. Mahwah, NJ: Erlbaum, S. 49–69.

Peytchev, Andy; Conrad, Frederick G.; Couper, Mick P. und Tourangeau, Roger (2010): Increasing respondents' use of definitions in web surveys. In: Journal of Official Statistics, 26, S. 633–650.

Peytchev, Andy; Peytcheva, E. und Groves, R. M. (2010): Measurement error, unit nonresponse, and self-reports of abortion experiences. In: Public Opinion Quarterly, 74, S. 319–327.

Prüfer, Peter; Rexroth, Margrit (2005): Kognitive Interviews (= ZUMA-How-to-Reihe, Nr. 15). Mannheim: ZUMA.

Prüfer, Peter; Stiegler, Angelika (2002): Die Durchführung standardisierter Interviews: Ein Leitfaden(= ZUMA-How-to-Reihe Nr. 11). Mannheim: ZUMA.

Redline, Cleo (2013): Clarifying categorical concepts in a web survey. In: Public Opinion Quarterly, 77, S. 89–105.

Reeve, Bryce B.; Mâsse, Louise C. (2004): Item response theory modelling for questionnaire Evaluation. In: Presser, Stanley; Rothgeb, Jennifer M.; Couper, Mick P.; Lessler, Judith T.; Martin, Elizabeth; Martin, Jean und Singer, Eleanor (Hg.): Methods for testing and evaluating survey questionnaires. Hoboken, NJ: John Wiley, S. 247–273.

Reinecke, Jost (2014): Strukturgleichungsmodelle in den Sozialwissenschaften, 2. Aufl. Berlin: De Gruyter Oldenbourg.

Rohrmann, Bernd (1978): Empirische Studien zur Entwicklung von Antwortskalen für die psychologische Forschung. In: Zeitschrift für Sozialpsychologie, 9, S. 222–245.

Ross, Brian H.; Murphy, Gregory L. (1999): Food for Thought: Cross-classification and category organization in a complex real-world domain. In: Cognitive Psychology, 38, S. 495–553.

Saris, Willem E.; Gallhofer, Irmtraud N. (2014): Design, evaluation, and analysis of questionnaires for survey research. 2. Aufl. New York: John Wiley.

Schaeffer, Nora C. (1991): Hardly ever or constantly? Group comparisons using vague quantifiers. In: Public Opinion Quarterly, 55, S. 395–423.
Schober, Michael F.; Conrad, Frederick G. (1997): Does conversational interviewing reduce survey measurement error? In: Public Opinion Quarterly, 61, S. 576–602.
Schober, Michael F.; Conrad, Frederick G. und Fricker, Scott S. (2004): Misunderstanding standardized language in research interviews. In: Applied Cognitive Psychology, 18, S. 169–188.
Schuman, Howard; Presser, Stanley (1981): Questions and answers in attitude surveys: Experiments in question form, wording and context. New York: Academic Press.
Schwind, Hans D.; Fetchenhauer, Detlef; Ahlborn, Wilfried und Weiß, Rüdiger (2001): Kriminalitätsphänomene im Langzeitvergleich am Beispiel einer deutschen Großstadt. Bochum 1975–1986–1998. Neuwied: Luchterhand.
Smith, Tom W. (2003): Developing comparable questions in cross-national surveys. In: Harkness, Janet A.; van de Vijver, Fons J. R. und Mohler, Peter P. (Hg.) (2003): Cross-cultural survey methods. Hoboken, NJ: John Wiley, S. 69–91.
Smith, Tom W. (2004): Developing and evaluating cross-national survey instruments. In: Presser, Stanley; Rothgeb, Jennifer M.; Couper, Mick P.; Lessler, Judith T.; Martin, Elizabeth; Martin, Jean und Singer, Eleanor (Hg.): Methods for testing and evaluating survey questionnaires. Hoboken, NJ: John Wiley, S. 431–452.
Suessbrick, Anna; Schober, Michael F. und Conrad, Frederick G. (2000): Different respondents interpret ordinary questions quite differently? In: Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association, S. 907–912.
Suppes, Patrick; Zinnes, Joseph L. (1963): Basic measurement theory. In: Luce, R. Duncan; Bush, Robert R. und Galanter, Eugene (Hg.): Handbook of mathematical psychology I. New York: John Wiley, S. 1–76.
Tourangeau, Roger (1984): Cognitive science survey methods: a cognitive perspective. In: Jabine, Thomas; Straf, Miron L.; Tanur, Judith und Tourangeau, Roger (Hg.): Cognitive aspects of survey methodology: Building a bridge between disciplines. Washington DC: National Academy Press, S. 73–100.
Tourangeau, Roger (1987): Attitude measurement: A cognitive perspective. In: Hippler, Hans-J.; Schwarz, Norbert und Sudman, Seymour (Hg.): Social information processing and survey methodology. New York: Springer, S. 149–162.
Tourangeau, Roger; Rasinski, Kenneth; Jobe, Jared B. und Smith, Tom W. (1997): Sources of error in a survey of sexual behavior. In: Journal of Official Statistics, 13, S. 341–365.

Tourangeau, Roger; Rips, Lance J. und Rasinski, Kenneth (2000): The psychology of survey response. Cambridge, MA: Cambridge University Press.

Tourangeau, Roger (2004): Experimental design considerations for testing and evaluating questionnaires. In: Presser, Stanley; Rothgeb, Jennifer M.; Couper, Mick P.; Lessler, Judith T.; Martin, Elizabeth; Martin, Jean und Singer, Eleanor (Hg.): Methods for testing and evaluating survey questionnaires. Hoboken, NJ: John Wiley, S. 209–224.

Tourangeau, R.; Conrad, Frederick G.; Arens, Zachary; Fricker, Scott; Lee, Sunghee und Smith, Elisha (2006): Everyday concepts and classification errors: Judgments of Disability and residence. In: Journal of Official Statistics, 22, S. 385–418.

Tourangeau, Roger und Yan, Ting (2007): Sensitive questions in surveys. In: Psychological Bulletin, 133, S. 859–883.

Tourangeau, Roger; Conrad, Frederick G.; Couper, Mick P. und Ye, Cong (2014): The effects of providing examples in survey questions. In: Public Opinion Quarterly, 78, S. 100–125.

Wegener, Bernd; Faulbaum, Frank und Maag, Gisela (1982): Die Wirkung adverbialer Antwortvorgaben. In: Psychologische Beiträge, 24, S. 343–345.

Willis, Gordon B. (2005): Cognitive interviewing: A tool for improving questionnaire design. London: Sage.

Willis, Gordon B.; Lessler, Judith T. (1999): Question Appraisal System. Research Triangle Institute.

Zeller, Richard A.; Carmines, Edward G. (1980): Measurement in the social sciences. Cambridge: Cambridge University Press.

Soziale Erwünschtheit in Viktimisierungsbefragungen

Berenike Waubert de Puiseau, Adrian Hoffmann und Jochen Musch

Viktimisierungsbefragungen beruhen auf Selbstberichten. Deren Validität ist jedoch durch systematische Antwortverzerrungen insbesondere aufgrund der sozialen Erwünschtheit bestimmter Selbstauskünfte bedroht; Unter- und Überschätzungen der Prävalenz sensibler Merkmale sind mögliche Folgen. Im vorliegenden Kapitel werden zunächst potenzielle dadurch bedingte Schwächen der Belastbarkeit der Ergebnisse von Viktimisierungsstudien diskutiert. Anschließend wird das Konzept der sozialen Erwünschtheit theoretisch und in Bezug auf Viktimisierungsstudien näher erläutert. Danach werden Methoden zur Messung und zur Kontrolle sozialer Erwünschtheit besprochen. Existierende Viktimisierungsstudien, die solche Methoden bereits angewendet haben, werden exemplarisch vorgestellt.

1 Viktimisierungsstudien und soziale Erwünschtheit

Amtliche Statistiken zu Straftaten beleuchten, per Definition, immer nur das Hellfeld. In Opferwerdungsbefragungen werden möglichst repräsentative Stichproben hinsichtlich erlebter Straftaten innerhalb eines bestimmten Zeitraums befragt. Solche Studien sind wichtige Instrumente zur Erfassung der Kriminalitätswirklichkeit. Bei Viktimisierungsstudien kann jedoch nicht davon ausgegangen werden, dass diese ein unverfälschtes Abbild der Wirklichkeit liefern (Birkel 2003; Görgen u. a. 2006; Heinz 2006; 2009; siehe auch die Beiträge von Wollmann, Kolmey und Oberwittler in Band 1). Um dies zu erreichen, müssten nicht nur alle Opfer alle Straftaten als solche wahrgenommen haben und korrekt erinnern; sie müssten darüber hinaus auch ausnahmslos die Bereitschaft besitzen, erlebte Taten im Rahmen der Befragung zu berichten (Cook u. a. 2011; Koss 1993; Skogan 1975) und sie dürften dabei keine Taten hinzuerfinden. Systematische Antwortverzerrungen aufgrund sozial erwünschten Antwortverhaltens der Opfer stellen deshalb stets eine mögliche Bedrohung der Validität von Viktimisierungsstudien dar.

1.1 Die Methodik von Viktimisierungsstudien

Groß angelegte Viktimisierungsstudien gab es bereits in den 1970er Jahren in den USA. Seit nunmehr einigen Jahren werden Viktimisierungsbefragungen

auch in Deutschland durchgeführt (Heinz 2009, Kury 2010; siehe auch die Beiträge von van Dijk, Mischkowitz und Obergfell-Fuchs in Band 1). Diese Studien haben gemeinsam, dass sie auf Selbstberichten beruhen und in der Regel ein direktes Frageformat verwenden. Das interessierende Konstrukt wird dabei explizit abgefragt, wie Beispiel 1 verdeutlicht:

> Hat Sie schon einmal jemand mit Gewalt oder unter Androhung von Gewalt gegen Ihren Willen zum Beischlaf oder zu beischlafähnlichen Handlungen gezwungen oder versucht das zu tun? (Wetzels/Pfeiffer 1995, 2)

Auf besonders sensible Themen wird teilweise Rücksicht genommen, indem Fragen vorsichtig eingeleitet werden, wie Beispiel 2 zeigt:

> Im Folgenden stellen wir Ihnen eine ziemlich persönliche Frage. Es kommt manchmal vor, dass Menschen andere Menschen aus sexuellen Motiven unangenehm berühren, anfassen oder angreifen. Solche Vorfälle ereignen sich zu Hause oder auch in der Öffentlichkeit, zum Beispiel in einer Kneipe, auf der Straße, in der Schule, im öffentlichen Nahverkehr, in Kinos, am Strand oder am Arbeitsplatz. Ist Ihnen in den letzten fünf Jahren so etwas zugestoßen? Bitte nehmen Sie sich Zeit, über diese Frage nachzudenken. (von den Autoren übersetzte Frage Q80 aus dem ICVS bzw. EU ICS 2005)

Das Antwortformat ist in der Regel geschlossen, die teilnehmenden Personen beantworten die Fragen also entweder anhand der Antwortoptionen „Ja“, „Nein“ und „Weiß nicht“ oder mittels einer Skala, auf der sie die erlebte Häufigkeit der erfragten Straftaten abtragen (van Dijk u. a. 2007). Die Erhebung der Daten erfolgt in Viktimisierungsstudien vorwiegend anhand von (computer-assistierten) Telefon- oder persönlichen Interviews (CATI bzw. CAPI). Beim International Crime Victims Survey (ICVS) 2004/2005 wurden persönliche Interviews lediglich dort eingesetzt, wo Telefone nicht flächendeckend vorhanden waren (van Dijk u. a. 2007). Schriftliche Befragungsformate werden auch, aber seltener verwendet (Block 1993; Koss 1993). In einigen Studien wurde ein persönliches oder telefonisches Interview durch einen Fragebogen ergänzt, der entweder zu einem vereinbarten Zeitpunkt wieder abgeholt oder mittels eines bereits frankierten Briefumschlags zurückgesendet wurde (z. B. Görgen u. a. 2006; Ohlemacher u. a. 1997). Die Verwendung einer zusätzlichen Erhebungsmethode wurde in diesem Fall damit begründet, dass man sich bei Fragebögen eine größere Offenheit der teilnehmenden Personen versprach (Birkel 2003; Görgen u. a. 2006). Mangels objektiver Vergleichszahlen und aufgrund des direkten Frageformats kann jedoch bei keiner der beschriebenen Viktimisierungsstudien ermittelt werden, ob und in welchem Ausmaß systematische Antwortverzerrungen vorliegen.

Um die Validität von Selbstberichten zu überprüfen, wurden im Rahmen einiger kleinerer, eher methodisch orientierter Studien neben (potenziellen) Opfern auch (potenzielle) Täterinnen bzw. Täter befragt. Besonders in Befragungen zu Gewalt in Partnerschaften zeigten sich dabei mitunter große

Diskrepanzen zwischen den Angaben der Partnerinnen und Partner. Es fanden sich Hinweise darauf, dass insbesondere bei Selbstberichten von Opfern systematische Antwortverzerrungen zu einer Unterschätzung der tatsächlichen Prävalenz oder der Schwere der Gewalt geführt haben könnten, während selbstberichtete Täterschaft weniger betroffen war. Heckert und Gondolf (2000) befragten 840 Paare, bei denen die männlichen Partner aufgrund einer Verurteilung wegen häuslicher Gewalt an einem Therapieprogramm teilnahmen. Frauen verschwiegen dabei ihre Opfererfahrungen häufiger als die Männer ihre Straftaten (29 vs. 19 %). In einer weiteren Befragung von 97 Paaren, die eine Klinik für Paartherapie aufgesucht hatten, machte ca. ein Drittel diskrepante Angaben über verübte Gewalt. Das durch das Opfer berichtete Ausmaß der Gewalt lag bei etwa 30–40 % dieser Paare unter dem vom Täter bzw. der Täterin berichteten Ausmaß (Langhinrichsen-Rohling/Vivian 1994). In einer Studie aus den Niederlanden wurde die Diskrepanz zwischen einer Viktimisierungsstudie und der polizeilichen Statistik quantifiziert und genauer untersucht. Anlass war unter anderem die Befürchtung, die untersuchten Maße könnten durch soziale Erwünschtheit verzerrt werden. Für 18 % aller Studienteilnehmerinnen ergab sich dabei eine Unvereinbarkeit von Viktimisierungsbericht und Statistik (Averdijk/Elffers 2012). Knapp die Hälfte der bei der Polizei registrierten Viktimisierungen wurden in der Studie nicht berichtet (48 %); umgekehrt war mehr als die Hälfte der in der Studie berichteten Viktimisierungen der Polizei nicht bekannt (65 %).

Während Verzerrungen infolge von Erinnerungsfehlern relativ häufig als systematische Fehlerquelle in Viktimisierungsstudien identifiziert werden (Averdijk/Elffers 2012), wird die aus sozial erwünschtem Antwortverhalten resultierende Fehlervarianz seltener diskutiert und oft gar nicht erwähnt (z. B. Cantor/Lynch 2000; Schneider 1981). Gleichwohl wurden durch soziale Erwünschtheit bedingte Messprobleme in polizeilichen Statistiken und Viktimisierungsstudien bereits vor mehreren Jahrzehnten (Skogan 1975) diskutiert – und schon damals hatte man die „angemessene“ Bearbeitung des Fragebogens als Fehlerquelle identifiziert („responded to survey appropriately“, Skogan 1975, 21). Das damit angesprochene Konzept der sozialen Erwünschtheit wird deshalb im Folgenden zunächst theoretisch aufgearbeitet und anschließend im Kontext von Viktimisierungsstudien genauer erläutert.

1.2 Definition der sozialen Erwünschtheit

Als sozial erwünschtes Antwortverhalten wird die Tendenz verstanden, sich selbst in einem möglichst positiven Licht darzustellen und vorhandenen sozialen Normen zumindest im Antwortverhalten gerecht zu werden, um so einer möglichen Missbilligung durch Dritte vorzubeugen (Borkenau/Amelang 1986; Edwards 1957; Paulhus 1991; 2002).

Durch sozial erwünschtes Antwortverhalten kann die Prävalenz der sozialen Norm entsprechender Verhaltensweisen überschätzt und die Häufigkeit mit der Norm in Konflikt stehender Verhaltensweisen unterschätzt werden. Soziale Erwünschtheit wird so zu einer systematischen Fehlerquelle, welche die Validität aller auf Selbstberichten beruhenden Untersuchungen bedroht. Verzerrungen infolge sozialer Erwünschtheit betreffen dabei nicht nur Persönlichkeitsfragebögen (Borkenau/Amelang 1986; Crowne/Marlowe 1960; Mummendey 1981; Paulhus 1991), sondern auch eine Vielzahl epidemiologisch angelegter Studien zur Häufigkeit sensibler Merkmale und Einstellungen, beispielsweise zur Häufigkeit von Drogenmissbrauch oder Schwangerschaftsabbrüchen (Tourangeau/Yan 2007). Dabei sind sozial erwünschte Antworten umso eher zu erwarten, je sensibler die erfragten Themen sind (Barnett 1998), was im Rahmen von Rational-Choice-Modellen[1] zur Beantwortung sensibler Fragen gut erklärt werden kann (Stocké 2004). Entscheidend für die Sensibilität sind dabei nach Tourangeau und Yan (2007) die „Aufdringlichkeit" einer Frage, die aus einer Antwort resultierende Bedrohung sowie die mangelnde soziale Akzeptanz möglicher Antworten. Neben einer möglichen Verzerrung infolge sozialer Erwünschtheit kommt es bei sensiblen Fragen auch vermehrt zu Abbrüchen bei der Bearbeitung und zu einer höheren Zahl fehlender Antworten (Tourangeau/Yan 2007).

Hinsichtlich der Adressatin bzw. des Adressaten der Tendenz, sich in einer positiven Weise darzustellen, wird von manchen Autorinnen und Autoren (z. B. Musch u. a. 2012; Paulhus 2002) zwischen Selbst- und Fremdtäuschung als Teilkomponenten sozialer Erwünschtheit unterschieden. Empirische Untersuchungen zu dieser Unterscheidung liegen im Bereich von Viktimisierungsstudien bislang allerdings nicht vor, weshalb sie im Folgenden nicht weiter vertieft wird.

1.3 Systematische Antwortverzerrungen in Viktimisierungsstudien

Da soziale Erwünschtheit überwiegend bei Selbstberichten relevant ist, werden im Folgenden nur solche Studien vorgestellt, in denen potenzielle Opfer direkt befragt wurden.[2] Untersucht werden zum einen mögliche Gründe für

---

1 Rational-Choice-Modelle gehen davon aus, dass Menschen sich rational verhalten mit dem Ziel, ihren Nutzen zu maximieren. Im Kontext sozial erwünschten Antwortverhaltens bedeutet dies, dass Menschen durch ihr Antwortverhalten die Wahrscheinlichkeit negativer Konsequenzen minimieren und die positiver Reaktionen maximieren (Stocké 2004).

2 Studien, die auf der Polizeilichen Kriminalstatistik (PKS) bzw. anderen amtlichen Statistiken, auf Aktenanalysen, auf Interviews von Expertinnen und Experten oder auf Stichproben aus bekannten Opfern beruhen, werden nicht berücksichtigt.

sozial erwünschtes Antwortverhalten in Viktimisierungsstudien. Zum anderen werden Methoden beleuchtet, mithilfe derer soziale Erwünschtheit in Opferwerdungsbefragungen begegnet werden kann. Dabei wird davon ausgegangen, dass Opfer sich ihres Opferstatus bewusst sind. Aufgrund der geringen Anzahl deutschsprachiger Studien zur sozialen Erwünschtheit in Viktimisierungsbefragungen werden zusätzlich englischsprachige Studien einbezogen.

Zur Beantwortung der Frage, in welche Richtung sich mögliche systematische Verzerrungen durch soziale Erwünschtheit auswirken, muss zunächst die soziale Erwünschtheit der Opferrolle untersucht werden. Einerseits kann es für das Opfer sozial erwünscht sein, die erlebte Viktimisierung zu berichten, um beispielsweise von Opferhilfeprogrammen profitieren zu können (Sugarman/Hotaling 1997); andererseits kann die Opferrolle aber auch als stigmatisierend erlebt und deshalb verschwiegen werden. Der erste Fall würde zu einer Überschätzung, der zweite zu einer Unterschätzung der Opferzahl führen.

Insbesondere die positivistische Viktimologie schrieb dem Opfer mindestens eine Teilschuld an den erlebten Straftaten zu (Kury 2010; Schneider 1979). Obgleich nach der Abkehr von der positivistischen Viktimologie das Opfer weniger negativ dargestellt wurde, wird in jüngster Zeit unter Jugendlichen der Begriff „Du Opfer“ häufig als Beleidigung verwendet. Dies suggeriert, dass Opfer schwach und hilfsbedürftig seien (Kilchling 2010).

Insgesamt lässt sich festhalten, dass sozial erwünschtes Antworten sowohl zu einer Unter- als auch zu einer Überschätzung von Viktimisierungsraten führen kann, auch wenn allgemeingültige Aussagen über die soziale Erwünschtheit von Viktimisierung kaum möglich sind (Kilchling 2010). Delikt- und möglicherweise auch opferspezifische Eigenschaften können sich nicht nur auf das Anzeigeverhalten (Averdijk/Elffers 2012; Görgen 2009; Heinz 2009), sondern auch auf Antworten in Viktimisierungsstudien auswirken. Leider fehlt es jedoch an Studien, welche die soziale Erwünschtheit der Opferrolle in Abhängigkeit spezifischer Straftaten gezielt ermitteln. Dies erschwert die verlässliche Vorhersage möglicher Verzerrungen durch soziale Erwünschtheit. Lediglich zu Straftaten gegen die sexuelle Selbstbestimmung sowie Delikten im sozialen Nahraum herrscht Konsens in der Literatur darüber, dass die Viktimisierung mit einem Stigma assoziiert und hochgradig sensibel ist (z. B. Cook u. a. 2011; Greve u. a. 1994; Koss 1993; 1996). Folglich werden die größten Dunkelfelder bei Nahraumgewalttaten und Vergewaltigungen vermutet (Koss 1993, 1996, Skogan 1975). Als ein mögliches Motiv, die eigene Opferrolle zu verschweigen oder das Ausmaß der Viktimisierung zu minimieren, wird hier die Angst vor Konflikten mit oder der Verlust von nahestehenden Personen genannt (z. B. Heckert/Gondolf 2000). Wegen ihrer offensichtlichen Sensibilität liegen zu diesen Themen empirische Studien vor, die sich mit sozial erwünschten Antwortverzerrungen beschäftigt haben. Diese wer-

den zusammen mit verschiedenen Möglichkeiten, der Problematik sozial erwünschten Antwortens zu begegnen, in den folgenden Abschnitten vorgestellt.

2 Methoden zur Erfassung sozialer Erwünschtheit

Zur Erfassung und Kontrolle sozialer Erwünschtheit sind grundsätzlich drei Ansätze unterscheidbar. Dazu gehört (1) die Identifikation von Konstrukten, deren Messung durch den Einfluss sozialer Erwünschtheit verzerrt wird, um ungeeignete Items oder Skalen notfalls auszuschließen. Im Falle sexueller Viktimisierung kann es dazu beispielsweise hilfreich sein, solche Fragestellungen zu wählen, deren Beantwortung so wenig wie möglich durch soziale Erwünschtheit verzerrt wird. Zu nennen ist (2) die Identifikation von Personen, die unter dem Einfluss sozialer Erwünschtheit stehen, um diese notfalls ebenfalls aus der Stichprobe auszuschließen. In diesem Sinne wurde beispielsweise von Lenhard u. a. (2005) die deutsche Fassung des „Balanced Inventory of Desirable Responding“ (Paulhus 1991) eingesetzt, um in einer Befragung der Mütter von Kindern mit Down-Syndrom diejenigen Befragten von der Auswertung auszuschließen, die ungewöhnlich hohe soziale Erwünschtheitswerte zeigten. Schließlich können (3) direkte Messungen der Tendenz, sozial erwünscht zu antworten, auch mithilfe geeigneter Skalen vorgenommen werden; anschließend kann man versuchen, diese als Grundlage der statistischen Bereinigung von durch soziale Erwünschtheit verzerrten Messwerten zu verwenden, beispielsweise mittels regressionsanalytischer Techniken. (z. B. Kerlinger/Pedhazur 1973; Tabachnick/Fidell 2012). Zu diesem Zweck infrage kommende Instrumente sind die *Marlowe-Crowne Social Desirability Scale* (MC-SDS/SDS-CM; Crowne/Marlowe 1960; deutsche Fassung von Grabitz-Gniech 1971, Lück/Timaeus 1969), das *Balanced Inventory of Desirable Responding* (BIDR; Paulhus 1991; deutsche Fassung von Musch u. a. 2002) und die Soziale-Erwünschtheits-Skala-17 (SES-17; Stöber 1999). Ob sich durch die Auspartialisierung so gemessener sozialer Erwünschtheitswerte tatsächlich validere Messungen erreichen lassen, ist allerdings keineswegs gesichert (Pauls/Crost 2004).

Empirische Befunde und theoretische Überlegungen legen nahe, dass soziale Erwünschtheit in Studien zu sexueller Viktimisierung eine besondere Rolle spielt. Bei deren Untersuchung werden deshalb häufig spezifische Fragebögen wie der *Sexual Experiences Survey* (Koss/Gidycz 1985; Koss/Oros 1982; Koss u. a. 2007) oder die *Conflict Tactics Scale* (Straus 1979) verwendet, in denen legale und illegale Verhaltensweisen beschrieben und dabei oft Kurzszenarien mit konkreten Verhaltensbeschreibungen verwendet werden, um Interpretationsspielräume zu verringern. Zu jeder Verhaltensweise soll auf einer Skala angegeben werden, wie häufig diese in einem bestimmten Zeitraum er-

lebt wurde. Das Frageformat ist direkt und geschlossen. Zusätzlich zur Skala über sexuelle Viktimisierung beantwortet jede Person einen Fragebogen zur Messung sozial erwünschten Antwortverhaltens. Über die Bestimmung der Korrelation mit den Angaben zur Viktimisierung soll der durch sozial erwünschtes Antworten bedingte Anteil der Varianz in den Selbstauskünften bestimmt werden. Weicht der Korrelationskoeffizient signifikant von null ab, wird vermutet, dass der Viktimisierungsselbstbericht durch sozial erwünschtes Antworten systematisch verzerrt wird.

Eine Metaanalyse von sieben Studien zu Gewalt in Paarbeziehungen ergab, dass die Tendenz, sozial erwünscht zu antworten, mit niedrigeren Angaben selbsterlebter Viktimisierung einherging (Sugarman/Hotaling 1997). In einer weiteren Studie zu Gewalt in der Partnerschaft wurde ebenfalls ein signifikanter negativer Zusammenhang zwischen sozial erwünschtem Antwortverhalten und der Wahrscheinlichkeit, sich selbst als Opfer physischer Aggression zu klassifizieren, festgestellt. Dies galt jedoch nur für die weibliche Substichprobe; die selbstberichtete Viktimisierung der Männer zeigte keine Zusammenhänge mit der Tendenz, sozial erwünscht zu antworten (Bell/Naugle 2007). In einer Stichprobe von 75 Frauen, die sich selbst als Opfer körperlicher und emotionaler Gewalt in einer Beziehung identifiziert hatten, wurde kein Zusammenhang zwischen dem Bericht über die Erlebnisse und zwei Maßen sozialer Erwünschtheit gefunden (Dutton/Hemphill 1992). Ebenfalls keinen Zusammenhang ermittelte eine Untersuchung selbstberichteter Viktimisierung von verheirateten Frauen und Männern in Paarbeziehungen, in der jedoch keine Paare, sondern nur Einzelpersonen untersucht wurden (Arias/Beach 1987).

Eine methodische Einschränkung erfahren Befunde dieser Art insofern, als ein Zusammenhang zwischen sozial erwünschtem Antwortverhalten und selbstberichteter Viktimisierung keine Rückschlüsse auf eine Kausalbeziehung dieser Variablen zulässt. Vielmehr gibt es verschiedene Gründe, weshalb Maße sozialer Erwünschtheit und Viktimisierungsangaben miteinander zusammenhängen könnten. So kann es sein, dass sozial erwünschtes Antwortverhalten tatsächlich zu einer Unterschätzung selbstberichteter Viktimisierung führt. Eine mögliche Alternativerklärung wäre jedoch, dass Personen, die hohe Werte auf Skalen sozialer Erwünschtheit erzielen, auch eher Verhaltensweisen an den Tag legen, die sie vor Viktimisierung schützen oder eine solche begünstigen. Denkbar wäre beispielsweise, dass eine Tendenz, sich an den Erwartungen anderer zu orientieren, die Wahrscheinlichkeit für soziale Konflikte reduziert. Anstelle korrelativer Untersuchungen sind insbesondere solche methodischen Ansätze von Interesse, die eine experimentelle Kontrolle sozial erwünschten Antwortverhaltens ermöglichen.

3 Methoden zur Kontrolle sozialer Erwünschtheit

Im folgenden Abschnitt werden Methoden vorgestellt, die über eine bloße Erfassung sozialer Erwünschtheit hinausgehen und auf eine Kontrolle dieses Einflussfaktors in Viktimisierungsstudien abzielen. Mit ihrer Hilfe sollen Selbstauskünfte ermittelt werden, die potenziell nicht oder weniger durch soziale Erwünschtheit verzerrt sind und folglich eine gesteigerte Validität aufweisen. Angelehnt an die Übersichtsarbeiten von Nederhof (1985) und Paulhus (1991) werden Methoden zur Kontrolle sozialer Erwünschtheit zunächst übergeordnet klassifiziert und kurz hinsichtlich ihrer Anwendbarkeit in Viktimisierungsbefragungen analysiert. Der Fokus liegt dabei auf indirekten Befragungstechniken, die eine nach unserer Meinung besonders vielversprechende Möglichkeit zur effektiven und effizienten sowie flexiblen Kontrolle sozial erwünschten Antwortverhaltens in Viktimisierungsstudien darstellen. Zu den verschiedenen Techniken werden, sofern vorhanden, jeweils einschlägige Viktimisierungsstudien zur Illustration vorgestellt.

3.1 Gestaltung der Befragungssituation

Bei Befragungen zu sensiblen Themen können verschiedene Aspekte der Befragungssituation so gewählt werden, dass ein Einfluss sozialer Erwünschtheit möglichst minimiert wird. Beispielsweise sind bezüglich der Wahl geeigneter Interviewerinnen und Interviewer ein professionelles Auftreten und eine höhere soziale Distanz zu den Befragten insgesamt mit einer erhöhten Validität der Ergebnisse in Verbindung gebracht worden (Nederhof 1985). Insbesondere bei sensiblen Themen wie sexueller Viktimisierung sollte ein möglicher Einfluss des Geschlechts der interviewenden Person beachtet werden. So wurde in eine Studie mit 3.125 Frauen und Männern gegenüber weiblichen Interviewern 1,27 Mal so häufig eine sexuelle Viktimisierung angegeben wie gegenüber einem männlichen Interviewer (Sorenson u. a. 1987).

Einige empirische Studien legen nahe, dass sich Befragungsmodi hinsichtlich ihrer Anonymität und somit ihrer Anfälligkeit für sozial erwünschtes Antworten unterscheiden (Kury 1994), wenngleich empirische Untersuchungen diesbezüglich ein uneinheitliches Bild zeichnen (Barnett 1998). Im Rahmen des National Crime Survey, der u. a. erlebte Sexualstraftaten erfasst, wurden Telefoninterviews mit persönlichen Interviews verglichen. Die in persönlichen Interviews ermittelten Prävalenzraten, u. a. von Tätlichkeiten im Nahbereich, waren höher als die in Telefoninterviews erfassten, jedoch waren die Unterschiede nur gering und überwiegend statistisch nicht signifikant (Turner 1984). Ähnliche Ergebnisse beobachtete Koss (1993) über verschiedene Stichproben hinweg bei der Untersuchung sexueller Viktimisierung. Anonyme schriftliche Befragungen ergaben allerdings besonders hohe Prävalenzra-

ten. An einer Stichprobe von 1.962 Personen aus der deutschen Allgemeinbevölkerung fand Kury (1994) Unterschiede im Antwortverhalten zwischen mündlichen und schriftlichen Befragungen – unabhängig davon ob sensible Viktimisierungsfragen zu Kontaktdelikten wie sexuellem Missbrauch gestellt oder allgemeine Persönlichkeitsfragebögen vorgegeben wurden. Mündlich befragte Personen antworteten eher im Sinne sozialer Erwünschtheit als schriftlich Befragte. Dies traf besonders auf Männer sowie Personen über 40 Jahren zu. In einer neueren Studie wurden Rücklaufquote und Datenqualität verschiedener Datenerhebungsverfahren untersucht. Dazu gehörten CATI, CAPI, das Computer-assistierte Onlineinterview (CAWI) und das Offlinefragebögen-assistierte Interview (PAPI; van Dijk u. a. 2010). Die Rücklaufquoten des CATI, also der Telefoninterviews, waren am besten, die des CAWI am schlechtesten. In einer Analyse der Daten des ICVS-2 zeigten sich Unterschiede zwischen Erhebungsmodi nur für die sensible Frage nach sexueller Viktimisierung (Guzy/Leitgöb 2015).

Problematisch ist, dass sich Erhebungsmethoden neben der sozialen Erwünschtheit in vielen anderen Dimensionen unterscheiden. So ist es bei persönlichen Interviews eher möglich, Unklarheiten zu beseitigen. CATIs wirken möglicherweise unpersönlich und stoßen bei den Befragten eventuell auf mangelnde Antwortbereitschaft. Auch stellt diese Art der Befragung höhere Ansprüche an das Arbeitsgedächtnis der interviewten Personen, da diese sich die Fragetexte merken müssen. Mittels schriftlich dargebotener Fragebögen kann dieses Problem umgangen werden, allerdings ist hier ausreichende Lesekompetenz erforderlich. Darüber hinaus spielen auch finanzielle und zeitliche Erwägungen bei der Wahl einer Erhebungsmethode eine wichtige Rolle (Kury 1994). Es wäre deshalb wünschenswert, nicht durch einen Wechsel des Befragungskontexts, sondern innerhalb des gleichen Befragungsmodus Möglichkeiten zu realisieren, durch eine Erhöhung der Anonymität und Vertraulichkeit sozial erwünschtes Antwortverhalten besser zu kontrollieren.

In jedem Fall ist es ratsam, der empfundenen Anonymität und Vertraulichkeit bei Befragungen zu sensiblen Delikten eine hohe Priorität einzuräumen, da diese beiden Variablen positiv mit der Validität der ermittelten Ergebnisse assoziiert zu sein scheinen. Folglich sollten Abfragen personenbezogener Daten auf das erforderliche Minimum reduziert und strenge Datenschutzrichtlinien gewissenhaft eingehalten werden (siehe den Beitrag von Hatt in diesem Band). Der Einfluss von Anonymität und Vertraulichkeit auf Viktimisierungsberichte wurde in einer groß angelegten Studie mit 11.195 Rekruten der *US Navy* untersucht, in der verschiedene Formen schriftlicher Befragungen zur selbstberichteten Prävalenz sexuellen, emotionalen und physischen Missbrauchs in der Kindheit miteinander verglichen wurden (Olson u. a. 2004). Die Fragebogenbedingungen unterschieden sich hinsichtlich ihrer Anonymität. Der erste Fragebogen war standardmäßig mit der Sozialversicherungs-

nummer zu versehen und für die Personalakte der Rekruten gedacht (nicht anonyme, nicht vertrauliche Bedingung). Der zweite Fragebogen diente wissenschaftlichen Zwecken. Den teilnehmenden Rekruten wurde mitgeteilt, dass die Vorgesetzten die Angaben in diesen Fragebögen nicht würden einsehen können. Zusätzlich wurde im zweiten Fragebogen manipuliert, ob die Rekruten instruiert wurden, ihre Sozialversicherungsnummer anzugeben (nicht anonyme, vertrauliche Bedingung) oder nicht (anonyme, vertrauliche Bedingung). Während in der nicht anonymen, nicht vertraulichen Bedingung nur fünf Prozent der Rekruten ihre Viktimisierung angaben, lag der Anteil bei der nicht anonymen, vertraulichen Bedingung mehr als sechs Mal (31 %) und bei der anonymen, vertraulichen Bedingung mehr als sieben Mal so hoch (36 %). Da sich die Frageformulierungen im ersten und zweiten Fragebogen deutlich unterschieden, obwohl Frageformulierungen in Studien zu Missbrauchserfahrungen einen erheblichen Einfluss auf ermittelte Prävalenzen haben können (z. B. Cook u. a. 2011; Koss 1993), sind die großen Unterschiede allerdings nicht zweifelsfrei auf soziale Erwünschtheit zurückzuführen. Lediglich der Unterschied zwischen der nicht anonymen, vertraulichen und der anonymen, vertraulichen Bedingung, eine Differenz von fünf Prozent, lässt sich eindeutig auf eine systematische Antwortverzerrung zurückführen. Das Ergebnis dieser Studie legt jedoch insgesamt gleichwohl nahe, dass sich hohe Standards bei Vertraulichkeit und Anonymität positiv auf die Validität der Befragungsergebnisse auswirken.

Schließlich wird empfohlen, den empfundenen Stress in der Befragungssituation so weit wie möglich zu reduzieren (beispielsweise durch die Wahl großzügiger Zeitvorgaben und eine Vermeidung von emotionaler Erregung und Ablenkung), da ein hohes Stressniveau zu einem erhöhten Anteil sozial erwünschter Antworten führen kann (Paulhus 1991). Gänzlich ungeeignet für eine Anwendung in Opferwerdungsbefragungen erscheinen folglich Methoden, die über einen erhöhten Druck auf die Befragten zu ehrlicheren Antworten führen sollen. Zu dieser Verfahrensklasse ist beispielsweise die Bogus-Pipeline-Technik (Jones/Sigall 1971) zu zählen. Dabei werden Befragte an eine als Lügendetektor ausgewiesene technische Apparatur angeschlossen. Wenngleich dieser Lügendetektor in Wirklichkeit keine Funktion erfüllt, führt seine Anwendung häufig zu ehrlicheren Antworten und damit valideren Ergebnissen als eine konventionelle Befragung (Roese/Jamieson 1993). Die ethische Vertretbarkeit von Bogus-Pipeline-Untersuchungen ist allerdings umstritten (Mummendey 1981). Bezüglich eines Einsatzes in Viktimisierungsstudien wiegt ein weiteres Gegenargument schwer: Ein massiver Druck, ehrlich zu antworten, könnte zu unfreiwilligen und schmerzhaften Selbsterkenntnissen führen und potenziell die psychologische Stabilität von Befragten gefährden (Aguinis/Handelsman 1997). Uns ist vermutlich aus diesen Gründen keine Viktimisierungsstudie bekannt, in der unter Anwendung der Bogus-Pipeline-Technik der Einfluss sozialer Erwünschtheit auf das Antwort-

verhalten kontrolliert oder gemessen werden sollte. Darüber hinaus wäre es aufgrund der Kosten der Face-to-face-Erhebung kaum möglich, unter Verwendung der Bogus-Pipeline-Technik eine ausreichend große Fallzahl zu erheben. Vielversprechende Ansätze zur Kontrolle sozialer Erwünschtheit zielen auf eine Optimierung der Gestaltung des Befragungsinstruments ab.

3.2 Gestaltung des Erhebungsinstruments

Bei der Konstruktion des einzusetzenden Messinstruments kann auf verschiedenen Wegen einem Einfluss sozialer Erwünschtheit auf die Befragungsergebnisse entgegengewirkt werden. Zunächst sollten Fragen so gewählt werden, dass ihre Beantwortung so weit wie möglich unabhängig vom Einfluss sozialer Erwünschtheit ist. Besonders in Bezug auf hochsensible Straftaten wie sexuellen Missbrauch und Vergewaltigung kann schon die Definition des Delikts einen erheblichen Einfluss auf die Höhe der ermittelten Prävalenzschätzungen und folglich auf die Validität der Ergebnisse nehmen (siehe Unterkapitel 3.1; Cook u. a. 2011; Hamby/Koss 2003). Die Verwendung stigmatisierender Begriffe kann darüber hinaus zu einer geringeren Bereitschaft zum Selbstbericht führen (Koss 1993). Die Wahl möglichst neutraler Fragen kann einerseits auf Basis rationaler Kriterien erfolgen. Andererseits erlauben spezielle statistische Verfahren wie die Faktorenanalyse die Identifikation von Fragen, welche die inhaltlich relevante Variable möglichst faktorrein – d. h. unabhängig vom Einfluss sozialer Erwünschtheit – messen (Paulhus 1991). In Viktimisierungsstudien sind jedoch sowohl der Auswahl als auch der Ausformulierung einzusetzender Fragen natürliche Grenzen gesetzt. Die Prävalenz eines spezifischen Delikts muss oft auf Basis einer einzelnen, deliktspezifischen Frage geschätzt werden, die sich meist an Gesetzestexten orientiert und somit wenig Gestaltungsspielraum bietet. Im gegebenen Rahmen ist es ratsam, der Gestaltung von Instruktionen, Fragen und begleitenden Texten möglichst hohe Aufmerksamkeit zukommen zu lassen und Formulierungen zu vermeiden, die einen Selbstbericht tatsächlich erlebter Opfererfahrungen unnötig erschweren. Schließlich ist eine Standardisierung von Frageformulierungen (und ggf. sogar ganzen Erhebungsinstrumenten) erstrebenswert, um eine bisher häufig nicht gegebene Vergleichbarkeit der Ergebnisse von Viktimisierungsbefragungen zukünftig zu ermöglichen (Cook u. a. 2011, Görgen 2009).

Neben einer gezielten Auswahl geeigneter Fragen können in Umfragen zu sensiblen Merkmalen spezielle Befragungstechniken zur Anwendung kommen, um die Bereitschaft zu ehrlichem Antwortverhalten aufseiten der Befragten zu steigern. Besonders geeignet für einen Einsatz in Viktimisierungsstudien erscheinen indirekte Befragungstechniken wie die Randomized-Response-Technik (Warner 1965), die im Folgenden näher dargestellt werden.

3.3 Indirekte Befragungstechniken

Das grundlegende Prinzip indirekter Befragungstechniken besteht darin, durch eine Zufallsverschlüsselung jegliche Verknüpfung zwischen den individuellen Antworten der Teilnehmerinnen und Teilnehmer und ihrem Merkmalsstatus in Bezug auf das untersuchte sensible Merkmal zu eliminieren. So sollen Befragungsteilnehmerinnen und -teilnehmer zu ehrlicheren Antworten motiviert werden. Während individuelle Antworten vertraulich bleiben, kann auf Stichprobenebene eine Schätzung für die Prävalenz des sensiblen Merkmals gewonnen werden, die durch den Einfluss sozial erwünschten Antwortverhaltens häufig weniger verzerrt ist. Die potenziell erhöhte Validität wird durch eine gegenüber konventionellen direkten Fragen reduzierte Stichprobeneffizienz erkauft, da die Zufallsverschlüsselung zusätzliche Varianz in den Messwerten erzeugt (Ulrich u. a. 2012). Eine reduzierte Effizienz wird dann – und nur dann – als akzeptabel erachtet, wenn der Einsatz indirekter Befragungstechniken eine höhere Validität der ermittelten Ergebnisse ermöglicht. Tatsächlich konnte in einer Metaanalyse gezeigt werden, dass indirekte Fragen im Mittel zu höheren Schätzungen für die Prävalenz sozial unerwünschter Merkmale führten als direkte Fragen (Lensvelt-Mulders u. a. 2005). Diese höheren Schätzungen werden nach dem „More is better"-Kriterium (Umesh/Peterson 1991) meist als valider angesehen, da sie vermutlich weniger durch den Einfluss sozial erwünschten Antwortverhaltens verzerrt sind. In derselben Metaanalyse lagen Schätzungen auf Basis indirekter Fragen außerdem insgesamt näher am wahren Wert, wenn dieser in der jeweiligen Stichprobe bekannt war. Diese Ergebnisse legen zusammenfassend nahe, dass indirekte Befragungstechniken einen geeigneten Ansatz zur Kontrolle sozialer Erwünschtheit in Umfragen darstellen.

Gerade in Viktimisierungsstudien ist eine Beschränkung auf kurze, effiziente Befragungsinstrumente zentral, um die (ggf. mehrfache) Erhebung großer Stichproben zu ermöglichen (Heinz 2006). Indirekte Befragungstechniken entsprechen diesem Kriterium, da ihr Einsatz für Umfrageteilnehmerinnen und -teilnehmer in der Regel mit keinem oder nur geringem Mehraufwand verbunden ist. Dabei bietet die Befragung großer, repräsentativer Stichproben optimale Rahmenbedingungen, um der verringerten Effizienz indirekter Befragungstechniken zu begegnen und zu ausreichend genauen Schätzern zu gelangen. Untersuchungen zur genauen Größe der erforderlichen Stichproben werden für verschiedene indirekte Befragungstechniken von Ulrich u. a. (2012) berichtet. In Deutschland werden für die Durchführung solcher Viktimisierungsbefragungen hinreichend große Stichproben in der Regel erreicht (Görgen 2009).

Der folgende Abschnitt führt anhand der *Randomized-Response*-Technik (RRT; Warner 1965) zunächst in das Themengebiet indirekter Befragungs-

techniken ein. Anschließend werden ausgewählte aktuelle Weiterentwicklungen der RRT und verwandte Verfahren diskutiert, die derzeit in der Umfrageforschung zu sensiblen Themen zur Anwendung kommen. Die beispielhafte Darstellung kann der großen Anzahl konkurrierender Modelle aus dieser Verfahrensklasse aus Platzgründen nicht gerecht werden; der bzw. die interessierte Lesende sei deshalb an dieser Stelle auf einschlägige Übersichtsarbeiten hingewiesen (z. B. Antonak/Livneh 1995; Chaudhuri/Christophides 2013; Fox/Tracy 1986; Tracy/Mangat 1996; Umesh/Peterson 1991). Im Folgenden werden jedoch exemplarisch Viktimisierungsstudien vorgestellt, die indirekte Fragetechniken zum Zwecke der genaueren Schätzung von Viktimisierungsraten angewendet haben.

3.3.1 Die Randomized-Response-Technik

Von Warner (1965) wurde als erste indirekte Befragungstechnik zur Kontrolle sozialer Erwünschtheit in Umfragen die Randomized-Response-Technik (RRT) vorgestellt. Im ursprünglichen *Related-Questions*-Modell (RQM) werden den Umfrageteilnehmerinnen und -teilnehmern gleichzeitig zwei sensible Aussagen präsentiert, von denen eine die Negation der anderen ist. Anhand eines Zufallsgenerators wird nun bestimmt, zu welcher der beiden Aussagen die Umfrageteilnehmerinnen und -teilnehmer Stellung nehmen sollen. Als Zufallsgeneratoren dienten in der Vergangenheit z. B. Würfel, Drehscheiben oder Lostrommeln. Alternativ kann auch eine Randomisierung auf der Basis eines persönlichen Merkmals vorgenommen werden, von dem die Verteilung in der Population bekannt ist. Dazu gehört z. B. der Geburtsmonat (z. B. Moshagen u. a. 2012; Ostapczuk/Musch 2011). So könnten die Instruktionen lauten, bei einem Geburtstag im November oder Dezember zur sensiblen Aussage A und bei einem Geburtstag in irgendeinem anderen Monat auf die Negation der sensiblen Aussage, Aussage B, zu antworten. *Abbildung 1* zeigt ein Beispiel einer möglichen Frage in einer Opferwerdungsbefragung zu einem potenziell hochsensiblen Merkmal (in Anlehnung an Wetzels/Pfeiffer 1995).

Abbildung 1:

Beispiel für eine Frage im RRT-Format in einer Opferwerdungsbefragung zum Thema Vergewaltigung

Unter „Vergewaltigung“ wird im Folgenden verstanden, wenn ein Täter sein Opfer mit Gewalt oder unter Androhung von Gewalt gegen den Willen des Opfers zum Beischlaf oder zu beischlafähnlichen Handlungen zwingt oder versucht, das zu tun.
Im Folgenden werden Ihnen hierzu zwei Aussagen präsentiert. Nehmen Sie bitte Stellung zu …
… Aussage A, wenn Sie *im November oder Dezember* geboren wurden.
… Aussage B, wenn Sie *in einem anderen Monat* geboren wurden.

Aussage A:	Ich wurde schon einmal vergewaltigt.
Aussage B:	Ich wurde noch nie vergewaltigt.

Antwort: [ ] Stimmt
[ ] Stimmt nicht

Sofern der Ausgang des Zufallsexperiments (der Geburtsmonat der Befragten) gegenüber der Umfrageleitung geheim gehalten wird, können aus den Antworten der Teilnehmerinnen und Teilnehmer keine Rückschlüsse über ihren wahren Merkmalsstatus gezogen werden: Eine „Stimmt“-Antwort kann – abhängig vom Ausgang des Zufallsexperiments – gleichermaßen von einer Merkmalsträgerin bzw. einem Merkmalsträger wie von einer Nicht-Merkmalsträgerin bzw. einem Nicht-Merkmalsträger stammen. Aus offiziellen Geburtsstatistiken kann jedoch die Randomisierungswahrscheinlichkeit p bestimmt werden, mit der Aussage A ausgewählt wird (im Beispiel etwa $p = ,158$; Moshagen u. a. 2012). Dies erlaubt die Aufstellung eines Gleichungssystems, das die beobachteten Antworthäufigkeiten als Funktion der bekannten Randomisierungswahrscheinlichkeit p (Geburtsmonat) und der unbekannten und zu schätzenden Prävalenz π des sensiblen Merkmals beschreibt. Unter Anwendung von Maximum-Likelihood-Verfahren kann so auf Gruppenebene eine Schätzung der Prävalenz sensibler Merkmale gewonnen werden. Aufgrund der garantierten Vertraulichkeit wird erwartet, dass diese Schätzung eine höhere Validität aufweist als eine Schätzung auf Basis einer konventionellen direkten Frage. Warner (1965) hat Formeln für die Schätzung der Prävalenz π und für den Standardfehler dieses Schätzers veröffentlicht. Alternativ bietet eine Reformulierung als multinomiales Verarbeitungsbaummodell, das kategoriale Daten mithilfe latenter Parameter beschreibt, eine verhältnismäßig einfache Möglichkeit für Prävalenzschätzungen und Hypothesentests (Batchelder 1998; Batchelder/Riefer 1999; Hu/Batchelder 1994; Moshagen u. a. 2012).

Seit Einführung der RRT durch Warner (1965) wurde eine Vielzahl weiterentwickelter Modelle vorgestellt, die beispielsweise eine verbesserte Effizienz (z. B. Boruch 1971; Dawes/Moore 1980; Mangat 1994), eine Berücksichtigung von Fragen mit mehr als zwei Antwortkategorien (z. B. Abul-Ela u. a. 1967; Liu/Chow 1976) oder eine Steigerung der Bereitschaft zu ehrlichen Antworten zum Ziel hatten (z. B. Greenberg u. a. 1969; Kuk 1990; Ostapczuk u. a. 2009). Besonderes Augenmerk soll im folgenden Abschnitt auf zwei Modelle gelegt werden, die ein mögliches Nichtbeachten der Instruktionen durch einen Teil der Befragten explizit in die Modellannahmen einbeziehen.

3.3.2 RRT-Modelle mit Verweigererdetektion

Obwohl RRT-Fragen die Vertraulichkeit individueller Antworten garantieren, haben Untersuchungen an Teilnehmerinnen und Teilnehmern mit bekanntem Merkmalsstatus gezeigt, dass Befragte sich nicht immer an die Instruktionen halten – vor allem wenn die abgefragten Inhalte besonders sensibel sind (z. B. Edgell u. a. 1992; Edgell u. a. 1982). Unehrliche Antworten können besonders in *Forced-Choice*-Varianten der RRT (Boruch 1971; Dawes/Moore 1980) beobachtet werden. Hier wird den Teilnehmerinnen und Teilnehmern nur eine einzelne sensible Aussage dargeboten. Ein Zufallsexperiment entscheidet dann, ob auf diese Aussage – unabhängig vom wahren Merkmalsstatus – mit einer Wahrscheinlichkeit von p mit „stimmt“ oder mit der Gegenwahrscheinlichkeit von $1-p$ ehrlich geantwortet werden soll (siehe *Abbildung 2*).

Abbildung 2:

Beispiel für eine mögliche RRT-Frage im Forced-Choice-Format in einer Opferwerdungsbefragung zum Thema Vergewaltigung

Im Folgenden wird Ihnen eine Aussage präsentiert. Falls Sie …

… *im November oder Dezember* geboren wurden, antworten Sie bitte mit „Stimmt“ (unabhängig von Ihrer wahren Antwort).

… *in einem anderen Monat* geboren wurden, antworten Sie bitte ehrlich.

Aussage:	Ich wurde schon einmal vergewaltigt.

Antwort: [ ] Stimmt
[ ] Stimmt nicht

Auch bei diesem Frageformat kann aus einer „Stimmt“-Antwort kein Rückschluss auf den wahren Merkmalsstatus der Befragten gezogen werden. Eine „Stimmt nicht“-Antwort bietet jedoch die Möglichkeit, eine Merkmalsträgerschaft sicher auszuschließen, da diese bei vollkommen ehrlichem Antwortver-

halten logisch nur von einer Nicht-Merkmalsträgerin bzw. einem Nicht-Merkmalsträger stammen kann. Folglich könnten Merkmalsträgerinnen und -träger den Drang verspüren, unehrlich mit „stimmt nicht“ zu antworten, um ihren wahren Status zu verbergen; ebenso könnten Nicht-Merkmalsträgerinnen und Nicht-Merkmalsträger statt einer ggf. durch den Zufallsgenerator geforderten „Stimmt“-Antwort die „sichere“ Antwortalternative wählen, um das Risiko einer fälschlichen Identifikation als Merkmalsträgerinnen bzw. -träger auszuschließen (Antonak/Livneh 1995). Um den Anteil unehrlicher Antworten auf RRT-Fragen quantifizierbar zu machen, stellten Clark und Desharnais (1998) das *Cheating Detection Model* (CDM) vor, das neben ehrlich antwortenden Merkmalsträgerinnen und -trägern (Anteil π) und ehrlich antwortenden Nicht-Merkmalsträgerinnen bzw. Nicht-Merkmalsträgern (Anteil β) eine dritte Klasse von Verweigerinnen und Verweigerern (Anteil γ) berücksichtigt, die ungeachtet der Instruktionen die Antwortoption „Stimmt nicht“ wählen. Zur Schätzung dieser drei Parameter müssen zwei unabhängige Stichproben mit unterschiedlichen Randomisierungswahrscheinlichkeiten $p1 \neq p2$ erhoben werden, wobei eine größere Differenz von $p1$ und $p2$ in einer höheren Effizienz des Modells resultiert (Clark/Desharnais 1998). Gleichungen für die Schätzung der Anteile π, β und γ sowie die Varianz der Schätzer können Clark/Desharnais (1998) entnommen werden; Beispiele für eine Umsetzung als multinomiales Verarbeitungsbaummodell finden sich u. a. in Ostapczuk/Musch (2011) und Ostapczuk u. a. (2011). In mehreren Studien wurde unter Verwendung dieses Modells ein substanzieller Verweigereranteil von bis zu 50 % nachgewiesen (z. B. Ostapczuk u. a. 2011; Pitsch u. a. 2007). Da über den wahren Merkmalsstatus von Verweigerern im CDM explizit keine Annahmen formuliert werden, ist eine Schätzung für die Prävalenz des sensiblen Merkmals in solchen Fällen nur im mitunter breiten Intervall von π (wenn kein Verweigerer Merkmalsträger ist) bis $\pi + \gamma$ (wenn alle Verweigerer Merkmalsträger sind) möglich.

Ein weiteres Modell zur Schätzung unehrlich antwortender Teilnehmerinnen und Teilnehmer wurde mit dem „Stochastischen Lügendetektor“ (SLD; Moshagen u. a. 2012) vorgestellt. Dieses Modell implementiert eine Modifikation des Modells von Mangat (1994). Allen Studienteilnehmenden werden zwei Aussagen präsentiert. Falls sie Merkmalsträgerinnen und -träger sind, werden die Befragten explizit angewiesen, zu Aussage A Stellung zu nehmen. Nicht-Merkmalsträgerinnen und Nicht-Merkmalsträger durchlaufen wie gewohnt das Zufallsexperiment und nehmen Stellung zu einer der sensiblen Aussagen A und B, von denen eine die Negation der anderen ist (siehe Abbildung 3).

Abbildung 3:

Beispiel für eine mögliche RRT-Frage im Format des SLD in einer Opferwerdungsbefragung zum Thema Vergewaltigung

Im Folgenden werden Ihnen zwei Aussagen präsentiert.
Falls Sie schon einmal Opfer einer Vergewaltigung geworden sind, nehmen Sie bitte in jedem Fall Stellung zu Aussage A.
Falls Sie noch nie Opfer einer Vergewaltigung geworden sind, nehmen Sie bitte Stellung zu …
… Aussage A, wenn Sie *im November oder Dezember* geboren wurden.
… Aussage B, wenn Sie *in einem anderen Monat* geboren wurden.

Aussage A:	Ich wurde schon einmal vergewaltigt.
Aussage B:	Ich wurde noch nie vergewaltigt.

Antwort: [ ] Stimmt
[ ] Stimmt nicht

Auch bei Anwendung des SLD lässt eine „Stimmt"-Antwort keinen eindeutigen Schluss auf den wahren Merkmalsstatus zu. Eine „Stimmt nicht"-Antwort bietet allerdings eine scheinbar sichere Antwortalternative. Im Gegensatz zum CDM ist in den Modellannahmen des SLD explizit formuliert, dass nur Merkmalsträgerinnen und -träger den Drang zu einer unehrlichen Antwort verspüren sollten; Nicht-Merkmalsträgerinnen bzw. Nicht-Merkmalsträger sollten ehrlich antworten. Entsprechend ist auf der Basis des SLD zusätzlich zur Schätzung des Anteils von Merkmalsträgerinnen und -trägern (π) eine Schätzung des Anteils ehrlich antwortender Merkmalsträgerinnen und -träger t (= true) möglich, der als Teilmenge von π begriffen wird. Wie beim CDM müssen auch bei Anwendung des SLD zwei unabhängige Stichproben mit unterschiedlichen Randomisierungswahrscheinlichkeiten erhoben werden. Entsprechende Gleichungen für die Schätzung von π und t, für die Varianzen der Schätzer und eine Umsetzung als multinomiales Modell finden sich in Moshagen u. a. (2012). Wenngleich mehrere Anwendungen vielversprechende Ergebnisse in Bezug auf die Validität des Modells und die Nützlichkeit des t-Parameters ermitteln konnten (z. B. Moshagen u. a. 2014; Moshagen u. a. 2012), deuten manche Befunde auf eine gegenüber Konkurrenzmodellen reduzierte Verständlichkeit der verhältnismäßig komplexen Instruktionen hin (z. B. Hoffmann/Musch, 2015; Hoffmann u. a., 2015b).

3.3.3 Das Crosswise-Modell

Kürzlich wurde eine neue Klasse indirekter Befragungstechniken, die sogenannten *Nonrandomized-Response*-Modelle (NRRT; Tian/Tang 2013), vor-

gestellt, die besonderes Gewicht auf die praktische Anwendbarkeit und die Verständlichkeit der Instruktionen legen. Ein vielversprechender Vertreter dieser Klasse ist das *Crosswise*-Modell (CWM; Yu u. a. 2008). Bei CWM-Fragen wird die Zufallsverschlüsselung individueller Antworten direkt in die Antwortoptionen eingebunden. Den Teilnehmerinnen und Teilnehmern werden zwei Aussagen präsentiert: eine sensible Aussage A mit unbekannter Prävalenz π und eine nicht sensible Aussage B mit bekannter Prävalenz p (z. B. eine Aussage zum Geburtsmonat). Befragte sollen nun jedoch nicht auf die einzelnen Aussagen antworten, sondern lediglich in einer verbundenen Antwort zu beiden Aussagen gleichzeitig Stellung nehmen (siehe *Abbildung 4*).

Abbildung 4:

Beispiel für eine mögliche Frage im CWM-Format in einer Opferwerdungsbefragung zum Thema Vergewaltigung

Im Folgenden werden Ihnen zwei Aussagen präsentiert, zu denen Sie in einer verbundenen Antwort Stellung nehmen sollen.

Aussage A:	Ich wurde schon einmal vergewaltigt.
Aussage B:	Ich wurde im November oder Dezember geboren.

Antwort:	[ ] Beide Aussagen stimmen oder keine der beiden Aussagen stimmt. [ ] Genau eine der beiden Aussagen (egal welche) stimmt.

Da das CWM mathematisch äquivalent mit Warners *Related-Questions*-Modell ist (Ulrich u. a. 2012), können die Schätzung für π und die Varianz von π anhand derselben Gleichungen bestimmt werden (Warner 1965; Yu u. a. 2008). Beispiele für Umsetzungen als multinomiales Modell sowie Hinweise auf eine gegenüber einer direkten Frage erhöhte Validität finden sich in Hoffmann u. a. (2015a, 2015b). Außerdem konnte gezeigt werden, dass das CWM gegenüber konkurrierenden Modellen bezüglich der Verständlichkeit der Instruktionen überlegen zu sein scheint und seine Anwendung mit einer gegenüber einer direkten Frage substanziell erhöhten subjektiv empfundenen Vertraulichkeit einhergeht (Hoffmann u. a. 2015b).

3.3.4 Die Unmatched-Count-Technik

Ähnlich wie die Klasse der *Nonrandomized-Response*-Modelle zielt die *Unmatched-Count*-Technik (UCT, manchmal auch *Item-Count*-, *Randomized-List*-Technik o. Ä.; Miller 1984) besonders auf ein hohes Instruktionsverständnis sowie auf eine direkte Sichtbarkeit des Vertraulichkeitsschutzes ab. In Befragungen mit der UCT werden zwei unabhängigen Gruppen jeweils

Listen mit mehreren Aussagen vorgelegt. Bei einer ersten (Experimental-)Gruppe besteht diese Liste aus mehreren Aussagen zu nicht sensiblen Merkmalen sowie einer Aussage zu dem interessierenden sensiblen Merkmal. Eine zweite (Kontroll-)Gruppe erhält nur die nicht sensiblen Aussagen. Versuchsteilnehmerinnen und -teilnehmer sollen nun lediglich die Anzahl der Aussagen angeben, denen sie zustimmen, unabhängig davon welcher der Aussagen zugestimmt wird (siehe *Abbildung 5*).

Abbildung 5:

Beispiel für eine mögliche Frage im UCT-Format in einer Opferwerdungsbefragung zum Thema Vergewaltigung

Im Folgenden werden Ihnen mehrere Aussagen präsentiert. Geben Sie bitte an, wie viele der Aussagen auf Sie persönlich zutreffen.

Gruppe 1 (Experimentalgruppe)		*Gruppe 2 (Kontrollgruppe)*	
Aussage 1:	Ich wurde im November oder Dezember geboren.	**Aussage 1:**	Ich wurde im November oder Dezember geboren.
Aussage 2:	Ich war schon einmal in London.	**Aussage 2:**	Ich war schon einmal in London.
Aussage 3:	Ich bin Vegetarier(in).	**Aussage 3:**	Ich bin Vegetarier(in).
Aussage 4:	Ich wurde schon einmal vergewaltigt.		

Antwort:

[ ] Keine Aussage trifft zu.
[ ] Eine Aussage trifft zu.
[ ] Zwei Aussagen treffen zu.
[ ] Drei Aussagen treffen zu.
[ ] Alle Aussagen treffen zu.

Sofern durch eine angemessene Auswahl der Art und Anzahl nicht sensibler Aussagen vermieden wird, dass Teilnehmerinnen bzw. Teilnehmer keiner oder allen Aussagen zustimmen müssen, bleibt die Vertraulichkeit individueller Antworten gewahrt (Erdfelder/Musch 2006; Fox/Tracy 1986). Auf Stichprobenebene liefert die Differenz zwischen Experimental- und Kontrollgruppe in der mittleren Anzahl zustimmender Aussagen eine Schätzung für die Prävalenz des sensiblen Merkmals. Auch für die UCT liegen mehrere Studien vor, die auf eine gegenüber einer direkten Frage erhöhte Validität hindeuten (z. B. Coutts/Jann 2011; LaBrie/Earleywine 2000; Wimbush/Dalton 1997). Darüber hinaus existieren Hinweise darauf, dass die UCT konkurrierenden Modellen bezüglich der subjektiv empfundenen Vertraulichkeit individueller Antworten überlegen sein könnte (Coutts/Jann 2011; Hoffmann u. a. 2015b).

3.3.5 Die Verwendung indirekter Fragetechniken in Viktimisierungsstudien

Die Anwendung indirekter Befragungstechniken in Viktimisierungsstudien wurde sowohl im deutschsprachigen Raum (Treibel/Funke 2004) als auch international gefordert (Fox/Tracy 1980). Der Literatur sind allerdings bislang entsprechende Studien nur für amerikanische Stichproben zu entnehmen. Diese werden im Folgenden dargestellt. Dabei garantieren die indirekten Fragetechniken Vertraulichkeit und sind deshalb aus ethischen Gründen für Viktimisierungs- (und auch für Täter-)Studien besonders geeignet (Fox/Tracy 1980). Indirekte Fragetechniken wurden erfolgreich in Onlinestudien und in Papier-Bleistift-Befragungen eingesetzt (z. B. Hoffmann u. a. 2015a; Hoffmann/Musch 2015; Moshagen u. a. 2012; Ostapczuk u. a. 2009). Inwieweit indirekte Fragetechniken auch in bevölkerungsrepräsentativen Telefonbefragungen mit wenig gebildeten Teilnehmerinnen und Teilnehmern einsetzbar sind, wurde bislang nur unzureichend erforscht.

Bereits 1986 wurde die RRT zur Erfassung der Prävalenz von Vergewaltigung eingesetzt (Soeken/Damrosch 1986). Befragt wurden insgesamt 368 Personen, wovon etwa die Hälfte eine Frage im RRT-Format beantwortete. Die ermittelte Prävalenz lag bei etwa 15 %. Da die Studie keine Kontrollgruppe vorsah, war allerdings kein unmittelbarer Vergleich zwischen Prävalenzschätzungen anhand herkömmlicher Methoden und der indirekten Befragung möglich. Ob die Verwendung der RRT-Frage zu einer höheren Prävalenzschätzung führte, ist demnach unklar. Einen direkten Vergleich zwischen verschiedenen Befragungsmethoden ermöglichte eine Studie an 331 Studierenden (Thornton/Gupta 2004). Die teilnehmenden Personen wurden zufällig einer der fünf Befragungsbedingungen zugewiesen. Im Vergleich zu einem Face-to-Face-Interview (15 %) und einer anonymen schriftlichen Befragung (22 %) wurde die Prävalenz von Gewalt in der Beziehung unter Verwendung der RRT am höchsten eingeschätzt (33 %); eine Bogus-Pipeline-Befragung ergab mit 19 % eine geringere Prävalenzschätzung als die RRT. In einer weiteren Studie wurden unter Verwendung der UCT anhand einer Stichprobe mit 5.446 Studierenden online zwei Arten sexueller Viktimisierung abgefragt (Krebs u. a. 2011). Die Studie deckte einen Unterschied zwischen den gemittelten Prävalenzen der direkten Befragung (4,7 %) und der UCT-Bedingung (5,3 %) in der erwarteten Richtung auf, der jedoch zufallskritisch nicht abgesichert werden konnte. In einer weiteren Studie wurde die Prävalenz aus Vorurteilen resultierender Übergriffe (sogenannten *Hate crimes*) in einer Stichprobe mit 287 Studierenden entweder unter Verwendung der UCT oder anhand eines direkten Frageformats untersucht (Rayburn u. a. 2003). Insgesamt wurden 15 verschiedene Formen von Viktimisierung abgefragt. In der

UCT-Bedingung wurden fast ausnahmslos höhere Prävalenzen gefunden. Die Unterschiede zwischen der direkten und der UCT-Befragung fielen für 13 der 15 Formen von Viktimisierung signifikant aus, u. a. für sexuelle Belästigung (15,8 % vs. 27,7 %) und sexuelle Nötigung (3,4 % vs. 23,6 %).

Trotz der Vorzüge indirekter Befragungstechniken wurden diese bislang in Viktimisierungsstudien kaum eingesetzt, obwohl sie im Vergleich zu direkten Befragungen eine ethisch unbedenklichere Erhebungsmethode darstellen. Die indirekte Formulierung einzelner Fragen schließt eine generelle Verwendung direkter Fragen nicht aus. Vielmehr kann durch einen Vergleich indirekter Befragungsbedingungen mit Kontrollgruppen, in denen direkte Fragen verwendet werden, eine Aussage darüber getroffen werden, ob eine Verzerrung durch soziale Erwünschtheit tatsächlich vorliegt. Nur dann ist der Aufwand des Einsatzes indirekter Fragetechniken gerechtfertigt. Wichtig ist bei deren Verwendung nämlich, ausreichend große Stichproben zu erheben, um ihre im Vergleich zu direkten Befragungen verringerte Effizienz zu kompensieren. Für die Erzielung hinreichend präziser Prävalenzschätzungen sind in der Regel drei- bis vierstellige Stichprobengrößen erforderlich (Ulrich u. a. 2012). Derartige Stichprobenumfänge sind in Viktimisierungsstudien allerdings ohnehin wünschenswert. Ein Nachteil der Verwendung indirekter Fragetechniken besteht darin, dass die Durchführung von Zusammenhangsanalysen aufgrund der verwendeten Randomisierungsverfahren nur mittels fortgeschrittener statistischer Verfahren möglich ist (Böckenholt u. a. 2009). Zu klären bleibt in diesem Kontext die Frage nach der Anwendbarkeit indirekter Fragetechniken und dem Instruktionsverständnis in bildungsfernen Stichproben (Hoffmann u. a. 2015b).

Zusammenfassend kann festgehalten werden, dass sozial erwünschtes Antwortverhalten potenziell eine fundamentale Bedrohung der Validität von Viktimisierungsstudien darstellt. Ansätze zur besseren Kontrolle sozialer Erwünschtheit in Viktimisierungsstudien existieren; sie wurden in diesem Kapitel dargestellt und erläutert. Indirekte Befragungstechniken erweisen sich als besonders vielversprechende Kandidaten für die Verbesserung der Validität von Viktimisierungsstudien. Wir empfehlen deshalb, auch im deutschsprachigen Raum den Nutzen indirekter Fragetechniken in Viktimisierungsstudien zu überprüfen und im Fall von Antwortverzerrungen die Vorteile dieser Fragetechniken verstärkt für die Viktimisierungsforschung zu nutzen, um den Einfluss sozialer Erwünschtheit besser als in bisherigen Studien kontrollieren zu können.

4 Zusammenfassung

Sozial erwünschtes Antwortverhalten ist die Tendenz, sich selbst in einem möglichst positiven Licht darzustellen und vorhandenen sozialen Normen zumindest im Antwortverhalten gerecht zu werden, um einer möglichen Missbilligung durch Dritte vorzubeugen.

- Sozial erwünschtes Antwortverhalten ist potenziell eine fundamentale Bedrohung der Validität von Viktimisierungsstudien und kann zu substanziellen Verzerrungen bei der Schätzung der Prävalenz sensibler Merkmale führen.

- Bislang wurden mögliche Verzerrungen durch soziale Erwünschtheit in Viktimisierungsstudien wenig berücksichtigt.

- Indirekte Befragungstechniken sind ein besonders vielversprechendes Mittel zur Kontrolle sozial erwünschten Antwortverhaltens.

- Eine dieser indirekten Befragungstechniken, das Crosswise-Modell, hat sich in Studien als besonders leicht verständlich und für die Kontrolle sozialer Erwünschtheit gut geeignet erwiesen.

- Keine der zur Verfügung stehenden indirekten Befragungstechniken kam jedoch bislang in deutschsprachigen Viktimisierungsstudien zum Einsatz.

- Eine häufigere Verwendung indirekter Befragungstechniken für die Viktimisierungsforschung auch im deutschsprachigen Raum wird deshalb empfohlen.

5 Literatur

Abul-Ela, Abdel-Latif A.; Greenberg, Gernard G. und Horvitz, Daniel G. (1967): A multi-proportions randomized response model. In: Journal of the American Statistical Association, 62, S. 990–1008.

Aguinis, Herman; Handelsman, Mitchell M. (1997): Ethical issues in the use of the bogus pipeline. In: Journal of Applied Social Psychology, 2 (7), S. 557–573.

Antonak, Richard F.; Livneh, Hanoch (1995): Randomized-response technique – a review and proposed extension to disability attitude research. In: Genetic, Social, and General Psychology Monographs, 12 (1), S. 97–145.

Arias, Ileana; Beach, Steven R. H. (1987): Validity of self-reports of marital violence. In: Journal of Family Violence, 2 (2), S. 139–149.

Averdijk, Margit; Elffers, Henk (2012): The discrepancy between survey-based victim accounts and police reports revisited. In: International Review of Victimology, 18 (2), S. 91–107.

Barnett, Julie (1998): Sensitive questions and response effects: An evaluation. In: Journal of Managerial Psychology, 13 (1-2), S. 63–76.

Batchelder, William H. (1998): Multinomial processing tree models and psychological assessment. In: Psychological Assessment, 10, S. 331–344.

Batchelder, William H.; Riefer, David M. (1999): Theoretical and empirical review of multinomial process tree modeling. In: Psychonomic Bulletin & Review, 6 (4), S. 57–86.

Bell, Kathryn M.; Naugle, Amy E. (2007): Effects of social desirability on students' self-reporting of partner abuse perpetration and victimization. In: Violence and Victims, 22 (2), S. 243–256.

Birkel, Christoph (2003): Die polizeiliche Kriminalstatistik und ihre Alternativen. In: Der Hallesche Graureiher, 2003, 1, S. 1–111.

Block, Richard (1993): A cross-national comparison of victims of crime: Victim surveys of twelve countries. In: Review of Victimology, 2 (3), S. 183–207.

Böckenholt, Ulf; Barlas, Sema; van der Heijden, Peter G. M. (2009): Do randomized-response designs eliminate response biases? An empirical study of non-compliance behavior. In: Journal of Applied Econometrics, 24 (3), S. 377–392.

Borkenau, Peter; Amelang, Manfred (1986): Zur faktorenanalytischen Kontrolle sozialer Erwünschtheitstendenzen. Eine Untersuchung anhand des Freiburger-Persönlichkeits-Inventars. In: Zeitschrift für Differentielle und Diagnostische Psychologie, 7 (1), S. 17–28.

Boruch, Robert F. (1971): Assuring confidentiality of responses in social research: A note on strategies. In: American Sociologist, 6 (4), S. 308–311.

Cantor, David; Lynch, James P. (2000): Self-report surveys as measures of crime and criminal victimization. In: Measurement and Analysis of Crime and Justice, 4, S. 85–138.

Chaudhuri, Arijit; Christofides, Tasos C. (2013): Indirect questioning in sample surveys. Berlin, Heidelberg: Springer.

Clark, Stephen J.; Desharnais, Robert A. (1998): Honest answers to embarrassing questions: Detecting cheating in the randomized response model. In: Psychological Methods, 3 (2), S. 160–168.

Cook, Sarah L.; Gidycz, Christine A.; Koss, Mary P. und Murphy, Megan (2011): Emerging issues in the measurement of rape victimization. In: Violence against Women, 17 (2), S. 201–18.

Coutts, Elisabeth; Jann, Ben (2011): Sensitive questions in online surveys: Experimental results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). In: Sociological Methods & Research, 40 (1), S. 169–193.

Crowne, Douglas P.; Marlowe, David (1960): A new scale of social desirability independent of psychopathology. In: Journal of Consulting Psychology, 24 (4), S. 349–354.

Dawes, Robyn M.; Moore, Michael (1980): Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen. In: Petermann, Franz (Hg.): Einstellungsmessung, Einstellungsforschung. Göttingen: Hogrefe, S. 117–133.

Dutton, Donald G.; Hemphill, Kenneth J. (1992): Patterns of socially desirable responding among perpetrators and victims of wife assault. In: Violence and Victims, 7 (1), S. 29–39.

Edgell, Stephen E.; Duchan, Karen L. und Himmelfarb, Samuel (1992): An empirical-test of the Unrelated Question Randomized-Response Technique. In: Bulletin of the Psychonomic Society, 30 (2), S. 153–156.

Edgell, Stephen E.; Himmelfarb, Samuel und Duchan, Karen L. (1982): Validity of forced responses in a Randomized-Response model. In: Sociological Methods & Research, 11 (1), S. 89–100.

Edwards, Allen L. (1957): The social desirability variable in personality assessment and research. New York: The Dryden Press.

Erdfelder, Edgar; Musch, Jochen (2006): Experimental methods of psychological assessment. In: Eid, Michael und Diener, Ed (Hg.): Handbook of Multimethod Measurement in Psychology. Washington DC: American Psychological Association, S. 205–220.

Fox, James A.; Tracy, Paul E. (1980): The Randomized Response approach: Applicability to criminal justice research and evaluation. In: Evaluation Review, 4 (5), S. 601–622.

Fox, James A.; Tracy, Paul E. (1986): Randomized Response: A method for Sensitive Surveys, Beverly Hills, CA: Sage.

Görgen, Thomas; Rabold, Susann und Herbst, Sandra (2006): Viktimisierung im Alter und in der häuslichen Pflege: Wege in ein schwieriges Forschungsfeld. Befragungsinstruments der Studie „Kriminalität und Gewalt im Leben alter Menschen“. Forschungsbericht Nr. 99. Hannover: Kriminologisches Forschungsinstitut Niedersachsen.

