Universität Konstanz
Fachbereich Politik- und Verwaltungswissenschaften

Interviewer- und Moduseffekte in Viktimisierungssurveys

Diplomarbeit

1. Betreuer: Prof. Dr. Rainer Schnell
2. Betreuer: Prof. Dr. Thomas Hinz

Inna Becher
Fürstengutweg 10

78462 Konstanz

Konstanz, 25. Januar 2007

Für meine Eltern
Ljubov Khusnullina und Wjatscheslav Khusnullin

Danksagung

Zuerst möchte ich meinem Professor, Herrn Dr. Rainer Schnell, danken, der mich zur gewählten Fragestellung ermutigte und von dem ich eine umfangreiche Betreuung bekommen habe.
Danken möchte ich auch Herrn Dr. Thomas Hinz, der sich die Zeit nahm, mich zu betreuen, und im Rahmen meiner Arbeit als Zweitprüfer fungiert.
Ich möchte mich auch herzlichst bei Frau Dr. Stefanie Eifler bedanken, die mich auf viele interessante Ideen und Problemlösungen brachte. Gedankt sei auch Frau Dr. Monika Schröttle für die Unterstützung bei der Beschaffung wichtiger Daten für meine Arbeit.
Als Letztes möchte ich meinem Mann, Mario Becher, für seine tagtägliche Unterstützung und viel Verständnis während der gesamten Diplomarbeit danken und meinem kleinen Sonnenschein, Tochter Vanessa Alexandra, die für mich eine große Motivationsquelle war.

Inhaltsverzeichnis

Tabellenverzeichnis

Abstract

The present work deals with the problems of biased results in social science inquiries. The main focus is placed on investigating distortions which can be caused by the interviewer's attributes as well as by the mode of data collection. These effects are examined in particular in the context of crime victimization surveys. Therefore, two German datasets were selected for the analyses - the DEFECT dataset and the dataset of the „Study of the Life Situation, Security and Health of Women in Germany“[1].

The theoretical basis for the explanation of these effects was provided by Esser (1985) who developed a model for explaining interviewee's behaviour as a rational action. The intraclass correlation coefficient as suggested by Kish (1962) was used for the mathematical analyses in this work. This coefficient was calculated by means of the one-factorial or two-factorial ANOVA or ANCOVA and with the help of the STATA-tool -gllamm-.

The analyses of the DEFECT dataset confirmed a strong biasing effect of the interviewer in the questioning process. It could be shown that male interviewers exerted significantly more influence on survey results than female interviewers. Moreover, this effect became even stronger with a rising age difference between the interviewer and the interviewee. The results of the women's study dataset approved the general biasing effect of the interviewer and the influence of the age difference between the interviewer and the interviewee.

Concerning the effects of data-collection method for the DEFECT study, a stronger impact of the interviewer was exercised in face-to-face interviews as compared to telephone ones. Additionally, in accordance with former investigations (cf. Schwarz *et al.* (1989)), the presence of recency effects in CATI interviews could be confirmed. For the women's study the higher item nonresponse was asserted in mail surveys as opposed to personal interviews. Furthermore the differences in the interviewee's behaviour between the modes were confirmed with regard to answering selected open questions.

The ascertained interviewer and mode effects in both studies lead to the renewed stress of the importance of high requirements for study designs which should strive for excision or at least for control of these effects.

[1] Here referred to as „Women's study“.

1 Einführung in die Problemstellung

Die empirische Sozialforschung hat im Laufe ihrer Geschichte verschiedene Instrumenten für die Datenerhebung entwickelt. Der überwiegende Teil der Daten wird mit Hilfe schriftlicher, mündlicher, telefonischer oder Internetbefragungen gewonnen. Die Entwicklung der letzten Jahre verzeichnet außerdem eine Verbreitung der sogenannten Mixed-Mode Designs, die die Vorteile verschiedener Modi in einer Untersuchung zum Einsatz bringen[1]. Jedes dieser Instrumente weist in seinen Eigenschaften Vor- und Nachteile auf. Die Nachteile drücken sich in erster Linie in den durch Besonderheiten des Instruments bedingten Verzerrungen der Ergebnisse aus.

Die vorliegende Arbeit beschäftigt sich gerade mit diesen Verzerrungen der Ergebnisse sozialwissenschaftlicher Befragungen. Diese lassen sich grundsätzlich in Sampling und Non-Sampling Errors unterteilen. Der Schwerpunkt der Arbeit soll auf der Untersuchung der Non-Sampling Errors liegen, die sich in a) Unit- sowie Item-Nonresponse und b) Fehler, die durch andere Faktoren, wie z.B. durch Lügen, Missverständnis der Fragen durch die Befragten etc. zustande kommen, unterteilen (vgl. Bradburn (1983)). In Anlehnung an Bradburn (1983: 289) werden die Fehler dieses zweiten Typs als Response Effekte bezeichnet und im Weiteren ausführlicher behandelt.

Die Grundlage für die Untersuchung der Response Effekte bildet die Varianz der Datenqualität. Bradburn (1983) unterscheidet *drei Ursachen* für die Varianz in der Datenqualität: 1) die Charakteristiken des Untersuchungsgegenstandes (wobei es dabei vor allem um den Inhalt der Fragen, deren Formulierung und Reihenfolge, den Befragungsmodus etc. geht); 2) Merkmale des Interviewers und 3) Charakteristiken des Befragten. Ich werde insbesondere auf die Einflüsse des Interviewers und die Interaktion der Interviewer- und Befragtenmerkmale sowie

[1] Es kann sich dabei z.B. um eine mündliche Befragung mit anschließendem Selbstausfüller handeln.

auf die Einflüsse des Befragungsmodus (Moduseffekte) eingehen.
Den Interviewer- und Moduseffekten wurde in der empirischen Sozialforschung viel Aufmerksamkeit gewidmet. Mein Anliegen ist es, diese in einem spezielleren Zusammenhang zu betrachten, nämlich ihre Ausprägung in Viktimisierungsbefragungen. Ich gehe davon aus, dass sowohl die Interviewer- als auch Moduseffekte in Viktimisierungssurveys von besonderer Bedeutung sind. Diese Vermutung stützt sich auf frühere Untersuchungen, insbesondere auf Ergebnisse der Studien zu Interviewereffekten bei sensitiven Fragen (vgl. Tourangeau (1996))[2].
Da solche Fragen bei Viktimisierungbefragungen in der Überzahl sind, dürfen die möglichen Interviewer- und Moduseffekte nicht vernachlässigt werden. Die Untersuchung dieser Effekte in Viktimisierungssurveys ermöglicht außerdem den Vergleich der interessierenden Effekte für sensitive gegenüber nicht sensitiven Fragen.
Eine weitere Besonderheit der Arbeit besteht in dem Versuch, der hierarchischen Struktur der Daten (nur im Falle des DEFECT-Datensatzes) Rechnung zu tragen und, wenn möglich, Sampling-Point- von Interviewereffekten zu trennen. Die sogenannten Mehrebenenmodelle werden für ausgesuchte Hypothesen am DEFECT-Datensatz getestet.

Es ergeben sich somit folgende Untersuchungsziele:

- Feststellung des (Nicht-)Vorliegens von Interviewer- sowie Moduseffekten in Viktimisierungssurveys,
- Untersuchung der Frage, ob die Intensität der Effekte mit der Frageart (sensitive vs. neutrale Fragen) variiert,
- Betrachtung der Einflussfaktoren, die für die Entstehung der genannten Effekte verantwortlich sind und
- Trennung der Interviewer- von den Sampling-Point-Effekten (soweit möglich).

Als Analysemethode wurde die einfaktorielle bzw. mehrfaktorielle Varianzanalyse gewählt, die mit STATA[3] gerechnet wurde.

[2] Als sensitiv können vor allem Fragen über Sexualität, Gewalterfahrungen, Einkommen, Krankheiten u.v.m. eingestuft werden.
[3] Version: STATA 8.2. Die Varianzanalyse ist als -loneway-, -oneway- oder -anova- implementiert.

Für die Trennung der Varianzen bei Mehrebenenmodellen wurde das von Rabe-Hesketh *et al.* (2004) geschriebene Programm -gllamm- (implementiert für STATA) verwendet. Es ermöglichte zudem bessere Schätzungen für logistische Modelle als dies mit ANOVA möglich wäre. Abhängig von den jeweiligen Itemcharakteristiken wurden lineare oder logistische Modelle analysiert.
Die Schätzung der Interviewereffekte wurde mit Hilfe des von Kish (1962) vorgeschlagenen Intraklassenkorrelationskoeffizienten durchgeführt, der stichprobenunabhängige Vergleiche ermöglichte. Bei den Untersuchungen der Moduseffekte wurden zusätzlich verschiedene Anteilswerte mit Hilfe von T-Tests verglichen.
Um eine bessere Vergleichbarkeit der Ergebnisse gewährleisten zu können, wurden für die Analysen zwei deutsche Datensätze ausgewählt[4]. Als weiteres Auswahlkriterium diente die Aktualität der Daten. Somit wurde zum einen die Studie zur „Lebenssituation, Sicherheit und Gesundheit von Frauen in Deutschland“[5] und zum anderen der DEFECT-Datensatz[6] ausgewählt. Der große Vorteil des zweiten Datensatzes war das Design mit interpenetrierenden Stichproben (vgl. dazu Bailar (1983)), das eine Trennung der Interviewer- und Sampling-Point-Effekte ermöglichte. Eine detaillierte Beschreibung der Datensätze mit Hinweisen bezüglich ihrer Vor- und Nachteile liefert Kapitel 5.

Im Weiteren wird der Ablauf meiner Arbeit kurz geschildert.
Das zweite Kapitel erläutert die theoretischen Grundlagen dieser Arbeit und den aktuellen Forschungsstand. Es beinhaltet eine Definition und die Klassifizierung der untersuchten Effekte sowie eine Darstellung ihrer Entstehungsursachen.
Kapitel 3 beschreibt die analysierten Datensätze mit ihren Vor- und Nachteilen für die Untersuchung. Das vierte Kapitel setzt sich mit dem Datenaufbereitungsprozess auseinander. Alle vorgenommenen Transformationen der Daten werden ausführlich behandelt und ihre Relevanz für die Analysen begründet.

[4]Die ursprüngliche Idee, die Effekte an mehreren Datensätzen zu testen und somit eine noch bessere Verallgemeinerbarkeit der Ergebnisse zu erlangen, wurde aufgegeben, da dies der zeitliche Rahmen meiner Diplomarbeit nicht zuließ. In Frage kamen folgende Datensätze: National Crime Victimization Survey, 1992-2004; International Victimization Survey, 1988-1992; International Crime Victimization Survey (ICVS), 1989-2000; British Crime Survey, 2002-2003; Violence and Threats of Violence Against Women and Men in the United States, 1994-1996.

[5]Dies ist die offizielle Bezeichnung der Studie. In meiner Arbeit wird diese Studie als „Frauenstudie“ und der Datensatz als „Frauendatensatz“ bezeichnet.

[6]Nur die erste Welle, Näheres hierzu im Kapitel 3.

Das fünfte Kapitel beschäftigt sich mit der mathematischen und inhaltlichen Beschreibung des eigentlichen Analysemodells. Im sechsten und siebten Kapitel erfolgen Formulierung und Prüfung der Hypothesen zu den Interviewer- und Moduseffekten. Das letzte Kapitel enthält ein Fazit über die Analyseergebnisse und erläutert deren Bedeutung für spätere Studien.

2 Theoretische Grundlagen

Das vorliegende Kapitel beschäftigt sich eingehend mit den theoretischen Konzepten, die dieser Arbeit zu Grunde liegen. Es behandelt unter anderem die Klärung der wichtigsten Begriffe, die Entstehungsursachen von Interviewer- und Moduseffekten und deren Konsequenzen für die Datenqualität sowie Beschreibung des Rational Choice Modells, das als theoretische Grundlage dieser Untersuchung dient. Des Weiteren werden die Besonderheiten der Ausprägung der Interviewer- und Moduseffekte in Viktimisierungssurveys angesprochen.

2.1 Verzerrungsursachen in Surveys

Das Interview ist als reaktiver sozialer Interaktionsprozess zu verstehen. Die Ergebnisse der Befragungen hängen vor allem von den Eigenschaften und dem Verhalten der Befragten und der Interviewer sowie von Rahmenbedingungen ab, die durch das Surveydesign (Itemauswahl, -formulierung, -abfolge, Modus der Befragung etc.) vorgegeben werden. Diese drei Faktorengruppen und deren Wechselwirkungen zählen daher zu den wichtigsten Verzerrungsursachen der Befragungsergebnisse (Bradburn (1983)).
Im Weiteren werden die Klassifizierung, Entstehung und Konsequenzen der genannten Verzerrungsursachen für die Datenqualität erläutert. Für die Befragten- und Interviewereffekte wird die Klassifizierung von (Reinecke (1991): 24ff.) übernommen.