Görgen, Thomas (2009): Viktimologie. In: Kröber, Hans-Ludwig; Dölling, Dieter; Leygraf, Norbert und Sass, Henning (Hg.): Handbuch der Forensischen Psychiatrie. Band 4. Kriminologie und Forensische Psychiatrie. Darmstadt: Steinkopff, S. 236–264.

Grabitz-Gniech, Gisla (1971): Some restrictive conditions for the occurrence of psychological reactance. In: Journal of Personality and Social Psychology, 19 (2), S. 188–196.

Greenberg, Gernard G.; Abul-Ela, Abdel-Latif A.; Simmons, Wait R. und Horvitz, Daniel G. (1969): Unrelated question randomized response model: Theoretical framework. In: Journal of the American Statistical Association, 64 (326), S. 520–539.

Greve, Werner; Strobl, Rainer und Wetzels, Peter (1994): Das Opfer kriminellen Handelns: Flüchtig und nicht zu fassen. Konzeptuelle Probleme und methodische Implikationen eines sozialwissenschaftlichen Opferbegriffes. Forschungsbericht Nr. 33. Hannover: Kriminologisches Forschungsinstitut Niedersachsen.

Guzy, Nathalie; Leitgöb, Heinz (2015): Assessing mode effects in online and telephone victimization surveys. In: International Review of Victimology, 21 (1), S. 101–131.

Hamby, Sherry L.; Koss, Mary P. (2003): Shades of gray: A qualitative study of terms used in the measurement of sexual victimization. In: Psychology of Women Quarterly, 27 (3), S. 243–255.

Heckert, D. Alex; Gondolf, Edward W. (2000): Assessing assault self-reports by batterer program participants and their partners. In: Journal of Family Violence, 15 (2), S. 181–197.

Heinz, Wolfgang (2006): Zum Stand der Dunkelfeldforschung in Deutschland. In: Obergfell-Fuchs, J. und Brandenstein, M. (Hg.): Nationale und internatione Entwicklungen in der Kriminologie. Frankfurt: Verlag für Polizeiwissenschaften, S. 241–263.

Heinz, Wolfgang (2009): Kriminalität und Kriminalitätskontrolle in Deutschland. In: Kröber, Hans-Ludwig; Dölling, Dieter; Leygraf, Norbert und Sass, Henning (Hg.): Handbuch der Forensischen Psychiatrie. Band 4. Kriminologie und Forensische Psychiatrie. Darmstadt: Steinkopff, S. 1–133.

Hoffmann, Adrian; Diedenhofen, Birk; Verschuere, Bruno und Musch, Jochen (2015a): A strong validation of the Crosswise Model using experimentally induced cheating behavior. In: Experimental Psychology (im Erscheinen).

Hoffmann, Adrian; Musch, Jochen (2015): Assessing the validity of two indirect questioning techniques: A stochastic lie Detector versus the Crosswise Model. In: Behavior Research Methods, DOI 10.3758/s13428-015-0628-6.

Hoffmann, Adrian; Schmidt, Alexander F.; Waubert de Puiseau, Berenike und Musch, Jochen (2015b): On the comprehensibility and perceived privacy protection of indirect questioning techniques, im Erscheinen.

Hu, Xiangen; Batchelder, William H. (1994): The statistical analysis of general processing tree models with the EM algorithm. In: Psychometrika, 59 (1), S. 21–47.

Jones, Edward E.; Sigall, Harold (1971): The Bogus Pipeline: A new paradigm for measuring affect and attitude. In: Psychological Bulletin, 76 (5), S. 349–364.

Kerlinger, Fred N.; Pedhazur, Elazar J. (1973): Multiple regression in behavioral research. New York u. a.: Holt, Rinehart and Winston.

Kilchling, Michael (2010): Veränderte Perspektiven auf die Rolle des Opfers im gesellschaftlichen, sozialwissenschaftlichen und rechtspolitischen Diskurs. In: Hartmann, Jutta und ado e. V. (Hg.): Perspektiven professioneller Opferhilfe. Theorie und Praxis eines interdisziplinären Handlungsfelds. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 39–50.

Koss, Mary P. (1993): Detecting the scope of rape – a review of prevalence research methods. In: Journal of Interpersonal Violence, 8 (2), S. 198–222.

Koss, Mary P. (1996): The measurement of rape victimization in crime surveys. In: Criminal Justice and Behavior, 23 (1), S. 55–69.

Koss, Mary P.; Abbey, Antonia; Campbell, Rebecca; Cook, Sarah; Norris, Jeanette; Testa, Maria; Ullman, Sarah; West, Carolyn und White, Jacquelyn (2007): Revising the Ses: A collaborative process to improve Assessment of sexual aggression and victimization. In: Psychology of Women Quarterly, 31 (4), S. 357–370.

Koss, Mary P.; Gidycz, Christine A. (1985): Sexual experiences survey: Reliability and validity. In: Journal of Consulting and Clinical Psychology, 53 (3), S. 422–423.

Koss, Mary P.; Oros, Cheryl J. (1982): Sexual Experiences Survey: A research instrument investigating sexual aggression and victimization. In: Journal of Consulting and Clinical Psychology, 50 (3), S. 455–457.

Krebs, Christopher P.; Lindquist, Christine H.; Warner, Tara D.; Fisher, Bonnie S.; Martin, Sandra L. und Childers, James M. (2011): Comparing sexual assault prevalence estimates obtained with direct and indirect questioning techniques. In: Violence against Women, 17 (2), S. 219–235.

Kuk, Anthony Y. C. (1990): Asking sensitive questions indirectly. In: Biometrika, 77, S. 436–438.

Kury, Helmut (1994): Zum Einfluß der Art der Datenerhebung auf die Ergebnisse von Umfragen. In: Monatsschrift für Kriminologie und Strafrechtsreform, 77 (1), S. 22–33.

Kury, Helmut (2010): Entwicklungslinien und zentrale Befunde der Viktimologie. In: Hartmann, Jutta und ado e. V. (Hg.): Perspektiven professioneller Opferhilfe. Theorie und Praxis eines interdisziplinären Handlungsfelds. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 51–72.

LaBrie, Joseph W.; Earleywine, Mitchell (2000): Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. In: Journal of Sex Research, 37 (4), S. 321–326.

Langhinrichsen-Rohling, Jennifer; Vivian, Dina (1994): The correlates of spouses' incongruent reports of marital aggression. In: Journal of Family Violence, 9 (3), S. 265–283.

Lenhard, Wolfgang; Breitenbach, Erwin; Ebert, Harald; Schindelhauer-Deutscher, Hans-Joachim und Henn, Wolfram (2005): Psychological benefits of diagnostic certainty for mothers of children with disabilities: Lessons from Down syndrome. In: American Journal of Medical Genetics Part A, 133 (2), S. 170–175.

Lensvelt-Mulders, Gerty J. L. M.; Hox, Joop J.; van der Heijden, Peter G. M. und Maas, Cora J. M. (2005): Meta-analysis of randomized response research thirty-five years of validation. In: Sociological Methods & Research, 33 (3), S. 319–348.

Liu, P. T.; Chow, L. P. (1976): A new discrete quantitative Randomized Response model. In: Journal of the American Statistical Association, 7 (3), S. 72–73.

Lück, Helmut E.; Timaeus, Ernst (1969): SDS-CM – Skala zur Erfassung sozialer Wünschbarkeit (CM-Skala). In: Diagnostica, 15, S. 134–141.

Mangat, Naurang S. (1994): An improved Randomized-Response strategy. In: Journal of the Royal Statistical Society: Series B (Statistical Methodology), 56, S. 93–95.

Miller, Judith D. (1984). A new survey technique for studying deviant behavior, unveröffentlichte Dissertation, George Washington University, Department of Sociology.

Moshagen, Morten; Hilbig, Benjamin E.; Erdfelder, Edgar und Moritz, Annie (2014): An experimental validation method for questioning techniques that assess sensitive issues. In: Experimental Psychology, 61 (1), S. 48–54.

Moshagen, Morten; Musch, Jochen und Erdfelder, Edgar (2012): A stochastic lie detector. In: Behavior Research Methods, 44 (1), S. 222–231.

Mummendey, Hans Dieter (1981): Methoden und Probleme der Kontrolle sozialer Erwünschtheit (Social Desirability). In: Zeitschrift für Differentielle und Diagnostische Psychologie, 2, S. 199–218.

Musch, Jochen; Brockhaus, Robbi und Bröder, Arndt (2002): Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. In: Diagnostica, 48 (3), S. 121–129.

Musch, Jochen; Ostapczuk, Martin und Klaiber, Yvonne (2012): Validating an inventory for the assessent of egoistic bias and moralistic bias as two separable cmponents of social desirability. In: Journal of Personality Assessment, 94 (6), S. 620–629.

Nederhof, Anton J. (1985): Methods of coping with social desirability bias – a review. In: European Journal of Social Psychology, 15, S. 263–280.

Ohlemacher, Thomas; Gabriel, Ute; Mecklenburg, Eberhard und Pfeiffer, Christian (1997). Die KFN-Geschäftsleute-Erhebung. Deutsche und ausländische Gastronomen in Konfrontation mit Schutzgelderpressung und Korruption: Erste Befunde der Hauptuntersuchung. Forschungsbericht Nr. 61. Hannover: Kriminologisches Forschungsinstitut Niedersachsen.

Olson, Cheryl B.; Stander, Valerie A. und Merrill, Lex L. (2004): The influence of survey confidentiality and construct measurement in estimating rates of childhood victimization among navy recruits. In: Military Psychology, 16 (1), S. 53–69.

Ostapczuk, Martin; Moshagen, Morten; Zhao, Zengmei und Musch, Jochen (2009): Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. In: Journal of Educational and Behavioral Statistics, 34 (2), S. 267–287.

Ostapczuk, Martin; Musch, Jochen (2011): Estimating the prevalence of negative attitudes towards people with disability: A comparison of direct questioning, projective questioning and randomised response. In: Disability and Rehabilitation, 33 (5), S. 1–13.

Ostapczuk, Martin; Musch, Jochen und Moshagen, Morten (2011): Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. In: Statistical Methods in Medical Research, 20 (5), S. 489–503.

Paulhus, Delroy L. (1991): Measurement and control of response bias. In: Robinson, John P.; Shaver, Phillip R. und Wrightsman, Lawrence S. (Hg.): Measures of personality and social psychological attitudes, Band 1. San Diego, CA: Academic Press, S. 17–59.

Paulhus, Delroy L. (2002): Socially desirable responding: The evolution of a construct. In: Braun, Henry I.; Jackson, Douglas N. und Wiley, David E. (Hg.): The role of constructs in psychological and educational measurement. Mahwah, New Jersey, London: L. Erlbaum Publishers, S. 49–69.

Pauls, Cornelia A.; Crost, Nicolas W. (2004): Jenseits von Werturteilen: Ein Plädoyer für eine empirische Erforschung sozial erwünschten Antwortverhaltens in Bewerbungskontexten. In: Zeitschrift für Personalpsychologie, 3 (2), S. 79–82.

Pitsch, Werner; Emrich, Elke und Klein, Markus (2007): Doping in elite sports in Germany: results of a www survey. In: European Journal of Sport and Society, 4 (2), S. 89–102.

Rayburn, Nadine R.; Earleywine, Mitchell und Davison, Gerald C. (2003): Base rates of hate crime victimization among college students. In: Journal of Interpersonal Violence, 18 (10), S. 1209–1221.

Roese, Neal J.; Jamieson, David W. (1993): 20 years of bogus pipeline research – a critical review and metaanalysis. In: Psychological Bulletin, 114 (2), S. 363–375.

Schneider, Anne L. (1981): Methodological problems in victim surveys and their implications for research in victiminology. In: The Journal of Criminal Law & Criminology, 72 (2), S. 818–838.

Schneider, Hans-Joachim (1979): Das Opfer und sein Täter. Partner im Verbrechen, München: Kindler.

Skogan, Wesley G. (1975): Measurement problems in official and survey crime rates. In: Journal of Criminal Justice, 3 (1), S. 17–32.

Soeken, Karen L.; Damrosch, Shirley P. (1986): Randomized response technique: Applications to research on rape. In: Psychology of Women Quarterly, 10 (2), S. 119–125.

Sorenson, Susan B.; Stein, Judith A.; Siegel, Judith M.; Golding, Jaqueline M. und Burnam, M. Audrey (1987): The prevalence of adult sexual assault. The Los Angeles epidemiologic catchment area project. In: American Journal of Epidemiology, 126 (6), S. 1154–1164.

Stöber, Joachim (1999): The Social Desirability Scale-17 (SDS-17). Convergent validity, discriminant validity, and relationship with age. In: European Journal of Psychological Assessment, 17 (3), S. 222–232.

Stocké, Volker (2004): Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Prognosen der Rational-Choice Theorie und des Modells der Frame-Selektion. In: Zeitschrift für Soziologie, 33 (4), S. 303–320.

Straus, Murray A. (1979): Measuring intrafamily conflict and violence: The conflict tactics (CT) scales. In: Journal of Marriage and the Family, 41 (1), S. 75–88.

Sugarman, David B.; Hotaling, Gerald T. (1997): Intimate violence and social desirability: A meta-analytic review. In: Journal of Interpersonal Violence, 12 (2), S. 275–290.

Tabachnick, Barbara G.; Fidell, Linda S. (2012): Using multivariate statistics, 6. Aufl. Cloth: Pearson.

Thornton, Bill; Gupta, Sat (2004): Comparative validation of a partial (versus full) randomized response technique: Attempting to control for social desirability reponse bias to sensitive questions. In: Individual Differences Research, 2 (3), S. 214–224.

Tian, Guo-Liang; Tang, Man-Lai (2013): Incomplete categorical data design: Non-Randomized Response techniques for sensitive questions in surveys. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Tourangeau, Roger; Yan, Ting (2007): Sensitive questions in surveys. In: Psychological Bulletin, 133 (5), S. 859–883.

Tracy, D. S.; Mangat, Naurang S. (1996): Some development in randomized response sampling during the last decade – a follow up of review by Chaudhuri and Mukerjee. In: Journal of Applied Statistical Science, 4, S. 147–158.

Treibel, Angelika; Funke, Joachim (2004): Die internetbasierte Opferbefragung als Instrument der Dunkelfeldforschung – Grenzen und Chancen. In: Monatsschrift für Kriminologie und Strafrechtsreform, 87 (2), S. 146–151.

Turner, Anthony G. (1984): An experiment to compare three interview procedures in the National Crime Survey. 2: Methodological Studies. In: Lehnen, Robert G.; Skogan, Wesley G. (Hg.): The National Crime Survey: Working Papers. Bureau of Justice Statistics, S. 49–53.

Ulrich, Rolf; Schröter, Hannes; Striegel, Heiko und Simon, Perikles (2012): Asking sensitive questions: A statistical power analysis of Randomized-Response models. In: Psychological Methods, 17 (4), S. 623–641.

Umesh, Uchila N.; Peterson, Robert A. (1991): A critical evaluation of the Randomized-Response method – applications, validation, and research agenda. In: Sociological Methods & Research, 20, S. 104–138.

van Dijk, Jan J. M.; Mayhew, Pat; van Kesteren, John; Aebi, Marcelo und Linde, Antonia (2010): Final report on the study on crime victimisation. Tilburg: Tilburg University.

van Dijk, Jan J. M.; van Kesteren, John und Smit, Paul (2007). Criminal victimisation in international perspective. Den Haag: Wetenschappelijk Onderzoek – en Documentatiecentrum.

Warner, Stanley L. (1965): Randomized-response – a survey technique for eliminating evasive answer bias. In: Journal of the American Statistical Association, 60 (309), S. 63–69.

Wetzels, Peter; Pfeiffer, Christian (1995): Sexuelle Gewalt gegen Frauen im öffentlichen und privaten Nahraum. Ergebnisse der KFN-Opferbefragung 1992. Forschungsbericht Nr. 37. Hannover: Kriminologisches Forschungsinstitut Niedersachsen.

Wimbush, James C.; Dalton, Dan R. (1997): Base rate for employee theft: Convergence of multiple methods. In: Journal of Applied Psychology, 82 (5), S. 756–763.

Yu, Jun-Wu; Tian, Guo-Liang und Tang, Man-Lai (2008): Two new models for survey sampling with sensitive characteristic: design and analysis. In: Metrika, 67 (3), S. 251–263.

Datenschutzrechtliche Grundlagen für die Durchführung repräsentativer Dunkelfeld-Opferbefragungen

Janina Hatt

1 Einleitung

Der Schutz personenbezogener Daten und das wissenschaftliche Erkenntnisinteresse bilden in der Dunkelfeldforschung ein komplexes Spannungsfeld: Einerseits ist das Forschungsinteresse unbegrenzt. Andererseits dürfen personenbezogene Daten nur dann erhoben werden, wenn es unbedingt erforderlich ist. Zudem unterliegt der Umgang mit personenbezogenen Daten zahlreichen Beschränkungen. Der Datenschutz setzt dem Forschungsinteresse also Grenzen. Ein effektiver Schutz personenbezogener Daten ist zugleich aber auch eine notwendige Bedingung für eine ergiebige Opferbefragung im Dunkelfeld, denn die Teilnahme- und Auskunftsbereitschaft der Probanden hängt entscheidend davon ab, ob ihr Anonymitätsinteresse gewahrt bleibt. Die Teilnehmer werden nur dann bereitwillig und wahrheitsgemäß Auskunft erteilen, wenn sie sich sicher sein können, dass ihre Angaben ausschließlich für das Forschungsprojekt und nicht anderweitig oder gar zu ihrem Nachteil verwendet werden. Sowohl die Forschenden als auch die Probanden haben damit ein gemeinsames Interesse an einem Rechtsrahmen, der verlässlich und transparent den Schutz personenbezogener Daten gewährleistet.

Der vorliegende Beitrag beschreibt diesen rechtlichen Rahmen im Überblick. Ob eine Erhebung oder Verarbeitung personenbezogener Daten zulässig ist, ist immer das Ergebnis einer Abwägungsentscheidung, in der alle konkreten Umstände des Einzelfalls zu berücksichtigen und ihrer Bedeutung entsprechend zu gewichten sind. Im Folgenden werden deshalb die wichtigsten in diesem Zusammenhang einzubeziehenden Erwägungen exemplarisch dargestellt.

2 Anwendungsbereich des Datenschutzrechts

Daten fallen nur dann in den Schutzbereich der Bundes- oder Landesdatenschutzgesetze, wenn sie personenbezogen sind. Wenn Daten dagegen von vornherein anonym erhoben werden, gelten die Restriktionen des Datenschutzrechts nicht. Sie gelten auch ab dem Zeitpunkt nicht mehr, ab dem ur-

sprünglich personenbezogene Daten anonymisiert sind. Die Abgrenzung anonymer von personenbezogenen Daten steht also im Zentrum der Frage, ob bei einem Forschungsprojekt Datenschutzgesetze zu beachten sind.

2.1 Abgrenzung personenbezogener von anonymisierten Daten

Nach § 3 Abs. 1 BDSG sind personenbezogene Daten Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (dazu ausführlich auch Simitis 2014, § 3 Rn. 20 ff.). Eine Person ist bestimmbar und nicht mehr anonym, sobald sie indirekt identifiziert werden kann. Dafür genügt es, dass eine Kennnummer oder ein oder mehrere spezifische Elemente, die Ausdruck der physischen, physiologischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität sind, einer Person zugeordnet werden können.[1] Personenbezogen sind neben dem Namen und der Adresse beispielsweise Kreditkarten-, Telefon- oder Personalnummern, aber auch Kontodaten oder Kfz-Kennzeichen. Aber auch z. B. über das Aussehen oder den Gang kann eine Person identifiziert werden. Personenbezogene Daten liegen außerdem vor, wenn eine Kombination von Merkmalen die Bestimmung einer Einzelperson ermöglicht. Für die Frage, ob ein Datum personenbezogen oder anonym ist, ist damit zunächst entscheidend, ob es überhaupt *theoretisch möglich* ist, dass dieses mit einer bestimmten Person verknüpft werden kann.

Das bloße Pseudonymisieren entfernt den Personenbezug der Daten deshalb nicht. Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren (z. B. § 3 Abs. 6a BDSG). Das bedeutet, dass der Personenbezug nur deshalb wesentlich erschwert oder ausgeschlossen ist, weil z. B. der Name durch eine Kennnummer ersetzt wurde. Der Bezug zu einer bestimmten Person kann aber jederzeit wiederhergestellt werden. Dabei ist nicht relevant, *wer oder was* z. B. die Kennnummern vergibt und die Schlüsselregel kennt, mithilfe derer der Personenbezug wiederhergestellt werden kann. Dies können beispielsweise die Probanden selbst, ein vertrauenswürdiger Dritter oder auch die Forschenden sein (Schaar 2014, 98 f.). Entscheidend ist, dass die Daten wieder zusammengeführt werden können. Deshalb sind auch pseudonymisierte Daten personenbezogene Daten, für die datenschutzrechtlichen Regelungen zu beachten sind.

---

[1] Definition in Art. 2 lit. a der zugrunde liegenden EG-Datenschutzrichtlinie 95/46/EG.

Daten sind dann nicht personenbeziehbar, wenn der Bezug zu einer Person unter keinen Umständen mehr hergestellt werden kann, d. h. wenn sie absolut anonymisiert sind. Aber die sich ständig fortentwickelnde Datenverarbeitungstechnik ermöglicht es, auch aus riesigen Datenmengen Informationen über bestimmte Personen herauszufiltern und Verknüpfungen herzustellen (Metschke/Wellbrock 2002, 21). Mithilfe einer entsprechenden technischen Ausstattung können in vielen Fällen Informationen wieder einer bestimmten Person zugeordnet werden, in denen dies vor der Erfindung und Weiterentwicklung der elektronischen Datenverarbeitung unmöglich erschien. Um den Anwendungsbereich des Datenschutzrechts nicht uferlos werden zu lassen, wird die Definition der Personenbeziehbarkeit deshalb relativiert.

Die Datenschutzgesetze sind auch dann nicht anwendbar, wenn die Daten zwar nicht absolut, aber im Sinne von § 3 Abs. 6 BDSG so anonymisiert sind, dass eine Person nur mit einem unverhältnismäßigen Aufwand reidentifiziert werden kann (sog. faktische Anonymität). Für die Frage, ob der Aufwand im konkreten Einzelfall unverhältnismäßig ist oder nicht, sind einerseits die für eine Identifizierung notwendigen zeitlichen, finanziellen, personellen oder sonstigen Ressourcen entscheidend. Auch das Risiko etwa einer Bestrafung für die Wiederherstellung des Personenbezugs spielt eine Rolle. Andererseits ist zu berücksichtigen, welchen Informationsgewinn der Gehalt der konkreten Daten verspricht, d. h. wie sensibel die Daten sind (Metschke/Wellbrock 2002, 21). Relevant ist auch, in welchem Rahmen die Daten zugänglich sind und welches Zusatzwissen frei verfügbar ist. Das Zusatzwissen kann dabei sowohl bei der verantwortlichen Stelle selbst vorhanden sein als auch bei Dritten eingeholt werden.

2.2 Besonderheiten bei Dunkelfeldstudien

Die Frage, ob ein Datum personenbezogen ist oder nicht, ist auf den konkreten Einzelfall bezogen zu beantworten. Je sensibler ein Datum ist, desto höhere Anforderungen sind daran zu stellen, dass die entsprechende Person nicht identifiziert werden kann.

Dunkelfeldstudien befassen sich mit Straftaten, die den Strafverfolgungs-Behörden verborgen geblieben sind. Auch für Dunkelfeldstudien gelten aber unterschiedliche Maßstäbe, je nachdem welches Deliktsfeld untersucht wird und welche Gründe die Opfer haben, die Straftat nicht anzuzeigen. Die Gründe, weshalb Opfer auf eine Strafanzeige verzichten, sind unterschiedlich: Bei Delikten, die einen eher geringen Schaden verursachen, wie z. B. der Diebstahl eines alten Fahrrads, mag ein Teil der Opfer von einer Anzeige absehen, weil damit ein bürokratischer Aufwand verbunden wäre, den das Opfer wegen einer Bagatelle nicht auf sich nehmen möchte. Dem Opfer kommt es in

diesem Fall also nicht darauf an, anonym zu bleiben. Bei schweren Straftaten liegen die Gründe dagegen tendenziell anders: Hier sieht das Opfer eher aus Gründen der Scham von einer Anzeige ab. Eine weitere Ursache kann darin bestehen, dass das Opfer in einer engen persönlichen Beziehung zum Täter steht und trotz allem nicht möchte, dass dieser stigmatisiert und bestraft wird.

Bei schweren Straftaten ist das Interesse an einer Deanonymisierung in der Regel größer als bei geringfügigen Delikten, da Interessierte hier tendenziell mehr Aufwand betreiben werden, um die Person – und damit möglicherweise auch den Täter – identifizieren zu können, als bei geringfügigen Delikten. Dementsprechend gelten dann auch strengere Anforderungen für die Frage, wann die Daten anonymisiert sind.

2.3 Beispiele für die Abgrenzung anonymisierter und personenbezogener Daten

Beispiel 1: Die Identität eines 105-jährigen Mannes, der in einer namentlich zitierten Gemeinde mit nur wenigen Einwohnern lebt, ist einfach zu ermitteln. Der Datensatz ist also personenbeziehbar. Auch Ortsfremde können verhältnismäßig einfach den Namen der Person erfahren. Nicht mehr identifizierbar ist der Mann aber, wenn Gruppen gebildet werden (Aggregation), z. B. indem mehrere Personen zusammengefasst und Aussagen über „Männer im Rentenalter“ oder „Männer über 70“ getroffen werden. Ergänzend oder alternativ können auch die Informationen von „Gemeinden in Niedersachsen“ oder „Gemeinden im Landkreis Cloppenburg“ zusammengefasst werden. Die Gruppen sind so zu bilden, dass der Rückschluss auf eine bestimmte Person kaum noch möglich ist (Weichert 2013).

Beispiel 2: Daten, die über Telefoninterviews erhoben werden, sind dann nicht personenbezogen, wenn weder Namen und Adressen noch Telefonnummern gespeichert werden. Zudem darf auch über den Informationsgehalt des ausgefüllten Fragebogens kein Rückschluss auf eine bestimmte Person möglich sein. Dazu sind die Fragen bzw. Antwortmöglichkeiten so auszugestalten, dass keine Zuordnung der Einzelmerkmale möglich ist. Das heißt, dass beispielsweise nach der Zugehörigkeit zu einer von mehreren angebotenen Berufsgruppen statt nach der konkreten beruflichen Tätigkeit zu fragen ist. Denn selbst wenn weder Name, Adresse noch Telefonnummer erhoben werden, können bestimmte Merkmalskombinationen oder Einzelmerkmale eine Identifikation zulassen. Ein besonders plakatives Beispiel wäre z. B. die Berufsangabe „Bundeskanzler“. Es ist außerdem sicherzustellen, dass auch über andere Informationen bzw. ihre Kombination kein Bezug auf eine bestimmte Person hergestellt werden kann. Dies wäre beispielsweise der Fall, wenn den gesammelten Informationen ein Vermerk über die Uhrzeit dieser Befragung

beigefügt und diese Uhrzeit ebenfalls bei der angerufenen Telefonnummer, z. B. zu Abrechnungszwecken gespeichert wird. Durch diese Verknüpfung wäre es möglich, den Bezug zu einer bestimmten Person herzustellen, und die Daten damit personenbezogen.

Beispiel 3: Bei der Datenerhebung über eine Onlinebefragung sind die Grundsätze, die für die telefonische Befragung gelten, analog anzuwenden: Die erhobenen Informationen sind anonym, wenn sie ohne Identifikationsmerkmale, wie z. B. Name oder Adresse erhoben werden, entsprechend aggregiert sind und kein Bezug zur IP-Adresse besteht. Die Internetseite sollte zudem vom Einsatz sogenannter Cookies absehen, die Informationen über das sonstige Surfverhalten der Person sammeln (dazu ausführlich Logemann 2014). Diese Kriterien erfüllte beispielsweise eine Dunkelfeldstudie des Vereins MOGiS e. V. zum Thema sexueller Missbrauch. Die Probanden füllten ein Onlineformular ohne Namensnennung aus. IP-Adressen wurden nicht gespeichert. Außerdem waren der Abbruch der Befragung und auch eine selektive Beantwortung der Fragen möglich. Die Probanden konnten dadurch gegebenenfalls selbst entscheiden, ob die Kombination der erteilten Informationen einen zu engen Bezug zu ihrer Person herstellen könnte (MOGiS e. V. 2011).

2.4 Hinweise für die Planungsphase von Forschungsprojekten

Der Umgang mit personenbezogenen Daten ist zahlreichen datenschutzrechtlichen Restriktionen unterworfen. In der Planungsphase eines Forschungsvorhabens sollten die Forschenden deshalb sorgfältig abwägen, ob personenbezogene Daten tatsächlich für den Erfolg des Forschungsvorhabens erforderlich sind oder das Vorhaben auch mit anonymen Daten durchgeführt werden kann.

Gelangen die Forschenden zu der Einschätzung, dass das Vorhaben nicht ohne personenbezogene Daten umgesetzt werden kann, dürfen die Daten jedoch nur so sparsam wie möglich erhoben werden. Anschließend sind sie so rasch wie möglich zu anonymisieren (§ 3a S. 1 BDSG).

Es ist immer der Weg zu wählen, der aus Sicht des Probanden den schonendsten Umgang mit seinen Daten bietet. Das kann auch bedeuten, dass seine Daten zunächst so anonymisiert werden, dass dies den Bezug zu seiner Person noch nicht so erschwert, dass dieser nur mit einem unverhältnismäßigen Aufwand herzustellen wäre, die Daten also *noch* nicht faktisch anonymisiert sind. Wenn auch eine abgeschwächte Anonymisierung nicht in Betracht kommt, ohne dass der Forschungszweck gefährdet wird, dann muss das Forschungsteam die Daten zumindest so weit wie möglich pseudonymisieren. Das For-

schungsteam ist dann aber weiterhin an die Vorgaben des jeweils einschlägigen Datenschutzrechts gebunden.

Da die Restriktionen des Datenschutzrechts nur so lange gelten, wie die Daten personenbezogen sind, hat also eine möglichst frühe Anonymisierung Vorteile. Denn sind die Daten anonymisiert, können die Informationen frei verwendet werden.

2.5 Anonymisierungs- und Pseudonymisierungstechniken

Die Möglichkeiten, bereits zum Zeitpunkt der Erhebung einen Personenbezug bei den Forschungsdaten zu vermeiden, sind je nach Erhebungsmethoden grundsätzlich breit gefächert:[2]

Beispiel 1: Beim Adressmittlungsverfahren wird die Verknüpfung konkreter Informationen mit einem vorhandenen Adressenbestand verhindert. Dies geschieht wie folgt: Es gibt eine Stelle, die über einen Adressbestand für das Forscherteam interessanter Personen verfügt. Das könnte für den Fall von Dunkelfeldstudien z. B. der Weiße Ring e. V. mit seiner Adressdatei sein. Die Stelle gibt ihre Adressen jedoch nicht weiter, sondern wird als Adressmittler tätig.

Das Forscherteam verfasst nun einen Fragebogen, der nach den oben erläuterten Maßstäben nicht geeignet ist, Rückschlüsse auf eine einzelne Person zuzulassen. Dazu müssen Frage- und Antwortkategorien so ausgestaltet sein, dass keine hervorstechenden Merkmale abgefragt werden, z. B. durch vorgegebene Multiple-Choice-Antworten. Natürlich darf der Fragebogen selbst kein Adress- oder Unterschriftenfeld o. Ä. vorsehen. Der Adressmittler verschickt den Fragebogen an die Personen in seiner Adressdatei mit einem Anschreiben, in dem das Vorgehen umfassend erläutert wird. Als Absender sollte die Adressmittlerstelle fungieren, um zu verhindern, dass etwa unzustellbare Briefe an das Forscherteam gesandt werden. Die ausgefüllten Fragebögen schicken die Probanden kuvertiert in einem beigefügten Rückumschlag ohne Absenderangabe zurück (dazu auch Metschke/Wellbrock 2002, 34).

Das Verfahren kann analog auf die Versendung per E-Mail oder den Gebrauch von Telefonnummern angewendet werden.

Beispiel 2: Für einige Dunkelfeldstudien kann sich die Randomized-Response-Technik anbieten. Sie stellt die Anonymität der Befragten in Umfragen si-

---

2 Hierzu auch ausführlich Häder 2009, 12 ff.

cher, indem – ebenso wie bei der Adressmittlermethode - bereits zum Zeitpunkt der Erhebung auf personenbezogene Daten verzichtet wird. Das ist dann besonders wichtig, wenn den Probanden Fragen gestellt werden, die sie als unangenehm empfinden, wie z. B. die Frage, ob jemand schon einmal einen Diebstahl begangen habe. Der größte Teil der Probanden wird diese Frage nur dann wahrheitsgemäß beantworten, wenn seine Anonymität gesichert ist. Dies wird gewährleistet, indem z. B. ein Würfel oder eine Münze entscheidet, ob der Befragte gebeten wird, ehrlich auf die kritische Frage zu antworten, oder ob er – unabhängig vom Frageninhalt – nur „Ja“ antworten soll. Der Ausgang des Zufallsexperiments ist dem Interviewer nicht bekannt. Er weiß also nicht, ob eine individuelle „Ja“-Antwort z. B. durch die Münze vorgegeben war oder der Proband damit das Begehen eines Diebstahls zugegeben hat. Mithilfe statistischer Verfahren kann aber der tatsächliche Anteil der Personen bestimmt werden, der auf die kritische Frage mit „Ja“ geantwortet hat. Studien belegen, dass durch dieses Verfahren kritische Verhaltensweisen jedenfalls häufiger eingestanden werden als ohne Gewährleistung der Anonymität (Musch 1999; ausführlich zum Ganzen: Ostapczuc 2008).

Wenn der Forschungszweck nicht ohne die Erhebung personenbezogener Daten erreicht werden kann, so verlangt das Datenschutzrecht, dass der Personenbezug jedenfalls so rasch wie möglich entfernt wird, die Daten also nach der Erhebung anonymisiert werden. In aller Regel sind mehrere Methoden in Kombination miteinander erforderlich, um das Stadium der faktischen Anonymisierung zu erreichen. Welche Methodenkombination beim jeweiligen Forschungsdesign zur Anwendung kommen sollte, richtet sich nach den konkreten Gegebenheiten der Studie (Metschke/Wellbrock 2002, 40 ff.). Beispiele für Anonymisierungsmethoden sind:

Aggregation: Unter Aggregation versteht man die Zusammenfassung mehrerer Einzelgrößen hinsichtlich eines gleichartigen Merkmals, um Zusammenhänge zu gewinnen (Weischer 2015a, 14). Die oben (2.3 – Beispiel 1) erläuterte Zusammenfassung aller Informationen der „Männer über 70“ stellt eine Aggregation dar, die zu einer Anonymisierung führen kann, wenn dadurch ein Rückschluss auf eine bestimmte Person ohne unverhältnismäßigen Aufwand nicht mehr möglich ist.

Zufallsfehler: Eine weitere Methode ist das Einstreuen von Zufallsfehlern. Dabei werden bei einer geringen Anzahl zufällig ausgewählter Datensätze ein oder mehrere Merkmale automatisiert verändert. Diese Änderungen fallen bei einer hohen Datensatzanzahl statistisch kaum ins Gewicht. Es ist aber nun nicht mehr mit Sicherheit feststellbar, ob das Merkmal in einem konkreten Datensatz zutrifft oder nicht.

Stichproben: Infrage kommen auch Stichproben oder Sub-Stichproben. Eine Stichprobe ist eine Teilmenge einer Grundgesamtheit, die für eine Unter-

suchung ausgewählt wird (Weischer 2015b, 396). Wenn diese das Ergebnis einer nach Zufallsauswahl durchgeführten Teilerhebung ist, spricht man von einer Zufallsstichprobe. Da diese nur zufallsabhängig sind, können ihre Kenngrößen mit Methoden der Inferenzstatistik auf die Grundgesamtheit übertragen werden (Hochrechnung). Eine Zufallsstichprobe wird daher als repräsentativ für die Grundgesamtheit bezeichnet (Jann/Farys 2015, 448).

Entscheidend für die Frage, wann die Daten z. B. mithilfe der dargestellten Techniken ausreichend anonymisiert sind, sind auch hier die oben dargestellten allgemeinen Grundsätze: Die Daten sind so lange personenbezogen, bis der Bezug zu einer Person nur noch mit unverhältnismäßig großem Aufwand hergestellt werden kann. Die stetige Weiterentwicklung des technischen Fortschritts verändert auch immer wieder die konkreten technischen Ausprägungen der Verfahren. Die Grundsätze bleiben jedoch unabhängig vom aktuellen Stand der Technik anwendbar.

Bei der Pseudonymisierung gibt es ebenfalls mehrere Verfahren. Allerdings unterscheiden diese sich jeweils nur danach, wer die Schlüsselregel generiert und nach welchem Schema diese erstellt wird: Die Codierung kann entweder durch die Probanden selbst, einen vertrauenswürdigen Dritten oder auch durch die Forscher erfolgen (Schaar 2014, 98 f.). Für die Auswahl der Zuordnungsvorschrift gibt es ebenfalls zahlreiche Methoden, häufig werden mathematische Algorithmen angewandt (Metschke/Wellbrock 2002, 19 f.).

3 Abgrenzung Bundes- und Landesdatenschutzgesetze

Kommt das Forschungsteam zu der Einschätzung, dass die Erhebung personenbezogener Daten für das Projekt unerlässlich ist, muss im nächsten Schritt festgestellt werden, welches der Datenschutzgesetze im konkreten Fall einschlägig ist. Dies hängt davon ab, wer an der Studie maßgeblich beteiligt ist. Dafür ist zunächst zu bestimmen, ob es sich um eine öffentliche Stelle oder eine nicht öffentliche Stelle im Sinne der Datenschutzgesetze handelt.

3.1 Datenerhebung und Verarbeitung durch öffentliche Stellen

Für öffentliche Stellen ist entscheidend, ob diese dem Bund oder einem Bundesland zuzuordnen sind. Öffentliche Stellen des Bundes sind Behörden, Organe der Rechtspflege und andere öffentlich-rechtlich organisierte Einrichtungen des Bundes, bundesunmittelbare Körperschaften, Anstalten und Stiftungen des öffentlichen Rechts sowie deren Vereinigungen ungeachtet ihrer Rechtsform (§ 2 Abs. 1 BDSG). Für sie gilt grundsätzlich das Bundesdatenschutzgesetz (§ 1 Abs. 2 BDSG), es sei denn, ein Spezialgesetz sieht

ebenfalls Regelungen zur Datenerhebung oder -verarbeitung für diese Stelle vor. Ist dies der Fall, wie z. B. bei der Datenverarbeitung durch das Bundeskriminalamt (BKA), dann sind diese speziellen Regelungen vorrangig und die Regelungen des BDSG gelten nur subsidiär.

Ist die öffentliche Stelle dagegen einem Land zuzuordnen,[3] gelten die Regelungen des entsprechenden Landesdatenschutzgesetzes (LDSG). Voraussetzung ist allerdings auch hier, dass keine vorrangige landesgesetzliche Spezialregelung, wie z. B. zur Datenerhebung und -verarbeitung durch die Landespolizeien,[4] einschlägig ist. Da die Mehrzahl der Hochschulen dem Landesrecht untersteht, sind für Forschungsprojekte häufig die Landesdatenschutzgesetze relevant.

3.2 Datenerhebung und Verarbeitung durch nicht öffentliche Stellen

Nicht öffentliche Stellen sind natürliche und juristische Personen, Gesellschaften und andere Personenvereinigungen des Privatrechts, soweit sie nicht als öffentliche Stellen zu qualifizieren sind.[5]

Für sie gelten die Regelungen des Bundesdatenschutzgesetzes, soweit sie die Daten unter Einsatz von Datenverarbeitungsanlagen verarbeiten, nutzen oder dafür erheben oder die Daten in oder aus nicht automatisierten Dateien verarbeiten, nutzen oder dafür erheben, es sei denn, die Erhebung, Verarbeitung oder Nutzung der Daten erfolgt ausschließlich für persönliche oder familiäre Tätigkeiten (§ 1 Abs. 2 Nr. 3 BDSG).

Der Begriff der Datenverarbeitungsanlagen ist dabei weit auszulegen. Erfasst werden neben der klassischen automatisierten Verarbeitung auch Datenerhebungen, die beispielsweise zunächst durch handschriftliche Aufzeichnungen fixiert werden, soweit eine spätere automatische Verarbeitung aus objektiver Sicht intendiert wird. Ob diese tatsächlich zur Anwendung kommt, ist nicht relevant. Auch die Nutzung muss nicht unmittelbar mit der Funktion einer Datenverarbeitungsanlage verbunden sein. Für die Verarbeitung von Daten in oder aus nicht automatisierten Dateien gelten im Ergebnis keine Einschränkungen in Bezug auf den Anwendungsbereich des BDSG, da es genügt, dass die Daten von einer nicht öffentlichen Stelle überhaupt in oder aus einer Datei verarbeitet oder genutzt werden (Dammann 2014, § 1 Rn. 140 ff.).

---

3 Vgl. z. B. § 2 LDSG Baden-Württemberg.

4 Vgl. z. B. §§ 19 ff. PolG Baden-Württemberg.

5 Vgl. § 2 Abs. 4 BDSG mit Bezugnahme auf § 2 Abs. 1–3 BDSG.

Nicht anwendbar sind datenschutzrechtliche Bestimmungen dagegen auf ausschließlich für objektive persönliche oder familiäre Tätigkeiten angelegte Datensammlungen (Dammann in Simitis 2014, § 1 Rn. 147 ff.). Für Forschungsprojekte bleibt diese Regelung ohne Auswirkung, da sie in aller Regel nicht nur rein privaten Interessen dienen.

3.3 Erhebung personenbezogener Daten durch beauftragte Dritte

Bei zahlreichen Forschungsprojekten werden die Daten nicht durch die verantwortliche Stelle, d. h. durch das Forschungsinstitut, selbst erhoben. Stattdessen werden Dritte wie beispielsweise kommerzielle Umfrageinstitute damit betraut, die für das Forschungsprojekt notwendigen personenbezogenen Daten zu erheben.

Die Datenschutzgesetze von Bund und Ländern[6] stellen bei solchen Konstellationen sicher, dass die im konkreten Fall das Vorhaben initiierende Stelle auch weiterhin die Verantwortung trägt. Diese Stelle, wie z. B. das Forschungsinstitut, muss auch im Fall der Auftragserteilung sicherstellen, dass das geltende Datenschutzrecht eingehalten wird. Die jeweils einschlägigen Datenschutzgesetze von Bund oder Ländern beschreiben deshalb im Detail, welche Rechte, Pflichten und Maßnahmen im Einzelnen vertraglich zwischen Auftraggeber und Auftragnehmer zu treffen sind.

Für die Auftragnehmer gelten die allgemeinen Datenschutzregelungen, die bei der Erhebung personenbezogener Daten zu beachten sind.[7]

4 Abgrenzung Datenschutz und Datensicherheit

Die Datenschutzgesetze regeln zum einen den rechtlichen Rahmen für die Erhebung und den weiteren Umgang mit personenbezogenen Daten. Dabei geht es darum, ob diese rechtmäßig erhoben, verarbeitet oder anderweitig genutzt wurden. Die datenerhebenden und -verarbeitenden Stellen werden aber zum anderen auch dazu verpflichtet, die technischen und organisatorischen Maßnahmen zu treffen, die erforderlich sind, um den Schutz der Daten zu gewährleisten.[8]

---

6 Zum Beispiel § 11 BDSG, § 7 LDSG BW, § 4 LDSG RP.

7 Siehe dazu ausführlich Kapitel 5.

8 Für Verarbeitungen im Bereich des BDSG § 9 S. 1 BDSG. Die im Gesetzestext genannte Anlage enthält Hinweise zu den technischen und organisatorischen Vorkehrungen, die zum Schutz personenbezogener Daten zu treffen sind.

Dabei sollen personenbezogene Daten technisch vor unbefugtem Zugriff Fremder geschützt werden. Das bedeutet für ihre Aufbewahrung beispielsweise, dass die Räume oder Computersysteme, in denen sie sich befinden, gesichert sein müssen. Die Sicherung muss geeignet sein zu verhindern, dass Personen, die nicht in das Forschungsprojekt involviert sind, auf diese Daten zugreifen können. Die Räume oder Computersysteme müssen also mit bestimmten Zugangsbarrieren versehen werden. Soweit es um diesen technisch-organisatorischen Schutz der Daten geht, spricht man von der Datensicherheit.

Über den Begriff der *erforderlichen* Maßnahmen soll sichergestellt werden, dass der Umfang der notwendigen Sicherheitsmaßnahmen die Grundsätze der Verhältnismäßigkeit wahrt. Die Entscheidung, welche Maßnahmen verhältnismäßig sind oder nicht, ist durch die Abwägung aller Einzelfallumstände zu treffen (dazu auch Schneider 2011, 6 ff.): Dabei können unter anderem finanzielle Aspekte eine Rolle spielen, aber auch die Art der aufbewahrten Informationen ist relevant. Ein Indikator für die Notwendigkeit relativ hoher Sicherheitsvorkehrungen wäre beispielsweise der Umgang mit als überdurchschnittlich sensibel einzustufenden personenbezogenen Daten.[9] Zur Einordnung, welche Vorkehrungen nach dem Verhältnismäßigkeitsprinzip erforderlich wären, kann das folgende *Beispiel* dienen:

Eine private Arbeitsvermittlung versandte Profile ihrer Mandanten an potenzielle Arbeitgeber per E-Mail. Arbeitnehmer-Profile bzw. Lebensläufe sind personenbezogene Daten, die als durchschnittlich sensibel einzustufen sind. Der Datenschutzbeauftragte hatte den Arbeitsvermittler aufgefordert, die E-Mails zu verschlüsseln oder zu pseudonymiseren. Die Pseudonymisierung bedeutet für die potenziellen Arbeitgeber aber einen Verlust von Informationen. Nur bei einer verschlüsselten Übersendung hätten die potenziellen Arbeitgeber die vollständigen Informationen einsehen können. Bei der verschlüsselten Versendung von E-Mails müssen sowohl Absender als auch Empfänger über bestimmte aufeinander abgestimmte technische Voraussetzung verfügen. Das bedeutet, dass der Arbeitsvermittler vor dem Versenden sicherstellen müsste, dass der Empfänger über die entsprechende Technik verfügt. Falls der Empfänger diese Technik nicht vorrätig hätte, müsste der Arbeitsvermittler die Technik auf seine Kosten zur Verfügung stellen. Der Arbeitsvermittler klagte gegen diese Entscheidung des Datenschutzbeauftragten beim Verwaltungsgericht Berlin. Das Gericht gab dem Arbeitsvermittler Recht, weil die Kosten für ihn im Einzelfall nicht zumutbar waren und auch

[9] Vgl. z. B. § 3 Abs. 9 BDSG: Besonders schützenswert sind Angaben über ethnische Herkunft, politische Meinungen, religiöse oder philosophische Überzeugungen, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben.

der Aufwand, jedem potenziellen Kunden eine entsprechende Entschlüsselungssoftware zur Verfügung zu stellen oder zu installieren, im konkreten Fall unverhältnismäßig wäre (VG Berlin, 2011).

Für ein Forschungsvorhaben bedeutet das Folgendes: Die Forschenden müssen einschätzen, wie sensibel die personenbezogenen Daten sind, die sie erheben. Sollen z. B. Opfer schwerer Straftaten intime Informationen offenbaren, handelt es sich um überdurchschnittlich sensible Daten. Die Informationen haben zugleich einen gesteigerten Nutzwert für potenzielle Datenangreifer. In einer solchen Konstellation wären relativ hohe technische und organisatorische Sicherheitsstandards notwendig.

5 Zulässigkeit der Erhebung personenbezogener Daten

Grundsätzlich kommen nur zwei Möglichkeiten einer zulässigen Erhebung personenbezogener Daten in Betracht. Entweder ist die Erhebung personenbezogener Daten *durch eine Rechtsgrundlage ausdrücklich erlaubt* oder der *Betroffene, d. h. der Proband, willigt in die Erhebung seiner personenbezogenen Daten wirksam ein* (§ 4 Abs. 1 BDSG). Dies gilt unabhängig davon, ob eine öffentliche oder eine nicht öffentliche Stelle tätig wird und ob landes- oder bundesrechtliche Regelungen eingreifen.

5.1 Erhebung personenbezogener Daten aufgrund einer Einwilligung

Für die Dunkelfeldforschung dürften nur wenige Datensammlungen verfügbar sein, die auf einer gesetzlichen Grundlage bzw. einer gesetzlich verankerten Auskunftspflicht der Probanden basieren.[10] Bei entsprechenden Forschungsprojekten muss die Datengrundlage in der Regel einwilligungsbasiert erhoben werden, ist also so zu gewinnen, dass Fragen an Probanden gerichtet werden und diese zuvor in die Befragung eingewilligt haben. Diese Einwilligung ist nur dann wirksam, wenn der Proband sie freiwillig erteilt und alle relevanten Umstände der Verarbeitung kennt (§ 4a Abs. 1 S. 1 BDSG). Der Proband darf also einerseits nicht in einer Zwangslage stecken oder eine Sanktion befürchten, wenn er seine Einwilligung verweigert.

Er muss andererseits Kenntnis über alle *relevanten Umstände* haben. Dazu gehören beispielsweise die Information, wer verantwortlicher Träger und Leiter des Forschungsvorhabens ist, wofür genau die Daten gebraucht, wie sie

[10] Für das „Hellfeld“ basieren die Datensammlungen z. B. auf den Erhebungsnormen der Polizeigesetze, die in den polizeilichen Kriminalstatistiken dargestellt werden.

weiter verwendet und wann sie gelöscht oder anonymisiert werden (dazu ausführlich Metschke/Wellbrock 2002, 26). Diese Informationen sollten einerseits möglichst detailliert, andererseits aber auch so übersichtlich wie möglich für den Probanden sein. Für den Betroffenen steht dabei die Frage im Mittelpunkt, ob die Daten möglicherweise zu seinem Nachteil verwendet werden könnten.[11]

Die Datenschutzgesetze schreiben als Regelfall vor, dass der Proband schriftlich in seine Befragung einwilligt. Dies bedeutet, dass der Einwilligende ein Schriftstück, in dem alle relevanten Umstände niedergelegt sind, eigenhändig unterschreibt (§ 4a Abs. 1 S. 3 BDSG i. V. m. § 126 BGB). Das hat zwei Gründe: Erstens soll sichergestellt werden, dass die einwilligende Person auch tatsächlich mit dem Befragten übereinstimmt (*Garantiefunktion*), und zweitens ist mit einer Unterschrift stets auch ein Reflexionsprozess beim Unterzeichner verbunden. Dem Befragten soll damit bewusst gemacht werden, dass er über seine personenbezogenen Daten verfügt (*Warnfunktion*).

Die Einwilligung kann aber beispielsweise auch per E-Mail erteilt werden. Dazu wäre dem Befragten zuvor ein entsprechendes Dokument zu übermitteln, in dem alle relevanten Umstände für die Erhebung und Verarbeitung beschrieben sind. Voraussetzung dafür, dass die Einwilligung per E-Mail ausreichend ist, wäre aber, dass der Proband über eine qualifizierte elektronische Signatur verfügt (§ 126a BGB).[12] Grund für dieses Erfordernis ist wiederum die Funktion der Unterschrift: Ebenso wie bei einem Schriftstück muss auch bei einer elektronischen Unterschrift sichergestellt sein, dass es sich tatsächlich um den Unterzeichner handelt. Würde man beispielsweise eine eingescannte Unterschrift genügen lassen, wäre die Missbrauchsgefahr zu groß.

Von der Regel des Schriftformerfordernisses sind Ausnahmen möglich. Die wissenschaftliche Forschung erfährt insoweit eine Privilegierung, als sie vom Schriftformerfordernis befreit ist, wenn dieses den konkreten Forschungszweck *erheblich* beeinträchtigen würde (§ 4a Abs. 2 S. 1 BDSG).

Um festzustellen, ob im konkreten Fall eine erhebliche Beeinträchtigung vorliegt, sind alle relevanten Einzelfallumstände mit- und gegeneinander abzuwägen. Eine erhebliche Beeinträchtigung der Forschung wird jedenfalls nicht vorliegen, wenn durch das Schriftformerfordernis lediglich zusätzlicher bürokratischer Aufwand oder Kosten entstehen.

---

11 Dazu ausführlich 7.3.

12 Weitere Voraussetzungen finden sich im Signaturgesetz (Gesetz über Rahmenbedingungen für elektronische Signaturen, kurz SigG oder SigG 2001). Die elektronische Signatur ist jedoch wenig verbreitet und vergleichsweise kostspielig in der Anschaffung.

Für Forschungsvorhaben bedeutet dies zunächst, dass im Rahmen einer Face-to-Face-Befragung das Unterschriftenerfordernis auch für die wissenschaftliche Forschung obligatorisch ist. Denn hier ist der Aufwand für eine schriftliche Einwilligung überschaubar und damit auch verhältnismäßig.

Ob personenbezogene Daten für Forschungsprojekte zulässigerweise über Telefoninterviews unter Umgehung des Schriftformprinzips erhoben werden können, entscheidet die Abwägung aller im konkreten Fall relevanten Umstände.

Aus Forschersicht dürfte vor allem der Kostenfaktor eine Rolle für die Entscheidung zur Durchführung der im Vergleich zur Face-to-Face-Befragung günstigeren Telefoninterviews spielen. Andererseits ist zu berücksichtigen, dass es bei Dunkelfeldbefragungen für die Probanden besonders wichtig ist, genau zu wissen, wofür und inwieweit ihre Daten weiter verwendet werden (Simitis 2014, § 4a Rn. 61; ausführlich hierzu auch: Häder 2009, 23 ff.). Die ausschließlich mündliche Aufklärung am Telefon ist im Vergleich zur Vorlage eines Schriftstücks aus Sicht des Probanden stets etwas weniger transparent. Daher sollte bei Telefoninterviews der Versuch unternommen werden, durch ausgleichende Maßnahmen eine Situation zu erzeugen, die für den Probanden mit der Face-to-Face-Befragung vergleichbar ist.

Beispiel: Das Erheben personenbezogener Daten über Telefoninterviews kann beispielsweise unter folgenden Bedingungen zulässig sein: Der Betroffene wird zuvor ausführlich über die Umstände aufgeklärt. Dies wird in der Interviewniederschrift vermerkt. Es werden keine Daten erhoben, die als überdurchschnittlich sensibel einzustufen sind, und der Personenbezug wird nach der Erhebung so rasch wie möglich entfernt. Positiv auswirken könnte sich zudem die Einrichtung einer Homepage mit Informationen über alle relevanten Informationen, auf die der Interviewer beim Erstkontakt verweist. Es ist allerdings wichtig, dass der Verweis *vor* der Befragung erfolgt und der Proband *vor* seiner Einwilligung die Möglichkeit hat, sich entsprechend zu informieren.

Für Onlinebefragungen gelten vergleichbare Grundsätze. Allerdings dürfte es in dieser Konstellation einfacher sein, den Probanden alle notwendigen Informationen in Textform zugänglich zu machen. Die Internetseite oder die Einführung sollte von den Forschern so ausgestaltet sein, dass alle Informationen verständlich und übersichtlich dargestellt sind. Informationen zu den relevanten Umständen sollten deshalb nicht z. B. in weiterführenden Links verborgen sein, weil dies erfahrungsgemäß dazu führt, dass ein großer Teil der Probanden diesen Schritt aus Bequemlichkeit unterlässt (ADM-Richtlinien 2007).

Fehlt eine erforderliche Komponente der Einwilligung, so ist die Erklärung nichtig (§ 125 BGB). Sie ist beispielsweise nicht vollständig, wenn der Pro-

band nicht alle relevanten Informationen bekommen, er sie zu spät erhalten oder seine Einwilligung nicht schriftlich erteilt hat, ohne dass ein oben beschriebener Ausnahmefall vorlag. Die Einwilligung darf aber auch keine Bedingung für eine andere Leistung oder etwa ein Geschenk sein, das der Proband nur erhält, wenn er seine Einwilligung erteilt. Solche Fehler haben für das Forschungsvorhaben zur Folge, dass die Angaben der Probanden nicht verwendet werden dürfen und umgehend vernichtet werden müssen.

Ob auch Kinder oder Jugendliche wirksam ihre Einwilligung in die Erhebung oder Verarbeitung ihrer Daten erteilen können, ist nicht abschließend geklärt. Einige stellen auf die zivilrechtlichen Regelungen zu Rechtsgeschäften ab, die die Einwilligung des gesetzlichen Vertreters verlangen. Vielfach wird auf die Einsichtsfähigkeit der Minderjährigen abgestellt, die ab einem Alter von 14 Jahren in vielen Fällen vermutet werden kann. Angesichts des Umstands, dass die Verwertbarkeit der Ergebnisse einer Studie in erster Linie von der Repräsentativität der Datengrundlage abhängt, sollte aus Gründen der Vorsicht und zur Steigerung der Akzeptanz jedoch die vorherige Zustimmung der gesetzlichen Vertreter eingeholt werden. So scheiterte etwa eine Dunkelfeldstudie an Berliner Schulen, weil der Landeselternausschuss erfolgreich die fehlende Einwilligung der Eltern bemängelte (Peiritsch 2010).

5.2 Erhebung personenbezogener Daten aufgrund einer Rechtsgrundlage

Im Datenschutzrecht gibt es keine Vorschriften, die speziell für die wissenschaftliche Forschung eine eigene Erhebungsgrundlage vorsehen. Weder § 13 Abs. 1 BDSG (für öffentliche Stellen des Bundes) noch § 28 BDSG (für nicht öffentliche Stellen) ist eine eigene Erhebungsgrundlage, sondern sie regeln nähere Ausführungen zur Zulässigkeit (Sokol 2014, § 13 Rn. 7; Simitis 2015, § 28 Rn. 5). Es gelten damit für die Erhebung personenbezogener Forschungsdaten die allgemeinen Grundsätze.

Wie unter 5.1 bereits ausgeführt basieren Datenerhebungen der kriminologischen bzw. sozialwissenschaftlichen Dunkelfeldforschung aber überwiegend auf einer Einwilligung der Probanden. Es sind Konstellationen denkbar, bei denen die Antworten der Probanden zwangsläufig auch Informationen über eine weitere Person preisgeben und diese dritte Person aufgrund der Informationen identifizierbar wird.

Beispiel: Es werden zu Forschungszwecken Opfer häuslicher Gewalt befragt. Durch bestimmte Fragen bzw. ihre Kombination, wie etwa nach dem Personenstand, der Häufigkeit von Übergriffen und dem Maß finanzieller Abhängigkeit vom Täter, kann auch der Täter zu einer bestimmbaren Person werden. Gibt der Proband z. B. bei dieser Fragenkonstellation an, dass er ver-

heiratet ist, die Übergriffe täglich stattfinden und er finanziell vom Täter abhängig ist, ist es sehr wahrscheinlich, dass es sich beim Täter um den Ehemann oder die Ehefrau handelt.

Sind die Fragen also so konstruiert, dass neben denen des Opfers auch personenbezogene Daten weiterer Personen erhoben werden, verlangt das Datenschutzrecht, dass *auch diese Erhebung* personenbezogener Daten *zulässig* sein muss. Eine Erhebung personenbezogener Daten ist zulässig, wenn sie entweder auf einer Einwilligung oder einer Rechtsgrundlage beruht.

Teilweise wird zwar vertreten, dass bei Daten, die zwangsläufig auch das nähere Umfeld des Betroffenen beschreiben (Doppelbezug) keine gesonderte Erhebungsgrundlage für die miterhobenen Informationen des Dritten notwendig sei (Gola 2015, § 4 Rn. 20). Dies kann aber grundsätzlich nur für Konstellationen gelten, in denen die ursprüngliche Erhebung auf einer Rechtsgrundlage beruht. Ansonsten wäre der Umgehung des Einwilligungserfordernisses Tür und Tor geöffnet, indem man über gezielte Fragen durch die Hintertür ganze Personenkreise des Betroffenen ausforscht.

Eine Einwilligung wird der Täter in solchen Fällen jedoch in der Praxis nur selten erteilen. Wenn personenbezogene Daten des Täters oder einer anderen Person aus dem Umfeld des Opfers erhoben werden, dann wird die Erhebung deshalb in den meisten Fällen nur durch eine einschlägige Norm zu rechtfertigen sein. Da zumindest der Täter zudem an der Datenerhebung nicht mitwirkt, muss diese Norm explizit auch eine Datenerhebung erlauben, die ohne Mitwirkung oder Kenntnis des Betroffenen erfolgt. An die Zulässigkeit einer solchen Erhebung knüpft das Datenschutzrecht aber strenge Voraussetzungen: So darf sie beim Betroffenen selbst nicht nur einen unverhältnismäßigen Aufwand bedeuten und es dürfen keine Anhaltspunkte dafür bestehen, dass seine überwiegenden schutzbedürftigen Interessen beeinträchtigt werden (§ 4 Abs. 2 Nr. 1 oder 2 BDSG). Sein Interesse wäre z. B. dann beeinträchtigt, wenn die Gefahr bestünde, dass seine personenbezogenen Daten im Zuge des Forschungsvorhabens zulässigerweise zweckentfremdet werden.[13]

Werden Daten eines Dritten ohne seine Mitwirkung auf diesem Wege erhoben, so ist dieser in der Regel nachträglich über die Erhebung seiner personenbezogenen Daten zu unterrichten. Ausnahmen sind allerdings möglich, wenn damit ein unverhältnismäßiger Aufwand einhergehen würde.[14]

---

13 Zur Frage, wann eine zulässige Zweckentfremdung vorliegen und wie diese ggf. verhindert werden könnte, ausführlich Kapitel 7.

14 § 19a Abs. 2 Nr. 2 oder § 33 Abs. 2 Nr. 5 BDSG.

6 Übermittlung von Daten an die wissenschaftliche Forschung

An der *Übermittlung* personenbezogener Daten aus den Datenvorräten *öffentlicher Stellen* dürfte die Dunkelfeld-Opferforschung nur in wenigen Konstellationen ein Interesse haben.

Dagegen könnten personenbezogene Daten einiger nicht öffentlicher Stellen, wie beispielsweise der Opferschutzorganisation der Weiße Ring e. V., als Datenbasis für die Dunkelfeldforschung interessant sein. Die Übermittlung dieser Datenbestände an die Forschenden müsste aber wiederum nach den allgemeinen Grundsätzen zulässig, d. h. entweder eine Einwilligung der Betreffenden oder durch eine Rechtsgrundlage gedeckt sein.

Für die einwilligungsbasierte Übermittlung heißt das, dass alle infrage kommenden Probanden zuvor in die Übermittlung ihrer personenbezogenen Daten an die Forschenden und anschließend ein weiteres Mal in die eigentliche Befragung einwilligen müssten.

Ohne Einwilligung des Betroffenen wäre die Übermittlung personenbezogener Daten an die Forschenden nur zulässig, wenn die Daten seitens des Übermittelnden nicht unter den Schutz eines Berufsgeheimnisses fallen und die Übertragung im Interesse der Forschenden zur Durchführung wissenschaftlicher Forschung erforderlich ist. Zusätzlich müsste das wissenschaftliche Interesse an der Durchführung des Forschungsvorhabens das Interesse des Betroffenen an der ursprünglichen Verarbeitungsintention erheblich überwiegen. Eine weitere Bedingung wäre, dass der Zweck des Forschungsvorhabens auf andere Weise nicht oder nur mit unverhältnismäßigem Aufwand erreicht werden kann (z. B. § 28 Abs. 2 Nr. 3 BDSG). Für die Entscheidung sind alle betroffenen Belange mit- und gegeneinander abzuwägen. Zugunsten der Probanden ist gegebenenfalls ihr gesteigertes Interesse an der Wahrung ihrer Anonymität zu berücksichtigen. Sofern durch die zulässige Übermittlung außerdem die Gefahr erhöht wird, dass die Dunkelfelddaten zum Nachteil der Probanden verwendet werden,[15] so ist auch dieser Umstand im Rahmen der Abwägung als schutzwürdiges Interesse des Betroffenen einzustufen.

7 Verarbeitung personenbezogener Daten durch die Forschenden

Welche Verarbeitungsbedingungen für die personenbezogenen Daten nach ihrer Erhebung gelten, hängt maßgeblich davon ab, *zu welchem Zweck die Daten erhoben wurden* und wer als verantwortliche Stelle fungiert.[16] Grundsätz-

---

15 Dazu ausführlich Unterkapitel 7.3.

16 Dazu ausführlich Kapitel 2.

lich gilt für die Verarbeitung, dass *entweder die Rechtsgrundlage* der Erhebung oder die *Einwilligung* die konkreten Vorgaben für die weitere Verarbeitung festlegen (vgl. z. B. § 4 Abs. 1 BDSG).

Da das Grundgesetz die Tätigkeit von Wissenschaft und Forschung besonders privilegiert (Art. 5 Abs. 3 GG), wird die Verarbeitung zu diesem Zweck in zahlreichen Vorschriften begünstigt. Der Anwendungsbereich dieser Forschungsklauseln variiert je nachdem, ob die verantwortliche Forschungseinrichtung eine öffentliche oder eine nicht öffentliche Stelle ist. Für nicht öffentliche Stellen gilt mangels Gesetzgebungskompetenz der Länder stets Bundesrecht.

7.1 Forschungsklausel des BDSG

Soweit es sich bei der konkreten Forschungseinrichtung um eine öffentliche Stelle handelt, greift die Forschungsklausel des BDSG ohne Rücksicht auf die Verarbeitungsform ein (§ 40 Abs. 1 i. V. m. § 1 Abs. 2 Nr. 1 BDSG). Auf nicht öffentliche Stellen ist die Regelung dagegen mit den geringen Einschränkungen anwendbar, die die allgemeinen Regelungen für die Anwendbarkeit des Datenschutzrechts vorsehen.[17]

Der Anwendungsbereich der Forschungsklausel des BDSG beschränkt sich allerdings ausschließlich auf Forschungsprojekte, die von einer dem Bundesrecht unterliegenden Stelle initiiert und durchgeführt werden. Da der Bereich der Hochschulforschung in den meisten Fällen der Landesgesetzgebung untersteht, ist die bundesgesetzliche Regelung deshalb nur für Forschungsprojekte der Hochschulen und öffentlichen Stellen des Bundes oder auf die privatrechtlich organisierte wissenschaftliche Forschung anwendbar.

Die Forschungsklausel ermächtigt nicht zur Erhebung, sondern regelt ausschließlich die weitere Verarbeitung, Nutzung etc. Sie begrenzt den Radius der Verarbeitung und Nutzung auf die Zwecke der wissenschaftlichen Forschung und verpflichtet zu einer schnellstmöglichen Anonymisierung. Bis zur Anonymisierung sollen die Merkmale gesondert gespeichert werden, sodass Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person zugeordnet werden können. Auch im Rahmen dieser Regelung sieht der Gesetzgeber zur Entscheidungsfindung vor, dass die widerstreitenden Interessen im Einzelfall miteinander abgewogen werden.

---

17 Unterkapitel 3.2.

7.2 Forschungsklauseln der Landesdatenschutzgesetze

Die Forschungsklauseln der Landesdatenschutzgesetze bieten insgesamt kein einheitliches Bild, vor allem weil die Länder von ihrer Gesetzgebungskompetenz unterschiedlich Gebrauch gemacht haben. Die Forschungsklauseln sind dadurch in verschiedene Kontexte eingebettet. Die große Mehrheit beschränkt die zulässige Verwendung jedoch nicht generell auf die Zwecke der wissenschaftlichen Forschung, sondern spricht lediglich einzelne Verarbeitungsbedingungen an. Einige Landesdatenschutzgesetze wollen der wissenschaftlichen Forschung eigentlich nur die Verwendung anonymisierter oder pseudonymisierter Daten zugestehen, lassen aber gleichzeitig breite Ausnahmen von diesem Grundsatz zu[18] (zum Ganzen ausführlich Simitis in Simitis, 2014, § 40 Rn. 18 sowie Rn. 88 ff.).

Andere stellen zusätzliche Voraussetzungen für das Eingreifen der Privilegierung für die wissenschaftliche Forschung auf, indem mit der Unabhängigkeit der Forschungseinrichtung die Zulässigkeit der Verarbeitung an ein weiteres Kriterium geknüpft wird.[19] Wieder andere Klauseln beschränken die Verarbeitung im Kontext wissenschaftlicher Forschung auf bestimmte, d. h. einzelne konkretisierte Forschungsvorhaben.[20]

7.3 Schlussfolgerungen für die Verarbeitung personenbezogener Daten durch die Dunkelfeldforschung

Allen Forschungsklauseln – sowohl in den allgemeinen als auch in den besonderen Datenschutzgesetzen des Bundes- und Landesrechts – ist gemein, dass die Frage, ob eine Verarbeitung oder andere Nutzung im konkreten Fall möglich ist, durch eine Abwägung aller betroffenen Belange im Einzelfall entschieden wird. Für die Forschenden hat das den Nachteil, dass jede Entscheidung im Einzelfall mit vielen Unsicherheiten belastet ist.

In der Abwägung haben alle Umstände des konkreten Falls Berücksichtigung zu finden, soweit sie für die Ausübung der betroffenen Grundrechte relevant sind.

Bei einer Dunkelfeldstudie ist zugunsten der Forschenden z. B. der Umstand zu berücksichtigen, dass bei Befragungen im Dunkelfeld kaum auf vorhandene Datensammlungen zurückgegriffen werden kann. Zudem gestaltet sich bereits die Ermittlung der infrage kommenden Probanden im Dunkelfeld auf-

---

18 Vgl. die §§ 34 Abs. 1 DSG M-V; 28 Abs. 1 DSG NW; 22 Abs. 1 LDSG SH.

19 Wie beispielsweise § 35 Abs. 1 BW LDSG.

20 Wie beispielsweise § 28 BbgDSG oder § 33 Abs. 1 S. 1 HDSG.

wendig. Die Datengrundlage muss deshalb stets mit einem verhältnismäßig großen Aufwand beschafft werden.

Zugunsten der Probanden sind folgende Faktoren zu berücksichtigen: Eine besondere Rolle spielt hier die Frage, inwieweit bei dem jeweiligen Forschungsprojekt die Gefahr besteht, dass die zum Zweck der wissenschaftlichen Forschung erhobenen, übermittelten bzw. verarbeiteten Daten zulässigerweise zum Nachteil des Betroffenen verwendet werden. Der Gesetzgeber hat Zweckentfremdungen nicht per se ausgeschlossen, vielmehr sind sie durch eine besondere gesetzliche Vorschrift sogar möglich (z. B. § 4 Abs. 1 BDSG). Wenn ein Gesetz es also erlaubt, dass personenbezogene Daten für einen anderen als den ursprünglich intendierten Zweck verwendet werden, so ist dies zugunsten der Probanden zu gewichten. Sie sind in diesem Fall besonders schutzwürdig.

Eine gesetzliche Vorschrift, die zu einer Zweckentfremdung der Daten berechtigen kann, ist z. B. die Beschlagnahme durch die Ermittlungsbehörden nach §§ 94 ff. StPO. Da Datensammlungen im Dunkelfeld stets in einem engen Zusammenhang mit Straftaten stehen, dürften die Strafverfolgungsbehörden grundsätzlich an diesen Informationen interessiert sein.

Die Frage, ob eine Beschlagnahme nach §§ 94 ff. StPO tatsächlich zu einer Zweckentfremdung berechtigt, ist zwar nicht abschließend geklärt.[21] So lange die Möglichkeit einer zulässigen Zweckentfremdung der Daten z. B. über eine Beschlagnahme personenbezogener Probandendaten jedoch nicht ausgeschlossen werden kann, ist dieser Umstand im Rahmen der Abwägung als Risikofaktor für den Betroffenen insofern zu berücksichtigen, als die Durchführung eines Forschungsvorhabens nicht zu einem verdeckten Vehikel der Strafverfolgung umgedeutet werden darf.

Um dieses Risiko für die Probanden auszuschließen, sollte ein entsprechendes Forschungsvorhaben möglichst so konzipiert werden, dass die Gefahr einer zulässigen Zweckentfremdung von vornherein vermieden wird. Eine Zweckentfremdung personenbezogener Daten z. B. durch die Ermittlungsbehörden ist nicht möglich, wenn die Daten vom Schutz eines Berufsgeheimnisses erfasst werden. Mögliche Berufsgeheimnisträger sind beispielsweise Ärzte, Rechtsanwälte, Notare, Geistliche oder Psychotherapeuten.[22]

---

21 In einer Entscheidung (Az. 2 BvR 988/75 vom 24. 05. 1977) deutet das Bundesverfassungsgericht an, dass die Beschlagnahme im konkreten Fall nicht mit der Verfassung vereinbar war. Gleichzeitig geht aus den Verhältnismäßigkeitserwägungen aber klar hervor, dass die Entscheidung z. B. bei etwas schwereren Delikten möglicherweise anders ausgefallen wäre.

22 § 53 StPO enthält eine Aufzählung der Berufsgeheimnisträger.

Der Gesetzgeber hat bislang davon abgesehen, der wissenschaftlichen Forschung selbst einen entsprechenden Berufsgeheimnisschutz einzuräumen. Die Verleihung eines solchen Geheimnisschutzes ist nicht beliebig (dazu ausführlich BVerfGE 1972, 383 f.) und direkt aus der Verfassung ebenfalls nur in absoluten Ausnahmefällen möglich.

Eine zulässige Zweckentfremdung durch Ermittlungsbehörden ist vor allem in denjenigen Konstellationen besonders schwierig, in denen bei der Befragung des Opfers gleichzeitig auch personenbezogene Daten des Täters erhoben werden. Da sich der Proband bei einer Befragung im Rahmen eines Forschungsprojekts in Sicherheit wähnt und sich offenbar den staatlichen Ermittlungsbehörden gegenüber gerade nicht öffnen möchte, stellt eine solche Zweckentfremdung aus der Sicht des Betroffenen im Ergebnis eine Täuschung dar. Dies wiegt umso schwerer, als dies häufig eine Hintergehung des rechtsstaatlichen Prinzips des Aussageverweigerungsrechts für Täter und nahe Angehörige bedeuten würde. An dieser Stelle verläuft die Argumentation jedoch im Kreis: Mit der Annahme, dass die personenbezogenen Daten der bzw. des Betroffenen zulässigerweise auch zu seinem Nachteil verwendet werden könnten, wird das Ergebnis der Abwägung in der Regel sein, dass seine Interessen das der wissenschaftlichen Forschung überwiegen.

Für die sozialwissenschaftliche bzw. die kriminologische Forschung resultiert daraus, dass Sammlung und Verarbeitung personenbezogener Daten aus dem Dunkelfeld mit einem Restrisiko der Zweckentfremdung durch die Ermittlungsbehörden behaftet sein können. Dieses Restrisiko kann nur dann ausgeschaltet werden, wenn im Rahmen des Forschungsprojekts ein Berufsgeheimnis seine Schutzwirkung entfalten kann. Das Berufsgeheimnis kann aber nur dann zur Geltung kommen, wenn es sich im konkreten Fall um „berufsbezogene Tätigkeiten“ handelt.

Beispiel: Das ärztliche Berufsgeheimnis kommt bei der Studie und Präventionstherapie „Kein Täter werden“ der Charité zur Anwendung. Das Angebot richtet sich zwar primär an pädophil veranlagte Probanden, die noch nicht übergriffig geworden sind. In diesem Deliktsfeld gibt es jedoch zahlreiche Graubereiche, deren Kenntnis die Ermittlungsbehörden grundsätzlich interessieren dürfte.[23]

Es genügt aber nicht, willkürlich z. B. einen Mediziner mit der Verwahrung personenbezogener Daten zu betrauen. Das Gesetz sieht vielmehr vor, dass die Informationen dem betreffenden Geheimnisträger in seiner *beruflichen Eigenschaft* anvertraut worden sein müssen.

---

23 Siehe https://www.kein-taeter-werden.de/.

8 Veröffentlichung

Die Frage, ob und in welcher Form Forschungsergebnisse veröffentlicht werden dürfen, ist wiederum eine Abwägungsentscheidung. Es gilt der Grundsatz, dass keine überwiegenden Interessen des Betroffenen entgegenstehen dürfen.

Grundsätzlich sehen die Forschungsklauseln auch Ausnahmen für die Veröffentlichung personenbezogener Daten durch die Forschung vor. Diese betreffen jedoch Konstellationen, die im Fall von Dunkelfeldstudien kaum zutreffen dürften. So erlaubt z. B. § 40 Abs. 3 BDSG eine Veröffentlichung personenbezogener Daten, wenn der Betroffene eingewilligt hat oder dies im Rahmen der Darstellung von Forschungsergebnissen über Ereignisse der Zeitgeschichte unerlässlich ist.

Für Dunkelfeldstudien bedeutet dies, dass die Studienergebnisse nicht mit personenbezogenen Daten veröffentlicht werden dürfen, sondern so zu anonymisieren sind, dass Rückschlüsse auf konkrete Personen nicht gezogen werden können, ohne dass ein unverhältnismäßiger Aufwand erforderlich wäre.