2.1.1 Befragteneffekte

Der Befragte mit seinen Charakteristiken und Einstellungen steht im Mittelpunkt der sozialwissenschaftlichen Befragungen. Er ist in seiner situationsspezifischer Strategiewahl frei. Der Befragte alleine schätzt die Interviewsituation mit den

für ihn relevanten Kosten und dem Nutzen ein und entscheidet sich für eine der vorhandenen Handlungsstrategien. Es sind vor allem seine individuell charakteristischen Merkmale, die für die Wahl der Strategien verantwortlich sind: „Kognitive und sprachliche Kommunikationsbarrieren, Stimulusambiquität etc. erhöhen einerseits deferentes Verhalten wie Desinteresse oder Meinungslosigkeit, andererseits führen hohe Situationsstrukturierung oder Stimuluseindeutigkeit zur Präsentation von Sicherheit, Überlegenheit und sozial erwünschten Verhaltensweisen“ (Reinecke (1991): 24).
Demnach unterscheidet Reinecke (1991) in Anlehnung an Cronbach (1946) und Messick (1968) folgende Arten von Fehlreaktionen (die sogenannten „Response Sets“):

- Tendenz zu raten,
- Tendenz zu lügen,
- Tendenz zur Vollständigkeit,
- Bevorzugung von mittleren und neutralen Antwortkategorien,
- Bevorzugung von Extremkategorien,
- Bevorzugung von Geschwindigkeit vor Genauigkeit,
- Beurteilungsunterschiede bezüglich der Kategorien,
- die inhaltsunabhängige Zustimmungstendenz,
- die Tendenz, sozial erwünscht zu antworten.

Es wird darauf hingewiesen, dass die ersten sechs Arten eher unsystematisch auftreten. Der inhaltsunabhängigen Zustimmungstendenz und der Tendenz, sozial erwünscht zu antworten, wird dagegen ein systematischer Charakter zugesprochen (vgl. Reinecke (1991)). Diese zwei Response Sets spielen sowohl für die Interviewer- als auch für Moduseffekte eine besondere Rolle. Ihre Bedeutung wird im Rahmen des Rational-Choice-Modells des Befragtenverhaltens erläutert.

2.1.2 Interviewereffekte

Der Interviewer ist neben dem Befragten die zweite mögliche Quelle von Verzerrungen der Ergebnisse sozialwissenschaftlicher Befragungen. Die Effekte, die durch die Anwesenheit des Interviewers, die Wahrnehmung seiner Merkmale oder seines Verhaltens entstehen, werden Interviewereffekte genannt.
In der empirischen Literatur werden diese auf verschiedene Weise konzipiert. So werden die Interviewereffekte als „Interviewer Bias“ oder „Interviewervarianz“ aufgefasst. Im Weiteren werde ich die unterschiedlichen Auffassungen dieser Konzepte für die Berechnung der Interviewereffekte erläutern.
In dieser Arbeit halte ich mich an die Definitionen von Kish (1962) und Freeman & Butler (1976). Demnach wird die Varianz der Interviewer definiert als der Anteil der totalen Varianz der Antworten, der durch die Unterschiede zwischen den Interviewern erklärt bzw. herbeigeführt wird (Kish (1962)). Jeder Interviewer kann die Befragten in einem gewissen Umfang beeinflussen. Damit produziert er einen Bias (systematischen Fehler) bei den Befragten. Dies ist nach Freeman & Butler (1976) die erste Bedeutung des Interviewer Bias.
Betrachtet für alle teilnehmenden Interviewer, können sich diese systematischen Fehler aufheben, wobei der Populationswert (z.B. der Populationsmittelwert) unverzerrt bleiben kann, wenn die Bias verschiedener Interviewer in unterschiedliche Richtungen vom „wahren“ Wert abweichen.
Wenn dies nicht der Fall ist, wird die Schätzung des Populationsparameters verzerrt. Diese Verzerrung nennt man ebenfalls Interviewer Bias (Freeman & Butler (1976): 79).
Kish (1962) verwendet in diesem Zusammenhang den Begriff Interviewervarianz. Diese Varianz entsteht ebenfalls aus dem individuellen Interviewer Bias und hat einen Einfluss auf die Varianz des Stichprobenmittelwerts.
Somit wird der Interviewereffekt im Weiteren als Anteil der Varianz des Interviewers an der totalen Varianz mit Hilfe des von Kish (1962) eingeführten Intraklassenkorrelationskoeffizienten *rho* definiert. Eine mathematische Darstellung des Koeffizienten erfolgt im Kapitel 5.

Ursachen der Interviewereffekte

Den Ausgangspunkt für die Entstehung der Interviewereffekte bilden die Funktionen des Interviewers im Befragungsprozess. Seiner zentralen Rolle in diesem Prozess ist der potentiell große Einfluss auf die Befragungsergebnisse zuzuschreiben.
Folgende Aufgaben können dem Interviewer übertragen werden: Auswahl der Untersuchungseinheiten auf der letzten Stufe, Förderung der Kooperationsbereitschaft der Befragten, Durchführung der Befragung, Aufnahme der Antworten, maschinelle Eingabe der Antworten, Datenbereinigung etc. (vgl. dazu Groves *et al.* (2004): 269). Bei der Durchführung dieser Aufgaben können systematische oder unsystematische Fehler die Ergebnisse beeinflussen.
Diese können auf folgende *Faktoren* zurückgeführt werden (Maccoby & Maccoby (1972))[1]:

- Auftreten und Gebaren des Interviewers,
- die Art, in der ein Interviewer Fragen formuliert,
- Einstellungen des Interviewers,
- Erwartungen des Interviewers in Bezug auf die Einstellung des Befragten,
- Variationen der Interviewer bei Sondierungsfragen,
- Unterschiede zwischen den Interviewern bei der Aufzeichnung der Antworten.

Diese Faktoren können wiederum in drei Gruppen aufgeteilt werden:

1. sichtbare Merkmale des Interviewers (Geschlecht, Rassenzugehörigkeit, Alter, Auftraggebereffekt etc.),
2. nicht sichtbare Merkmale des Interviewers (z.B. Einstellungen, Erwartungen etc.),
3. Verhalten des Interviewers (Art der Fragestellung, Tonart, korrekte Aufnahme der Antworten etc.).

[1] zitiert nach Reinecke (1991).

Der Einfluss vieler der genannten Faktoren kann mittels eines durchdachten Designs reduziert werden, das auf intensives Interviewertraining und umfangreiche Kontrollen sowie eine weitgehende Standardisierung des Designs setzt (Groves *et al.* (2004)). Dies betrifft vor allem die zweite und dritte Faktorengruppe.
Der Effekt der sichtbaren Merkmale lässt sich hingegen nur schwer reduzieren, da es dabei vor allem um die Interaktion dieser Merkmale mit entsprechenden Merkmalen der Befragten und des Befragungsinstruments geht. Wenn die sichtbaren Merkmale, wie z.B. Geschlecht, Alter oder sozialer Status[2] etwas mit der konkreten Fragestellung oder dem Untersuchungsinhalt sowie mit den Eigenschaften der Befragten zu tun haben, können Interviewereffekte auftreten.
Die im Kapitel 6 formulierten und untersuchten Hypothesen zu Interviewereffekten konzentrieren sich daher auf den Einfluss der sichtbaren Merkmale des Interviewers (Geschlecht und Alter) und deren Interaktion mit den Befragtenmerkmalen sowie den verschiedenen Itemtypen.

Forschungsstand

Die Befunde der vorhandenen empirischen Untersuchungen zu Interviewereffekten sind nicht eindeutig. Unabhängig davon, welche Faktorengruppe in ihrem Einfluss untersucht wird, kommen verschiedene Forscher zur Bestätigung oder zur Ablehnung von Interviewereinflüssen.
Sudman & Bradburn (1974) weisen in ihrem Überblick der Literatur auf widersprüchliche Studienergebnisse hin. Einige Arbeiten finden keine bedeutenden Effekte der Befragten- oder Interviewercharakteristiken auf das Befragtenverhalten. Andere verbinden das Vorhandensein der Effekte mit dem Studieninhalt.
Vielzählige Untersuchungen kommen zur Schlussfolgerung, dass die Interviewereffekte tatsächlich existieren. Die Grundlage für eine solche Schlussfolgerung ist es, dass die Varianz der abhängigen Variable groß ist und nicht allein durch die Varianz, die durch die Art der Stichprobenziehung entsteht (sampling variance) erklärt werden kann (vgl. Hanson & Marks (1958), Stock & Hochstim (1951) und Kish (1962)).
Andere Studien gehen einen Schritt weiter und stellen unterschiedliche Intensitäten der Interviewereffekte für verschiedene Itemtypen fest. So bestätigt z.B. Tucker (1983) Interviewereffekte bei ausgewählten Items der von CBS News und

[2]Sozialer Status lässt sich aus den sichbaren Merkmalen des Interviewers ableiten.

The New York Times durchgeführten nationalen Befragungen. Ebenso stellen Groves & Magilavy (1986) anhand von neun Surveys des Survey Research Centers fest, dass es gewisse Unterschiede bei den Messungen der Interviewereffekte zwischen offenen und geschlossenen sowie zwischen faktischen oder Einstellungsfragen gibt. Diese Effekte hängen vor allem von den Einflussmöglichkeiten, die dem Interviewer zur Verfügung stehen, ab. Bei offenen, schwierigen oder sensitiven Fragen steigt sein potentieller und realer Einfluss auf die Befragten[3].

Auch dem Einfluss der Interviewer- und Befragtenmerkmale und deren Interaktion für die Entstehung der Interviewereffekte wurde eine Reihe von Studien gewidmet. So bestätigen Groves & Magilavy (1986) und Hanson & Marks (1958) eine größere Anfälligkeit der älteren Respondenten gegenüber den Interviewereinflüssen.
Auf die Interviewereffekte als Ergebnis der Interaktion von Befragten-, Interviewer- und Itemmerkmalen weisen in ihren Untersuchungen unter anderem Freeman & Butler (1976), Tucker (1983) und Williams (1964) hin. Diese Effekte sind dann besonders ausgeprägt, wenn sich die wahrgenommenen Charakteristiken der Interviewer, wie z.B. Geschlecht, Rasse oder Alter mit dem Untersuchungsgegenstand in Verbindung bringen lassen (vgl. dazu Sudman & Bradburn (1974)). Wenn der Befragte keinen Zusammenhang zwischen seinen Antworten und den sichtbaren Charakteristiken des Interviewers sieht, sollten diese frei von Interviewereffekten sein (Groves (1989)).
Eine wichtige Rolle spielen auch Interviewererfahrung und -erwartungen. Der Einfluss der Interviewererfahrung lässt sich oft von dem des -alters nicht trennen, da die Erfahrung mit dem Alter stark korrelliert. In der Regel haben jüngere Interviewer weniger Erfahrung und sind somit für stärkere Interviewereffekte verantwortlich. Laut Sudman & Bradburn (1974) sind Response Effekte bei unerfahrenen Interviewern doppelt so hoch wie bei erfahrenen.
Den Einfluss des Interviewergeschlechts untersuchen unter anderem Huddy *et al.* (1997), Sczesny & Stahlberg (1999) und Groves & Fultz (1985). Viele Studien dieser Art haben Sexualität im weitesten Sinne zum Gegenstand der Untersuchung. Bradburn (1983) gibt in diesem Zusammenhang an, dass sich die Effekte, die sich auf das Geschlecht des Interviewers beziehen, überwiegend in Studien

[3]Siehe dazu Kapitel 2.1.1 zum Befragtenverhalten.

mit geschlechtsspezifischen Themen zeigen lassen. Im Allgemeinen findet man die größten Effekte bei den Dyaden „männlicher Interviewer - männlicher Befragte“, sehr kleine Effekte dagegen bei den Frauen, die von Männern oder Frauen befragt wurden. Robins (1974) sowie Johnson & Delameter (1976) konnten dagegen keinen Einfluss des Interviewergeschlechts feststellen.

2.1.3 Moduseffekte

Neben den Interviewereinflüssen wird hier die verzerrende Wirkung der Befragungsmodi untersucht. Im Folgenden werde ich Moduseffekte definieren, Unterschiede zwischen verschiedenen Modi beschreiben und auf die Konsequenzen dieser Unterschiede für die resultierenden Daten eingehen[4].

Forschungsstand

Manche Forscher sehen das größte Verzerrungspotential im Interviewer selbst (vgl. Stock & Hochstim (1951)), andere betonen hingegen den überragenden Einfluss der „Aufgabe“ („task itself“): „task definition is primarily a matter of what questions are asked, how they are asked - that is, their form and wording - the order in which they are asked, and the mode of administration of the questionnaire“ (Bradburn (1983)). Sudman & Bradburn (1974) sehen in diesen Variablen die wichtigste Ursache der Response Effekte.
Eine in der früheren Forschung verbreitete Herangehensweise zielt auf den Vergleich von Befragungsmodi unter den Gesichtspunkten der Kosten, der Antwortraten und der Qualität der Antworten. So findet z.B. Hochstim (1967) heraus, dass die drei klassischen Befragungsmodi fast identische Qualität der Daten, insbesondere in Bezug auf die Validität der Daten und die Antwortraten liefern. De Leeuw (1992) dagegen weist auf höhere Antwortraten bei persönlichen Befragungen im Vergleich zu den beiden anderen Modi hin.
Moduseffekte nach De Leeuw (1992: 21) werden als „systematic differences between data collected by means of mail, telephone, and face to face surveys“ definiert. Er vergleicht Modi unter anderem im Hinblick auf die Antwortraten und verschiedene Aspekte der Datenqualität.

[4]Hier und in der weiteren Arbeit geht es um die drei klassischen Modi - die persönliche, telefonische und schriftliche Befragung.