9 Aufbewahrungs- und Löschungsfristen

Auch für die zulässige Dauer der Aufbewahrung und die Verpflichtung zur Löschung gilt der Grundsatz „nur so wenig personenbezogene Daten wie unbedingt nötig“. Für den Verlauf des Forschungsvorhabens bedeutet dies, dass die Daten fortlaufend so intensiv wie möglich zu anonymisieren sind. Sobald der Personenbezug für die Fortführung des Forschungsvorhabens nicht mehr erforderlich ist, sind die Daten auf eine Weise zu vernichten, dass keine denkbare Möglichkeit mehr besteht, einen Personenbezug wiederherzustellen.

10 Zusammenfassung

– Die Forschenden haben in der Planungsphase des Forschungsprojekts die Entscheidung zu treffen, ob personenbezogene Daten für ihre Datengrundlage unerlässlich sind. Damit verknüpft ist auch die Frage, ob Datenschutzrecht zu beachten ist, da es nur zur Anwendung kommt, wenn personenbezogene Daten erhoben oder anderweitig verwendet werden.

– Nur dann, wenn das Forschungsprojekt nicht ohne personenbezogene Daten umgesetzt werden kann, dürfen sie erhoben werden. Die Forschenden dürfen dann auch nur so viele personenbezogene Daten erheben, wie unbedingt erforderlich ist. Anschließend sollen diese schnellstmöglich anonymisiert werden.

– Für die Dunkelfeldforschung spielt die Frage, ob personenbezogene Daten erhoben werden oder nicht, eine besondere Rolle, da die Probanden in vielen Fällen ein gesteigertes Anonymitätsinteresse haben. Sie werden nur dann zu einer wahrheitsgemäßen und umfassenden Auskunft bereit sein, wenn sie entweder anonym bleiben oder sich sicher sein können, dass diese Informationen nur für Zwecke der Forschung verwendet werden.

– Welches Datenschutzrecht einschlägig ist, hängt davon ab, welche Stelle das Forschungsprojekt durchführt: Handelt es sich um eine öffentliche Stelle, die entweder dem Bundesrecht oder dem Landesrecht untersteht? Oder wird die Studie durch eine nicht öffentliche Stelle vorgenommen?

– Jede Erhebung personenbezogener Daten muss außerdem zulässig sein. Dazu muss der Proband entweder in die Erhebung seiner personenbezogenen Daten einwilligen oder es muss eine Rechtsvorschrift bestehen, die den Forschenden diese Datenerhebung erlaubt. Im Dunkelfeld werden personenbezogene Daten in aller Regel mit der Einwilligung des Probanden erhoben.

– Bei Gestaltung der Befragung des Probanden sollten sich die Forschenden bewusst machen, ob über die Fragen bzw. die Antworten nicht auch personenbezogene Daten anderer Personen aus dem Umfeld des Probanden, etwa dem Täter, miterhoben werden. Sollte dies der Fall sein, so müsste auch diese Datenerhebung zulässig, d. h. wiederum durch eine Rechtsvorschrift erlaubt oder durch die Einwilligung dieser betroffenen Person gerechtfertigt sein. Da eine Einwilligung in der Praxis selten erteilt werden dürfte, müsste eine Rechtsvorschrift die Erhebung erlauben. Einschlägige Normen verbinden die Erhebung personenbezogener Daten ohne Mitwirkung der betroffenen Person aber mit hohen Anforderungen.

– Das Datenschutzrecht überlässt die Entscheidung darüber, ob z. B. eine bestimmte Erhebung, Nutzung oder Weitergabe personenbezogener Daten zulässig ist, überwiegend der Entscheidung im Einzelfall. Das heißt, dass alle im konkreten Fall betroffenen Belange von Forschenden und Probanden mit- und gegeneinander abgewogen werden müssen. Für das Forschungsteam bedeutet dies, dass in jeder Situation personenbezogener Datenverarbeitung überlegt werden muss, welche Faktoren zugunsten oder zulasten jeweils der Forschenden und Probanden zu berücksichtigen sind.

11 Literaturverzeichnis

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (2007): Richtlinie für Online-Befragungen. URL: www.adm-ev.de/fileadmin/user_upload/PDFS/R08_D_07_08.pdf – Download vom 27. 02. 2015.

Gola, Peter; Schomerus, Rudolf (Hg.) (2015): Kommentar zum Bundesdatenschutzgesetz (BDSG), 12. Auflage. München: Beck.

Häder, Michael (2009): Der Datenschutz in den Sozialwissenschaften, Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland, Rat für Sozial- und Wirtschaftsdaten. URL: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf – Download vom 27. 02. 2015.

Institut für Sexualwissenschaft und Sexualmedizin, Zentrum für Human- und Gesundheitswissenschaften, Universitätsklinikum Charité Campus Mitte. URL: https://www.kein-taeter-werden.de/ – Download vom 27. 02. 2015.

Jann, Ben; Farys, Ralf (2015): Zufallsstichprobe. In: Diaz-Bone, Rainer; Weischer, Christoph (Hg.) (2015): Methodenlexikon für die Sozialwissenschaften. Wiesbaden: Springer Fachmedien, S. 448.

Logemann, Thorsten (2014): Datenschutzbeauftragter INFO. IP-Adressen – personenbezogene Daten. URL: https://www.datenschutzbeauftragter-info.de/ip-adressen-personenbezogene-daten/ – Download vom 27. 02. 2015.

Metschke, Rainer; Wellbrock, Rita (2002): Datenschutz in Wissenschaft und Forschung (herausgegeben vom Berliner Beauftragter für Datenschutz und Informationsfreiheit). 3. Auflage. Berlin: Druckerei Conrad GmbH.

MOGiS e. V. – Eine Stimme für Betroffene (2011): Sexueller Missbrauch an Kindern und Jugendlichen und Beeinträchtigung der Lebensqualität – Ergebnisse einer von Betroffenen initiierten Online-Befragung. URL: https://mogis-verein.de/wp-content/uploads/2011/05/MOGiS_Studie_Web2.pdf – Download vom 10. 12. 2014.

Musch, Jochen (1999): Ehrliche Antworten auf peinliche Fragen: Die Randomized-Response Technik. In: Reips, Ulf-Dietrich: Aktuelle Online-Forschung: Trends, Techniken, Ergebnisse. Online-Tagungsband zur German Online Research Tagung am 28. und 29. 10. 1999 in Nürnberg. URL: http://www.psychologie.uni-bonn.de/sozial/staff/musch/gor99b.pdf – Download vom 27. 02. 2015.

Ostapczuk, Martin S. (2008): Experimentelle Umfrageforschung mit der Randomized-Response-Technik. URL: http://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-7960/Dissertation%20Ostapczuk_PDF-a.pdf – Download vom 27. 02. 2015.

Peiritsch, Günther (2010): Stopp für diese kriminologische Forschung an unseren Schulen! Offene Aufforderung des Berliner Landeselternausschusses. URL: http://www.lea-berlin.de/index.php?option=com_content&view=article&id=435:offene-aufforderung-des-landeselternausschusses&catid=1:aktuelle-nachrichten – Download vom 27. 02. 2015.

Schaar, Peter (2014): Anonymisieren und Pseudonymisieren als Möglichkeit der Forschung mit sensiblen, personenbezogenen Daten. In: Lenk, Christian; Duttge, Gunnar und Fangerau, Heiner (Hg.): Ethik und Recht der Forschung am Menschen. Heidelberg: Springer, S. 95–99.

Schneider, Jochen (2011): Die Datensicherheit – eine vergessene Regelungsmaterie? – Ein Plädoyer für Aufwertung, stärkere Integration und Modernisierung des § 9 BDSG, In: Zeitschrift für Datenschutz, S. 6 ff.

Simitis, Spiros (Hg.) (2014): Kommentar zum Bundesdatenschutzgesetz, 8. Auflage. Baden-Baden: Nomos.

Urteil des Bundesverfassungsgerichts (BVerfGE) vom 19. 07. 1972 (Az. 2 BvL 7/71). In: Entscheidungssammlung des Bundesverfassungsgerichts, 33. Band, S. 367–394.

Urteil des Bundesverfassungsgerichts (BVerfGE) vom 24. 05. 1977 (Az. 2 BvR 988/75). In: Neue Juristische Wochenschrift (1977), S. 1489–1493 (Leitsatz und Gründe).

Urteil des Verwaltungsgerichts (VG) Berlin, 1. Kammer, vom 24. 05. 2011. In: Computer und Recht (2012), S. 191–193 (red. Leitsatz und Gründe).

Weichert, Thilo (2013): Virtuelles Datenschutzbüro. Anonymisierung und Pseudonymisierung im Sinne des Bundesdatenschutzgesetzes. URL: http://www.datenschutz.de/feature/detail/?featid=101 – Download vom 27. 02. 2015.

Weischer, Christoph (2015a): Aggregation. In: Diaz-Bone, Rainer; Weischer, Christoph (Hg.): Methodenlexikon für die Sozialwissenschaften. Wiesbaden: Springer Fachmedien, S. 14.

Weischer, Christoph (2015b): Stichprobe. In: Diaz-Bone, Rainer; Weischer, Christoph (Hg.): Methodenlexikon für die Sozialwissenschaften. Wiesbaden: Springer Fachmedien, S. 396.

2 Amtliche Daten der Kriminalstatistik versus Daten aus Opferbefragungen – Vergleichsschwierigkeiten und Kombinationsmöglichkeiten

Comparing Difficulties and Combination Possibilities: Experience in the United Kingdom

Paul Norris

1 Introduction

The growth of victimisation surveys over the last 30 years can be seen as part of a wider pluralisation of crime statistics. It owes much to a desire to understand the 'dark figure' of crime; victimisation which is not recorded in official crime statistics either because it is not reported to the police, or due to police actions when they are informed of an event. Despite increased methodological sophistication, and wider substantive coverage, victimisation surveys remain subject to limitations meaning that, like recorded crime statistics, they provide only a partial picture of victimisation.

This chapter aims to illustrate how data collected through victimisation surveys can complement, and help to contextualise, data recorded in official statistics. Furthermore, the comparison of victimisation as recorded through surveys and recorded crime statistics helps to illustrate several important methodological issues around the construction, and analysis, of survey data. Understanding how, and why, the measurement of crime varies between sources can help those interested in policy and practice to identify the appropriate indicator for different purposes.

2 A Brief History of Recorded Crime Statistics

The systematic presentation of crime statistics has a long history within Europe, first coming to prominence in mid-19th century France. While early attempts to collect data on patterns of victimisation were driven by academics, they soon became a focus for governments in various nations (Maguire 2012). The first national crime statistics in the England and Wales were compiled by the Home Office in 1857, and the domination of police recorded data in official publication remained largely unaltered until the early 2000s. The publication "Criminal Statistics, England and Wales" provided tables of aggregate numbers of offenses classified in accordance with the legal definition of different crimes. These tables were typically provided at both a national level and the level of individual police forces, while data on trends over time was also included. Since the early 2000s, a new annual publication from the

Home Office, "Crime in England and Wales", has published data on crimes recorded by the police, alongside estimates from data from the National Crime Survey for England and Wales (formally referred to as the British Crime Survey). This shift in publication strategy illustrates how the alternative methods of estimating victimisation can be seen as complementary, exhibiting different strengths and weaknesses, both capable of making an important contribution to understanding patterns of crime. Reflecting how the United Kingdom comprises three separate jurisdictions, separate publications of crime recorded by the police are published for Scotland and Northern Ireland.

Until the mid-20th century recorded crime stayed consistently low across most Western democracies. This was followed by a period of notable growth. Trends in recent years exhibit less consistency, characterised by fluctuations which vary between nations, crime type and over time (for a detailed review of evidence in Europe, see Aebi/Linde 2012).

3 Methodological Critiques of Recorded Crime Statistics

A range of issues exist which limit the use of recorded crime statistics for understanding patterns of victimisation. Briefly outlining these issues helps to understand not only the motivations for the development of victimisation surveys, but also how the two data sources can complement each other to present a fuller picture of crime.

Recorded crime figures are typically presented as aggregates, either a total number of crimes, or a rate of crime per 100,000 people. Yet counts of crime take no account of changes in the size of the population of interest, for example, if the number of potential victims doubles, it seems reasonable to expect that the amount of crime will similarly increase. Furthermore, neither counts nor rates provide any contextualisation of the crime they report. For instance, has the make-up of the population changed due societal shifts to include more of those at risk of a particular type of victimisation? Finally, it has long been argued (for example, by McClintock and Avison 1968) that the broad categories of crime used in statistical reporting group together incidents which are qualitatively different in terms of victim experience.

The recording of a crime in official statistics represents not simply the occurence of an event, but also reflects the public's willingness to engage with the criminal justice system and the working of an administrative process. This results in recorded crime underestimating overall victimisation, and means that apparent changes in the level of victimisation, might be caused by changes in reporting and recording practices rather than shifts in victimisation. Understanding how changes in reporting behavior and recording practice affect re-

corded crime levels is likely to prove useful for policy makers and practitioners, as it will provide evidence as to how changes in policy and practice influence the actions of the public and police.

The vast majority of crime recorded by the police has come to their attention as the result of a report from the public. An individual's willingness to report crime is influenced by many issues, such as the social acceptability of particular types of behaviour, the perceived seriousness of the incident, perceptions of the police, the practicalities of reporting, and any compulsion to do so (for instance as a requirement of an insurance claim). These issues cause variation in the extent to which different types of crime come to the attention of the police (see *Table 1* later in this chapter). Such issues are also likely to vary over time, reducing the usefulness of recorded crime statistics for identifying trends in victimisation. For instance, Maguire (2012, p. 216) identifies how the apparent erosion of informal social control mechanisms has led to an increase in the use of the police to address minor incidence involving children, while publicity around improvements in the way rape victims were treated by the police in the mid-1980s appears to have led to an increase in the reporting of such incidents.

When the police become aware of an incident, an administrative process (which involves elements of discretion and decision making on the part of individual officers) is played out before the crime is formally recorded. In England and Wales, only crimes on the "Notifiable Offence List" feature in recorded crime statistics. This list contains offences which are typically considered more serious, those which must, or can be, tried in a Crown Court, along with "a few closely related summary offences (dealt with by a magistrate)" (ONS 2014). While a focus on more serious crime might be desirable to avoid a situation in which data are dominated by minor issues which might have little impact on everyday life (for instance it excludes many millions of minor traffic and parking offences), the boundary between serious and minor has varied over time, and is a continued source of debate. For instance, Maguire (2012, p. 212) notes that the decision to include the previously excluded offences of common assault, harassment and assault on a constable in 1998/9 added approximately a quarter of million incidents to recorded crime. Drink-driving remains excluded from the list despite the substantial policing focus on it over the last 20 years, and the modern-day widespread social unacceptability of such behaviour. The Notifiable Offence List also changes to reflect new legislation and this can further limit comparability over time. An indication of how the focus on specific serious offences can cause an underestimation of crime can be seen in how, in 2011, approximately 1.2 million convictions were handed out by Magistrates Courts in England and Wales (Sentencing Council 2014), the majority of which would not appear on the Notifiable List.

Even if a classification is agreed for recording crimes, variation may still exist in how this is implemented in practice. A single event may involve multiple individual crimes, or it may be unclear exactly what crime, if any, has been committed in a particular situation. For instance, is an individual who assaults someone and steals their wallet, guilty of assault or theft, one offence or two? For a long time there was little, if any, attention paid to how police recorded particular incidents, or how the process varied between regional forces; who may have attempted to influence figures to enhance clear-up rates, or to support arguments for greater resources. Only towards the end of the 20th century were substantial efforts made to harmonise the recording processes across England and Wales, leading to the introduction of the National Crime Recording Standard (NCRS) in 2002. Indeed, this process saw a shift towards recording crime in a manner consistent with the rules of the Crime Survey for England and Wales, both in terms of how different offences should be classified, and in terms of recording crime in terms of the number of victims rather than events. Maguire (2012, p. 215) argues that introduction of the NCRS saw the percentage of incidents reported to the police which were recorded as crime increase from 62% in 2000/1 to 75% in 2003/4 – a finding arrived at by comparing recorded crime figures with those estimated through the national crime survey. This serves to illustrate how evidence collected via surveys can provide a contextualisation of figures presented in recorded crime statistics (a topic covered in further detail below).

4 Early Crime Surveys

Beginning in the United States in the mid-1960s, victimisation surveys are intended to address many of the concerns raised above with regards to recorded crime statistics. Most notably they were seen as a mechanism to account for the 'dark figure of crime', the underestimation of crime due to incidents not coming to the attention of the police, or the police choosing not to record them as crimes.

The first National Crime Survey (NCS) was conducted in the United States in 1972, following the recommendation of the Commission on Law Enforcement and Administration of Justice which was established by President Johnson in 1965. The Commission's interest was in developing a measure comparable with Uniform Crime Reports (recorded crime statistics in the US) but which would be independent of the criminal justice system, and gave an indication of the true level of crime experienced by the population (as opposed to the amount of crime recorded by the police).

Originally conceived as a series of surveys which would concentrate respectively on victimisation in specific cities, the experience of businesses and a

representative sample of the national population, the NCS evolved to concentrate only on the final of these objectives. Questions in the NCS covered the eight most serious crimes included in the UCR, reflecting the belief that the two datasets should be comparable and complementary.

The development of the NCS was soon followed by the creation of similar surveys in other nations, notably across Europe, North America and Oceania. 1972 also saw the inclusion of several crime related questions in the British 1972 General Household Survey. Despite this, it was not until the work of Sparks et al. (1977), who conducted surveys in specific areas of London that victimisation surveys began to develop a strong position in British criminology. This was followed in 1982 by the first British Crime Survey run by the Home Office.

In addition to national crime surveys, there are several examples of surveys intended to allow for comparisons between countries as well as over time. The most established of these is the International Crime Victim Survey (see chapter by van Dijk and Castelbajac in the first volume), which began in the late 1980s, and has been conducted six times (in various forms) since then. In addition, the European Social Survey has included questions around crime and justice (most notably during round 5 in 2010). Similarly, the Gallup World Poll, a survey conducted annually since 2005 in over 150 nations includes several justice related questions (two of which concern victimisation). In all these cases, the primary objective is provide comparable data, irrespective of differences in legal definitions, reporting behaviour or recording practices between nations.

Although early crime surveys were often seized upon by many as providing a true estimate of the level of crime, researchers soon began to exploit their wider potential. Specifically, because they asked about a range of demographic issues, as well as responses to victimisation, and attitudes towards criminal justice institutions, they opened the opportunity of contextualising knowledge about levels of victimisation, for instance understanding who in the population was most likely to be a victim, and investigating how recorded crime statistics might be influenced by different groups of victims having different propensities to report crime to the police. This contextualisation of victimisation can play an important role in providing an evidence base for those wishing to shape policing policy. For instance, understanding how the risk of victimisation varies across the population can help the targeting of crime prevention activities towards those most likely to benefit, similarly knowledge of who does not report victimisation to the police can provide important insights as to how providing alternative routes for individuals to report crime to the police may encourage previously marginalised groups to more often interact with the police. More generally, the inclusion of questions around general

preceptions of the police, and around indivduals' specific interactions with the criminal justice system, can help to situate how the public's willingness to accept police decisions is influenced by wider perception of legitimacy (see for instance Hough et al 2013, who use data from the European Social Survey).

5 Crime Surveys in the UK

A detailed history of UK crime surveys will not be provided here. A short overview is provided in Hough and Norris (2009), while a fuller discussion of the first 30 years of the British Crime Survey is provided in the edited volume of Hough and Maxfield (2010). This section will highlight several aspects of survey design which are particularly pertinent when attempting to compare survey based estimates of crime with recorded crime statistics.

The population of the UK are covered by three separate 'national' crime surveys; The Crime Survey for England and Wales[1], The Scottish Crime and Justice Survey[2] and the Northern Ireland Crime Survey. Large overlaps exist between these surveys in terms of both their methodology and the questions they employ. However, the continued use of three different surveys reflects how the UK consists of three separate criminal justice jurisdictions. This point is particular relevant when making comparisons between recorded crime figures and estimates from survey data, as it allows for differences in the definitions of crimes between jurisdictions to be mirrored in the creation of survey based estimates.

Early crime surveys in the UK were sporadic. In recent years, all three crime surveys in the UK have moved towards continuous interviewing. Interviews are conducted all the time with 12 months' worth of interviews then combined together to make a dataset for a particular year. Continuous interviewing can present some methodological challenges since different respondents within the same dataset will have been asked about victimisation in different 12 month periods. With continuous interviewing, victimisation estimates are typically based on rolling averages. Such estimates involve a more complex calculation (compared to the simple averages constructed for early surveys with fixed reference periods) and it can be argued they are less transparent in their construction since changes in the methodology used for calculating the rolling average will influence the estimates provided (see Lynn/Elliot 2000,

---

1 Known as the British Crime Survey until 2012.

2 Known variously as the Scottish Crime Survey and the Scottish Crime and Victimisation Survey.

pp. 16–17 for a more detailed discussion). However, the move to continuous interviewing means it is possible to construct estimates for any time period of interest, and should aid the comparison of trends over time.

While their sampling frames have varied over time, often reflecting the sampling frame which gives the most complete coverage of a nation's population, all three surveys have, until recently, focussed on individuals aged 16+ and living in private households. More recently, the National Crime Survey for England and Wales has included a sample of children aged 10–15.

The questionnaires used for the three surveys follow a broadly similar structure. They begin by asking a set of screener questions to establish if the respondent has experienced victimisation in the previous 12 months.[3] These questions are generally consistent across the three surveys. Where a respondent reports having experienced victimisation during the reference period, they are asked a series of follow-up questions via 'victim forms'. These questions are intended to establish if an incident was a crime, appropriate for counting within rules of the survey, and if so which type of crime has occurred. The number of victim forms a respondent is asked to complete has varied across surveys and over time. Furthermore, each survey restrict the total number of victim forms a respondent is asked to complete, with those crime types perceived as most serious, or least common, taking precedent if a respondent has reported a large number of incidents. Each victim form asks if the incident was reported to the police. This question is particularly important for understanding the relationship between survey estimates and recorded crime statistics.

Post data collection, the answers to the questions included on the victim forms are used to classify incidents in terms of the type of offence that occurred. The rules for classifications vary between the three surveys, and are intended to reflect the definitions of crimes employed in the relevant criminal justice system. This coding of offences is therefore a crucial stage in establishing comparability between recorded crime statistics and victimisation survey results. Ensuring comparability between these different measures within a jurisdiction is, to an extent, at the expense of comparability between jurisdictions. Hence it is important that the extent to which a survey is intended to provide estimates which are either consistent between jurisdictions, or consistent with the definitions used for recorded crime statistics, is established early in the design process.

---

[3] Until the surveys moved to continuous interviewing, a 12 month reference period, often the previous calendar year was employed.

6 Methodological Critiques of Survey Data

While crime surveys were originally envisaged as a tool to overcome the short-comings of recorded crime statistics for understanding levels of crime, they are themselves subject to several methodological limitations (discussed in detail elsewhere in this volume). Several of these are worthy of a brief mention here, since they impact on the complementarity of crime surveys and recorded crime statistics. They also help to illustrate the limitations of survey data which practitioners and researchers wishing to understand patterns of victimisation will need to be aware of and address in reporting their analysis.

Recorded crime statistics include crimes irrespective of the nature of the victim. In contrast, crime surveys typically focus on individuals, their experiences, and those of their household members. This means they often exclude crimes committed against businesses, or those crimes where an individual may not be aware they have been victimised (see Hoare 2010, with regards to measuring incidence of fraud). Furthermore, most national crime surveys have focused on individuals aged 16+, and living in private households, excluding particular groups of the population from their analysis (for instance children and those living in care homes) who might experience patterns of victimisation different from the wider population.

Definitional issues can limit the comparability of different datasets. Firstly, individuals may not report events within a crime survey if they do not view them as crimes. Therefore, the prevalence with which particular incidents are reported in a survey will, to some extent, reflect individual perceptions, expectations and social norms, rather than giving a neutral measure of criminal victimisation. Furthermore, the definitions of crime employed in constructing a survey will vary depending on the purpose of the survey. A survey which is intended to provide data comparable to recorded crime statistics must therefore define crimes in a way which matches the legal definition within the relevant jurisdiction. As Coleman and Moynihan (1996, p. 81) note, "surveys do not necessarily use legalistic definitions of crime; this means that had the incident been reported to the police, it would not necessarily have been recorded." All three of the UK surveys attempt to match respondents' responses to legal definitions when classifying an event as a crime.

Even where definitions appear to match, further difficulties are still likely to remain. Mayhew et al. (1993, p. 4) observe that when legal definitions are applied to survey data, they provide "a nominal definition of crime: a count of incidents which according to the letter of the law could be punished." This may not match the 'operational definition' employed by the police, whose recording practices might vary depending on the specific circumstances of the report they receive. While such differences might limit the direct comparability of recorded crime statistics and survey based estimates, investigating

whether such matters are consistent over time and between crimes, can help illuminate the extent to which police decision making may contribute to the 'mismeasurement' of crime in recorded crime statistics (a topic covered further below).

Victimisation surveys are typically conducted retrospectively. That is to say they ask respondents about past experience of victimisation. As mentioned above, all three national crime surveys in the UK now employ a rolling 12 month recall period, asking questions, which begin similar to "Since [INSERT DATE 12 MONTHS AGO] have you . . .". This contrasts with recorded crime statistics which are dominated by contemporaneous reports, provided soon after an incident occurred. A wide literature exists on respondents' ability to recall victimisation when responding to a survey (for instance Gottfredson/Hindelang 1976, Murphy/Cowan 1976, Cantor/Lynch 2000). Typically, respondents are more likely to include serious incidents in their responses, when they in fact occurred before the period with which the survey is concerned, and correspondingly more likely to exclude more minor offences which occurred early in the reference period (believing they occurred earlier).

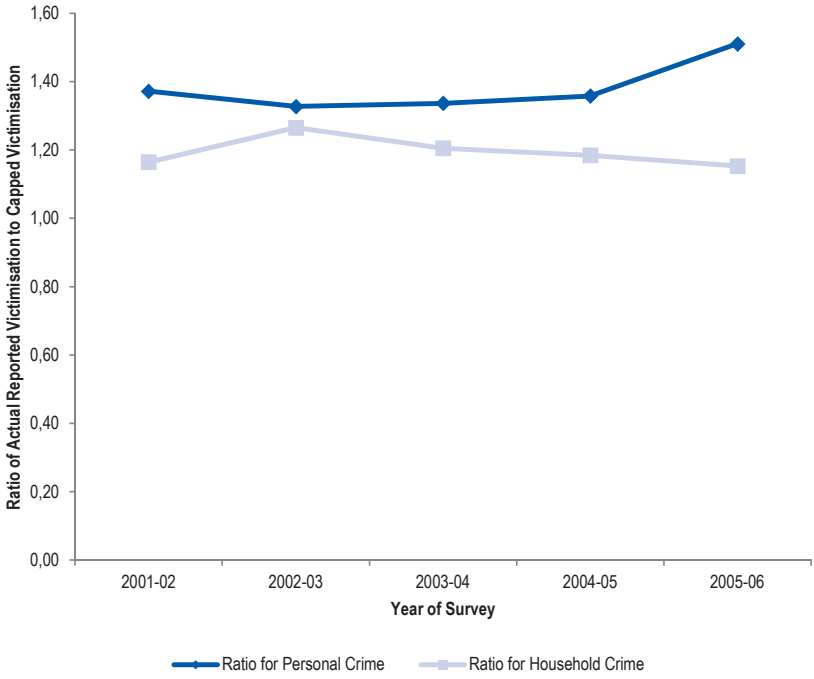
One of prime purposes of crime surveys was to attempt to address the underestimation of victimisation which is present in recorded crime statistics. Yet the design of most crime surveys means that they are still likely to provide an underestimation of the true level of crime, notably underrepresenting the experience of those who experience high levels of victimisation. All three crime surveys within the UK cap the amount of victimisation any single respondent can contribute to the survey. As outlined previously, a respondent who reports been a victim of crime, is asked a series of detailed follow-up questions. However, the number of such questions is limited. Furthermore, if a respondent reports repeat victimisation of the same type which was likely by the same perpetrator then these events contribute a maximum of 5 to any count of victimisation, even if the actual number of incidents reported is much higher. This counting behaviour is likely to see crime surveys in the UK underestimate the true level of victimisation. Farrell and Pease (2007, p. 43) note that in the 2005-06 Crime Survey of England and Wales, these counting ruleslikely saw the exclusion of around one third of personal victimisation from estimates of the level of crime. This suggests that crime surveys only go partway to uncovering victimisation not present in recorded crime statistics.

Considering both personal and household crime between 2001-02 and 2005-06, *Figure 1* shows the ratio of total victimisation reported to the Crime Survey for England and Wales with the capped rate reported in the official report (based on Farrell/Pease 2007, p. 43). *Figure 1* suggests that while the impact of capping may be relatively consistent, it does fluctuate both between crime

types and over time. Understanding how counting rules might influence survey based estimates of crime is important to contextualise any comparisons with recorded crime statistics.

Figure 1:

Ratio of Total Victimisation Reported to the National Crime Survey for England and Wales versus Capped Estimates



7 Comparing Recorded Crime and Estimates from Survey Data

The anonymous nature of victimisation survey data means that it is not possible to link specific survey respondents to specific incidents in police recorded crime data. Instead, comparisons are made at the aggregate level. Indeed, while the crime surveys conducted in the UK employ sample sizes which could be considered large, by the standards of most pieces of survey based research, comparisons are often restricted to jurisdiction level as it has been argued that the sample size does not allow for robust comparisons of smaller

geographic units (Flatley/Bradley 2013, p. 7)[4]. It is possible to undertake several different types of comparison, when considering the aggregate level relationship between recorded crime statistics and survey based estimates. At the most basic level, comparisons can be made concerning the total amount of crime at a single point in time (spot estimates). Alternatively, comparisons can concern the mix of victimisation across different types of crime (compositional comparisons). Finally, comparisons of change in the levels of victimisation over time can be made (trend analysis).

The process of comparing estimates of crime from recorded crime statistics with those from a survey involves two, apparently simple, tasks:

(1) Identifying comparable offences, both in terms of types of crime and the population of victims

(2) Multiplying up survey based estimates of victimisation to represent the whole population from which the sample was drawn.

Recorded crime statistics cover all types of victimisation, while crime surveys typically concentrate on a limited range of crimes committed against individuals and their personal property[5]. It is therefore necessary, as a first step to comparing recorded crime statistics with crime survey estimates, to identify a common set of crimes which are recorded in both datasets. Within the UK, notable crimes included in police statistics but not covered by national crime surveys include, shoplifting, fraud and damage to commercial properties (all of which are committed against institutions rather than individuals), as well as murder, drug crime and burglary of properties in which no one lives.

Once a comparable set of crimes is identified, it is necessary to adjust the police recorded counts for these crimes to ensure they are based on the same population (and time period) as the survey data. For instance, a downward estimate of the number of recorded assaults might be needed to exclude those which have involved victims under the age of 16 (who have been traditionally excluded from UK victimisation surveys). Such adjustments are often made on the basis of additional information collected from police forces (which may not be uniformly collected), and employ a range of assumptions (as an example, details for England and Wales in 2005/06 can be found in Hough/Norris 2009, pp. 120–126).

---

[4] For instance, the National Crime Survey for England and Wales has, over recent years, involved a sample of 35,000+ respondents in each 12 month period. However, the jurisdiction consists of 43 separate police force areas.

[5] Crime surveys concerning crimes against commercial interests can also be run, although they are less common. For an example, see ONS2014.

Estimating the level of victimisation experienced by a population from the results of a sample survey is conceptually straight-forward. It simply requires accurate knowledge of the size of the population that the sample was intended to represent. However, precise knowledge of the size of a given population (for instance all adults over the age of 16) is difficult to locate. In the UK, estimates of populations take the decennial census, with adjustments made year on year to try and account for births, deaths and migration. Since such adjustments are based on various estimated data sources, their accuracy is likely to vary over time. Any inaccuracies in estimates of the size of the population will hinder the creation of population based estimates of victimisation, while variation in accuracy of the population estimates over time will limit the ability to undertake robust analysis of trends in victimisation. The complications of estimating population size become greater if further levels of granularity are required, for instance calculating the level of vehicle crime in a jurisdiction is likely to require not only knowledge of the population size, but also of the vehicle ownership rate.

With regards to the adjustments needed to create comparable statistics based on recorded crime and survey data in England and Wales, Flatley and Bradley (2013, p. 7) note "it should be recognised that this 'comparable' series remains broadly rather than directly comparable and that the offence classification system used in the survey can only approximate that used by the police."

Possibly because the comparable crimes are seen only as broadly comparable, recent reports covering the National Crime Survey for England and Wales, and the Scottish Crime and Justice Survey, typically include comparisons between survey data and recorded crime statistics only in the broadest terms – a total estimate of comparable crimes. In Scotland, the 2012/13 victimisation survey provides an estimate of 527,000 incidents for those crimes which are comparable with recorded crime statistics. In contrast, recorded crime statistics for the comparable period identify approximately 145,000 incidents, just less than 30% of the incidents estimated via the crime survey (Scottish Government 2014, p. 19).

While current UK survey reports typically provide only an overall estimate of how the level of crime reported by respondents compares to police recorded crime, it is possible to create estimates by sub-types of crime, or indeed for specific comparable crimes. Since any specific type of victimisation is only likely to have been experienced by a small number of respondents, such estimates are likely to be more volatile than a simple overall figure. *Table 1* gives estimates from the 2006 Scottish Crime and Justice Survey to show how the relationship between survey estimates of crime and police recorded crime figure varies depending on the type of crime considered. Across all three categories of crime, recorded crime levels are substantially below those reported

in the survey. While absolute levels of victimisation vary, there is evidence that the ordering of crimes (in terms of frequency) is consistent across the different data sources. Violence is consistently the most prevalent type of crime (320 incidents per 100,000 people in the survey data and 132 incidents per 100,000 people in the recorded crime data). Similarly acquisitive crime appears the least common in both measures of crime (97 and 61 incidents per 100,000 people respectively). This suggests that while absolute levels of victimisation might vary between different sources, they may often present similar broad trends, and considering multiple datasets might provide useful triangulation of data, increasing the robustness of any analysis.

Although, in terms of incident rates, the ordering of crime types is the same when considering survey based estimates and recorded crime statistics, the relatively prevalence of crimes appears to vary. Hence while violence accounts for around 46% of crime in both datasets, vandalism represents just under 40% of survey based crime but only 32% of recorded crime. In contrast, acquisitive crime appears more prevalent in recorded crime statistics (21%) compared to the survey data (14%). This suggests that while survey based data and police recorded crime statistics might give similar pictures in terms of the most common crimes in absolute terms, they vary in their representation of the composition of victimisation. The difference in the relative prevalence of different types of crime, across the two datasets, likely reflects differences in the reporting behaviour of victims and the recording practices of the police (for instance people might need to report acquisitive crime to the police to support insurance claims). The apparent lack of consistency between the two datasets when considering the relative prevalence of different types of crime, contrasts with the stability suggested in the previous paragraph (which considered absolute levels of victimisation) and illustrates the importance of establishing how the data collection process associated with any dataset will affect the results achieved for a specific piece of analysis.

Table 1:

Crimes Reported to, and Recorded by, the Police - Broken-down by Crime Type in the 2006 Scottish Crime and Justice Survey (Brown/Bolling 2007, p. 78)[6]

	Vandalism	Acquisitive	Violence	All Comparable Crimes
Survey Measured Crimes (per 100,000)	274	97	320	691
Percent Survey Crimes Reported to Police	33.4	62.7	41.3	41.2
Number of Survey Crimes Reported to Police (per 100,000)	92	61	132	285
Police Recorded Crimes (per 100,000)	42	50	75	167
Percent of Crimes Reported to Police that are Recorded	45.9	82.5	56.7	58.7
Percent of all Survey Crimes Recorded by Police	15.3	51.5	23.4	24.2

Considering data from the 2012/13 Scottish Crime and Justice Survey (Scottish Government 2014, p. 19) suggests that the composition of victimisation reported in the crime survey and recorded crime statistics is now more consistent than it was in 2006. In 2012/13, violence still accounted for 46% of police recorded crime in Scotland, in contrast to 45% of survey reported crime. Acquisitive crime accounted for 15% of recorded crime and 14% of crime recorded in the crime survey, while Vandalism made up 39% of recorded crime and 42% of survey crime. That the composition of crime presented in the survey data and recorded crime statistics in 2012/13 is more consistent than that presented in 2006, suggests that the effect of differences in reporting and recording of victimisation has varied over time (a point which will be considered with reference to England and Wales below).

Returning to the detailed breakdown of data for Scotland in 2006 (*Table 1*), it is notable that acquisitive crime exhibits the highest levels of reporting and recording, although even in this case only around one in every two incidents is recorded by the police. In contrast, just less than one in six incidents of vandalism appear to have been included in the police recorded crime statistics. Variation in the relationship between survey based estimates and police

[6] Vandalism is a single crime category. Acquisitive crimes are Housebreaking, Theft of a Motor Vehicle and Bicycle Theft. Violence includes Assault and Robbery.

recorded crime statistics reflects differences in both the likelihood of victims reporting crime to the police, and the probability that the police will classify an incident as a crime once they are made aware of it. Acquisitive crimes have the highest rate of reporting to the police (possibly due to the requirements of insurance claims), and the highest rate of police recording (maybe reflecting how the theft of property is relatively easy to define as a crime). In contrast, many minor incidents of vandalism go unreported to the police, likewise for many assaults which make up the majority of the comparable violent crime subset. Similarly, it may be harder for the police to establish the circumstances around many minor incidents of vandalism and violence, meaning it is less clear that any specific incident should be recorded as a crime. The variation in reporting and recording behaviour across crime types suggests that looking at the overall relationship between survey based estimates of crime and police recorded statistics may be insufficient for understanding how victims and the police respond to victimisation.

The comparisons presented above are spot estimates, whether concerning levels of victimisation or the composition of crime, they involve data measured at one point in time. An alternative research question involves whether, even if absolute levels of victimisation vary between survey data and recorded crime statistics, the two sources show similar patterns in terms of changes in the level of crime over time. Reflecting the longer history of national crime surveys in England and Wales (compared to the rest of the UK), this section will concentrate on comparisons within that jurisdiction.

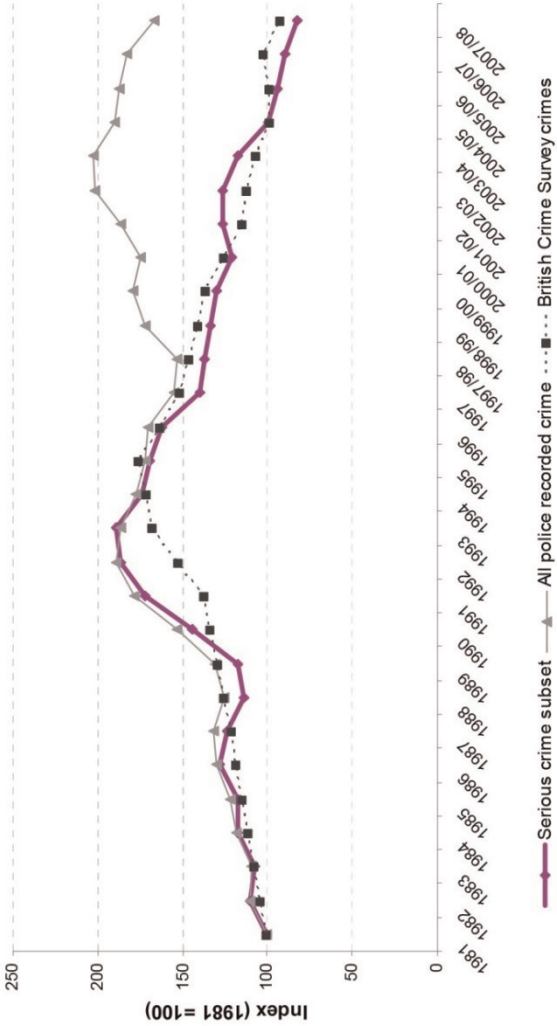
Figure 2 shows changes in victimisation in England and Wales (indexed to 1981 levels) as measured by the national crime survey, through total police recorded crime, and in terms of recorded crime across a subset of more serious victimisation[7]. In interpreting *Figure 2*, it is important to take account of methodological changes which have affected both recorded crime statistics, and the victimisation survey estimates over the period. Firstly, the change of reference period employed in the victimisation survey, following the introduction of continuous interviewing in 2001, means that the periods covered by this time series are not constant across the period (see Kershaw et al. 2008, p. 28). Furthermore, the estimates of recorded crime are affected by a change from presenting figures based on calendar years to financial years, which occurred in 1997, a change to counting rules in 1998, and the introduction of the National Crime Recording Standard in 2002 (discussed above).

[7] Crimes included are most serious violence against the person; most serious sexual offences; robbery; burglary, theft or unauthorised taking of a motor vehicle; and aggravated vehicle taking.

All three time-series follow a similar trend until the mid-1990s, suggesting that shifts in victimisation were consistent across the alternative measures. The mid to late 1990s show a fall in crime as measured by the victimisation survey, and police recorded crime when considering only serious offences. In contrast, total police recorded crime saw a jump following the introduction of the new counting rules in 1998, and a similar discontinuity following the introduction of the National Police Recording Standard in 2002. This serves to illustrate how changes in the mechanism for collecting trend based victimisation data can influence the pattern that is presented. That the police recorded time-series referring to serious offences has continued to mirror the pattern shown by the survey data suggests that the impact of changes in policing practice appear to have predominantly affected how the police record more minor criminal matters. Although now at a higher absolute level, since around 2000 the rises and falls in total police recorded crime have largely mirrored those shown in the survey data. This could provide evidence that while changes in police practice might provide points of discontinuity in a time-series, once these are accounted for both types of data show broadly similar trends.

Figure 2:

Trends in Survey Based Estimates of Victimisation and Police Recorded Crime Statistics for England and Wales between 1981 and 2007/08 (Kershaw et al. 2008, p. 30)

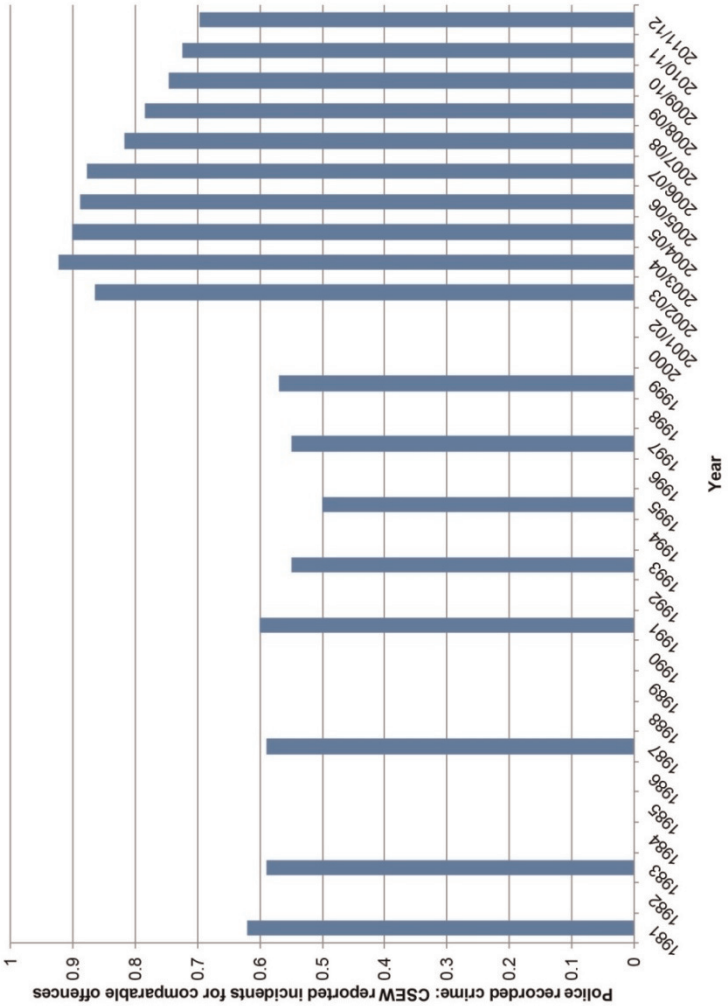


As highlighted in *Table 1*, variation in estimates of crime between recorded crime statistics and victimisation surveys may reflect victims' willingness to report crime to the police, or the police's willingness to record an incident as a crime. *Figure 3* shows how the number of crimes recorded by the police in England and Wales has varied as a proportion of crimes which victimisation survey respondents say came to the attention of the police. As such, it gives some indication as to how the likelihood of crimes been recorded by the police has changed over time (cancelling out any changes in victims' propensity to report crime to the police). Throughout the 1980s and 1990s, the proportion of crimes which the police ended-up recording, once they were aware of them, varied between 50% and 62%. As might be expected, the introduction of the National Crime Recording Standard seems to have been associated with an increased likelihood of the police recording an incident as a crime. However, since the mid-2000s, it appears that the proportion of crime recorded by the police has fallen. This pattern mirrors the changes in overall police recorded crime shown in *Figure 2*. Detailed discussion of what might explain the drop in police recording since the mid-2000s is beyond the scope of this chapter. One speculative explanation might be that as the police become more accustomed to what is expected of them under new guidance, their feelings of compulsion to record an incident as a crime reduce. It is, however, worth noting that divergence between measures can complicate discussions of trends in victimisation, as different protagonists are able to draw on different measures to support their arguments, while the relative technical nature of the reasons for differences between various data sources do not easily lend themselves to easy explanation to a non-specialist audience.

Beyond any substantive interpretation, *Figure 3* serves to illustrate two points about combining police recorded crime statistics with survey estimates. Firstly, the sample error associated with surveys means that small year on year changes should not be over interpreted. However, conducting trend based analysis allows for more long-term patterns to be noticed, which can be considered more robust. Hence, it seems reasonable to note a change in police recording practice when comparing the 2000s to the 1980s and 1990s, and to note that a decrease has occurred since the mid-2000s. Secondly, the analysis underpinning *Figure 3* illustrates the importance of including questions about victim-police interaction in a victimisation survey. In order to identify changes in police recording practice, it is necessary to separate this effect from changes in the public's reporting behaviour. This is only possible because when a survey respondent mentions they have experienced an incident of victimisation, they are asked if the incident came to the attention of the police.

Figure 3:

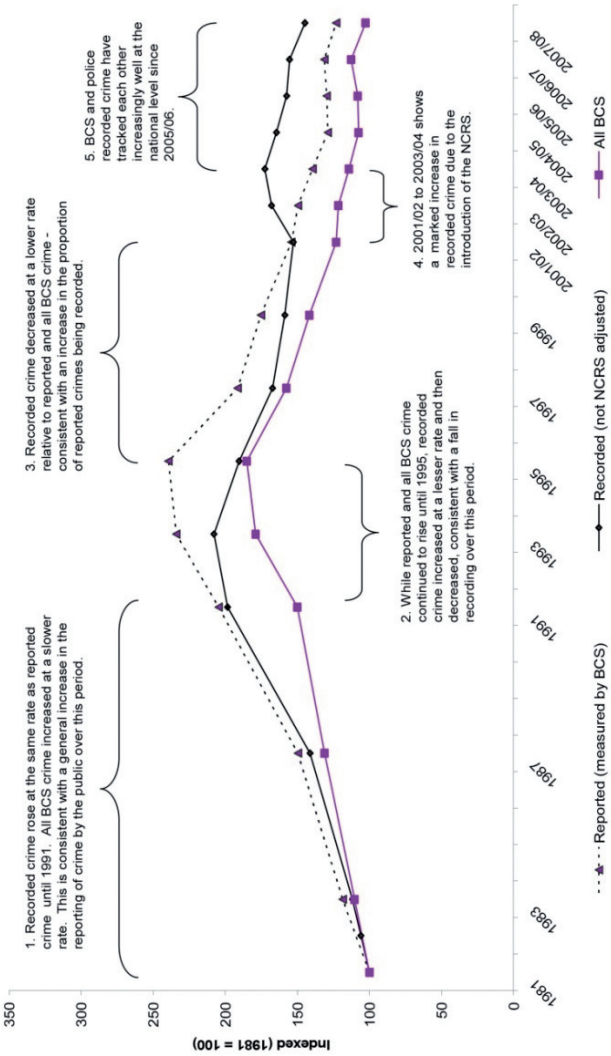
The Ratio of Police Recorded Crime to Victimisation Survey Estimates for England and Wales between 1981 and 2011/12 (Flatley/Bradley 2013, p. 9)



If differences between survey based estimates of crime and police recorded crime statistics can be attributed to a victim's willingness to report to the police and the likelihood of the police recording an incident as a crime, then understanding the relative impact of these different mechanisms over time can help to illustrate important points with regards to criminal justice policy. *Figure 4* illustrates how reporting and recording rates varied in England and Wales between 1981 and 2007/08. The shifts identified in *Figure 4* can be linked to the wider policy context within the jurisdiction. Hence the early 1990s saw recorded crime rise less steeply than either total victimisation, or the number of incidents that survey respondents claim to have informed the police about. This relative fall in the police's recording of crimes occurred at a time when there was a concerted political effort to claim that crime was falling. In contrast, the period from the late 1990s saw a range of policy changes intended to push the police towards full recording of crime. This period saw police recorded crime increase, while the other two measures decreased. All three series appear to have followed a consistent trend since the mid-2000s, possibly reflecting the bedding down of new police recording practices, as also suggested by the trends shown in *Figures 2* and *3*.

Figure 4:

Victimisation in England and Wales 1981-2007/08 (Kershaw et al. 2008, p. 41)



Using examples from England and Wales as well as Scotland, the analyses above have illustrated a range of ways in which survey based estimates of victimisation can be compared to those presented in police recorded crime statistics (similar examples from a range of European countries can be found in Robert 2009). Before any comparisons are made, it is important to establish that both measures are measuring comparable crimes. The restricted range of crimes typically covered by victimisation surveys means any such comparisons will be restricted to a subset of total victimisation. While comparisons can be made at a single point in time, and for specific types of victimisation, the inherent short-term variability of survey based estimates means more robust analysis is likely to be achieved by looking at trends over time, and considering a more general concept of victimisation. Beyond considering the level, and composition of victimisation, useful insights can be gained by considering how the relationship between survey-based estimates of victimisation and recorded crime statistics varies over time.

8 Contextualising Crime Statistics – Understanding How Crime Comes to the Attention of the Police

The aggregate level analysis discussed in the previous section highlighted how police recorded crime statistics are affected by the willingness of the public to report crime to the police. Furthermore, *Table 1* suggested that such behaviour varied between crimes, while *Figure 4* suggests variation over time. Understanding patterns of victimisation in recorded crime statistics therefore requires knowledge of which incidents of victimisation come to the attention of the police. Such insight cannot be gained from the recorded crime statistics themselves, as it requires those incidents of which the police are aware to be considered within the context of those that they were not informed about. Victimisation surveys include information about victims of crime, irrespective of whether or not the police were informed, as such, they can help understand patterns of reporting behaviour across victims, and hence help to contextualise recorded crime statistics. A detailed discussion of how surveys can help understand the likelihood of specific crimes coming to the attention of the police is provided in Chapter 2 in volume 1 of this compendium by Uwe Kolmey and in a wider literature (for instance, Baumer/Lauritsen 2010; Goudriaan, et al. 2006; MacQueen/Norris 2014; Tarling/Morris 2010).

Table 2:

Most Common Reasons for not Reporting Crime to the Police in Scotland in 2011/12 (Scottish Government 2014, p. 48)

	Unreported Property Crime	Unreported Violent Crime	All Unreported Crime
Police could not have done anything about it	41	24	37
Incident was too trivial, not worth reporting	43	17	36
Police would not have been interested	14	16	15
Victims dealt with the matter themselves	5	23	9
Incident was considered a personal or family matter	5	14	7
Inconvenient / too much bother to report	7	4	6
Crime was reported to other authorities or organisations	5	2	4
Fear of reprisals by offenders	2	3	3
There was no loss or damage	2	2	2
Previous bad experience of the police or courts	1	3	2
Dislike / fear of the police	0	6	2
Number of Crimes	1,190	170	1,360

Figures are percentage of victims who did not report to the police (multiple reasons could be given)

With regards to Scotland in 2011/12, *Table 2* presents a breakdown of the reasons crime survey respondents gave for not reporting crime to the police. It highlights that the reasons for crime not coming to the attention of the police vary across crime types (note the high prevalence of victims dealing with the matter themselves and incidents being considered a personal or family matter with regards to violent crime compared to property crime). This provides an example of how the supplementary data collected in a crime survey can help to understand the biases that might exist in police recorded crime statistics.

9 Understanding Patterns of Individual Victimisation

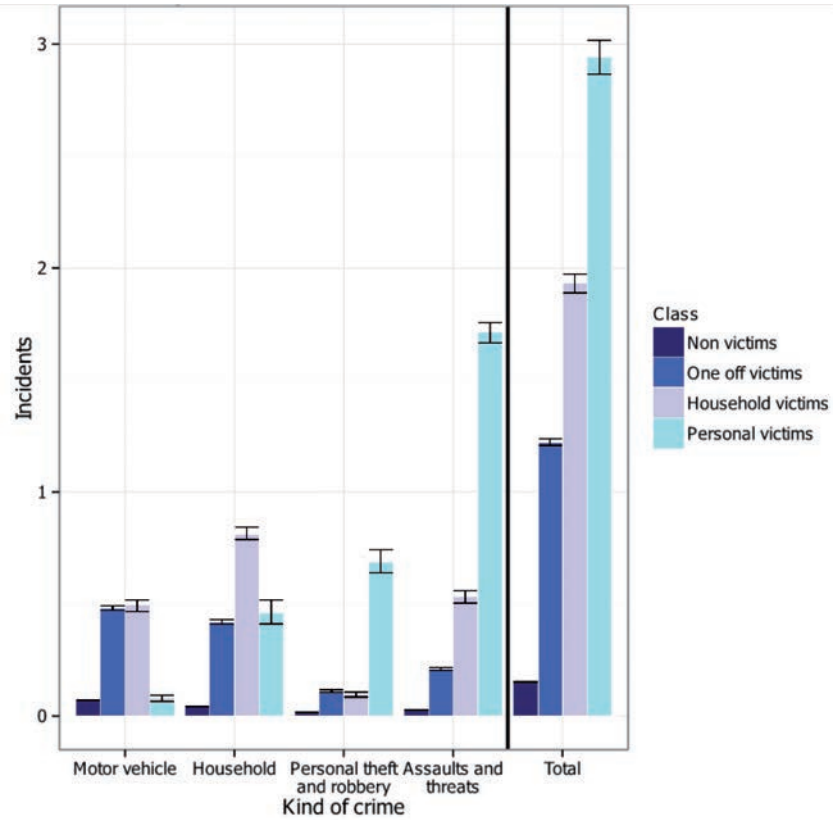
It is beyond the scope of this chapter to consider how the explanatory variables collected as part of a victimisation survey can help understand the differing risk of victimisation across a population. An example of such analysis

using bivariate statistical methods can be found in Scottish Government (2014, pp. 25–35), while the work of Tseloni et al. (2002) provides a detailed example concerning property crime victimisation.

One area where survey data outperforms police recorded statistics is in the understanding of repeat victimisation, an issue which can have important policy implications as it illustrates how victimisation is often focused on specific individuals (or groups) who may benefit from specific police interventions. Police statistics are typically only provided in aggregation and, as such, it is not possible to identify if an individual has been the victim of multiple crimes. Applying Latent Class Analysis to survey responses from Scotland between 1992 and 2011 suggests that the population can be split into four groups with varying experiences of victimisation (*Figure 5*). The majority of people (just under 80%) can be considered non-victims (they have only a small risk of experiencing any victimisation). One-off victims experience an average of one incident per year, typically motor vehicle or household theft or vandalism (15% of respondents). A third group, so called Household Victims (5%), experience around two incidents of victimisation a year, while a very small proportion of the population (0.5%) experience multiple incidents of victimisation across different crime types, but notably assaults and threats (Personal Victims).

Figure 5:

Latent Class Representation of Groups of Victims in Scotland 1992–2011



Bars represent average number of incidents per person – with 95% confidence intervals.

Changes in the aggregate level of victimisation (which can be identified from both survey data and recorded crime statistics) can be better understood by considering how the prevalence and average level of victimisation for the groups identified above has varied over time. Overall victimisation has fallen in Scotland since 1992. However, this drop has not been experienced equally by the four groups identified in *Figure 5*. In line with the overall drop in victimisation, the probability of being a non-victim has increased consistently over time (from 76% to 80% between 1992 and 2011). Yet while the number of respondents appearing in the One-off and Household victim groups has fallen, the number of Personal Victims (who experience the highest levels of

repeat victimisation) has remained constant. Similarly, while the average number of crimes experienced by those classified as Household or One-Off Victims fell between 1992 and 2011, Personal Victims only experienced a fall in the first part of the 1990s (and this was only restricted to motor vehicle and household crime). This suggests that the fall in crime experienced in Scotland since the early 1990s is not evenly shared across the population (see Norris et al. 2014 for full details). Since this analysis relies on information about an individual experience of repeat victimisation, as well as those who have not experienced any crime, this is a conclusion which cannot be arrived at through the use of recorded crime data.

10 Summary

This chapter has considered how the comparability of victimisation survey data and police recorded crime statistics can be influenced by a range of decisions around the process of data collection. The strength of victimisation surveys lies not just in their (partial) ability to address, and understand, the underrepresentation of victimisation in recorded crime statistics, but also in the additional analysis that their data allow to be undertaken, for instance, understanding an individual's pattern of repeat victimisation. The following key points summarise this chapter:

- Recorded crime statistics underestimate victimisation because victims might choose not to inform the police of their experience, and because not all incidents are defined as crimes by the police.

- Victimisation surveys aim to overcome the limitations of recorded crime statistics by asking individuals about their experience of crime.

- Like recorded crime statistics, victimisation surveys do not measure the full extent of victimisation. They are restricted in the crimes they cover, the population whose experience they represent, and through the counting rules they employ.

- The ability to compare estimates of victimisation between recorded crime statistics and victimisation surveys is dependent on both sources using a consistent classification of events.

- Comparisons over time (either within one data source, or between recorded crime statistics and victimisation surveys) are often limited by changes in methodology, data collection processes, and the process used to classification crimes.

- Comparing estimates of victimisation between sources can help highlight how changes in victim behaviour, and policing practice, influence the level of recorded crime.

- For surveys to help contextualise recorded crime statistics it is necessary to collect data on whether particular events were bought to the attention of the police and background data on the event and victim.

- There are a range of research question around victimisation, such as understanding the characteristics of victims, and patterns of repeat victimisation, which cannot be understood through the analysis of aggregate recorded crime data. Addressing these issues is a strength of victimisation surveys.

11 Bibliography

Aebi, Marcelo F.; Linde, Antonia (2012): Crime Trends in Western Europe According to Official Statistics from 1990 to 2007. In: van Dijk, Jan; Tseloni, Andromachi and Farrell, Graham (Ed.): The International Crime Drop: New Directions in Research. London: Palgrave Macmillan, pp. 37–75.

Baumer Eric P.; Lauritsen, Janet L. (2010): Reporting Crime to the Police, 1973–2005: A Multivariate Analysis of Long-Term Trends in the National Crime Survey (NCS) and National Crime Victimization Survey (NCVS). In: Criminology, Vol. 48, pp. 131–185.

Brown, Matthew; Bolling, Keith (2007): 2006 Scottish Crime and Victimisation Survey: Main Findings. Edinburgh: Scottish Government.

Cantor, David; Lynch, James P. (2000): Self-Report Surveys as Measures of Crime and Criminal Victimization. In: Duffee, David (Ed.): Measurement and Analysis of Crime and Justice: Criminal Justice 2000. Washington: National Institute of Justice, pp. 85–138.

Coleman, Clive; Moynihan, Jenny (1996): Understanding Crime Data: Haunted by the Dark Figure. Buckingham: OUP.

Farrell, Graham; Pease, Ken (2007): The Sting in the Tail of the British Crime Survey: Multiple Victimisation. In: Hough, Mike; Maxfield, Mike (Ed.): Surveying Crime in the 21st Century. London: Lynne Rienner Publishers, pp. 33–53.

Flatley, John; Bradley, Jenny (2013): Analysis of variation in crime trends: A study of trends in 'comparable crime' categories between the Crime Survey for England and Wales and the police recorded crime series between 1981 and 2011/12, Office of National Statistics. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/methodological-note–analysis-of-variation-in-crime-trends.pdf – cited 07. 11. 2014.

Goudriaan, Heike; Wittebrood, Karin and Nieuwbeerta, Paul (2006): Neighbourhood Characteristics and Reporting Crime: Effects of Social Cohesion, Confidence in Police Effectiveness, and Socio-economic Disadvantage. In: British Journal of Criminology, Vol. 46, pp. 719–742.

Gottfredson, Michael R.; Hindelang, Michael J. (1976): A Consideration of Telescoping and Memory Decay Biases in Victimization Surveys. In: Journal of Criminal Justice, Vol. 5, pp. 205–216.

Hoare, Jacqueline (2010): Drug Misuse Declared: Findings from the 2009/10 British Crime Survey, Home Office Statistical Bulletin, 13, 10. London: Home Office.

Hough, Mike; Jackson, Jonathan and Bradford, Ben (2013): Legitimacy, Trust and Compliance: An Empirical Test of Procedural Justice Theory using the European Social Survey. In: Tankebe, Justice; Liebling, Alison

(Ed.): Legitimacy and Criminal Justice: an International Exploration. Oxford: Oxford University Press, pp. 326–352.
Hough, Mike; Maxfield, Mike (Ed.)(2007): Surveying Crime in the 21st Century. London: Lynne Rienner Publishers.
Hough, Mike; Norris, Paul (2009): Comparisons between Survey Estimates of Crime and Crimes Recorded by the Police: The UK Position. In: Robert, Philippe (Ed.): Comparing Crime Data in Europe. Brussels: VUB-Press, pp. 103–126.
Kershaw, Chris; Nicholas, Sian and Walker, Alison (2008): Crime in England and Wales 2007/08. London: Home Office.
Lynn, Peter; Elliot, Dave (2000): The British Crime Survey: A Review of Methodology, URL: http://webarchive.nationalarchives.gov.uk/20110218135832/http:/rds.homeoffice.gov.uk/rds/pdfs08/bcs-methodology-review-2000.pdf – Download from 28. 02. 2015.
MacQueen, Sarah; Norris, Paul (2014): Police awareness and involvement in cases of domestic and partner abuse, Policing and Society. DOI: 10108010439463201492208 4 – Download from 07. 11. 2014.
Maguire, Mike (2012): Criminal Statistics and the Construction of Crime. In: Maguire, Mike; Morgan, Rod and Reiner, Robert (Ed.): Oxford Handbook of Criminology, 5th Edition. Oxford: OUP, pp. 206–244.
Mayhew, Pat; Aye Maung, Natalie; Mirrless-Black, Catriona (1993): The 1992 British Crime Survey. London: Home Office.
McClintock, Frederick H.; Avison, Howard (1968): Crime in England and Wales. London: Heinemann.
Murphy, Linda R.; Cowan, Charles D. (1976): Effects of Bounding on Telescoping in the National Crime Survey. In: Lehnen, Robert G.; Skogan, Wesley G. (Ed.): The National Crime Survey: Working papers, Volume II: Methodological studies. Washington: U.S. Bureau of Justice Statistics, pp. 83–89.
Norris, Paul; Pillinger, Rebecca and McVie, Susan (2014): Changing patterns of victimisation in Scotland 1993–2011, AQMeN Research Briefing. Edinburgh: AQMeN.
ONS (2014): Statistical Bulletin: Crime in England and Wales, Year Ending March 2014. URL: http://www.ons.gov.uk/ons/dcp171778_371127.pdf – Download from 07. 11. 2014.
Robert, Philippe (Ed.) (2009): Comparing Crime Data in Europe, Brussels: VUBPress.
Sentencing Council (2014): Facts and Figures. URL: http://sentencingcouncil.judiciary.gov.uk/facts/facts-and-figures.htm – Download from 07. 11. 2014.
Sparks, Richard; Glenn, Hazel and Dodd, David (1977): Surveying Victims. London: Heinemann.

Tarling, Roger; Morris, Katie (2010): Reporting Crime to the Police. In: British Journal of Criminology, Vol. 50, pp. 474–490.
Tseloni, Andromachi; Osborn, Denise R.; Trickett, Alan and Pease, Ken (2002): Modelling Property Crime Using the British Crime Survey. In: British Journal of Criminology, Vol. 42, pp. 109–128.
van Dijk, Jan; Castelbajac, Matthieu de (2015): The hedgehog and the fox; the history of victimisation surveys from a Trans-Atlantic perspective. In: Guzy, Nathalie; Birkel, Christoph and Mischkowitz, Robert (Ed.): Viktimisierungsbefragungen in Deutschland, Band 1: Ziele, Nutzen und Forschungsstand, Wiesbaden: Bundeskriminalamt, pp. 10–28.

Vergleichsschwierigkeiten und Kombinationsmöglichkeiten

Wolfgang Heinz

1 Auf der Suche nach einem Messinstrument für „Kriminalitätswirklichkeit“

Aktuelle, umfassende und verlässliche Daten zu Umfang, Struktur und Entwicklung von Kriminalität sind eine notwendige (wenngleich keine hinreichende) Bedingung für rationale Kriminalpolitik, für organisatorische Planungen, für die Konzeption von Präventions- und Interventionsansätzen, für Kontrolle und für Evaluationsmessungen sowie für eine zutreffende Unterrichtung der Öffentlichkeit über die innere Sicherheit. Generationen von Kriminalstatistikerinnen bzw. Kriminalstatistikern schienen die Daten der amtlichen Kriminal- und Strafrechtspflegestatistiken diesen Zwecken zu genügen. Sie glaubten, von der bereits 1835 von Quetelet postulierten „stillschweigenden Annahme“ ausgehen zu können, zwischen der statistisch erfassten Kriminalität und der „Totalsumme begangener Verbrechen“ bestehe „ein beinahe unveränderliches Verhältnis“.[1] Diese später zum „Gesetz der konstanten Verhältnisse“[2] hochstilisierte Annahme ist indes empirisch widerlegt. Die Eignung einer Verbrechensrate zu Zwecken der Indexbildung sinkt, „je weiter sich das Stadium ihrer statistischen Erfassung von der Begehung der Straftaten entfernt“ (Sellin 1931, 589). Allerdings erwies sich die Annahme als unzutreffend, Daten der Polizeilichen Kriminalstatistik (PKS) seien wegen ihrer (im Vergleich zur Strafverfolgungsstatistik größeren) Tatnähe ein geeigneter Kriminalitätsindex, denn auch die polizeilich registrierten Fälle stellen nur einen und überdies in mehrfacher Hinsicht verzerrten Ausschnitt der „Kriminalitätswirklichkeit“ dar. Fehlschlüsse hinsichtlich Struktur und Entwicklung der Kriminalität sind auf dieser Basis unvermeidbar (Prätor 2014, 33 ff.). Die delikt- und opfergruppenspezifisch sowie regional und zeitlich unterschied-

---

1 Quetelet 1835, Bd. 2, 173 f. (zitiert nach: Quetelet 1921, 253).

2 Danach sollen sich „unter normalen Verhältnissen“ die wirkliche Kriminalität (K), die zur Anzeige gelangende Kriminalität (A), die abgeurteilte Kriminalität (U) und die zur Verurteilung führende Kriminalität (V) „ziemlich nahe kommen. Auf jeden Fall werden dann die Größen A, U und V symptomatische Begleiter von K bilden und so ziemlich alle Veränderungen, denen dieser Faktor unterworfen ist, proportional mitmachen. Man könnte diese Regelmäßigkeit in den Beziehungen füglich das ‚Gesetz der konstanten Verhältnisse‘ nennen“ (Wadler 1908, 15).

lich hohe Anzeigewahrscheinlichkeit[3] ist der größte Verzerrungsfaktor; Struktur und Entwicklung der Kriminalität können „fast als direkte Funktion der Anzeigebereitschaft der Bevölkerung definiert werden“ (Pudel 1978, 205). Es lag deshalb nahe, durch eine noch größere Nähe zur Tat, also durch Erhebung von Daten beim Täter oder beim Opfer, „zu einer gültigen Schätzung der tatsächlichen Verbreitung der [...] Delikte zu gelangen“ (Schwind u. a. 2001, 113).

Von den Pionieren der Dunkelfeldforschung wurde angenommen, durch Opferbefragungen ließe sich die Zahl der nicht angezeigten Delikte ermitteln. Verkannt wurde hierbei, dass „Kriminalität“ – und zwar sowohl im Hellfeld als auch im Dunkelfeld – kein naturalistischer, objektiv zu messender Sachverhalt ist. Deskriptive, einen Beobachtungssachverhalt feststellende Aussagen sind von askriptiven, ihn bewertenden Aussagen zu trennen. Was als „Kriminalität“ wahrgenommen wird, ist sowohl das Ergebnis vorgängiger gesellschaftlicher Festlegungen als auch (zumeist) mehrstufig erfolgender Prozesse der Wahrnehmung von Sachverhalten, deren Interpretation und Bewertung. Dies heißt, dass dasselbe Ereignis von unterschiedlichen Akteurinnen und Akteuren unterschiedlich definiert werden, es also mehrere Realitätsdefinitionen[4] geben kann – die der befragten Opfer, dritter Personen bzw. des/der Interviewers/-in bzw. Wissenschaftlers/-in usw. –, die wiederum abweichen können von den Definitionen der jeweiligen Instanzen sozialer Kontrolle. Das Dunkelfeld[5] der Taten bezeichnet danach die Zahl der Fälle, die zwar von den Befragten (oder von den sie interviewenden Wissenschaftlern/-innen) als „Straftat“ kategorisiert wurden, nicht aber von der Polizei.[6] Entsprechend lässt sich von einem Dunkelfeld der Täter, der Verurteilten usw. sprechen.[7] Dieser Ausgangspunkt sollte immer mitbedacht werden, wenn aus Gründen sprachlicher Vereinfachung im Folgenden auch reobjektivierende Termini verwendet werden.

---

3 Ausführlich zu den Determinanten der Anzeigebereitschaft Baier u. a. 2009, 43 ff.; Köllisch 2009, 28 ff.

4 Da es kein Kriterium gibt, um eine bestimmte Definition als „richtig“ bzw. „wahr“ zu bestimmen, gibt es bei Definitionskonflikten verschieden große Dunkelfelder.

5 Wird nur die Definitionsleistung der Reaktionsinstanzen als relevant anerkannt, dann gibt es freilich weder begrifflich noch erkenntnistheoretisch ein „Dunkelfeld“ (so Ditton 1979, 20 f.; zutreffend dagegen Dellwing 2010).

6 Nur erwähnt, nicht weiter erörtert werden kann, dass ein Teil der Ereignisse zwar der Polizei bekannt ist, von ihr aber nicht als „kriminell“ bewertet bzw. registriert wird (Antholz 2010; Kürzinger 1978, 159).

7 Zu verschiedenen Dunkelfeldbegriffen Coleman/Moynihan 1996, 1 ff., insbesondere 16 f.; Kreuzer u. a. 1993, 14 f.; Prätor 2014, 32 f.

Ob und inwieweit die in Opferbefragungen erfassten Definitionsleistungen[8] mit den Definitionsleistungen der Polizei verglichen werden können, ist davon abhängig, inwieweit sich Ziele, Deliktsgruppen, Opfergruppen, Erfassungsregeln und Referenzzeiträume überschneiden, inwieweit die jeweiligen Daten "valide" sind und ob und inwieweit vergleichbare Belastungszahlen berechnet werden können.

2 (Teil-)Überschneidung der Ziele von Polizeilicher Kriminalstatistik und von Opferbefragungen

2.1 Polizeiliche Kriminalstatistik

Die Polizeiliche Kriminalstatistik (PKS) ist ein Tätigkeitsbericht der Polizei, der die polizeilichen Bewertungen aufgrund der dieser verfügbaren Sachinformationen entsprechend den Erfassungsrichtlinien abbildet – sie ist eine „Konstruktion polizeilich registrierter Kriminalität".[9] Sie soll „im Interesse einer wirksamen Kriminalitätsbekämpfung zu einem überschaubaren und möglichst verzerrungsfreien Bild der angezeigten Kriminalität führen" (PKS 2012, 6).

2.2 Opferbefragungen

Ziel der ersten, in den USA durchgeführten Opferbefragung war eine alternative „Kriminalitätsmessung", um das „wahre" Kriminalitätsaufkommen ermitteln und die offiziellen Kriminalstatistiken „korrigieren" zu können.[10] Sehr rasch wurde indes erkannt, dass keine „objektive" Messung möglich ist, da die Befragungsergebnisse Bewertung, Vorerfahrungen, Auskunftsbereitschaft usw. der Opfer widerspiegeln. Deshalb wurde die auf kriminalstatistische Funktionen beschränkte Zielsetzung von Opferbefragungen („crime sur-

---

8 In Opferbefragungen werden zumeist nur Ereignisse abgefragt. Die Einordnung als „kriminell" stammt dann entweder von der Interviewerin bzw. dem Interviewer, die bzw. der dieses Ereignis bewertet und entsprechend erfasst oder – werden nur vorgegebene Fallgruppen abgefragt – von der bzw. dem Befragten und von der Konstrukteurin bzw. vom Konstrukteur des Fragebogens.

9 Zugespitzt formuliert Kunz: „Sie drückt nicht registrierte Kriminalität, sondern Registrierungsverhalten der strafrechtlichen Kontrollinstanzen aus. Die Grundeinheit, welche in Kriminalstatistiken aufgezeichnet wird, ist nicht das raumzeitliche Geschehen einer kriminellen Handlung, sondern die amtliche Registrierung und Rekonstruktion des angenommenen Verdachts eines solchen Geschehens" (Kunz 2011, § 19 Rn. 5, ebenso Kunz 2008, 16).

10 Zu Entwicklung und Stand der Dunkelfeldforschung statt vieler Heinz 2006; Prätor 2014.

vey“) verändert und erweitert auf die Gewinnung opferbezogener Erkenntnisse („victim survey“).[11]

1. Die Erhebung von Informationen über Viktimisierungserlebnisse in einem bestimmten Referenzzeitraum dient dazu, einen weiteren Indikator für „Kriminalität“ zu gewinnen, der unabhängig vom selektiven und im Zeitverlauf sich wandelnden Anzeigeverhalten ist.

2. Die mitgeteilten Viktimisierungserfahrungen erlauben es, nach Selbstwahrnehmungen und -bewertungen Risikogebiete sowie besonders gefährdete Gruppen zu identifizieren (Risikopopulationen), die in der gegenwärtigen PKS noch nicht mit Opferkennungen versehen sind.

3. Durch Ermittlung des Anzeigeverhaltens sollen die deliktspezifischen Größenordnungen von Dunkelfeldanteilen sowie deren Veränderungen im zeitlichen Längsschnitt bestimmt und auf diese Weise (mögliche) unterschiedliche Trends im Dunkelfeld gegenüber denjenigen im Hellfeld erklärt werden. Bei vorliegenden Informationen über eine zunehmende Anzeigebereitschaft würde der durch vermehrte Anzeigen verursachte Anstieg der PKS-Daten nicht (wie gegenwärtig zumeist) als Indikator der Verschärfung eines gesellschaftlichen Problems, sondern z. B. als Ausdruck einer Verbesserung des Verhältnisses zwischen Bürgern/-innen und Polizei sowie einer vermehrten Erfassung entsprechender Vorfälle und damit gegebenenfalls als Erfolg kriminalpolitischer Interventionen gesehen werden können.

4. Durch die Erhebung der Gründe für Anzeige bzw. Nichtanzeige wird eine zentrale Hintergrundvariable für die Ausprägung des Anzeigeverhaltens erfasst. Über die Zeit hinweg liefert eine Veränderung dieser Gründe Anhaltspunkte dafür, weshalb und in welchen Bereichen sich die Hellfeld-/Dunkelfeldrelationen verändert haben.

5. Die Erfassung sowohl des objektiven Schweregrads (materielle und immaterielle Schäden) als auch der subjektiven Seite der Opfererfahrungen (unmittelbare psychische Folgen sowie langfristige psychosoziale Auswirkungen) sowie der Verarbeitung derartiger Ereignisse und der daraufhin getroffenen Vorsichtsmaßnahmen liefert Informationen über die Bedeutsamkeit von Viktimisierungserfahrungen aus Sicht der Opfer. Im zeitlichen Längsschnitt lässt eine etwaige Veränderung nicht nur erkennen, wo-

[11] Die folgende Darstellung der Ziele von Opferuntersuchungen orientiert sich am Abschlussbericht des Verfassers (BUKS 2002).

von sich die Bürgerinnen und Bürger vor allem betroffen fühlen, sondern gibt auch Anhaltspunkte darüber, wie sich die Schwereeinschätzung entwickelt, was wiederum Rückwirkungen auf das Anzeigeverhalten und damit auf Ausmaß und Struktur des Hellfelds haben kann. Nicht zuletzt liegen hier auch Ansatzpunkte, um einen möglicherweise sich im Zeitverlauf ändernden Unterstützungs- und Hilfebedarf erkennen zu können.

6. Gegenstand von Opferbefragungen können ferner spezielle Themen sein, so insbesondere die Strafbedürfnisse von Opfern, die genaue Charakterisierung der Täter-Opfer-Beziehung, von Opfererfahrungen im familiären/häuslichen Kontext, alltagsweltliche Möglichkeiten der informellen Regulierung strafrechtlich relevanter Konflikterlebnisse bzw. die Verfügbarkeit informeller sozialer Unterstützungssysteme zur Bewältigung von Opfererlebnissen.