Klassifizierung der Einflussfaktoren

In der einschlägigen Literatur werden außerdem drei Hauptfaktoren genannt, die für die Unterschiede zwischen den Modi verantwortlich sein können[5]:

1. *Mediumverbundene Faktoren („media related factors").* Darunter versteht man z.B. den *Verbreitungsgrad* des Mediums oder die *Vertrautheit* der Befragten mit dem Medium. Dieser Faktor ist für den sogenannten Coverage Error[6] verantwortlich und soll in dieser Arbeit nicht weiter behandelt werden.

   Der zweite Faktor dieser Gruppe betrifft die *Zuweisung der Kontrollmöglichkeiten* in der Befragung. In einem Face-to-Face-Interview teilen sich der Interviewer und der Befragte die Kontrolle, in einem Telefoninterview übernimmt der Interviewer die Kontrolle (er ist der Initiator und bestimmt weitgehend das Tempo), bei einem Mailsurvey liegen alle Kontrollmöglichkeiten beim Befragten (z.B. Bestimmung des Tempos, der Reihenfolge der Fragenbeantwortung, schließlich auch die Auswahl der beantwortenden Person).

   Weitere Faktoren der Gruppe sind die *Akzeptanz von Konversationspausen* und die Möglichkeiten des Mediums, die Befragten zur *Aufrichtigkeit und Ehrlichkeit* zu überzeugen (De Leeuw (1992): 15). So kann in einem Face-to-Face-Interview durch persönliche Überzeugung oder bei einer schriftlichen Befragung mit Hilfe von Logos des Erhebungsinstituts und Unterschriften der Projektführenden ein besserer Überzeugungsgrad als in einer telefonischen Befragung erreicht werden. All diese Faktoren spielen für die Formulierung sowie Prüfung der Hypothesen eine zweitrangige Rolle und werden daher im Weiteren nicht ausführlich untersucht.

2. *Informationsübertragung.* Als erstes sind in diesem Zusammenhang die *Kommunikationskanäle* zu nennen. Man unterscheidet verbale, nonverbale und paralinguistische Kommunikation[7]. Nur Face-to-Face-Befragungen sind in

[5] Die hier beschriebene Klassifizierung der Faktoren findet sich bei De Leew (1992: 13). Bei der Formulierung und Prüfung der Hypothesen zu Moduseffekten werde ich auf diese Klassifizierung der Faktoren zurückgreifen.

[6] Coverage Error entsteht wenn Beobachtungen doppelt vorkommen oder in die Analyse fälschlicherweise eingeschlossen werden bzw. im Sample fehlen, obwohl sie ein Teil der Zielpopulation wären.

[7] „[...] paralinguistic communication is concerned with (non verbal) auditive signals, like emotional tone, timing, emphasis, and utterances like „mhm-hmm" (De Leeuw (1992)).

der Lage, von allen drei Kommunikationskanälen zu profitieren. Die Telefoninterviews basieren dagegen nur auf verbaler und paralinguistischer Kommunikation. Bei den Mail-Surveys wird nicht zwischen den drei Formen unterschieden, da nur schriftliche Ausdrucksmöglichkeiten zur Verfügung stehen. Die eingesetzten Graphiken und Layouts können jedoch zum Teil die Rolle der nonverbalen und paralinguistischen Kommunikationskanäle übernehmen.

Der zweite Faktor dieser Gruppe bezieht sich auf die *Stimuluspräsentation*. Diese kann visuell oder als Audiopräsentation erfolgen. Bei diesem Vergleich schneiden ebenfalls persönliche Befragungen am besten ab, da diese von beiden Arten Gebrauch machen können. Mailsurveys beschränken sich weitestgehend (jedoch nicht ausschließlich) auf die visuelle und telefonische Befragungen auf die Audiopräsentation. Die Präsentationsart verbunden mit dem zeitlichen Aspekt der Befragung (Zeitdruck vorhanden oder nicht) führt zur Entstehung der Response Order Effekte. Unter diesen Begriff fallen Primacy und Recency Effekte[8].

Schwarz *et al.* (2002) bestätigen in ihrer Untersuchung folgende bereits etablierte Hypothese: „mail surveys or face-to-face interviews with the help of show-cards may render results that are quite different from the results of telephone interviews without the use of show-cards, given that the primacy effects that emerge in one mode combine with the recency effects that emerge in the other“ (Schwarz *et al.* (2002): 204). Diese Effekte sollen am DEFECT-Datensatz untersucht werden[9].

Der dritte Faktor dieser Gruppe ist die *zeitliche Reihenfolge der Itempräsentation*. Face-to-Face und telefonische Befragungen beruhen auf der sequentiellen Präsentation der Items. Die Reihenfolge wird durch das Design bestimmt und vom Interviewer kontrolliert. Diese Kontrolle ist bei einer schriftlichen Befragung nicht möglich, somit können Respondenten selber bestimmen, welche Fragen wann beantwortet werden. Dieser Faktor ist für die Ausschaltung der Itemreihenfolge- und Anwesenheit der Itemkontexteffekte bei schriftlichen Befragungen und für die umgekehrte Konstellation

[8] „[...] *primacy effects*, that is, higher endorsements of items presented early in the list, [...] *recency effects*, that is, higher endorsements of items presented late in the list [...] (Schwarz *et al.* (2002)).

[9] Die Frauenstudie enthält keine CATI-Befragung.

der genannten Effekte in telefonischen und persönlichen Interviews verantwortlich. Diese wurden u.a. von Schwarz *et al.* (2002) und Bishop *et al.* (1987) untersucht. Beide Untersuchungen bestätigen den postulierten Zusammenhang: Itemreihefolgeeffekte sind bei schriftlichen Befragungen nicht signifikant, Itemkontexteffekte haben bei mündlichen und telefonischen Befragungen geringe Bedeutung.
Außerdem variiert die *Kommunikationsgeschwindigkeit* zwischen den Modi. Da in persönlichen Interviews nonverbale sowie paralinguistische Kommunikation zusätzlich eingesetzt werden, besteht ein geringer Zeitdruck. Die Telefonbefragungen dagegen zeichnen sich durch einen starken Zeitdruck aus, was zu bestimmten Effekten führen kann. Für die schriftlichen Befragungen hat dieser Faktor keine besondere Bedeutung.

3. *Interviewereinflüsse.* Verschiedene Modi bieten unterschiedliche Möglichkeiten, den Einfluss des Interviewers zu beschränken. In Mailsurveys sind diese abwesend und können weder einen positiven noch einen negativen Einfluss ausüben. Bei telefonischen Befragungen haben Interviewer einen im Vergleich zu Face-to-Face-Interviews geringeren Einfluss auf die Befragten.
Zu den möglichen positiven Einflüssen gehören u.a. die Motivation der Befragten durch die Interviewer, Klärung verschiedener Fragen der Respondenten während des Interviews und eine bessere Kontrolle über den Ablauf der Befragung (De Leeuw (1992)). Die Abwesenheit des Interviewers bei einem Mailsurvey dagegen kann eine bessere Anonymität gewährleisten, was insbesondere für Befragungen zu sensitiven Themen von Bedeutung sein kann. Dieser Zusammenhang wird am Datensatz der Frauenstudie überprüft.
Der negative Einfluss resultiert aus den „Klumpeneffekten“, da der jeweilige Interviewer als eine gruppierende Variable betrachtet werden kann, deren Einfluss auf die Befragten in spezifischer Weise erfolgt. Diese Effekte sind in Face-to-Face-Befragungen größer, da die vielfältigen Kommunikationskanäle eine Vermittlung sichtbarer sowie nichtsichtbarer Merkmale der Interviewer ermöglichen. Eine Überprüfung findet im Rahmen des Hypothesentestens am DEFECT-Datensatz statt.
Man geht davon aus, dass die Präsenz der Interviewer zu Effekten führt, die auf die soziale Wünschbarkeit zurück zu führen sind. Das Entstehen der

Effekte der sozialen oder kulturellen Wünschbarkeit[10] hängt von der wahrgenommenen Anonymität (größer in einem Mailsurvey) und dem Itemtyp ab. Bei bedrohlichen, sensitiven oder Einstellungsfragen sind diese Effekte in einem Face-to-Face-Interview am stärksten ausgeprägt. Die Befragten können aus den wahrgenommenen Interviewercharakteristiken bestimmte Erwartungen der Interviewer ableiten und sich dementsprechend verhalten (Sudman & Bradburn (1974)). Diese Hypothese wird an beiden Datensätzen getestet.

Von allen angesprochenen Möglichkeiten wird in der vorliegenden Untersuchung besonders dem Vergleich der Interviewereffekte in verschiedenen Modi Aufmerksamkeit geschenkt.

2.2 Rational Choice Modell zur Erklärung des Befragtenverhaltens

Bei der Suche nach den Verzerrungsursachen, die in sozialwissenschaftlichen Befragungen relevant sind, stehen der Befragte und sein Verhalten als wichtigste Erklärungsfaktoren im Mittelpunkt. Es wird dementsprechend nach passenden Modellen zur Erklärung des Befragtenverhaltens gesucht.

Die moderne empirische Sozialforschung wendet sich von Modellen, die auf den Annahmen der klassischen Testtheorie basieren ab, da das Verhalten der Befragten nicht nur durch seine stabilen latenten Merkmale erklärt werden kann (Esser (1986)). Esser (1986) stellt dem normativen Paradigma, das von fixen Rollenerwartungen und sich stets wiederholenden Typen von Handlungskonstellationen ausgeht, das interaktionistische Konzept gegenüber. Demnach sind Personen „Situationen interpretierende, Bedeutungen stets neu aushandelnde und definierende Akteure, die Rollen nicht bloß passiv-konform ausfüllen, sondern nach ihren Zielsetzungen und Deutungen aktiv ausgestalten“ (Esser (1986)).

Das nachfolgend beschriebene Konzept von Esser (1985) wird in dieser Arbeit als Grundlage der Erklärung des Befragtenverhaltens und damit der Antwortverzerrungen in Interviews verwendet.

[10] Näheres dazu weiter in diesem Kapitel unter „Rational Choice Modell zur Erklärung des Befragtenverhaltens“.

2.2.1 Systematische Fehler in Befragungen

Vor der eigentlichen Modellformulierung definiert Esser (1986: 4) zwei wichtige Arten systematischer Fehler: die *inhaltsunabhängigen Verzerrungen* und die *inhaltsbezogenen Reaktionen.*

Zu den *inhaltsunabhängigen* Verzerrungen zählt vor allem die sogenannte *Zustimmungstendenz*. Sie wird definiert als „systematische Reaktion in Situationen hoher Diffusität bei Personen [...], die solche Sutuationen nicht auf andere Weise zu steuern gewohnt sind und „Deferenz“ und Anpassung als einzige Behauptungsstrategie kennen“ (Esser 1986: 5).

Als eine der wichtigsten Formen der *inhaltsbezogenen* Verzerrungen gilt die *soziale Erwünschtheit*, die wie auch die Zustimmungstendenz entweder als ein Merkmal der Persönlichkeit oder als eine ausgearbeitete Strategie betrachtet werden kann. Die soziale Erwünschtheit als Persönlichkeitsmerkmal kann als kulturelle oder situationale Erwünschtheit aufgefasst werden.

Im ersten Fall geht es um „internalisierte Rollenerwartungen“, die aus dem Zusammenspiel der äußeren Merkmale der Befragten (Geschlecht, Alter etc.) und anderer Merkmale (z.B. Einstellungen, Erwartungen) entstehen.

Wenn die normativen Erwartungen, die die Antworten beeinflussen, erst in einer spezifischen Situation wirksam werden, dann spricht man von der situationalen sozialen Erwünschtheit. Diese Art der Erwünschtheit führt zu Effekten, die hier von besonderem Interesse sind: Effekte von Merkmalen der Interviewer, der Anwesenheit Dritter, Sponsoreneffekte etc..

Anschließend soll das Problem der heiklen und bedrohlichen Fragen angesprochen werden, die DeMaio (1984) als einen Spezialfall des Problems der sozialen Erwünschtheit bezeichnet. Solche Fragen werden in meiner Untersuchung als „sensitive Fragen“ bezeichnet. Da es in Viktimisierungssurveys vor allem um bedrohliche, heikle Fragen oder Fragen zu tabuisierten Themen geht, erwarte ich starke Erwünschtheitseffekte.

Im Weiteren werden ältere Erklärungsmodelle des Befragtenverhaltens dargestellt und eine ihrer Weiterentwicklungen, das Modell des Befragtenverhaltens als „rationales Handeln“ ausführlich besprochen.

2.2.2 Ältere Erklärungen des Befragtenverhaltens

Esser (1985) bietet einen Überblick über die älteren Konzepte des Befragtenverhaltens und erwähnt in diesem Zusammenhang vor allem die sogenannten Orientierungstheorien. Diese beschreiben die Datenerhebung als einen wechselseitigen Prozess, in dem die Beteiligten die Situation wahrnehmen und reflektieren, was zum Versuch führen kann, sich in einem möglichst günstigsten Licht darzustellen. Die Orientierungstheorie von Kahn & Cannell (1968) erklärt das Interviewergebnis aus Merkmalen, Wahrnehmungen und dem wechselseitig orientierten Verhalten vom Befragten und Interviewer. Die Theorie liefert jedoch keinen Mechanismus zur Bestimmung des Befragtenverhaltens.
Eine Weiterentwicklung der Orientierungshypothese stammt von Phillips (1971), der auf den folgenden Grundmechanismus aller sozialen Prozesse hinweist: das alltägliche Handeln wird durch das Eigeninteresse des jeweiligen Akteurs bestimmt, seinen Nutzen zu maximieren und dabei vor allem nach sozialer Anerkennung zu streben. Der Befragte nimmt die Situation in ihrem Gesamtbild wahr, deutet dabei die vermuteten Absichten des Interviewers und handelt schließlich entsprechend seiner eigenen Zielsetzungen und den vermuteten Erwartungen des Interviewers. Seine Kosten-Nutzen-Überlegungen sind für die Wahl einer „wahren“ oder verzerrten Antwort verantwortlich: eine „wahre“ Antwort wird nur gewählt, wenn die Konsequenzen einer falschen Antwort als unwichtig oder als sehr unwahrscheinlich eingeschätzt werden (vgl. dazu Phillips (1971): 89f.). Es sind also nicht allein die Erwartungen, die Eigenschaften des Interviewers und des Befragten sowie die erwarteten Sanktionen, vielmehr geht es bei der Wahl der Handlungsstrategie um den Vergleich verschiedener alternativer Reaktionen und die dabei vollzogenen Kosten-Nutzen-Erwägungen.
Außerdem geht Esser (1985) auf eine Reihe von spezielleren Ansätzen ein, wie z.B. die „Theorie der Frage“ von Holm (1974) oder den Ansatz von Atteslander & Kneubühler (1975). Diese spezielleren Modelle bewerten das Befragtenverhalten ebenfalls als Ergebnis einer nach Kosten-Nutzen-Überlegungen erfolgten Entscheidung zwischen Handlungsalternativen. Diese Idee wird in dem Ansatz von Esser (1985) weiterentwickelt und um präzisere Mechanismen erweitert.

2.2.3 Befragtenverhalten als „rationales Handeln“

Nachfolgend soll das Modell des Befragtenverhaltens nach Esser (1985) dargestellt werden.
Das Grundmodell besteht aus drei grundlegenden Elementen. Das erste Element bilden die unterschiedlichen *Ziele* eines Individuums in einer konkreten Befragungssituation. Diese Ziele sind Ergebnisse gewisser variabler Bedürfnisse, z.B. nach sozialer Anerkennung, personaler Identität, materiellem Wohlergehen etc.. Jedes Ziel wird vom Individuum in seiner *Intensität* mit U_1, U_2,..., U_n bewertet. Zweitens hat ein Individuum einen Satz von einigen in der Situation vorstellbaren *Handlungsalternativen* A_1, A_2,..., A_m, die mit den Zielen über *subjektive Erwartungen* p_{11}, ..., p_{ij}, p_{mn}, dass eine bestimmte Handlung A_i zum Ziel (der Bewertung U_j) führt, verbunden sind. Für jede Handlungsalternative A_i nimmt der Akteur in Bezug auf jedes Ziel eine Gewichtung mit der subjektiven Wahrscheinlichkeit vor. Dafür wird das Produkt der Zielbewertung U_j mit der subjektiven Wahrscheinlichkeit p_{ij} gebildet. Dieses Produkt heißt die „Relevanz“ der Handlung oder die *Handlungstendenz*.
Im letzten Schritt wird für jede Handlung A_i die Summe der Produkte $p_{ij}U_j$ gebildet, die als „subjective expected utility“ bezeichnet wird. Gewählt wird die Handlung, die in der gegebenen Situation die höchste subjektive Nutzenerwartung aufweist.
Das Schema kann durch Hinzunahme der negativ bewerteten Ziele (der „Kosten“) erweitert werden. Dementsprechend wird es Handlungen geben, die sowohl Kosten als auch Nutzen aufweisen und Handlungen, die ausschließlich Nutzen hervorbringen können. Beim Vergleich der Handlungsalternativen wird es dann nicht nur um den höchsten Netto-Nutzen gehen. Handlungen, die gegenüber Alternativen Nutzen ohne Kosten versprechen, werden präferiert (Esser (1986)).
Dieser Aspekt ist für die Entstehung sowie den Umgang mit Interviewer- und Moduseffekten unentbehrlich. Bei den erwähnten Kosten kann es sich um den Anonymitätsgrad der Befragung (modusabhängig), den Effekt der Anwesenheit Dritter etc. handeln. Durch den gezielten Einfluss auf diese störenden Faktoren (Kosten) können die subjektiven Wahrscheinlichkeiten für sozial erwünschte Antworten beeinflusst werden: „die befürchtete Reaktion einer Bezugsumwelt [schlägt] dann nicht mehr als Kosten einer an der „wahren“ persönlichen Einstellung orientierten Antwort zur Buche“ (Esser (1986)).

Nun muss auch der Einfluss der Situtationsvariabilität im Konzept aufgenommen werden. Im Einklang mit dem Interaktionismus wird davon ausgegangen, dass sich die subjektiven Wahrscheinlichkeiten und die Zielbewertungen in neuen Situationen erst herausbilden müssen. Solche neuen oder undefinierten Situationen kennzeichnen sich durch die Ungewissheit über die Konsequenzen aller Handlungsalternativen. Gerade bei dieser Konstellation können die Interviewer- und Moduseinflüsse auftreten: abhängig vom Modus der Befragung kann die Transparenz der Situation durch nonverbale Signale, paralinguistische Konversation, Erläuterungen u.a. erhöht und dadurch die fehlenden Werte der Matrix der subjektiven Wahrscheinlichkeiten aufgefüllt werden. Abhängig davon, in welche Richtung die Situation definiert wird[11], bekommt die subjektive Wahrscheinlichkeit für die „wahre“ oder die verzerrte Einstellung einen höheren Wert (Esser (1985)). Esser (1985) geht davon aus, dass die Zielbewertungen relativ stabil bleiben, die subjektiven Wahrscheinlichkeiten dagegen sehr variabel sind. Die Werte der subjektiven Wahrscheinlichkeiten werden durch die kurzfristigen Einflüsse der Situation verändert[12]. Vor allem über diese wirken die Eigenschaften und Handlungen anderer Akteure auf die Wahl der Strategie eines Befragten. Der situationsbedingte Einfluss ist insbesondere bei hoher Ambiguität des Fragestimulus und bei starker Variabilität der Kognition[13] nicht zu unterschätzen. In diesen Situationen werden die kurzfristigen Einflüsse das Ergebnis weitgehend bestimmen.

Der entscheidende Punkt in der Theorie des Befragtenverhaltens ist die Klärung der Frage, wann sich der Befragte für die „wahre“ Antwort entscheidet und wann eine verzerrte Einstellung abgegeben wird. Zur Beantwortung dieser Frage untergliedert Esser (1985) die Handlungsziele des Befragten in drei Dimensionen:
1) die *„wahre“ Einstellung*, die als latentes Merkmal ein Teil der Persönlichkeit darstellt (mit der Intensität U_t), 2) die Bedeutung von kulturellen Normen und der Anerkennung in der gewohnten sozialen Umgebung, die sogenannte *kulturelle Identität* (mit der Intensität U_c) und 3) die *situationale Erwünschtheit* einer

[11] So kann z.B. eine unklare Fragestellung verdeutlicht oder die Einstellung des Interviewers zum Thema sichtbar werden.

[12] Zu den kurzfristigen Einflüssen der Situation gehören z.B. Vermutungen über die Einstellungen des Interviewers, der wahrgenommene Anonymitätsgrad und die gelernten „Alltagstheorien“.

[13] Z.B. keine Einstellung des Befragten zum Thema, bedeutungslose oder lange zurückliegende Ereignisse.

Antwort, die für situationsbezogenen Verzerrungen verantwortlich ist (mit der Intensität U_s).
Für die hier untersuchten Effekte ist die situationale Erwünschtheit, insbesondere sozial erwünschte Reaktionen der Befragten von besonderem Interesse. Für diese nennt Esser (1986: 19) drei Bedingungen, die gleichzeitig erfüllt werden müssen. Erstens muss die Situation vom Befragten typisiert werden können, so dass die Folgen bestimmter Handlungen abgeschätzt werden können. Dabei geht es um die sichtbaren Merkmale der Interviewer. Zweitens müssen diese Situationsmerkmale mit bestimmten typisierten Konsequenzerwartungen verbunden werden. So können z.B. die Befragten bei bestimmten Themen unterschiedliche Einstellungen seitens der männlichen und weiblichen Interviewer erwarten. Der dritte Aspekt ist die wahrgenommene Öffentlichkeit der Situation[14]. Erst wenn die genannten drei Bedingungen vorliegen, kann das Verhalten durch Erwünschtheitseffekte beeinträchtigt werden - die subjektive Wahrscheinlichkeit für eine „wahre“ Antwort wird geringer als die für eine sozial erwünschte Antwort sein.
Esser (1986) weist auf einen weiteren interessanten Aspekt hin: nicht jede sozial erwünschte Antwort ist automatisch als eine verzerrte Antwort aufzufassen. Eine Verzerrung liege demnach erst dann vor, wenn die situational bedingte Reaktion A_i von der dem „wahren“ Wert entsprechenden Reaktion A_j abweicht. Somit sind systematische Effekte sozialer Erwünschtheit nur bei Vorliegen einer Differenz zwischen A_i und A_j vorhanden.

Fazit

In diesem Kapitel wurde auf die drei wichtigsten Faktoren der Verzerrungen in sozialwissenschaftlichen Befragungen eingegangen - den Befragten, den Interviewer und die Rahmenbedingungen, die durch das Untersuchungsdesign definiert werden. Von diesen Faktoren ausgehend wurden die möglichen Verzerrungen klassifiziert und die Mechanismen ihrer Entstehung erläutert.
Weiterhin wurden ältere sowie aktuelle Erklärungsmodelle des Befragtenverhaltens angesprochen und das in dieser Arbeit verwendete Modell von Esser (1985) ausführlich dargestellt.

[14] Die Situation kann durch ein eingeschränktes Vertrauen dem Interviewer gegenüber oder durch die Anwesenheit Dritter als öffentlich und somit als nicht sicher wahrgenommen werden.

3 Beschreibung der Datensätze

Eine detaillierte Beschreibung der Datensätze ist für die Hypothesenprüfung sowie die statistischen Analysen entscheidend, da dies eine Legitimationsbasis für die Anwendung bestimmter Verfahren und die Schlussfolgerungen bildet. Die für die hier durchgeführten Analysen wichtigen Charakteristiken der Datensätze werden im Weiteren dargestellt.

3.1 DEFECT

Design

Der DEFECT-Datensatz ist das Ergebnis eines DFG-Projekts, das als Hauptziel die Schätzung der Effekte, die durch die räumliche Klumpung der Untersuchungseinheiten entstehen, hatte[1]. Diese Effekte führen zur Unterschätzung der Standardfehler und Verfälschung der Konfidenzintervalle, die auf den in der empirischen Sozialforschung verbreiteten mehrstufigen Klumpenauswahlen beruhen. Der Datensatz erlaubte erstmalig die Berechnung der Designeffekte für die Bundesrepublik Deutschland sowie die Schätzung der Modus- und Interviewereffekte, da auf Grund des Designs eine Trennung verschiedener Effekte möglich war (Schnell & Kreuter (2000)).
Das Design basierte auf den sogenannten „interpenetrierenden Stichproben“ (Bailar (1983)). Die Implementation des Designs für die DEFECT-Studie sah den Einsatz eines Interviewers in nur einem Sampling-Point und die gleichzeitige Arbeit zweier Institute (somit auch zweier Interviewer) unabhängig voneinander in einem Sampling-Point vor, die die Face-to-Face-Befragungen durchführten. Zusätzlich führte ein dritter Interviewer in den gleichen Sampling-Points

[1]Das DEFECT-Projekt lief unter der Projektnummer SCHN 586/2-1 bei der Deutschen Forschungsgemeinschaft.

mit Quotenauswahl eine weitere Face-to-Face-Befragung durch. Um eine Berechnung der Moduseffekte zu ermöglichen, führte ein viertes Institut in den gleichen Sampling-Points eine CATI-Befragung[2] durch. Außerdem wurde in den gleichen Sampling-Points eine Mailbefragung von den Mitarbeitern des Projekts erhoben. Alle Befragungen wurden mit einem identischen Fragebogen in den gleichen Sampling-Points zum gleichen Zeitpunkt von voneinander unabhängigen Instituten vollzogen (Schnell & Kreuter (2000)).
Die Feldzeiten der Haupterhebungen[3] erstreckten sich von Oktober 1999 bis Februar 2000.

Stichprobenziehung und Auswahl der Untersuchungseinheiten

Die erste Stufe der Stichprobenziehung bildete die Auswahl der Sampling-Points aus dem ADM-Mastersample[4] basierend auf der Datei des Bundeswahlleiters (160 Sampling-Points für Random-Erhebungen und 160 für die Quoten-Stichprobe). Die Stichproben wurden nach politischen Gemeindegrößenklassen geschichtet. Für die Quotenstichprobe wurde ausgehend von den 160 Sampling-Points aus der ADM-Stichprobe der Random-Erhebungen eine überschneidungsfreie, strukturgleiche Parallelstichprobe gezogen[5].
Der zweite Schritt für die Random-Erhebungen bestand in der Auswahl der Zielhaushalte mit Hilfe des „Address-Random“-Verfahrens (Behrens & Löffler (1999)). Gegen eine Einwohnermeldeamtstichprobe sprachen folgende Gründe: höhere Kosten, längere Vorlaufzeiten, Datenschutzprobleme, die durch die Weitergabe der Adressen von einem Institut an andere Institute entstehen würden.
Pro Sampling-Point wurden 110 Adressen gesammelt und nach der maschinellen Bereinigung zufällig auf fünf unabhängige Stichproben verteilt. Drei der Stichproben bekamen die Institute[6], eine Stichprobe diente der Durchführung des Mail-Surveys und die letzte Adressenstichprobe als Reserve.