7. Die Erfassung der opferseitigen Wahrnehmung und Bewertung polizeilicher und gegebenenfalls justizieller Reaktionen dient der Bestimmung der Akzeptanz, Nutzung und Bewertung von (polizeilichen und anderen) Angeboten an Hilfe und Beratung bei der Bewältigung der Folgen der Straftat; sie gibt des Weiteren Hinweise auf etwaige Akzeptanzhindernisse. Fragen über die Kenntnis und Inanspruchnahme von Hilfeeinrichtungen dienen dazu, Hinweise auf Erwartungen und Bedürfnisse der Opfer von Straftaten hinsichtlich derartiger Unterstützungseinrichtungen zu erhalten.

8. Durch die Erhebung der verschiedenen Dimensionen von „Kriminalitätsfurcht“ sollen nicht nur deren Ausmaß und Entwicklung in verschiedenen Opfergruppen erfasst, sondern auch die relative Bedeutsamkeit von (unmittelbaren oder mittelbaren) Viktimisierungserfahrungen auf das Ausmaß von „Kriminalitätsfurcht“ festgestellt werden.

9. Fragen zur Schwereeinschätzung von Kriminalitätsformen dienen dazu, Informationen über die moralische Bewertung von Verhaltensweisen in der Bevölkerung, folglich über die Akzeptanz unterschiedlicher Strafnormen in der Bevölkerung und den Stellenwert eines Delikts in Relation zu anderen Delikten zu gewinnen.

10. Erfolgreiche Kriminalpolitik setzt voraus, dass die Bevölkerung die Institutionen der Strafrechtspflege, namentlich Polizei und Gerichte, nicht ablehnt, sondern möglichst positiv einschätzt. Geeignete Einstellungsfragen erlauben z. B. spezifische Bewertungen der Polizeiarbeit oder die Einschätzung der Strafpraxis der Gerichte.

11. Fallbezogene Fragen zur Strafe bzw. Sanktionierung dienen dazu, nicht nur das relative Maß der Übereinstimmung mit den Strafnormen und deren Anwendung festzustellen, sondern auch Gestaltungsspielräume der Kriminal- und Strafrechtspolitik aufzuzeigen.

Die Gewinnung von Informationen über Umfang und Veränderung von Viktimisierungserfahrungen ist folglich nur eines unter mehreren Zielen von Opferbefragungen. Deren Potenzial wird erst dann ausgeschöpft, wenn Opferwerdung, Opfererleben, dessen Verarbeitung und Folgen erfasst werden – und zwar nicht nur im Querschnitt, sondern vor allem im zeitlichen Längsschnitt durch Befragungswiederholung.

3 „Kriminalitätsmessung“ durch Opferbefragungen und PKS im Vergleich

3.1 (Teil-)Überschneidung von Delikten bzw. Deliktsgruppen

Die PKS erfasst fast alle[12] der Polizei bekannt gewordenen und von ihr als strafrechtlich relevant bewerteten Sachverhalte (Hellfeld der polizeilich registrierten „Kriminalität“) mit inländischem Tatort. Die Erfassung erfolgt unabhängig davon, ob das Opfer zur Wohnbevölkerung gehört oder nicht, ob es In- oder Ausländer ist, ob es sich um eine natürliche oder um eine juristische Person handelt oder ob ein sogenanntes opferloses Delikt vorliegt.

Im Unterschied zur PKS kann Gegenstand von Opferbefragungen nur ein Deliktsspektrum sein, von dem die Befragen persönlich betroffen sein können.[13] Es scheiden deshalb solche Sachverhalte aus, die im strengen Sinn kein Opfer haben (*victimless crimes*) bzw. sich nicht unmittelbar gegen Privatpersonen richten, ferner Delikte, bei denen das Opfer naturgemäß keine Angaben (mehr) machen kann, wie z. B. vollendete Tötungsdelikte. Relativ gut erfassbar sind also vor allem Eigentums- und Körperverletzungsdelikte, die sich gegen Privatpersonen richten, wobei zwischen haushalts- und personenbezogenen Delikten unterschieden werden sollte,[14] d. h. zwischen Delikten, durch die nur der Befragte selbst, und solchen, durch die alle Mitglieder des Haushalts geschädigt wurde(n), wie z. B. Wohnungseinbruch oder Autodiebstahl.

---

12 Von den polizeilich bearbeiteten Straftaten sind in der PKS statistisch nicht erfasst vor allem Staatsschutzdelikte und Straßenverkehrsdelikte.

13 Auf spezielle Befragungen bei Unternehmen, beispielsweise zur Wirtschafts- und Unternehmenskriminalität, kann hier nur hingewiesen, aber nicht eingegangen werden.

14 Zum Beispiel BUKS 2002, 35 f. oder die „Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage“ (Brings u. a. 2010; Fuhr/Guzy 2010, 637).

Erhebungseinheiten der PKS sind der „Fall“, der bzw. die „Tatverdächtige“ und – für einen eingeschränkten Deliktskreis[15] – das „Opfer“. Die PKS orientiert sich an Straftatbestimmungen, die unter kriminologisch-kriminalistischen Gesichtspunkten weiter differenziert werden. So existieren derzeit für die Erfassung des einfachen Diebstahls 57, des schweren Diebstahls 86, der Körperverletzung 13, des Raubes, der Erpressung und des räuberischen Angriffs auf Kraftfahrer 46 selbstständige Schlüsselzahlen.[16]

Eine derart differenzierte Aufschlüsselung ist in Opferbefragungen unmöglich. Hier muss ein Ereignis, das einen Kernbereich eines Straftatbestands erfassen soll, für möglichst alle Gruppen von Befragten verständlich beschrieben werden. Eine zu differenzierte Befragung würde schon wegen des dann erforderlichen Zeitaufwands die Mitwirkungsbereitschaft der Befragten deutlich senken. Deshalb wird z. B. zur Erfassung eines als Raub bewerteten Ereignisses regelmäßig nur gefragt, ob jemand dem Befragten „persönlich mit Gewalt oder unter Androhung von Gewalt etwas weggenommen“ oder ihn „gezwungen hat, etwas herzugeben“, bzw. dies versucht hat.[17]

Während in der PKS nur Ereignisse mit inländischem Tatort erfasst werden, werden in Opferbefragungen Viktimisierungen in der Regel unabhängig vom Tatort erfasst, also auch dann, wenn sie z. B. während eines Auslandaufenthalts erfolgt sind.

3.2 (Teil-)Überschneidung der Opfergruppen

3.2.1 Polizeiliche Kriminalstatistik

In der PKS werden die der Polizei bekannt gewordenen Fälle unabhängig davon erfasst, welche Merkmale das Opfer aufweist, also unabhängig davon, wie alt es ist, ob es sich in deutscher Sprache verständigen kann, wo es lebt (z. B. in einer geschlossenen Einrichtung), ob es seinen Wohnsitz in Deutschland hat oder Tourist/-in, Durchreisende/r usw. ist, ob es sich um eine natürliche Person, um eine Institution/juristische Person oder um die Allgemeinheit (überindividuelles Rechtsgut) handelt und wann sich der angezeigte Sachverhalt ereignet hat. Nicht erfasst werden hingegen im Ausland verübte Straftaten, selbst wenn sie sich gegen Inländer richten.

---

15 Die bundesweite Opfererfassung in der PKS beschränkt sich derzeit noch auf Delikte, bei denen Leib oder Leben bzw. Gesundheit eines Menschen unmittelbar gefährdet bzw. geschädigt wurde.

16 Bundeskriminalamt (Hg.): Polizeiliche Kriminalstatistik PKS 2012, Straftatenschlüsselverzeichnis.

17 So z. B. im „Viktimisierungssurvey 2012“.

3.2.2 Opferbefragungen

Allgemeine Opferbefragungen gewinnen ihre Daten aus Befragungen einer repräsentativen Stichprobe der Bevölkerung, die ihren Wohnsitz in Deutschland hat. Im Unterschied zur PKS werden deshalb Viktimisierungen der Befragten in der Regel auch dann berücksichtigt, wenn sie sich im Ausland ereignet haben.[18] Andererseits wird ein nicht unerheblicher Teil der in der PKS erfassten Fälle nicht von Opferbefragungen erfasst, weil bestimmte Personengruppen aus erhebungstechnischen Gründen[19] zumeist nicht in der Grundgesamtheit berücksichtigt werden:[20]

– In einer repräsentativen Stichprobe der Bevölkerung, die ihren Wohnsitz in Deutschland hat, sind alle diejenigen nicht erfasst, die entweder nicht meldepflichtig sind, wie z. B. Touristen, Durchreisende, Berufspendler usw., oder sich nicht gemeldet haben, wie z. B. sich illegal aufhaltende Personen.[21]

– Ausgeschlossen werden ferner zumeist Kinder oder unter 16 Jahre alte Personen,[22] vielfach wird auch eine obere Altersgrenze eingeführt.

– Aus Kostengründen ist die Stichprobe zumeist auf Personen beschränkt, die die deutsche Sprache hinreichend gut verstehen.[23] In einigen neueren

---

18 In der „Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage“ wurde festgestellt, dass „knapp 15 % aller Raubopfer und 10 % aller Diebstahlsopfer nicht in Deutschland viktimisiert“ (Fuhr/Guzy 2010, 640) worden waren. Im „Viktimisierungssurvey 2012“ wurden deshalb Opfererlebnisse in Deutschland identifiziert (Birkel 2014b, 85), um Vergleiche mit der PKS zu ermöglichen.

19 Je nach gewählter Erhebungsmethode werden u. U. andere bzw. noch weitere Einheiten ausgeschlossen, z. B. bei CATI-Befragungen Personen, die (weil sie z. B. schon Opfer geworden sind) sich nicht mehr in die öffentlichen Telefonregister eintragen lassen, die über kein Telefon oder nur über ein Mobiltelefon (*Mobile Onlys*) verfügen usw. Dies ist nicht ergebnisneutral, denn es gibt Grund zur Annahme, dass die *Mobile Onlys* ein erhöhtes Viktimisierungsrisiko aufweisen (Guzy 2014, 154).

20 Diese in allgemeinen Stichproben regelmäßig ausgeschlossenen Teilgruppen können aber u. U. durch Spezialmodule erfasst werden.

21 In den USA wurden (auch) deshalb die anfänglich durchgeführten *City Surveys* wieder eingestellt, weil in der Grundgesamtheit des *Uniform Crime Report* alle Fälle erfasst waren, die in der jeweiligen Stadt geschahen. Im *City Survey* wurden aber nur Fälle erfasst, bei denen Einwohner/-innen der Stadt Opfer waren.

22 So im „Viktimisierungssurvey 2012“. Zu neueren Befragungen von Kindern bzw. Jugendlichen siehe die Beiträge von Baier sowie Stadler/Kemme im ersten Teil des Sammelbands.

23 Mangels Dokumentation der Zahl der kontaktierten, aber mangels Sprachkenntnissen nicht befragten Personen ist aus früheren Opferuntersuchungen nicht bekannt, wie groß der Anteil der deshalb ausgeschlossenen Personen ist. Bezogen auf die Gesamtstichprobe dürfte er gering, bezogen auf die Teilgruppe der Personen mit Migrationshintergrund dagegen groß sein (Birkel 2014b, 72 Anm. 5 m. w. N.).

deutschen Opferbefragungen wurden die Fragebögen zumindest für einige der großen Migranten/-innengruppen übersetzt.[24]

– In der Regel bilden Personen, die in Privathaushalten leben, die Grundgesamtheit. Ausgeschlossen sind damit Obdachlose sowie Personen in geschlossenen Institutionen (Strafanstalten, Krankenhäuser, Alters- und Pflegeheime), wobei es sich hierbei teilweise um Gruppen mit überdurchschnittlich hohem Viktimisierungsrisiko handelt.[25]

– Der regionale Vergleich der Ergebnisse ist dadurch beeinträchtigt, dass für die PKS der Tatort, für Opferbefragungen aber der Wohnort maßgeblich ist.

– Schließlich bleiben alle Delikte unberücksichtigt, die kein persönliches Opfer haben, also insbesondere alle Vorfälle, durch die Unternehmen, Institutionen, juristische Personen usw. sowie die Allgemeinheit geschädigt worden sind.[26]

Je nach Erhebungsmethode werden in der Stichprobe einige Personengruppen systematisch unterrepräsentiert erfasst, insbesondere überdurchschnittlich mobile Personen, die zumeist ein hohes Viktimisierungsrisiko aufweisen. Eine weitere Verzerrung besteht möglicherweise dann, wenn intensiv viktimisierte Personen überdurchschnittlich häufig die Teilnahme an der Befragung verweigern.

Ein exakter Vergleich der Ergebnisse von Opferbefragungen mit Daten der PKS ist deshalb selbst dann nicht möglich, wenn sich die Opfererfassung der PKS künftig auch auf Eigentums- und Vermögensdelikte erstreckt, Institutionen und natürliche Personen getrennt erfasst, der Aufenthaltsstatus der Opfer sowie das Land der Viktimisierung erhoben und schließlich nicht Inzidenzen, also Fallzahlen, sondern Prävalenzen, also Personenzahlen, miteinander verglichen werden würden. Wegen der fehlenden Information zu den hinreichend guten deutschen Sprachkenntnissen der Opfer ist nämlich keine Übereinstimmung der Grundgesamtheiten herstellbar. Mildern lässt sich diese partielle

---

24 Im „Viktimisierungssurvey 2012“ wurde in die russische und türkische Sprache übersetzt. Die computergestützte telefonische Befragung wurde bei Bedarf von zweisprachigen Interviewern durchgeführt.

25 In Spezialmodulen oder durch spezielle Befragungen ist – freilich begrenzt – auch die Viktimisierung in derartigen Einrichtungen erfassbar (siehe die Beiträge von Görgen sowie von Neubacher/Hunold im ersten Teil des Sammelbands). Zur Viktimisierung im Justizvollzug zuletzt m. w. N. Neubacher 2014, 485 ff.; zur Viktimisierung älterer Menschen Görgen 2010.

26 In der PKS lassen sich – wegen der auf wenige Deliktsgruppen beschränkten Opfererfassung – die Fälle ohne persönliches Opfer nur teilweise bestimmen, z. B. über eine Deliktskategorie wie „Ladendiebstahl“ (Birkel 2014b, 86).

Nichtübereinstimmung, wenn künftig auch große Migrantengruppen durch entsprechend übersetzte Fragebögen einbezogen werden würden.

Für den Vergleich von Fallzahlen und Opferinzidenzen gibt es freilich keine entsprechende Lösung. Auf absehbare Zeit wird aber in Deutschland vornehmlich eine Kontrastierung von Fallzahlen und Inzidenzen in Betracht kommen, weil für die Mehrzahl der in Viktimisierungsstudien berücksichtigten Delikte keine Angaben zu den Opfern erhoben werden. Für den absehbaren Zeitraum ist lediglich für Delikte mit Opfererfassung, wie Körperverletzungsdelikte, über eine PKS-Sonderauswertung nach Opferalter eine begrenzte Vergleichbarkeit herstellbar.

3.3 (Teil-)Überschneidung der Erfassungsregeln

Die Erfassungsregeln der PKS weichen von den Erfassungsregeln einer Opferbefragung ab. In der PKS ist der Fall die Zähleinheit,[27] in der Opferbefragung dagegen die Viktimisierungserfahrung eines individuellen Opfers. Daraus ergeben sich Unterschiede der Zählung vor allem in folgenden Fallgestaltungen, die freilich insgesamt nicht sehr häufig praktisch relevant werden dürften:

– Jede bekannt gewordene rechtswidrige Handlung ist in der PKS ohne Rücksicht auf die Zahl der Geschädigten als ein Fall zu erfassen. Bei einem Wohnungseinbruch wird nur ein Fall erfasst, unabhängig davon, wie viele Familienmitglieder, Untermieter/-innen oder gar Besucher/-innen geschädigt sind. In der Opferbefragung kommt es darauf an, ob eine haushaltsbezogene oder eine personenbezogene Betrachtungsweise zugrunde gelegt wird.[28]

– Besteht zwischen mehreren Handlungen ein enger räumlicher und zeitlicher Zusammenhang i. S. „natürlicher Handlungseinheit“, ist in der PKS ein Fall zu erfassen, und zwar auch dann, wenn mehrere Opfer/Geschädigte betroffen sind. Erfasst wird nur der Straftatbestand mit der schwersten Strafdrohung. Werden z. B. bei einem Gaststätteneinbruch der Vermieter (Sachbeschädigung an der Eingangstür), der Wirt (Diebstahl aus der Kas-

---

27 Auf die Opferzählung der PKS wird hier nicht eingegangen, weil für die Mehrzahl der Delikte, die in Opferbefragungen erfasst werden – Eigentums- und Vermögensdelikte – in der PKS derzeit noch keine Opfererfassung stattfindet.

28 Bei einer haushaltsbezogenen Befragung wird nur ein Fall gezählt, weil nur ein Haushalt betroffen ist. Bei einer personenbezogenen Betrachtung werden alle in der Stichprobe befindlichen Personen, die durch diesen Wohnungseinbruch geschädigt wurden, gezählt, also z. B. ein Familienmitglied und ein Besucher.

se) und der Aufsteller des Spielautomaten (Sachbeschädigung und Diebstahl) geschädigt, dann wird in der PKS nur der Einbruchsdiebstahl, der sich gegen den Wirt richtet, erfasst. In der Opferbefragung wären es dagegen drei Opfer – sofern alle drei Personen per Zufall in der Stichprobe sind.

– Gleichartige Serientaten zum Nachteil desselben Opfers sind in der PKS ebenfalls als ein Fall zu erfassen. Entwendet z. B. über mehrere Monate hinweg der Tatverdächtige immer wieder Weinflaschen aus dem Weinkeller desselben Geschädigten, dann ist nur ein Fall zu erfassen. In der Opferbefragung wären es dagegen mehrere Viktimisierungen, die sich in den Inzidenzen auswirken. Vergleichbares gilt z. B. für innerfamiliäre Gewalttaten.

– Sind von mehreren selbstständigen Handlungen desselben Tatverdächtigen verschiedene Personen geschädigt, zählt in der PKS jede Handlung als ein Fall, der beim schwersten Straftatbestand erfasst wird. Werden z. B. zehn Kraftfahrzeuge aufgebrochen und Gegenstände entwendet, dann werden zehn Fälle (Einbruch in Kfz) erfasst, und zwar auch dann, wenn die Eigentümer der Kfz nicht identisch sind mit den Eigentümern der gestohlenen Gegenstände. In der Opferbefragung würden dagegen so viele Fälle gezählt, wie Opfer von Sachbeschädigung und von Einbruch vorliegen.

3.4 (Teil-)Überschneidung der Referenzzeiträume

In der PKS werden sämtliche im jeweiligen Kalenderjahr abschließend bearbeiteten strafrechtlichen Sachverhalte erfasst, und zwar unabhängig vom Jahr der Tatbegehung. So beruhte z. B. die deutliche Steigerung der registrierten vorsätzlichen Tötungsdelikte Anfang der 1990er Jahre auf den von der Zentralen Ermittlungsgruppe Regierungs- und Vereinigungskriminalität (ZERV) erfassten Fällen von Mord und Totschlag, deren Tatzeiten zwischen 1951 und 1989 lagen.

In Opferbefragungen werden dagegen alle Ereignisse erfasst, die sich nach Erinnerung des Befragten innerhalb des Referenzzeitraums ereigneten. Da die Verfälschung durch Erinnerungsfehler[29] mit der Länge des Referenzzeit-

[29] Die methodische Auswertung der „Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage“ ergab z. B., dass ein Teil der Befragten Schwierigkeiten hatte, sich an Delikte zu erinnern, die ihnen innerhalb der letzten fünf Jahre widerfahren waren (Fuhr/Guzy 2010, 639). Deshalb sollten „einheitliche bzw. möglichst wenig unterschiedliche Referenzzeiträume abgefragt werden“ (Brings u. a. 2010, 740).

raums zunimmt,[30] wird häufig nur nach Viktimisierungsereignissen gefragt, die sich in den letzten zwölf oder gar nur letzten sechs Monaten vor dem Interview ereignet haben. Da sich bei großen Stichproben die Interviews über mehrere Monate erstrecken, sind die Referenzzeiträume nicht für alle Befragten identisch. Ohnedies ist der Zwölfmonatszeitraum nicht identisch mit dem Kalenderjahr der PKS.

Da das Datum der letzten Tat für die PKS erfasst ist, könnte durch eine Sonderauswertung der PKS begrenzt Vergleichbarkeit hergestellt werden, limitiert freilich durch die Fälle mit unbekannter oder geschätzter Tatzeit sowie bei Anzeigen, die erst längere Zeit nach der Tatzeit erfolgen.

3.5 Validität von Daten der Polizeilichen Kriminalstatistik und von Opferbefragungen

3.5.1 Polizeiliche Kriminalstatistik

Dass jeder zu erfassende Fall auch statistisch (und gleichsam den Erfassungsrichtlinien entsprechend) in der PKS erfasst wird, ist nicht gesichert; unterschiedliche „Erfassungstraditionen“ in den Ländern oder auch in örtlichen Dienststellen sind nicht auszuschließen. Eine systematische Fehlerquellenanalyse wurde zwar noch nicht durchgeführt. Einzeluntersuchungen belegen aber sowohl Über- als auch Untererfassungen.[31] Speziell zur Deliktsdefinition ergab die Untersuchung von Gundlach/Menzel in Hamburg eine Fehlerquote von 18 % (Gundlach/Menzel 1993, 122), während Stadler/Walser (2000, 72, Abbildung 2, 81 f.) bei einzelnen Delikten eine Übererfassung zwischen 16 % und 50 % feststellten und bei konstruierten Fällen (Körperverletzung, Ladendiebstahl) eine Fehlerfassung von bis zu 36 % experimentell ermittelten.

Die Erfassung in der PKS gibt die Bewertung zum Zeitpunkt des Abschlusses der polizeilichen Ermittlungen wieder. Aus Sicht der Bewertung nachfolgender Instanzen – Staatsanwaltschaft und/oder Gericht – tendiert die Polizei zur „Überschätzung“ – und zwar sowohl hinsichtlich der Zahl der „Taten“ und der „Tatverdächtigen“ als auch hinsichtlich der Schwere des Sachverhalts,

---

30 Trotz (oder gerade wegen) der Vorgabe eines Referenzzeitraums kann es zu einer Überschätzung der Häufigkeiten relevanter Ereignisse kommen, wenn Ereignisse, die nicht in der Referenzperiode stattgefunden haben, irrigerweise in diese Periode verlegt werden (*Forward Telescoping*). Es kann – wohl seltener – zu einer Unterschätzung kommen, wenn Ereignisse irrigerweise als vor der Referenzperiode stattgefunden berichtet werden (*Backward Telescoping*). Zu den möglichen Methoden, um derartige Telescoping-Effekte zu verringern, siehe BUKS 2002, 24 f., 119 ff.

31 Mit weiteren Nachweisen 1. PSB, 20 ff.; Birkel 2014a, 26.

d. h. im Zweifel wird der als schwerer zu beurteilende Sachverhalt angenommen (Überbewertungstendenz). Diese Überbewertung wird, wenn sie im weiteren Fortgang des Verfahrens geändert wird, im statistischen Ausweis der PKS nicht zurückgenommen. Insbesondere bei schweren Delikten findet häufig eine Umdefinition im weiteren Verfahrensgang statt, und zwar regelmäßig ein „Herunterdefinieren". Deren Ausmaß und Art lassen freilich die gegenwärtigen Kriminalstatistiken nicht erkennen.[32]

3.5.2 Opferbefragungen

Die Grenzen von Dunkelfeldforschungen werden zum einen bestimmt durch die allgemeinen methodischen Probleme von Stichprobenbefragungen, zum anderen durch spezielle Probleme dieses Befragungstyps. Hierzu zählen die beschränkte Erfragbarkeit von Delikten, die Verständlichkeit der Deliktsfragen, die Erinnerungsfähigkeit der Befragten und der „Wahrheitsgehalt" der Aussagen.

– Zu den allgemeinen methodischen Problemen jeder Befragung zählt vor allem, dass bestimmte Personengruppen der Grundgesamtheit typischerweise nicht oder nicht repräsentativ erfasst werden, wie z. B. in bestimmten subkulturellen Milieus lebende Personen (z. B. Rotlichtmilieu, Drogenszene) sowie Angehörige überdurchschnittlich mobiler Personengruppen, die aus Gründen des beruflichen oder des privaten Lebensstils schwieriger an ihrer Wohnanschrift anzutreffen sind als andere, d. h weniger mobile Personengruppen.[33] Ferner werden wie bereits erwähnt aus erhebungstechnischen Gründen bestimmte Gruppen der Wohnbevölkerung mehr oder weniger systematisch ausgeschlossen. Die Art der Datenerhebung bestimmt sowohl die Ausschöpfungsquote als auch das Antwortverhalten.[34]

– Nicht auszuschließen ist, dass Opfer seltener die Teilnahme verweigern als Nichtopfer und deshalb systematisch überrepräsentiert sind.

---

32 Aus definitionstheoretischer Sicht ist dies freilich keine Frage der Validität. Denn jede Definition ist „wahr".

33 Viktimisierung, insbesondere wiederholte Viktimisierung, ist nicht gleichmäßig in der Bevölkerung verteilt. Gerade Gruppen in bestimmten subkulturellen Milieus oder überdurchschnittlich mobile Gruppen dürften ein überdurchschnittlich hohes Viktimisierungsrisiko aufweisen. Ihre Untererfassung in Opferbefragungen dürfte deshalb zu einer systematischen Verzerrung der Ergebnisse führen. Eine Gewichtung der Ergebnisse hilft nicht weiter, weil die Größe der Verzerrung unbekannt ist.

34 Guzy 2014; Prätor 2014, 54 f.

– Bei repräsentativen Opferbefragungen sind, wenn Telescoping-Effekte, also fehlerhafte zeitliche Zuordnung hinsichtlich des Referenzzeitraums, möglichst vermieden und Erinnerungsverluste möglichst beschränkt werden sollen, relativ kurze Referenzzeiträume geboten. Wegen der dann gegebenen kleinen Prävalenzrate muss die Stichprobe relativ groß sein. Ab einer bestimmten Größe ist aber nicht mehr jede methodisch gewünschte Befragungsart, z. B. Face-to-face-Interview, durchführbar, weil nicht genügend geschulte Interviewkapazität verfügbar ist, von den Kosten ganz abgesehen.

– Opferbefragungen setzen die Wahrnehmung des Vorgangs und dessen Bewertung als Straftat voraus. Manche Opfer bemerken den Verlust der gestohlenen Sache nicht, manche Opfer bewerten den scheinbar „günstigen" Kauf nicht als Betrug („absolutes Dunkelfeld").

– Opferbefragungen setzen ferner die Erinnerung der Befragten an zurückliegende Ereignisse voraus. Neben individuellen Eigenschaften und Dispositionen wird die Erinnerungsfähigkeit vor allem beeinflusst von der Länge des Referenzzeitraums sowie durch die subjektive Bedeutsamkeit der fraglichen Ereignisse. Täter- und Opferbefragungen haben ergeben, dass schwerere Delikte eher erinnert werden als leichte und ein Teil der länger zurückliegenden schweren Delikte in den Befragungszeitraum hinein zeitlich verschoben wird (sogenannter Telescoping-Effekt).[35] Bei wiederholten Viktimisierungen im Referenzzeitraum fällt es den Befragten nicht selten schwer, jedes Ereignis in Erinnerung zu rufen und zeitlich korrekt einzuordnen, insbesondere wenn es sich um mehrere gleichartige Ereignisse handelt. Selbst die methodisch beste Dunkelfeldforschung kann nicht das „doppelte" Dunkelfeld der nicht oder der fehlerhaft wahrgenommenen/bewerteten Sachverhalte überwinden. Aus definitionstheoretischer Sicht besteht freilich kein „doppeltes" Dunkelfeld, weil Opfererleben und -bewertung zutreffend wiedergegeben werden.

– Das in regelmäßig wiederholten Opferbefragungen erfassbare Deliktspektrum bilden vor allem Eigentums- und einige Vermögensdelikte, die sich gegen Private richten, sowie Körperverletzungsdelikte. Bei anderen Delikten gegen Private, wie z. B. Raub- oder Sexualdelikte, hängt die Aussagekraft zum einen davon ab, dass die Stichprobe hinreichend groß genug ist, um noch genügend Opfer zu finden, zum anderen davon, dass durch geeignete Befragungstechniken keine Beeinträchtigung der Auskunftsbereitschaft der Befragten erfolgt.

---

35 Zu einem Überblick Lynch/Addington 2010; Prätor 2014, 55 f.

– Ein allgemeines, aber sich bei Täter- und Opferbefragungen in besonderer Schärfe stellendes Problem besteht in der Schwierigkeit, strafrechtliche Tatbestände adäquat in die Umgangssprache umzusetzen.[36] „Unterschiede in Bildungsniveau und sozioökonomischem Status dürften das korrekte Verständnis von Deliktsdefinitionen wie das Vertrauen in die zugesicherte Anonymität der Auswertung beeinflussen“ (Kunz 2011, § 21 Rn. 26).

– Auch wenn die Befragten die Frage richtig verstehen und sich zutreffend erinnern, können ihre Angaben „fehlerhaft“ sein. Zu einer Überschätzung der Viktimisierungsereignisse bzw. Fehlbewertung des Ereignisses führt es, wenn das Opfer die Frage bejaht, der Vorfall aber entweder nicht oder jedenfalls so nicht stattgefunden hat bzw. ein anderes Begriffsverständnis zugrunde liegt.[37] Zu einer Unterschätzung kommt es dann, wenn das Opfer eine individuelle Schädigung oder Beeinträchtigung überhaupt nicht wahrgenommen oder das erkannte Geschehen nicht bzw. fälschlich als strafbar bewertet hat.[38] Die „Lösung“, statt einer Einordnung durch die

---

36 Der „Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage“ zufolge konnte eine Vergleichbarkeit der Deliktsbeschreibungen in der Opferbefragung mit der Straftatendefinition der PKS erreicht werden bei „Diebstahl in und aus Kraftfahrzeugen, Diebstahl an Kfz, von Kraft- oder Fahrrädern und Raub“. „Relativ gut“ ließ sich „Gewaltkriminalität“ abbilden. Nur „sehr eingeschränkt“ vergleichen ließen sich „Wohnungseinbruch, Autodiebstahl, Beschädigung von Autos, Sachbeschädigung, Sonstiger Diebstahl, Warenbetrug, Dienstleistungsbetrug, Bestechung“, weil entweder in der PKS entsprechende Schlüsselzahlen fehlten oder unklar war, ob die Befragten ein den PKS-Definitionen entsprechendes Deliktsverständnis hatten. Nicht vergleichen ließen sich „moderne Delikte wie Identitätsdiebstahl, Phishing, Computervirenverbreitung oder Hacking“ (Fuhr/Guzy 2010, 641), da entsprechende Schlüsselzahlen der PKS fehlten.

37 In der „Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage“ wurde bei den in der zweiten Befragungswelle geführten deliktspezifischen Interviews mit viktimisierten Personen festgestellt, dass mehr als die Hälfte (35 von 60 Personen) nicht – wie in der ersten schriftlichen Befragung angegeben – von einem Raub, sondern „nur“ von einem Diebstahl betroffen war (Fuhr/Guzy 2010, 640). Von ähnlichen Problemen wurde hinsichtlich der Unterscheidung von Wohnungseinbruch und Diebstahl, bei Gewalt und Belästigung sowie Gewalt und Raub berichtet. „Mehr als ein Drittel aller Gewalttaten [war] bereits im Raubteil angegeben worden“ (Fuhr/Guzy 2010, 640).

38 Vereinzelt wurde versucht, diesen Fehler bei den Bewertungen der Befragten zu vermeiden, indem z. B. in einigen Studien die Befragten gebeten wurden, die jeweiligen Viktimisierungen mit ihren eigenen Worten zu berichten. Diese Schilderungen wurden anschließend von geschulten Juristen den einzelnen Straftatbeständen zugeordnet (Schwind u. a. 2001, 22 f.; 110, zum entsprechenden Vorgehen in der Bochumer Dunkelfeldstudie). Eine zweite Methode, die u. a. in der „Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage“ verwendet wurde, besteht darin, in einer Screeningfrage die juristischen Kriterien alltagssprachlich zu benennen und durch gezielte Folgefragen zu ermitteln, ob die Subsumtion zutreffend war. Derartige Methoden in einer bundesweit repräsentativen Opferbefragung zu realisieren, dürfte indes zu aufwendig und zu kostspielig sein (Birkel 2014b, 88, hinsichtlich des „Victimsurvey 2012“). Überdies stellt sich das Problem, dass in der Opferbefragung nur die Sichtweise des Opfers zum Tragen kommt, in der PKS-Bewertung dagegen das Ergebnis der Ermittlungen unter Einschluss auch von Zeugen und Sachbeweisen. Ferner ist nicht auszuschließen, dass die Bewertung der Forscher nicht übereinstimmt mit der Bewertung der po-

Befragten Sachverhaltsschilderungen zu erbitten und diese durch Interviewer/-innen entsprechend einer „durchschnittlichen Subsumtionspraxis der Polizei“ (Birkel 2014b, 78) bewerten zu lassen, ersetzt *eine* Bewertung durch eine *andere*, die ebenfalls fraglich ist.

– Nicht auszuschließen ist, dass eine in bestimmten Deliktsbereichen, z. B. Gewaltkriminalität, erfolgende allgemeine Sensibilisierung nicht nur das Anzeigeverhalten, sondern auch die soziale Wahrnehmung beeinflusst – mit der Folge, dass bei Wiederholungsbefragungen Ereignisse berichtet werden, über die früher nicht berichtet worden ist.[39]

– Kaum zuverlässig erfassbar sind Delikte, bei denen Täter und Opfer einverständlich zusammenwirken, Delikte, an denen das Opfer selbst beteiligt oder interessiert ist, Delikte, bei deren Offenbarung Repressalien zu befürchten sind.[40] Furcht vor einer möglichen Bestrafung, Schamgefühle, übergroßes Geltungsstreben bis hin zur Verfälschung in Richtung auf die vermeintlich erwartete Antwort[41] können Gründe für unbewusst oder bewusst unwahre Angaben sein.[42] Speziell innerfamiliäre Vorfälle, Tätlichkeiten, sexueller Missbrauch, sexuelle Gewalt in der Familie usw. werden aus Gründen der Scham oder weil sie nicht als Straftat, sondern als Privatsache angesehen werden, zu einem erheblichen Anteil nicht mitgeteilt. Teilweise wurde durch spezielle Befragungsmodule oder Befragungsarten versucht, auch hier verlässliche Auskunft zu erhalten.[43]

---

lizeilichen Sachbearbeiter, die aus ermittlungstaktischen oder sonstigen Gründen eher vom schwersten Tatvorwurf ausgehen (Birkel 2014b, 77).

39 Mansel/Hurrelmann 1998, 85 f.

40 Hierzu Ohlemacher 1998.

41 Prätor 2014, 49 f.

42 Durch einen „reverse record check“, also eine Befragung nur solcher Opfer, die Anzeige bei der Polizei erstattet hatten, wurde z. B. festgestellt, dass gerade bei innerfamiliärer Gewalt ein erheblicher Teil der angezeigten Fälle in der Befragung nicht angegeben wurde (Block/Block 1984, 147).

43 Vielfach wird ein Fragebogen zum Selbstausfüllen verwendet („Drop-off-Fragebogen“), häufig kombiniert mit einem vorgeschalteten Interview (z. B. Wetzels u. a. 1995, 179 ff.; Hellmann 2014). Um Verweigerungs- und Rücklaufquoten zu ermitteln, muss freilich durch geeignete Vorgehensweisen sichergestellt werden, dass nur die Zielperson den Fragebogen erhält.

3.6 Unterschiede in der Berechnung von Belastungszahlen in der Polizeilichen Kriminalstatistik und in Opferbefragungen

Um Ergebnisse aus verschiedenen Grundgesamtheiten im Längs- oder im Querschnitt miteinander vergleichen zu können, bedarf es einer standardisierten Größe. In der PKS werden hierfür Belastungszahlen (hier: Häufigkeitszahl bezüglich der Fälle und Opfergefährdungszahl bezüglich der Opfer) verwendet, die pro 100 000 der Wohnbevölkerung berechnet werden.

> "Die Aussagekraft der Häufigkeitszahl wird dadurch beeinträchtigt, [...] dass u. a. Stationierungsstreitkräfte, ausländische Durchreisende, Touristinnen bzw. Touristen, Besucherinnen oder Besucher und grenzüberschreitende Berufspendlerinnen bzw. Berufspendler sowie Nichtdeutsche, die sich illegal im Bundesgebiet aufhalten, in der Einwohnerzahl der Bundesrepublik Deutschland nicht enthalten sind. Straftaten, die von diesem Personenkreis begangen wurden, werden aber in der Polizeilichen Kriminalstatistik gezählt." (PKS 2012, 353)

In einer Opferbefragung werden Prävalenz- bzw. Inzidenzraten berechnet und auf die Grundgesamtheit der Stichprobe, also z. B. der 14- bis unter 80-jährigen Deutschen, bezogen. Eine analoge Berechnung der PKS-Daten ist nicht möglich, weil für die Mehrzahl der in Opferbefragungen erfassten Delikte in der PKS keine Opfererfassung erfolgt. Ein weiteres Hindernis stellt die in neueren Studien übliche Differenzierung in haushalts- und personenbezogene Delikte dar. Die entsprechenden Raten werden teils auf Haushalte, teils auf Personen bezogen. Eine entsprechende Berechnung ist mit den PKS-Daten nicht möglich.

3.7 Ermittlung von Umfang und Veränderung der Relationen zwischen Ergebnissen von Opferbefragungen und der PKS

3.7.1 Methoden zur Bestimmung der Relationen

Durch Kontrastierung der Ergebnisse von Opferbefragungen mit jenen der PKS soll deren Relation abgeschätzt, also eine sogenannte Dunkelziffer ermittelt werden. Dies kann mittels Hochrechnung der angegebenen Vorfälle auf die Grundgesamtheit und Vergleichs dieses Ergebnisses mit den entsprechenden Daten der PKS realisiert werden.[44] Die Vergleichbarkeit dieser beiden Werte ist freilich, von allem anderen abgesehen, beeinträchtigt durch die Unterschiede in den Grundgesamtheiten, den Referenzzeiträumen, den Erfas-

[44] So z. B. die „vereinfachte Hochrechnung" von Liebl 2014, 82, Tabelle 25, 190 f., Tabellen 67 und 68. Der weitere Ansatz von Liebl, Anteile der Delikte im *Victim Survey* zu vergleichen mit Anteilen in der PKS (a. a. O., 76 ff.), ist verfehlt, weil, wie der Autor selbst einräumt, völlig unterschiedliche Grundgesamtheiten der Berechnung von Anteilen zugrunde liegen.

sungsregeln sowie in der Berechnung der Belastungszahlen. „Einfache“ Hochrechnungen, ohne diese Unterschiede auch nur ansatzweise zu berücksichtigen, sind für eine Kontrastierung der Ergebnisse ungeeignet. Ziel sollte sein, eine „höchstmögliche Vergleichbarkeit“ zu erreichen, „um Nutzern, die keine Expertinnen oder Experten sind, einen methodisch adäquaten Umgang mit diesen Daten zu erleichtern“ (Birkel 2014b, 69).

Eine andere Methode zur Ermittlung einer Dunkelzifferrelation ist die Berechnung des Verhältnisses angezeigter zu nicht angezeigten Straftaten in der Annahme, die (hochgerechneten) Angaben der Befragten zu den angezeigten Straftaten entsprechen dem sogenannten Hellfeld. In der Regel werden in Opferuntersuchungen diejenigen Befragten, die angeben, Opfer geworden zu sein, um die Mitteilung gebeten, ob Anzeige erfolgt sei. Voraussetzung hierfür ist freilich zunächst, dass Inzidenzen erfasst werden und das Anzeigeverhalten für jedes Viktimisierungsereignis erhoben wird. Verzerrte Ergebnisse sind zu erwarten, wenn das Anzeigeverhalten nur für das letzte oder nur für das schwerste Delikt ermittelt wird.[45]

Bei validen Angaben der Befragten sowohl zum Delikt als auch zur Anzeige sollte, so die Annahme, die hochgerechnete Zahl der angezeigten Delikte in etwa der Größenordnung der für Individualopfer im Referenzzeitraum registrierten Fälle entsprechen. In einigen Untersuchungen ergaben die Schätzungen der Opferbefragung aber das Mehrfache der PKS-Werte. Dies kann darauf beruhen, dass die Frage nach der Anzeige eine sozial erwünschte Antwort provoziert,[46] also eine Anzeige bejaht wird, die tatsächlich nicht stattgefunden hat, ferner auch darauf, dass die Polizei die Anzeige nicht aufgenommen hat oder die Opfer irrtümlich meinen, eine Anzeige aufgegeben zu haben, dass z. B. eine telefonische Mitteilung bei einem Strafantragsdelikt eine Anzeige sei.[47] Um die Validität der Antworten zu erhöhen, wird deshalb in neueren Opferbefragungen auch danach gefragt, ob bei der Polizei ein Protokoll unterschrieben worden sei.[48] Freilich ergeben sich auch dann noch teilweise erhebliche Differenzen zwischen den auf Basis der Angaben über angezeigte Delikte hochgerechneten und den polizeilich registrierten Fallzahlen.[49]

---

45 Eingehend BUKS 2002, 37 ff.

46 In einer vom BKA 1984 zu Testzwecken durchgeführte Befragung zur Opfereigenschaft bei Diebstahl und Körperverletzung haben „ca. zwei Drittel der Befragten – und damit unglaubhaft viele – behauptet, die Tat bei der Polizei mit Unterschrift angezeigt zu haben“ (Dörmann 1988, 404). Siehe ferner m. w. N. Schwind u. a. 2001, 113 f., 134 ff.

47 Zu diesen Möglichkeiten Schwind u. a. 2001, 114 f.

48 Schwind u. a. 2001, 115. Im „Viktimisierungssurvey 2012“ wurde danach gefragt, wer die Polizei informiert hat, wie (mit einer Reihe von Antwortvorgaben) die Polizei informiert wurde, ob Anzeige erstattet wurde, wer die Anzeige erstattet hat und ob bei der Polizei ein Schriftstück unterzeichnet wurde.

49 Schwind u. a. 2001, 134 ff.

Schwind u. a. haben deshalb vorgeschlagen, in künftigen Untersuchungen die Probandinnen und Probanden der Opferbefragung um ihr „Einverständnis zu bitten, ihre Angaben zum Anzeigeverhalten mit den polizeilichen Daten abzugleichen“,[50] was freilich nur bei einer Befragung auf lokaler Ebene realisierbar sein wird.[51]

3.7.2 Messung der Veränderung der Relationen

Kriminalität ist ein „seltenes“ Ereignis. Um statistisch signifikant nicht nur die Häufigkeit, sondern vor allem die Veränderung von Prävalenz- und Inzidenzraten (d. h. den Opferanteil in der Bevölkerung sowie die Häufigkeiten von Viktimisierungsereignissen) messen zu können, sind deshalb relativ große Stichproben erforderlich, jedenfalls wenn sich die Befragung nicht nur auf „Massendelikte“, wie Sachbeschädigung oder Diebstahl, beschränken soll.[52]

4 Zusammenfassung

– Dunkelfeldforschung ist nicht, wie einst angenommen, der „Königsweg“ zur Messung der „Kriminalitätswirklichkeit“. Dunkelfeldforschung misst nicht die „Kriminalitätswirklichkeit“, sondern immer nur die Selbstbeurteilung und Selbstauskunft der Befragten (oft in einer zumeist vorstrukturierten Befragungssituation), d. h., es wird erfasst, wie Befragte bestimmte Handlungen definieren, bewerten, kategorisieren, sich daran erinnern und bereit sind, darüber Auskunft zu geben. Opferbefragungen wie PKS reproduzieren die ihren unterschiedlichen Forschungsdesigns jeweils zugängliche Kriminalitätswahrnehmung von Bevölkerung und Instanzen.

> “Opferbefragungen erheben verbalisierte Erinnerungen an Handlungen, die entweder nach Einstufung der viktimisierten Personen nach groben, im Erhebungsinstrument implementierten Kriterien einen bestimmten Straftatbestand erfüllen könnten oder dies nach Anwendung strafrechtlicher Kriterien durch geschulte Kodierer tun sollten. Polizeiliche Kriminalstatistiken messen auf Grundlage verschiedener Informationsquellen rekonstruierte Handlungsabläufe, welche nach der Beurteilung eines Ermittlers oder polizeilichen Statistiksachbearbeiters jeweils einen bestimmten Straftatbestand erfüllen und daher einer bestimmten Kategorie der Kriminalstatistik zuzuordnen sind.“ (Birkel 2014b, 78)

---

50 Schwind u. a. 2001, 140.

51 Dörmann 1988, 404: „[. . .] auf Bundesebene (fehlt) die Möglichkeit, die oft falschen Angaben z. B. zum Anzeigeverhalten anhand von Polizeiunterlagen zu überprüfen“.

52 Zu Einzelheiten siehe den Beitrag von Schnell/Noack in diesem Band.

– Opferbefragungen und PKS sind komplementäre Datenquellen.[53] Es gibt nicht *das* Messinstrument, mit dem *die* Kriminalität gemessen werden könnte, sondern (durchaus unterschiedliche) Wahrnehmungen und (durchaus unterschiedliche) Bewertungen auf jeder Tätigkeitsstufe.

– Die Daten aus Opferbefragungen und PKS überschneiden sich in allen relevanten Punkten nur partiell. Der Deliktsbereich von Opferbefragungen erfasst nicht alle in Hellfelddaten detektierten Ereignisse; im Überschneidungsbereich sind die PKS-Daten keine Teilmenge der Opferbefragungsdaten. Unterschiede bestehen ferner hinsichtlich der Grundgesamtheiten, der Referenzzeiträume, der Erfassungsregeln sowie der Berechnung von Belastungszahlen. Beide Datengruppen weisen schließlich unterschiedliche Validitätsprobleme auf. Die Ergebnisse können deshalb auch nicht exakt zueinander in Beziehung gesetzt werden. Vergleichbarkeit lässt sich nur annäherungsweise und nur für bestimmte Wirklichkeitsausschnitte erreichen.

– Opferbefragungen sind kein Ersatz für die PKS, sie sind aber eine notwendige Ergänzung und Erweiterung, denn

  – sie liefern *erstens* opferbezogene Erkenntnisse auf mehreren kriminalpolitisch wichtigen Feldern,[54] die für die PKS nicht erhoben werden,

  – sie informieren *zweitens* über Viktimisierungen, die im Dunkelfeld geblieben sind, über deren Folgen und deren Verarbeitung,

  – sie beugen *drittens* einer „naiven Gleichsetzung von Hellfelddaten und Kriminalitätswirklichkeit“[55] vor,

  – sie erlauben *viertens* – und zwar weitaus besser als die eine nationale Strafrechtsordnung widerspiegelnde PKS – internationale Vergleiche,

  – sie ermöglichen schließlich *fünftens* – jedenfalls für Teilbereiche –, die hinsichtlich der PKS stattfindenden Selektionsprozesse, insbesondere die Anzeige betreffend, abschätzen, quantifizieren und in ihrer Bedeutung für das kriminalstatistische Bild bewerten zu können. Die in Op-

---

53 Mit weiteren Nachweisen zur Einschätzung des Verhältnisses beider Datenquellen Birkel 2014a, 30 f.; Birkel 2014b.

54 Opfererleben und -verarbeitung, Kriminalitätsfurcht, Hilfe- und Beratungsbedarf, Akzeptanz von Polizei und Justiz, Bewertung von Straftaten, Gestaltungsspielräume von Kriminalpolitik

55 Birkel 2014a, 33.

ferbefragungen gewonnenen Erkenntnisse zum Anzeigeverhalten sowie der Deliktsschwereeinschätzung bieten Anhaltspunkte für die Erklärung etwaiger Divergenzen.

– Die Zusammenschau der Ergebnisse aus beiden Datenquellen verbessert die Erkenntnisbasis, weil empirische Befunde zur Frage vorliegen, ob Veränderungen bei den der Polizei bekannt gewordenen Fällen eher auf Veränderungen von Ereignissen beruhen, die wahrgenommen und bewertet werden, oder eher auf Veränderungen des Anzeigeverhaltens.[56] Allerdings ist dieser Erkenntnisgewinn nur möglich, wenn regelmäßige, gleichartige und repräsentative Opferuntersuchungen durchgeführt und geeignete Maßnahmen zur Optimierung der Vergleichbarkeit getroffen werden.[57]

---

56 Damit ist freilich nur ein Ausschnitt der in Betracht kommenden Faktoren benannt, denkbar ist z. B. auch eine Änderung der Bewertung von Sachverhalten durch die Polizei. Ostendorf hat aus seiner praktischen Erfahrung als Generalstaatsanwalt den Eindruck gewonnen, dass auch „Anforderungen aus der Öffentlichkeit und Politik“ nicht ohne Einfluss sind. „Das Zündeln im Keller eines Mietshauses, in dem auch Ausländer wohnen, ist z. T. ohne weiteres als Mordversuch eingestuft worden, um ja nicht den Eindruck einer ausländerfeindlichen Einstellung aufkommen zu lassen. Ich kenne einen Fall, wo es anschließend eine Verfahrenseinstellung wegen Geringfügigkeit gegeben hat“ (Ostendorf 1998, 182 f.).

57 Siehe die von Birkel 2014a, 35 ff. genannten Maßnahmen im Rahmen des Projekts „Barometer Sicherheit in Deutschland“.

5 Literaturverzeichnis

Antholz, Birger (2010): Dämmerfeld. Anteil der polizeigemeldeten, aber nicht förmlich in der Polizeilichen Kriminalstatistik registrierten Kriminalität. In: MSchrKrim 93, S. 409–423.

Baier, Dirk; Pfeiffer, Christian; Simonson, Julia und Rabold, Susann (2009): Jugendliche in Deutschland als Opfer und Täter von Gewalt. Erster Forschungsbericht zum gemeinsamen Forschungsprojekt des Bundesministeriums des Innern und des KFN. KFN-Forschungsbericht Nr. 107, Hannover: KFN.

Birkel, Christoph (2014a): Gefährdungen durch Kriminalität in „offiziellen" Zahlen und subjektivem Erleben der Menschen: Polizeiliche Kriminalstatistik und Dunkelfeldbefragungen. In: Röllgen, Jasmin (Hg.): 5. SIRA Conference Series, S. 23–43.

Birkel, Christoph (2014b): Hellfeld vs. Dunkelfeld: Probleme statistikbegleitender Dunkelfeldforschung am Beispiel der bundesweiten Opferbefragung im Rahmen des Verbundprojektes „Barometer Sicherheit in Deutschland" (BaSiD). In: Eifler, Stefanie; Pollich, Daniela (Hg.): Empirische Forschung über Kriminalität. Wiesbaden: Springer VS, S. 67–94.

Birkel, Christoph; Guzy, Nathalie; Hummelsheim, Dina; Oberwittler, Dietrich und Pritsch, Julian (2014): Der Deutsche Viktimisierungssurvey 2012. Erste Ergebnisse zu Opfererfahrungen, Einstellungen gegenüber der Polizei und Kriminalitätsfurcht, Freiburg i. Br. URL: http://www.bka.de/nn_233148/SharedDocs/Downloads/DE/Publikationen/Publikationsreihen/SonstigeVeroeffentlichungen/2014DeutscherViktimisierungssurvey2012.html – Download vom 07. 01. 2015.

Block, Carolyn Rebecca; Block, Richard L. (1984): Crime Definition, Crime Measurement, and Victim Surveys. In: Journal of Social Issues 40, S. 137–160.

Brings, Stefan; Fuhr, Gabriele; Guzy, Nathalie; Hanefeld, Ute und Mischkowitz, Robert: Kriminalität und Sicherheitsempfinden. Testerhebung zur Vorbereitung einer europaweiten Bevölkerungsumfrage (Viktimisierungsbefragung), Wirtschaft und Statistik 2010, S. 735–744.

Bundesministerium des Innern; Bundesministerium der Justiz (Hg.) (2001): Erster Periodischer Sicherheitsbericht, Berlin: Bundesministerium des Innern; Bundesministerium der Justiz (zitiert: 1. PSB).

Coleman, Clive; Moynihan, Jenny (1996): Understanding Crime Data. Haunted by the Dark Figure, Buckingham u. a.: Open University Press.

Dellwing, Michael (2010): Dunkelfeldforschung als Definitionsaktivität. In: MSchrKrim 93, S. 180–197.

Ditton, Jason (1979): Controlology. Beyond the New Criminology, London u. a.: Macmillan.

Dörmann, Uwe (1988): Dunkelfeldforschung im Dunkeln. Zum Problem der statistikbegleitenden Dunkelfeldforschung: Eine vergleichende Betrachtung. In: Kriminalistik, 42, S. 403–405.

Fuhr, Gabriela; Guzy, Nathalie (2010): Europäische Dunkelfeldforschung in Deutschland. Ergebnisse der EU-Testerhebung „Translating and Testing a Victimisation Survey Module“. In: Kriminalistik, 11, 636–643.

Görgen, Thomas (Hg.) (2010): Sicherer Hafen oder gefahrvolle Zone? Kriminalitäts- und Gewalterfahrungen im Leben alter Menschen. Frankfurt/M.: Verlag für Polizeiwissenschaften.

Gundlach, Thomas; Menzel, Thomas (1993): Polizeiliche Kriminalistik: Fehlerquellen der PKS und ihre Auswirkungen am Beispiel Hamburgs. In: Kriminalistik, 47, S. 121–125.

Guzy, Nathalie (2014): International vergleichende Viktimisierungssurveys. Aktuelle Herausforderungen und Ergebnisse des Methodentests „ICVS-2“. In: Eifler, Stefanie; Pollich, Daniela (Hg.): Empirische Forschung über Kriminalität. Wiesbaden: Springer VS, S. 149–182.

Heinz, Wolfgang (2002): Abschlussbericht der Arbeitsgruppe „Regelmäßige Durchführung von Opferbefragungen“ für das BMJ und BMI (unveröffentlicht) (zitiert: BUKS 2002).

Heinz, Wolfgang (2006): Zum Stand der Dunkelfeldforschung in Deutschland. In: Festschrift für Helmut Kury. Frankfurt/M.: Verlag für Polizeiwissenschaften, S. 241–263.

Hellmann, Deborah F. (2014): Repräsentativbefragung zu Viktimisierungserfahrungen in Deutschland, KFN-Forschungsbericht, Nr. 122, Hannover: KFN.

Kreuzer, Arthur; Görgen, Thomas; Krüger, Ralf; Münch, Volker und Schneider, Hans (1993): Jugenddelinquenz in Ost und West, Vergleichende Untersuchungen bei ost- und westdeutschen Studienanfängern in der Tradition Gießener Delinquenzbefragungen. Bonn: Forum Verlag.

Kunz, Karl-Ludwig (2008): Die wissenschaftliche Zugänglichkeit von Kriminalität. Ein Beitrag zur Erkenntnistheorie der Sozialwissenschaften. Wiesbaden: Dt. Universitäts-Verlag.

Kunz, Karl-Ludwig (2011): Kriminologie, 6. Aufl., Bern u. a.: UTB.

Kürzinger, Josef (1978): Private Strafanzeige und polizeiliche Reaktion. Berlin: Duncker & Humblot.

Liebl, Karlhans (2014): Viktimisierung, Kriminalitätsfurcht und Anzeigeverhalten im Freistaat Sachsen. Frankfurt/M.: Verlag für Polizeiwissenschaften.

Lynch, James P.; Addington, Lynn A. (2010): Identifying and Addressing Response Errors in Self-Report Surveys. In: Piquero, Alex R.; Weisburd, David (Hg.): Handbook of Quantitative Criminology. New York, NY: Springer Science+Business Media, LLC, 2010, S. 251–272.

Mansel, Jürgen; Hurrelmann, Klaus (1998): Aggressives und delinquentes Verhalten Jugendlicher im Zeitvergleich. In: KZfSS 50, S. 78–109

Neubacher, Frank (2014): Aktuelle empirische Befunde der deutschen Kriminologie zur Gewalt unter Gefangenen. In: Festschrift für Ch. Pfeiffer. Baden-Baden: Nomos-Verlagsgesellschaft, S. 485–501.

Ohlemacher, Thomas (1998): Verunsichertes Vertrauen? Gastronomen in Konfrontation mit Schutzgelderpressung und Korruption. Baden-Baden: Nomos-Verlagsgesellschaft.

Ostendorf, Heribert (1998): Wachsende Kriminalität – Verschärfung des Strafrechts? In: Zentralblatt für Jugendrecht und Jugendwohlfahrt, S. 180–186.

Prätor, Susann (2014): Ziele und Methoden der Dunkelfeldforschung. Ein Überblick mit Schwerpunkt auf Dunkelfeldbefragungen im Bereich der Jugenddelinquenz. In: Eifler, Stefanie; Pollich, Daniela (Hg.): Empirische Forschung über Kriminalität. Wiesbaden: Springer VS, S. 31–66.

Pudel, Volker (1978): Motivanalyse des Anzeigeverhaltens. In: Schwind, Hans-Dieter; Ahlborn, Wilfried und Weiß, Rüdiger: Empirische Kriminalgeographie. Bestandsaufnahme und Weiterführung am Beispiel Bochum („Kriminalitätsatlas Bochum“). Wiesbaden: Bundeskriminalamt, S. 205–210.

Quetelet, Adolphe (1835): Sur l’homme et le développement de ses facultés ou essai de physique sociale, Paris: Bachelier, Bd. 2 (zitiert nach: Quetelet, A.: Soziale Physik oder Abhandlung über die Entwicklung der Fähigkeiten des Menschen, 2. Band, Jena: Fischer 1921).

Schwind, Hans-Dieter; Fetchenhauer, Detlef; Ahlborn, Wilfried und Weiß, Rüdiger (2001): Kriminalitätsphänomene im Langzeitvergleich am Beispiel einer deutschen Großstadt, Neuwied/Kriftel: Luchterhand.

Stadler, Willi; Walser, Werner (2000): Fehlerquellen der Polizeilichen Kriminalstatistik. In: Liebl, Karlhans; Ohlemacher, Thomas (Hg.): Empirische Polizeiforschung. Interdisziplinäre Perspektiven in einem sich entwickelnden Forschungsfeld, Herbolzheim: Centaurus-Verlags-GmbH, S. 68–89.

Wetzels, Peter; Greve, Werner; Mecklenburg, Eberhard; Bilsky, Wolfgang und Pfeiffer, Christian (1995): Kriminalität im Leben alter Menschen. Eine altersvergleichende Untersuchung von Opfererfahrungen, persönlichem Sicherheitsgefühl und Kriminalitätsfurcht. Ergebnisse der KFN-Opferbefragung 1992. Schriftenreihe des Bundesministeriums für Familie, Senioren, Frauen und Jugend, Bd. 105, Stuttgart u. a.

3 Analyse der Ergebnisse von Viktimisierungsbefragungen

Statistische Analyseverfahren

Michael Hanslmaier und Dirk Baier

1 Einleitung

Bei der Analyse von Daten aus Opferbefragungen ist man oft mit der Tatsache konfrontiert, dass die lineare Regression mittels der Methode der kleinsten Quadrate (*Ordinary Least Squares* – OLS) aufgrund der Beschaffenheit der Daten nicht angewendet werden kann. Dies ist der Fall, wenn die abhängige Variable nicht metrisch, sondern nur nominalskaliert ist, wie etwa Prävalenzen, also die Angabe, ob eine Person in einem bestimmten Zeitraum Opfer geworden ist oder nicht. Auch kann die abhängige Variable sehr asymmetrisch, d. h. schief verteilt sein. Dies trifft beispielweise auf Inzidenzen zu, also die Häufigkeit der Opferwerdung. Darüber hinaus weisen Daten aus Opferbefragungen häufig eine hierarchische Struktur auf, d. h., die Befragten sind Teil übergeordneter Kontexte, etwa Schülerinnen und Schüler in Schulklassen oder Bewohnerinnen und Bewohner in Stadtvierteln. Diese auch als Cluster bezeichnete Datenstruktur kann die Folge eines mehrstufigen Stichprobenverfahrens sein, bei dem z. B. zunächst Stadtviertel ausgewählt werden und innerhalb dieser Stadtviertel dann Personen. Darüber hinaus kann es auch bei einfachen Zufallsstichproben von Interesse sein, Kontexte zu berücksichtigen, wenn diese die abhängige Variable beeinflussen (z. B. Eigenschaften von Stadtvierteln auf das Opferrisiko). Die Elemente eines Kontexts sind sich in der Regel ähnlicher als Elemente aus verschiedenen Kontexten, u. a. da diese den gleichen Umwelteinflüssen unterliegen. Infolgedessen sind die Beobachtungen nicht mehr statistisch unabhängig voneinander – eine Basisannahme der meisten statistischen Verfahren (Rabe-Hesketh/Skrondal 2012, 1–2; Windzio 2008, 113–114; ausführlicher zur Thematik komplexer Stichproben Lee/Forthofer 2006).

Der vorliegende Beitrag verfolgt zwei Ziele. Zum einen sollen statistische Verfahren für die multivariate Analyse von binären Daten (Prävalenzen), Zähldaten (Inzidenzen) und Mehrebenendaten (Personen in Kontexten) vorgestellt werden. Der Vorteil multivariater Analysen liegt in der Möglichkeit, Drittvariablen kontrollieren und den jeweiligen Effekt verschiedener unabhängiger Variablen simultan schätzen zu können. Die Darstellung soll den Leserinnen und Lesern ein grundlegendes Verständnis der Verfahren liefern. Zum anderen sollen die Verfahren anhand von Beispielanalysen anschaulich

gemacht werden. Zu diesem Zweck werden auf der Basis einer Dunkelfeldbefragung von Schülern aus Niedersachsen aus dem Jahr 2013 Prädiktoren von Viktimisierungserfahrungen untersucht.

2 Einflussfaktoren von Viktimisierungserfahrungen Jugendlicher

Im Gegensatz zur Erklärung abweichenden Verhaltens existieren keine eigenständigen Theorien, die sich mit der Erklärung von Viktimisierung im Allgemeinen oder auch speziell von Jugendlichen beschäftigen. Jedoch lässt sich aus verschiedenen kriminologischen Theorien, wie etwa dem *Routine-Activity*-Ansatz (Cohen/Felson 1979), der *General Theory of Crime* (Gottfredson/Hirschi 1990) und dem Ansatz der sozialen Desorganisation (Shaw/McKay 1969; Sampson u. a. 1997) sowie aus der bisherigen Forschung eine Reihe von Prädiktoren von Viktimisierung ableiten.

So zeigen Studien (u. a. Baier/Prätor im Druck; Gruszczynska u. a. 2012), dass Jungen ein höheres Risiko für Viktimisierung durch Gewalt- und Eigentumsdelikte, Mädchen dagegen ein höheres Viktimisierungsrisiko für sexuelle Gewalt aufweisen. Der *Migrationshintergrund* beeinflusst ebenfalls das Risiko der Opfererfahrung. Neben rassistischen Delikten und Diskriminierung, die nur von Personen mit Migrationshintergrund erlebt werden können, gilt dies auch für andere Delikte. So finden etwa Baier und Prätor (im Druck) etwas höhere Prävalenzraten für Viktimisierung durch Gewaltdelikte von Jugendlichen mit Migrationshintergrund verglichen mit deutschen Jugendlichen. Allerdings variiert die Prävalenz stark über verschiedene Migrantengruppen hinweg.

In der empirischen Forschung hat sich zudem gezeigt, dass *intensiveres elterliches Monitoring* das Risiko für Diebstahl reduziert (Gruszczynska u. a. 2012). Als wichtiger Prädiktor hat sich auch *elterliche Gewalt* erwiesen: Jugendliche, die vor ihrem zwölften Lebensjahr schwere elterliche Gewalt erlebt haben, werden häufiger Opfer von Gewalt (Baier/Prätor im Druck).

Das *Freizeitverhalten* Jugendlicher spielt ebenfalls eine Rolle. Baier und Prätor (im Druck) berichten, dass Schülerinnen und Schüler, die Zeit in Bars, Diskotheken etc. verbringen, ein höheres Risiko aufweisen, Opfer von Gewaltdelikten zu werden. Auch der Kontakt mit *delinquenten Freundinnen und Freunden* erhöht das Viktimisierungsrisiko. Jugendliche mit delinquenten Freundinnen und Freunde werden durch diese in risikoreiche Aktivitäten verwickelt, die etwa mit gewalttätigen Auseinandersetzungen einhergehen und auch das Viktimisierungsrisiko für andere Delikte erhöhen können. Empirische Studien konnten signifikant positive Effekte sowohl für Diebstahl, Körperverletzung und Raub (Gruszczynska u. a. 2012) als auch für Gewaltdelikte (Baier/Prätor im Druck) insgesamt zeigen.

Jugendliche verbringen große Teile ihre Zeit in der *Schule*. Aus dem Desorganisationsansatz, der sich auch auf die Schule übertragen lässt (Hanslmaier 2014), kann abgeleitet werden, dass das Viktimisierungsrisiko von der Fähigkeit des Kontexts Schule abhängt, informelle soziale Kontrolle auszuüben. So zeigt etwa die Arbeit von Sapouna (2010), dass Viktimisierung durch Bullying in Klassen mit höherer kollektiver Wirksamkeit schwächer ist. Zudem berichten Gruszczynska u. a. (2012), dass Schülerinnen und Schüler in desorganisierten Schulen ein höheres Risiko aufweisen, Opfer von Diebstahl, Körperverletzung oder Raub zu werden.

3 Datensatz und Operationalisierung

Die empirischen Analysen im vorliegenden Beitrag basieren auf dem Niedersachsensurvey 2013. Dieser Datensatz erlaubt die Analyse einer großen Bandbreite verschiedener Prädiktoren der Inzidenz und Prävalenz von Viktimisierung und weist zudem eine hierarchische Datenstruktur auf. Der Niedersachsensurvey 2013[1] wurde im Jahr 2013 vom Kriminologischen Forschungsinstitut Niedersachsen durchgeführt und ist repräsentativ für die 9. Jahrgangsstufe in Niedersachsen. Für die Stichprobe wurden zunächst aus einer Liste aller Schulklassen in Niedersachen (geschichtet nach sieben Schultypen) die zu befragenden Klassen gezogen. Alle am Befragungstag anwesenden Schülerinnen und Schüler der ausgewählten Klassen wurden schriftlich und anonym im Klassenverband im Rahmen des Schulunterrichts im Beisein einer/eines geschulten Testleiterin bzw. Testleiters und einer Lehrperson befragt. Dementsprechend stellt die Stichprobe eine Klumpenstichprobe dar und die Daten weisen eine hierarchische Struktur auf. Die Rücklaufquote betrug 64,4 %, sodass die Stichprobe 9.512 Schülerinnen und Schüler aus 485 Klassen umfasst. Die Auswertungen in diesem Beitrag beschränken sich auf die 8.411 Schüler, die für alle im Folgenden betrachteten Variablen gültige Werte aufweisen.[2]

Für die Auswertungen werden drei Indikatoren der Viktimisierung herangezogen: Die *Prävalenz von Gewaltdelikten* gibt an, ob die Jugendlichen in den letzten zwölf Monaten Opfer mindestens eines von sechs Delikten (Raub, Erpressung, sexuelle Gewalt, Körperverletzung mit Waffen, Körperverletzung durch einzelne Personen und Körperverletzung durch mehrere Personen) geworden sind (siehe *Tabelle 1* für deskriptive Statistiken aller Variablen). Die

---

1 Für eine genauere Beschreibung der Studie siehe Baier 2015, 28 ff.

2 Insbesondere Befragte aus Förderschulen fallen dadurch aus der Stichprobe, da diese eine kürzere Version des Fragebogens beantwortet haben, die einige der betrachteten Variablen nicht enthält.

Inzidenz für Eigentumsdelikte (zwölf Monate) ergab sich aus der Summe der Opfererfahrungen für vier Delikte (Fahrraddiebstahl, anderer Fahrzeugdiebstahl, Diebstahl und Sachbeschädigung). *Opfer von Schulgewalt* sind Befragte, die mindestens einmal im letzten Schulhalbjahr mindestens eine von zwei Formen von Gewalt erlebt haben („Ich wurde von anderen Schülern absichtlich geschlagen oder getreten“ bzw. „Andere Schüler haben mich erpresst und gezwungen, Geld oder Sachen herzugeben“).

Ein *Migrationshintergrund* lag vor, wenn die/der Befragte oder deren/dessen leibliche Mutter oder leiblicher Vater eine andere als die deutsche Staatsangehörigkeit besaß oder nicht in Deutschland geboren war. Die sieben Schultypen wurden zu drei *Schulformen* zusammengefasst: Hauptschule, Realschule (inkl. integrierte Haupt- und Realschule) und Gymnasium (inkl. Gesamtschule).

Selbstkontrolle wurde über die Subdimension *Risikosuche* mit vier Items (Beispielitem „Ich teste gerne meine Grenzen, indem ich etwas Gefährliches mache“) auf einer vierfach gestuften Skala erhoben (Cronbachs $\alpha = 0{,}859$).

Das *elterliche Monitoring* wurde über drei Items zu elterlichem Erziehungsverhalten vor dem zwölften Lebensjahr erfasst (Beispielitem Mutter/Vater hat „genau gewusst, wo ich in meiner Freizeit bin“). Die Befragten machten jeweils getrennte Angaben für Mutter und Vater, die zunächst gemittelt wurden bevor der Mittelwert aus den drei Items gebildet wurde. Hohe Werte geben ein intensives elterliches Monitoring an (Cronbachs $\alpha = 0{,}693$).

Überdies wurde die erlebte *elterliche Gewalt* vor dem zwölften Lebensjahr erfasst. Die Schülerinnen und Schüler beantworteten sechs Items zur Häufigkeit erlebter Gewalt getrennt für Mutter und Vater auf einer sechsstufigen Skala. Zunächst wurde für jedes Item der Maximalwert aus den Angaben für Mutter und Vater und daraus anschließend drei Kategorien gebildet. Jugendliche, die keine der sechs Gewaltformen erlebt hatten, wurden als „keine Gewalt“ klassifiziert, wer mindestens eine der leichteren Gewaltformen („mir eine runtergehauen“/„mich hart angepackt oder gestoßen“/„mit einem Gegenstand nach mir geworfen“), aber keine der drei schwereren Gewaltformen („mich mit einem Gegenstand geschlagen“/„mich mit der Faust geschlagen oder mich getreten“/„mich geprügelt, zusammengeschlagen“) erlebt hatte, wurde mit „leichte Gewalt“ bezeichnet. In der Gruppe „schwere Gewalt“ sind Jugendliche, die mindestens eine der drei schwereren Gewaltformen erlebt haben.

Für sechs abweichende Verhaltensweisen (Ladendiebstahl, Schwänzen, Raub, Körperverletzung, Sachbeschädigung und Verkauf von Drogen) wurde getrennt erhoben, wie viele Freundinnen bzw. Freunde die Jugendlichen haben, die diese Verhaltensweisen in den letzten zwölf Monaten ausgeführt haben.

Die *Anzahl der delinquenten* Freundinnen und Freunde wurde aus diesen Angaben mittels Maximalwertbefehl (höchster Wert aus allen sechs Angaben) gebildet und in drei Kategorien eingeteilt.

Die *Routineaktivitäten der Jugendlichen* wurden über die Zeit erfasst, die diese pro Woche damit verbringen, „in Kneipe, Disco, Kino, zu Veranstaltungen“ zu gehen. Die Jugendlichen wurden dann in vier Gruppen anhand der empirischen Verteilung der Variable eingeteilt. Die erste Gruppe umfasst die Jugendlichen, die keine Zeit in Kneipen etc. verbringen, die weiteren Grenzen waren der Median und das dritte Quartil.

Darüber hinaus wurden Eigenschaften der Schule erhoben. Diese stellen Kollektivmerkmale dar, da es sich um Eigenschaften des Kontexts Schule handelt. Für diese Merkmale werden die Angaben der einzelnen Schülerinnen und Schüler aus derselben Klasse aggregiert, d. h. der Mittelwert über alle Schüler einer Klasse gebildet.[3] Einschränkend ist an dieser Stelle anzumerken, dass es sich bei den Eigenschaften um Eigenschaften der Klasse und nicht der Schule handelt. Dies ist auch der Datenstruktur geschuldet, da 485 Klassen aus 379 Schulen befragt wurden, d. h. zumeist nur eine Klasse pro Schule. *Schulische Desorganisation* wurde mit zwei Items erhoben (Beispielitem „An meiner Schule gibt es viel Gewalt“; $r = 0{,}551$), die die Schüler auf einer vierfach gestuften Skala von „stimmt nicht“ bis „stimmt genau“ bewerten sollten. Auf der gleichen Skala wurden acht Items verwendet, um *schulische Kohäsion* zu erfassen (Beispielitem „Wir halten in meiner Klasse fest zusammen“; Cronbachs $\alpha = 0{,}790$). Für beide Variablen wurde zunächst der Mittelwert über die jeweiligen Items für jeden Befragten gebildet und dann auf der Ebene der Klasse aggregiert, wobei mindestens vier gültige Individualangaben pro Klasse zur Verfügung standen.

[3] Für eine weiterführende Diskussion der Bildung von Kontexteigenschaften aus aggregierten Individualangaben v. a. im Hinblick auf Validität und Reliabilität siehe u. a. Oberwittler 2003.

Tabelle 1:

Deskriptive Statistiken der Variablen

		MW	SD	Min	Max
Individualvariablen (Level-1) n = 8.411 Schüler					
Prävalenz Gewaltdelikte	*0 „Nichtopfer“ 1 „Opfer“*	0,13	0,33	0	1
Inzidenz Eigentumsdelikte	*Anzahl der Opfererfahrungen*	0,59	2,78	0	120
Prävalenz für Schulgewalt	*0 „Nichtopfer“ 1 „Opfer“*	0,18	0,38	0	1
Geschlecht	*0 „weiblich“ 1 „männlich“*	0,50	0,50	0	1
Migrationshintergrund	*0 „kein Migrationshintergrund“ 1 „Migrationshintergrund“*	0,24	0,42	0	1
Schulform *Dummyvariablen*	*„Hauptschule“*	0,07	0,25	0	1
	„Realschule“	0,45	0,50	0	1
	„Gymnasium/Gesamtschule“	0,48	0,50	0	1
Risikosuche	*Mittelwertskala aus vier Items*	2,10	0,76	1	4
elterliches Monitoring	*Mittelwertskala aus drei Items*	4,02	0,72	1	5
elterliche Gewalt *Dummyvariablen*	*„keine Gewalt“*	0,57	0,50	0	1
	„leichte Gewalt“	0,31	0,46	0	1
	„schwere Gewalt“	0,12	0,33	0	1
delinquente Freunde *Dummyvariablen*	*„keine delinquenten Freunde“*	0,36	0,48	0	1
	„1 bis 5 delinquente Freunde“	0,55	0,50	0	1
	„6 und mehr delinquente Freunde“	0,09	0,29	0	1
Zeit pro Woche in Kneipe, Disco, Kino etc. *Dummyvariablen*	*„0 Minuten pro Woche“*	0,39	0,49	0	1
	„1 bis unter 240 Minuten pro Woche“	0,10	0,29	0	1
	„240 bis unter 480 Minuten“	0,24	0,43	0	1
	„480 und mehr Minuten“	0,27	0,45	0	1
	Eigenschaften auf Klassenebene (Level-2) n = 454 Klassen				
Desorganisation	*Mittelwertskala aus zwei Items*	2,09	0,34	1,32	3,22
Kohäsion	*Mittelwertskala aus acht Items*	2,67	0,24	1,90	3,27

4 Empirische Analysen und Ergebnisse

Bevor die logistische Regression sowie Modelle für Zähl- und Mehrebenendaten vorgestellt werden, soll das einfache lineare Regressionsmodell (OLS) rekapituliert werden, da dies als Ausgangspunkt für die Erklärung der weiteren Verfahren dient.[4] Ziel des linearen Regressionsmodells ist es, eine abhängige Variable y durch eine oder – im multivariaten Fall – k unabhängige Variablen x_j (mit $j = 1, \ldots k$) zu erklären. Als Gleichung lässt sich der Wert der abhängigen Variable y_i für die Beobachtung i darstellen als:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad (1)$$

Mit x_{1i} und x_{2i} sind die Werte der (in diesem Fall zwei) unabhängigen Variablen für die Beobachtung i bezeichnet. Der Koeffizient β_0 bezeichnet den y-Achsenabschnitt (Intercept), die Koeffizienten (Slopes) β_1 und β_2 geben jeweils den Zusammenhang zwischen den Ausprägungen (genauer: den linearen Einfluss) der unabhängigen Variable x_1 bzw. x_2 und der abhängigen Variable an. Der Fehlerterm ε_i beinhaltet alle nicht im Modell enthaltenen Einflüsse und wird als zufällig und normalverteilt mit einem Mittelwert von null angenommen. Wenn dies zutrifft, kann dieser zufällige Faktor vernachlässigt werden und der mit $\hat{y}_i$ bezeichnete geschätzte Wert der abhängigen Variable der Beobachtung i (z. B. Einkommen) ergibt sich aus den Werten des Befragten für x_1 (z. B. Bildung) und x_2 (z. B. berufliche Stellung) und den entsprechenden Werten der drei Koeffizienten. Die Abweichung zwischen dem tatsächlichen Wert y_i und dem geschätzten Wert $\hat{y}_i$ wird als Residuum $\varepsilon_i = y_i - \hat{y}_i$ ausgedrückt. Für die Schätzung der Koeffizienten aus Gleichung (1) werden diese so gewählt, dass die quadrierten Residuen minimal sind (Methode der kleinste Quadrate = *Ordinary Least Squares* = OLS).