[2] CATI - Computer Assisted Telephone Interviewing.
[3] Unter Haupterhebungen sind hier die drei Face-to-Face-Befragungen, der Random-CATI-Survey und die Mail-Befragung zu verstehen.
[4] ADM - Arbeitskreis deutscher Marktforschungsinstitute. Eine Erklärung des ADM-Mastersamples anhand des ALLBUS 1980 findet sich bei Schnell *et al.* (2005).
[5] Alle Einzelheiten dazu finden sich bei Schnell & Kreuter (2000).
[6] Entsprechend für zwei Random-Face-to-Face-Befragungen und eine Random-CATI-Befragung. Das vierte Institut führte die Quotenbefragung durch, die im Weiteren erläutert wird.

Die Untersuchungseinheiten der letzten Stufe waren Individuen mit folgenden Merkmalen: deutschsprachig, älter als 18 Jahre, lebend in Privathaushalten. Für die Face-to-Face-Befragungen erfolgte die Auswahl der Personen über einen „Schwedenschlüssel“, bei der telefonischen und postalischen Befragung nach der „Last-Birthday“-Methode[7].

In der Quotenstichprobe kamen zwei Quotentabellen zum Einsatz. Die Festlegung der Quotenmerkmale orientierte sich an den Ergebnissen bisheriger Kriminalitätsfurchtsurveys, die untersucht hatten, bei welchen Merkmalen mit Unterschieden in der abhängigen Variable gerechnet werden muss.

Pro Stichprobe wurde eine Mindestanzahl von acht Interviews in jedem Sampling-Point angestrebt. Dafür bekamen die Institute zunächst 16 Adressen und vier Ersatzadressen pro Point und später weitere Adressen, falls die vereinbarte Mindestanzahl nicht realisiert werden konnte. Für die beiden Random-Face-to-Face-Erhebungen sowie den Mail-Survey wurden im Anschluss an die Erhebung CATI-Nonresponse-Studien durchgeführt.

Für die CATI-Erhebung wurden ebenfalls 20 Adressen pro Sampling-Point vergeben, die aus einer Telefon-CD-Rom stammten. Um eine Mindestanzahl von acht Interviews pro Point realisieren zu können, wurden zusätzlich 4.465 RLD-Nummern erzeugt[8].

Auswahl der Interviewer

Für die Auswahl der Interviewer und deren Schulung waren die Institute[9] zuständig. Die Interviewer der Face-to-Face-Erhebungen wurden von den Instituten schriftlich, die Interviewer der CATI-Erhebung im Telefonstudio geschult. Das Design der Studie erforderte, dass jeder Interviewer nur in einem Sampling-Point eingesetzt wurde[10] und dass in maximal 10% der Sampling-Points der Einsatz eines zweiten Interviewers erlaubt war, wenn dieser in keinem anderen Point tätig

[7] Zur Erklärung der beiden Verfahren siehe Schnell *et al.* (2005)

[8] RLD - Randomized Last Digit. Bei diesem Auswahlverfahren wird eine Zufallszahl zur einer zufällig aus dem Telefonbuch gewählten Nummer addiert. Siehe dazu Schnell *et al.* (2005).

[9] Beteiligt waren folgende Institute: ACADEMIC DATA Gesellschaft für Umfragen, Methodenberatung und Analysen mbH (Essen), foerster & thelen Marktforschung Feldservice GmbH (Bochum), infas - Institut für angewandte Sozialwissenschaft GmbH (Bonn) und INRA Deutschland Gesellschaft für Markt- und Sozialforschung mbH (Mölln).

[10] In einigen Fällen, in denen die Einhaltung der Bedingung unmöglich erschien, sollte der zweite Interviewer desselben Instituts erst dann eingesetzt werden, wenn der Erste aufgehört hat zu arbeiten.

war. Weiterhin gab es keine Anforderungen an das Geschlecht des Interviewers.

Befragungsinstrument

Der Fragebogen der Studie hatte Kriminalitätsfurcht zum Thema. Er enthielt 71 Fragen, die mit Berücksichtigung aller Filter 135 Items ergaben. Außer der Fragen zur Kriminalitätsfurcht, subjektiven Viktimisierungswahrscheinlichkeit und allgemeinen Lebenssituation enthielt der Fragebogen auch sehr sensitive Fragen zur tatsächlich erlebten Gewalt. Vor allem an diesen Fragen im Vergleich zu neutralen Fragen werden später die aufgestellten Hypothesen getestet. Fragen zur sexuellen Gewalt, die in dieser Untersuchung ebenfalls von großem Interesse sind, wurden ausschließlich Frauen gestellt.
Sehr hilfreich sind für das Testen der Interviewereffekte detaillierte Angaben zu den Interviewermerkmalen: eindeutige Identifikationsnummer, Alter, Geschlecht, Schulabschluß etc.. Diese Angaben sind in den in Deutschland zur Verfügung stehenden Datensätzen keine Selbstverständlichkeit.

Antwortraten

In den beiden Random-Face-to-Face-Befragungen wurden 1345 und 1326 Interviews durchgeführt, was einer Antwortrate von 39.3% und 41.5% entsprach. Im Rahmen der telefonischen Befragung konnten 1350 Interviews realisiert werden, was eine Antwortrate von nur 29.4% bedeutete. Schnell & Kreuter (2000) weisen in diesem Zusammenhang auf die deutlich kürzere Feldzeit der CATI-Befragung als Ursache der niedrigeren Ausschöpfungsquote hin. Die postalische Befragung lieferte 1161 gültige Fragebögen und verzeichnete somit eine Rücklaufquote von 49.1%. Die Quotenbefragung lieferte 1276 Interviews, in denen die Quotenvorgaben erfüllt werden konnten.

3.2 Frauenstudie

Design

Das wesentliche Ziel der Untersuchung war die Erstellung eines umfassenden Bildes der Gewalterfahrungen und Lebenssituation von Frauen in Deutschland. Die Studie sollte die erste repräsentative Befragung dieser Art in Deutschland sein und gleichzeitig den Vergleich der Ergebnisse mit anderen nationalen Untersuchungen in europäischen Ländern ermöglichen. Die Untersuchung wurde von März 2002 bis September 2004 durch das Interdisziplinäre Zentrum für Frauen- und Geschlechterforschung (IFF) der Universität Bielefeld im Auftrag des Bundesministeriums für Familie, Frauen, Senioren und Jugend in Kooperation mit *infas* durchgeführt (Müller & Schröttle (2004)).
Die Studie bestand aus drei Teilen: 1) einer repräsentativen Hauptuntersuchung (Februar bis Oktober 2003), die in Form eines standardisierten Face-to-Face-Interviews durchgeführt wurde und einem schriftlichen Selbstausfüller, der im Anschluss an das persönliche Interview in Anwesenheit der Interviewerin ausgefüllt wurde; 2) einer Zusatzbefragung bei türkischen und osteuropäischen/russischen Migrantinnen mit je 250 Interviews in türkischer und russischer Sprache und 3) einer Teilpopulation-Zusatzbefragung bei Asylbewerberinnen, Frauen in Haft und Prostituierten.
Meine Arbeit basiert ausschließlich auf den Daten der Hauptuntersuchung und des schriftlichen Selbstausfüllers.

Stichprobenziehung und Auswahl der Untersuchungseinheiten

In der ersten Auswahlstufe wurden bundesweit insgesamt 250 Gemeinden mit 278 Sampling-Points per Zufall ausgewählt[11]. Entsprechend dem nach alten und neuen Bundesländern disproportionalen Design wurden 188 Sample-Points in 175 Gemeinden für die alten Bundesländer inklusive West-Berlin und 90 Sample-Points in 75 Gemeinden für die neuen Bundesländer inklusive Ost-Berlin zufallsgesteuert ausgewählt[12].

[11] Leider enthält der Public-Use-File keine Angaben zu den Sampling-Points.

[12] Zufallsgesteuert innerhalb von Schichtungszellen, die sich aus der Kombination von Kreisen mit zehn BIK-Gemeindegrößenklassen ergaben (Müller & Schröttle (2004)). Die BIK-Gemeindegrößenklassen sind ein Modell zur Einteilung von Gemeinden nach der Zahl der Einwohner.

In jeder Gemeinde wurde die gleiche Anzahl von Personenadressen per Zufallsauswahl gezogen. In Großstädten, die mehrfach in die Stichprobe gelangten, wurde ein Vielfaches dieser Anzahl ausgewählt. In den alten Bundesländern waren somit aus den 1.076 Schichten 188 Sampling-Points (175 Gemeinden), in den neuen Bundesländern 90 Sampling-Points (75 Gemeinden) aus 435 Schichten ausgewählt.
Die Ziehung der Personenadressen für die Zielgruppe in den 250 Gemeinden basierte auf einer systematischen Zufallsauswahl (Intervallziehung). Ausgehend von einer zufälligen Startadresse wurden über eine feste Schrittweite die übrigen Adressen systematisch ausgewählt (Müller & Schröttle (2004)).
Die angestrebte Anzahl der Interviews pro Sampling-Point betrug durchschnittlich 36. Um den vielfältigen möglichen neutralen Ausfällen Rechnung zu tragen, wurden 144 Adressen pro Sample-Point angefordert.
Aus dieser Netto-Stichprobe wurde im letzten Schritt eine Brutto-Einsatzstichprobe nach einem vorgegebenen Verfahren gezogen (Müller & Schröttle (2004): 763).
Der Vorteil der Personenstichprobe für die Studie hing aus der Sicht der Forscher mit den Besonderheiten des Untersuchungsthemas zusammen. Die Personenstichprobe sollte eine bessere Ansprache der Zielpersonen (womöglich auch höhere Antwortraten) ermöglichen.
Als Zielpersonen kamen alle in der Bundesrepublik Deutschland lebende Frauen im Alter von 16 bis 85 Jahren, auch Ausländerinnen (sofern deutsch sprechend) in Frage. Insgesamt wurden 10.264 Frauen im Alter von 16 bis 85 Jahren befragt.

Auswahl der Interviewer

Für die Befragungen sollten ausschließlich weibliche Interviewer eingesetzt werden. Diese Entscheidung wurde damit begründet, dass sich weibliche Befragte gegenüber Frauen bei sehr sensiblen Themen - insbesondere bei Fragen zu sexueller Gewalt - leichter und vertrauensvoller öffnen sollten.
Bei der Auswahl der Interviewerinnen war die fachliche Qualifikation entscheidend. Die für diese Studie eingesetzten Interviewerinnen sollten Erfahrungen mit komplexen Erhebungsdesigns und gleichzeitig mit sensiblen Themenbereichen besitzen (Müller & Schröttle (2004)).
Zu den Vorbereitungen der Interviewerinnen gehörten eine eintägige persönliche Schulung durch *IFF* und *infas* sowie ein schriftliches Interviewerhandbuch

für jede Interviewerin. Den Großteil der Interviewerinnen bildeten die Stamm-Interviewerinnen von *infas*.
Um den Einfluss einer wichtigen Verzerrungsquelle - des Interviewerworkloads - zu minimieren, wurden statt der 100 geplanten 213 Interviewerinnen eingesetzt. Die Durchschnittsbelastung betrug 48 Interviews pro Interviewerin (Fredebeul *et al.* (2004)). Leider enthielt der Public-Use-File keine Angaben zur eindeutigen Identifikation der Interviewerinnen in beiden Befragungen sowie keine Angaben zum Alter und Bildungsstand der Interviewerinnen[13].

Befragungsinstrument

Im Verlauf des persönlich-mündlichen Interviews wurden zwei Erhebungsinstrumente eingesetzt: ein mündlicher Fragebogen, der durch ein Listenheft ergänzt wurde und ein schriftlicher Fragebogen zum Selbstausfüllen (Drop-off). Der Selbstausfüller wurde nach Beendigung des mündlichen Interviews im Beisein der Interviewerin von den Befragten ausgefüllt und der Interviewerin in einem verschlossenen Briefumschlag übergeben. In Ausnahmefällen konnte er auch per Post zurückgeschickt oder später von der Interviewerin abgeholt werden.
Zwar konzentrierten sich beide Fragebögen im Großen und Ganzen auf das gleiche Thema, tatsächlich unterschieden sie sich jedoch hinsichtlich Länge, Items und Schwerpunkte. Während der mündliche Fragebogen allgemeine demographische Merkmale, Wohnsituation, Gesundheit, aktuelle sowie ehemalige Partnerschaften und allgemeine Gewalterfahrungen erhebte, behandelte der schriftliche Fragebogen insbesondere Gewalterlebnisse in den Partnerschaften.
Durch diese Varianz in den Fragebögen muss auf die Prüfung einiger Hypothesen zu Moduseffekten, deren Prüfung am DEFECT-Datensatz möglich ist, verzichtet werden.

Antwortraten

Für den Kontakt mit der Zielperson war mit den Interviewerinnen eine Mindestzahl von vier Kontaktversuchen pro Adresse vereinbart worden, sofern nicht bereits zu einem früheren Kontakt ein Interview realisiert werden konnte. Im Durchschnitt waren 2.3 Kontaktversuche für die Realisierung eines Interviews

[13] An dieser Stelle sei Frau Dr. Monika Schröttle für ihre Kooperation und Hilfsbereitschaft bei der Beschaffung dieser fehlenden Angaben gedankt.

notwendig.
Die Ausschöpfungsquote der von den neutralen Ausfällen bereinigten Brutto-Stichprobe lag bei insgesamt 51.6 Prozent. Die Realisierungsquote sank mit zunehmendem Alter. Bei den jungen Zielpersonen unter 24 Jahren lag diese über 57 Prozent, bei den ältesten Zielpersonen über 75 Jahre sank sie auf 39 Prozent. Im Anschluss an den mündlichen Teil konnte in 9.905 Fällen (96.5%) der schriftliche Fragebogen an die Befragten übergeben werden. In 624 Fällen fehlte der Drop-off oder wurde nicht vollständig bzw. korrekt ausgefüllt (6.1%). Bezogen auf die Gesamtheit aller auswertbaren Fälle (mündliches Interview) betrug die Ausschöpfung für die schriftliche Befragung 93.9% (Müller & Schröttle (2004)).

Zusammenfassend kann man sagen, dass beide Datensätze die Prüfung der allgemeinen Interviewereffekte sowie der Interaktionseffekte des Intervieweralters mit dem Befragtenalter ermöglichen. Weiterhin können verschiedene Arten der Effekte des Interviewgeschlechts am DEFECT-Datensatz getestet werden, da die Auswahl der Interviewer nicht auf die weiblichen Personen beschränkt war. Somit werden für den DEFECT-Datensatz im sechsten Kapitel sieben, für die Frauenstudie drei Hypothesen formuliert und getestet.
Die Vorteile des DEFECT-Datensatzes können ebenfalls bei der Untersuchung der Moduseffekte genutzt werden. Die formulierten Hypothesen spiegeln die im Kapitel 2 definierten Auffassungen der Moduseffekte wieder - als 1) Ähnlichkeit der Beobachtungen innerhalb eines Befragungsmodus, 2) unterschiedliche Ausprägung der Interviewereffekte in verschiedenen Modi und 3) Moduseffekte als unterschiedliche Datenqualität. Das Design der Frauenstudie erlaubt nur eine eingeschränkte Prüfung der Moduseffekte im Hinblick auf die Datenqualität.

Fazit

Hier wurden die Datensätze mit den für die Untersuchung relevanten Charakteristiken dargestellt. Es wurden Besonderheiten des Designs, der Auswahl der Untersuchungseinheiten und der Interviewer sowie die Antwortraten in beiden Datensätzen angesprochen. Anschließend wurde die Relevanz dieser Charakteristiken für die Formulierung und Prüfung der Hypothesen im Kapitel 6 und 7 erläutert.
Im nächsten Kapitel wird näher auf den Datenaufbereitungsprozess eingegangen.

4 Datenaufbereitung

Der Datenanalyseprozess kann generell in drei Schritten beschrieben werden: Datenaufbereitung, Datenanalyse und Interpretation der Ergebnisse. Der erste Schritt zielt auf die Datenbereinigung und auf die Untersuchung der Datenstruktur ab. Bei einer Sekundäranalyse, wie sie hier vorgenommen wurde, geht es vor allem um das Auffinden von Mustern in den Daten, die Beschreibung der Verteilungen der Variablen, die Behandlung der fehlenden Werte (missings) und der Ausreißer. Dieser Schritt beinhaltet zudem auch Datentransformationen sowie Erstellung analytischer Variablen, Umkodierungen und Kategorisierungen (vgl. Marczyk *et al.* (1964)).
Die Datenaufbereitung wird im Folgenden getrennt nach Datensätzen beschrieben.

4.1 DEFECT

Die überwiegende Anzahl der im Kapitel 6 und 7 formulierten Hypothesen wird anhand von 16 Items getestet, die die abhängigen Variablen des Modells darstellen.
Bei der Itemauswahl sind folgende Überlegungen ausschlaggebend: die meisten von ihnen sollen Viktimisierungserfahrungen, insbesondere sexuelle Gewalterfahrungen erheben und gleichzeitig keine sehr schiefe Verteilungen aufweisen. Außerdem sollen zum Vergleich der Effekte auch nicht sensitive Items in die Analyse aufgenommen werden.
Für die Analysen werden alle Items in zwei Gruppen aufgeteilt:

- **Typ1-Items**: Items mit erhöhter Wahrscheinlichkeit der Interviewereffekte. Dazu zählen sensitive/emotional geladene (bedrohliche oder Fragen mit hohem Grad der sozialen Erwünschtheit), Einstellungsfragen, schwierige und

offene Fragen[1].

- **Typ2-Items**: Items mit geringerer Wahrscheinlichkeit der Interviewereffekte. Hierzu gehören nicht sensitive (neutrale), faktische, einfache, geschlossene Fragen[2].

In der Literatur finden sich kompliziertere Klassifizierungen der Items nach Typen, die sich in ihrem „Anfälligkeitsgrad“ gegenüber den Interviewereffekten unterscheiden sollten (vgl. dazu Hanson & Marks (1958)). So beinhaltet die von Mangione *et al.* (1992) vorgeschlagene Klassifizierung folgende vier Dimensionen: schwierig/einfach, sensitiv/nicht sensitiv, faktisch/nicht faktisch und offen/geschlossen[3].
Für die genannten Itemtypen werden verschiedene Hypothesen bezüglich der Ausprägung der Interviewereffekte aufgestellt und getestet, deren Ergebnisse bei verschiedenen Forschern zum Teil widersprüchlich ausfallen. So finden z.B. Collins & Butcher (1982) heraus, dass faktische Fragen weniger beeinflussbar sind als die Einstellungsfragen. Zum gegenteiligen Ergebnis kommen Kish (1962) und O'Muircheartaigh & Campanelli (1998). Ebenfalls als „anfällig“ für die Einflüsse des Interviewers werden auf Grund weiterer Studien sensitive, schwierige oder offene Fragen eingestuft (Kish (1962)).
Da der Schwerpunkt dieser Untersuchung die Analyse der Interviewer- und Moduseffekte in Viktimisierungssurveys und nicht die Feinheiten der Ausprägungen dieser Effekte bei verschiedenen Itemtypen aller Arten ist, wird auf die aufwendige Klassifizierung der Items verzichtet und lediglich der Vergleich von sensitiven mit nicht sensitiven Items vorgenommen.

Folgende drei Hypothesen dieser Arbeit werden jedoch an fast allen Items des Datensatzes untersucht: „Der Befragungsmodus hat einen Einfluss auf die Varianz der Befragungsergebnisse der Viktimisierungssurveys“ (getestet an 112 von 135 Items), „Männliche Interviewer haben bei Viktimisierungsbefragungen einen

[1] Folgende Items fallen in diese Kategorie: f10, f13_2, f13_4, f14_2, f14_3, f25, f33a, f34, f35, f35a, f42, f44. Für die Zuordnung des Items zu dieser Kategorie ist ausreichend, wenn es eine der genannten Charakteristiken aufweist. Die Ausformulierung der Fragen findet sich im Anhang A.2.1.

[2] Diese Kategorie beinhaltet folgende Items: f1, f31, f38_4, f43. Zur Ausformulierung siehe den Anhang A.2.1.

[3] Eine der Anwendungen dieser Klassifikation mit anschließenden Hypothesentests findet sich bei Schnell & Kreuter (2005).

größeren Einfluss auf weibliche Befragte als weibliche Interviewer“ und „Weibliche Interviewer haben bei Viktimisierungsbefragungen einen größeren Einfluss auf männliche Befragte als männliche Interviewer“ (getestet an 104 von 135 Items). Die Reduzierung der Itemanzahl in diesen drei Fällen ergibt sich aus verschiedenen Gründen. Erstens werden nominalskalierte Variablen ausgeschlossen[4]. Zweitens werden sehr schiefe binär kodierte Variablen bei allen Hypothesentests ausgeschlossen. Dazu zählen alle Items, bei denen eine der Antwortkategorien von mehr als 90% der Befragten gewählt wurde (siehe dazu Schnell & Kreuter (2005: 399)). Dadurch werden weitere 14 Items aus der Analyse entfernt. Drittens werden die Variablen, die mit Charakteristiken der Sampling-Points zu tun haben (z. B. „Die Entfernung zum nächsten Bahnhof“), nicht analysiert, da diese u.U. große Varianz in den Antworten aufweisen, die jedoch weniger auf die Interviewer-, sondern vielmehr auf die Sampling-Point-Effekte zurückzuführen wäre.
Weitere Datentransformationen beziehen sich auf die Interviewerauslastung (interviewer workload). Es werden in Anlehnung an Schnell & Kreuter (2005) alle Interviewer gelöscht, die weniger als sechs Interviews durchgeführt haben, was eine bessere Vergleichbarkeit der Ergebnisse gewährleisten sollte.

Die für die Untersuchung ausgewählten Items werden auf folgende Weise behandelt bzw. transformiert[5]:

- Einige Items haben drei Antwortkategorien - „Ja“, „Nein“ und eine weitere Kategorie, die sich einer der beiden zuordnen lässt (z.B. „Nein, hatte noch nie ein Telefon/Fahrrad/Auto“.) Solche dritte Kategorie wird als „Nein“ bzw. “Ja“-Antwort behandelt. In manchen Fällen kann diese dritte Kategorie nicht als „Ja“ oder „Nein“-Antwort, jedoch als Verweigerung bzw. eine „Weiß-Nicht“-Antwort aufgefasst werden. In diesen Fällen wird die dritte Kategorie zu einem fehlenden Wert umkodiert. So können die erwähnten Items auf zwei Arten dichotomisiert werden. Diese Transformation ist bei

[4]Schnell & Kreuter (2005) in Anlehnung an Davis & Scott (1995) haben in ihrer Untersuchung die Kategorien der nominalskalierte Variablen dichotom umkodiert. Da diese Variablen den Annahmen des ANOVA-Modells, das hier vorrangig verwendet wird, nicht genügen und in hoher Anzahl im Datensatz vorhanden sind, verzichte ich auf diese Umkodierung, um den Einfluss dieser Variablen auf die Ergebnisse nicht zu groß werden zu lassen. Die Gesamtzahl der Items reduziert sich dadurch um vier.

[5]Ausführliche Beschreibung zu allen in diesem Kapitel beschriebenen Datentransformationen des DEFECT-Datensatzes im Do-File „cranalysedefect.do“.

fünf Items notwendig.

- Ordinalskalierte Variablen mit mehreren Antwortkategorien werden als metrischskalierte behandelt.

- Intervall- sowie ordinalskalierte Variablen mit sehr vielen Antwortkategorien werden auf die Verletzung der Normalverteilungsannahme hin untersucht. Bei einigen Items können wenig besetzte (meist die letzten 4-5 Kategorien) zu einer Kategorie zusammengefasst werden. Zwar gehen dadurch Informationen verloren, dafür können die Variablen den Annahmen des Modells besser entsprechen. Diese Veränderungen finden bei acht Items statt.
  Für andere können graphische Verfahren, wie z.B. der Normal-Probability Plot oder statistische Tests (-sktest- in STATA) die Verletzung der Normalverteilungsannahme feststellen. Mit Hilfe des Befehls -ladder- sowie -gladder- lassen sich für einige Items Transformationsmöglichkeiten feststellen. So werden fünf Variablen logarithmiert und für eine weitere Variable wird eine Transformation als Quadratwurzel der ursprünglichen Werte durchgeführt.

- Für einige Variablen werden Kodierungsfehler festgestellt. Diese werden entsprechend korrigiert. So gibt es z.B. fehlerhafte Angaben zum Alter des Interviewers (negative Werte), die zu Missings umkodiert werden.

- Ein weiteres Problem stellen die Kategorien „weiß nicht“ und „verweigert“ dar. Diese Kategorien sind bei den ausgewählten Items sehr spärlich besetzt. In vielen Fällen gibt es pro ca. 2000 Befragte einen Verweigerer. Da diese beiden Kategorien mit hohen Werten kodiert sind (7, 8, 97, 98, 99 etc) und ANOVA u.a. auf Mittelwertdifferenzen basiert, ergibt sich ein zu hoher Einfluss der wenigen Verweigerer, die für keine bedeutenden systematischen Effekte in Frage kommen. Auf Grund dieser Überlegung werden diese Kategorien zu fehlenden Werten umkodiert.

- Einige Variablen müssen für die Analysen erstellt werden. So erfordern z.B. die Tests der Hypothesen zu den Alterseffekten die Erstellung der Variablen „Alter des Befragten“, „Absolute Altersdifferenz zwischen dem Interviewer und dem Befragten“ (gruppiert und ungruppiert) etc..