4.1 Logistische Regression: Analyse von Prävalenzdaten

4.1.1 Lineares Wahrscheinlichkeitsmodell und logistische Regression

Daten zur Prävalenz von Opferschaft nehmen *qua definitione* nur zwei Zustände an: Opfer und Nichtopfer. Derartige Merkmale werden in den Sozialwissenschaften als binäre oder dichotome Merkmale bezeichnet und für die statistischen Analysen mit 0 und 1 codiert. Im vorliegenden Fall würde man Nichtopfer mit 0 und Opfer mit 1 codieren und versuchen, die Varianz dieser

[4] Die Darstellung lehnt sich an Windzio 2013, 17–21 an.

Variable mithilfe der unabhängigen Variablen in einem multivariaten Modell zu erklären.

Generell besteht die Möglichkeit, ein Modell zur Erklärung von Prävalenzen in Abhängigkeit diverser unabhängiger Variablen mittels OLS-Regression zu schätzen. Dieses Verfahren wird als lineares Wahrscheinlichkeitsmodell (*Linear Probability Model* = LPM) bezeichnet. Die tatsächliche abhängige Variable ist in diesem Fall allerdings nicht binär, sondern metrisch und wird als Wahrscheinlichkeit P(Y = 1) dessen interpretiert, dass die abhängige Variable den Wert 1 annimmt (Best/Wolf 2010, 828; Windzio 2013, 39–40).

Allerdings führt das LPM zu einer Reihe von Problemen (Best/Wolf 2010, 830; Windzio 2013, 40–42; u. a. Long 1997, 38–40):

– Die durch das Regressionsmodell vorhergesagten Werte sind möglicherweise kleiner als 0 oder größer als 1 und liegen somit außerhalb des für Wahrscheinlichkeiten definierten Bereichs.

– Da die abhängige Variable nur die Werte 0 und 1 annehmen kann, sind die Residuen heteroskedastisch. Heteroskedastizität bedeutet, dass die Varianz der Residuen systematisch von den Werten der unabhängigen Variable abhängt. Dies führt zu verzerrten Standardfehlern und beeinflusst somit die Inferenzstatistiken der Koeffizienten, die Rückschlüsse über die Existenz von Effekten in der Grundgesamtheit erlauben.

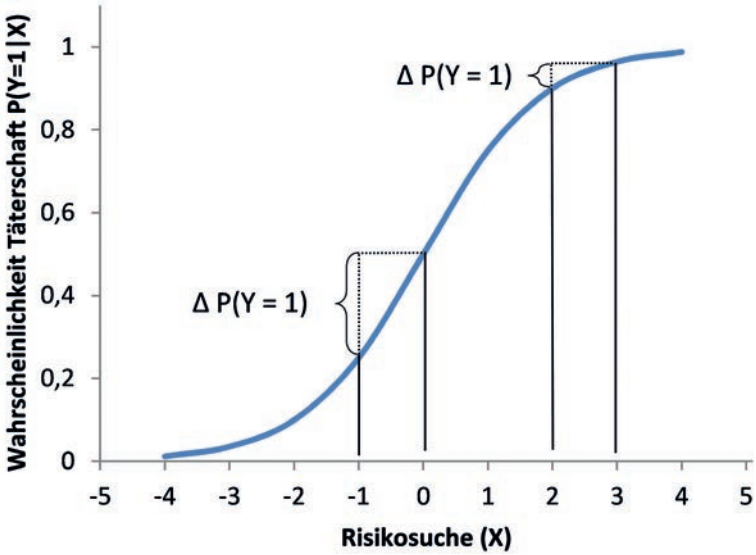
– Die Normalverteilungsannahme der Residuen ist verletzt. Diese können für jede Konstellation der unabhängigen Variablen nur zwei Werte annehmen.

– Darüber hinaus stellt sich die Frage, ob der lineare Zusammenhang zwischen den unabhängigen Variablen und der abhängigen Variable, von dem das LPM ausgeht, funktional angemessen ist. So ist davon auszugehen, dass sich die Wahrscheinlichkeiten den Werten 0 und 1 in Abhängigkeit des Prädiktors nicht linear, sondern asymptotisch annähern, d. h., die Steigung der Funktion, die die Wahrscheinlichkeit von Y = 1 in Abhängigkeit der Ausprägungen der Prädiktoren angibt, sollte am unteren und oberen Bereich geringer als in der Mitte ausfallen. Das bedeutet, dass sich der Effekt eines zusätzlichen Anstiegs der unabhängigen Variable reduziert, wenn sich die Wahrscheinlichkeit den Randbereichen 0 und 1 annähert. Geht man beispielsweise davon aus, dass die Wahrscheinlichkeit eines Jugendlichen, Gewalttäter zu werden, vom Ausmaß seiner Risikobereitschaft abhängt, dann wird der Effekt einer Veränderung der Risikobereitschaft um einen bestimmten Betrag für Jugendliche mit einer sehr geringen oder sehr hohen Risikobereitschaft nur geringen Einfluss auf de-

ren Täterwahrscheinlichkeit haben, da diese ohnehin bereits sehr hoch (hohe Risikobereitschaft) oder sehr niedrig ist (niedrige Risikobereitschaft). Bei Jugendlichen mit mittlerem Niveau an Risikobereitschaft kann deren Veränderung um den gleichen Betrag einen größeren Einfluss auf die Täterwahrscheinlichkeit haben. *Abbildung 1* zeigt dies exemplarisch. So ist der Effekt des Anstiegs der Risikosuche um eine Einheit auf die Veränderung der Täterwahrscheinlichkeit $\Delta P(Y = 1)$ davon abhängig, wo man sich auf der x-Achse befindet. Steigt die Risikosuche x von −1 auf 0, wächst die Wahrscheinlichkeit stärker, als wenn x von 2 auf 3 steigt. In der Abbildung wurde die Skalierung so gewählt, dass die Risikosuche einen Wertebereich von −4 bis 4 hat. Somit entspricht der Anstieg von −1 auf 0 einem Anstieg im mittleren Wertebereich, während der Anstieg von 2 auf 3 sich im höheren Wertebereich abspielt.

Abbildung 1:

Effekt des Anstiegs von X um eine Einheit auf die Veränderung der Wahrscheinlichkeit von Y = 1 in Abhängigkeit von X



Die logistische Regression kann diese Probleme des LPM lösen. Zu diesem Zweck wird die Schätzgleichung transformiert. Die erste Umformung wandelt die Wahrscheinlichkeit, dass Y den Wert 1 annimmt, also $P(Y = 1)$, in Odds (= Chancen) um. Odds stellen Verhältnisse von Wahrscheinlichkeiten

dar – genauer die Beziehung von Eintrittswahrscheinlichkeit zu Gegenwahrscheinlichkeit:

$$Odds = O = \frac{P(Y=1)}{1-P(Y=1)} \quad (2)$$

Liegt etwa die Wahrscheinlichkeit, Opfer einer Straftat zu werden, bei 20 %, dann entspricht dies Odds von $^{20}/_{80} = 0{,}25$. Für eine Wahrscheinlichkeit von 50 % liegen die Odds bei $^{50}/_{50} = 1$; für eine Wahrscheinlichkeit von 80 % liegen die Odds bei $^{80}/_{20} = 4$. Odds haben einen Wertbereich von 0 bis +∞; allerdings ist zu beachten, dass die Transformation von Wahrscheinlichkeiten in Odds nicht linear verläuft. Im Weiteren wird von Odds auch als Chance oder Risiko gesprochen; wenn es um Wahrscheinlichkeiten geht, wird der Begriff Wahrscheinlichkeit verwendet. Diese begriffliche Festlegung ist notwendig, da die Begriffe Risiko bzw. Chance im Deutschen nicht ganz eindeutig sind.

Um die Begrenzung der Odds nach unten ebenfalls zu eliminieren, werden sie logarithmiert. Diese sogenannten Logits haben einen Wertebereich von –∞ bis +∞, da Odds zwischen 0 und 1 durch das Logarithmieren negativ werden (Best/Wolf 2010, 829; Kohler/Kreuter 2008, 262–267; Long 1997, 50–54).

Die durch diese zweifache Transformation aus den Wahrscheinlichkeiten berechneten Logits dienen dann als abhängige Variable einer Linearkombination der k unabhängigen Variablen x_i. Diese Linearkombination ist dem linearen Wahrscheinlichkeitsmodell (LPM) zwar ähnlich, wenngleich nicht die Wahrscheinlichkeit selbst, sondern deren Logit linear von den unabhängigen Variablen abhängt. Für die Interpretation der Ergebnisse bedeutet dies, dass eine Erhöhung einer unabhängigen Variable x_i um eine Einheit nicht die Wahrscheinlichkeit, sondern die logarithmierten Odds der Wahrscheinlichkeit $P(Y = 1)$ um β_i Einheiten ändert. Aufgrund dieser Transformationen ist der Zusammenhang zwischen den unabhängigen Variablen und der Wahrscheinlichkeit nicht linear. Dies bringt Herausforderungen für die Interpretation mit sich, auf die im nächsten Abschnitt eingegangen wird. Durch Transformationen der Gleichung wird jedoch erreicht, dass die abhängige Variable einen nicht begrenzten Wertebereich hat und die funktionale Form des Zusammenhangs zwischen den Prädiktoren und der Auftrittswahrscheinlichkeit eine sinnvolle Form[5] annimmt (Best/Wolf 2010, 829), wie im vorangegangen Beispiel für den Zusammenhang zwischen Risikosuche und Täterschaft ausgeführt. Die Schätzung dieser Gleichung erfolgt im Gegensatz zum linearen Wahrscheinlichkeitsmodell (LPM) nicht mit der Methode der kleinsten Qua-

[5] Für eine ausführlichere Diskussion hierzu siehe Windzio 2013, 39–47.

drate (OLS) sondern mittels Maximum-Likelihood[6] (Best/Wolf 2010, 830; Kohler/Kreuter 2008, 267).[7]

4.1.2 Durchführung und Interpretation der logistischen Regression

Durchführung und Interpretation der logistischen Regression sollen am Beispiel der Prävalenz von Gewaltkriminalität dargestellt werden. Für diese und alle weiteren Analysen in diesem Beitrag wurde auf die Statistiksoftware Stata zurückgegriffen. Im Weiteren wird in Fußnoten jeweils auf die konkreten Stata-Befehle hingewiesen.

In *Tabelle 2* ist das Ergebnis einer logistischen Regression[8] für die Prävalenz der Opferschaft von Gewaltdelikten dargestellt. In der ersten Spalte sind die Koeffizienten der logistischen Regression als Logits dargestellt. Diese können analog zur linearen Regression interpretiert werden: Ein Anstieg einer unabhängigen Variablen (z. B. der Risikosuche) um eine Einheit erhöht die Logits (also die logarithmierten Odds), Opfer einer Gewalttat zu werden, um 0,271. Analog lassen sich Dummyvariablen interpretieren: Hier gibt der jeweilige Koeffizient den Unterschied der Logits zur Referenzgruppe an. Positive Koeffizienten stehen somit für einen positiven Effekt, negative Koeffizienten für einen negativen Effekt des jeweiligen Prädiktors auf die abhängige Variable. Allerdings lassen sich die Logits aufgrund der nicht linearen Transformation der Wahrscheinlichkeiten nicht inhaltlich, sondern nur hinsichtlich Vorzeichen und Signifikanz interpretieren (Best/Wolf 2010, 831).

Die abgebildeten Logits aus *Tabelle 2* (Spalte 1) zeigen ein bei Hauptschülerinnen und Hauptschülern höheres Risiko für Gewaltviktimisierung als bei Realschülerinnen und Realschülern und Gymnasiastinnen und Gymnasiasten, wobei nur der Unterschied zu Letzteren signifikant ist. Ein wichtiger Einflussfaktor ist daneben das Elternhaus: Elterliches Monitoring reduziert das Viktimisierungsrisiko signifikant. Demgegenüber erhöht in der Kindheitspha-

---

6 Bei der OLS-Regression werden die Parameter geschätzt, indem die quadrierten Abweichungen zwischen den vorhergesagten und den beobachteten Werten minimiert werden. Beim Maximum-Likelihood-Verfahren hingegen wird eine Annahme über die durch die unabhängigen Variablen bedingte Verteilung der abhängigen Variable getroffen. Dann werden die unbekannten Parameter β_i so geschätzt, dass die Wahrscheinlichkeit, die beobachteten Werte zu erhalten, maximal ist (Kohler/Kreuter 2008, 267–271; ausführlich Gautschi 2010).

7 Das logistische Regressionsmodel lässt sich auch als latentes Variablenmodell herleiten. Da die Herleitung über die Transformation von Wahrscheinlichkeiten aber besser verständlich ist, wurde darauf verzichtet. Für eine ausführlichere Darstellung siehe Long 1997, 40–50. Kürzere Darstellungen finden sich in Long und Freese 2003, 110–113 sowie Best und Wolf 2010, 834–836.

8 Zu diesem Zweck wurde der Stata-Befehl *logit* verwendet. Die Ergebnisse lassen sich über die entsprechende Option des Befehls entweder als Logits oder als Odds Ratios ausgeben.

se bis zum Alter von 12 Jahren erlebte elterliche Gewalt das Risiko der Jugendlichen, Opfer einer Gewalttat zu werden. Ein weiterer wichtiger Prädiktor ist das Freizeitverhalten: Jugendliche, die delinquente Freunde haben und viel Zeit in Kneipen, Discotheken etc. verbringen, haben ein erhöhtes Risiko, Opfer von Gewalt zu werden.

Tabelle 2:

Logistische Regression der Prävalenz für Gewaltdelikte

	Logit	OR	AME
männlich (1 = ja)	0,042	1,043	0,004
Migrationshintergrund (1 = ja)	0,003	1,003	0,000
Schulform			
Hauptschule	Referenz	Referenz	Referenz
Realschule	-0,182	0,833	-0,021
Gymnasium	-0,470***	0,625***	-0,050***
Risikosuche	0,271***	1,311***	0,027***
elterliches Monitoring	-0,232***	0,793***	-0,023***
elterliche Gewalt			
keine Gewalt	Referenz	Referenz	Referenz
leichte Gewalt	0,725***	2,065***	0,070***
schwere Gewalt	1,275***	3,580***	0,150***
delinquente Freunde			
keine delinquenten Freunde	Referenz	Referenz	Referenz
1 bis 5 delinquente Freunde	0,537***	1,711***	0,049***
6 und mehr delinquente Freunde	1,153***	3,166***	0,130***
Freizeitverhalten			
keine Zeit in Kneipe etc.	Referenz	Referenz	Referenz
wenig Zeit in Kneipe etc.	-0,032	0,969	-0,003
moderat Zeit in Kneipe etc.	0,114	1,121	0,011
viel Zeit in Kneipe etc.	0,269**	1,309**	0,028**
Konstante	-2,403***	0,090***	
Pseudo R^2	0,102		

n = 8.411; *** p < 0,001, ** p < 0,01, * p < 0,05

Eine größere inhaltliche Aussagekraft der Koeffizienten ergibt sich durch eine Umwandlung der Logits durch Entlogarithmierung in *Odds Ratios* (OR). OR sind Verhältnisse von Odds; für die Interpretation von Dummyvariablen bedeutet dies im vorliegenden Beispiel etwa, dass die Chance bzw. das Risiko (Odds), Opfer einer Gewalttat zu werden, für Jungen 1,043-mal so groß ist wie für Mädchen (nicht signifikant). Dabei gilt es zu beachten, dass sich die OR auf die Verhältnisse von Odds und nicht von Wahrscheinlichkeiten beziehen.

Bei der Interpretation der Odds Ratios kontinuierlicher Variablen ist die multiplikative Verknüpfung zwischen der unabhängigen Variablen und den Odds zu beachten. So erhöht der Anstieg der Risikosuche um eine Einheit das Risiko, Opfer einer Straftat zu werden auf das 1,311-Fache; ein Anstieg der Risikosuche um zwei Einheiten hingegen steigert das Risiko auf das $1{,}311 \times 1{,}311 = 1{,}719$-Fache. OR kleiner als 1 geben einen negativen Effekt an, OR größer als 1 einen positiven Effekt (Kohler/Kreuter 2008, 274–275; Best/Wolf 2010, 831–832).

Allerdings ist auch die Interpretation der OR nur bedingt aussagekräftig, wie Best und Wolf (2010, 832–833) ausführen. So handelt es sich um Verhältnisse von Odds (= Wahrscheinlichkeitsverhältnisse). Ohne die Basiswahrscheinlichkeit (bzw. die Odds), als Mädchen Opfer von Gewalt zu werden, zu kennen, kann man keine Aussage über das absolute Risiko etwa der Jungen, Opfer von Gewalt zu werden, treffen. Auch ist die Frage nach der Bedeutung eines Effekts nicht auf Basis der OR zu klären, da z. B. eine Verdoppelung des Risikos je nach Höhe des Ausgangsrisikos einen unterschiedlich großen Effekt auf die Wahrscheinlichkeit hat (auch Windzio 2013, 53–54).

Anschaulicher als OR sind Interpretationen auf der Basis prognostizierter Wahrscheinlichkeiten. Wahrscheinlichkeiten sind darüber hinaus das eingangs dargestellte eigentliche Ziel der logistischen Regression. Möchte man den Effekt einer oder mehrerer unabhängiger Variablen auf die vorhergesagte Wahrscheinlichkeit darstellen, ergibt sich die Schwierigkeit, dass das Verhältnis von Logits und Wahrscheinlichkeiten nicht linear ist. Dies bewirkt, dass der Effekt einer Zunahme der unabhängigen Variable um eine Einheit auf die vorhergesagte Wahrscheinlichkeit über den Wertebereich der unabhängigen Variable variiert. Darüber hinaus hängt der Effekt des Anstiegs einer unabhängigen Variable um eine Einheit im multivariaten Modell auch von den Ausprägungen der übrigen unabhängigen Variablen ab (Windzio 2013, 62–64; auch Best/Wolf 2010; Kohler/Kreuter 2008).

Folgendes Beispiel soll den Effekt des Anstiegs einer unabhängigen Variablen auf die Wahrscheinlichkeit, z. B. Opfer einer Straftat zu werden, in Abhängigkeit des Basisrisikos bei gegebenem OR verdeutlichen. Hat beispielweise eine Variable einen OR von zwei, verdoppelt sich das Risiko,

wenn die Variable um eine Einheit ansteigt. Liegt das Basisrisiko bei $^{5}/_{100}$, führt eine Verdoppelung des Risikos zu Odds von $^{10}/_{100}$. Dementsprechend steigt die Wahrscheinlichkeit von $^{5}/_{105} = 0{,}048$ auf $^{10}/_{110} = 0{,}091$. Würde das Basisrisiko bei $^{50}/_{50} = 1$ liegen, würde die Wahrscheinlichkeit von $^{50}/_{100} = 0{,}500$ auf $^{100}/_{150} = 0{,}667$ steigen. Folglich erhöht der Anstieg der unabhängigen Variablen die Wahrscheinlichkeit je nach Basisrisiko um 0,043 bzw. 0,167. Das Basisrisiko wiederum hängt von den Ausprägungen der anderen Variablen ab.

Dementsprechend kann der Effekt einer unabhängigen Variablen auf die Wahrscheinlichkeit nicht ohne Weiteres in einer Zahl ausgedrückt werden. Gleichwohl bestehen diverse Möglichkeiten, um den Einfluss der unabhängigen Variablen auf die Wahrscheinlichkeit darzulegen.

Eine Möglichkeit ist die *grafische Darstellung*. Hierbei wird die vorhergesagte Wahrscheinlichkeit für verschiedene Ausprägungen einer unabhängigen Variablen berechnet (ggf. getrennt für verschiedene Subgruppen). Da die vorhergesagte Wahrscheinlichkeit auch von den Ausprägungen der anderen unabhängigen Variablen abhängt, muss für alle anderen Prädiktoren, also unabhängigen Variablen, ein Wert festgelegt werden.

Abbildung 2:

Vorhergesagte Wahrscheinlichkeiten

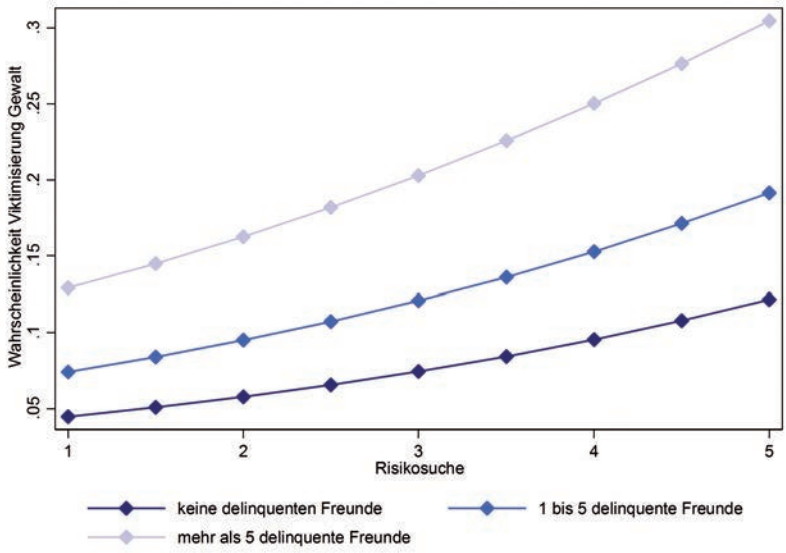


Abbildung 2[9] stellt die vorhergesagte Wahrscheinlichkeit für Gewaltviktimisierung in Abhängigkeit von der Risikosuche und der Zahl der delinquenten Freunde dar. Alle anderen Prädiktoren aus *Tabelle 2* sind am Mittelwert (kontinuierliche Variablen) bzw. an der Referenzkategorie (Dummyvariablen) konstant gehalten. Es wird deutlich, dass die Wahrscheinlichkeit, Opfer von Gewalt zu werden, mit zunehmender Risikosuche steigt. Jedoch hängt der Betrag des Anstiegs davon ab, wie viele delinquente Freunde man hat und welchen Ausgangswert man zugrunde legt, d. h. ob die Risikosuche von 1 auf 2 oder von 4 auf 5 steigt. Dies verdeutlicht noch einmal grafisch die nicht linearen Zusammenhänge zwischen den Variablen und den vorhergesagten Wahrscheinlichkeiten.

Eine zweite Möglichkeit der Darstellung sind sogenannte *marginale Effekte*. Diese geben die Veränderung der Wahrscheinlichkeit an, wenn sich die betreffende Variable um einen infinitesimal kleinen (d. h. gegen null strebenden) Betrag ändert, d. h., es geht um die Steigung der logistischen Funktion (Windzio 2013, 65–66). Auch der marginale Effekt hängt von Ausprägungen aller Variablen im Modell ab (Long/Freese 2003, 139). Demensprechend lassen sich verschiedene marginale Effekte unterscheiden (Windzio 2013, 66–67).[10] Der *Marginal Effect at the Mean* (MEM) berechnet den marginalen Effekt, wenn alle unabhängigen Variablen den Mittelwert aufweisen. Im Gegensatz dazu berechnet der *Average Marginal Effect* (AME)[11] den Mittelwert der marginalen Effekte aus allen im Modell enthaltenen Beobachtungen (Best/Wolf 2010; Long 1997, 74; Windzio 2013, 66). Die AME haben, wie Best und Wolf (2010, 840) ausführen, gegenüber den MEM den Vorteil, dass sie Koeffizienten von Modellen vergleichen können, in die schrittweise mehr Variablen aufgenommen werden. Die letzte Spalte in *Tabelle 2* zeigt die AME für die Prävalenz von Gewaltkriminalität.[12] Dabei tritt zutage, dass die Wahrscheinlichkeit, Opfer einer Gewalttat zu werden, durchschnittlich um 2,7 Prozentpunkte steigt, wenn die Risikosuche um eine Einheit zunimmt; betrachtet

---

9 Die vorhergesagten Wahrscheinlichkeiten, die die Basis der Grafik bilden, lassen sich mit dem Befehl *prgen* berechnen, der Teil des *spost9*-Pakets ist. Das Paket *spost9* wurde von J. Scott Long und Jeremy Freese programmiert und bietet eine Reihe hilfreicher Befehle für die Analyse kategorialer Daten. Dieses Zusatzpaket kann über den Befehl *findit spost9* installiert werden. Eine gute Darstellung des Paktes findet sich in Long und Freese 2003.

10 Für eine mathematische Herleitung der marginalen Effekte siehe Windzio 2013, 65–66 und ausführlicher Long 1997, 72–75.

11 Die AME können mit dem Befehl *margins* ausgegeben werden. Dieser Befehl ist ab Version 12 implementiert. Alternativ kann der Befehl *margeff* (Bartus 2005) genutzt werden.

12 Möchte man Aussagen über die Grundgesamt treffen, sollte dies bei der Berechnung der Inferenzstatistiken (Standardfehler etc.) der AME berücksichtigt werden. Dies ist in Stata als Option des *margins*-Befehls verfügbar (StataCorp 2013, 1172–1173). Für die vorliegende Darstellung wurde darauf verzichtet.

man den Einfluss der Schulform, so zeigt sich, dass Realschülerinnen und Realschüler eine um 2,1 Prozentpunkte (Gymnasiastinnen und Gymnasiasten 5,0 Prozentpunkte) geringere Wahrscheinlichkeit aufweisen, Opfer zu werden, als Hauptschülerinnen und Hauptschüler (= Referenzkategorie). Den größten Effekt auf die Wahrscheinlichkeit, Opfer einer Gewalttat zu werden, haben elterliche Gewalt und der Kontakt zu delinquenten Freunden.

4.1.3 Modellfit, Modellvergleich und Mediationsanalysen

In der linearen Regression steht mit R^2, also dem Anteil der durch das Modell erklärten Varianz an der Gesamtvarianz der abhängigen Variable, ein Maß für die Anpassungsgüte des Modells zur Verfügung. Für die logistische Regression existieren verschiedene Kennzahlen, die die Anpassungsgüte eines Modells angeben und zum Vergleich der Erklärungskraft von Modellen herangezogen werden können. Diese Maße sollten einen Wertebereich zwischen 0 und 1[13] aufweisen, wobei ein Modell mit einem Wert von 0 keinerlei Erklärungskraft hat und Modelle, die höhere Werte erzielen eine größere Erklärungskraft besitzen.

Ein Maß, das auf dem Vergleich der Likelihood des geschätzten Modells mit der Likelihood eines Modells ohne erklärende Variablen (Nullmodell) basiert, ist *McFaddens Pseudo-R^2*. Andere Maße wie *Cox&Snell R^2*, *Cragg&Uhler R^2* oder *Nagelkerke-R^2* basieren auf demselben Prinzip, enthalten aber zusätzlich eine Normierungs- oder Korrekturkomponente.[14] Bei der Verwendung der Maße muss beachtet werden, dass es 1.) keinen einheitlichen Standard gibt, 2.) mehr unabhängige Variablen prinzipiell zu höheren Werten führen und 3.) verschiedene Maße nicht vergleichbar sind. So liefert Nagelkerke-R^2 immer höhere Werte als andere Maße (Best/Wolf 2010, 843–844).[15]

In multivariaten Modellen ist man oft daran interessiert, Mediationseffekte zu untersuchen, d. h. man beschäftigt sich mit der Frage, wie sich Koeffizienten verändern, wenn man weitere Variablen in das Modell aufnimmt. Das Problem in nicht linearen Modellen ist, dass die Regressionskoeffizienten nicht

---

13 Allerdings können nicht alle Maße (z. B. McFaddens Pseudo R^2, Cox&Snell R^2) auch tatsächlich den Wert 1 erreichen (Best/Wolf 2010, 843).

14 Stata gibt standardmäßig *McFaddens Pseudo-R^2* aus; der Befehl *fitstat* aus dem *spost9*-Paket berechnet zahlreiche weitere Fitmaße.

15 Neben diesen Maßen der Anpassungsgüte können auch Informationsmaße wie *Akaike's Information Criterion* (AIC) und das *Bayesian Information Criterion* (BIC) zur Modellselektion herangezogen werden (Best/Wolf 2010, 844; Long/Freese 2003, 94–95).

miteinander über Modelle hinweg verglichen werden können (Best/Wolf 2010, 838; Long 1997, 70; Mood 2009; Windzio 2013, 69–71 dort auch genauere Diskussion). Um dennoch Koeffizienten zwischen Modellen vergleichen zu können, bestehen zwei Möglichkeiten: Zum einen können wie bereits erwähnt die AME über Modelle hinweg verglichen, zum anderen voll- oder teilstandardisierte Koeffizienten berechnet werden, die Verzerrungen reduzieren, aber nicht komplett eliminieren (Best/Wolf 2010, 838–839). [16]

In *Tabelle 3* sind zwei Modelle für die Prävalenz von Gewaltdelikten dargestellt. Der Unterschied zwischen beiden Modellen besteht darin, dass in Modell 2 der Einfluss des elterlichen Erziehungsverhaltens hinzugefügt wurde (Modell 2 ist somit identisch mit dem Modell aus *Tabelle 2*). Betrachtet man zunächst den Modellfit, zeigt sich ein Anstieg der Erklärungskraft des Modells. Ein Likelihood-Ratio-Test[17] (LR $chi^2_{df=3} = 250{,}56$; $p < 0{,}001$) verdeutlicht zudem, dass die neu hinzugenommenen Parameter zu einer signifikanten Modellverbesserung führen.

16 Eine andere Möglichkeit ist die Reskalierung von Koeffizienten (Karlson u. a. 2012).

17 Ein Likelihood-Ratio-Test kann in Stata mit dem Befehl *lrtest* durchgeführt werden.

Tabelle 3:

Logistische Regression der Prävalenz für Gewaltdelikte: unstandardisierte und standardisierte Koeffizienten

Modell	(1)		(2)	
	Logit	stdXY	Logit	stdXY
männlich (1 = ja)	–0,008	–0,002	0,042	0,010
Migrationshintergrund (1 = ja)	0,174*	0,038	0,003	0,001
Schulform				
Hauptschule	Referenz		Referenz	
Realschule	–0,211	–0,055	–0,182	–0,045
Gymnasium	–0,548***	–0,142	–0,470***	–0,118
Risikosuche	0,345***	0,135	0,271***	0,103
delinquente Freunde				
keine delinquenten Freunde	Referenz			
1 bis 5 delinquente Freunde	0,673***	0,174	0,537***	0,134
6 und mehr delinquente Freunde	1,365***	0,203	1,153***	0,165
Freizeitverhalten				
keine Zeit in Kneipe etc.	Referenz		Referenz	
wenig Zeit in Kneipe etc.	–0,050	–0,008	–0,032	–0,005
moderat Zeit in Kneipe etc.	0,086	0,019	0,114	0,024
viel Zeit in Kneipe etc.	0,237**	0,055	0,269**	0,060
elterliches Monitoring			–0,232***	–0,083
elterliche Gewalt				
keine Gewalt			Referenz	
leichte Gewalt			0,725***	0,168
schwere Gewalt			1,275***	0,211
McFaddens R^2	0,063		0,102	

n = 8.411; *** p < 0,001, ** p < 0,01, * p < 0,05

Zusätzlich sind in *Tabelle 3* die standardisierten Koeffizienten aufgeführt.[18] Standardisierte Koeffizienten β^S geben die Veränderung der Logits um β^S Standardabweichungen an, wenn die entsprechende unabhängige Variable *x* um eine Standardabweichung verändert wird, und erlauben einen modellübergreifenden Vergleich der Koeffizienten. Im vorliegenden Beispiel haben die

[18] Diese lassen sich mit dem Befehl *listcoef* im Anschluss an die Modellschätzung ausgeben, der ebenfalls zum *spost9*-Paket gehört.

meisten Variablen durch die Hinzunahme der weiteren Prädiktoren geringere standardisierte Koeffizienten, wobei sich Richtung und Signifikanz nicht ändern. Ein interessanter Mediationseffekt zeigt sich im Hinblick auf den Migrationsstatus. Dessen Effekt verschwindet unter Kontrolle des elterlichen Erziehungsverhaltens. Eine Erklärung lautet, dass Jugendliche mit Migrationshintergrund häufiger elterliche Gewalt erleben (Baier/Pfeiffer 2007), die einen positiven Effekt auf Gewaltviktimisierung hat.

4.1.4 Logistische Regression: Fazit und Ausblick

Die logistische Regression erlaubt die Analyse dichotomer Variablen, wie Viktimisierungsprävalenzen. Im Gegensatz zum linearen Wahrscheinlichkeitsmodell (LPM), das sich aus der OLS-Regression ableitet, spezifiziert die logistische Regression einen nicht linearen Zusammenhang zwischen der vorhergesagten Wahrscheinlichkeit und der unabhängigen Variable, sodass die Interpretation der Ergebnisse schwieriger ist als in der OLS-Regression. Während Logits und Odds Ratios (OR) nur im Hinblick auf Signifikanz und Richtung des Effekts aussagekräftig sind, erlauben Kennzahlen auf Basis vorhergesagter Wahrscheinlichkeiten anschaulichere Aussagen.

Ein Aspekt, der aus Platzgründen nicht ausführlich besprochen werden konnte und vor allem die praktische Durchführung logistischer Regressionen betrifft, ist die *Modelldiagnostik*. Insbesondere die Frage nach dem funktionalen Zusammenhang (Kohler/Kreuter 2008, 283–287) und die Analyse von Residuen und einflussreichen Fällen stehen dabei im Vordergrund (Windzio 2013, 71–72; Long/Freese 2003, 124–127; ausführlich Hosmer/Lemeshow 2000).

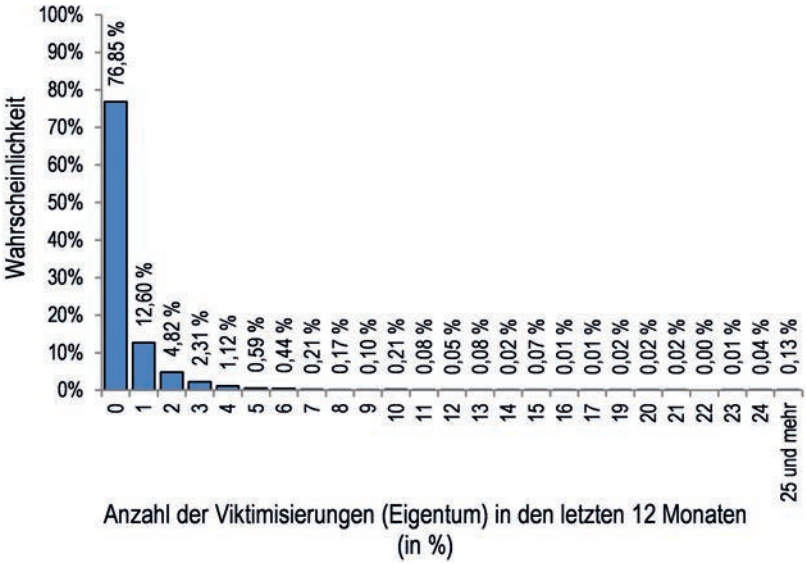
4.2 Zähldaten – Inzidenzen

Im Gegensatz zu Prävalenzen geht es bei Inzidenzen um die Frage, wie häufig ein bestimmtes Ereignis in einem Zeitintervall auftritt. Bezogen auf Opfererfahrungen steht also die Anzahl der Opfererlebnisse in einem bestimmten Zeitraum im Fokus. In der Empirie zeigt sich, dass Inzidenzen oftmals sehr (rechts-)schief verteilt sind, da zahlreiche Personen keine Opfererfahrungen machen. Darüber hinaus haben die meisten Opfer auch nur wenige Viktimisierungen erlebt. Im vorliegenden Beispiel weist die Inzidenz für Eigentumsdelikte eine rechtsschiefe Verteilung auf (*Abbildung 3*). Diese schiefe Verteilung der Inzidenzen kann zu verzerrten Ergebnissen einer OLS-Schätzung führen (Windzio 2013, 193–194).[19]

[19] Die Tatsache, dass Inzidenzen diskrete Variablen (nur ganzzahlige Werte) und keine stetig normalverteilten Variablen sind, führt nicht notwendigerweise zu verzerrten Ergebnissen einer OLS-Schätzung (Windzio 2013, 193–194).

Abbildung 3:

Inzidenzen der Viktimisierung durch Eigentumsdelikte (n = 8.411)[20]



4.2.1 Poisson- und negative Binomialregression

Um dennoch eine Regression schätzen zu können, ist eine Verteilungsfunktion zu verwenden, die der empirischen Verteilung der zu erklärenden Variable entspricht. Im vorliegenden Fall muss die Verteilungsfunktion also diskret (d. h. abzählbar) und asymmetrisch sein. In der Regel wird hierfür die *Poisson-Verteilung* verwendet, die durch einen einzigen, sowohl den Mittelwert als auch die Varianz der Verteilung beschreibenden Parameter μ bestimmt wird. Diese Eigenschaft wird als Equidispersion bezeichnet. Die Poisson-Verteilung gibt die Wahrscheinlichkeit an, dass eine zufällig ausgewählte Person eine bestimmte Zahl von Ereignissen (hier: Viktimisierungen) erlebt hat (Cameron/Trivedi 1998, 1–18; Long/Freese 2003, 245–251; Windzio 2013, 195). Angenommen die empirische Verteilung der Inzidenz von Diebstahl folgt einer Poisson-Verteilung, dann gibt der Mittelwert μ die durchschnittliche Zahl

[20] Personen mit 25 und mehr Opfererlebnissen wurden für die Darstellung der Übersichtlichkeit halber zusammengefasst.

der Viktimisierungen in der jeweiligen Stichprobe an. Darüber hinaus ergibt sich aus der Poisson-Verteilung mit dem spezifischen Wert μ die Wahrscheinlichkeit dafür, dass eine zufällig ausgewählte Person aus der Stichprobe z. B. vier Viktimisierungen erlebt hat. Je höher μ ist, desto mehr verschiebt sich der Modalwert (d. h. der häufigste Wert) der Poisson-Verteilung nach rechts, d. h. die Kategorie mit der höchsten Wahrscheinlichkeit hat eine immer höhere Anzahl an Ereignissen (Windzio 2013, 196–197).

Bei der *Poisson-Regression* (PRM) ergibt sich die Wahrscheinlichkeit einer Beobachtung i (d. h. der Person i), eine bestimme Zahl von Viktimisierungen erlebt zu haben, aus einer Poisson-Verteilung mit dem Parameter μ_i. Dieser Parameter resultiert aus den Ausprägungen von i der unabhängigen Variablen und den Koeffizienten[21] und lässt sich auch als Inzidenzrate interpretieren, d. h. als die durchschnittliche Zahl an Viktimisierungen für die jeweilige Beobachtung (Long/Freese 2003, 251–252; Windzio 2013, 196–197).

Ein Problem der PRM ist allerdings, dass die Annahme der Equidispersion, also der Äquivalenz von Mittelwert und Varianz, empirisch nicht immer erfüllt ist (Windzio 2013, 198). Auch im vorliegenden Beispiel zur Inzidenz von Eigentumskriminalität ist dies der Fall. So beträgt die Varianz 7,73, fällt also deutlich größer als der Mittelwert (0,59) aus; man spricht hier von einer Überdispersion, die in der Regel angesichts zahlreicher Fälle, die im Referenzzeitraum kein Ereignis erlebt haben, entsteht. Allerdings ist zu fragen, ob die Abweichung von der Equidispersion auch unter Kontrolle der unabhängigen Variablen weiter fortbesteht (Windzio 2013, 198).

Eine Möglichkeit, um auch in Fällen von Überdispersion Regressionsschätzungen vornehmen zu können, ist die *negative Binomialregression* (NBRM). Zu diesem Zweck wird die Poisson-Verteilung um eine empirisch geschätzte Varianz erweitert. Das PRM ist im NBRM *genested*,[22] sodass beide Modelle miteinander verglichen und mittels statistischen Tests überprüft werden kann, ob tatsächlich eine Abweichung von der Equidispersion vorliegt (Windzio 2013, 198–199). Liegt eine Überdispersion vor, dann sind die Schätzer des PRM ineffizient und die Standardfehler zu klein (Long/Freese 2003, 269).

---

21 Konkret wird der logarithmierte Wert von μ_i als Linearkombination der unabhängigen Variablen geschätzt (Windzio 2013, 195)

22 Genested bedeutet, dass Modell A in Modell B enthalten ist. Dies ist der Fall, wenn Modell B durch das Nullsetzen von Parametern in Modell A überführt werden kann.

4.2.2 Durchführung und Interpretation von Poisson- und negativer Binomialregression

Bei der Analyse von Zähldaten sollte zunächst geprüft werden, ob eine Abweichung von der Equidispersion unter Kontrolle der unabhängigen Variablen vorliegt, d. h. ob der konditionale Mittelwert der konditionalen Varianz entspricht. Das Statistikprogramm Stata berechnet hierzu den Parameter α: Wenn α null ist, dann entspricht die konditionale Varianz dem konditionalen Mittelwert und somit ist das NBRM mit dem PRM identisch (Long/Freese 2003, 268–270; Windzio 2013, 199). In *Tabelle 4* sind für die gleichen Prädiktoren beide Modelle für die Erklärung der Inzidenz von Viktimisierung durch Eigentumskriminalität geschätzt worden. Für α zeigt sich ein Wert von 4,345, der sich signifikant von 0 unterscheidet.[23] Folglich liegt eine Überdispersion vor und es sollte ein NBRM verwendet werden.

Für die Interpretation wurden in *Tabelle 4* die exponierten Koeffizienten angegeben (*Incidence Rate Ratios* = IRR), die sich ähnlich wie die Odds Ratios (OR) interpretieren lassen.[24] So rangiert die Viktimisierungsrate (also die durchschnittliche Zahl der erwarteten Ereignisse im gegebenen Zeitraum) von Jungen im Vergleich zu Mädchen um den Faktor 1,16 höher. Analog lassen sich die anderen Dummyvariablen auslegen. Bei den kontinuierlichen Variablen können die IRR als Anstieg der erwarteten Anzahl an Ereignissen interpretiert werden, wenn die unabhängige Variable um eine Einheit steigt. So führt etwa ein Anstieg der Risikosuche um eine Einheit zu einer Erhöhung der erwarteten Zahl an Viktimisierungen um den Faktor 1,186. Ähnlich den OR der logistischen Regression hängt die tatsächliche Rate der Viktimisierungen von den Ausprägungen aller anderen Variablen ab (Long 1997, 224–228; Windzio 2013, 200–201). Die inhaltliche Interpretation der Ergebnisse zu Prädiktoren der Inzidenz von Eigentumsviktimisierung erfolgt unter 4.2.4.

---

23 Stata berechnet dazu einen Likelihood-Ratio-Test für die Nullhypothese $\alpha = 0$ (StataCorp 2013, 1397).

24 Das PRM wird in Stata mit dem Befehl *poisson* geschätzt, das NBRM mit *nbreg*. Die Option *irr* gibt die Koeffizienten als IRR aus.

Tabelle 4:

Poisson- (PRM) bzw. negatives Binomialmodell (NBRM) für die Inzidenz von Eigentumskriminalität (*Incidence Rate Ratios* = IRR)

Modell	(1) PRM	(2) NBRM
männlich (1 = ja)	1,295***	1,160*
Migrationshintergrund (1 = ja)	1,043	1,197**
Schulform		
Hauptschule	Referenz	Referenz
Realschule	1,028	0,870
Gymnasium	1,128*	0,957
Risikosuche	1,104***	1,186***
elterliches Monitoring	0,802***	0,833***
elterliche Gewalt		
keine Gewalt	Referenz	Referenz
leichte Gewalt	1,201***	1,308***
schwere Gewalt	1,873***	1,932***
delinquente Freunde		
keine delinquenten Freunde	Referenz	Referenz
1 bis 5 delinquente Freunde	1,912***	1,824***
6 und mehr delinquente Freunde	4,736***	4,607***
Freizeitverhalten		
keine Zeit in Kneipe etc.	Referenz	Referenz
wenig Zeit in Kneipe etc.	0,772***	0,830
moderat Zeit in Kneipe etc.	1,039	1,019
viel Zeit in Kneipe etc.	1,148***	1,156*
Konstante	0,386***	0,337***
Alpha		4,345
McFaddens R^2	0,106	0,040

n = 8.411; *** p < 0,001, ** p < 0,01, * p < 0,05

4.2.3 „Exzess“ von Nullen: *zero-inflated* Modelle

Im vorhergehenden Beispiel wurde deutlich, dass die Angemessenheit des Poisson-Modells (PRM) nicht immer gegeben ist (häufig aufgrund einer sehr hohen Anzahl von Nullen). In manchen Fällen ist die Zahl der Nullen so groß, dass man von einem „Exzess“ der Nullen spricht, der auch durch die negative Binomialverteilung nicht mehr abgebildet werden kann. Ein Exzess von Nullen kann auf soziale Prozesse hindeuten, denen zwei latente Gruppen zugrunde liegen. Diese beiden Gruppen unterscheiden sich signifikant in Bezug auf die abhängige Variable, jedoch gibt es keinen manifesten Faktor, der beide Gruppen identifiziert. Sie sind deshalb unbeobachtet, d. h. latent (Windzio 2013, 201). Bezogen auf die Erklärung von Inzidenzen bedeutet dies zweierlei: Erstens ist zu untersuchen, welche Faktoren die Zugehörigkeit zur Gruppe der Nichtopfer, d. h. Personen die keinerlei Risiko einer Opferwerdung aufweisen, bzw. der potenziellen Opfer, d. h. Personen, die einem Risiko unterliegen, Opfer zu werden, determinieren. Zweitens muss man sich bei den potenziellen Opfern damit befassen, welche Faktoren für die Anzahl der Opfererlebnisse verantwortlich sind. Denkbar ist, dass sich die jeweils relevanten Faktoren hinsichtlich ihres Einflusses unterscheiden. Andere Beispiele aus der Kriminologie wenden diese Idee auf die Erklärung von Gewalttäterschaft an (Windzio/Baier 2009).

Für die Erklärung solcher Phänomene können *zero-inflated* Poisson-Modelle (ZIP) oder *zero-inflated* negative Binomialmodelle (ZINB) verwendet werden. Diese Modelle schätzen den Einfluss der unabhängigen Variablen sowohl auf die Wahrscheinlichkeit, zu einer der beiden Gruppen zu gehören, als auch auf die Anzahl der Ereignisse. Hierfür werden simultan ein binäres Modell zur Erklärung der Viktimisierung (wobei gilt: 0 = Opfer und 1 = Nichtopfer) und ein Modell zur Erklärung der Häufigkeit der Ereignisse berechnet. Aus den vorhergehenden Ausführungen wurde bereits deutlich, dass sich für Ersteres Logitmodelle anbieten, während für die Erklärung der Inzidenz Poisson- oder negative Binomialmodelle infrage kommen (Windzio 2013, 202–203).

Bisher wurden vier Modelle (PRM, NBRM, ZIP und ZINB) zur Erklärung der Inzidenz von Viktimisierung durch Eigentumskriminalität vorgestellt. Alle vier Modelle lassen sich auf die gleichen Daten, im vorliegenden Fall auf die Erklärung von Inzidenzen, anwenden. Daher stellt sich die Frage, welches der vier das angemessene Modell ist, zu deren Beantwortung alle Modelle auf identische Daten angewendet und miteinander verglichen werden:

– Für den Vergleich von PRM und NBRM wurde bereits der Test des Dispersionsparameters α vorgestellt, der im vorliegenden Fall signifikant von 0 verschieden ist, es liegt also keine Equidispersion vor; somit ist das NBRM dem PRM vorzuziehen.

– Um zu überprüfen, ob der Exzess an Nullen so groß ist, dass ein zero-inflated Modell notwendig ist, kann der Vuong-Test[25] verwendet werden. Dieser vergleicht das PRM mit dem ZIP bzw. das NBRM mit dem ZINB (Windzio 2013, 204–205). Im vorliegenden Fall wird sowohl für die Poisson-Modelle (nicht dargestellt) als auch für die negativen Binomialmodelle das jeweilige Standardzählmodell zugunsten der zero-inflated Version verworfen. Dies ist aus den signifikanten positiven Werten des Voung-Tests[26] ersichtlich.

– Schließlich kann noch verglichen werden, welches der beiden zero-inflated Modelle ZIP und ZINB angemessener ist. Da beide Modelle genested sind, kann dies mittels eines Likelihood-Ratio-Tests überprüft werden. Analog zum Vergleich von PRM und NBRM wird getestet, ob sich der Überdispersionsparameter α signifikant von 0 unterscheidet (Long/Freese 2003, 285). Im vorliegenden Fall verwirft der Likelihood-Ratio-Test[27] die Nullhypothese H_0: $\alpha = 0$ mit $p < 0{,}001$, sodass das ZINB dem ZIP vorgezogen wird.[28]

4.2.4 Modellschätzung und Interpretation

Tabelle 5 gibt die Ergebnisse des zero-inflated negativen Binomialmodells (ZINB) für die Inzidenz von Eigentumsviktimisierung wieder. Hierbei werden für jede Variable zwei Koeffizienten ausgegeben. Der erste Koeffizient veranschaulicht den Einfluss der Variable auf die Wahrscheinlichkeit, zur Gruppe der Nichtopfer ($Y = 0$) zu gehören. Die Koeffizienten sind hier als Odds Ratio (OR) dargestellt und werden analog zur logistischen Regression interpretiert. Im Ergebnis zeigt sich, dass Jungen eine höhere Wahrscheinlichkeit haben, zur Gruppe der *Nichtopfer* zu gehören (OR von 2,025). Gleichzei-

---

25 Für eine detailliertere Darstellung des Tests siehe Windzio 2013, 204–205 und Long 1997, 248–249. In Stata kann der Test als Option (*vuong*) bei der Schätzung eines ZIP- bzw. ZINB-Modells angefordert werden.

26 Die Teststatistik für den Vergleich von NBRM und ZINB ist in *Tabelle 5* angegeben.

27 Siehe Long und Freese 2003, 285 für Details und Teststatistik.

28 Daneben existiert auch eine grafische Methode, die die auf Basis des jeweiligen Modells vorhergesagten Wahrscheinlichkeiten mit den beobachteten Wahrscheinlichkeiten vergleicht (hierzu Long 1997, 247–249; Long/Freese 2003, 283–284; Windzio 2013, 205–206).

tig beeinflusst das Geschlecht auch die Inzidenz signifikant: Jungen haben eine höhere Inzidenzrate als Mädchen (IRR von 1,265). Eine Erklärung könnte lauten, dass Jungen eher dazu neigen, Eigentumskriminalität nicht als solche wahrzunehmen und dementsprechend nicht im Fragebogen angeben. Diejenigen, die aber derartige Erlebnisse als Kriminalität interpretieren und angeben, erleben Derartiges tatsächlich häufiger. Andere Prädiktoren zeigen demgegenüber konsistente Effekte. So weisen Jugendliche, die delinquente Freunde haben, eine höhere Wahrscheinlichkeit auf, in der Gruppe der Opfer zu sein, und gleichzeitig erhöht die Präsenz delinquenter Freunde die Inzidenzrate. Risikosuche hat nur einen Effekt auf die Gruppenzugehörigkeit von Opfern und Nichtopfern, nicht jedoch auf die Häufigkeit der Opferwerdung. Elterliches Monitoring hat keinen Einfluss auf die Gruppenzugehörigkeit, reduziert aber die Zahl der Viktimisierungen signifikant. Elterliche Gewalt erhöht die Wahrscheinlichkeit, zur Gruppe der Opfer zu gehören (nur leichte Gewalt signifikant), und die Zahl der Opferwerdungen (nur schwere Gewalt signifikant).

Abschließend sei auf weitere Möglichkeiten von Interpretation und Darstellung der Ergebnisse von Zähldatenmodellen verwiesen. Analog zum logistischen Modell lassen sich verschiedene Kennziffern des Zusammenhangs auf Basis vorhergesagter Wahrscheinlichkeiten berechnen und grafisch darstellen (Long/Freese 2003).

Tabelle 5:

Zero-inflated negatives Binomialmodell zur Erklärung der Inzidenz von Eigentumsviktimisierung

	ZINB	
	OR (Y = 0)	IRR
männlich (1 = ja)	2,025*	1,265***
Migrationshintergrund (1 = ja)	0,753	1,140
Schulform		
Hauptschule	Referenz	Referenz
Realschule	1,232	0,927
Gymnasium	1,166	0,998
Risikosuche	0,450**	1,080
elterliches Monitoring	1,002	0,835***

	ZINB	
	OR (Y = 0)	IRR
elterliche Gewalt		
keine Gewalt	Referenz	Referenz
leichte Gewalt	0,165*	1,091
schwere Gewalt	0,580	1,764***
delinquente Freunde		
keine delinquenten Freunde	Referenz	Referenz
1 bis 5 delinquente Freunde	0,686	1,711***
6 und mehr delinquente Freunde	0,982	4,365***
Freizeitverhalten		
keine Zeit in Kneipe etc.	Referenz	Referenz
wenig Zeit in Kneipe etc.	0,105	0,645**
moderat Zeit in Kneipe etc.	0,274*	0,839
viel Zeit in Kneipe etc.	0,464	1,002
Konstante	1,776	0,565*
Vuong		2,698**
Alpha		3,584
McFaddens R^2		0,043

n = 8.411; *** p < 0,001, ** p < 0,01, * p < 0,05

4.3 Mehrebenenmodelle

In den vorherigen Abschnitten bestand die Problematik darin, dass Skalierung bzw. Verteilung der abhängigen Variable eine Anwendung der OLS-Regression ausschließt. In diesem Abschnitt steht demgegenüber die Struktur der Daten selbst im Fokus. Konkret geht es darum, dass die Beobachtungen in Kontexten enthalten sind; man spricht in diesem Fall auch von geklumpten, geclusterten oder hierarchischen Daten. Als Beispiele für hierarchische Datenstrukturen lassen sich etwa Schülerinnen und Schüler in Schulklassen, Bewohnerinnen und Bewohner einer Nachbarschaft oder auch Straftäterinnen und Straftäter in Gefängnissen anführen. Dabei wird die kleinste Ebene (z. B. Schülerinnen und Schüler) als Level-1 bezeichnet und die Kontextebene als Level-2 (z. B. Schulklasse). Dieses Design lässt sich auch erweitern. So wäre

bspw. die zusätzlich eingefügte Ebene der Schule Level-3. In den im Folgenden diskutierten Modellen liegt die abhängige Variable immer auf der untersten Ebene (Level-1).

Weisen Daten eine solche hierarchische Struktur auf, dann werden alle Beobachtungen dieses Kontexts durch identische Umweltbedingungen beeinflusst. In der Folge – vorausgesetzt diese Umweltbedingungen beeinflussen die abhängige Variable – sind sich die Beobachtungen eines Kontexts ähnlicher als Beobachtungen aus verschiedenen Kontexten. Dies führt aber zur Verletzung der Annahme statistischer Unabhängigkeit der Beobachtungen, die die meisten statistischen Verfahren voraussetzen (Windzio 2008, 113–114).

Eine Maßzahl zur Beschreibung der Ähnlichkeit von Einheiten desselben Kontexts ist der *Intraclass Correlation Coefficient* (ICC), definiert als der Quotient aus der Varianz zwischen den Kontexten und der Gesamtvarianz. Die Maßzahl lässt sich interpretieren als der Anteil der Varianz der abhängigen Variable, der durch die Gruppenstruktur erklärt wird. Je höher der ICC ist, desto ähnlicher sind sich Beobachtungen aus demselben Kontext (für die Berechnung siehe Snijders/Bosker 2012, 17–23).

Die hierarchische Datenstruktur kann dabei einerseits ein unerwünschter Nebeneffekt des Stichprobendesigns sein. So wird aus praktischen Erwägungen und aus Kostengründen oft ein mehrstufiges Stichprobendesign verwendet, bei dem zunächst die Level-2-Einheiten gezogen werden und aus diesen dann die zu untersuchenden Level-1-Einheiten. Andererseits kann eine hierarchische Datenstruktur aber auch helfen, interessante soziale Phänomene insofern aufzudecken, als Eigenschaften des Kontexts direkt oder indirekt Eigenschaften der Individuen beeinflussen (Snijders/Bosker 2012, 6–9; Windzio 2008).

Allgemein betrachtet lassen sich mittels eines Mehrebenendesigns drei Arten von Hypothesen überprüfen. So können zwei Level-1-Variablen in Beziehung gesetzt werden (Mikrohypothese), um z. B. den Einfluss der elterlichen Erziehung auf das Risiko der Opferwerdung zu untersuchen. In diesem Fall wäre die Multilevelstruktur eher als „Nuisance“ (Störfaktor)[29] zu betrachten, den es bei der Analyse zu berücksichtigen gilt. Daneben können aber auch substanzielle Hypothesen zum Einfluss des Kontexts geprüft werden. Makro-Mikro-Hypothesen untersuchen den Effekt von Level-2-Variablen (z. B. Klassenklima) auf Level-1-Variablen (Viktimisierungsrisiko). Cross-Level-Interaktionen schließlich befassen sich mit dem Effekt von Level-2-Variablen auf

[29] Ist man nicht an der Aufdeckung von Kontexteffekten interessiert, können statt Mehrebenenmodellen auch normale Regressionsverfahren mit Standardfehlern geschätzt werden, die gegenüber Klumpenstichproben robust sind („geclusterte“ Standardfehler).

die Beziehung zwischen zwei Level-1-Variablen. So ist es denkbar, dass das Klassenklima den Zusammenhang zwischen der elterlichen Erziehung der Schülerinnen und Schüler und dem Viktimisierungsrisiko beeinflusst, da insbesondere in einem positiven Umfeld Defizite von benachteiligten Schülern ausglichen werden können (Snijders/Bosker 2012, 9–12).

4.3.1 Das Mehrebenenmodell

Die nachfolgende Darstellung versucht möglichst ohne mathematische Formeln auszukommen und lehnt sich an Windzio (2008) an. Es wird im Folgenden von Personen in Kontexten die Rede sein, die Ausführungen lassen sich aber allgemein auf hierarchische Daten übertragen. Ausgangspunkt der Darstellung der Mehrebenenanalyse ist das oben vorgestellte lineare Regressionsmodell. Der Wert der abhängigen Variablen y_i für Person i hängt von einem Intercept β_0, dem Koeffizienten β_1, der Ausprägung der unabhängigen (Level-1-)Variable x_{1i} für das Individuum i und dem Fehlerterm ε_i ab.

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \tag{3}$$

Im Fall von Personen, die in Kontexten geclustert sind, kann Gleichung (3) um einen eigenen Intercept β_{0j} für jeden Kontext erweitert werden (4). Die abhängige Variable y_{ij} gibt somit den Wert der $i$ten Person in Kontext j an, der neben dem Fehlerterm ε_{ij} von den Werten von x_{ij}, dem Koeffizienten β_1 und dem Intercept β_{0j} des Kontexts j, dem die jeweilige Person angehört, abhängt.

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \varepsilon_{ij} \tag{4}$$

Der spezifische Intercept β_{0j} für einen Kontext j (z. B. für eine bestimmte Klasse) ergibt sich aus dem Mittelwert der Intercepts über alle Kontexte β_{00} und den Abweichungen μ_{0j} der einzelnen Kontexte von diesem Mittelwert. Ist die Abweichung μ_{0j} negativ, dann weisen die Mitglieder dieses Kontexts im Durchschnitt geringere Werte für y auf als der Durchschnitt aller Beobachtungen (kontrolliert um den Effekt der unabhängigen Variablen x). Die kontextspezifischen Abweichungen μ_{0j} entstammen dabei einer Zufallsverteilung, die aus den Daten geschätzt wird. Die Varianz dieser Verteilung gibt daher an, wie stark sich die Mittelwerte der Kontexte voneinander unterscheiden. Wäre die Varianz 0, würden wir keine Unterschiede zwischen den Kontexten beobachten. Das Modell in Gleichung (4) wird als *Random-Intercept*-Modell bezeichnet.

Die Logik der Varianz des Intercepts lässt sich auch auf die Koeffizienten der unabhängigen Variablen übertragen. Hierzu erweitert man die zu schätzende Gleichung um einen eigenen Koeffizienten (Slope) β_{1j} für jeden Kontext. Das

bedeutet im Fall der linearen Regression, dass die Steigungen der Regressionsgeraden zwischen den Kontexten variieren (*Random-Slope*-Modell).

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \varepsilon_{ij} \tag{5}$$

Analog zum Random-Intercept-Modell ergibt sich der Koeffizient β_{1j} für den Kontext j aus dem Mittelwert β_{10} über alle kontextspezifischen Slopes und der Abweichung μ_{1j} des Kontexts j von diesem Mittelwert. μ_{1j} entstammt ebenfalls einer Zufallsverteilung, die aus den Daten geschätzt wird. Je größer deren Varianz ist, desto stärker unterscheiden sich die Slopes zwischen den Kontexten. Variieren die Slopes signifikant (d. h. die Varianz von μ_1 ist größer 0), dann bedeutet dies, dass sich die Zusammenhänge zwischen der unabhängigen und abhängigen Variable zwischen den Kontexten unterscheiden.

Die Mehrebenenanalyse ist demensprechend in der Lage, Unterschiede sowohl in den Intercepts als auch in den Slopes zu modellieren. Es ist jedoch in beiden Fällen eine empirische Frage, ob beide signifikant zwischen den Kontexten variieren, d. h. ob sich die Varianzen von μ_0 und μ_1 signifikant von null unterscheiden.

4.3.2 Die hierarchische Datenstruktur als erklärender Faktor

Der Nutzen der Mehrebenenanalyse beschränkt sich gleichwohl nicht darauf festzustellen, ob sich die Intercepts oder die Slopes zwischen Kontexten unterscheiden. Die Analyse kann zugleich zur Erklärung sozialer Prozesse genutzt werden. Zeigen Analysen, dass die Intercepts signifikant zwischen den Modellen variieren, kann versucht werden, diese Unterschiede durch Eigenschaften der Kontexte zu erklären. Im eingangs genannten Beispiel (Makro-Mikro-Hypothese) könnte etwa das Klassenklima das Viktimisierungsrisiko beeinflussen. Um dies zu überprüfen, kann eine Gruppenvariable z (die den gleichen Wert für alle Beobachtungen eines Kontexts hat) in das Modell aufgenommen werden. Diese Variable kann die Varianz der Abweichungen der Kontexte μ_0 reduzieren, d. h. erklären. Im Fall eines Random-Slope-Modells würde sich Gleichung (5) erweitern zu:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \beta_{01} z_j + \varepsilon_{ij} \qquad \text{mit } \beta_{0j} = \beta_{00} + \mu_{0j} \tag{6}$$

Variieren in einem Modell die Slopes einer (oder mehrerer) Level-1-Variablen signifikant (Random-Slope-Modell), kann auch versucht werden, diese zu erklären. Hierfür wird ein Interaktionseffekt zwischen der Level-1-Variablen mit dem Random Slope und einer Level-2-Variablen spezifiziert. Dies bedeutet, dass die Varianz des Koeffizienten der Level-1-Variable durch die Level-2-Variable erklärt wird (ausführlich Snijders/Bosker 2012; Windzio 2008).

4.3.3 Durchführung und Interpretation einer logistischen Mehrebenenanalyse

Im Folgenden soll die Durchführung und Interpretation einer Mehrebenenanalyse am Beispiel der Erklärung von Viktimisierung durch Schulgewalt verdeutlicht werden. Die Betrachtung des Schulkontexts erlaubt dabei, Merkmale der Klasse als substanzielle Einflussfaktoren zu betrachten. Für die Erklärung der Prävalenz von Viktimisierung durch Schulgewalt wird ein logistisches Mehrebenenmodell[30] verwendet. Durch die nicht lineare Linkfunktion kommen etwas andere Formeln zur Anwendung als eben für das lineare Modell dargestellt, die Mehrebenenlogik bleibt aber bestehen. Zusätzlich sind die Schwierigkeiten bei der Interpretation der Koeffizienten zu beachten, auf die in Unterkapitel 4.1 hingewiesen wurde.

Tabelle 6 zeigt die Ergebnisse der Modellschätzung.[31] In einem ersten Schritt wird das sogenannte leere Modell, das nur Random Intercepts und keine erklärenden Variablen enthält, geschätzt, um den Grad der Ähnlichkeit zwischen Individuen derselben Klasse zu bestimmen (Modell 0). Im Ergebnis zeigt sich ein Intraklassenkorrelationskoeffizient (ICC) von 0,074. Dementsprechend gehen 7,4 % der Varianz der Viktimisierung durch Schulgewalt auf die Klassen zurück. Die Varianz der Intercepts ist signifikant, wie ein Likelihood-Ratio-Test der Nullhypothese, dass die Varianz der Random Intercepts gleich null ist, zeigt.[32]

In Modell 1 wird der Einfluss der Level-1-Variablen auf das Risiko, Schulgewalt zu erleben, geschätzt. Die metrischen Variablen wurden am Gesamtmittelwert (*Grand Mean*) zentriert. Dies ist bei Mehrebenenmodellen aus zwei Gründen sinnvoll. Zum einen erhält in diesem Fall die Konstante einen sinnvollen Wert und zum anderen hängt die Varianz der Random Slopes davon ab, ob die Variablen zentriert sind oder nicht (Windzio 2008, 126–128).[33] In diesen Modellen wurde auf die Indikatoren für Alkoholkonsum und den Besuch von „Zeit pro Woche in Kneipe, Disco, Kino etc.“ verzichtet, da diese im Sinne des Routine-Activity-Ansatzes eher Prädiktoren für Viktimisierung

---

30 Ausführlicher hierzu Windzio 2008, 129–134 sowie Snijders und Bosker 2012, 290–309.

31 Alle Modelle wurden in Stata mit dem Befehl *xtmelogit* geschätzt. Es wird ein *Unit-specific*-Modell geschätzt. Für eine Diskussion der Unterschiede zwischen *Unit-specific*- und *Population-Average*-Modell siehe Raudenbush u. a. 2004, 109–11 und Neuhaus u. a. 1991.

32 Dieser Test wird im Anschluss an die Schätzung von Stata standardmäßig ausgegeben.

33 Eine andere Form der Zentrierung, die im Rahmen von Mehrebenenanalysen anzutreffen ist, ist die *Group-Mean*-Zentrierung, die das Testen sogenannter Froschteichhypothesen erlaubt Windzio 2008, 127–128.

außerhalb des Schulkontexts darstellen. In der Tabelle sind die OR dargestellt, die sich analog zur einfachen logistischen Regression interpretieren lassen. Es zeigt sich, dass Jungen ein mehr als dreimal höheres Risiko haben, Opfer von Schulgewalt zu werden, als Mädchen. Migrantinnen und Migranten haben demgegenüber ein geringeres Risiko als Einheimische. Der besuchte Schultyp hat keinen signifikanten Einfluss auf das Viktimisierungsrisiko, wenngleich eine höhere elterliche Supervision der Jugendlichen mit einem geringeren Risiko für Gewaltviktimisierung einhergeht. Elterliche Gewalt und delinquente Freundinnen und Freunde erhöhen demgegenüber das Risiko, Opfer von Schulgewalt zu werden. Der ICC sinkt durch Hinzunahme der Level-1-Variablen leicht auf 0,072 ab, sodass ein Teil der Varianz zwischen den Klassen durch eine unterschiedliche Zusammensetzung der Klassen im Hinblick auf diese Level-1 Variablen erklärt wird (Kompositionseffekt).

Modell 2 berücksichtigt zusätzlich zu den Level-1-Variablen den Einfluss von Kontextmerkmalen. Es werden das Klassenniveau an Desorganisation und Kohäsion in das Modell aufgenommen. Während eine höhere schulische Desorganisation mit einem erhöhten Viktimisierungsrisiko einhergeht, ist der Einfluss der Kohäsion nicht signifikant. Der ICC sinkt gegenüber Modell 1 leicht ab, d. h., ein Teil der Varianz zwischen den Klassen kann durch die Kontextmerkmale erklärt werden. Die Varianz des Intercepts $Var(\mu_{oj})$ ist in Modell 2 signifikant größer als null.

Im nächsten Schritt (Modell 3) wird neben dem Random Intercept ein Random Slope für den Einfluss des elterlichen Monitorings geschätzt, d. h., es wird überprüft, ob der Effekt des Monitorings der Eltern auf das Viktimisierungsrisiko zwischen den Klassen variiert. Die Idee dahinter ist, dass der Effekt elterlichen Monitorings von den Eigenschaften des Klassenkontexts abhängt. So konnte in der bisherigen Forschung (Hanslmaier 2014) gezeigt werden, dass der Einfluss familialen Sozialkapitals auf Gewalttäterschaft vom schulischen Sozialkapital abhängt. Allerdings ist zu überprüfen, ob der Zusammenhang zwischen elterlichem Monitoring und dem Viktimisierungsrisiko (also die Slopes) tatsächlich über die Klassen signifikant variiert. Wenn die Slopes signifikant variieren, dann bedeutet dies, dass die Varianz der Abweichungen der Slopes für elterliches Monitoring von der mittleren Steigung, $Var(\mu_{1j})$, signifikant ist. Dies wird mit einem Deviance-Test überprüft, der das Modell ohne Random Slope mit dem Modell mit Random Slope vergleicht.[34] Der Test zeigt ein nicht signifikantes Ergebnis. Das bedeutet, dass der Zusam-

[34] Der in diesem Fall von Stata ausgegebene Test ist konservativ. Deshalb empfiehlt es sich, den in Snijders und Bosker 2012, 98–99 beschriebenen Test, der auf der Differenz der von Stata berechneten *Deviances* der Modelle beruht, zu verwenden. Die kritischen Werte ergeben sich dabei aus einer speziellen Tabelle.

menhang zwischen elterlicher Supervision und Viktimisierung durch Schulgewalt nicht signifikant zwischen den Schulklassen variiert. In Modell 3 steigt auch der ICC gegenüber Modell 2 an.

Tabelle 6:

Mehrebenenmodell zur Erklärung der Prävalenz der Viktimisierung durch Schulgewalt (OR)[35]

	(0)	(1)	(2)	(3)	(4)
Level-1-Variablen					
männlich (1 = ja)		3,073***	3,085***	3,100***	3,079***
Migrationshintergrund (1 = ja)		0,735***	0,724***	0,724***	0,724***
Schulform					
Hauptschule					
Realschule		1,135	1,250	1,276	1,240
Gymnasium		0,956	1,232	1,253	1,228
Risikosuche		1,096*	1,096*	1,098*	1,097*
elterliches Monitoring		0,852***	0,854***	0,872**	0,841**
elterliche Gewalt					
keine Gewalt					
leichte Gewalt		1,468***	1,468***	1,470***	1,465***
schwere Gewalt		2,060***	2,036***	2,037***	2,029***
delinquente Freunde					
keine delinquenten Freunde					
1 bis 5 delinquente Freunde		1,590***	1,576***	1,574***	1,572***
6 und mehr delinquente Freunde		1,665***	1,625***	1,636***	1,626***
Level-2-Variablen					
schulische Desorganisation			1,626**	1,657***	1,666***
schulische Kohäsion			0,972	0,976	0,972
elterliches Monitoring x schulische Desorganisation					1,226+

[35] Die ICCs in der Tabelle wurden mit dem Paket *xtmrho* von Lars E. Kroll berechnet. Für die Berechnung des R^2 wurde das Paket *r2_mz* von Dirk Enzmann verwendet. Alternativ können beide Maßzahlen auch von Hand auf Basis der entsprechenden Formeln in Snijders und Bosker 2012 berechnet werden.

	(0)	(1)	(2)	(3)	(4)
Var(μ_{0j})	0,263	0,256	0,237	0,232	0,237
Var(μ_{1j})				0,039	
Cov (μ_{0}j; μ_{1j})				–0,036	
ICC	0,074	0,072	0,067	0,076	0,067
McKelvey & Zavoina's R^2	–	0,136	0,141	0,140	0,143

n = 8.411 Schüler aus 454 Klassen. *** p < 0,001, ** p < 0,01, * p < 0,05, + p < 0,10; alle metrischen Variablen sind Grand-Mean-zentriert

Eingangs wurde argumentiert, dass ein Cross-Level-Interaktionseffekt spezifiziert wird, um die Varianz des Slopes einer Level-1-Variable zu erklären. Allerdings ist es mitunter auch ratsam, von dieser Strategie abzuweichen, nämlich dann, wenn ein Cross-Level-Interaktionseffekt aus theoretischen Gründen zu erwarten ist. Dies ist insbesondere deshalb sinnvoll, da die statistische Power zur Aufdeckung einer Cross-Level-Interaktion höher als die den korrespondierenden Random Slope aufdeckende ist (wenn tatsächlich ein solcher Interaktionseffekt existiert) (Snijders/Bosker 2012, 106).

Modell 4 schätzt dementsprechend einen Cross-Level-Interaktionseffekt zwischen elterlichem Monitoring und dem Level schulischer Desorganisation. Der Koeffizient der Cross-Level-Interaktion ist marginal signifikant (p < 0,10) und hat ein OR > 1. Für die Interpretation bedeutet dies, dass in Kontexten mit einem mittleren Niveau[36] an Desorganisation ein Ansteig der elterlichen Supervision zu einem geringeren Risiko der Viktimisierung führt. In Kontexten mit einer geringeren Desorganisation führt der Anstieg elterlichen Monitorings zu einer größeren Reduzierung des Viktimisierungsrisikos. Demgegenüber ist der Effekt elterlichen Monitorings in Kontexten mit hoher Desorganisation geringer. Allerdings sollte der Interaktionseffekt aufgrund der relativ hohen Irrtumswahrscheinlichkeit eher zurückhaltend interpretiert werden. Zudem ist zu berücksichtigen, dass in logistischen Modellen bereits implizit Interaktionseffekte enthalten sind, da der Einfluss einer Variablen von den Ausprägungen der anderen Variablen abhängt. Dies erschwert den Nachweis von Interaktionseffekten (Best/Wolf 2010, 840–842; ausführlich Ai/Norton 2003). Auch ein Blick auf die Modellfitparameter (McKelvey & Zavoina's R^2) zeigt kaum Verbesserungen zwischen Modell 2 und Modell 4. Der ICC bleibt gleich, d. h., Unterschiede zwischen den Kontexten können

[36] Da alle metrischen Variablen am Gesamtmittelwert (Grand-Mean) zentriert wurden, bedeutet der Wert Null bei dieser Variable, dass ein mittleres Niveau an Desorganisation vorliegt.

auch nicht besser erklärt werden, wenn man den Interaktionsterm mit in das Modell aufnimmt. Auch das McKelvey & Zavoina's R^2 steigt nur gering an.

Für die Beispielanalyse standen 8.411 Schülerinnen und Schüler aus 454 Klassen zur Verfügung. Generell gilt im Rahmen der Mehrebenenanalyse, dass die Anzahl der Level-1-Einheiten für Schätzungen von Level-1-Effekten am wichtigsten ist und die Zahl der Level-2-Einheiten für die Schätzung von Level-2-Effekten. Die Größe der Cluster spielt keine besondere Rolle. Die Schätzung von Random Slopes hingegen ist von der Größe der Cluster abhängig (Snijders 2005). Weiterführende Diskussionen zu statistischer Power, d. h. zum Aufdecken eines tatsächlich vorhandenen Effekts, in Abhängigkeit des Studiendesigns finden sich bei Snijders und Bosker (2012).

5 Zusammenfassung

– Daten aus Opferbefragungen weisen bestimmte Eigenschaften auf, die die Anwendung einfacher linearer Regressionen (OLS-Regressionen) verhindern, da bestimmte statistische Annahmen verletzt werden. Dies betrifft die Skalierung und Verteilung der abhängigen Variable. So stellen Prävalenzen von Opfererfahrungen dichotome Merkmale dar und erfordern daher spezielle Regressionsverfahren, die diesem Umstand Rechnung tragen. Bei der Analyse von Inzidenzen von Opfererfahrungen tritt das Problem auf, dass diese in der Regel sehr rechtsschief verteilt sind. Darüber hinaus kann als Folge der Stichprobenziehung eine hierarchische Datenstruktur oder spezielles Interesse an Kontexteinflüssen vorliegen.

– Die *logistische Regression* löst das Problem der multivariaten Analyse dichotomer Merkmale. Hierfür wird eine Linkfunktion verwendet, die dazu führt, dass die Beziehung zwischen den unabhängigen Variablen und der Wahrscheinlichkeit, eine Viktimisierung zu erleben, nicht mehr linear ist. Dies erschwert die Interpretation der Koeffizienten im Vergleich zur linearen Regression.

– *Zähldatenmodelle* kommen bei der Analyse von Inzidenzen zur Anwendung. Die Poisson- und die Negativbinomialregression verwenden hierfür eine Verteilungsfunktion, die der schiefen und diskreten Verteilung der Daten angemessen ist. Wenn die Zahl der Nullen sehr groß ist, d. h. es gibt viele Personen, die überhaupt nicht Opfer geworden sind, dann kommen zero-inflated Modelle zur Anwendung. Bei diesen Verfahren werden simultan zwei Modelle geschätzt, die einerseits den Einfluss der unabhängigen Variablen auf die Wahrscheinlichkeit, zur Gruppe der Nichtopfer zu

gehören, und andererseits den Einfluss dieser Variablen auf die Häufigkeit der Opfererfahrung angeben (Windzio 2013).

– Zudem weisen Daten aus größeren Surveys oftmals eine *hierarchische Struktur* auf, d. h., die Beobachtungen auf der Individualebene sind in bestimmten Kontexten (z. B. Schulen, Nachbarschaften) geschachtelt. Dies führt zu einer Verletzung der Annahme der statistischen Unabhängigkeit von zwei Beobachtungen.

– *Mehrebenenmodelle* dienen der Analyse geclusterter Daten. Ihre Struktur (z. B. Schülerinnen und Schüler in Klassen) kann dabei entweder als Störfaktor betrachtet werden, den es zu kontrollieren gilt, um unverzerrte Ergebnisse zu erhalten, oder dazu genutzt werden, soziale Prozesse, wie etwa den Einfluss von Kontexten, aufzudecken (Windzio 2008).

6 Weiterführende Literatur

Der Aufsatz von Best und Wolf (2010) liefert eine kurze, aber umfassende Darstellung der logistischen Regression. Auch das entsprechende Kapitel bei Windzio (2013) ist zu empfehlen. Eine ausführlichere Diskussion findet sich bei Long (1997). Windzio und Long behandeln zudem in ihren Büchern die hier besprochenen Zähldatenmodelle. Im Hinblick auf die praktische Durchführung von logistischen und Zähldatenanalysen mit Stata bietet sich insbesondere das Buch von Long und Freese (2003) an, das Beispiele aufführt und den Möglichkeiten der Ergebnisdarstellung ausführlich Raum gibt. Kohler und Kreuter (2008) liefern ebenfalls eine gute Einführung in die Durchführung logistischer Regressionen mit Stata.

Einen kurzen, sehr anschaulichen Überblick über die Logik der Mehrebenenanalyse liefert Windzio (2008). Ein ausführlicheres „Standardwerk“ ist das Buch *Multilevel Analysis* von Snijders und Bosker (2012). Dort werden auch die Voraussetzungen des Modells, Strategien zur Modellbildung und verschiedene Regressionsverfahren (OLS, Logit, Zähldatenmodelle) erläutert. Das zweibändige Werk *Multilevel and Longitudinal Modeling Using Stata* von Rabe-Hesketh und Skrondal (2012) bietet eine Einführung für Stata mit zahlreichen Übungsbeispielen und konkreten Hinweisen für die Durchführung der Analysen.

8 Literatur

Ai, Chunrong; Norton, Edward C. (2003): Interaction Terms in Logit and Probit Models. Economics Letters, 80, S. 123–129.

Baier, Dirk (2015): Sicherheit und Kriminalität in Niedersachsen, Ergebnisse einer Repräsentativbefragung, KFN-Forschungsbericht Nr. 127. Hannover: KFN.

Baier, Dirk; Pfeiffer, Christian (2007): Gewalttätigkeit bei deutschen und nichtdeutschen Jugendlichen – Befunde der Schülerbefragung 2005 und Folgerungen für die Prävention, KFN-Forschungsbericht Nr. 100. Hannover: KFN.

Baier, Dirk; Pfeiffer, Christian; Simonson, Julia und Rabold, Susann (2009): Jugendliche in Deutschland als Opfer und Täter von Gewalt. Erster Forschungsbericht zum gemeinsamen Forschungsprojekt des Bundesministeriums des Innern und des KFN. Hannover: KFN.

Baier, Dirk; Prätor, Susann (im Druck): Adolescents as Victims of Violence. In: Representative Studies on Victimization. Research Findings from Germany.

Bartus, Tamás (2005): Estimation of marginal effects using margeff. In: Stata Journal, 5, S. 309–329.

Best, Henning; Wolf, Christof (2010): Logistische Regression. In: Wolf, Christof; Best, Henning (Hg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag, S. 827–854.

Cameron, Colin A.; Trivedi, Pravin K. (1998). Regression Analysis of Count Data. Cambridge: Cambridge University Press.

Cohen, Lawrence E.; Felson, Marcus (1979): Social Change and Crime Rate Trends: A Routine Activity Approach. In: American Sociological Review, 44, S. 588–608.

Enzmann, Dirk (2010): Chapter 4. Germany. In: Junger-Tas, Josine; Marshall, Ineke Haen; Enzmann, Dirk; Killias, Martin; Steketee, M. und Gruszczynska, Beata (Hg.): Juvenile Delinquency in Europe and Beyond. Results of the Second International Self-Report Delinquency Study. New York, NY: Springer New York, S. 47–64.

Gautschi, Thomas (2010): Maximum-Likelihood Schätztheorie. In: Wolf, Christof; Best, Henning (Hg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 205–235.

Gottfredson, Michael R.; Hirschi, Travis (1990): A General Theory of Crime. Stanford California: Stanford University Press.

Gottfredson, Michael R.; Hirschi, Travis (2001): Self-Control Theory. In: Paternoster, Raymond; Bachman, Ronet (Hg.): Explaining Criminals and Crime. Los Angeles, CA: Roxbury Publishing Company, S. 81–96.

Gruszczynska, Beata; Lucia, Sonia und Killias, Martin (2012): Chapter 4. Juvenile Victimization from an International Perspective. In: Junger-Tas, Josine; Marshall, Ineke Haen; Enzmann, Dirk; Killias, Martin; Steketee, Majona und Gruszczynska, Beata (Hg.): The Many Faces of Youth Crime. New York, NY: Springer New York, S. 95–116.

Hanslmaier, Michael (2014): Soziales Kapital und Jugendgewalt. Die Wechselwirkungen von Schule und Familie. In: Zeitschrift für Soziologie der Erziehung und Sozialisation, 34, S. 314–330.

Hosmer, David W.; Lemeshow, Stanley (2000): Applied Logistic Regression, 2. Auflage. New York: John Wiley.

Karlson, Kristian B.; Holm, Anders und Breen, Richard (2012): Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit: A New Method. In: Sociological Methodology, 42, S. 286–313.

Kohler, Ulrich; Kreuter, Frauke (2008): Datenanalyse mit Stata, 3. Auflage. München: Oldenbourg.

Lee, Eun S.; Forthofer, Ronald N. (2006): Analyzing Complex Survey Data, 2. Auflage. Sage: Thousand Oaks.

Long, J. Scott (1997): Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks: Sage.

Long, J. Scott; Freese, Jeremy (2003): Regression Models for Categorical Dependent Variables Using Stata. College Station, Texas: Stata Corporation.

Mood, Carina (2009): Logistic regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. In: European Sociological Review, 26, S. 67–82.

Neuhaus, J., Kalbfleisch, J. und Hauck, W. (1991): Comparison of Cluster-Specific Approaches for Population-Averaged Data Analyzing Correlated. International Statistical Review, 59, 1, S. 25–35.

Oberwittler, Dietrich (2003): Die Messung und Qualitätskontrolle kontextbezogener Befragungsdaten mithilfe der Mehrebenenanalyse – am Beispiel des Sozialkapitals von Stadtvierteln. In: ZA-Informationen, 53, S. 11–41.

Paternoster, Raymond; Bachman, Ronet (2001a): Classical and Neuve Classical Schools of Criminology. In: Paternoster, Raymond; Bachman, Ronet (Hg.): Explaining Criminals and Crime. Los Angeles, CA: Roxbury Publishing Company, S. 11–22.

Paternoster, Raymond; Bachman, Ronet (2001b): Control Theories of Crime. In: Paternoster, Raymond; Bachman, Ronet (Hg.): Explaining Criminals and Crime. Los Angeles, CA: Roxbury Publishing Company, S. 73–80.

Rabe-Hesketh, Sophia; Skrondal, Anders (2012): Multilevel and Longitudinal Modeling Using Stata, 3. Auflage. College Station, TX: Stata Press.

Raudenbush, Stephen W.; Bryk, Anthony S.; Cheong, Yuk F.; Congdon, Richard und du Toit, Mathilda (2004): HLM6: Hierarchical Linear and Nonlinear Modeling. Lincolnwood, IL: Scientific Software International.

Sampson, Robert J.; Raudenbush, Stephen W. und Earls, Felton (1997): Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. In: Science, 277, S. 918–924.

Sapouna, Maria (2010): Collective efficacy in the school context: does it help explain victimization and bullying among Greek primary and secondary school students? In: Journal of Interpersonal Violence, 25, S. 1912–1927.

Shaw, Clifford R.; McKay, Henry D. (1969): Juvenile Delinquency in Urban Areas, überarbeitete Auflage. Chicago: University of Chicago Press.

Snijders, Tom A. B. (2005): Power and Sample Size in Multilevel Modelling. In: Everitt, B. S.; Howell, D. C. (Hg.): Encyclopedia of Statistics in Behavioral Science. Band 3. Chichester, West Sussex: Wiley, S. 1570–1573.

Snijders, Tom A. B.; Bosker, Roul J. (2012): Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling, 2. Auflage. London: Sage.

StataCorp (2013): Stata Base Reference Manual. Release 13. College Station, TX: StataCorp LP.

Windzio, Michael (2008): Social Structures and Actors: The Application of Multilevel Analysis in Migration Research. In: Romanian Journal of Population Research, 2, S. 113–138.

Windzio, Michael (2013): Regressionsmodelle für Zustände und Ereignisse. Eine Einführung. Wiesbaden: VS Verlag.

Windzio, Michael; Baier, Dirk (2009): Violent Behavior of Juveniles in a Multiethnic Society: Effects of Personal Characteristics, Urban Areas, and Immigrants' Peer Networks. In: Journal of Ethnicity in Criminal Justice, 7, S. 237–270.

Designs für Viktimisierungsbefragungen und die Grundprinzipien von Kausalität

Heinz Leitgöb und Daniel Seddig

1 Einleitung

Die auf repräsentativen Viktimisierungsbefragungen begründete empirische Befundlage stellt eine wertvolle Arbeitsgrundlage sowie handlungsleitende Informationsquelle für eine Vielzahl von Interessensgruppen in den relevanten Tätigkeitsfeldern (z. B. Politik, Sicherheitsbehörden, Justiz, Wissenschaft, soziale Arbeit) dar (Cantor/Lynch 2000). Die daraus abzuleitende Bedeutsamkeit verpflichtet im Rahmen der Durchführung zur kompromisslosen Gewährleistung der Einhaltung von Qualitäts- bzw. Gütekriterien hinsichtlich der Erhebungskonzeption (z. B. Groves u. a. 2009; einschlägig: National Research Council 2008) sowie der darin enthaltenen Messkonzepte (z. B. Moosbrugger/Kelava 2008; einschlägig: Mosher u. a. 2011). Weiterhin gilt es – unter gegebenen Budgetrestriktionen – das Design so anzulegen, dass der aus Viktimisierungssurveys zu erzielende Informationsgehalt maximiert wird und so eine Vielzahl unterschiedlicher Fragestellungen geklärt werden kann.

Dem Beitrag liegt die Absicht zugrunde, zunächst die Möglichkeiten der Implementierung von Viktimisierungsbefragungen im Quer- und Längsschnitt und die daraus jeweils resultierenden Datenformate vor- bzw. einander gegenüberzustellen (Kapitel 2). Des Weiteren soll eine Einführung in die Grundprinzipien von Kausalität als Kernkonzept des quantitativ-erklärenden Forschungsparadigmas in den Sozialwissenschaften vorgelegt werden (Kapitel 3). Aus methodischer Perspektive lässt sich eine logische Anknüpfung der Kausalitätsthematik an die Ausführungen in Kapitel 2 über die Frage herstellen, welchen Kriterien viktimisierungsbezogene Daten genügen müssen, um die im Rahmen eines nicht experimentellen Settings bestmögliche Klärung kausaler Fragestellungen zu erlauben. Kapitel 4 dient der kompakten Zusammenfassung des Beitrags und das finale Kapitel 5 enthält weiterführende Literaturhinweise.

2 Designs im Quer- und Längsschnitt

Da die Qualität der Ergebnisse aus empirischen Studien bekanntermaßen nur so hoch sein kann wie jene des zugrunde liegenden Designs, gilt es auch im

Zuge der wissenschaftlichen Beantwortung kriminologischer bzw. viktimologischer Fragestellungen, die unter gegebenen Bedingungen optimale methodische Vorgehensweise zu wählen. Aus der von Toon (2000) vorgelegten Typologie longitudinaler (= längsschnittlicher) Designs sollen in der Folge jene vorgestellt werden, die auch als konzeptioneller Rahmen für Viktimisierungssurveys infrage kommen: das Trend- (Unterkapitel 2.2), das Panel- (Unterkapitel 2.3) und das retrospektive Design (Unterkapitel 2.4). An den Beginn wird zunächst jedoch die Erläuterung des Querschnittsdesigns gestellt (Unterkapitel 2.1).

2.1 Querschnittsdesign

Liegt das Erkenntnisinteresse auf der Abbildung der Kriminalitätsbelastung bzw. der Identifikation potenzieller Korrelate von Viktimisierungserfahrungen in einer interessierenden Population *N* (auch als Grundgesamtheit bezeichnet) zu einem bestimmten Zeitpunkt *t*, so erscheint die Realisierung einer Viktimisierungsbefragung im Rahmen eines Querschnittsdesigns als hinreichend. Der spezifische Charakter dieses Ansatzes manifestiert sich in der einmaligen Ziehung und Befragung einer Zufallsauswahl von *n* Elementen[1] aus der Grundgesamtheit *N* zu Zeitpunkt *t*.[2] Während sich die viktimisierungsbezogenen Messungen auf Ereignisse innerhalb eines zeitlich in der Vergangenheit verorteten Referenzzeitraums (z. B. die letzten zwölf Monate oder fünf Jahre vor der Befragung) beziehen, repräsentieren die Messungen der Begleitmerkmale (z. B. Kriminalitätsfurcht, Erfahrungen mit und Einstellungen gegenüber der Justiz bzw. Polizei; für eine Liste relevanter Dimensionen siehe United Nations 2010, 76 ff.) den jeweiligen Status der ausgewählten Personen direkt zu Zeitpunkt *t*. Folglich liegt den Daten bereits im Querschnitt eine inhärente temporäre Struktur – eine zeitliche Ordnung in den gemessenen Dimensionen – zugrunde, die auf die retrospektive Erfassung der Viktimisierungserfahrungen zurückzuführen ist und im Rahmen der Erläuterung der Kausalitätsprinzipien noch von Bedeutung sein wird.

---

1 Repräsentative Viktimisierungsbefragungen werden oftmals als Haushaltssurveys konzipiert. Im Rahmen eines mehrstufigen Verfahrens werden in einem ersten Schritt zufällig der definierten Grundgesamtheit angehörende Haushalte ausgewählt, die auch als Primäreinheiten (*primary sampling units*; PSU) bezeichnet werden. In einem zweiten Schritt erfolgt die ebenfalls zufallsbasierte Ziehung von Personen aus den PSU. Da sich die viktimisierungsbezogenen Messungen teilweise auf die Haushalte (z. B. Wohnungseinbruch) und teilweise auf die Personen (z. B. Gewaltdelikte) beziehen, wird in der Folge neutral von „Elementen" gesprochen, ausgenommen die Argumentation bezieht sich explizit auf Haushalte bzw. Personen. Zu den Details sei auf den Beitrag von Schnell und Noack in diesem Band verwiesen.

2 Für eine Einführung in die Verfahren der Zufallsauswahl sei auf Kish (1965) verwiesen.

Als aktuelles Beispiel für eine groß angelegte Viktimisierungsbefragung im Querschnitt kann der 2012 vom deutschen Bundeskriminalamt (BKA) in Kooperation mit dem Max-Planck-Institut für ausländisches und internationales Strafrecht aus Freiburg im Zuge des Projekts „Barometer Sicherheit in Deutschland“ (BaSiD) realisierte *deutsche Viktimisierungssurvey* (Birkel u. a. 2014) genannt werden.

2.2 Trenddesign

Im Rahmen eines Trenddesigns werden für eine Zielpopulation N zu T verschiedenen Zeitpunkten jeweils auf unabhängigen Zufallsstichproben basierende Querschnittsbefragungen realisiert, denen K identische Messungen bzw. das gleiche Erhebungsinstrument zugrunde liegen.[3] Ist das Zeitintervall zwischen den einzelnen Erhebungen per Design fixiert (z. B. halbjährlich, jährlich, alle zwei Jahre), wird von einem „periodischen Survey“ gesprochen (Duncan/Kalton 1987, 99). Unter Einnahme einer programmevaluativen Perspektive der Kriminal-, Sicherheits- bzw. Justizpolitik kann eine auf einem periodischen Trenddesign basierende Viktimisierungsbefragung auch als kriminalitätsbezogenes Monitoring – nach Rossi u. a. (2007, 171) allgemein definiert als „the systematic and continual documentation of key aspects of program performance that assesses whether the program is operating as intended or according to some appropriate standard“ – verstanden werden (siehe einschlägig die Ausführungen zu *Crime Monitoring* in United Nations 2010). Als prominentes Beispiel lässt sich hierfür der *Crime Survey for England and Wales* (CSEW; vormals *British Crime Survey*) anführen, der von 1982 bis 2000 alle zwei Jahre (mit den Ausnahmen 1986 und 1990) durchgeführt wurde und seit 2001 als kontinuierliche Erhebung[4] konzipiert ist (Office for National Statistics 2014).

Weiterhin existieren trenddesignbasierte Viktimisierungsbefragungen, die unregelmäßig – mit variierenden zeitlichen Abständen zwischen den einzelnen Erhebungen – wiederholt werden. Der gelegentliche Charakter ist oftmals auf

---

[3] Es gilt darauf hinzuweisen, dass die Komposition der interessierenden Population N zeitlichen Änderungen unterworfen ist, da durch den natürlichen Alterungsprozess sukzessive (alte) Personen aus N ausscheiden und neue (junge) Personen nachrücken. Dies führt zu Veränderungen in der Sozialstruktur von N (= kohortenbezogener sozialer Wandel). Folglich müsste N im Grunde mit dem Subskript t versehen werden. Da im vorliegenden Fall N allerdings abstrakt als die definierte Population an sich (z. B. die deutsche Gesamtbevölkerung ab 18 Jahren) und nicht als die konkrete Menge an Personen, die zu einem bestimmten Zeitpunkt t dieser Population angehört, verstanden wird, soll auf das Subskript t verzichtet werden.

[4] Das zentrale Kennzeichen einer kontinuierlichen Erhebung ist die Aufteilung der Stichprobe nach Referenzwochen, die sich gleichmäßig auf die Kalenderwochen des Jahres verteilen (siehe dazu einschlägig United Nations 2010, 46).

das Fehlen einer gesicherten langfristigen Finanzierung zurückzuführen. In diese Kategorie fällt etwa der *International Crime Victims Survey* (ICVS), der zwischen 1989 und 2010 insgesamt sechs Mal im Abstand zwischen drei und fünf Jahren in über 80 Ländern aus allen Regionen der Erde realisiert wurde (z. B. van Kesteren u. a. 2014).

Der zentrale Vorteil, den Trenddaten gegenüber Querschnittsdaten besitzen, liegt in der Möglichkeit zur Abbildung zeitlicher Veränderung bzw. Entwicklung auf Aggregatebene der Population (Toon 2000, 6). So lässt sich für eine interessierende Population die dunkelfeldbezogene Kriminalitätsentwicklung im Zeitverlauf darstellen und Trends wie etwa der internationale *Crime Drop* der letzten Jahrzehnte (für einen umfassenden Überblick siehe z. B. Tseloni u. a. 2010) identifizieren. Demgegenüber erlauben Trenddaten weder die Betrachtung intraindividueller Entwicklungsverläufe noch die analytische Lösung des Problems der kausalen Ordnung von Phänomenen und die Spezifikation reziproker (= wechselseitiger) Beziehungsmuster (Menard 2002, 29). Zur Klärung dieser Fragestellungen muss – wie es nachfolgend zu zeigen gilt – auf Paneldaten rekurriert werden.

2.3 Paneldesign

Ein Paneldesign lässt sich allgemein charakterisieren durch die Kombination aus der einmaligen Ziehung einer zufallsbasierten Stichprobe n zum Zeitpunkt $t = 1$ und der wiederholten Befragung[5] dieser Stichprobe zu insgesamt t (es gilt $T > 1$) verschiedenen Zeitpunkten mit dem gleichen Erhebungsinstrument bzw. zumindest K identischen Messungen.[6] Die Stichprobe der n zufällig ausgewählten Elemente verbleibt somit per Design über die gesamte Laufzeit des Panels unverändert und jedem Element i ($i = 1, \ldots, n$) kann – bei vollständiger Informationslage – zu jedem Zeitpunkt t ($t = 1, \ldots, T$) für jedes erhobene Merkmal k ($k = 1, \ldots, K$) ein Messwert (x_{ikt}) zugewiesen werden. Demgegenüber ist im Rahmen von Trenddesigns lediglich die Realisierung

---

5 Die zeitlichen Abstände zwischen den einzelnen Erhebungswellen sollten idealerweise äquidistant (= in gleich langen zeitlichen Abständen) sein und so gewählt werden, dass eine lückenlose Beobachtung von Viktimisierungsereignissen ohne die Überschneidung der Referenzzeiträume zweier Messungen möglich ist (United Nations 2010, 46 ff.).

6 Dieses Design wird auch als *Multiple-Cohort Panel Design* bezeichnet, das die simultane Verfolgung mehrerer Geburtskohorten vorsieht. Demgegenüber ist das *Single-Cohort Panel Design* durch die Verfolgung lediglich einer Kohorte gekennzeichnet. Als Beispiel kann etwa die Duisburger Panelerhebung im Rahmen des von der Deutschen Forschungsgesellschaft (DFG) geförderten Projekts „Kriminalität in der modernen Stadt“ (CRIMOC) ab der dritten Welle angeführt werden (Boers u. a. 2014). Für eine Einführung in weitere Varianten von Paneldesigns sei z. B. auf Duncan/Kalton (1987) sowie Schnell u. a. (2013, 233 ff.) verwiesen.

von x_{itk}, dem Wert des Elements i aus der Stichprobe n_t zum Zeitpunkt t im Merkmal k, möglich und in Querschnittsdesigns liegt ausschließlich x_{ik}, die Ausprägung des Elements i im Merkmal k, vor. Dies offenbart, dass Paneldesigns im Vergleich zu Querschnitts- und Trenddesigns Daten mit einem größeren Informationsgehalt hervorbringen.

Als Beispiel für eine der wenigen paneldesignbasierten Viktimisierungsbefragungen kann mit dem *National Crime and Victimization Survey* (NCVS) gleichsam der wohl bekannteste Survey dieser Art erwähnt werden. Der NCVS wurde 1973 eingeführt und repräsentiert in der gegenwärtigen Form ein rotierendes Haushaltspanel – die Panelstichprobe wird zufällig in g gleichmäßig besetzte Gruppen von Haushalten unterteilt und bei jeder Welle erfolgt die Ersetzung einer der bisherigen Gruppen durch eine neu gezogene Stichprobe, die dann wiederum g Wellen in der Stichprobe verbleibt – mit etwa 49.000 Haushalten bzw. ~100.000 Personen pro Erhebungswelle. Die Erhebungsfrequenz ist halbjährlich angelegt und die Haushalte verbleiben insgesamt jeweils sechs Wellen im Panel.

Paneldaten bilden die angemessene Datengrundlage, um die Entwicklung der Lebenslaufviktimologie (*life-course victimology*; z. B. Averdijk 2014; Farrell u. a. 2001) durch die Bereitstellung empirischer Informationen zu intraindividuellen Viktimisierungsverläufen entscheidend voranzubringen. Das Datenformat erlaubt etwa die Abbildung individueller „Age-Victimization Curves“[7] für aufeinanderfolgende Geburtskohorten, die Identifikation „typischer“ Viktimisierungsverläufe (z. B. Averdijk 2014; Farrell u. a. 2001; Lauritsen 1998) sowie – unter Anwendung angemessener Identifizierungsstrategien[8] – die (zumindest partielle) analytische Trennung viktimisierungsbezogener Alters-, Perioden- und Kohorteneffekte. Ferner lässt sich durch die dreidimensionale Struktur von Paneldaten (Elemente × Merkmale × Zeit) die empirische Prüfung theoretischer Annahmen realisieren, die aufgrund des zeitlichen Bezugs der Messungen im Querschnitts- bzw. Trenddesign nicht möglich ist. So kann unter Berücksichtigung der temporären Ordnung etwa der Effekt von Kriminalitätsfurcht (gemessen zum Zeitpunkt t_1) auf Viktimisierungserfahrungen im Zeitraum zwischen t_1 und t_2 (gemessen zu t_2) über ein Zwei-Wellen-Modell spezifiziert werden. Zugleich lassen sich die reziproken Effekte von Viktimisierungserfahrungen im Zeitraum zwischen t_0 und t_1 (gemessen zu t_1) auf Kriminalitätsfurcht (gemessen zu den Zeitpunkten t_1 und t_2) in das Modell in-

---

7 Die *Age-Victimization Curve* repräsentiert die viktimologische Entsprechung zur *Age-Crime Curve* in der täterinnen- und täterorientierten Entwicklungs- bzw. Lebenslaufkriminologie (z. B. Loeber/Farrington 2014).

8 Diesbezüglich sei auf Yang/Land (2013) verwiesen.

tegrieren. Diese Vorgehensweise erlaubt somit einerseits die Identifikation des Beziehungsgefüges und andererseits die Schätzung der Nettoeffekte – befreit von der Konfundierung durch zeitlich vorangelagerte Interdependenzen bzw. einander überlagernde Effekte – zwischen interessierenden Phänomenen (siehe diesbezüglich auch Kapitel 3).

Das beachtliche Analysepotenzial von Paneldaten ist allerdings vor dem Hintergrund damit einhergehender Probleme zu sehen, die in der Folge kurz erläutert werden sollen.

2.3.1 Sicherstellung der Äquivalenz der Messkonzepte über die Zeit

Diese notwendige Eigenschaft wiederholter Messungen[9] wird auch unter dem Begriff ‚Messinvarianz' *(measurement invariance)* diskutiert und bezeichnet (vereinfacht ausgedrückt) die Fähigkeit von Messinstrumenten bzw. Tests, die zugrunde liegenden Konstrukte[10] immer in der gleichen Qualität zu messen (z. B. Meredith 1993). Ist dies nicht der Fall, so können die zu unterschiedlichen Zeitpunkten realisierten Messungen bzw. die aus den daraus resultierenden Daten gewonnenen Parameter nicht direkt miteinander verglichen bzw. zueinander in Beziehung gesetzt werden. Während bei manifesten (direkten) Messungen wenige bis keine Möglichkeiten zur Identifikation, Berücksichtigung und somit auch Korrektur zeitbezogener Messinäquivalenzen zur Verfügung stehen, können bei über multiple Indikatoren gemessenen latenten Konstrukten im Rahmen des Ansatzes der Strukturgleichungsmodellierung (*structural equation modeling*, SEM; z. B. Bollen 1989; Reinecke 2014) die Messmodelle zum Teil entsprechend spezifiziert und somit angepasst werden (allgemein z. B. Meredith 1993; Vandenberg/Lance 2000; spezifisch für longitudinale Messungen siehe Widaman u. a. 2010). Prinzipiell bleibt allerdings festzuhalten, dass zur Maximierung der Datenqualität Messinvarianz auch in den latenten Konstrukten zwischen den einzelnen Erhebungswellen unter allen Umständen anzustreben ist.

---

9 Folglich gilt die Eigenschaft auch für Trenddesigns.

10 In der Mess- bzw. Testtheorie bezeichnet ein Konstrukt eine latente Variable, die ein nicht direkt beobachtbares Merkmal (z. B. Einstellungen, Wertorientierungen, Emotionen, psychologische und kognitive Eigenschaften von Personen) repräsentiert und folglich mittelbar über ein Set an Indikatoren (= einen Test) gemessen werden muss (z. B. Moosbrugger/Kelava 2008, 136 ff; Rost 2004, 30).

2.3.2 Panelausfälle bzw. -mortalität

Panelmortalität stellt eine spezifische Form des Ausfalls ganzer Befragungseinheiten (*Unit Nonresponse*[11]; z. B. Little/Rubin 2002) in Paneldesigns dar. Hierbei handelt es sich um den permanenten Drop-out aus der Stichprobe, etwa durch die Verweigerung der weiteren Teilnahme (mögliche Ursachen: fehlende Bereitschaft, Unterbrechung in der Teilnahmegewohnheit, Teilnahmemüdigkeit, durch lebensverändernde Ereignisse hervorgerufene Schocks; siehe Lugtig 2014) oder (umzugsbedingte) Nichterreichbarkeit. Im günstigsten aller Fälle ist der zugrunde liegende Ausfallprozess vollkommen zufälliger Natur, sodass die Ausfälle als *Missing Completely at Random* (MCAR)[12] zu bezeichnen sind. In diesem Fall führt die sukzessive Reduktion der Stichprobe lediglich zu einem über die fortlaufenden Erhebungswellen zunehmenden Genauigkeitsverlust der Parameterschätzer (z. B. Cohen 1988). Demgegenüber kann systematische bzw. selektive Panelmortalität, insbesondere im Fall von *Missing Not at Random* (NMAR; die Ausfallgründe sind auf die zu messenden Merkmale selbst und somit etwa auf Viktimisierungserlebnisse zurückzuführen) in stark verzerrten Schätzern der interessierenden Populationsparameter resultieren. Mit Blick auf die bestehende empirische Befundlage (z. B. Averdijk 2014; Ybarra/Lohr 2000) muss davon ausgegangen werden, dass in panelbasierten Viktimisierungssurveys systematische Panelmortalität in der Form vorliegt, dass „the individuals most prone to victimization also being most prone to attrition, leading to increasing bias over subsequent panel waves“ (Averdijk 2014, 266). So konnten etwa Xie/McDowall (2008) einen Effekt direkter und indirekter Gewaltviktimisierung auf die Wohnmobilität bzw. die Entscheidung zum Wohnortswechsel nachweisen. Dieser Umstand führt – insbesondere bei adressbasierten Haushaltspanels wie dem NCVS – zu systematischer Panelmortalität von Gewaltopfern.

Um Panelmortalität entgegenzuwirken werden in der Regel Strategien der Panelpflege zum Einsatz gebracht. Hierzu zählen Maßnahmen wie die sorgfältige Pflege der Kontraktadressen sowie die Adressrecherche bei verzogenen Panelteilnehmerinnen und -teilnehmern, die Bereitstellung von Informationen zu ausgewählten Ergebnissen, die Vergabe von Präsenten bzw. monetären Anreizen als extrinsische Motivation zur weiteren Teilnahme und die Administration der wiederholten Befragungen durch dieselben Interviewerinnen und Interviewer (Weischer 2015, 303). Zur Anwendungspraxis sei auf die weiterführenden Werke in Kapitel 5 verwiesen.

---

11 *Unit Nonresponse* stellt eines der zentralen Probleme in der Umfrageforschung (z. B. de Leeuw/de Heer 2002) und somit in allen vorgestellten Erhebungsdesgins dar.

12 Zur klassischen Typologie der fehlenden Werten zugrunde liegenden Ausfallprozesse siehe Rubin (1976).

2.3.3 Paneleffekte

Unter Paneleffekten (oder *Panel-Conditioning*) wird „die Veränderung der Teilnehmer durch die wiederholte Befragung“ (Schnell u. a. 2013, 233) im Rahmen des Paneldesigns verstanden (siehe weiterführend z. B. Cantwell 2008; Sturgis u. a. 2009; Warren/Halpern-Manners 2012; Watson/Wooden 2009; Waterton/Lievesley 1989). Nach Biderman/Cantor (1984, 709) ist das Auftreten von Paneleffekten in Viktimisierungssurveys auf „respondent motivational decline – ’respondent fatigue‘, loss of interest, wearing out of welcome, accumulation of burden, decay of novelty“ zurückzuführen. *Respondent Fatigue* bezeichnet in diesem Zusammenhang die – trotz gegebener Bereitschaft zur Befragungsteilnahme – auftretende „Ermüdung“ von Personen, wiederholt von erlebten Viktimisierungserlebnissen zu berichten und resultiert mit zunehmender Paneldauer in einem systematischen Anstieg von viktimisierungsbezogenem Underreporting (Averdijk 2014). Als mögliche Ursache können Lerneffekte aus vorangegangenen Erhebungswellen angeführt werden, die aus Strategien zur Vermeidung der auf die Angabe von Viktimisierungserlebnissen üblicherweise folgenden Follow-up-Fragen resultieren. Empirische Befunde auf Basis des *US National Youth Survey* (NYS) indizieren allerdings, dass die Abnahme des Berichtens von Opfererfahrungen über den Panelverlauf auch dann zu beobachten ist, wenn keine Follow-up-Fragen gestellt werden (Lauritsen 1998).

Averdijk (2014, 276) stellt vier weitere (Panel-)Effekte vor, die als methodische Alternativerklärungen für die (theoretisch unerwartete) konsistent für alle Geburtskohorten zu beobachtende Abnahme der Viktimisierungsraten über den Panelverlauf infrage kommen. *(1)* Zunächst wird in Anlehnung an Biderman/Cantor (1984) die Annahme formuliert, dass die auf die Angabe eines Viktimisierungserlebnisses folgenden detaillierten Follow-up-Fragen (insbesondere das Inzidenz-Item)[13] zu einer Anpassung der „Definition von Viktimisierung“ bei den Respondentinnen und Respondenten führen, die wiederum Einfluss auf das viktimisierungsbezogene Antwortverhalten in den nachfolgenden Erhebungswellen nimmt:

> [. . .] we find evidence that more “burdened“ respondents are more inclined to report incidents than are less burdened ones. Exposure to detailed questioning about any incident mentioned in screening, we believe may change a respondent’s definition of the interview. The incident questioning impresses on respondents the high concern of the survey with temporal and other accuracy and may reduce tendencies toward loose and expressive incident mentions in subsequent screening. (Biderman/Cantor 1984, 709)

[13] Als viktimisierungsbezogene Inzidenz wird in der Kriminologie die Häufigkeit der Opferwerdung von Elementen innerhalb eines definierten Zeitraums bezeichnet.

Der zuvor dargelegte Befund von Lauritsen (1998) spricht allerdings gegen die Existenz eines solchen Effekts. *(2)* Weiterhin erscheint es plausibel, dass – insbesondere bei viktimisierungsbezogenen Befragungen von Kindern und Jugendlichen – der „alterungsbedingte Reifeprozess" der Befragten über die Zeit einen Einfluss darauf ausübt, ob bestimmte Situationen noch länger als Viktimisierungserlebnis gedeutet und folglich auch berichtet werden. Beispielsweise könnte mit zunehmendem Alter, hervorgerufen durch die vermehrte medienbasierte Wahrnehmung der Omnipräsenz von Gewalt sowie durch eigene bzw. stellvertretende Gewalterfahrungen, die Sensitivität gegenüber interpersonaler Gewalt abnehmen. *(3)* Als typischer „Reaktivitätseffekt" kann der aufgrund der wiederholten Befragung initiierte Prozess einer verstärkten Bewusstseinsbildung bezüglich des persönlichen Viktimisierungsrisikos bezeichnet werden, der Verhaltensänderungen zur Vermeidung erneuter Opfererlebnisse bewirkt und somit die zukünftige Viktimisierungswahrscheinlichkeit reduziert. *(4)* Letztlich – und in Verbindung mit dem zuletzt genannten Effekt – ist durchaus zu erwarten, dass Opfer als allgemeine Reaktion auf ein bzw. mehrere Opfererlebnis/se ihren Alltag derart umgestalten, dass die Viktimisierungsprävention zum zentralen handlungsleitenden Prinzip erhoben wird (z. B. Averdijk 2011).

Während die unter *(3)* und *(4)* diskutierten Effekte eine tatsächliche Abnahme der Viktimisierungswahrscheinlichkeit zu späteren Messzeitpunkten implizieren, stellen die in *(1)* und *(2)* enthaltenen Argumente Erklärungsansätze für Underreporting dar.

2.3.4 Aufwand und Kosten

Abschließend gilt es noch kurz auf den mit einem Paneldesign verbundenen administrativen Aufwand und die damit einhergehenden Kosten hinzuweisen. So bedarf die Kombination aus der Umsetzung wiederholter Erhebungen, der Kontaktpflege zu den Elementen der Stichprobe zwischen den einzelnen Panelwellen und der Komplexität der Datenaufbereitung einer entsprechenden administrativen Infrastruktur und folglich auch finanzieller Mittel, die jene von Querschnitts- und Trenddesigns deutlich übersteigen.

2.4 Retrospektives Design

Eine Möglichkeit, mehreren der oben ausgeführten Problematiken bzw. Schwächen von (prospektiven) Paneldesigns entgegenzuwirken, liegt in der Implementierung eines retrospektiven Designs. Dieses lässt sich definieren als die zu einem Zeitpunkt *t* realisierte Messung von (ereignisbezogenen) Merkmalen im Querschnitt, die sich allerdings auf Zeitpunkte bzw. -räume

bezieht, die zeitlich vor *t* zu verorten sind (z. B. Viktimisierungserlebnisse in jedem Kalenderjahr von *t* – 1 bis *t* – 10). Die auf diese Weise generierten Daten werden auch als Recall-Daten (Powers u. a. 1978) bezeichnet, da sie auf spezifischen kognitiven Prozessen beruhen, deren Grundlage das Erinnerungsvermögen von Personen darstellt. Für Details sei auf das von Schwarz (1990) vorgeschlagene fünfstufige Modell der kognitiven Aufgaben im Zuge der Beantwortung retrospektiver Fragen hingewiesen.

Während im Vergleich zum (prospektiven) Paneldesign die Vorteile der Einmalmessung im Rahmen des retrospektiven Designs in der Absenz von Panelmortalität und Paneleffekten, der schnellen Datenverfügbarkeit sowie im erheblich reduzierten administrativen Aufwand und der damit einhergehenden Kostenersparnis liegen, ist der Ansatz allerdings auch mit schwerwiegenden Problemen behaftet. Zunächst muss festgehalten werden, dass der retrospektive Charakter des Designs es nicht erlaubt, dieses als Instrument für das Kriminalitätsmonitoring einzusetzen. Weiterhin ist – vor allem bei sehr langen Referenzzeiträumen – das Auftreten eines *Memory Bias* im Zuge des Abrufens der relevanten Informationen aus dem Langzeitgedächtnis ein bekanntes Phänomen, das die Validität der retrospektiven Messungen maßgeblich beeinträchtigen kann (z. B. Schwarz/Sudman 1994; einschlägig z. B. Turner 1984). In diesem Zusammenhang sind nicht nur das völlige Vergessen von Opfererlebnissen oder die kognitive Unfähigkeit zu deren Erinnerung während der Erhebungssituation, sondern auch die inkorrekte zeitliche Verortung zu erinnernder Viktimisierungserlebnisse von Relevanz. Dieses auch als *Telescoping* bezeichnete Phänomen kann in beide Richtungen auftreten (*Forward*: Ein Ereignis wird fälschlicherweise in einen bestimmten Zeitraum „teleskopiert“, obwohl es bereits früher aufgetreten ist; *Backward*: Das Ereignis wird zeitlich vor einem Zeitraum platziert, obwohl es in diesem eingetreten ist) und wurde in Viktimisierungssurveys wiederholt empirisch nachgewiesen (z. B. Woltman u. a. 1984; für eine knappe Beschreibung der Thematik sei auf Guzy/Leitgöb 2015, 105 verwiesen). Als weiteres Defizit des retrospektiven Designs sei erwähnt, dass es (beinahe) ausgeschlossen ist, nicht konkret fassbare und somit ohne Probleme wiederabrufbare Zustände wie Einstellungen, Emotionen sowie andere psychologische Faktoren valide für mehrere Zeitpunkte innerhalb eines langen Referenzzeitraums retrospektiv zu erfassen. Dieser Umstand schränkt die Möglichkeiten zur longitudinalen Kausalanalyse aufgrund der völligen Absenz bzw. der zu erwartenden geringen Messqualität eines Teils der relevanten Begleitmerkmale erheblich ein.

Als Beispiel für einen stark auf retrospektiven Elementen basierenden Viktimisierungssurvey kann der 1996 vom *Netherlands Institute for the Study of Criminality and Law Enforcement* (NISCALE) durchgeführte *Netherlands Survey on Criminality and Law Enforcement* angeführt werden. Für die insgesamt 1.939 face-to-face befragten Personen wurden auf Basis der *Life*

Event Calendar Method[14] (z. B. Roberts/Horney 2010; Sutton 2010) Informationen über die Viktimisierungs-, Familien-, Wohn-, Bildungs- und Erwerbshistorien (insgesamt 88.060 Personen-Jahre) generiert.

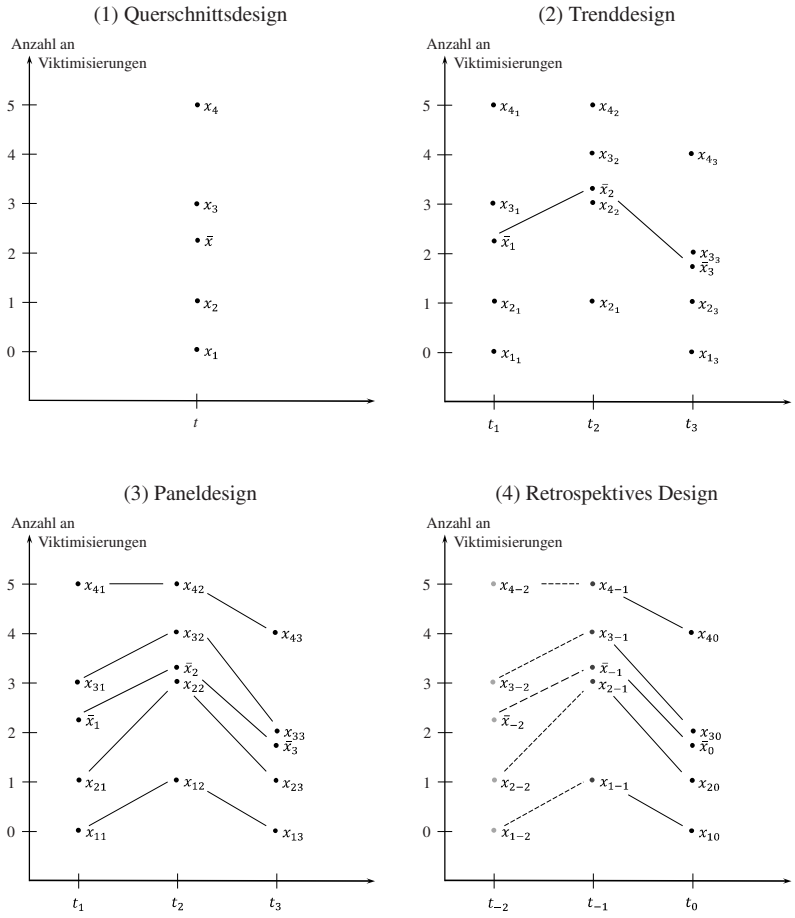
2.5 Zusammenfassende Übersicht

Eine grafische Gegenüberstellung der behandelten Designs wird abschließend in *Abbildung 1* zur Verfügung gestellt. Während bei Befragungen im Querschnitt (*Abbildung 1 (1)*) die Daten lediglich für einen Zeitpunkt t vorliegen, stehen bei Anwendung der anderen Designs Informationen zu multiplen Zeitpunkten zur Verfügung. Das Trenddesign (*Abbildung 1 (2)*), charakterisiert durch wiederholte Befragung einer jeweils anderen Stichprobe (x_{it} repräsentiert die zum Erhebungszeitpunkt t gemessene Viktimisierungshäufigkeit der Person i im Viktimisierungsindikator x), erlaubt den zeitbezogenen Vergleich auf Aggregatebene (z. B. von Mittelwerten, Anteilen, Streuungen). Ferner ermöglichen auf dem (prospektiven) Paneldesign (*Abbildung 1 (3)*) sowie dem retrospektiven Design (*Abbildung 1 (4)*) basierende Daten die Abbildung der intraindividuellen Entwicklung im Zeitverlauf. Die gestrichelten Linien im retrospektiven Design sollen die mit den Messungen der Merkmale für bereits länger zurückliegende Zeitpunkte bzw. Referenzzeiträume ($t - 1$, $t - 2$, ...) verbundene Abnahme an Validität illustrieren.

[14] Unter der Life-Event-Calendar-Methode werden Datenerhebungsansätze subsumiert, die der Sammlung von Informationen über den Zeitpunkt des Eintritts interessierender Ereignisse mittels Kalender dienen.

Abbildung 1:

Designs im Vergleich[15]



[15] *Abbildung 1* ist angelehnt an die Abbildungen 3.1 bis 3.3 aus Reinecke (2005) sowie die Abbildungen 1 und 2 aus Wittenberg (2015).

3 Die Grundprinzipien von Kausalität

Gemäß einer Einschätzung von Opp (2010, 9) hat das Schrifttum, das sich mit Kausalität – dem Ursache-Wirkungs-Gefüge zwischen sozialen Phänomenen – auseinandersetzt, ein nicht länger überschaubares Ausmaß erreicht. Diese Entwicklung ist nicht zuletzt dem Umstand geschuldet, dass Fragen nach den konstituierenden Ursachen interessierender Phänomene in den (Sozial-)Wissenschaften eine herausragende Bedeutung zukommt: „Most quantitative empirical analyses are motivated by the desire to estimate the causal effect of an independent variable on a dependent variable" (Winship/Morgan 1999, 659). Dennoch ist es bislang nicht in hinreichendem Maße gelungen, eine konsensuale allgemeine Definition des Kausalitätsbegriffs sowie ein generalisiertes formales (mathematisch-statistisches) Framework zur Identifikation kausaler Effekte auf der Grundlage „nicht experimenteller Daten"[16] vorzulegen. Aus diesem Grund werden in Anlehnung an Goldthorpe (2001) drei unterschiedliche Konzepte von Kausalität – als robuste Abhängigkeit (Unterkapitel 3.1), als konsequente Manipulation (Unterkapitel 3.2) und als generativer Prozess (Unterkapitel 3.3) – vorgestellt.

3.1 Kausalität als robuste Abhängigkeit

Das Kernargument dieses Kausalitätskonzepts beruht auf der Annahme, dass ein Merkmal *X* genau dann eine genuine Ursache eines anderen Merkmals *Y* repräsentiert, wenn die Abhängigkeit des Merkmals *Y* vom Merkmal *X* robust ist. In Anlehnung an Blalock (1964; siehe auch Menard 2002; Toon 2000) müssen hierfür die folgenden vier Bedingungen erfüllt sein:

(I) Kovariation: Das Vorliegen einer Kovariation – allgemeiner: einer Assoziation – zwischen *X* und *Y* ist eine notwendige, keinesfalls jedoch eine hinreichende Bedingung für eine Kausalbeziehung zwischen den beiden Merk-

[16] Unter dem Begriff ‚nicht experimentelle Daten' werden jene Datenmassen subsumiert, deren Generierungsprozess nicht den Kriterien experimenteller Designs (insbesondere des *Randomized Control Trials* (RCT); z. B. Shadish u. a. 2002) und somit dem Goldstandard für den Nachweis kausaler Effekte genügt (z. B. Rubin 2007; einschlägig Welsh u. a. 2013). Cook (2002, 275) verweist in diesem Zusammenhang auf „the well-nigh universal acknowledgement that experiments provide the best justification for causal conclusions" (für eine einschlägige kritische Perspektive siehe z. B. Sampson 2010). Viktimisierungsbefragungen sowie sämtliche anderen viktimisierungsbezogenen Datenquellen fallen in die Kategorie der nicht experimentellen Daten, da eine zufallsbedingte Zuweisung von Personen in eine Experimentalgruppe, die einem wie auch immer ausgestalteten realen Viktimisierungsszenario ausgesetzt wird, schon aus forschungsethischen und rechtlichen Gründen nicht legitimiert werden kann.

malen.[17] Dies wird in Anlehnung an Barnard (1982, 387) in zahlreichen einschlägigen Beiträgen mit der Aussage „Correlation is not causation!“ auf den Punkt gebracht. Aus statistischer Perspektive legt Pearl eine klare Abgrenzung zwischen den beiden Konzepten vor:

> An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. (Pearl 2010, 79)

Die Assoziation zwischen den beiden Merkmalen muss allerdings nicht linearer Natur sein.

(II) Beständigkeit des Effekts unter Kontrolle von Drittvariablen (non-spuriousness): Diese Bedingung repräsentiert den Kern des Konzepts der „robusten Abhängigkeit“ und impliziert, dass die Beziehung zwischen den Merkmalen *X* und *Y* auch unter der statistischen Kontrolle aller möglichen Drittvariablen *Z* Bestand haben muss. Es gilt somit auszuschließen, dass es sich bei dem Effekt von *X* auf *Y* um einen auf Alternativerklärungen zurückzuführenden „Scheineffekt“ handelt. Dieser Fall liegt vor, wenn ein Merkmal *Z* – auch als konfundierende Variable (*confounder*) bezeichnet – eine gemeinsame Ursache (*common cause*) für Variation in *X* und *Y* repräsentiert (*Abbildung 2 (2)*). Von einer partiellen Konfundierung ist die Rede, wenn sich der Effekt von *X* auf *Y* nach der Kontrolle von *Z* (substanziell) reduziert, aber nicht vollständig verschwindet (*Abbildung 2 (3)*).

Weiterhin kann ein bivariat existenter Effekt (*Abbildung 2 (1)*) ausschließlich darauf zurückzuführen sein, dass ein drittes Merkmal *Z* den Effekt von *X* auf *Y* vermittelt (*Abbildung 2 (4)*; z. B. Iacobucci 2008). In diesem Fall wird von einem indirekten bzw. von *Z* mediierten kausalen Effekt von *X* auf *Y* gesprochen, dessen Stärke sich aus dem Produkt der Effekte von *X* auf *Z* und weiter von *Z* auf *Y* ergibt. Bleibt unter Berücksichtigung von *Z* ein direkter Effekt bestehen (*Abbildung 2 (5)*), so setzt sich der kausale Gesamteffekt von *X* auf *Y* additiv aus den direkten und indirekten (über *Z* vermittelten) Komponenten zusammen.

Da keine analytische Möglichkeit zur Klärung besteht, ob *Z* eine konfundierende oder mediierende Wirkung auf den Effekt von *X* auf *Y* ausübt (das Common-Cause-Modell und das Mediationsmodell besitzen die gleiche Anzahl an Modellparametern und sind den empirischen Daten gleich gut angepasst), müssen hierfür theoretische Argumente herangezogen werden. Eine Differen-

---

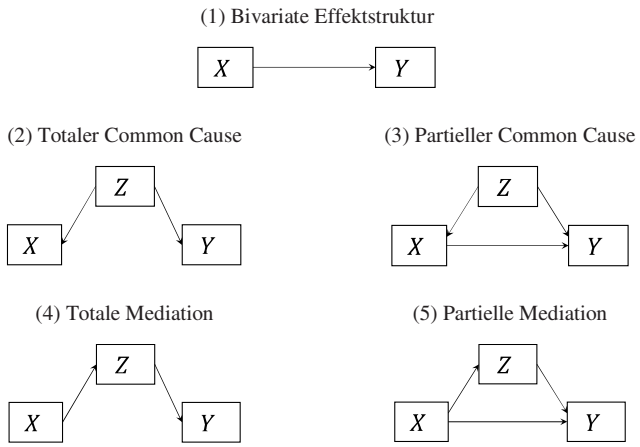
[17] Eine Ausnahme stellt das Phänomen der „Suppression“ dar. So ist es prinzipiell möglich, dass sich ein kausaler Effekt von *X* auf *Y* erst nach der statistischen Kontrolle einer Drittvariable *Z* manifestiert, da dieser durch die Auspartialisierung der Varianzanteile von *Z* aus *X* freigelegt wird (z. B. Urban/Mayerl 2011, 94 f.).

zierung der beiden Situationen ist jedenfalls von zentraler Bedeutung, da unter Geltung eines Common-Cause-Beziehungsgefüges gemäß *Abbildung 2 (2)* kein kausaler Effekt von *X* auf *Y* vorliegt, während unter Geltung des Mediationsmodells gemäß *Abbildung 2 (4)* *X* zumindest einen indirekten kausalen Einfluss auf *Y* ausübt.

Konkludierend kann somit festgehalten werden, dass die statistische Kontrolle aller konfundierenden Merkmale Z_k zur Identifikation des kausalen Bruttoeffekts (= Summe des direkten und aller indirekten Effekte) und die weitere Kontrolle aller mediierenden Merkmale Z_m zur Identifikation des direkten kausalen Effekts von *X* auf *Y* führt. Die Schwierigkeit des empirischen Nachweises kausaler Effekte im Forschungsalltag liegt nun darin, Kenntnis über die den Sets Z_k und Z_m angehörenden Merkmale zu erlangen und diese valide zu messen.

Abbildung 2:

Common Cause und Mediation[18]



Zur Verdeutlichung sollen die bisherigen Ausführungen am Beispiel des Effekts sexueller Viktimisierungserlebnisse (V_s) auf die deliktspezifische Krimi-

18 Die Abbildung enthält mit dem Common-Cause- und dem Mediations-Modell lediglich zwei mögliche Effektstrukturen zwischen den drei Merkmalen. Für die Darstellung weiterer potenzieller Drei-Variablen-Modelle siehe etwa Cohen u. a. (2003, 458) sowie Reinecke (2014, 51). Für die Diskussion von Moderationseffekten sowie die Kombination von Moderations- und Mediationseffekten sei aktuell auf Hayes (2013) verwiesen.

nalitätsfurcht (F_s) durchdekliniert werden.[19] So könnte eine substanzielle Reduktion (bzw. das völlige Verschwinden) eines Effekts von V_s auf F_s mit der Berücksichtigung des konfundierenden Merkmals Geschlecht (G) verbunden sein: (1) Frauen sind im Vergleich zu Männern einem höheren sexualdeliktspezifischen Viktimisierungsrisiko ausgesetzt (Effekt von G auf V_s; aktuell z. B. Breiding u. a. 2014) und (2) weisen – unabhängig von einschlägigen Viktimisierungserfahrungen – ein höheres Ausmaß an sexualdeliktspezifischer Kriminalitätsfurcht auf (z. B. May 2001), etwa weil sie sich dessen bewusst sind, einer besonderen Risikogruppe anzugehören (Effekt von G auf F_s). Demgegenüber könnte dem Merkmal des perzipierten Risikos einer zukünftigen einschlägigen Opferwerdung (R_s) die Rolle eines totalen Mediators zwischen V_s und F_s zukommen (z. B. Boers 2003; Hirtenlehner/Meško 2011), wie von Hirtenlehner und Farrall treffend in der Verdichtung eines weit verbreiteten Befunds der Kriminalitätsfurchtforschung zum Ausdruck gebracht wird:

> Personal victimization increases the perceived likelihood of future victimization (especially for the same offense), which then elevates fear of crime. As soon as differences in risk assessment are taken into account, no direct fear-enhancing effect of prior victimization is left. (Hirtenlehner/Farrall 2014, 13 f.)

Abschließend muss noch angemerkt werden, dass die Bedingung (II) entgegen der Behauptung von Menard (2002, 15) nicht abschließend auf der Basis von Querschnittdaten geprüft werden kann. Dies ist insbesondere dann nicht möglich, wenn simultan reziproke Effekte und zeitbezogene Stabilitätseffekte auftreten (*Abbildung 3 (1)*) und somit Merkmal Y zum Zeitpunkt $t-1$ eine gemeinsame Ursache für Variation in X und Y zum Zeitpunkt t darstellt. Die im Querschnitt fehlende Beobachtung von Y_{t-1} führt in diesem Fall auch dann zu einer verzerrten Schätzung des direkten kausalen Effekts von X auf Y zum Zeitpunkt t, wenn theoretisch alle anderen konfundierenden und mediierenden Drittvariablen (Z_k und Z_m) kontrolliert werden. Diese Problematik lässt sich somit nur über den Rückgriff auf Paneldaten lösen.

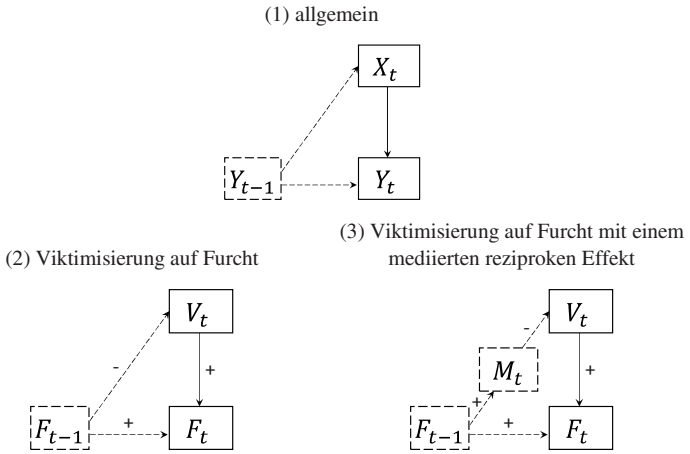
So wird etwa der kausale Effekt eines Viktimisierungserlebnisses im Referenzzeitraum zwischen $t-1$ und t (gemessen zum Zeitpunkt t) auf das Ausmaß an Kriminalitätsfurcht zum Zeitpunkt t unterschätzt, wenn ein positiver Stabilitätseffekt für Kriminalitätsfurcht und ein negativer reziproker Effekt von Kriminalitätsfurcht zu $t-1$ auf die Wahrscheinlichkeit einer Opferwerdung zwischen $t-1$ und t vorliegen (*Abbildung 3 (2)*). Der letztgenannte negative Effekt könnte auf einen totalen Mediationseffekt zurückzuführen sein

[19] Es gilt anzumerken, dass den angeführten Beispielen nicht der Anspruch zugrunde liegt, in vollkommener Kongruenz mit der empirischen Befundlage zu stehen. Vielmehr dienen sie der plastischen Darstellung der behandelten kausalen Effektstrukturen.

(*Abbildung 3 (3)*): Ein hohes Ausmaß an kriminalitätsbezogener Furcht führt zur Entwicklung von Vermeideverhaltensstrategien bzw. zur Übernahme protektiver Handlungsmuster im Alltag (M_t; diese werden analog zu V_t zum Zeitpunkt t gemessen, beziehen sich allerdings auf den Zeitraum zwischen $t-1$ und t), die wiederum die Wahrscheinlichkeit zukünftiger Viktimisierungen reduzieren (z. B. Boers 1991).

Abbildung 3:

Durch einen reziproken Effekt und einen Stabilitätseffekt konfundierter Querschnittseffekt (gestrichelte Linien indizieren nicht beobachtete Merkmale bzw. Effekte)



(III) Temporäre Ordnung von Ursache und Wirkung: Das Eintreten einer Wirkung als Konsequenz aus einer Ursache muss dieser zwangläufig zeitlich zumindest infinitesimal nachgelagert sein. Eine Wirkung kann niemals vor deren Ursache eintreten. Während bislang – in Übereinstimmung mit der in den Sozialwissenschaften vorherrschenden Auffassung von Kausalität – implizit davon ausgegangen wurde, dass Merkmale die konstituierenden Entitäten von Ursachen und Wirkungen repräsentieren (z. B. $X \rightarrow Y$), gilt es, an dieser Stelle auf die Feststellung Hedströms (2005, 14) hinzuweisen, dass „most philosophers of science insist that causes and effects must be events“ (siehe diesbezüglich auch die in Sosa/Tooley 1993 gesammelten Beiträge). Ganz im Sinne dieser Tradition definiert Lewis in seinem richtungsweisenden Artikel Kausalität wie folgt:

> If *c* and *e* are two actual events such that *e* would not have occurred without *c*, than *c* is a cause of *e*. (Lewis 1973, 563)

Sei ein Ereignis definiert als die Zustandsänderung eines Elements in einem Merkmal (Blossfeld/Rohwer 1997, 361), dann manifestiert sich ein kausaler Effekt gemäß der Ereignislogik in der Zustandsänderung in einem Merkmal *Y*, die ursächlich auf eine (zeitlich vorangegangene) Zustandsänderung im Merkmal *X* zurückzuführen ist. Der Kern dieses Kausalitätsverständnisses kann knapp auf den Punkt gebracht werden: Veränderung ist die Ursache für zeitlich nachgelagerte Veränderung, die sich allerdings nicht zwangsläufig einstellt, sondern lediglich mit einer bestimmten Wahrscheinlichkeit.[20]

Als inhaltsbezogenes Beispiel soll wiederholt der Effekt einer erlebten Viktimisierung auf Kriminalitätsfurcht dienen. Aus der ereignisorientierten Perspektive liegt ein kausaler Effekt vor, wenn ein Viktimisierungserlebnis die Ursache für eine darauffolgende Erhöhung der Eintrittswahrscheinlichkeit einer positiven Änderung des Kriminalitätsfurchtniveaus (= Zunahme der Furcht) von Personen repräsentiert.

Hinsichtlich der Bedingung der korrekten temporären Anordnung von Ursache und Wirkung lassen sich die folgenden Implikationen für die Konzeption von Designs für Viktimisierungsbefragungen ableiten: Aus merkmalsorientierter Perspektive ist Bedingung (III) für die Abbildung eines kausalen Effekts von Viktimisierung auf Furcht im Rahmen des Querschnittsdesigns erfüllt, da die viktimisierungsbezogene Messung zwar zum Erhebungszeitpunkt t realisiert wird, sich jedoch auf das in der Vergangenheit gelegene Zeitintervall $[t-1, t]$ bezieht. Somit werden zwangsläufig Viktimisierungserfahrungen erfasst, die dem Zeitpunkt t zeitlich vorgelagert und – bei Bestehen eines kausalen Effekts – ursächlich für (zumindest einen Teil der) Streuung im Merkmal Kriminalitätsfurcht zum Zeitpunkt t verantwortlich sind. Im Gegensatz dazu bedarf es zur empirischen Prüfung kausaler Effekte nach der Ereignislogik eines Paneldesigns, da das Ereignis einer Änderung des Kriminalitätsfurchtniveaus als Reaktion auf eine Viktimisierung nicht valide über eine Indikatorvariable gemessen werden kann, sondern sich in der Differenz aus wiederholten Messungen abbildet.[21]

(IV) Theoretische Begründbarkeit: Die vierte Bedingung für das Bestehen eines kausalen Effekts von *X* auf *Y* gemäß dem Prinzip der robusten Abhängig-

---

20 Wie etwa Blossfeld/Rohwer (1997, 369) anmerken, erscheint die probabilistische Formulierung kausaler Annahmen in den Sozialwissenschaften angemessener als der deterministische Ansatz, der auf dem Prinzip der deduktiv-nomologischen Erklärung nach dem Hempel-Oppenheim Schema (Hempel/Oppenheim 1948) fundiert.

21 Im Idealfall wird das einer Person inhärente Ausmaß an Kriminalitätsfurcht ganz knapp vor einer Viktimisierungserfahrung mit jenem kurz nach dem Ereigniseintritt kontrastiert, um eine Änderung (z. B. einen Anstieg der Furcht) tatsächlich auf die Opferwerdung zurückführen zu können. Zu diesem Zweck ist es erforderlich, möglichst kurze Zeiträume zwischen den Erhebungswellen zu planen.

keit ist nicht von statistischer Natur und wird aus diesem Grund oftmals nicht explizit angeführt. Sie besagt, dass die kausale Interpretation empirisch identifizierter Effekte immer einer theoretischen Fundierung bzw. einer Rechtfertigung auf der Grundlage theoretischer Argumente bedarf. So stellt Hedström fest:

> If it proves impossible to specify how the phenomenon to be explained could have been generated by the actions of individuals, or if the account must be based on highly implausible assumptions, one's faith in the proposed causal account is sharply reduced. (Hedström 2005, 29)

Die Bedeutsamkeit dieser vierten Bedingung kann durch eine knappe Auswahl an weiteren Aussagen verdeutlicht werden:

> Causal inference is theoretically driven; causal statements need a theoretical argument specifying how the variables affect each other in a particular setting across time [...]. Thus, a causal process cannot be demonstrated directly from the data; the data can only present relevant empirical evidence serving as a link in a chain of reasoning about causal mechanisms. (Toon 2000, 4)

> Blind empiricism unguided by a theoretical framework for interpreting facts leads nowhere. (Heckman 2005, 5)

Solange jene Mechanismen, die einen empirisch identifizierten Effekt hervorgebracht haben, einer plausiblen theoretischen Grundlage entbehren, darf selbst dann nicht von einem kausalen Effekt ausgegangen werden, wenn der hypothetische Fall der Gültigkeit der Bedingungen (I) bis (III) vorliegt.

Aus analytischer Sicht entspricht die der Logik von Kausalität als robuster Abhängigkeit angemessene Identifikationsstrategie kausaler Effekte dem Prinzip der statistischen Kontrolle von Drittvariableneffekten aus dem regressionsanalytischen Ansatz. Dies trifft in besonderem Maße auf das *Fixed Effects Model* (z. B. Allison 2009) zu, das die Kontrolle sämtlicher zeitinvarianter (= sich über die Zeit nicht verändernder) Merkmale erlaubt – und zwar unabhängig davon, ob diese explizit im Modell berücksichtigt werden oder nicht.

3.2 Kausalität als konsequente Manipulation

Die Auffassung von Kausalität als konsequente Manipulation soll trotz bzw. gerade wegen ihres bislang äußerst mäßigen Verbreitungsgrads in der viktimologischen Forschung (ganz im Gegensatz zur beständig wachsenden Popularität des Ansatzes in den Sozialwissenschaften im Allgemeinen; z. B. Morgan/Winship 2007) in der gebotenen Kürze dargestellt werden. Das Fundament des in der statistischen Literatur auch als „kontrafaktischer Ansatz" firmierenden Kausalitätskonzepts geht auf die Arbeiten von Neyman

(1990 [1923]) und Rubin (1974) zurück.[22] Aus diesem Grund wird oftmals vom *Neyman-Rubin Model of Causal Inference* bzw. vom *Potential Outcomes Model* gesprochen.

Die Ausgangslage, von Holland (1986, 947) auch als „fundamental problem of causal inference" bezeichnet, wird von Morgan und Winship treffend auf den Punkt gebracht:

> The key assumption of the counterfactual framework is that each individual in the population of interest has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time. (Morgan/Winship 2007, 5)

Übertragen auf das bereits vertraute Beispiel mit dem Effekt einer Viktimisierung auf das Ausmaß an Kriminalitätsfurcht kann der Ansatz wie folgt verdeutlicht werden: Es ist unmöglich, dass eine Person *i* im Zeitraum zwischen $t-1$ und *t* SOWOHL keinem ($V_{it} = 0$) ALS AUCH zumindest einem ($V_{it} = 1$) Viktimisierungsereignis ausgesetzt war. Daraus folgt, dass nicht simultan für beide Möglichkeiten das jeweils daraus resultierende Kriminalitätsfurchtniveau (= Outcome) beobachtet werden kann. Vielmehr liegt für jede Person *i* nur ein beobachteter Outcome vor (F_{it}^0 falls $V_{it} = 0$; F_{it}^1 falls $V_{it} = 1$). Die beiden nicht beobachtbaren Outcomes – auch als *Counterfactuals* bezeichnet – sind nun

(1) das Ausmaß an Kriminalitätsfurcht, falls Person *i* viktimsiert worden wäre, obwohl sie tatsächlich nicht viktimisiert wurde (F_{it}^1 falls $V_{it} = 0$) und

(2) das Ausmaß an Kriminalitätsfurcht, falls Person *i* nicht viktimisiert worden wäre, obwohl sie tatsächlich viktimisiert wurde ((F_{it}^0 falls $V_{it} = 1$).

Eine Zusammenfassung der Ausführungen ist *Tabelle 1* zu entnehmen.[23]

Tabelle 1:

Beobachtete und nicht beobachtbare Outcomes

	Kriminalitätsfurcht	
Viktimisierung	F_{it}^0	F_{it}^1
nein ($V_{it} = 0$)	beobachtet	counterfactual
ja ($V_{it} = 1$)	counterfactual	beobachtet

22 Für eine detaillierte Darstellung der Genese des Ansatzes sei auf Morgan/Winship (2007, 4 ff.) sowie Barringer u. a. (2013) verwiesen.

23 *Tabelle 1* ist an Table 2.1 aus Morgan/Winship (2007, 35) angelehnt.

Unter der hypothetischen Annahme, dass die simultane Beobachtung von F_{it}^0 und F_{it}^1 möglich ist, könnte der kausale Effekt eines Viktimisierungserlebnisses auf die kriminalitätsbezogene Furcht einer Person *i* zum Zeitpunkt *t* als Differenz der Furchtniveaus (ΔF_{it}) *der beiden möglichen Szenarien – keine Viktimisierung (V_{it} = 0) bzw. zumindest eine Viktimisierung (V_{it} = 1) im Zeitraum zwischen t* – 1 und *t* – definiert werden:

$$\Delta F_{it} = F_{it}^1 - F_{it}^0 \tag{1}$$

Im Fall von $\Delta F_{it} > 0$ *wäre nun das Ausmaß an Kriminalitätsfurcht der Person i* zum Zeitpunkt *t* größer, wenn sie zumindest einmal Opfer einer Straftat geworden wäre, verglichen mit der komplementären Situation, dass sie keinem Viktimisierungserlebnis ausgesetzt gewesen wäre. Daraus lässt sich die kausale Schlussfolgerung ableiten, dass (zumindest für die Person *i* im beobachteten Zeitraum) ein Viktimisierungserlebnis kriminalitätsfurchterhöhend wirkt.

Da kausale Inferenzen sich in aller Regel nicht auf einzelne Personen beziehen, sondern einen höheren Grad an Verallgemeinerung besitzen sollten, kann Gleichung (1) durch die Verwendung von Erwartungswerten auch für eine beliebige Population *N* generalisiert werden.

$$ATE = E(\Delta F_t) = E(F_t^1 - F_t^0) \tag{2a}$$

$$= E(F_t^1) - E(F_t^0) \tag{2b}$$ [24]

Der kausale Effekt – auch als *Average Treatment Effect* (ATE) bezeichnet[25] – wird über $E(\Delta F_t)$ abgebildet und entspricht dem Erwartungswert der Differenz der beiden potenziellen Outcomes (Gleichung (2a)) bzw. der Differenz der beiden Erwartungswerte (Gleichung (2b)) in der Population. Anders als in Gleichung (1) enthalten (2a) und (2b) nicht länger das personenbezogene Subskript *i*. Daraus folgt eine zentrale statistische Konsequenz: Die zuvor getroffene hypothetische Annahme der simultanen Beobachtung der beiden potenziellen Outcomes F_{it}^0 und F_{it}^1 für jede Person *i* muss nicht länger auf-

---

24 Die Reformulierung von Gleichung (2a) in der Form von (2b) ist darauf zurückzuführen, dass der Erwartungswert einer Differenz der Differenz der beiden Erwartungswerte entspricht.

25 Wie F_t^0 und F_t^1 repräsentiert auch ΔF_t *eine Zufallsvariable, sodass* $E(\Delta F_t)$ dem erwarteten kausalen Effekt einer zufällig aus der Population gezogenen Person *i* entspricht. Dies bedeutet nicht, dass ΔF_t *für jede Person i* den gleichen Wert aufweisen und somit konstant sein muss (Winship/Morgan 2007, 36). Für eine Einführung in die dem ATE zugrunde liegenden Annahmen (z. B. *Stable Unit Treatment Value Assumption*, SUTVA – die Unabhängigkeitsannahme) siehe Winship/Morgan (2007, 31 ff.). Weiterhin sei erwähnt, dass neben dem ATE noch eine Reihe weiterer äußerst bedeutsamer kausaler Effekte existiert: z. B. der *Average Treatment Effect for the Treated* (ATT), der *Average Treatment Effect for the Untreated* (ATUT) sowie der *Local Average Treatment Effect* (LATE).

rechterhalten werden, da die Kenntnis der beiden Erwartungswerte $E(F_t^1)$ und $E(F_t^0)$ für die Identifikation des kausalen Effekts $E(\Delta F_t)$ hinreichend ist. Folglich gilt es, über ein entsprechendes Design zu gewährleisten, dass $E(F_t^1)$ und $E(F_t^0)$ unverzerrt geschätzt werden. Unter method(olog)ischen Gesichtspunkten würde sich hierfür insbesondere der *Randomized Controlled Trial* (RCT) als geeignet erweisen. Aus den bereits in Fußnote 14 erläuterten Gründen muss die Anwendung experimenteller Designs zur Klärung viktimisierungsbezogener kausaler Fragestellungen jedoch kategorisch ausgeschlossen werden.

Als Alternativen zum RCT kommen in erster Linie jene surveybasierten Designs infrage, die in Kapitel 2 vorgestellt wurden, wenngleich mit jenen Problemen behaftet, die nicht experimentelle Daten im Rahmen der Identifikation kausaler Effekte allgemein mit sich bringen. Treten Viktimisierungsereignisse in einer interessierenden Population nicht vollkommen zufällig auf, sondern setzen sich spezifische Gruppen von Elementen hinsichtlich ihres Viktimisierungsrisikos von der Masse ab, wird analytisch von einer systematischen Selektion in den Opferstatus gesprochen. Repräsentieren in weiterer Folge jene Merkmale, die für die Selektion in den Opferstatus verantwortlich sind, zugleich auch Ursachen von Kriminalitätsfurcht, kann der kausale Effekt eines Viktimisierungserlebnisses nicht länger unverzerrt über die naive Differenz der Schätzer der Erwartungswerte $E(F_t^1)$ und $E(F_t^0)$ gemäß Gleichung (2b) abgebildet werden.

Übertragen auf das Beispiel zur Sexualviktimisierung lässt sich dies wie folgt verdeutlichen: Unter der Annahme, dass Frauen – *ceteris paribus* – ein höheres Ausmaß an Kriminalitätsfurcht aufweisen als Männer, führt das deutlich gehobene sexualdeliktspezifische Viktimisierungsrisiko von Frauen zu einer weiblichen Überrepräsentanz im Opferstatus V_{st}^1 (bei simultaner Unterrepräsentanz in V_{st}^0), die sich wiederum positiv in $\bar{F}_{st}^1$ niederschlägt und so in der systematischen Überschätzung von $E(F_{st}^1)$ und letztlich auch des kausalen Effekts $E(\Delta F_{st})$ resultiert.

Lösungsansätze für die Problematik dieser Form systematischer Selektion entstammen unterschiedlichen Traditionen. Während die statistischen Ansätze verstärkt auf Matchingverfahren fokussieren (z. B. Rubin 1974; Rosenbaum 2002; für eine umfassende Einführung siehe Guo/Fraser 2010), zielt die ökonometrische Perspektive auf die explizite Berücksichtigung des Selektionsprozesses in parametrischen Modellen (z. B. Breen 1996; Heckman 1979, 2005) und die Verwendung von Instrumentalvariablen (z. B. Angrist u. a. 1996) ab.

3.3 Kausalität als generativer Prozess

Während die beiden zuvor behandelten Ansätze in erster Linie auf die Identifikation der *Effects of the Causes* abzielen, liegt der Fokus des Prinzips von Kausalität als generativer Prozess auf den *Causes of Effects* und somit auf jenen Mechanismen, die das Ursache-Wirkungs-Gefüge zwischen sozialen Phänomenen konstituieren (siehe dazu Hedström/Swedberg 1998b, 1; Opp 2004, 364).[26] Dementsprechend rückt die statistische Perspektive, die in den ersten beiden Kausalitätskonzepten eine dominierende Stellung eingenommen hat, zugunsten einer theoretisch-analytisch ausgerichteten Perspektive in den Hintergrund. Ganz in diesem Sinne hält Hedström (2005) in „Dissecting the Social", seinem Manifest zur analytischen Soziologie, unter Bezugnahme auf Boudon (1976, 117) fest, dass die wissenschaftlichen Bemühungen darauf konzentriert werden sollten, jene generativen Prozesse aufzudecken, die für die mittels statistischer Verfahren identifizierten Zusammenhänge zwischen sozialen Phänomenen verantwortlich zeichnen. Dies impliziert die Erhellung jener Blackbox, die beispielsweise im Rahmen des zuvor erörterten Prinzips von Kausalität als konsequente Manipulation durch die exklusive Fokussierung auf die Differenz der potenziellen Outcomes verbleibt, und soll noch weit darüber hinausgehen (Hedström/Swedberg, 1998b, 9 ff.).[27] So begreift Little die analytische Substanz sozialwissenschaftlicher Erklärungen wie folgt:

> „I maintain that social explanation requires discovery of the underlying causal mechanisms that give rise to outcomes of interest." (Little 2009, 167)

Im Mittelpunkt dieses Kausalitätsverständnisses steht folglich das Konzept der sozialen Mechanismen, das sich nach Hedström wie folgt definiert lässt:

> „A social mechanism is a precise, abstract, and action-based explanation which shows how the occurrence of a triggering event regularly generates the type of outcome to be explained." (Hedström 2005, 25 in Anlehnung an Hedström/Swedberg 1998b)[28]

---

26 Holland argumentiert aus der Perspektive der kontrafaktischen Logik für eine konzeptionelle Präferenz der *Effects of the Causes*: „[...] I believe that formal theories of causation must begin with the effects of the causes than vice versa" (Holland 1986, 659). Die Gegenposition wird etwa von Heckman vertreten: „Science is all about constructing models of the causes of effects" (Heckman 2005, 2). Für eine aktuelle Diskussion dieser Thematik sei auf Dawid u. a. (2014) und die dazugehörigen Kommentare und Repliken verwiesen.

27 Der kontrafaktische Ansatz leistet somit keinen Beitrag zur Erklärung der tiefer gelegenen Ursachen eines identifizierten kausalen Effekts. Vielmehr zielt er ausschließlich auf die Identifikation kausaler Bruttoeffekte ab (siehe dazu Morgan/Winship 2007, 280 ff.).