4.2 Frauenstudie

Die Itemauswahl für die Analyse anhand der Frauenstudie gestaltet sich etwas komplizierter, was auf das Design der Untersuchung zurück zu führen ist.
Die Interviewereffekte werden nur am mündlichen Datensatz untersucht. Obwohl man diese Effekte im schriftlichen Drop-Off nicht ausschließen kann (da der Interviewer beim Ausfüllen anwesend war und den ausgefüllten Fragebogen in den meisten Fällen persönlich überreicht bekam), werden diese für den Selbstausfüller als gering betrachtet und hier nicht weiter behandelt.
Daraus folgt, dass die Hypothesen zu den Interviewereffekten an den ausgewählten Items des mündlichen Fragebogens (35 Items) und die zu den Moduseffekten an weiteren Items beider Fragebögen getestet werden, die sehr ähnlich oder identisch formuliert sind und deren Antwortvorgaben weitgehend (fast oder vollständig) übereinstimmen.
Die Items für die Interviewereffekte werden analog zum DEFECT-Datensatz in Typ1-[6] und Typ2-Items[7] aufgeteilt.
Die Fragenauswahl wird zusätzlich zum Itemtyp durch die inhaltliche Bedeutung der Items eingeschränkt. Es werden also in erster Linie Fragen zur sexuellen Gewalt und anderen Gewalterfahrungen gewählt und mit berechneten Statistiken der Kontrollitems verglichen.
Bei der Prüfung der Hypothesen zu Moduseffekten wird versucht, identische oder sehr ähnliche Items in beiden Fragebögen zu finden. Da der Fragenkontext, teilweise Fragenformulierung oder Reihenfolge der Antwortkategorien zwischen den gematchten Items variiert, kann kein direkter Vergleich der Items erfolgen, da eine Kontrolle der zusätzlichen Effekte nicht gewährleistet werden kann. Daher werden verschiedene Anteilswerte zum Testen der Hypothesen verwendet. Weiterhin wird beim Untersuchen der Moduseffekte die bereits bekannte Klassifizierung der Items nach Typen beachtet (relevant für die Hypothese 1b).
Die zweite Hypothese erfordert einen Vergleich der Anteile der beantworteten offenen Fragen. Dadurch werden in beiden Fragebögen offene Fragen jeglicher Art ausgewählt.

[6] 18 Typ1-Items: f110_3, f110_4, f111_3, f111_4, f203_1, f203_3, f204, f300, f303, f309, f310, f501_a1, f602, f604_1, f605, f800, f802, f913. Zur Ausformulierung siehe Anhang A.2.2.
[7] 17 Typ2-Items: f100, f101, f102, f112_a, f112_c, f205, f207, f208, f211, f213, f408, f600, f732, f908, f914, f920_1, f928. Zur Ausformulierung siehe Anhang A.2.2.

Die explorative Datenanalyse zeigt, dass die numerische Kodierung der Antwortkategorien eine wichtige Rolle für die Ergebnisse der Analysen spielt. Den Kategorien „weiß nicht“, „verweigert“ und „keine Angabe“ wurden im Ausgangsdatensatz immer die höchsten Werte zugewiesen (z.B. 6-8, 96-99, 9996-9999). Die hier verwendeten Modelle zielen auf die Vergleiche der Gruppenmittelwerte und der berechneten Varianzanteile. Durch eine solche Kodierung bekommen die unspezifischen Kategorien[8] einen großen Einfluss auf die Ergebnisse, obwohl sie meist nur geringfügig besetzt sind und ein systematischer Effekt oft nicht zu erwarten wäre. Um dieser Besonderheit Rechnung tragen zu können, werden die Analysen einmal ohne Berücksichtigung der unspezifischen Kategorien und einmal mit Berücksichtigung deren (alle Items werden binär umkodiert: gültige Antworten als „1“, unspezifische als „0“) berechnet. Diese Datenveränderungen sind für die Berechnung der Interviewereffekte notwendig[9].

Bei der Untersuchung von Moduseffekten ist außerdem der Item-Nonresponse wichtig. Für die Analyse werden die uneinheitlich kodierten Werte für die Kategorien „keine Angabe“ und „verweigert“ zum Wert „99“ umkodiert und dann die entsprechenden Anteile von Item-Nonresponse ausgerechnet. Da aus den Informationen über den Datensatz kein Unterschied zwischen der inhaltlichen Interpretation der beiden Antwortkategorien festgestellt werden kann (bei den Antwortvorgaben wird nicht immer zwischen „verweigert“ und „keine Angabe“ unterschieden, oft enthalten die Items zusätzlich zu inhaltlichen Kategorien nur „verweigert“ als Antwortmöglichkeit), werden sie als gleichwertig behandelt.
Für einige Analysen werden neue Variablen generiert, z.B. diverse Altersvariablen (analog zu den analytischen Variablen im DEFECT- Datensatz) u.a.[10].
Die abhängigen Variablen des Modells werden in Bezug auf ihre Skalenniveaus auf die gleiche Weise behandelt wie solche im DEFECT-Datensatz.
Ebenfalls analog zu der für den DEFECT-Datensatz durchgeführten explorativen Datenanalyse wird versucht, nicht normalverteilte ordinal- und intervallskalierte Variablen zu transformieren, um eine annähernd normale Verteilung zu erreichen.

[8] Als unspezifische Kategorien werden folgende Antwortvorgaben zusammengefasst: „weiß nicht“, „keine Angabe“, „verweigert“.
[9] Siehe dazu das Do-File „cranalysefrauen.do“.
[10] Ausführlich dazu das Do-File „cranalysefrauen.do“.

Leider ist dies für die von mir untersuchten Items nicht möglich.

Fazit

Dieses Kapitel beschrieb die einzelnen Schritte der Datenaufbereitung, die vor dem Testen der Interviewer- und Moduseffekte vorgenommen wurde. Es wurde unter anderem auf die Auswahl der Items, die Transformationen der Variablen und die eigene Klassifizierung der Items in „Items mit erhöhter Wahrscheinlichkeit der Interviewereffekte“ und „Items mit geringerer Wahrscheinlichkeit der Interviewereffekte“ eingegangen. Alle Datentransformationen wurden in Do-Files beschrieben, auf die an den entsprechenden Stellen verwiesen wird.

5 Das Analysemodell

In diesem Kapitel soll die mathematische und inhaltliche Darstellung der hier verwendeten Analysemodelle getrennt nach untersuchten Effekten erfolgen. Nach einem kurzen Überblick über die Möglichkeiten der Messung von Interviewereffekten wird der Intraklassenkorrelationskoeffizient vorgestellt, der in diesem Zusammenhang als maßgebliches Konzept gilt (Groves et al. (2004): 275). Im Anschluss daran werden die Varianzanalyse und das STATA-Tool -gllamm- beschrieben. Danach werden die Möglichkeiten der Messung von Moduseffekten aufgezeigt.

5.1 Interviewereffekte

5.1.1 Möglichkeiten der Messung von Interviewereffekten

Zur Messung der Interviewereinflüsse werden verschiedene Ansätze verwendet. So konzipieren Groves & Couper (1998) die Interviewereffekte in Form von multivariaten Regressionsmodellen, die die Antwortraten auf der Interviewerebene vorhersagen und aus den signifikanten Werten der Koeffizienten die Bestätigung der Effekte ableiten. Auch Becker & Günther (2004) führen multivariate Analysen des Antwortverhaltens mit Hilfe der logistischen binären Regression durch (Item beantwortet vs. Item nicht beantwortet) und betrachten bestimmte Koeffizientenwerte als eine Bestätigung der gesuchten Effekte.
Huddy *et al.* (1997) untersuchen insbesondere die Einflüsse des Interviewergeschlechts. Zu diesem Zweck transformieren sie alle Itemskalen zu einer 10-Punkte-Skala und vergleichen die standardisierten Antworten der Respondenten miteinander. Um die Interviewereffekte zu berechnen, wird der Mittelwert jeder abhängigen Variable (jedes Items) der Respondenten, die von einem Mann befragt wurden, vom Mittelwert der von den Frauen befragten Respondenten substrahiert. Eine Differenz von 1.0 signalisiert dann z.B., dass die von Frauen befragten Per-

sonen um 10% mehr Äußerungen bejahen, in denen Charakteristiken und Eigenschaften der Frauen positiv bewertet werden.
Johnson & Parsons (1993) dagegen verwenden in diesem Zusammenhang Strukturgleichungsmodelle. Die Vorteile der Methode begründen sie mit der Möglichkeit, multiple Indikatoren für latente Konstrukte zu verwenden und Reliabilität sowie Validität schätzen zu können.
Eine andere Sichtweise der Interviewereffekte, die in der empirischen Forschung eine breite Anwendung gefunden hat und die Grundlage der Berechnungen in dieser Arbeit bildet, ist auf die durch den Interviewer bedingte Variation der Befragungsergebnisse gerichtet. Die Varianzanalyse ermöglicht die Schätzung der Varianz der abhängigen Variable, die durch den Interviewer und weitere Faktoren herbeigeführt wird und die anschließende Berechnung des Intraklassenkorrelationskoeffizienten. Der Koeffizient wurde von Kish (1962) vorgeschlagen und bietet im Vergleich zu anderen Messungen, z.B. der Nutzung von F-Ratios als Messung der Interviewervarianz, die Möglichkeit, Vergleiche zwischen Studien mit unterschiedlichen Stichprobengrößen durchzuführen (Freeman & Butler (1976)). Auf diese Weise berechnete Interviewereinflüsse finden sich u.a. bei Tucker (1983), Groves & Magilavy (1986) sowie Schnell & Kreuter (2005).

5.1.2 Der Intraklassenkorrelationskoeffizient

Die Interviewervarianz wird definiert „as a component of the total variance per respondent; thus $s^2 = s_a^2 + s_b^2$, where s_a^2 is the „between interviewer“ and s_b^2 the „whithin interviewer“ component; $p = s_a^2/s^2$ is the proportion of the interviewer effect“ (Kish (1962): 92). Die Interviewervarianz kann zusätzlich zur Stichprobenvarianz die Varianz der abhängigen Variable beeinflussen. Dieser Einfluss ist im Gegensatz zu dem der unkorrelierten Zufallsfehlern (random errors) in den gewöhnlichen Berechnungen der Varianz der abhängigen Variable nicht enthalten.
Das Modell von Kish (1962) basiert auf der Annahme der Zufallsauswahl der Interviewer aus einem sehr großen Pool von potentiellen Interviewern. Jeder Interviewer produziert seinen durchschnittlichen „Interviewer Bias“, der die Antworten der von ihm befragten Personen beeinflusst. Das gesamte Modell betrachtet den Einfluss der Zufallsauswahl dieser Verzerrungen (biases) auf die Varianz des Stichprobenmittelwerts.

Der Intraklassenkorrelationskoeffizient[1] berechnet sich als

$$p = s_a^2/(s_a^2 + s_b^2) \tag{5.1}$$

In komplexen Stichproben ist außerdem die Varianz des Sampling-Points bzw. der Region zu berücksichtigen $(s_p^2)^2$. Diese wird bei der Berechnung des Intraklassenkorrelationskoeffizienten für die Interviewervarianz zur totalen Varianz addiert. Der Koeffizient berechnet sich dementsprechend als:

$$p = s_a^2/(s_a^2 + s_b^2 + s_p^2) \tag{5.2}$$

Die Trennung von Varianzanteilen (z.B. der Stichprobenvarianz von der Interviewervarianz) kann nur mit Hilfe komplizierter Forschungsdesigns erfolgen. Da der DEFECT-Datensatz das dafür erforderliche Design mit interpenetrierenden Stichproben aufweist, kann diese Trennung der Effekte bei einigen Tests im Gegensatz zur Frauenstudie durchgeführt werden.
Die Berechnung des Intraklassenkorrelationskoeffizienten erfordert die Schätzung der entsprechenden Varianzanteile. In der vorliegenden Untersuchung werden dafür Varianzanalysen verschiedener Art verwendet. Abhängig davon, ob die Einflussfaktoren als eine Zufallsauswahl aller möglichen Faktoren oder als feste Faktoren bzw. Vollerhebung der Faktoren betrachtet werden, wird Random Effects oder Fixed Effects ANOVA berechnet. So wird z.B. der Interviewer als Zufallseffekt (random effect) betrachtet, da die Ergebnisse auf weitere Interviewer als Einflussfaktor verallgemeinert werden und das Interviewergeschlecht als fester Effekt (fixed effect) definiert, da es nur die zwei verwendeten Ausprägungen aufweist.

5.1.3 Varianzanalyse (ANOVA)

ANOVA (Analysis of Variance) gehört zu den Techniken der GLM (generalized linear models). Dieses Verfahren erlaubt die Bestimmung der Wahrscheinlichkeit, dass die unabhängige Variable andere Ergebnisse liefert, als durch Zufallsschwankungen hervorgerufen (Iversen & Norpoth (1976)). Das Verfahren schätzt mit Hilfe der F-Tests die statistische Signifikanz der Beziehung zwischen kategorialen

[1] Hier die Notation von Kish (1962).
[2] s_p^2 ist die eigene Notation für die Varianz des Sampling-Points.

unabhängigen Variablen und einer kontinuierlichen abhängigen Variable. Die Hypothesenprüfung erfolgt durch den Vergleich der Mittelwerte zwischen den Gruppen der Beobachtungen, die auf unterschiedliche Weise beeinflusst werden. Dafür werden die Beobachtungen entsprechend den Ausprägungen der unabhängigen Variable gruppiert.
Die Differenzen *zwischen* den Gruppenmittelwerten entstehen laut den Annahmen der ANOVA als Ergebnis des systematischen Einflusses der gruppierenden Variable und der unsystematischen Gruppendifferenzen (random error). Die Differenzen zwischen den Beobachtungen *innerhalb* jeder Gruppe sind hingegen auf unsystematische individuelle Differenzen (random error) zurück zu führen (vgl. Newton (1999): 202). Diese Abweichungen der beobachteten Werte vom Gruppenmittelwert einerseits und vom totalen Mittelwert (grand mean) andererseits sowie die Abweichungen der Gruppenmittelwerte vom totalen Mittelwert werden für Berechnungen der Varianz *zwischen* den und *innerhalb* der Gruppen und somit auch für die Berechnungen des Intraklassenkorrelationskoeffizienten verwendet.