28 Weitere Definitionsversuche sozialer Mechanismen stellen etwa Hedström (2005, 25) und Little (2009, 167 ff.) zur Verfügung.

Eine Strömung, die sich dem mechanismenbasierten Ansatz verschrieben hat, ist die analytische Soziologie. Da es den Umfang des vorliegenden Beitrags bei Weitem sprengen würde, auch nur ansatzweise das mittlerweile äußerst umfangreiche Schrifttum zur analytischen Soziologie vorzustellen, sei lediglich auf einige der zentralen Werke verwiesen (z. B. Demeulenaere 2011; Hedström 2005; Hedström/Bearman 2009; Hedström/Swedberg 1998a) und der Versuch unternommen, deren Grundprinzipien am bewährten Beispiel des Effekts von Viktimisierungserfahrungen auf Kriminalitätsfurcht herauszuarbeiten. Zu diesem Zweck soll auf die kompaktere Definition sozialer Mechanismen als kausale Prozesse bzw. Wirkungsketten (z. B. Opp 2004, 362; Hedström/Swedberg 1998b, 9) rekurriert werden.

Die Aufgabe besteht nun darin, die sozialen Mechanismen und somit jene zentralen kausalen Verbindungslinien zwischen dem Erleben einer Viktimisierungssituation und Kriminalitätsfurcht aufzudecken. Als theoretische Basis kann etwa das von Ferraro (1995) vorgeschlagene *Risk Interpretation Model* herangezogen und erweitert werden.[29] Demzufolge resultieren – wie bereits an anderer Stelle erläutert – Viktimisierungserfahrungen in einer gehobenen Risikoeinschätzung, die wiederum einerseits zu einem direkten Anstieg kriminalitätsbezogener Ängste führt und sich andererseits in der Entwicklung und Umsetzung von Vermeideverhaltens- und Copingstrategien niederschlägt. Diese manifestieren sich in konkreten Handlungen bzw. in der Unterlassung als risikobehaftet wahrgenommener Handlungen und limitieren so die Ausübung der obligatorischen und insbesondere der optionalen Routineaktivitäten viktimisierter Personen (siehe z. B. Averdijk 2011; Rengifo/Bolton 2012). Die spezifische Änderung der alltäglichen Verhaltensmuster wirkt im weiteren Zeitverlauf letztlich furchtreduzierend (Rengifo/Bolton 2012), falls sich diese tatsächlich als effektiv erweist bzw. ihr zumindest Effektivität unterstellt wird und so ein „neues Gefühl von Sicherheit“ entsteht.

Zur empirischen Prüfung der auf Basis der explizierten theoretischen Annahmen postulierten kausalen Mechanismen drängt sich zunächst die Strukturgleichungsmodellierung (z. B. Reinecke 2014) als analytischer Ansatz der ersten Wahl auf (siehe dazu insbesondere Bollen/Pearl 2013), der eine simultane Identifikation der direkten und vermittelten Effektstrukturen erlaubt. Allerdings lässt sich in der Community zur analytischen Soziologie ein deutlicher Trend der Abkehr von gleichungsbasierten hin zu agentenbasierten Modellen und zu Mikrosimulationen (z. B. Epstein 2006) erkennen (z. B. Hedström 2005; Macy/Flache 2009).

---

29 Auf die von Ferraro (1995) vorgeschlagene Berücksichtigung von Kontextfaktoren wie etwa Kriminalitätsaufkommen auf der Makroebene der Gesamtgesellschaft sowie *Incivilities* und Ausmaß sozialer Kohäsion auf der Mesoebene der Nachbarschaft bzw. Wohnumgebung wird zugunsten der Klarheit des Fallbeispiels verzichtet.

4 Zusammenfassung

Die zentralen Befunde des entlang der beiden Themenbereiche „Research Designs“ und „Grundprinzipien von Kausalität“ strukturierten Beitrags lassen sich wie folgt verdichten:

– Das Querschnittsdesign ist charakterisiert durch die einmalige Ziehung und Befragung einer repräsentativen Stichprobe aus der interessierenden Grundgesamtheit. Der Zweck der Durchführung liegt in der Schätzung zentraler Populationsparameter (z. B. viktimisierungsbezogener Prävalenzen bzw. Inzidenzen) für einen definierten Zeitraum zu einem (unmittelbar) nachgelagerten Erhebungszeitpunkt *t*, mit dem deskriptiven Primärziel, das Ausmaß des gesamten (im Hell- und Dunkelfeld) gegenwärtigen Kriminalitätsaufkommens zu *t* abzubilden.

– Das Trenddesign basiert auf der wiederholten Ziehung und Befragung jeweils unabhängiger Stichproben zu unterschiedlichen Zeitpunkten mit dem gleichen Erhebungsinstrument. Dies erlaubt die Abbildung von (kriminalitäts- bzw. viktimisierungsbezogenen) Trends auf Aggregatebene und eignet sich somit etwa für das *Crime Monitoring*.

– Im Rahmen eines Paneldesigns erfolgt die einmalige Ziehung einer repräsentativen Stichprobe. Diese wird in regelmäßigen Abständen mit dem gleichen Erhebungsinstrument befragt, sodass die Beobachtung intra- und interindividueller Entwicklungen über die Zeit möglich ist. Allerdings ergibt sich der Preis des hohen Informationsgehalts des Paneldesigns (so können z. B. viktimisierungsbezogene intra- und interindividuelle Entwicklungsverläufe über die Zeit beobachtet werden) aus den spezifischen methodischen Problemen, die zu erheblichen Verzerrungen in den interessierenden Effekten führen können. Zudem sind die Kosten und der administrative Aufwand als erheblich zu bezeichnen.

– Das retrospektive Design lässt sich definieren als die zu einem Zeitpunkt *t* realisierte einmalige Messung von Merkmalen, die sich auf bereits vergangene Zeitpunkte bzw. -räume beziehen. Während retrospektive Designs einige der Defizite von Paneldesigns vermeiden, ist ihre zentrale Schwäche in der generell geringen Validität der retrospektiven Fragen bzw. Items zu verorten.

– Der Frage nach den Ursache-Wirkungs-Beziehungen zwischen sozialen Phänomenen – im vorliegenden Fall z. B. zu den Ursachen von (wiederholten) Viktimisierungserlebnissen, Kriminalitätsfurcht, Vermeideverhalten – kommt in den Sozialwissenschaften eine zentrale Bedeutung zu, da

nur Erkenntnisse über die kausalen Beziehungsstrukturen, die das „Zustandekommen“ interessierender Phänomen erklären, Anhaltspunkte zur effektiven systematischen Einflussnahme über die Implementierung geeigneter Interventionen zur Verfügung stellen (Stichwort: *Evidence-Based Policy*).

– Gegenwärtig liegt kein einheitliches Kausalitätskonzept vor. Vielmehr lassen sich in Anlehnung an Goldthorpe (2001) drei Prinzipien unterscheiden: Kausalität als (1) robuste Abhängigkeit, (2) konsequente Manipulation und (3) generativer Prozess.

– Aus der zeitlichen Diskrepanz zwischen dem Auftreten einer Ursache und der daraus resultierenden Wirkung ergibt sich die Notwendigkeit, Designs zur validen Beantwortung kausaler Fragestellungen derart zu konzeptualisieren, dass dieser *Time Lag* in angemessener Weise berücksichtigt werden kann. Diese Anforderung wird auf Individualebene lediglich vom Paneldesign und – mit deutlichen Abstrichen hinsichtlich der Validität der Messungen – vom retrospektiven Design erfüllt.

– Abschließend lässt sich die folgende Konklusion formulieren: Zur angemessenen Beantwortung kausaler Fragestellungen bedarf es im Allgemeinen zumindest (1) eines konkreten theoretischen Ansatzes, der einer empirischen Prüfung zugänglich ist, (2) komplexer – in aller Regel longitudinaler – Designs sowie (3) elaborierter statistischer Verfahren, die weit über die Deskriptivstatistik hinausgehen. Ist zumindest eine dieser Anforderungen nicht erfüllt, sollte *kausalen* Schlussfolgerungen ein hohes Maß an Skepsis entgegengebracht werden!

5 Weiterführende Literatur

Als allgemeine Einführung in die Grundlagen, die konkrete Planung und die praktische Durchführung longitudinaler Erhebungsdesigns sei neben Toon (2000) auf Bijleveld/van der Kamp (1998), Kasprzyk u. a. (1989), Lynn (2009) sowie Menard (2002, 2008) verwiesen. Für eine Darstellung des Analysepotenzials von Trenddaten können einführend Firebaugh (1997, 2010) sowie weiterführend McLaren/Steel (im Erscheinen) empfohlen werden. Hinsichtlich der Analyse von Paneldaten lassen sich grundlegend zwei Richtungen unterscheiden: der ökonometrische und der strukturgleichungsbasierte Ansatz. Für erstgenannte Perspektive sei auf Andreß u. a. (2013), Baltagi (2013) und Wooldridge (2002) sowie auf die zahlreichen weiteren ökonometrischen Textbücher verwiesen. Für den strukturgleichungsbasierten Ansatz kann eine Empfehlung für die aktuellen Werke von Little (2013) sowie McArdle/Nesselroade (2014) ausgesprochen werden.

Unter der umfassenden Literatur zu den Grundprinzipien von Kausalität fällt es schwer, einige wenige Arbeiten für eine explizite Empfehlung auszuwählen. Für anwendungsorientierte Leserinnen und Leser dürften die umfassenden und analytisch ausgerichteten Werke von Morgan (2013), Pearl (2009) sowie Morgan/Winship (2007) wohl den größten Nutzen stiften.

6 Literatur

Allison, Paul D. (2009): Fixed Effects Regression Models. Thousand Oaks: Sage.

Andreß, Hans-Jürgen; Golsch, Katrin und Schmidt, Alexander W. (2013): Applied Panel Data Analysis for Economic and Social Surveys. Berlin: Springer.

Angrist, Joshua D.; Imbens, Guido W. und Rubin, Donald B. (1996): Identification of Causal Effects Using Instrumental Variables. In: Journal of the American Statistical Association, 91, S. 444–455.

Averdijk, Margit (2011): Reciprocal Effects between Victimization and Routine Activities. In: Journal of Quantitative Criminology, 27, 2, S. 125–149.

Averdijk, Margit (2014): Methodological Challenges in the Study of Age-Victimization Patterns. Can We Use the Accelerated Design of the NCVS to Construct Victim Careers? In: International Review of Victimology, 20, 3, S. 265–288.

Baltagi, Badi H. (2013): Econometric Analysis of Panel Data. Hoboken: Wiley.

Barnard, George A. (1982): Causation. In: Kotz, Samuel; Johnson, Norman L. (Hg.): Encyclopedia of Statistical Sciences, Band 1. New York: Wiley, S. 387–389.

Barringer, Sondra N.; Eliason, Scott R. und Leahey, Erin (2013): A History of causal Analysis in the Social Sciences. In: Morgan, Stephen L. (Hg.): Handbook of Causal Analysis for Social Research. New York: Springer, S. 9–26.

Biderman, Albert D.; Cantor, David (1984): A Longitudinal Analysis of Bounding, Respondent Conditioning and Mobility as Sources of Panel Bias in the National Crime Survey. Proceedings of the Survey Methods Research Section. Alexandria: American Statistical Association, S. 708–713.

Bijleveld, Catrien C. J. H.; van der Kamp, Leo. J. T. (1998): Longitudinal Data Analysis. Designs, Models and Methods. Thousand Oaks: Sage.

Birkel, Christoph; Guzy, Nathalie; Hummelsheim, Dina; Oberwittler, Dietrich und Pritsch, Julian (2014): Der deutsche Viktimisierungssurvey 2012. Erste Ergebnisse zu Opfererfahrungen, Einstellungen gegenüber der Polizei und Kriminalitätsfurcht (= Arbeitsbericht A7 10/2014 aus der Schriftenreihe des Max-Planck-Instituts für ausländisches und internationales Strafrecht). Freiburg: Max-Planck-Instituts für ausländisches und internationales Strafrecht.

Blalock, Hubert M. (1964): Causal Inference in Nonexperimental Research. New York: Norton.

Blossfeld, Hans-Peter; Rohwer, Götz (1997): Causal Inference, Time and Observation Plans in the Social Sciences. In: Quality & Quantity, 31, 4, S. 361–384.

Boers, Klaus (1991): Kriminalitätsfurcht. Über den Entstehungszusammenhang und die Folgen eines sozialen Problems. Pfaffenweiler: Centaurus.

Boers, Klaus (2003): Fear of Violent Crime. In: Heitmeyer, Wilhelm; Hagan, John (Hg.): International Handbook of Violence Research. Dordrecht: Kluwer Academic Publishers, S. 1131–1150.

Boers, Klaus; Reinecke, Jost; Bentrup, Christina; Daniel, Andreas; Kanz, Kristina-Maria; Schulte, Philipp; Seddig, Daniel; Theimann, Maike; Verneuer, Lena und Walburg, Christian (2014): Vom Jugend- zum frühen Erwachsenenalter. Delinquenzverläufe und Erklärungszusammenhänge in der Verlaufsstudie „Kriminalität in der modernen Stadt“. In: Monatsschrift für Kriminologie und Strafrechtsreform, 97, 3, S. 183–202.

Bollen, Kenneth A. (1989): Structural Equations with Latent Variables. New York: Wiley.

Bollen, Kenneth A.; Pearl, Judea (2013): Eight Myths About Causality and Structural Equation Models. In: Morgan, Stephen L. (Hg.): Handbook of Causal Analysis for Social Research. New York: Springer, S. 301–328.

Boudon, Raymond (1976): Comment on Hauser's Review of Education, Opportunity, and Social Inequality. In: American Journal of Sociology, 81, 5, S. 1175–1187.

Breen, Richard (1996): Regression Models. Censored, Sample Selected, or Truncated Data. Thousand Oaks: Sage.

Breiding; Matthew J.; Smith, Sharon G.; Basile, Kathleen C.; Walters, Mikel L.; Chen, Jieru und Merrick, Melissa T. (2014): Prevalence and Characteristics of Sexual Violence, Stalking, and Intimate Partner Violence Victimization – National Intimate Partner and Sexual Violence Survey, United States, 2011. In: Morbidity & Mortality Weekly Report, 63, 8, S. 1–18.

Cantor, David; Lynch, James P. (2000): Self-Report Surveys as Measures of Crime and Criminal Victimization. In: Duffee, David (Hg.): Measurement and Analysis of Crime and Justice. Criminal Justice 2000. Washington DC: National Institute of Justice, S. 85–138.

Cantwell, Patrick J. (2008): Panel Conditioning. In: Lavrakas, Paul J. (Hg.): Encyclopedia of Survey Research Methods, Band 2. Thousand Oaks: Sage, S. 567–568.

Cohen, Jacob (1988): Statistical Power Analysis for the Behavioral Sciences. Mahwah: Lawrence Erlbaum Associates.

Cohen, Jacob; Cohen, Patricia; West, Stephen G. und Aiken, Leona S. (2003): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Mahwah: Lawrence Erlbaum Associates.

Cook, Thomas D. (2002): Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for Not Doing Them. In: Educational Evaluation and Policy Analysis, 23, 4, S. 175–199.

Dawid, Philip A.; Faigman, David L. und Fienberg, Stephen E. (2014): Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects. In: Sociological Methods & Research, 43, 3, S. 359–421.

Demeulenaere, Pierre (Hg.) (2011): Analytical Sociology and Social Mechanisms. Cambridge: Cambridge University Press.

de Leeuw, Edith; de Heer, Wim (2002): Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In: Groves, Robert M.; Dillman, Don A.; Eltinge, John L. und Little, Roderick J. A. (Hg.): Survey Nonresponse. New York: Wiley, S. 41–54.

Duncan, Greg J.; Kalton, Graham (1987): Issues of Design and Analysis of Surveys across Time. In: International Statistical Review, 55, 1, S. 97–117.

Epstein, Joshua M. (2006): Generative Social Science. Studies in Agent-Based Computational Modeling. Princeton: Princeton University Press.

Farrell, Graham; Tseloni, Andromachi; Wiersema, Brian und Pease, Ken (2001): Victim Careers and 'Career Victims'? Towards a Research Agenda. In: Farrell, Graham; Pease, Ken (Hg.): Repeat Victimization. Monsey: Criminal Justice Press, S. 241–254.

Ferraro, Kenneth F. (1995): Fear of Crime. Interpreting Victimization Risk. Albany: State University of New York Press.

Firebaugh, Glenn (1997): Analyzing Repeated Surveys. Thousand Oaks. Sage.

Firebaugh, Glenn (2010): Analyzing Data from Repeated Surveys. In: Marsden, Peter W.; Wright, James D. (Hg.): Handbook of Survey Research. Howard House: Emerald Group Publishing Limited, S. 795–812.

Goldthorpe, John H. (2001): Causation, Statistics, and Sociology. In: European Sociological Review, 17, 1, S. 1–20.

Gray, Emily; Jackson, Jonathan und Farrall, Stephen (2008): Reassessing the Fear of Crime. In: European Journal of Criminology, 5, 3, S. 363–380.

Groves, Robert M.; Fowler, Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor und Tourangeau, Roger (2009): Survey Methodology. Hoboken: Wiley.

Guo, Shenyang; Fraser, Mark W. (2010): Propensity Score Analysis: Statistical Methods and Applications. Thousand Oaks: Sage.

Guzy, Nathalie; Leitgöb, Heinz (2015): Assessing Mode Effects in Online and Telephone Victimization Surveys. In: International Review of Victimology, 21, 1, S. 101–131.
Hayes, Andrew F. (2013): Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York: Guilford Press.
Heckman, James (1979): Sample Selection Bias as a Specification Error. In: Econometrica, 47, 1, S. 153–161.
Heckman, James (2005): The Scientific Model of Causality. In: Sociological Methodology, 35, S. 1–97.
Hedström, Peter (2005): Dissecting the Social. On the Principals of Analytical Sociology. Cambridge: Cambridge University Press.
Hedström Peter; Bearman, Peter (Hg.) (2009): The Oxford Handbook of Analytical Sociology. Oxford: Oxford University Press.
Hedström Peter; Swedberg, Richard (Hg.) (1998a): Social Mechanisms. An Analytical Approach to Social Theory. Cambridge: Cambridge University Press.
Hedström Peter; Swedberg, Richard (1998b): Social Mechanisms. An Introductory Essay. In: Hedström Peter; Swedberg, Richard (Hg.): Social Mechanisms. An Analytical Approach to Social Theory. Cambridge: Cambridge University Press, S. 1–31.
Hempel, Carl G.; Oppenheim, Paul (1948): Studies in the Logic of Explanation. In: Philosophy of Science, 15, 2, S. 135–175.
Hirtenlehner, Helmut; Farrall, Stephen (2014): Is the "Shadow of Sexual Assault" Responsible for Women's Higher Fear of Burglary? In: British Journal of Criminology. DOI:10.1093/bjc/azu054 – Download vom 02. 08. 2014.
Hirtenlehner, Helmut; Meško, Gorazd (2011): The Linking Mechanisms between Personal Victimization and Fear of Crime: Testing a Theory of Psychological Incapacitation. In: Hasselm, Alicia F. (Hg.): Crime: Causes, Types and Victims. New York: Nova Science Publishers, S. 65–86.
Holland, Paul W. (1986): Statistics and Causal Inference. In: Journal of the American Statistical Association, 81, S. 945–960.
Iacobucci, Dawn (2008): Mediation Analysis. Thousand Oaks: Sage.
Kasprzyk, Daniel; Singh, M. P.; Duncan, Greg und Kalton, Graham (Hrsg.) (1989): Panel Surveys. New York: Wiley.
Kish, Leslie (1965): Survey Sampling. New York: Wiley.
Lauritsen, Janet L. (1998): The Age-Crime Debate: Assessing the Limits of Longitudinal Self-Report Data. In: Social Forces, 77, 1, S. 127–154.
Lauritsen, Janet L. (1999): Limitations in the Use of Longitudinal Self-Report Data: A Comment. In: Criminology, 37, 3, S. 687–694.
Lewis, David (1973): Causation. In: The Journal of Philosophy, 70, 17, S. 556–567.

Little, Daniel (2009): The Heterogeneous Social: new Thinking About the Foundations of the Social Sciences. In: Mantzabinos, Chrysostomos (Hg.): Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice. Cambridge: Cambridge University Press, S. 154–178.

Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical Analysis with Missing Data. Hoboken: Wiley.

Little, Todd D. (2013): Longitudinal Structural Equation Modeling. New York: Guilford Press.

Loeber, Rolf; Farrington, David P. (2014): Age-Crime Curve. In: Bruinsma, Gerben; Weisburd, David (Hg.): Encyclopedia of Criminology and Criminal Justice. New York: Springer, S. 12–18.

Lugtig, Peter (2014): Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers. In: Sociological Methods & Research, 43, 4, S. 699–723.

Lynn, Peter (Hrsg.) (2009): Methodology of Longitudinal Surveys. Chichester: Wiley.

Macy, Michael; Flache, Andreas (2009): Social Dynamics from the Bottom Up. Agent-Based Models of Social Interaction. In: Hedström, Peter; Bearman, Peter (Hg.): The Oxford Handbook of Analytical Sociology. Oxford: Oxford University Press, S. 245–268.

May, David C. (2001): The Effect of Fear of Sexual Victimization on Adolescent Fear of Crime. In: Sociological Spectrum, 21, 2, S. 141–174.

McArdle, John J.; Nesselroade, John R. (2014): Longitudinal Data Analysis Using Structural Equation Models. Washington DC: American Psychological Association.

McLaren, Craig; Steel, David (im Erscheinen): Design and Analysis of Repeated Surveys. Hoboken: Wiley.

Menard, Scott (2002): Longitudinal Research. Thousand Oaks: Sage.

Menard, Scott (Hg.) (2008): Handbook of Longitudinal Research. Design, Measurement, and Analysis. Burlington: Elsevier.

Meredith, William (1993): Measurement Invariance, Factor Analysis and Factorial Invariance. In: Psychometrika, 58, 4, 525–543.

Moosbrugger, Helfried; Kelava, Augustin (2008): Testtheorie und Fragebogenkonstruktion. Berlin und Heidelberg: Springer.

Morgan, Stephen L. (Hg.) (2013): Handbook of Causal Analysis for Social Research. New York: Springer.

Morgan, Stephen L.; Winship, Christopher (2007): Counterfactuals and Causal Inference. Methods and Principles for Social Research. New York: Cambridge University Press.

Mosher, Clayton J.; Miethe, Terance D. und Hart, Timothy C. (2011): The Mismeasure of Crime. Thousand Oaks: Sage.

National Research Council (2008): Surveying Victims. Options for Conducting the National Crime Victimization Survey. Washington DC: National Academic Press.

Nyman, Jerzy S. (1990 [1923]): On the Application of Probability theory to Agricultural Experiments. Essay on Principals. Section 9. In: Statistical Science, 5, 4, S. 465–472.

Office for National Statistics (2014): User Guide to Crime Statistics in England and Wales. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/guide-to-finding-crime-statistics/crime-survey-for-england-and-wales–csew-/index.html – Download vom 17. 12. 2014.

Opp, Karl-Dieter (2004): Erklärung durch Mechanismen: Probleme und Alternativen. In: Kecskes, Robert; Wagner, Michael und Wolf, Christof (Hg.): Angewandte Soziologie. Wiesbaden: VS Verlag, S. 361–379.

Opp, Karl-Dieter (2010): Kausalität als Gegenstand der Sozialwissenschaften und der multivariaten Statistik. In: Wolf, Christof; Best, Henning (Hg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag, S. 9–38.

Pearl, Judea (2009): Causality. Models, Reasoning, and Inference. New York: Cambridge University Press.

Pearl, Judea (2010): The Foundations of Causal Inference. In: Sociological Methodology, 40, S. 75–149.

Powers, Edward A.; Goudy, Willis J. und Keith, Pat M. (1978): Congruence between Panel and Recall Data in Longitudinal Research. In: Public Opinion Quarterly, 42, 3, S. 380–389.

Reinecke, Jost (2005): Strukturgleichungsmodelle in den Sozialwissenschaften, 1. Aufl. München: Oldenbourg.

Reinecke, Jost (2014): Strukturgleichungsmodelle in den Sozialwissenschaften, 2. Aufl. München: De Gruyter Oldenbourg.

Rengifo, Andres F.; Bolton, Amanda (2012): Routine Activities and Fear of Crime: Specifying Individual-Level Mechanisms. European Journal of Criminology, 9, 2, S. 99–119.

Roberts, Jennifer; Hroney, Julie (2010): The Life Event Calendar Method in Criminological Research. In: Piquero, Alex R.; Weisburd, David (Hg.): Handbook of Quantitative Criminology. New York: Springer, S. 289–312.

Rosenbaum, Paul R. (2002): Observational Studies. New York: Springer.

Rossi, Peter H.; Lipsey, Mark W. und Freeman, Howard E. (2007): Evaluation. A Systematic Approach. Thousand Oaks: Sage.

Rost, Jürgen (2004): Lehrbuch Testtheorie – Fragebogenkonstruktion. Bern: Verlag Hans Huber.

Rubin, Donald B. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. In: Journal of Educational Psychology, 66, 5, S. 688–701.
Rubin, Donald B. (1976): Inference and Missing Data. In: Biometrika, 63, 3, S. 581–592.
Rubin, Donald B. (2007): The Design *versus* the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trails. In: Statistics in Medicine, 26, 1, S. 20–36.
Sampson, Robert J. (2010): Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology. In: Journal of Quantitative Criminology, 26, 4, S. 489–500.
Schnell, Rainer; Hill, Paul B. und Esser, Elke (2013): Methoden der empirischen Sozialforschung. München: Oldenbourg.
Schwarz, Norbert (1990): Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction. In: Hendrick, Clyde; Clark, Margaret S. (Hg.): Research Methods in Personality and Social Psychology (Review of Personality and Social Psychology, Bd. 11). Beverly Hills: Sage, S. 98–119.
Schwarz, Norbert; Sudman, Seymour (Hg.) (1994): Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer.
Shadish, William R.; Cook, Thomas D. und Campbell, Donald T. (2002): Experimental und Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin Company.
Sosa, Ernest; Tooley; Michael (Hg.) (1993): Causation. New York: Oxford University Press.
Sutton, James E. (2010): A Review of the Life-Calendar Method for Criminological Research. In: Journal of Criminal Justice, 38, 5, S. 1038–1044.
Sturgis, Patrick; Allum, Nick und Brunton-Smith, Ian (2009): Attitudes Over Time: The Psychology of Panel Conditioning. In: Lynn, Peter (Hg.): Methodology of Longitudinal Surveys. Chichester: Wiley, S. 113–126.
Toon, Taris W. (2000): A Primer in Longitudinal Data Analysis. Thousand Oaks: Sage.
Tseloni, Andromachi, Mailley, Jen; Farrell, Graham und Tilley, Nick (2010): Exploring the International Decline in Crime Rates. In: European Journal of Criminology, 7, 5, S. 375–394.
Turner, Anthony G. (1984): The Effect of Memory Bias on the Design of the National Crime Survey. In: Lehnen, Robert G.; Skogan, Wesley G. (Hg.): The National Crime Survey: Working Papers. Band 2: Methodological Studies, S. 80–82.
United Nations (2010): Manual on Victimization Surveys. Genf: United Nations Office on Drugs and Crime; United Nations Economic Commission for Europe.

Urban, Dieter; Mayerl, Jochen (2011): Regressionsanalyse: Theorie, Technik und Anwendung. Wiesbaden: VS Verlag.
Vandenberg, Robert J.; Lance, Charles E. (2000): A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. In: Organizational Research Methods, 3, 1, S. 4–70.
van Kesteren, John; van Dijk, Jan und Mayhew, Pat (2014): The International Crime Victims Surveys. A Retrospective. In: International Review of Victimology, 20, 1, S. 49–69.
Warren, John R.; Halpern-Manners, Andrew (2012): Panel Conditioning in Longitudinal Social Science Surveys. In: Sociological Methods & Research, 41, 4, S. 491–534.
Waterton, Jennifer; Lievesley, Denise (1989): Evidence of Conditioning Effects in the British Social Attitudes Panel. In: Kasprzyk, Daniel; Singh, M. P.; Duncan, Greg und Kalton, Graham (Hg.): Panel Surveys. New York: Wiley, S. 319–339.
Watson, Nicole; Wooden, Mark (2009): Identifying Factors Affecting Longitudinal Survey Response. In: Lynn, Peter (Hg.): Methodology of Longitudinal Surveys. Chichester: Wiley, S. 157–181.
Widaman, Keith F.; Ferrer, Emilio und Conger, Rand D. (2010): Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. In: Child Development Perspectives, 4, 1, S. 10–18.
Weischer, Christoph (2015): Panelsterblichkeit. In: Diaz-Bone, Rainer; Weischer, Christoph (Hg.): Methoden-Lexikon für Sozialwissenschaften. Wiesbaden: VS Verlag, S. 303.
Welsh, Brandon C.; Braga, Anthony A. und Bruinsma, Gerben J. N. (Hg.) (2013): Experimental Criminology. Prospects for Advancing Science and Public Policy. New York: Cambridge University Press.
Winship; Christopher; Morgan, Stephen L. (1999): The Estimation of Causal Effects from Observational Data. In: Annual Review of Sociology, 25, S. 659–707.
Wittenberg, Jochen (2015): Erhebungsdesigns für kriminologische Befragungen und Experimente. In: Eifler, Stefanie; Pollich, Daniela (Hg.): Empirische Forschung über Kriminalität. Methodologische und methodische Grundlagen. Wiesbaden: VS Verlag, S. 95–122.
Woltman, Henry; Bushery, John M. und Carstensen, Larry (1984): Recall Bias and Telescoping in the National Crime Survey. In: Lehnen, Robert G.; Skogan, Wesley G. (Hg.): The National Crime Survey: Working Papers. Band 2: Methodological Studies, S. 90–93.
Wooldridge, Jeffrey M. (2002): Econometric Analysis of Cross Section and Panel Data. Cambridge: MIT Press.

Xie, Min; McDowall, David (2008): Escaping Crime: The Effects of Direct and Indirect Victimization on Moving. In: Criminology, 46, 4, S. 809–840.

Yang, Yang; Land, Kenneth C. (2013): Age-Period-Cohort Analysis. New Models, Methods, and Empirical Applications. Boca Raton: CRC Press.

Yberra, Lynn M. R.; Lohr, Sharon L. (2000): Effects of Attrition in the National Crime Victimization Survey. In: Proceedings of the Survey Methods Research Section. Alexandria: American Statistical Association, S. 870–874.

4 Grenzen von Opferbefragungen

Grenzen von Opferbefragungen

Helmut Kury

1 Einleitung

Opferbefragungen gehören heute international zu einem festen Bestandteil empirischer kriminologischer Forschung. Nach dem starken Aufkommen von Meinungsumfragen und der „Entdeckung“ dieser Möglichkeit zur Informationsgewinnung auch zu politischen, kommerziellen oder mehr und mehr auch wissenschaftlichen Themen nach dem Zweiten Weltkrieg vor allem in den USA, „war es nur noch eine Frage der Zeit, bis diese Methode zur Erforschung der Erfahrungen und Ansichten der Bevölkerung im Zusammenhang mit der Kriminalität zur Anwendung kommen würde“ (Killias 2002, 68). Dabei zeigte der Einsatz von Umfragen zu damals weitgehend noch tabuisierten aber umso mehr Aufmerksamkeit auf sich ziehenden Themen wie dem „Sexuellen Verhalten der Frau“ oder des Mannes von Kinsey u. a. (1963, 1964) bis dahin ungeahnte Möglichkeiten auf, mehr Wissen auch über weitgehend abgeschirmte und tabuisierte Bereiche des menschlichen Verhaltens, gerade auch im Zusammenhang mit Kriminalität, zu erlangen. Diese frühen Studien zu sensiblen Themen wiesen gleichzeitig bereits deutlich auch auf methodische Probleme und Grenzen solcher Umfragen hin und gaben Beispiele für ein differenziertes Vorgehen bei der Gewinnung aussagekräftiger Daten (Kinsey u. a. 1963, 23ff.). Deutlich wurde bereits hier, dass die Erlangung valider Daten vor allem eine gute Planung, die Auswahl einer repräsentativen Stichprobe und ein differenziertes Vorgehen bei der Instrumentenentwicklung und insbesondere der Durchführung von Umfragen voraussetzen, damit auch Zeit und Geld kosten.

Neben den Kinsey-Reports, bei denen es vorwiegend um selbstberichtetes sexuelles Verhalten ging, sowie einzelnen Vorläufern (s. z. B. Wallerstein u. Wyle 1947) hat die Forschung in diesem Bereich mit Self-Report-Surveys an Studentinnen und Studenten und delinquenten Jugendlichen begonnen (vgl. Porterfield 1946; Short/Nye 1957; zusammenfassend Killias 2002, 58; Schneider 2007a, 311; Amelang 1986, 101). Die gefundenen Ergebnisse, etwa über die Höhe des Dunkelfeldes, waren bei Opfer- und Täterbefragungen beeindruckend, Methodenprobleme blieben in diesem Kontext zunächst vielfach wenig beachtet, obwohl sie vereinzelt bereits diskutiert wurden (Schneider 2007a, 311). Vor dem Hintergrund der mit fortschreitender Forschung sich mehr und mehr zeigenden Methodenprobleme (vgl. Sparks 1981, 24ff.; Reuband 1989; Ewald u. a. 1994; Fattah 1991, 35ff.) wurde schon

damals klar, „dass das ursprüngliche Anliegen, das Dunkelfeld der Kriminalität bis in den letzten Winkel auszuleuchten und so die ‚wahre' Kriminalität, die schweigend und unentdeckt hinter den offiziellen Statistiken steht, vollständig und korrekt zu erfassen, zum Scheitern verurteilt war" (Greve u. a. 1994, 5).

Während es bei den Crime Surveys vor allem um die Erfassung des Ausmaßes an Kriminalität geht, Opferaspekte mehr im Hintergrund stehen (Feldmann-Hahn 2011, 4), sind Victim Surveys, welche inzwischen die Mehrzahl der Befragungen ausmachen, Untersuchungen, „die von einem viktimologischen Standpunkt aus auch die tiefer liegenden Zusammenhänge der Opferwerdung mit den vielfältigen Aspekten einer Opferperspektive erforschen wollen" (Sautner 2010, 146). Gerade sie haben dann einen enormen Input in die kriminologische und viktimologische Forschung gebracht, bis hin zu Auswirkungen in Kriminalpolitik und Gesetzgebung.

Die anfängliche Beschränkung auf die Befragung leicht zu erreichender und zu motivierender junger Menschen, auf „Bequemlichkeitsstichproben", wurde der Opferforschung dann bei Diskussionen zur Methodik mehr und mehr vorgeworfen, etwa mit dem berechtigten Hinweis, die so gewonnenen Ergebnisse seien nicht verallgemeinerbar (Schneider 2007a, 316). So betonte etwa Kaiser (1996, 394) in diesem Zusammenhang noch, Dunkelfeldstudien würden sich „nahezu ausnahmslos auf Kinder, Jugendliche und Heranwachsende beschränken oder auf die frühere Jugendkriminalität der jetzt befragten Erwachsenen. Demgegenüber ist das Delinquenzverhalten des unbekannten erwachsenen Straftäters, der schwere Verbrechen begeht, so gut wie unbekannt".

Eine erste umfassende Darstellung internationaler empirisch-viktimologischer Forschung aus Anlass des VIIth International Symposium on Victimology der 1979 in Münster gegründeten World Society of Victimology enthält über 100 Beiträge zu verschiedenen Fragestellungen der Viktimologie und unterschiedlichen Opfergruppen (Kaiser u. a. 1991). Schon hier wurde deutlich auf Einschränkungen der Methode, etwa hinsichtlich wesentlicher Opfergruppen, wie Machtmissbrauch (Dussich 1991; Neuman 1991) oder Wirtschaftskriminalität (Titus 1991; Niemi 1991) hingewiesen.

Trotz aller methodischen Probleme, die Opferbefragungen bei der Komplexität des zu untersuchenden Sachverhaltes zwangsläufig mit sich bringen, besteht heute weitgehend Einigkeit darüber, dass „eine rationale Kriminal- und Strafrechtspolitik [. . .] ohne eine solide empirische Grundlage nicht möglich" ist und dazu gehören vor allem Ergebnisse aus Victim Surveys (Heinz 2006, 241).

Im Folgenden sollen zunächst Möglichkeiten und Grenzen der Methode der Opferbefragung erörtert werden. Hierbei gehen wir vor allem auf Probleme der Definition von „Opfer“ ein, auf solche des Zugangs zu Stichproben und Schwierigkeiten in der Forschungsmethodologie. Anschließend werden Möglichkeiten der Weiterentwicklung der Methode stichwortartig diskutiert.

2 Grenzen von Opferbefragungen

Die Vielschichtigkeit des zu untersuchenden Gegenstandes Kriminalität bzw. Viktimisierungen bringt zwangsläufig eine Vielzahl von methodischen Problemen mit sich. Der Erste Periodische Sicherheitsbericht der Bundesregierung (Bundesministerium des Innern, Bundesministerium der Justiz 2001, 599) stellt fest: „Erkenntnisse über Ausmaß, Struktur und Entwicklung der Kriminalität einerseits, über Strafverfolgung, Strafvollstreckung und Strafvollzug andererseits müssen in ausreichendem Umfang vorhanden sein, um kriminal- und strafrechtspolitische Maßnahmen erfolgreich gestalten und in ihren Auswirkungen überprüfen zu können“. Hierbei werden ausgesprochen komplexe Untersuchungsbereiche angesprochen, die erhebliche methodische Probleme aufwerfen. So wird dann auch im Zweiten Periodischen Sicherheitsbereit (Bundesministerium des Innern – Bundesministerium der Justiz 2006, 9) ausgeführt: „Kriminalität ist kein Sachverhalt, der einfach gemessen werden könnte, wie etwa die Länge, das Gewicht oder die Temperatur eines Gegenstandes. Kriminalität ist vielmehr ein von Struktur und Intensität strafrechtlicher Sozialkontrolle abhängiger Sachverhalt. Die Bezeichnung als ‚Kriminalität‘ ist einerseits das Ergebnis vorgängiger gesellschaftlicher Festlegungen, andererseits die Folge von zumeist mehrstufig verlaufenden Prozessen der Wahrnehmung von Sachverhalten und deren Bewertung.“

Der Zweite Periodische Sicherheitsbericht (Bundesministerium des Innern, Bundesministerium der Justiz 2006, 17) weist vor dem Hintergrund unterschiedlicher methodischer Vorgehensweisen weiterhin auf eine mangelnde Vergleichbarkeit der seit den 1990er Jahren auch bundesweit durchgeführten Opferbefragungen hin. Auch werden in Umfragen stets nur ausgewählte Fall- und Tätergruppen erfasst, andere würden sich „mit dieser Methode entweder nicht oder nur mit großem Aufwand untersuchen lassen“ (2006, 16f.). „Über Umfang, Struktur und Entwicklung der Kriminalität in ihrer Gesamtheit ist deshalb – empirisch belegt – nichts bekannt. Aber auch bezüglich der Eigentums- und Vermögensdelikte, dem gegenwärtigen Hauptbefragungsgebiet von Opferbefragungen, kann das Dunkelfeld weder vollständig noch verzerrungsfrei aufgehellt werden. Im Übrigen wird auch mit Dunkelfeldforschungen nicht Kriminalitätswirklichkeit gemessen, sondern die Selbstbeurteilung und Selbstauskunft der Befragten. Es wird mithin erfasst, wie Befragte bestimmte Handlungen definieren, bewerten, kategorisieren, sich daran erinnern und be-

reit sind, darüber Auskunft zu geben. Dunkelfeldforschungen sind deshalb kein Ersatz für Kriminalstatistiken. Sie stellen aber eine notwendige Ergänzung der Kriminalstatistiken dar, um – jedenfalls für Teilbereiche – die stattfindenden Selektionsprozesse, insbesondere hinsichtlich der Anzeige, erkennen, quantitativ einordnen und in ihrer Bedeutung für das kriminalstatistische Bild bewerten zu können“ (2006, 17).

Hier wird somit deutlich auf Einschränkungen der Möglichkeiten von Opferbefragungen und deren Aussagekraft hingewiesen, gleichzeitig jedoch auch deren Nutzen hervorgehoben. Im Folgenden sollen wesentliche Grenzen der Methode kurz dargestellt werden.

In der internationalen Literatur wird, trotz aller immer wieder aufkommenden Kritik zu verschiedenen Aspekten der Umfragen, kaum daran gezweifelt, dass Opferbefragungen die kriminologische Forschung erheblich bereichert und einen enormen Erkenntnisgewinn gebracht haben. Ergebnisse von Umfragen, etwa zu speziellen Opfergruppen wie Kindern, Frauen oder alten Menschen, oder zu spezifischen Umfeldern, wie Familie, Heimen, in diesem Zusammenhang in den letzten Jahren vor allem auch kirchlichen Einrichtungen, oder Gefängnissen haben den Blick für bestehende Problematiken zunehmend geschärft, auch auf Mangel an Informationen hingewiesen, und Opferhilfsorganisationen in ihren Bemühungen, den Betroffenen zu helfen, unterstützt. Erst die Ergebnisse aus Opferbefragungen machten oft deutlich, dass es in den untersuchten Bereichen „ein Problem“ gibt. Zahlreiche gesetzliche Entwicklungen, etwa zum Opferschutz, zur Opferhilfe und -entschädigung oder zu den Rechten des Opfers im Strafverfahren (vgl. Schwind 2013, 411ff.), wären ohne diese viktimologische Forschung kaum denkbar.

Opferbefragungen zeigen allerdings auch, wie bereits von Anfang an betont wurde, erhebliche Einschränkungen, die mit dieser Methode der Befragung letztlich auch nur teilweise zu überwinden sind. In diesem Kontext betonen etwa Greve u. a. (1994, 8) vor dem Hintergrund methodischer Aspekte von Opferstudien: „Insgesamt betrachtet wäre es [...] sicher unangemessen, die Ergebnisse von Opferbefragungen gegenüber den vermeintlich höhergradig fehlerbehafteten offiziellen Statistiken als ein valideres, ‚wirklichkeitsnäheres‘ Bild des Kriminalitätsgeschehens zu bezeichnen“. Hierbei ist allerdings auch zu berücksichtigen, dass die inzwischen präziser vorliegenden Ergebnisse aus Umfragen zum Dunkelfeld der Kriminalität, die, auch bei Beachtung aller Unsicherheiten und Zweifel, begründet annehmen lassen, dass dieses im Hinblick auf alle Straftaten, bei mindestens 90 % liegen dürfte, selbst bei schweren Delikten auf 50 % geschätzt wird (Kürzinger 1996, 181; Scheib 2002; Kury 2001). Die Polizeiliche Kriminalstatistik kann als „Tätigkeitsstatistik der Polizei“ nur ein rudimentäres Bild des Kriminalitätsgeschehens abgeben, beeinflusst von zahlreichen Faktoren, das gilt allerdings auch für Op-

ferstudien, wenngleich diese deutlich mehr Licht in das Dunkel bringen können. Im Folgenden sollen vor allem drei wesentliche Bereiche angesprochen werden, welche die Erkenntnismöglichkeiten von Opferstudien über das Kriminalitätsgeschehen einschränken:

– Definitionsprobleme hinsichtlich Opfer und auch Kriminalität,

– Zugangsprobleme zu möglichen Opfern und

– Methodenprobleme bei der Erfassung von Opferwerdungen.

2.1 Definitionsprobleme – wer ist ein Opfer?

Wie Steffen (2013, 56) zu recht betont, gibt es „das Opfer“ nicht, „Opferwerdung, Opferverhalten wie auch Opferwünsche sind höchst individuelle Geschehnisse. Nicht jedes Opfer leidet, einige Opfer leiden aber ihr Leben lang“. In der Kriminologie bzw. Viktimologie werden heute als „Opfer“ in aller Regel ausschließlich Opfer von gesetzlich definierten Straftaten verstanden. Zu Beginn der Viktimologie wurde teilweise noch eine breitere Definition von „Opfer“ diskutiert. So hatte etwa Mendelsohn (1974) ein umfassenderes Verständnis von Viktimologie, wollte etwa auch Opfer von Naturkatastrophen oder Unfällen mit eingeschlossen sehen, also nicht nur Opfer von strafrechtlich verbotenen Verhaltensweisen (Schneider 2007b, 396).

Wie Greve u. a. (1994, 3) betonen, ist der „für die Viktimologie zentrale Opferbegriff [...] schillernd“, was sich auch auf die Operationalisierung im Rahmen von empirischen Untersuchungen und damit auf Forschungsergebnisse auswirkt. Ein Verbrechensopfer kann man nach den Autoren nur dann werden (S. 31), „wenn die zugrundeliegende Handlung nach konsensuellen, expliziten oder sonstigen Kriterien ein Verbrechen war“, dadurch soll eine Abgrenzung vom alltagssprachlichen Opferbegriff erfolgen. Gleichzeitig betonen sie aber (S. 34), dass der Opferbegriff auch von gesellschaftlichen moralischen Rahmenregeln abhänge, er solle sich nicht auf Legalkriterien beschränken, „vor allem dann nicht, wenn man die Folgen von Opfererfahrungen für die betroffenen Personen oder für die Gesellschaft untersuchen möchte“ (S. 35). Der Opferbegriff müsse „auch im Interesse kreativer Forschung – hinreichend flexibel bleiben ...“ (S. 36). Wie Mitscherlich (1999) betont, ist der Opferbegriff auch aus psychologischer und politischer Sicht problematisch.

Nach Baurmann/Schädler (1999, 25f.) wird der Opferbegriff nicht nur „in den Medien, in Kriminologie, Viktimologie, Kriminalstatistik sowie bei der

Strafverfolgung und bei der Opferarbeit [. . .] undifferenziert gebraucht“, sie mussten bei ihren „Befragungen feststellen, dass der Begriff ‚Opfer‘ von den Betroffenen selbst aus mehreren Gründen als problematisch empfunden wurde, dass sie oftmals Begriffe wie ‚Geschädigte(r)‘ oder ‚Verletzte(r)‘ eher annehmen konnten.“ Es könne leicht zu einem Opfer-Labeling kommen, Personen, die sich nicht selbst als Opfer erleben, sollten deshalb auch nicht dazu erklärt werden (1999, 27).

Haas (2014, 245) betont in diesem Zusammenhang: „Hinsichtlich der Bezeichnung eines viktimisierten Menschen als Opfer gibt es in der Literatur und seitens der praktischen Opferhilfe Tendenzen, ihn durch die Substantive ‚Verletzter‘, ‚Geschädigter‘ oder ‚Betroffener‘ zu ersetzen, um einer emotionalen Implikation und einer historischen Determination entgegen zu wirken“. Opfer assoziiere eine „Degradierung zum Objekt, das – passiv dem Geschehen ausgeliefert – scheinbar zu keinem aktiven Beitrag mehr fähig ist“. Frauenunterstützungseinrichtungen würden vor diesem Hintergrund statt Opfer den Begriff „betroffene Frauen“ bevorzugen (S. 246).

Sessar (2012, 264) führt aus: „Die Verwendung des Begriffs ‚Opfer‘ ist [. . .] hoch problematisch, da eine abstrakte strafrechtliche Definition einem individuellen Erlebnis ‚ohne zu fragen‘ aufgepfropft wird. Kriminologisch liegt es näher, zwischen selbstdeklarierten und fremddeklarierten Opfern zu differenzieren, wodurch die Autonomie der von einer Straftat betroffenen Person, sich als Opfer zu verstehen oder nicht, anerkannt werden würde. Etwas Ähnliches intendiert der Begriff ‚Opfererleben‘, gemeint als Ausdruck der Entscheidung, eine negative Erfahrung als Opfer erlebt zu haben oder nicht“ (Ewald u. a. 1994, 79; Wetzels 1996).

Auch Steffen (2013, 60) weist auf die „Ambivalenz des Opferbegriffes“ hin, Opfer zu sein werde den Betroffenen „zugeschrieben“ (Görgen 2012, 90). In der öffentlichen Wahrnehmung gebe es einen „Widerspruch zwischen den ‚idealen Opfern‘ und den ‚wirklichen‘ Opfern“ (Steffen 2013, 60). Die „idealen“ Opfer stellen nur einen kleinen Teil aller Opfer von Straftaten dar, sie sind schwach, hilfe- und schutzbedürftig, sie haben keinerlei Schuld an ihrer Viktimisierung, das sind vor allem Kinder, Frauen (wenn sie sich nichts zuschulden kommen ließen), alte Menschen oder Pflegebedürftige. „Die zahlenmäßig bedeutendsten Gruppen von Kriminalitätsopfern werden dagegen immer noch übersehen bzw. nicht als Opfer wahrgenommen: Männer und Jungen“ (Steffen 2013, 64; vgl. a. Baurmann 2000, 3). „Dass diese Wahrnehmung nicht der Realität der Opferwerdung in unserer Gesellschaft entspricht, das zeigen die Befunde der kriminologischen und viktimologischen Forschung zur Verbreitung und Häufigkeit von Viktimisierungen im Hell- und Dunkelfeld“ (Steffen 2013, 64).

Der Begriff „Opfer“ ist inzwischen, vor allem unter Jugendlichen, teilweise zu einem Schimpfwort geworden (Voß 2003, 58; s. a. Barton 2012, 117; Steffen 2013, 60). Opfer geworden zu sein, erhöht nicht unbedingt den sozialen Status, eher das Gegenteil ist der Fall, vor allem etwa bei Frauen, die Opfer einer Sexualstraftat geworden sind (vgl. in diesem Zusammenhang etwa zu den „Rape Myths“ Kury 2003). Greve u. a. (1994, 15) bemängeln das Fehlen einer Theorie der Opferwerdung (Fattah 1991, 220ff.). Sie benennen Merkmale eines sozialwissenschaftlich und kriminologisch sinnvollen Opferbegriffs, wie Individuierbarkeit, negative Bewertung, Widerfahrnis, Zurechenbarkeit und Verletzung normativer Erwartungen (S. 24).

Vor diesem Hintergrund verwundert es nicht, dass auch Betroffene sich unterschiedlich als „Opfer“ sehen bzw. definieren oder nicht, einerseits etwa aufgrund einer Ablehnung der „Opferrolle“, was etwa bei (sexueller) Gewalt in der Familie nicht selten der Fall sein dürfte, andererseits aufgrund einer Unkenntnis strafrechtlicher Gesetze, die zu einer falschen Zuordnung führen können. So betont etwa Schneider (2007a, 318), dass Vergewaltigungen, die vielfach im sozialen Nahraum geschehen, in Opferbefragungen oft nicht angegeben, als Unfall oder Unglück umdefiniert werden (Vito u. a. 2007, 280). „Wahrscheinlich können nur auf Vergewaltigung und andere Sexual- und Gewaltdelikte spezialisierte Viktimisierungsstudien das wahre Ausmaß dieser Delikte ermitteln. Wegen ihres höchstpersönlichen Charakters passen sie nicht in einen Fragenkatalog“ (Schneider 2007a, 318; vgl. a. Lamnek/Luedtke 2006). Dass Frauen in weit höherem Maße als in klassischen Opferstudien gefunden Opfer von Sexualdelikten und Körperverletzung, vor allem im familiären Bereich, werden, konnte etwa auch die „Frauenstudie“ zeigen (Müller/Schröttle 2004; Zedner 2002, 426).

Steffen (2013, 71) betont weiter zurecht: „Victim Surveys sind in der Regel als Bevölkerungsbefragungen angelegt, erfassen also nicht nur Opfer, sondern auch Nicht-Opfer. Damit ergibt sich das Problem der Abgrenzung bzw. Notwendigkeit einer Selbstdeklaration der Probanden als Opfer“. Reine Opferbefragungen wurden nur selten durchgeführt, etwa von Baurmann u. Schädler (1999) oder Richter (1997). Auch Sautner (2010, 165ff.) weist auf Probleme der Selbstdeklaration als Opfer hin, es sei von fehlerhaften Zuordnungen auszugehen. Nichtopfer könnten angeben, Opfer zu sein, was allerdings relativ selten sei. Viel bedeutsamer sei die Nichtdefinition als Opfer, Ursache hierfür könne sein:

– Vergessen,

– Verdrängung des Opfererlebnisses,

– bewusstes Verschweigen (Scham, Privatangelegenheit, Ablehnung der Opferrolle),

– Einstufung des Ereignisses als irrelevant (s. a. Feldmann-Hahn 2011, 44).

So stellte etwa Gold (1970) in seiner Untersuchung fest, dass 28 % der berichteten Sachverhalte, die von den Befragten als Delinquenz bewertet wurden, nach den Kriterien des Strafrechts keine Delikte waren, die zu einer Strafverfolgung hätten führen können. Auch Levine (1976) berichtet über ein „Crime Overreporting in Criminal Victimization Surveys“ (vgl. a. O'Brien u. a. 1979).

Nur wer sich somit als Opfer erlebt, diese Zuschreibung annimmt, sich auch vor dem Hintergrund gesetzlicher Vorgaben richtig zuordnet, kann bzw. wird in einer Opferstudie entsprechende Angaben zu einer Viktimisierung korrekt machen. Hier spielen „Sensibilitäten“, die sich in den letzten Jahrzehnten vor dem Hintergrund einer öffentlichen einschlägigen Diskussion, etwa zu „Gewalt“ (Kury 2015), deutlich verändert haben und die gruppen- und länderspezifisch sind, eine wesentliche Rolle, was sich gerade auch auf internationale Vergleiche auswirken wird. Die befragte Person muss die Viktimisierung als solche überhaupt erst erkannt haben, was beispielsweise bei Anlagebetrügereien oder neuerdings bei Internetkriminalität schwer sein kann, sie muss sich weiterhin an die Viktimisierung erinnern, einige Autoren weisen darauf hin, dass unangenehme Ereignisse eher vergessen werden (Greve u. a. 1994, 7), und weiterhin muss sie bereit sein, bei einer Umfrage Angaben hierzu zu machen. Schwerere Ereignisse, die in aller Regel seltener vorkommen, erfordern große Stichproben, um zu aussagekräftigen Resultaten zu gelangen. Gerade bei, teilweise auch schweren, Straftaten im sozialen Nahbereich, so in der eigenen Familie, wird etwa aus Schamgefühlen oder aus Angst vor noch schlimmeren Entwicklungen, oder weil die Angelegenheit als „Privatsache“ angesehen wird und die Polizei „sowieso nicht weiterhelfen kann“, vielfach keine Anzeige erstattet. So gab etwa bei der ersten Deutsch-deutschen Opferstudie bei den einzelnen abgefragten Delikten ein relativ hoher Prozentsatz als Grund für eine Nichtanzeige bei der Polizei an, diese sei „unnötig gewesen, kein Fall für die Polizei“ oder diese hätte „doch nichts machen können“, da man keine Beweise gehabt habe (Kury u. a. 1996, 45ff.). Wie Greve u. a. (1994, 7) betonen, können institutionelle Filterungsprozesse durch Opferbefragungen ausgeschlossen werden, nicht jedoch auf Opferseite vorhandene Interpretationen oder Filterungen.

Auf der anderen Seite kann Geltungsstreben, etwa bei Jugendlichen, zu einer Überbetonung eines Ereignisses bis hin zu falschen Angaben führen. Da in Opferbefragungen Viktimisierungsereignisse in der Regel für einen vorgegebenen Zeitraum abgefragt werden, etwa im letzten Monat, Jahr oder in den

letzten fünf Jahren, ist weiterhin die zeitlich korrekte Einordnung etwa länger zurückliegender entsprechender Ereignisse wichtig. Hier kann es zu „Telescoping-Effekten“, einer zeitlichen Fehleinordnung der Geschehnisse, kommen. Hinzu kommt weiterhin, dass die befragte Person das erlittene Ereignis auch richtig als Straftat einordnen können muss, was bei der Komplexität der Straftatbestände für Laien sehr schwer sein kann, was etwa auch daran zu ersehen ist, dass ein erheblicher Teil der von der Polizei, die in solchen Fragen ja geschult ist, registrierten Straftaten später von den Gerichten „umdefiniert“ wird. Es ist bei Opferbefragungen in aller Regel schwierig, strafrechtliche Tatbestände angemessen in die Umgangssprache zu übersetzen. Es kann so leicht vorkommen, dass vor allem im minderschweren Bereich von Straftaten Ereignisse als kriminelle Viktimisierungen gesehen werden, somit „Vorfälle, die rechtlich noch nicht die Grenzen der Strafbarkeit überschreiten, in Opferbefragungen als Viktimisierungserfahrungen registriert werden“ (Bundesministerium des Innern/Bundesministerium der Justiz 2001, 15).

2.2 Zugangsprobleme – wird nur Straßenkriminalität erfasst?

Einer der zentralen Vorwürfe gegenüber Opferbefragungen war von Anfang an die Auswahl der Stichproben, der Mangel an Repräsentativität und damit eine eingeschränkte Verallgemeinerbarkeit der Resultate (Killias 2002, 60; Schneider 2007a, 316), ein Vorwurf, der zumindest zu Beginn der Opferforschung berechtigt war (vgl. oben). So betont etwa der Erste Periodische Sicherheitsbericht (Bundesministerium des Innern/Bundesministerium der Justiz 2001, 14): „Zu den allgemeinen methodischen Problemen einer jeden Befragung zählt vor allem, dass bestimmte Personengruppen typischerweise nicht oder nicht repräsentativ erfasst werden, wie zum Beispiel Obdachlose, Internierte (etwa in Heimen oder in Strafvollzugsanstalten Untergebrachte) oder in bestimmten subkulturellen Milieus lebende Personen. Ferner werden aus erhebungstechnischen Gründen bestimmte Einheiten der Grundgesamtheit mehr oder weniger systematisch ausgeschlossen werden, wie zum Beispiel der deutschen Sprache nicht mächtige Gruppen, zu junge oder zu alte Personen, ferner Angehörige überdurchschnittlich mobiler Personengruppen, die, sei es aus Gründen des beruflichen oder des privaten Lebensstils, schwieriger an ihrer Wohnanschrift anzutreffen sind als andere, weniger mobile Personengruppen“.

Hinsichtlich dieser Personengruppen ist zu beachten, dass es sich hierbei zumindest teilweise um solche mit einem besonders hohen Viktimisierungsrisiko handelt, wie etwa Obdachlose, Flüchtlinge, in Heimen oder im Strafvollzug Untergebrachte. Erst in den letzten Jahren werden Viktimisierungen bei diesen Gruppen zumindest teilweise untersucht. So betont etwa Harth (2015, 1), in den Flüchtlingsheimen würden „Selbstjustiz und Hass“ vorherrschen.

„Verschiedenste Ethnien, Religionen und Kulturen sind immer noch im Krieg und werden hier auf engstem Raum zusammengepfercht“. Informationen hierüber gibt es bisher im Wesentlichen nur aus den Medien anhand von Einzelfällen. Untersuchungen zu „Hasskriminalität“ oder terroristischen Straftaten, liegen vor allem auch aufgrund der schweren Zugänglichkeit, bisher kaum vor (McDevitt/Williamson 2002; Arnold/Zoche 2014). Erst in den letzten Jahren ist etwa auch das Thema Gewalt an Polizeibeamtinnen und -beamten, Polizistinnen und Polizisten als Opfer, insbesondere durch Medienberichte, zu einem Thema geworden (Jager u. a. 2013).

Dass in Strafvollzugsanstalten, vor allem im Jugendstrafvollzug, Gewalt im Kontext einer besonderen Subkultur, einer vorherrschenden Prisonisierung, weit verbreitet ist, wird seit Jahrzehnten diskutiert (vgl. Sykes 1958; Goffman 1977; Ortmann 2002, 198ff.), Daten zu entsprechenden Vorkommnissen in Deutschland liegen inzwischen aus einem Forschungsprojekt der Universität Köln vor (Ernst/Neubacher 2014; Neubacher 2014). Wolter und Häufle (2014) weisen auf das enorme Dunkelfeld von Gewalthandlungen im Jugendstrafvollzug hin, die in Gefangenenpersonalakten registrierten entsprechenden Vorfälle erfassen nur die Spitze eines Eisbergs (vgl. zu Problemen der Aktenanalyse Dölling 1984; Hermann 1988). Im Rahmen ihrer Befragung war es erstmals möglich, „in Akten registrierte Vorkommnisse über Gewalthandlungen mit selbstberichteten Angaben über Gewalthandlungen gegenüber Mithäftlingen zu vergleichen“ (280). Die Resultate des „Hell-Dunkelfeldabgleichs zeigen eine deutliche Diskrepanz zwischen registrierten und selbstberichteten Gewalthandlungen. Auf der Täterebene entspricht die Hell-Dunkelfeld-Relation 1:5,3 [...] auf der Fallebene sogar 1:6,5 [...]“ (280; vgl. zu internationalen Ergebnissen Neuman 1991; Dyson u. a. 1997, 135; Kury/Smartt 2002). Bereits frühere Studien aus Nordrhein-Westfalen (Wirth 2006), Sachsen (Hinz u. Hartenstein 2010) oder Hessen (Heinrich 2002) weisen ebenfalls auf ein erhebliches Ausmaß von Gewalt im (Jugend-)Strafvollzug hin.

Erst in den letzten Jahren gibt es aussagekräftige Untersuchungen zu weiteren, bisher kaum berücksichtigten Gruppen, wie sexuelle Viktimisierungen bei Frauen (Müller/Schröttle 2004) und Kindern (Stadler u. a. 2012) bzw. Opfersituationen bei alten Menschen (Görgen u. a. 2002). Gerade auch alte Menschen werden aufgrund ihrer leichten Verletzbarkeit, insbesondere wenn sie pflegebedürftig sind und etwa in Altersheimen bzw. vergleichbaren Einrichtungen leben, aber auch in der eigenen Familie, relativ häufig Opfer von Misshandlungen bzw. Vernachlässigung, ein Thema, das in Zusammenhang mit dem ansteigenden Durchschnittsalter der Bevölkerung immer mehr in den Vordergrund rückte und zur Durchführung von Untersuchungen beitrug (Ahlf 1994; Heisler 2007). Vor diesem Hintergrund hat hier das Problem der Misshandlung in den letzten Jahren zugenommen, da die meisten älteren

Menschen in Familien gepflegt werden, liegen auch hier international die meisten Probleme (Fattah/Sacco 1989). Die Erreichbarkeit dieser Gruppe, etwa über Opferbefragungen, ist teilweise ausgesprochen schwierig, vielfach nur über die Angehörigen möglich, die unter Umständen jedoch die Täterinnen und Täter sind. Opfer in kirchlichen Einrichtungen wurden erst in den letzten Jahren ein Thema, Veröffentlichungen liegen bisher kaum vor (John Jay College 2006; Terry u. Smith 2006; Terry 2008; Erzdiözese Freiburg 2014), gegenwärtig werden auch in Deutschland, etwa im Auftrag der Deutschen Bischofskonferenz, umfangreichere Projekte zu dem Bereich durchgeführt (vgl. Deutsche Bischofskonferenz 2014).

Ein weiterer relativ großer Bereich von Viktimisierungen, der bisher nur ansatzweise untersucht wurde, sind etwa Geschäftsbetriebe (Heinz 2006, 246), obwohl gerade auch hier Schäden durch Kriminalität, wie erste Studien zeigen, erheblich sind. „Business are the driving force for economic development. Amongst the factors determining investment climate and private sector developments, a company's exposure to crime plays a significant role. Crime may cause high costs and damage to business. As a consequence, it may seriously hamper their activities" (United Nations Office on Drugs and Crime 2015, 1). In Anlehnung an die International Crime Victim Survey – ICVS führte Alvazzi del Frate (2004) in neun zentral- und osteuropäischen Hauptstädten eine Befragung bei Geschäftsleuten durch, wobei vor allem Korruption, Betrug und Erpressung erfasst wurden (vgl. a. Van Dijk/Terlouw 1996; Aromaa/Lehti 1996; Taylor/Mayhew 2002). Nieuwbeerta u. a. (2002, 172) fanden in ihrer Befragung in Industrie- und Entwicklungsländern enorme Unterschiede hinsichtlich Viktimisierungen durch Korruption mit einem Durchschnitt von etwa 20 % an Opfern in Asien und Lateinamerika und 10 % bis 15 % in Ländern Zentral- und Osteuropas. Solche Angaben sind allerdings mit großer Vorsicht zu betrachten, da etwa Täterinnen bzw. Täter und Opfer nicht immer klar voneinander zu trennen sind. Der Schaden durch Schwarzarbeit bzw. Schattenwirtschaft wird in Deutschland auf 12,2 % des Bruttoinlandprodukts geschätzt, was zu enormen wirtschaftlichen Schäden führt (ZEIT Online 2015). Untersuchungen sind auch hier ausgesprochen schwierig, da weder Täterinnen bzw. Täter noch Opfer an einer Offenlegung interessiert sein dürften.

Während in diesen Bereichen Umfragen grundsätzlich noch möglich und sinnvoll sind, wenn teilweise auch mit erheblichen Schwierigkeiten, etwa was die Erreichbarkeit der Betroffenen bzw. deren Mitarbeitsbereitschaft betrifft, stößt man auf weiteren Gebieten schnell an Grenzen der Anwendbarkeit von Victim Surveys. Opferstudien sind nur dann möglich, wenn die Opfer ihre Viktimisierungen selbst wahrgenommen haben und als Straftat richtig zuordnen. Je schwerwiegender eine Straftat wird, vor allem im Bereich Korruption, Drogenhandel, Menschenhandel, Handel mit Schusswaffen, Straftaten im

Spitzensport, Abrechnungsbetrug bei Ärzten, Cybercrime, Steuervergehen, Geldwäsche oder Staatskriminalität, um nur wenige Beispiele zu nennen, je mächtiger damit vielfach auch die Täterinnen und Täter werden, umso schwieriger wird es, durch Umfragen Licht in das Dunkel zu bringen (Adler u. a. 2007). Vielfach spielen hier die freien Medien eine Rolle, denen mehr oder weniger geheim Informationen zugespielt werden, die zur Aufdeckung solcher Taten – oft nach Jahren, führen. So überlegen inzwischen teilweise Betriebe, ein System von „Whistleblowing" einzuführen, um betriebsschädigende Straftaten besser aufdecken und unterbinden zu können. Opferstudien stoßen hier an ihre Grenzen.

Wenn etwa sowohl Täterinnen oder Täter als auch Opfer letztlich von Straftaten profitieren, dürften sie wenig zu einer Mitarbeit motiviert sein bzw. valide Angaben zu machen. Bereits bei „üblichen" Umfragen zeigt sich, dass bei Crime Surveys Befragte, die am wenigsten mitarbeiten, in aller Regel gleichzeitig am meisten belastet sind (Killias 2002, 61). Bei der Cambridge-Studie hat sich etwa die Erhöhung der Ausschöpfungsquote von 76 % auf 94 % massiv auf die Ergebnisse ausgewirkt (Farrington u. a. 1990, 136, 142).

2.3 Methodenprobleme – Erhebungsinstrument, Befragungsmethode

Während zu Beginn der Opferforschung Methodenprobleme vielfach wenig beachtet wurden, die Erhebungsinstrumente oft aus einer Ansammlung von Fragen zu „interessierenden" Bereichen bestanden, deren Einflüsse auf das Antwortverhalten der Befragten kaum überprüft wurden, repräsentative Stichproben kaum gewählt wurden und teilweise hohe Ausfälle zu verzeichnen waren, hat sich die Methodologie der Umfragen vor dem Hintergrund entsprechender Kritik inzwischen deutlich verfeinert, wenngleich nach wie vor Fragen und Probleme offen bleiben.

So konnte etwa in inzwischen vorliegenden Studien ein deutlicher Nachweis eines möglichen Einflusses des Erhebungsinstruments auf das Antwortverhalten der Rezipienten gezeigt werden. In der Psychologie, die bei Untersuchungen vielfach mit standardisierten (Persönlichkeits-)Fragebogen arbeitet, liegen seit Jahrzehnten zahlreiche Studien über den Einfluss der Fragebogengestaltung auf die gefundenen Ergebnisse vor (vgl. Fahrenberg u. a. 1978, 62ff; Kury 2002). Vielfach enthalten Persönlichkeitsfragebogen, teilweise auch Opferfragebogen, sogenannte „Lügenitems", welche Verfälschungstendenzen, etwa im Sinne der sozialen Erwünschtheit, aufdecken sollen, allerdings sind diese oft ebenfalls leicht zu verfälschen. Hoeth/Köbler (1967) versuchten, durch Zusatzinformationen, die darüber informierten, man würde „frisierte" Antworten erkennen, die Befragten zu ehrlichen Antworten zu motivieren, allerdings ohne wesentlichen Erfolg.

Im Rahmen einer Opferstudie in Jena (Kury u. a. 2000) wurde die per Zufall erfasste Stichprobe von 4.000 Personen ab 14 Jahren in einem zweiten Schritt wiederum per Zufall in zwei Unterstichproben aufgeteilt: 3.000 Personen bekamen das vollstandardisierte Erhebungsinstrument postalisch zugesandt, die zweite Gruppe von 1.000 Personen wurde in derselben Zeit von speziell hierfür ausgebildeten und trainierten Interviewerinnen und Interviewern anhand desselben Fragebogens – einschließlich des Freiburger Persönlichkeitsinventars (FPI) – persönlich befragt. Der Rücklauf betrug bei der schriftlichen Befragung 48,9 %, bei der mündlichen Datenerhebung dagegen 57,8 %, lag damit ca. 9 % höher. Erwartungsgemäß war somit – wie immer wieder festgestellt – die Ausschöpfungsquote bei der mündlichen Befragung höher als bei der schriftlichen. Die Ergebnisse zeigten deutliche Unterschiede zwischen den beiden Datenerhebungsmethoden derart, dass die postalisch Befragten weniger sozial erwünschte Antworten als persönlich Befragte gaben, sie schilderten sich somit offensichtlich ehrlicher, was sich auch auf die Antworten hinsichtlich Viktimisierungen auswirken dürfte (vgl. Kury 1983a, 1983b, 1994; Kury/Würger 1993; Rushton/Christjohn 1981; Kaiser 1996, 395).

Walker (2006, 3) berichtet über Erfahrungen aus der British Crime Survey (BCS) mit unterschiedlichen Erhebungsmethoden, die in dieselbe Richtung deuten. „The BCS includes a combination of face to face and self completion for more sensitive topics. These include drug use and domestic violence and sexual assault. [...] The self-completion module of the 2001 BCS produces substantially higher estimates than the main face-to-face BCS. A broad comparison between the prevalence measures (per cent victim once or more) shows that the self-completion finds a rate of approximately 5 times that of the face-to-face BCS“ (vgl. a. Müller/Schröttle 2004). Wie der Autor weiter berichtet, hat Schottland, um die Stichprobe erhöhen zu können und um so besser Regionalanalysen zu ermöglichen, die Datenerhebungsmethode von schriftlicher Face-to-face-Befragung in telefonische Befragungen geändert. „Calibration work has shown a marked difference between the victimisation levels found by both methods.“ Bei der Telefonsurvey waren die Opferraten größer. „One of the conclusions is that there is self selection bias. This results mainly from a greater proportion of non-victims refusing on the phone than in the face to face situation.“ Farrall u. a. (1997) fanden einen erheblichen Einfluss des Erhebungsmodus auf die Umfrageergebnisse hinsichtlich Verbrechensfurcht, Kury u. a. (2004) diskutieren Probleme der Definition und Messung hinsichtlich Sanktionseinstellungen (Punitivität).

Einen ausgesprochen deutlichen Hinweis auf einen Einfluss der Fragebogengestaltung auf die gefundenen Ergebnisse in Opferstudien zeigt sich in einer Methodenstudie zu der Hamburger Untersuchung von Sessar (1992) zum Thema Wiedergutmachung statt Strafe. Hierbei wurden Items aus der Studie,

in denen Straftaten vorgegeben wurden, auf die mit jeweils fünf Antwortmöglichkeiten unterschiedlicher Strafhärte (von „private Aussöhnung“ bis „Strafe ohne Anrechnung einer Entschädigung“) in eine Freiburger Untersuchung übernommen (Kury 1995). Die Freiburger Stichprobe wurde per Zufall in drei Untergruppen aufgeteilt: eine bekam die Hamburger Version der Items, bei der zweiten wurden die Antwortalternativen in umgekehrter Reihenfolge vorgegeben, bei der dritten wurden zu den fünf ursprünglichen Antwortalternativen drei weitere hinzugefügt. Die Ergebnisse zeigen, dass durch die Fragebogenveränderungen deutlich andere Ergebnisse erzielt wurden, die letztlich kaum noch mit den ursprünglichen Studienergebnissen vergleichbar waren, was den enormen Einfluss der Gestaltung eines Erhebungsinstruments auf die damit erzielten Resultate deutlich macht (vgl. a. Amelang 1986, 105f.; Amelang/Rodel 1970; Kury 1994a, 1994b).

Wie bereits angedeutet, kann die Art, wie die Daten erhoben wurden, ebenfalls einen Einfluss auf die Resultate haben. Walker (2006, 2) fand in seiner 2005 durchgeführten internationalen Umfrage in 33 Ländern, bei welcher er Informationen zu 78 Surveys sammeln konnte, dass als Datenerhebungsmethode vor allem eingesetzt wurden:

– Face to face Interviews mit schriftlichem Fragebogen,

– Face to face Interviews mit elektronischem Fragebogen (CAPI),

– selbst ausgefüllte Fragebogen (CASI),

– postalische Umfragen,

– telefonische Befragungen (CATI),

– über das Internet organisierte Surveys,

– eine Kombination unterschiedlicher Vorgehensweisen bzw.

– andere Erhebungsmethoden.

Immerhin 41 Surveys wurden durch Telefonbefragungen durchgeführt. Zusammenfassend stellt der Autor fest (S. 5): „The only marked difference between the overall situation and the ‚current‘ surveys is the move away from paper but not from face to face.“

Bereits Scheuch (1967, 136) hat das persönliche Interview als „das wichtigste Instrument der Sozialforschung“ bezeichnet, stuft dagegen die „Anwendungsbreite schriftlicher Befragungen“ als „beschränkter“ ein, so fehle die Kontrol-

le über die Situation der Befragung, die Interviewerin bzw. der Interviewer könne bei auftauchenden Unklarheiten nicht helfen, die stimulierende Wirkung der Anwesenheit der Interviewerin bzw. des Interviewers falle weg. Er weist jedoch gleichzeitig bereits auf den Vorteil der geringeren Kosten schriftlicher Befragungen hin. König (1957, 27) betont: „Wenn es [...] methodischer Kontrolle unterliegt, wird das Interview in seinen verschiedenen Formen [...] immer der Königsweg der praktischen Sozialforschung bleiben." Allerdings ist als Nachteil persönlicher Interviews neben den erheblich höheren Kosten, die letztlich zu einem Rückgang zugunsten computerunterstützter Vorgehensweisen geführt haben, auch ein möglicher Einfluss der Interviewerin bzw. des Interviewers auf das Antwortverhalten zu beachten.

Die national und international teilweise deutlich unterschiedlichen Vorgehensweisen bei Opferbefragungen sowie bei der Stichprobenziehung und den Antwortquoten führen dazu, dass die einzelnen Studienergebnisse in aller Regel wenn überhaupt nur eingeschränkt miteinander vergleichbar sind. Das gilt vor allem für internationale, aber auch nationale Studien, was etwa für die Situation in Deutschland immer wieder angeführt wird. Im Zusammenhang mit der zunehmenden Grenzöffnung und Internationalisierung, einer wachsenden Politisierung von Kriminalpolitik, wird mehr und mehr gefordert, validere internationale Vergleiche hinsichtlich Kriminalitätsbelastung durchzuführen, wobei das Anliegen so alt ist wie die Kriminalstatistik. Zurecht wird betont, dass sich Polizeistatistiken, etwa aufgrund unterschiedlicher rechtlicher Regelungen und Registrierungsmodalitäten, nicht für internationale Vergleiche eignen (vgl. van Dijk 2009, 234). Sehr viel Hoffnung wird demgegenüber in Opferstudien gelegt (Newman 1999), vor allem auch der seit 1988 inzwischen mehrfach durchgeführten ICVS (van Dijk u. a. 1990; van Dijk 2009; Cantor/Lynch 2000) sowie der International Violence Against Women Survey – IVAWS (Ollus u. Nevala 2005). So betont etwa von Hofer (2009, 128): „Internationale Vergleiche ermöglichen ..., Kenntnisse und Einsichten zu erwerben, die einer rein nationalen Sichtweise verschlossen bleiben können." Der Autor weist jedoch zurecht gleichzeitig darauf hin, man frage sich nach wie vor, „ob die Daten denn auch wirklich vergleichbar und anwendbar sind. Die Diskussion darüber schwankt gewöhnlich zwischen Optimismus und Pessimismus" (vgl. auch Young 2005).

So werden im Zusammenhang mit dem zunehmenden politischen Interesse an internationalen Kriminalitätsvergleichen vor allem Bereiche genannt, die für Opferbefragungen kaum zugänglich sind, wie Korruption, Menschen- und Drogenhandel, Terrorismus, Geldwäsche oder Wirtschaftskriminalität (Albrecht 2007, 257). Wenn von Hofer (2009, 127) als Vorteil internationaler Studien betont, dass die dadurch möglichen Vergleiche weit mehr ermöglichen als eine Antwort auf die Frage „ob die Kriminalitätsraten in verschiedenen Ländern höher oder niedriger sind. Verglichen werden kann, in welchem

Ausmaß Bevölkerungen über Kriminalität berichten und deren Einstellungen zu Fragen von Kriminalität, Sanktionen und Behörden“, so werden hier sensible Einstellungsbereiche angesprochen, die Kenntnisse über die Lage in einem Land erfordern, um die Ergebnisse aussagekräftig einordnen zu können. So berichtet beispielsweise Krajewski (2014) über ein punitiveres Klima in den früheren Ostblockländern oder Yoshida (2006, 169) über unterschiedliche Sichtweisen der japanischen Bevölkerung über Gewalt in der Familie. Beides dürfte sich jeweils auch spezifisch auf erfasste Ergebnisse von Umfragen auswirken.

3 Entwicklungsmöglichkeiten

Die heutige kriminologische Diskussion ist ohne die Ergebnisse von Opferstudien und deren inzwischen vorliegenden umfangreichen Ergebnissen, trotz vieler Widersprüche und Unklarheiten, nicht mehr denkbar. Heinz (2006, 263) betont in diesem Zusammenhang: „Trotz aller Vorbehalte gibt es keine Alternative zu Dunkelfeldforschungen als notwendige und unverzichtbare Ergänzung der amtlichen Kriminalstatistiken.“ Umso wichtiger ist es, den Forschungsansatz weiterzuentwickeln. Was die Forschungssituation in Deutschland betrifft, wird weitgehend übereinstimmend und zurecht immer wieder auf die Einführung einer einheitlichen regelmäßigen bundesweiten Opferstudie, wie sie etwa in den USA, Großbritannien oder den Niederlanden seit Jahren durchgeführt werden, hingewiesen. Heinz (2006, 263) drückte vor dem Hintergrund der damaligen Entwicklung noch die Hoffnung aus, dass die Durchführung solcher Opferstudien „in Deutschland inzwischen in greifbare Nähe gerückt“ sei, die Entwicklung zeigt, dass er sich offensichtlich getäuscht hat. Offensichtlich besitzen solche Studien kriminalpolitisch kein besonders hohes Gewicht. Die Bedeutung aussagekräftiger Opferstudien kann auch darin gesehen werden, dass bisherige Untersuchungen gerade vor dem Hintergrund von Informationen, „die durch die Daten der amtlichen Kriminal- und Strafrechtspflegestatistiken weder gewonnen werden noch werden können“ einen wesentlichen und konstruktiven Einfluss auf kriminalpolitische Entwicklungen hatten und haben, etwa was Fragen des Opferschutzes bzw. alternative Sanktionen betrifft (Heinz 2006, 245).

Eine kontinuierlich durchgeführte Opferbefragung und in diesem Zusammenhang Längsschnittstudien können wesentliche Informationen zu Bedingungen und Hintergründen von straffälligem Verhalten, insbesondere aber zu den Auswirkungen auf die Opfer geben (Boers 2009, 578). Standardisierte Umfragen können mit spezifischen Modulen zu aktuellen Fragestellungen, die im Laufe der Zeit wechseln können, ergänzt werden. Da die Erforschung der Geschehensabläufe und Auswirkungen straffälligen Verhaltens auf die Opfer kriminalpräventiv besonders wichtig ist, können durch Zusatzerhebungen etwa

Informationen aus dem Umfeld erhoben werden (Schneider 2007a, 322). Die Entwicklung von Kriminalität über das Lebensalter ist ohne Dunkelfeldbefragungen nicht aussagekräftig möglich (Farrington 2007). Wir wissen bisher viel über Einzelfaktoren, die kriminelles Verhalten etwa bei Kindern und Jugendlichen begünstigen, „however, the causal mechanisms linking these risk factors with antisocial outcomes are less well established“ (Farrington 2007, 199).

Ein wesentliches Thema zukünftiger Forschung, das besonderer Sensibilität hinsichtlich des methodischen Vorgehens bedarf, auch erhebliche kriminalpolitische Fragen aufwirft, ist die Erfassung der Auswirkungen des strafrechtlichen Umgangs mit Opfern. Bisherige Studien zeigen deutlich, dass sich die Opfer verständlicherweise vielfach wenig effektiv behandelt fühlen. So betont etwa Schneider (2007b, 409): „Durch Instrumentalisierung in einem täterorientierten Ermittlungs- und Strafverfahren macht die Kriminaljustiz das Opfer zum zweiten Mal zum Objekt. Zusätzlich zu dieser grundsätzlichen Sekundär-Viktimisierung kann die Kriminaljustiz das Opfer durch formalistische Routine und Gleichgültigkeit schädigen.“ Nach van Dijk (1999, 39) werden vor allem Frauen als Gewaltopfer international weniger rücksichtsvoll vom Kriminaljustizsystem behandelt. Das Opfer will im Strafverfahren eine Teilnehmerin bzw. ein Teilnehmer sein und anerkannt werden (Kilchling 1995).

Vor allem sollten auch Studiendesigns für besonders schwer erreichbare Gruppen, wie Ausländerinnen und Ausländer, Gefängnisinsassen, alte Menschen, etwa in Heimen, Obdachlose, Opfer von Wirtschaftsstraftaten um nur einige zu nennen, entwickelt werden. Die Erhebungsinstrumente sollten hinsichtlich ihres Einflusses auf die gefundenen Ergebnisse gründlicher überprüft werden, etwa auch der Einfluss des Kontextes, in welchem solche Studien stattfinden (vgl. etwa die Ergebnisse zur Verbrechensfurcht aus dem R+V-Infocenter 2015 im Vergleich zu entsprechenden Ergebnissen aus Opferstudien). Um einen vertieften Einblick in Opfererfahrungen und deren Auswirkungen zu erhalten, wird es nötig sein, standardisierte Erhebungen durch qualitative Interviews zu ergänzen, wie das teilweise bisher bereits erfolgte. So betont etwa Steffen (2013, 56): „Mehr quantitativ wie auch qualitativ orientierte Opferbefragungen – nicht nur zur Opferwerdung und zur Anzeigebereitschaft, sondern auch zu den Folgen der Viktimisierung, zu den Erwartungen und Wünschen der Opfer an die Hilfesysteme und die Strafrechtspflege, sind dringend erforderlich.“ Auch Greve u. a. (1994, 7) betonen die Bedeutung einer Erfassung der subjektiven Opfererfahrungen, nur sie „sollten den Anstoß und den Ausschlag geben für die Verarbeitungs- und Bewältigungsprozesse, denen eine zukünftige Opferforschung besondere Aufmerksamkeit zu widmen hat.“

4 Schluss

- International besteht weitgehend Einigkeit darüber, dass die Opferforschung in Form von Befragungen trotz aller teilweise nach wie vor bestehenden methodischen Probleme einen erheblichen Gewinn für Kriminologie und Kriminalpolitik gebracht hat und auch weiterhin bringen wird, nicht nur hinsichtlich des Kriminalitätsaufkommens, des Dunkelfeldes, sondern vor allem auch in Bezug auf Bereiche wie Opferschäden, Verbrechensfurcht, Vorstellungen und Wünsche der Opfer was Kriminalsanktionen und Hilfsmaßnahmen betrifft.

- Es steht „nicht mehr die Aufhellung des Dunkelfeldes, das ‚wahre Ausmaß' der Kriminalität, die Ermittlung der ‚Kriminalitätswirklichkeit' – was ohnehin nicht möglich ist – im Mittelpunkt des Forschungsinteresses, sondern das Opfer und die Folgen der Opferwerdung selbst" (Steffen 2013, 73; Heinz 2006, 245).

- Heute können wir hinsichtlich einer Weiterentwicklung von Kriminologie und Viktimologie auf Opferbefragungen nicht mehr verzichten. Deshalb ist es umso wichtiger, die Methodologie systematisch weiterzuentwickeln, etwa auch bisher nicht erfasste Bereiche und Gruppen mit einzubeziehen.

- Die Kritik, die Opferbefragungen hätten die Konzentration auf die „Straßenkriminalität" zu sehr forciert, Straftaten des „kleinen Mannes" in den Vordergrund gerückt, damit von der wesentlich gesellschaftsschädlicheren „Großkriminalität", etwa Korruption, Wirtschafts- oder politischen Straftaten, abgelenkt, nicht von der Hand zu weisen. Teilweise sind solche Kriminalitätsbereiche für Opferbefragungen kaum oder nicht zugänglich, teilweise können hier allerdings noch Fortschritte erzielt werden.

- Auch der Vorwurf, die Opferforschung habe zum Teil Opferbelange einseitig zu sehr betont, etwa was das Strafverfahren betrifft, ist zumindest für den Beginn in Teilen richtig, wobei allerdings zu berücksichtigen ist, dass Anliegen der Opfer bis zum Aufblühen einer Viktimologie auch kaum berücksichtigt wurden, das Opfer im Strafprozess auch heute noch eine randständige Rolle spielt, sich Alternativen gegenüber den klassischen Sanktionen nur mühsam durchsetzen, obwohl sie, wie die Forschung deutlich zeigen konnte, vielfach einen erheblichen Beitrag zur Wiederherstellung des Rechtsfriedens leisten können.

– Wieweit die Strafrechtspflege den Bedürfnissen von Opfern von Straftaten überhaupt gerecht werden kann, welche Veränderungen etwa nötig sind, bedarf weiterer differenzierter Forschung, in diesem Zusammenhang auch hinsichtlich der Rolle des sozialen Umfeldes.

– Bei Opferbefragungen hat sich der Fokus im Laufe der Entwicklung mehr und mehr von der bloßen Erfassung des Vorkommens von Straftaten wegentwickelt hin zu den Auswirkungen und Folgen einer Viktimisierung, den Vorstellungen und Wünschen hinsichtlich einer Reduzierung der Schäden. Dadurch ergaben sich auch neue forschungsmethodische Probleme. Nach Steffen (2013, 69) ist es auch heute noch „bemerkenswert, wie gering das empirisch gesicherte Wissen über die Opfer von Straftaten ist".

– Opferbefragungen werden auch in Zukunft, gerade durch eine Weiterentwicklung der Methodologie, ein wichtiges und unverzichtbares Korrektiv für die offiziellen Kriminalstatistiken sein, vor allem aber auch wesentliche zusätzliche Informationen zu einem konstruktiveren Umgang mit Kriminalität liefern.

5 Literatur

Adler, Freda; Mueller, Gerhard O. und Laufer, William S. (2007): Criminology and the Criminal Justice System, 6. Aufl., Boston, Burr, Ridge/IL: Dubuque.

Ahlf, Ernst H. (1994): Alte Menschen als Opfer von Gewaltkriminalität. In: Zeitschrift für Gerontologie, 27, S. 289–298.

Albrecht, Hans-Jörg (2007): Vergleichende Kriminologie. In: Schneider, Hans J. (Hg.): Internationales Handbuch der Kriminologie. Band 1: Grundlagen der Kriminologie. Berlin: De Gruyter, S. 255–288.

Alvazzi Del Frate, Anna (2004): The International Crime Business Survey: Findings from Nine Central-Eastern European Cities. In: European Journal on Criminal Policy and Research, 10, S. 137–161.

Amelang, Manfred (1986): Sozial abweichendes Verhalten. Entstehung – Verbreitung – Verhinderung. Berlin u. a.: Springer.

Amelang, Manfred; Rodel, Gerd (1970): Persönlichkeits- und Einstellungskorrelate krimineller Verhaltensweisen. Eine Untersuchung zur Dunkelziffer strafbarer Handlungen. In: Psychologische Rundschau, 21, 157–179.

Arnold, Harald; Zoche, Peter (Hg.)(2014): Terrorismus und Organisierte Kriminalität. Theoretische und methodische Aspekte komplexer Kriminalität. Berlin u. a.: LIT Verlag.

Aromaa, Kanko; Lehti, Martti (1996): Foreign Companies and Crime in Eastern Europe. Helsinki: National Research Institute of Legal Policy, Publication 135.

Barton, Stephan (2012): Strafrechtspflege und Kriminalpolitik in der viktimären Gesellschaft. In: Barton, Stephan; Kölbel, Ralf (Hg.): Ambivalenzen in der Opferzuwendung des Strafrechts. Baden-Baden, S. 111–137.

Baurmann, Michael C. (2000): Opferbedürfnisse, Mitschuldgefühl und Strafbedürfnis sowie die Erwartungen von Kriminalitätsopfern an Politik, Justiz und Polizei. In: Deutsches Polizeiblatt, 2, S. 2–5.

Baurmann, Michael C.; Schädler, Wolfram (1999): Das Opfer nach der Straftat – seine Erwartungen und Perspektiven. Eine Befragung von Betroffenen zu Opferschutz und Opferunterstützung sowie ein Bericht über vergleichbare Untersuchungen. Wiesbaden: Bundeskriminalamt.

Boers, Klaus (2009): Die kriminologische Verlaufsforschung. In: Schneider, Hans J. (Hg.): Internationales Handbuch der Kriminologie. Band 2: Besondere Probleme der Kriminologie. Berlin: De Gruyter, S. 577–616.

Bundesministerium des Innern; Bundesministerium der Justiz (Hg.)(2001): Erster Periodischer Sicherheitsbericht. Berlin.

Bundesministerium des Innern; Bundesministerum der Justiz (Hrsg.)(2006): Zweiter Periodischer Sicherheitsbericht. Berlin.

Cantor, David; Lynch, James P. (2000): Self-Report Surveys as Measures of Crime and Criminal Victimization. National Institute of Justice: Criminal Justice 2000. Band 4, S. 85–138.

Deutsche Bischofskonferenz (2014). Interdisziplinäres Forschungskonsortium führt Studie zum Thema „Sexueller Missbrauch an Minderjährigen“ durch. Bonn. URL: http://www.dbk.de/presse/details/?suchbegriff=Studie%20Sexueller%20Missbrauch&presseid=2517&cHash=4c5570e728fa938d07b1cfe974362d47 – Download vom 11. 07. 2015.

Dölling, Dieter (1984): Probleme der Aktenanalyse in der Kriminologie. In: Kury, H., (Hg.): Methodologische Probleme in der kriminologischen Forschungspraxis. Köln, S. 265–286.

Dussich, John P.J. (1991): Some Theoretical and Pragmatic Observations on the Abuse of Power. In: Kaiser, Günther; Kury, H. und Albrecht, Hans-Jörg (Hg.): Victims and Criminal Justice. Particular Groups of Victims. Part 2. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht, S. 677–688.

Dyson, Graham P.; Power, Kevin G. und Wozniak, Edward (1997): Problems with using official records from Young Offender Institutions as indexes of bullying. In: International Journal of Offender Therapy and Comparative Criminology, 41, 121–138.

Ernst, André; Neubacher, Frank (2014): Kontinuität oder Diskontinuität? – Was erklärt Gewaltverhalten im Jugendstrafvollzug. In: Niggli, Marcel A.; Marty, Lukas (Hg.): Risiken der Sicherheitsgesellschaft – Sicherheit, Risiko und Kriminalpolitik. In: Neue Kriminologische Schriftenreihe der Kriminologischen Gesellschaft, Band 115, Mönchengladbach: Forum Verlag Godesberg, S. 170–182.

Erzdiözese Freiburg (2014): Auswertung der Vorwürfe des sexuellen Missbrauchs und der körperlichen Gewalt in der Erzdiözese Freiburg von 1942 bis 31. Mai 2013. Freiburg: Erzdiözese.

Ewald, Uwe; Hennig, Carmen und Lautsch, Erwin (1994): Opfererleben in den neuen Bundesländern. In: Boers, Klaus; Ewald, Uwe; Kerner, Hans-Jörg; Lautsch, Eerwin und Sessar, Klaus (Hg.): Sozialer Umbruch und Kriminalität. Bd. 2: Ergebnisse einer Kriminalitätsbefragung in den neuen Bundesländern. Mönchengladbach: Forum Verlag, S. 75–170.

Fahrenberg, Jochen; Hampel, Rainer und Selg, Herbert (1978): Das Freiburger Persönlichkeitsinventar – FPI. Revidierte Fassung FRP-R und teilweise geänderte Fassung FPI-A1. Göttingen: Hogrefe.

Farrall, Stephen; Bannister, Jon; Ditton, Jason und Gilchrist, Elizabeth (1997): Questioning the measurement of the ‚fear of crime‘. In: British Journal of Criminology, 37, S. 658–679.

Farrington, D.P. (2007): Developmental and Life-Course Criminology. In: Schneider, H. J. (Hg.): Internationales Handbuch der Kriminologie. Band 1: Grundlagen der Kriminologie. Berlin: De Gruyter, S. 183–207.

Farrington, David P.; Gallagher, Bernhard; Morley, Lynda; Ledger, Raymond J. und West, Donald J. (1990): Minimizing Attrition in Longitudinal Research: Methods of Tracing and Securing Cooperatin in a 24-year Follow-up Study. In: Magnusson, David; Bergman, Lars R. (Hg.): Data Quality in Longitudinal Analysis. Cambridge: Cambridge University Press, S. 122–147.

Fattah, Ezzat A. (1991): Understanding criminal victimization. Scarborough: Prentice Hall.

Fattah, Ezzat A.; Sacco, Vincent F. (1989): Crime and victimization of the elderly. New York: Springer.

Feldmann-Hahn, Felix (2011): Opferbefragungen in Deutschland. Bestandsaufnahme und Bewertung. Holzkirchen/Obb.: Felix Verlag.

Görgen, Thomas, Kreuzer, Arthur, Nägele, Barbara, Krause, Sabine (2002). Gewalt gegen Ältere im persönlichen Nahraum. Wissenschaftliche Begleitung und Evaluation eines Modellprojektes. Schriftenreihe des Bundesministeriums für Familie, Senioren, Frauen und Jugend, Band 217. Stuttgart u. a.: Kohlhammer.

Görgen, Thomas (2012): Zum Stand der internationalen viktimologischen Forschung. In: Barton, Stephan; Kölbel, Ralf (Hg.): Ambivalenzen der Opferzuwendung des Strafrechts. Zwischenbilanz nach einem Vierteljahrhundert opferorientierter Strafrechtspolitik in Deutschland. Baden-Baden, S. 89–109.

Goffman, Erving (1977): Asyle. Über die soziale Situation psychiatrischer Patienten und Insassen. Frankfurt/M.

Gold, Martin (1970): Delinquent Behaviour in an American City. Belmont/Ca.

Greve, Werner; Strobl, Rainer und Wetzels, Peter (1994): Das Opfer kriminellen Handelns: Flüchtig und nicht zu fassen. Konzeptionelle Probleme und methodische Implikationen eines sozialwissenschaftlichen Opferbegriffes (=KFN Forschungsbericht Nr. 33). Hannover: Kriminologisches Forschungsinstitut Niedersachsen e. V..

Haas, Ute I. (2014): Das Kriminalitätsopfer. In: AK HochschullehrerInnen Kriminologie/Straffälligenhilfe in der Sozialen Arbeit (Hg.): Kriminologie und Soziale Arbeit. Ein Lehrbuch. Weinheim, Basel: Beltz Juventa, S. 242–262.

Harth, Peter (2015): Deutschlands Flüchtlingsproblem: Das Schweigen über die importierte Gewalt. Kopp Online. URL: http://info.kopp-verlag.de/hintergruende/deutschland/peter-harth/deutschlands-fluechtlingsproblem-das-schweigen-ueber-die-importierte-gewalt.html – Download vom 11. 07. 2015.

Heinrich, Wilfried (2002): Gewalt im Gefängnis – eine Untersuchung der Entwicklung von Gewalt im hessischen Justizvollzug (1989 – 1998). In: Bewährungshilfe, 49, S. 369–383.

Heinz, Wolfgang (2006): Zum Stand der Dunkelfeldforschung in Deutschland. In: Obergfell-Fuchs, Joachim; Brandenstein, Martin (Hg.): Nationale und internationale Entwicklungen in der Kriminologie. Festschrift für Helmut Kury zum 65. Geburtstag. Frankfurt/M: Verlag für Polizeiwissenschaft, S. 241–263.

Heisler, Candace J. (2007): Elder Abuse. In: Davis, Randy C.; Lurigio, Arthur J. und Herman, Susan (Hg.): Victims of Crime. Los Angeles, S. 161–188.

Hermann, Dieter (1988): Die Aktenanalyse als kriminologische Forschungsmethode. In: Kaiser, Günther; Kury, Helmut und Albrecht, Hans-Jörg (Hg.): Kriminologische Forschung in den 80er Jahren. Freiburg, S. 863–877.

Hinz, Sylvette; Hartenstein, Sven (2010): Jugendgewalt im Strafvollzug. Eine retrospektive Untersuchung im sächsischen Jugendstrafvollzug. In: Zeitschrift für Jugendkriminalrecht und Jugendhilfe, 21, S. 176–182.

Hoeth, Friedrich; Köbler, Viktoria (1967): Zusatzinformationen gegen Verfälschungstendenzen bei der Beantwortung von Persönlichkeitsfragebogen. In: Diagnostica, 13, S. 117–130.

Jager, Janine; Klatt, Thimna; Bliesener, Thomas (2013): NRW-Studie Gewalt gegen Polizeibeamtinnen und Polizeibeamte. Die subjektive Sichtweise zur Betreuung und Fürsorge, Aus- und Fortbildung, Einsatznachbereitung, Belastung und Ausstattung. Kiel: Christian-Albrechts-Universität.

John Jay College of Criminal Justice (2006): The nature and Scope of the Problem of Sexual Abuse of Minors by Catholic Priests and Deacons in the United States 1950 – 2002. New York: John Jay College. URS: http://www.philvaz.com/ABUSE.PDF – Download vom 11. 07. 2015.

Kaiser, Günther (1996): Kriminologie. Ein Lehrbuch. Heidelberg: C.F. Müller.

Kaiser, Günther; Kury, Helmut und Albrecht, Hans-Jörg (Hg.)(1991): Victims and Criminal Justice. Particular Groups of Victims. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht.

Kilchling, Michael (1995): Opferinteressen und Strafverfolgung. Freiburg/Br.: Max-Planck-Institut für ausländisches und internationales Strafrecht.

Killias, Martin (2002): Grundriss der Kriminologie. Eine europäische Perspektive. Bern: Stämpfli.

Kinsey, Alfred C.; Pomeroy, Wardell B.; Martin, Clyde E. und Gebhard, Paul H. (1963): Das sexuelle Verhalten der Frau. Berlin; Frankfurt/M.: Fischer.

Kinsey, Alfred C.; Pomeroy, Wardell B. und Martin, Clyde E. (1964): Das sexuelle Verhalten des Mannes. Berlin; Frankfurt/M.: Fischer.

König, René (1957): Das Interview. Köln.

Krajewski, Krzysztof (2014): Different penal climates in Europe. Kriminologijos studijos. University of Vilnius, 1, S. 86–111.
Kunz, Franziska (2014): Kriminalität älterer Menschen. Beschreibung und Erklärung auf der Basis von Selbstberichtsdaten. Berlin: Duncker & Humblot.
Kürzinger, Josef (1996): Kriminologie. Stuttgart u. a.: Boorberg.
Kury, Helmut (1983a): Zur Verfälschbarkeit von Persönlichkeitsfragebogen bei jungen Strafgefangenen. In: Zeitschrift für Strafvollzug und Straffälligenhilfe, 32, S. 323–332.
Kury, Helmut (1983b): Verfälschungstendenzen bei Persönlichkeitsfragebogen im Strafvollzug. In: Monatsschrift für Kriminologie und Strafrechtsreform, 66, S. 72–74.
Kury, Helmut (1994a): The influence of the specific formulation of questions on the results of victim studies. In: European Journal on Criminal Policy and Research, 2–4, S. 48–68.
Kury, Helmut (1994b): Zum Einfluss der Art der Datenerhebung auf die Ergebnisse von Umfragen. In: Monatsschrift für Kriminologie und Strafrechtsreform, 77, S. 22–33.
Kury, Helmut (1995): Wie restitutiv eingestellt ist die Bevölkerung? Zum Einfluss der Frageformulierung auf die Ergebnisse von Opferstudien. In: Monatsschrift für Kriminologie und Strafrechtsreform, 78, S. 84–98.
Kury, Helmut (2001): Das Dunkelfeld der Kriminalität. Oder: Selektionsmechanismen und andere Verfälschungsstrukturen. In: Kriminalistik, 55, S. 74–84.
Kury, Helmut (2002): Das Freiburger Persönlichkeitsinventar und sein Einsatz bei kriminologischen Fragestellungen. Das Problem der Verfälschungstendenzen. In: Myrtek, Michael (Hg.): Die Person im biologischen und sozialen Kontext. Göttingen: Hogrefe, S. 249–270.
Kury, Helmut (2003): Wie werden Opfer von Straftaten gesehen? Zur Stigmatisierung von Verbrechensopfern. In: Lamnek, Siegfried; Boatca, Manuela (Hg.): Geschlecht – Gewalt – Gesellschaft. Opladen: Lese + Budrich, S. 418–443.
Kury, Helmut (2015): Physische und psychische Gewalt. In: Melzer, Wolfgang; Hermann, Dieter; Sandfuchs, Uwe; Schäfer, Mechthild; Schubarth, Wilfried und Daschner, Peter (Hg.): Handbuch Aggression, Gewalt und Kriminalität bei Kindern und Jugendlichen. Bad Heilbrunn: Klinkhardt, S. 162–168.
Kury, Helmut; Würger, Michael (1993): The influence of the type of data collection method on the results of the victim surveys. In: Alvazzi del Frate, Anna; Zvekic, Uglijesa und van Dijk, Jan J.M. (Hg.): Understanding crime. Experiences of crime and crime control. Rome: UNICRI, S. 137–152.

Kury, Helmut; Dörmann, Uwe; Richter, Harald und Würger, Michael (1996): Opfererfahrungen und Meinungen zur Inneren Sicherheit in Deutschland. Ein empirischer Vergleich von Viktimisierungen, Anzeigeverhalten und Sicherheitseinschätzung in Ost und West vor der Vereinigung. Wiesbaden: Bundeskriminalamt.

Kury, Helmut; Obergfell-Fuchs, Joachim und Würger, Michael (2000): Gemeinde und Kriminalität. Eine Untersuchung in Ost- und Westdeutschland. Freiburg im Br.: edition iuscrim.

Kury, Helmut; Smartt, Ursula (2002): Prisoner-on-prisoner violence: Victimization of young offenders in prison. Some German Findings. In: Criminal Justice Journal, 2, 4, S. 411–437.

Kury, Helmut; Kania, Harald und Obergfell-Fuchs, Joachim (2004): Worüber sprechen wir, wenn wir über Punitivität sprechen? Versuch einer konzeptionellen und empirischen Begriffsbestimmung. In: Kriminologisches Journal, 36, S. 51–88.

Lamnek, Siegfried; Luedtke, Jens (2006): Opfer elterlicher Gewalt – Opfer von Gewalt in der Schule. In: Obergfell-Fuchs, Joachim; Brandenstein, Martin (Hg.): Nationale und internationale Entwicklungen in der Kriminologie. Festschrift für Helmut Kury zum 65. Geburtstag. Frankfurt/M.: Verlag für Polizeiwissenschaft, S. 139–167.

Levine, James P. (1976): The Potential for Crime Overreporting in Criminal Victimization Surveys. In: Criminology, 14, S. 307–330.

McDevitt, Jack; Williamson, Jennifer (2002): Hate Crimes: Gewalt gegen Schwule, Lesben, bisexuelle und transsexuelle Opfer. In: Heitmeyer, Wilhelm; Hagan, John (Hg.): Internationales Handbuch der Gewaltforschung. Wiesbaden, S. 1000–1019.

Mendelsohn, Beniamin (1974): The Origin of the Doctrine of Victimology. In: Drapkin, Israel, Viano, Emilio (Hg.): Victimology. Lexington/Mass.. S. 3–11.

Mitscherlich, Margarete (1999): Der irrationale Umgang der Gesellschaft mit ihren Opfern. Frauen und Minderheiten als Opfer krimineller Gewalt. In: Baurmann, Michael C.; Schädler, Wolfram (Hg.): Das Opfer nach der Straftat – seine Erwartungen und Perspektiven. Eine Befragung von Betroffenen zu Opferschutz und Opferunterstützung sowie ein Bericht über vergleichbare Untersuchungen. Wiesbaden: Bundeskriminalamt. S. 211–223.

Müller, Ursula; Schröttle, Monika (2004): Lebenssituation, Sicherheit und Gesundheit von Frauen in Deutschland. Eine repräsentative Untersuchung zu Gewalt gegen Frauen in Deutschland. Berlin: Bundesministerium für Familie, Senioren, Frauen und Jugend.

Neubacher, Frank (2014): Gewalt im Jugendstrafvollzug – Ein Überblick über Ergebnisse des Kölner Forschungsprojekts. In: Forum Strafvollzug, 63, S. 320–326.

Neuman, Elias (1991): Victimization and Abuse of Power in the Latin American Prisons. In: Kaiser, Günther; Kury, Helmut und Albrecht, Hans-Jörg (Hg.): Victims and Criminal Justice. Particular Groups of Victims, Part 2. Freiburg: MPI, S. 747–762.

Newman, Graeme (Hg.)(1999): Global Report on Crime and Justice. New York, Oxford: Oxford University Press.

Niemi, Hannu (1991): A Victimological Approach to Insurance Fraud: An Example of Powerful Victims. In: Kaiser, Günther; Kury, Helmut und Albrecht, Hans-Jörg (Hg.): Victims and Criminal Justice. Particular Groups of Victims, Part 1. Freiburg: MPI, S. 205–229.

Nieuwbeerta, Paul; Geest, Gerrit de und Siegers, Jacques (2002): Corruption in industrialized and developing countries. A test of law & economics hypotheses. In: Nieuwbeerta, Paul (Hg.): Crime victimization in comparative perspective. Results from the International Crime Victims Survey, 1989-2000. Den Haag: Boom Juridische uitgevers, S. 163–182.

O'Brien, Robert M.; Shichor, David S. und Decker, David L. (1979): An Empirical Comparison of the Validity of UCR and NCS Crime Rates. San Bernardino.

Ollus, Natalia; Nevala, Sami (2005): Challenges of Surveying Violence Against Women: Development of Research Methods. In: Smeenk, Wilma; Malsch, Marijke (Hg.): Family Violence and Police Response. Learning from Research, Policy and Practice in European Countries. Aldershot: Ashgate Publishing, S. 9–34.

Ortmann, Rüdiger (2002): Sozialtherapie im Strafvollzug. Eine experimentelle Längsschnittstudie zu den Wirkungen von Strafvollzugsmaßnahmen auf Legal- und Sozialbewährung. Freiburg: edition iuscrim.

Porterfield, Austin L. (1946): Youth in Trouble. Forth Worth, Texas.

Reuband, Karl-Heinz (1989): On the Use of Self-Reports in measuring Crime among Adults. Methodological problems and prospects. In: Klein, Malcolm (Hg.): Cross-national Research of Self-Reported Crime and Delinquency. Dordrecht u. a., S. 89–106.

Richter, Harald (1997): Opfer krimineller Gewalttaten. Individuelle Folgen und ihre Verarbeitung. Mainz: Weisser Ring.

Rushton, J. Philippe; Chrisjohn, Roland D. (1981): Extraversion, neuroticism, psychoticism and self-reported delinquency: Evidence from eight separate samples. In: Personality and Individual Differences, 2, S. 11–20.

R+V-Infocenter (2015): Die Ängste der Deutschen 2014. URL: https://www.ruv.de/de/presse/r_v_infocenter/studien/aengste-der-deutschen.jsp – Download vom 11. 07. 2015.

Sautner, Lyane (2010): Opferinteressen und Strafrechtstheorien. Zugleich ein Beitrag zum restaurativen Umgang mit Straftaten. Viktimologie und Opferrechte (VOR) (=Schriftenreihe der Weisser Ring Forschungsgesellschaft, Band 6). Innsbruck.

Scheib, Klaus (2002): Die Dunkelziffer bei Tötungsdelikten aus kriminologischer und rechtsmedizinischer Sicht. Berlin: Logos Verlag.
Scheuch, Erwin K. (1967): Das Interview in der Sozialforschung. In: König, René (Hg.): Handbuch der Empirischen Sozialforschung. Stuttgart: Enke, Band I, S. 136–196.
Schneider, Hans J. (2007a): Kriminalitätsmessung: Kriminalstatistik und Dunkelfeldforschung. In: Ders. (Hg.): Internationales Handbuch der Kriminologie. Band I: Grundlagen der Kriminologie. Berlin: De Gruyter, S. 289–332.
Schneider, Hans J. (2007b): Viktimologie. In: Ders. (Hg.): Internationales Handbuch der Kriminologie. Band I: Grundlagen der Kriminologie. Berlin: De Gruyter, S. 395–433.
Schwind, Hans-Dieter (2013): Kriminologie. Eine praxisorientierte Einführung mit Beispielen. Heidelberg: Kriminalistik Verlag.
Sessar, Klaus (1992): Wiedergutmachen oder Strafen? Einstellungen in der Bevölkerung und der Justiz. Pfaffenweiler: Centaurus.
Sessar, Klaus (2012): Kriminalitätswirklichkeit im Licht des Dunkelfeldes. In: Hilgendorf, Eric; Rengier, Rudolf (Hg.): Festschrift für Wolfgang Heinz zum 70. Geburtstag. Baden-Baden: Nomos, S. 262–274.
Short, James F.; Nye, F. Ivan (1957): Reported Behavior as a Criterion of Deviant Behavior. In: Social Problems, 5, S. 207–213.
Sparks, Richard F. (1981): Surveys of victimization – An optimistic assessment. In: Tonry, Michael; Morris, Norval (Hg.): Crime and justice: An annual review of research. Band 3. Chicago: University of Chicago Press, S. 1–60.
Stadler, Lena; Bieneck, Steffen und Pfeiffer, Christian (2012): Repräsentativbefragung Sexueller Missbrauch 2011 (=Forschungsbericht Nr. 118). Hannover: Kriminologisches Forschungsinstitut Niedersachen e. V..
Steffen, Wiebke (2013): Opferzuwendung in Gesellschaft, Wissenschaft, Strafrechtspflege und Prävention: Stand, Probleme, Perspektiven. Gutachten für den 18. Deutschen Präventionstag. In: Marks, Erich; Steffen, Wiebke (Hg.): Mehr Prävention – weniger Opfer. Ausgewählte Beiträge des 18. Deutschen Präventionstages 22. und 23. April 2013 in Bielefeld. Forum Verlag Godesberg, S. 51–121.
Sykes, Gresham M. (1958): The society of captives. New Jersey: University Press.
Taylor, Natalie; Mayhew, Pat (2002): Patterns of Victimisation among Small Retail Businesses. Canberra: Trends & Issues in Crime and Criminal Justice No. 221, Australian Institute of Criminology.
Terry, Karen, Smith, Margaret Leland (2006). The Nature and Scope of Sexual Abuse of Minors by Catholic Priests and Deacons in the United States. Supplementary Data Analysis. New York: John Jay College of Criminal Justice. URL: http://www.usccb.org/issues-and-action/child-

and-youth-protection/upload/Nature-and-Scope-supplemental-data-2006.pdf – Download vom 11. 07. 2015.
Terry, Karen (2008). Introduction To the Special Issue. Criminal Justice and Behavior 35, 545-548.
Titus, Richard M. (1991): Criminal Fraud: The Hidden Crime. In: Kaiser, Günther; Kury, Helmut und Albrecht, Hans-Jörg (Hg.): Victims and Criminal Justice. Particular Groups of Victims. Part 1. Freiburg: Max-Planck-Institut für ausländisches und internationales Strafrecht, S. 195–203.
United Nations Office on Drugs and Crime – UNODC (2015): Crime and Corruption Business Surveys (CCBS). URL: https://www.unodc.org/unodc/en/data-and-analysis/Crime-and-Corruption-Business-Surveys.html – Download vom 11. 07. 2015.
van Dijk, Jan J.M. (1999): The Experience of Crime and Justice. In: Newman, G. (Hg.): Global Report on Crime and Justice. New York, Oxford, S. 25–42.
van Dijk, Jan J.M. (2009): Criminological Research in the Framework of the United Nations. In: Schneider, Hans J. (Hg.): Internationales Handbuch der Kriminologie. Band 2: Besondere Probleme der Kriminologie. Berlin: De Gruyter, S. 227–253.
van Dijk, Jan J.M.; Mayhew, Pat und Killias, Martin (1990): Experiences of Crime across the World. Key findings from the 1989 International Crime Survey. Deventer: Kluwer.
van Dijk, Jan J.M.; Terlouw, Gert J. (1996): An international perspective of the business community as victims of fraud and crime. In: Security Journal, 7, 3, S. 157–167.
Vito, Gennaro F.; Maahs, Jeffrey R. und Holmes, Ronald M. (2007): Criminology – Theory, Research, and Policy. Boston u. a.
Voß, Stephan (2003): „Du Opfer . . .!“. In: Berliner Forum Gewaltprävention, 12, S. 56–59.
von Hofer, Hanns (2009): Der internationale Kriminalitätsvergleich mit Hilfe der Statistik. In: Schneider, Hans J. (Hg.): Internationales Handbuch der Kriminologie. Band 2: Besondere Probleme der Kriminologie. Berlin: De Gruyter, S. 121–144.
Walker, Alison (2006): Victim Survey Methodology: Mode, Sample Design and other Aspects – Results from the Inventory of Victimisation Surveys. UN Statistical Commission and UN Economic Commission for Europe, United Nations Office on Drugs and Crime, Conference of European Statisticians, Working Paper No. 6. URL: http://www.unece.org/stats/documents/2006.01.crime.htm – Download vom 11. 07. 2015.
Wallerstein, James S.; Wyle, Clemet J. (1947): Our Law-Abiding Law-Breakers. Probation, 25, S. 107–112.

Walsh, Anthony; Ellis, Lee (2007): Criminology – An Interdisciplinary Approach. Thousand Oaks: London; New Delhi.

Wetzels, Peter (1996): Wider den naiven Realismus kriminologischer Opferforschung. Plädoyer für einen subjektiven, konstruktivistischen Opferbegriff. In: Ewald, Uwe (Hg.): Kulturvergleichende Kriminalitätsforschung und sozialer Wandel in Mittel- und Osteuropa. Bonn, S. 117–143.

Wirth, Wolfgang (2006): Gewalt unter Gefangenen. Kernbefunde einer empirischen Studie im Strafvollzug des Landes Nordrhein-Westfalen. Abschlussbericht. Düsseldorf.

Wolter, Daniel; Häufle, Jenny (2014): Wie aussagekräftig sind Gefangenenpersonalakten als Entscheidungshilfe im Strafvollzug? Ergebnisse eines Hell-Dunkelfeld-Vergleichs am Beispiel von Gewalt unter Inhaftierten im Jugendstrafvollzug. In: Monatsschrift für Kriminologie und Strafrechtsreform, 97, S. 280–293.

Yoshida, Toshio (2006): Gewalt gegen Frauen in der japanischen Familie. In: Obergfell-Fuchs, Joachim; Brandenstein, Martin (Hg.): Nationale und internationale Entwicklungen in der Kriminologie. Festschrift für Helmut Kury zum 65. Geburtstag. Frankfurt/M.: Verlag für Polizeiwissenschaft, S. 169–192.

Young, Peter (2005). The Use of National Crime Statistics in Comparative Research: Ireland and Scotland Compared. In: Sheptycki, James; Wardak, Ali (Hg.): Transnational and Comparative Criminology. London: GlassHouse Press.

ZEIT-Online (2015). Schattenwirtschaft besonders im Süden stark. http://www.zeit.de/wirtschaft/2015-02/schwarzarbeit-schattenwirtschaft-oecd-deutschland – Download vom 11. 07. 2015.

Zedner, Lucia (2002): Victims. In: Maguire, Mike; Morgan, Rod und Reiner, Robert (Hg.): The Oxford Handbook of Criminology. Oxford, S. 419–456.

5 Zusammenfassung und Implikationen für die Praxis

Zusammenfassung und Implikationen für die Praxis

Nathalie Guzy, Christoph Birkel und Robert Mischkowitz

Die vorliegenden Beiträge haben deutlich gemacht, welche immense Bedeutung die methodische Herangehensweise bei der Durchführung und Auswertung von Opferbefragungen hat und welche Vielzahl an Methodeneffekten bei der Interpretation ihrer Ergebnisse berücksichtigt werden muss.

Im Folgenden kann und soll es nicht darum gehen, die zentralen Erkenntnisse der einzelnen Beiträge zusammenzufassen. Hierfür wird auf die Spiegelstriche am Ende sämtlicher Beiträge verwiesen, die – mit Blick auf die breite Zielgruppe des Sammelbands – sowohl für Forschende, Anwender und Anwenderinnen als auch Praktiker und Praktikerinnen formuliert wurden. In dem vorliegenden abschließenden Kapitel soll es vielmehr darum gehen, die Erkenntnisse sowie den aktuellen Forschungsstand zur Methodik und Methodologie von Opferbefragungen insgesamt zu reflektieren und hinsichtlich ihrer praktischen Bedeutung zu diskutieren.

Nach genauem Studium dieses Sammelbands überrascht es nicht, wenn der oder die eine oder andere interessierte Leser bzw. Leserin zu dem Ergebnis kommt, dass Opferbefragungen (oder Umfragen insgesamt) aufgrund der zahlreichen methodischen Einflüsse kaum zu validen und reliablen Ergebnissen ohne methodische Artefakte führen.[1] Dieser Eindruck ist zu einem bestimmten Grad (zumindest für einzelne Fragen) leider auch zutreffend. Allerdings muss gleichzeitig bedacht werden, dass Befragungen einmalige und einzigartige Informationsquellen darstellen, um systematisiert und auf Basis von (zumindest weitgehend repräsentativen) Bevölkerungsstichproben Daten über soziale Sachverhalte zu generieren, über die sonst meist *keinerlei* Informationen vorliegen. Dazu gehören bspw. Informationen über Einstellungen, Werte oder eben auch bestimmte Erfahrungen, über die in der Regel lediglich

[1] *Reliabilität und Validität* stellen zentrale Gütekriterien einer Messung dar. Unter *Reliabilität* versteht man die Zuverlässigkeit und Stabilität eines Messinstruments. Dabei wird gefordert, dass die Messergebnisse bei wiederholter Messung reproduzierbar sind, d. h., derselbe Befragte zu verschiedenen Messzeitpunkten identisch antworten (wenn sich das interessierende Merkmal nicht verändert hat) bzw. verschiedene Befragte mit gleichen Eigenschaften ähnlich antworten sollten. Unter *Validität* wird das Ausmaß verstanden, in dem das Messinstrument tatsächlich das misst, was es messen sollte. Die Validität einer Messung bezieht sich darauf, wie gut die Antwort auf eine Frage mit dem wahren Wert korrespondiert (Schnell u. a. 2005; Groves u. a. 2009). Unter *Messartefakten* versteht man scheinbar inhaltliche Ergebnisse, die auf Effekte der Methoden der Datenerhebung und/oder -auswertung zurückzuführen sind.

die Befragten selbst Auskunft geben können. Die Wissenschaft – und hier insbesondere auch die Kriminologie – haben in den letzten Jahrzehnten enorm von Ergebnissen der Umfrageforschung profitiert, was zu einer signifikanten Verbesserung des Verständnisses kriminologisch relevanter Prozesse geführt hat. Dazu gehören beispielsweise Erkenntnisse zu den Entstehungsbedingungen und Risikofaktoren von Kriminalität (vor allem im Zusammenhang mit der Überprüfung von Kriminalitätstheorien), zu längsschnittlichen Fragestellungen wie die Entwicklung kriminalitätsbezogener Einstellungen oder krimineller Karrieren, zur Verbreitung und Entstehung von Kriminalitätsfurcht und eben auch zur Verbreitung und Verteilung von Opfererfahrungen. Unterdessen ist auch zu bedenken, dass für zahlreiche sozialwissenschaftliche und/oder kriminologische Fragestellungen häufig keine zuverlässigeren und gültigeren Informationen als diejenigen, die über Befragungen gewonnen werden können, verfügbar sind.

Insbesondere Opferbefragungen, die nicht selten als thematisch breiter aufgestellte *Victim Surveys* konzipiert werden, stellen innerhalb der Kriminologie wohl eine der prominentesten Umfragen dar. Dabei dürfte insbesondere der erste Band dieses Sammelwerks deutlich gemacht haben, welche zentralen Erkenntnisse aus Opferbefragungen zu Kriminalitätsaufkommen (inkl. der Abschätzung von Hell-Dunkelfeld-Relationen), Sicherheitsgefühl, den Folgen von Opferwerdungen oder zu Polizeivertrauen gewonnen werden konnten.

Zusammenfassend scheint somit durchaus die Meinung vertretbar, dass Umfragen – bzw. die aus ihnen gewonnenen Informationen – trotz der verschiedenen „methodischen Empfindlichkeiten“ eine unerlässliche, weil einzigartige Datenquelle darstellen. Doch inwiefern kann man aus den in diesem Band detailliert ausgeführten Methodenproblemen und -effekten nun lernen und die Erkenntnisse praktisch nutzbar machen? Wie bereits in der Einleitung beschrieben lag die Intention dieses Sammelbands in zwei Bereichen: Die Leserinnen und Leser sollten einerseits eine umfängliche und fundierte Grundlage erhalten, um „gute“ Opferbefragungen selbst durchführen zu können, andererseits sollten sie in die Lage versetzt werden, Ergebnisse aus Opferbefragungen adäquat, d. h. vor dem Hintergrund ihrer methodischen Entstehungsgeschichte, zu interpretieren. Wie bereits angedeutet erscheint es aus methodologischer Sicht hilfreich, Opferbefragungen bzw. deren Ergebnisse als Resultat eines mehrstufigen Prozesses zu betrachten, in dessen Verlauf durch die Auswahl bestimmter Methoden jeweils gewisse Effekte auf die Ergebnisse wirksam werden (Groves u. a. 2009). Dabei erscheint es unerlässlich, dass sich sowohl Anwenderinnen und Anwender als auch Nutzerinnen und Nutzer von Opferbefragungen darüber im Klaren sind, ob und in welcher Form die der Befragung zugrunde liegenden Methoden die jeweilig interessierenden Ergebnisse beeinflussen. Nur auf dieser Basis können bspw. Anwenderinnen und Anwender (methodische) Probleme erkennen und das für die jeweilig interessierende

Forschungsfrage „optimale“ methodische Design auswählen. Dies kann freilich nur unter Abwägung verschiedener Methodeneffekte und unter Berücksichtigung der jeweiligen Fragestellung (und natürlich auch der zur Verfügung stehenden finanziellen und zeitlichen Mittel) geschehen. Eine vollständige Eliminierung von Methodeneffekten wird dabei vermutlich niemals möglich sein – gute Kenntnisse über die vorliegenden Effekte sind dagegen unverzichtbar, auch um in einem nächsten Schritt einschätzen zu können, ob und in welcher Form interessierende Ergebnisse inhaltlich interpretiert werden können.

Eine wertvolle Grundlage für diese Bewertung soll der vorliegende Sammelband liefern – auch wenn auf seiner Basis sicherlich nicht jede methodische Fragestellung zufriedenstellend beantwortet werden kann. Wenngleich für viele methodische Rahmenbedingungen zwar Effekte auf Umfrageergebnisse naheliegen (zumindest für Methodikerinnen und Methodiker oder methodisch interessierte Forscherinnen und Forscher bzw. Anwenderinnen und Anwender) – so z. B. dass durch das Ziehen einer Stichprobe auf Basis von Festnetznummern jüngere und somit in der Regel häufiger viktimisierte Personen unterrepräsentiert sind oder dass die Abfrage zur Kriminalitätsfurcht *nach* der Abfrage von Opfererlebnissen zu höheren Furchtanteilen führt – ist es der Komplexität und Wechselwirkung verschiedener methodischer Ansätze und inhaltlicher Fragestellungen geschuldet, dass für zahlreiche Erhebungs- und Messinstrumente (noch) keine zuverlässigen Aussagen über deren relevante Effekte auf inhaltlich interessierende Umfrageergebnisse vorhanden sind. Es versteht sich von selbst, dass hier die methodische und methodologische Forschung weiterhin gefordert ist.

In diesem Zusammenhang sei abschließend noch auf die Bedeutung von Replikationen einzelner Untersuchungsergebnisse, auch unter Verwendung diverser methodischer Vorgehensweisen verwiesen: So existieren Forschungsfragen, die sich jeweils durch einen ausgesprochen kontroversen Forschungsstand auszeichnen. Als Beispiele können z. B. die Befundlage zum Zusammenhang zwischen Viktimisierungserfahrungen und Kriminalitätsfurcht oder zu den verschiedenen Korrelaten von Strafeinstellungen genannt werden. Vieles spricht dafür, dass die unterschiedlichen methodischen Herangehensweisen der vorliegenden Beiträge eine zentrale Rolle spielen (Simonson 2011; Kury u. a. 2004; Hale 1996).[2] Derartige Studien bzw. deren Zusammenschau

[2] An dieser Stelle sei nur am Rande erwähnt, dass es andererseits auch eine Vielzahl an Ergebnissen gibt, die in nahezu allen zu diesem Themenbereich existierenden Umfragen unter Anwendung der unterschiedlichsten Methoden repliziert werden können. Hierzu zählt beispielsweise für den Bereich von Opferbefragungen das hohe Maß an Sicherheitsgefühl in der Bevölkerung, die höhere durchschnittliche Furcht von Frauen, die Zusammenhänge zwischen einzelnen Opferrisiken und dem Freizeitverhalten. Dadurch wird deutlich, dass methodische Herangehensweisen nicht zwingend gravierende Einflüsse auf Ergebnisse haben müssen.

können ebenfalls Hinweise auf die Sensibilität der Ergebnisse einzelner Methoden geben und zu einem besseren Verständnis von Methodeneffekten führen. Nicht nur deswegen, sondern auch zur Validierung einzelner Untersuchungsergebnisse erscheint es unerlässlich, empirische Forschungserkenntnisse über mehrere Studien und/oder Befragungen (auch mit unterschiedlichen Methoden) zu replizieren. Ergebnisse, die bisher noch nie in der vorliegenden Form festgestellt werden konnten, sollten daher stets mit einer gewissen Vorsicht interpretiert werden – was nicht bedeutet, dass sie falsch sind!

Noch wichtiger als die Replikation von Forschungsergebnissen ist es jedoch, ein fundiertes Verständnis für die unzähligen methodischen Einflüsse bei Umfragen zu entwickeln. Nach der Lektüre dieses Werks sollten sich die Leserinnen und Leser daher folgender zentraler methodischer Probleme und Einflüsse im Rahmen von Opferbefragungen bewusst sein: 1.) die zentrale Rolle der Stichprobenziehung (Design und Stichprobengröße) sowie die damit zusammenhängende Wahl einer Auswahlgrundlage, 2.) mögliche Verzerrungen durch Nonresponse, 3.) Einflüsse durch die Festlegung auf einen Erhebungsmodus (sowohl durch Effekte auf die Messung selbst als auch durch Zusammenhänge mit der Stichprobenqualität), 4.) Einflüsse durch die Fragebogenkonstruktion, insbesondere die Frage- und Antwortformulierung sowie die Fragereihenfolge, 5.) die Bedeutung der verwendeten Analyseform (insbesondere bei der Nutzung bi- oder multivariater Verfahren), 6.) die Bedeutung des Studiendesigns für die Ziehung von Schlussfolgerungen. Mit diesem Wissen ausgestattet sind die Leserinnen und Leser gut für einen sachkundigen Umgang mit Opferbefragungen gerüstet.

Es liegt in der Natur eines Kompendiums wie des vorliegenden, dass Forschungsdesiderate nur benannt werden, ohne dass ihnen abgeholfen werden kann. Wie oben angedeutet und in verschiedenen Kapiteln des vorliegenden Bands thematisiert, bestehen im Bereich der Methodik und Methodologie von Opferbefragungen derartige Lücken durchaus – und mit dem Aufkommen neuer methodischer Ansätze (etwa Onlinebefragungen oder der Einsatz spezieller Stichprobenverfahren zum Erreichen der Bevölkerung mit Migrationshintergrund) entstehen immer wieder neue Fragestellungen. Zudem ist es an der Zeit, dass die in der Einleitung zu diesem Band erwähnten älteren US-amerikanischen Untersuchungen im Kontext des *National Crime Victimization Surveys*, auf denen nach wie vor ein großer Teil der Erkenntnisse zur Methodologie von Opferbefragungen beruht, repliziert werden. Somit mangelt es nicht an Aufgaben für methodische Untersuchungen im Bereich der Umfragemethodologie allgemein wie auch spezifisch im Hinblick auf Opferbefragungen. Es steht zu hoffen, dass sich die *Scientific Community* – auch in Deutschland – dieser Herausforderungen annimmt.

Literatur

Groves, Robert M.; Fowler, Floyd J. Jr.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor und Tourangeau, Roger (2009): Survey Methodology, 2. Aufl. Hoboken, NJ: Wiley.

Hale, Chris (1996): Fear of Crime: A Review of the Literature. In: International Review of Victimology, 4, S. 79–150.

Kury, Helmut; Lichtblau, Andrea; Neumaier, André und Obergfell-Fuchs, Joachim (2004): Zur Validität der Erfassung von Kriminalitätsfurcht. In: Soziale Probleme, 15, S. 141–165.

Schnell, Rainer, Hill, Paul B. und Esser, Elke (2005): Methoden der empirischen Sozialforschung, 7. Aufl. München: Oldenbourg.

Simonson, Julia (2011): Problems in measuring punitiveness – results from a German study. In: Kury, Helmut; Shea, Evelyn (Hg.): Punitivity. International developments, Bd. 1: Punitiveness – a global Phenomenon? Bochum: Universitätsverlag Dr. N. Brockmeyer, S. 73–95.

Autorenverzeichnis

Dirk Baier

Dr., Leiter des Instituts für Delinquenz und Kriminalprävention an der Zürcher Hochschule für Angewandte Wissenschaften

Arbeitsschwerpunkte: Jugendkriminalität, Ausländerfeindlichkeit und Rechtsextremismus, Methoden der empirischen Sozialforschung

Kontakt: ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Departement Soziale Arbeit, Pfingstweidstrasse 96, CH - 8037 Zürich ,
E-Mail: baid@zhaw.ch

Christoph Birkel

Dr. phil., Soziologe, wissenschaftlicher Mitarbeiter am Bundeskriminalamt, Fachbereich KI 12 - Forschungs- und Beratungsstelle Polizeiliche Kriminalstatistik (PKS), Dunkelfeldforschung

Arbeitsschwerpunkte: Gewaltkriminalität, Dunkelfeldforschung, Viktimologie, Polizeiliche Kriminalstatistik, Methoden der empirischen Sozialforschung

Kontakt: Bundeskriminalamt, 65173 Wiesbaden,
E-Mail:Christoph.Birkel@bka.bund.de

Thomas Bliesener

Dr., Direktor des Kriminologischen Forschungsinstituts Niedersachsen in Hannover und Professor für Interdisziplinäre kriminologische Forschung an der Universität Göttingen

Arbeitsschwerpunkte: Aggression und Gewalt, Jugenddelinquenz und Kriminalität, Risiko- und Schutzfaktoren der Entwicklung von Störungen des Sozialverhaltens, Evaluation von Maßnahmen zur Kriminalprävention und -intervention, Evaluationsmethodologie

Kontakt: Kriminologisches Forschungsinstitut Niedersachsen, Lützerodestr. 9, 30161 Hannover, E-Mail: Bliesener@kfn.de

Kai Bussmann

Prof. Dr., Professor für Strafrecht und Kriminologie an der Martin-Luther-Universität Halle-Wittenberg, Leiter des Economy & Crime Research Centers, Geschäftsführender Vorsitzender der Vereinigung für Rechtssoziologie, Vorsitzender des Vorstandes der Landesgruppe Sachsen-Anhalt der Deutschen Vereinigung für Jugendgerichte und Jugendgerichtshilfe e. V.

Arbeitsschwerpunkte: Wirtschaftskriminalität, Gewalt in der Erziehung, Evaluation von kriminalpräventiven Maßnahmen

Kontakt: Martin-Luther-Universität Halle-Wittenberg, Juristische und Wirtschaftswissenschaftliche Fakultät, Lehrstuhl für Strafrecht und Kriminologie, Universitätsplatz 6, 06108 Halle (Saale),
E-Mail: kai.bussmann@jura.uni-halle.de

Marc Coester

Dr., Professor für Kriminologie an der Hochschule für Wirtschaft und Recht Berlin

Arbeitsschwerpunkte: Kriminalprävention, Rechtsextremismusforschung, Evaluationsforschung, Jugendstrafvollzug und Rückfallforschung, Jugendarbeit

Kontakt: Hochschule für Wirtschaft und Recht, Alt-Friedrichsfelde 60, 10315 Berlin, E-Mail: marc.coester@hwr-berlin.de

Matthieu de Castelbajac

Promovierter wissenschaftlicher Mitarbeiter am Institut für Soziologie an der Los Andes Universität Bogotá

Arbeitsschwerpunkte: Opferbefragungen, Soziologie der Gewalt, Geschichte der Kriminalstatistiken

Kontakt: Universidad de los Andes, Bogotá – Colombia | Carrera 1 Este, No 18A – 10, E-Mail: mh.decastelbajac@uniandes.edu

Bettina Doering

Dr. rer. nat., wissenschaftliche Mitarbeiterin am Institut für Psychologie, Fachbereich Klinische Psychologie und Psychotherapie der Philipps-Universität Marburg

Forschungsinteressen: delinquentes Verhalten im Kindes- und Jugendalter, Entwicklung von Stereotypen und Vorurteilen, Moralische Identität und moralische Motivation, Moralentwicklung

Kontakt: Fachbereich Psychologie, Philipps-Universität Marburg, Gutenbergstraße 18, 35037 Marburg,
E-Mail: bettina.doering@staff.uni-marburg.de

Dirk Enzmann

Dr., Wissenschaftlicher Mitarbeiter an der Fakultät für Rechtswissenschaft der Universität Hamburg

Arbeitsschwerpunkte: International vergleichende Jugenddelinquenzforschung, Surveyforschung, analytische Kriminologie

Kontakt: Universität Hamburg, Institut für Kriminalwissenschaften, Rothenbaumchausee 33, 20148 Hamburg,
E-Mail: dirk.enzmann@uni-hamburg.de

Frank Faulbaum

Dr., Professor für Sozialwissenschaftliche Methoden/Empirische Sozialforschung i. R., Vorstandsvorsitzender der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute (ASI) e. V.; Geschäftsführung und wiss. Leitung des Sozialwissenschaftlichen Umfragezentrums in Duisburg

Arbeitsschwerpunkte: Methoden der Umfrageforschung, komplexe Datenanalyse

Kontakt: Institut für Soziologie; Universität Duisburg-Essen; Lotharstraße 65; 47057 Duisburg; E-Mail: frank.faulbaum@uni-due.de

Thomas Görgen

Dr., Professor an der Deutschen Hochschule der Polizei in Münster; Leiter des Fachgebiets Kriminologie und interdisziplinäre Kriminalprävention

Arbeitsschwerpunkte: Kriminalität und demografischer Wandel, Jugendkriminalität, Gewalt im sozialen Nahraum, schwere Gewaltkriminalität, Kriminalprävention

Kontakt: Deutsche Hochschule der Polizei, Zum Roten Berge 18-24, 48165 Münster; E-Mail: thomas.goergen@dhpol.de

Werner Greve

Dr., Professor für Entwicklungspsychologie an der Universität Hildesheim

Arbeitsschwerpunkte: Entwicklungspsychologie, Kriminal- und Rechtspsychologie, Bewältigungsforschung, Evolutionäre Psychologie

Kontakt: Institut für Psychologie, Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim, E-Mail: wgreve@uni-hildesheim.de

Nathalie Guzy

Wissenschaftliche Mitarbeiterin am Bundeskriminalamt, Referat Forschungs- und Beratungsstelle Polizeiliche Kriminalstatistik (PKS), Dunkelfeldforschung

Arbeitsschwerpunkte: Viktimisierungsbefragungen, Vertrauen in die Polizei und Strafeinstellungen, Methoden der Umfrageforschung (insb. im Zusammenhang mit Viktimisierungsbefragungen)

Kontakt: Bundeskriminalamt, 65173 Wiesbaden,
E-Mail: Nathalie.Guzy@bka.bund.de

Michael Hanslmaier

Dr., Früherer wissenschaftlicher Mitarbeiter am Kriminologischen Forschungsinstitut Niedersachsen, jetzt Referat für Stadtplanung und Bauordnung der Landeshauptstadt München

Arbeitsschwerpunkte: Soziale Desorganisation, Kriminalitätswahrnehmung und Strafeinstellungen, Stadtsoziologie, Migration

Kontakt: Landeshauptstadt München, Referat für Stadtplanung und Bauordnung, I/21, Blumenstraße 31,80331 München,
E-Mail: michael.hanslmaier@muenchen.de

Janina Hatt

Doktorin der Rechtswissenschaft, Regierungsdirektorin

Referentin im Sekretariat des Nationalen Normenkontrollrates beim Bundeskanzleramt

Tätigkeitsschwerpunkt: Bessere Rechtsetzung, insbesondere Überprüfung von ex ante-Folgenabschätzungen

Kontakt: Nationaler Normenkontrollrat, Willy-Brandt-Str. 1, 10557 Berlin; E-Mail: janina.hatt@bk.bund.de

Wolfgang Heinz

Dr., Emeritierter Professor für Kriminologie und Strafrecht an der Universität Konstanz

Forschungsinteressen: Kriminalstatistik, Konstanzer Inventar, evidenzbasierte Kriminalpolitik

Kontakt: Holdersteig 13, 78465 Konstanz,
E-Mail: wolfgang.heinz@uni-konstanz.de

Helmut Hirtenlehner

Dr., Assoziierter Universitäts-Professor, Leiter des Zentrums für Kriminologie der Johannes Kepler Universität Linz

Forschungsschwerpunkte: Kriminalitätsfurcht, Abschreckung, Situational Action Theory

Kontakt: Johannes Kepler Universität Linz, Zentrum für Kriminologie, Altenberger Straße 69, A-4040 Linz, E-Mail: helmut.hirtenlehner@jku.at

Adrian Hoffmann

Dr. rer. nat., Diplom-Psychologe, Wissenschaftlicher Mitarbeiter in der Abteilung für Diagnostik und Differentielle Psychologie an der Heinrich-Heine-Universität Düsseldorf

Forschungsschwerpunkte: Indirekte Befragungstechniken, Soziale Erwünschtheit, Individuelle Unterschiede

Kontakt: Institut für Experimentelle Psychologie, Gebäude 23.03, Universitätsstr. 1, D-40225 Düsseldorf,
E-Mail: adrian.hoffmann@uni-duesseldorf.de

Edith Huber

Dr., Senior Researcherin im Fachbereich der Soziologie mit dem Fokus auf Sicherheitsforschung an der Donau-Universität Krems.

Forschungsschwerpunkte: Fehlverhalten im Internet, Cybercrime, Cyberstalking, Neue Medien, Sozialwissenschaft, Kriminologie, Cybersecurity und Saftey

Kontakt: Donau-Universität Krems, Dr. Karl-Dorrekstr. 30, 3500 Krems an der Donau, Austria, E-Mail: edith.huber@donau-uni.ac.at.

Dina Hummelsheim

Dr., Leiterin des Zentrums für Sozialindikatorenforschung, GESIS – Leibniz-Institut für Sozialwissenschaften, Mannheim

Arbeitsschwerpunkte: Kriminalitätserfahrungen und Kriminalitätswahrnehmungen im europäischen Ländervergleich, Sicherheit und Lebensqualität, Wohlfahrtsstaatenforschung

Kontakt: GESIS – Leibniz-Institut für Sozialwissenschaften, Zentrum für Sozialindikatorenforschung, B2,1, 68159 Mannheim,
E-Mail: dina.hummelsheim@gesis.org

Daniela Hunold

Wissenschaftliche Mitarbeiterin an der Deutschen Hochschule der Polizei in Münster; Fachgebiet Kriminologie und interdisziplinäre Kriminalprävention

Arbeitsschwerpunkte: Kriminalgeografie, Kriminalprävention, Polizeiforschung

Kontakt: Deutsche Hochschule der Polizei, Zum Roten Berge 18-24, 48165 Münster; E-Mail: daniela.hunold@dhpol.de

Janine Jager

Diplom-Psychologin, ehem. Mitarbeiterin am Institut für Psychologie der Christian-Albrechts-Universität zu Kiel

Cathleen Kappes

Dr. phil., Psychologin (Diplom), Wissenschaftliche Angestellte, Universität Hildesheim, Institut für Psychologie, Abteilung für Entwicklungspsychologie

Forschungsschwerpunkte: Entwicklung von Regulationsmechanismen in der Zielverfolgung und -ablösung, Emotionsentwicklung, Bewältigungsforschung

Kontakt: Institut für Psychologie, Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim, E-Mail: kappes@uni-hildesheim.de

Stefanie Kemme

Prof. Dr. iur., Professorin für Strafrecht und Kriminologie an der Akademie der Polizei Hamburg

Forschungsschwerpunkte: Jugendstrafrecht und Strafvollzugsrecht, Jugenddelinquenz, Interkulturelle Kriminologie, Kriminalitätsprognosen, Punitivität, Rechtspsychologie

Kontakt: Akademie der Polizei, Braamkamp 3b, 22297 Hamburg, E-Mail: stefanie.kemme@polizei-studium.org

Martin Killias

Prof. Dr. iur. h. c., Eigentümer und Geschäftsführer von Killias Research Consulting, Gastprofessor für Strafrecht an der Law School der Universität St. Gallen, Professor für Strafrecht an der Fernuniversität Schweiz, Dozent für Kriminologie in Wirtschaftskriminalität an der Hochschule Luzern

Forschungsschwerpunkte: Viktimisierungsstudien, Jugendkriminalität, Evaluation neuer Strafformen, Gewalt gegen Frauen, vergleichende und experimentelle Forschung im Bereich der Delinquenz

Kontakt: Killias Research & Consulting, Rathausgässli 27, Postfach 2094, 5600 Lenzburg, E-Mail: info@krc.ch

Thimna Klatt

Dipl.-Psych., M.Sc., Wissenschaftliche Mitarbeiterin am Kriminologischen Forschungsinstitut Niedersachsen

Arbeitsschwerpunkte: Jugenddelinquenz, Evaluation von Maßnahmen zur Kriminalprävention und -intervention, Gewalt gegen Polizeibeamte, Identifizierung von Tatverdächtigen durch Augenzeugen

Kontakt: Kriminologisches Forschungsinstitut Niedersachsen, Lützerodestr. 9, 30161 Hannover, E-Mail: thimna.klatt@kfn.de

Uwe Kolmey

Präsident des Landeskriminalamts Niedersachsen, Kriminalbeamter seit 1975

Forschungsschwerpunkte des LKA NI: Dunkelfeldforschung, Korruptionsanfälligkeit in der Polizei, Predictive Policing, Kriminalprävention im Städtebau

Kontakt: Uwe Kolmey, LKA Niedersachsen, Am Waterlooplatz 11, 30169 Hannover, E-Mail: uwe.kolmey@polizei.niedersachsen.de

Helmut Kury

Professor Dr., Dr. h. c. mult., Früherer Direktor des Kriminologischen Forschungsinstituts Niedersachsen (1980-1988), Mitglied der Forschungsgruppe Kriminologie am Max-Planck-Institut für ausländisches und internationales Strafrecht, Freiburg (pension.)

Arbeitsschwerpunkte: Strafvollzug und Resozialisierung von Straftätern, Sozialwissenschaftliche Forschungsmethoden, Verbrechensfurcht, Sanktionseinstellungen

Kontakt: Waldstraße 3, 79194 Heuweiler, E-Mail: helmut.kury@web.de

Heinz Leitgöb

Wissenschaftlicher Mitarbeiter an der Katholischen Universität Eichstätt-Ingolstadt; Leitung der „Working Group on Quantitative Methods in Criminology“ der European Society of Criminology (gemeinsam mit Daniel Seddig)

Arbeitsschwerpunkte: quantitative Methodenforschung, Soziologie des abweichenden Verhaltens & Kriminologie, Bildungsungleichheitsforschung & Evaluation des Bildungssystems

Kontakt: Katholische Universität Eichstätt-Ingolstadt, Kapuzinerkloster, Kapuzinergasse 2, 85072 Eichstätt, E-Mail: heinz.leitgoeb@ku.de

Robert Mischkowitz

Dr., Sozialwissenschaftler, Leiter des Fachbereichs KI 12 – Forschungs- und Beratungsstelle Polizeiliche Kriminalstatistik (PKS), Dunkelfeldforschung im Bundeskriminalamt

Forschungsschwerpunkte: Kriminelle Karrieren, kriminalistisch-kriminologische Analysen, Dunkelfeldforschung

Kontakt: Bundeskriminalamt, 65173 Wiesbaden,
E-Mail: Robert.Mischkowitz@bka.bund.de

Jochen Musch

Dr., Professor für Diagnostik und Differentielle Psychologie an der Heinrich-Heine-Universität Düsseldorf

Arbeitsschwerpunkte: Forschungsmethoden, Diagnostik, Individuelle Unterschiede

Kontakt: Institut für Experimentelle Psychologie, Gebäude 23.03, Universitätsstr. 1, D-40225 Düsseldorf,
E-Mail: jochen.musch@uni-duesseldorf.de

Frank Neubacher

Dr. iur., Professor für Kriminologie und Strafrecht an der Universität zu Köln; Direktor des Instituts für Kriminologie, 2014/15 Präsident der Kriminologischen Gesellschaft (KrimG)

Arbeitsschwerpunkte: Jugendkriminalität, Gewalt im Strafvollzug, Staatskriminalität und Völkerstrafrecht

Kontakt: Institut für Kriminologie der Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, E-Mail: f.neubacher@uni-koeln.de

Marcel Noack

Dr., Mitarbeiter am Lehrstuhl für empirische Sozialforschung an der Universität Duisburg-Essen

Arbeitsschwerpunkte: Grafisch gestützte Datenanalyse, Survey Methodology, Kriminalitätsfurcht

Kontakt: Universität Duisburg-Essen, Institut für Soziologie, Lotharstr. 65, 47057 Duisburg, E-Mail: marcel.noack@uni-due.de

Paul Norris

Dr., Dozent für Sozialpolitik, Institut für Sozial- und Politikwissenschaften

Forschungsinteressen: Vergleichende Kriminologie, Opferbefragungen, Muster der Opferwerdung, Öffentliche Wechselwirkung mit dem Strafrechtssytem, Quantitative sozialwissenschaftliche Methoden

Kontakt: Chrystal Macmillan Building, 15a George Square, Edinburgh, EH8 9LD, E-Mail: P.Norris@ed.ac.uk

Joachim Obergfell-Fuchs

Dr. phil., Oberpsychologierat, Leiter der Justizvollzugsschule und des Kriminologischen Dienstes Baden-Württemberg

Arbeitsschwerpunkte: Strafvollzug, Sexualstraftaten, Forensische Psychologie, Kriminalprävention

Kontakt: Justizvollzugsschule Baden-Württemberg, Pflugfelderstraße 21, 70439 Stuttgart,
E-Mail: Joachim.Obergfell-Fuchs@jvsbaden-wuerttemberg.justiz.bwl.de

Dietrich Oberwittler

Dr., Forschungsgruppenleiter am Max-Planck-Institut für ausländisches und internationales Strafrecht in Freiburg i. Brsg., Abteilung Kriminologie, und Privatdozent für Soziologie an der Albert-Ludwigs-Universität Freiburg

Arbeitsschwerpunkte: Kriminalsoziologie, Gewaltforschung, quantitative Forschungsmethoden

Kontakt: Max-Planck-Institut für ausländisches und internationales Strafrecht, Günterstalstraße 73, 79100 Freiburg,
E-Mail: d.oberwittler@mpicc.de

Lena Posch, geb. Stadler

Dr., Diplom-Psychologin, Fachpsychologin für Rechtspsychologie BDP/DGPs, psychologische Sachverständige im Straf- und Familienrecht am Bremer Institut für Gerichtspsychologie

Forschungsschwerpunkte: Viktimologie, sexueller Missbrauch an Kindern und Jugendlichen, Kindesmisshandlung und -vernachlässigung, Stalking

Kontakt: Bremer Institut für Gerichtspsychologie, Bürgermeister-Smidt-Str. 82, 28195 Bremen, E-Mail: posch@big-bremen.eu

Farina Rühs

BSc Psychologie, Geprüfte Wissenschaftliche Hilfskraft

Arbeitsschwerpunkte: Entwicklungspsychologie, Allgemeine Psychologie, Bewältigungsforschung

Kontakt: Institut für Psychologie, Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim, E-Mail: ruehsf@uni-hildesheim.de

Rainer Schnell

Dr., Professor für Methoden empirischer Sozialforschung an der Universität Duisburg-Essen, Director of the Centre for Comparative Social Surveys, City University London

Arbeitsschwerpunkte: Stichproben, Datenerhebungsmethoden, Record-Linkage

Kontakt: Prof. Dr. Rainer Schnell, City University London, Northampton Square, EC1V 0HB, London, United Kingdom
E-Mail: rainer.schnell@city.ac.uk

Monika Schröttle

Dr., Vertretungsprofessorin an der Fakultät für Rehabilitationswissenschaften der TU Dortmund, Koordinatorin des European Network on Gender and Violence (ENGV), Vorstand der wissenschaftlichen Fachgesellschaft Gender e. V.

Forschungsinteressen: Interdisziplinäre Gender-, Gewalt und Behinderungsforschung

Kontakt: Technische Universität Dortmund, Fakultät für Rehabilitationswissenschaften, Emil-Figge-Straße 50, 44227 Dortmund,
E-Mail: monika.schroettle@tu-dortmund.de

Daniel Seddig

Dr. phil.; Oberassistent am Soziologischen Institut der Universität Zürich; Chair der „European Working Group on Quantitative Methods in Criminology (EQMC)“

Arbeitsschwerpunkte: Jugendsoziologie und -entwicklung, Werte, Einstellungen, soziale Kontrolle und Devianz im Jugend- und Heranwachsendenalter, Quantitative Methoden, Strukturgleichungsmodelle und Statistik in den Sozialwissenschaften

Kontakt: Universität Zürich, Soziologisches Institut, Andreasstrasse 15, CH-8050 Zürich, E-Mail: seddig@soziologie.uzh.ch

Jan van Dijk

Dr., Emeritierter Professor der Viktimologie, Universität Tilburg/Gastprofessor an der Universität von Lausanne

Forschnungsinteressen: Opferbefragungen, Menschenhandel, Victim-Labeling-Theorie

Kontakt: Staalkade 3, 1011 JN Amsterdam, the Netherlands,
E-Mail: jan.vandijk@uvt.nl

Berenike Waubert de Puiseau

Diplom-Psychologin, Wissenschaftliche Mitarbeiterin in der Abteilung für Diagnostik und Differentielle Psychologie an der Heinrich-Heine-Universität Düsseldorf

Forschungsschwerpunkte: Rechtspsychologie, Diagnostik, Forschungsmethodik

Kontakt: Institut für Experimentelle Psychologie, Gebäude 23.03, Universitätsstr. 1, D-40225 Düsseldorf, E-Mail: bwdp@uni-duesseldorf.de

Polizeiliche Kriminalstatistiken geben leider nur ein ungenaues Bild der Kriminalitätswirklichkeit wieder, da sie stark an das Anzeigeverhalten der Bürgerinnen und Bürger sowie polizeiliche Schwerpunktsetzungen gebunden sind. Viele Straftaten verbleiben in einem kriminalstatistischen Dunkelfeld der Kriminalität, das aber vor allem mittels Opferbefragungen untersucht werden kann. Diese ergänzen das Gesamtbild der Kriminalität nun nicht nur im Hinblick auf das Ausmaß der berichteten Straftaten, die der Polizei nicht bekannt (gemacht) werden, sondern liefern überdies wichtige Informationen zur Kriminalitätsfurcht, dem Anzeigeverhalten und den Strafeinstellungen.

Im vorliegenden Sammelband werden sowohl der aktuelle Forschungsstand im Bereich von Opferbefragungen systematisch zusammengetragen als auch die methodischen und methodologischen Grundlagen und Probleme bei der Durchführung und Bewertung solcher Erhebungen in Deutschland beschrieben und diskutiert. Ein besonderer Fokus liegt dabei auch auf dem Aspekt der praktischen Relevanz, wobei sowohl die Notwendigkeit, als auch insbesondere die Ziele und Nutzungsmöglichkeiten der Ergebnisse von Opferbefragungen herausgearbeitet werden.

Der erste Band konzentriert sich auf die Erkenntnisse bisheriger Viktimisierungsbefragungen und thematisiert u. a.:

- Geschichte und Forschungsüberblick,
- Ziele und Nutzen von Opferbefragungen und
- Erkenntnisse zu delikt- und gruppenspezifischen Viktimisierungserfahrungen, hier u. a.: sexuelle Gewalt in Paarbeziehungen, Hasskriminalität, Gewalt gegen Ältere, Erfahrungen mit und Reaktionen auf Kriminalität.

Der zweite Band konzentriert sich auf die method(olog)ischen Grundlagen von Opferbefragungen und geht u. a. ein auf:

- Stichprobenziehung, Gewichtung und Nonresponse,
- datenschutzrechtliche Grundlagen,
- Effekte des Erhebungsmodus und
- soziale Erwünschtheit.

Die in diesem Werk diskutierten methodischen Probleme und praktischen Anwendungs- und Interpretationshilfen geben den Leserinnen und Lesern das für die (kritische) Interpretation und Würdigung der Ergebnisse von Viktimisierungsbefragungen benötigte Wissen an die Hand, so dass die Publikation eine Bereicherung für Vertreterinnen und Vertreter aus Wissenschaft, Politik und – insbesondere polizeilicher – Praxis darstellt.