Das formale Modell für die einfaktorielle Varianzanalyse sieht wie folgt aus:

$$Y_{ij} = \mu + U_i + R_{ij} \tag{5.3}$$

Laut diesem Modell kann man die beobachteten Werte der abhängigen Variable (Y_{ij}) als die Summe aus verschiedenen Komponenten betrachten: einer Konstante (dem Mittelwert über alle Beobachtungen, μ), dem Effekt der unabhängigen Variable (Effekt durch die Gruppenzugehörigkeit, U_i) und dem Effekt aller anderen, im Modell nicht ausdrücklich genannten Variablen (residuellem Effekt, R_{ij})[3].
Das Modell basiert auf folgenden Annahmen:

- μ, U_i und R_{ij} sind voneinander unabhängig
- E(U_i)=0 und Var(U_i)=δ_a^2
- E(R_{ij})=0 und Var(R_{ij})=δ_R^2

Die Populationsvarianz zwischen den Gruppen (den Makroeinheiten, z.B. zwischen den Interviewern) kann nicht einfach der beobachteten Varianz zwischen

[3]Notationen entnommen Snijders & Bosker (1999).

den Gruppen gleichgesetzt werden, da die Residuen der untersten Ebene (R_{ij}) einen gewissen Beitrag zur Varianz zwischen den Gruppen leisten. Die beobachtete Varianz ist jedoch ein guter Schätzer der Populationsvarianz. Die geschätzten Werte für die Varianzen fließen in die Berechnung des Intraklassekorrelationskoeffizienten ein (Snijders & Bosker (1999)).
Das Modell kann um weitere Einflussfaktoren oder Faktorenarten erweitert werden, so dass man von einer faktoriellen ANOVA oder ANCOVA[4] spricht. Diese Faktoren können sowohl der bereits enthaltenen Ebene, als auch einer niedrigeren bzw. höheren Ebene angehören[5].

Formulierung und Prüfung der Annahmen der ANOVA

Bevor jedoch die Varianzanalyse durchgeführt werden kann, müssen die für einen F-Test erforderlichen Annahmen der Normalverteilung und Unabhängigkeit der Residuen sowie der Varianzhomogenität überprüft werden.
Die *Normalverteilungsannahme* ist für die statistische Herleitung und die Geltung der Prüfgröße F notwendig (Schnell (1994)). Dies kann u.a. mit einem Kolmogorov-Smirnov-Test, Skewness-Kurtosis-Test[6] oder graphisch mit einem Normal-Probability-Plot[7] erfolgen. In meiner Untersuchung fanden der Normal-Probability-Plot und Skewness-Kurtosis-Test Anwendung.
Die Überprüfung der *Varianzhomogenitätsannahme* ist für sozialwissenschaftliche Erhebungen noch wichtiger, da diese oft ungleiche Zellenbesetzung auf Grund vielfältiger Faktoren, z.B. wegen Item-Nonresponse haben (die sogenannten unbalanced designs). Diese Annahme lässt sich z.B. mit einem Bartlett-Test oder einem Test auf Gleichheit der Standardabweichungen (-sdtest- in STATA) überprüfen.
Die Verletzung der Annahme der *Residuenunabhängigkeit* ist für die Ergebnisse der Varianzanalysen kritischer als die Verletzung anderer Annahmen. Die Annahme gilt als erfüllt, wenn die Stichprobe eine Zufallsstichprobe aus der Po-

[4] ANCOVA wird verwendet, falls eine Erweiterung um kontinuierliche Variablen stattfindet.
[5] Auszug aus der interaktiven STATA-Hilfe zur ANOVA: „anova fits analysis-of-variance (ANOVA) and analysis-of-covariance (ANCOVA) models for balanced and unbalanced designs, including designs with missing cells; for repeated measures ANOVA; and for factorial, nested, or mixed designs“.
[6] In STATA als -ksmirnov- und -sktest- implementiert. Weitere Prüfmöglichkeiten bieten der Shapiro-Wilk-Test (-swilk-) oder Shapiro-Francia-Test (-sfrancia-) in STATA.
[7] bei STATA als -qnorm- implementiert

pulation darstellt und die Zuweisung zu den experimentellen Gruppen zufällig geschieht (Schnell (1994)). Besonders die zweite Bedingung kann in den sozialwissenschaftlichen Befragungen nicht immer erfüllt werden. Obschon dies für CATI-Befragungen relativ leicht zu handhaben ist, kann eine zufällige Zuweisung der Befragten zu den Interviewern in Face-to-Face-Befragungen schnell sehr teuer werden und findet daher selten Anwendung. In den Sozialwissenschaften kann die Zuordnung zu Kategorien meist nicht durch Randomisierung erfolgen (z.B. können die Befragten nicht zufällig den Geschlechtskategorien zugeordnet werden). Auf Grund dessen kann man nicht definitiv sagen, dass die Differenz duch die jeweilige Gruppenzugehörigkeit entstanden ist (Newton & Rudestam (1999)).
Für die Überprüfung der Annahme eignet sich u.a. ein Residual*Predicted-Plot (Schnell (1994)) oder ein Histogram der Residuen (Iversen & Norpoth (1976)).
Trotz der Gefahren, die durch die Verletzungen der Annahmen entstehen, wird die Varianzanalyse in Bezug auf diese Verletzungen als robustes Verfahren gesehen (Iversen & Norpoth (1976), Schnell (1994)). Am unkritischsten ist dabei die Verletzung der Normalverteilungsannahme. Inwiefern eine Verletzung die Ergebnisse systematisch verzerrt, ist nicht einfach abzuschätzen. Eine festgestellte Verletzung kann man jedoch durch Variablentransformationen oder Anwendung nichtparametrischer Verfahren beheben (Schnell (1994))[8].
Die Schätzung der Varianzanteile mit ANOVA kann zudem weitere Schwierigkeiten bereiten. Falls es sich um eine zwei- oder mehrfaktorielle ANOVA mit Interaktionseffekten handelt, ist die Interpretation dieser Interaktionseffekte nicht einfach. Wenn die Einflussvariablen zusätzlich nicht orthogonal sind, lässt sich der gemeinsame Varianzanteil nicht eindeutig zuordnen. Die Lösung für dieses Problem liegt in den hierarchischen Analysen (Newton & Rudestam (1999)), die im nächsten Abschnitt erläutert werden.

5.1.4 Gllamm

Der Grund für die Nutzung von Mehrebenenmodellen (hierarchischen Modellen) ist in erster Linie die verletzte Annahme der Unabhängigkeit von Beobachtungen, die zu einem Klumpen gehören. Mit diesen Analysemodellen soll die hierarchische Struktur der Daten und die Abhängigkeit der Beobachtungen eines

[8]Siehe dazu Kapitel 4 „Datenaufbereitung“.

Klumpens berücksichtigt werden. Die Respondenten bilden dabei die niedrigste Ebene, Interviewer die mittlere und Sampling-Points die höchste Ebene. Außerdem ermöglichen die hierarchischen Modelle eine Trennung der Varianzen nach Analyseebenen.
Zwar kann auch ANOVA in STATA hierarchische Strukturen modellieren, jedoch sind ihre Möglichkeiten, die Verteilungen der abhängigen Variablen zu berücksichtigen, eingeschränkt. Viele der getesteten abhängigen Variablen in dieser Arbeit sind dichotom oder ordinalskaliert. Durch die Benutzung von -gllamm- müssen die ordinalskalierten Variablen nicht als intervallskalierte behandelt werden; für die binären Variablen enthält das Programm ebenfalls eine Option, die das Skalenniveau berücksichtigt.
Das oben beschriebene formale Modell der einfaktoriellen Varianzanalyse[9] bezeichnet man im Rahmen der Mehrebenenmodelle als „leeres Modell“ (vgl. Snijders & Bosker (1999)). Es bildet die Grundlage der komplizierteren Modelle, die um weitere Ebenen oder Kovariaten erweitert werden. Der Intraklassenkorrelationskoeffizient berechnet sich für das leere Modell analog zur Berechnung in ANOVA.
Die hier verwendeten hierarchischen Modelle unterscheiden sich von gewöhnlichen Regressionsmodellen dadurch, dass sie mehr als einen Fehlerterm für jede Ebene enthalten. In Mehrebenenmodellen enthält die abhängige Variable demnach den individuellen und den Gruppenaspekt. Dasselbe gilt für unabhängige Variablen, die ebenfalls außer der Aspekte der eigenen Ebene diejenigen der höheren Ebenen enthalten können (Snijders & Bosker (1999)).

Eine klassische multiple Regression kann als ein einfaches Mehrebenenmodell aufgefasst werden, in dem die abhängige Variable völlig durch die Einflussfaktoren der individuellen und der Gruppenebene erklärt werden. Die eigentliche hierarchische Struktur jedoch kann auf diese Weise nicht abgebildet werden. Die zusätzlichen Effekte der Datenstruktur sollen durch die zwischen den Gruppen variierenden Regressionskoeffizienten wiedergegeben werden (Snijders & Bosker (1999)). Es kann sich dabei um variierende Intercepts oder Slopes handeln.

[9] Siehe Gleichung 5.3.

Das *Random-Intercept-Modell* wird mit folgender Formel[10] beschrieben:

$$Y_{ij} = \gamma_{00} + \gamma_{10} * x_{ij} + U_{0j} + R_{ij} \tag{5.4}$$

In dieser Gleichung ist Y_{ij} die Ausprägung der abhängigen Variable beim Individuum i in der Gruppe j, x_{ij} die Erklärungsvariable auf der individuellen Ebene, γ_{00} der durchschnittliche Intercept (grand mean), γ_{10} der Regressionskoeffizient von X, U_{0j} der Effekt der Zugehörigkeit zur Gruppe (Zufallseffekt der zweiten Ebene) und R_{ij} das Residuum der ersten Ebene.
Für die Zufallsvariablen R_{ij} und U_{0j} wird angenommen, dass sie jeweils Erwartungswerte von 0 und Varianzen von $\mathrm{Var}(R_{ij})=\delta^2$ bzw. $\mathrm{Var}(U_{0j})=\tau_0^2$ sowie jeweils unabhängige Verteilungen gegeben die Werte der unabhängigen Variable X aufweisen. Außerdem sollen Residuen R_{ij} und U_{0j} aus normalverteilten Populationen stammen. Die Varianzen von R_{ij} sollen in den jeweiligen Gruppen konstant sein. Diese zwei Arten von Residuen erklären einen Teil der Varianz der abhängigen Variable auf zwei (ggf. mehr) Ebenen, der nicht durch die im Modell explizit enthaltenen Faktoren erklärt wird. „This partition of unexplained variability over the various levels is the essense of hierarchical random effects models“ (Snijders & Bosker (1999): 48).
Für diese beiden Ebenen lassen sich zwei residuale Varianzen berechnen: für die erste Ebene als $\delta^2/(\delta^2 + \tau_0^2)$ und für die zweite Ebene als $\tau_0^2/(\delta^2 + \tau_0^2)$.
Die Kovarianz oder Korrelation zwischen zwei Individuen aus einer Gruppe lassen sich zum Teil durch ihre Werte auf der unabhängigen Variable X erklären, die restliche Varianz bleibt unerklärt. Diese residuale unerklärte Korrelation zwischen den Beobachtungen eines Klumpens wird durch den *residualen Intraklassenkorrelationskoeffizienten* ausgedrückt. Im Unterschied zum bereits bekannten Korrelationskoeffizienten wird bei dieser Berechnung für den Einfluss von X kontrolliert.
Es können weitere Variablen auf allen Ebenen sowie zusätzliche Ebenen in das Modell aufgenommen werden. Für meine Untersuchung ist vor allem die Berücksichtigung der dritten Ebene, der Sampling-Points, und verschiedener unabhängiger Variablen auf der ersten und zweiten Ebene wichtig (z.B. Geschlecht des In-

[10]Notationen entnommen Snijders & Bosker (1999): 41.

terviewers, Altersdifferenzen zwischen Interviewern und Befragten etc.)[11].
Durch die Einbeziehung der dritten Ebene erweitert sich das Modell u.a. um die Residuen der dritten Ebene und ihre Varianz φ^2. Diese wird bei der Berechnung der residualen Varianzen sowie des Intraklassenkorrelationskoeffizienten auf die totale Varianz aufaddiert[12]. Somit „übernimmt“ die neue Ebene die Erklärung eines Teils der Varianz der abhängigen Variable, der früher durch die erste oder zweite Ebene (oder durch beide) erklärt wurde.
Es gibt eine weitere Art der hierarchischen Modelle, das *Random Slope Modell*, das eine Heterogenität der Regressionsgeraden zwischen den Gruppen, die sogenannte „group-by-covariate interaction“ (Snijders & Bosker (1999)) modelliert. Dieses Modell wird hier aus theoretischen Gründen nicht verwendet, da die möglichen Interaktionen der höheren Ebenen mit Erklärungsvariablen der unteren Ebenen nicht im Mittelpunt der Untersuchung stehen.
Die Gllamm-Modelle können bedingt durch die Datenqualität nur für den DEFECT-Datensatz gerechnet werden. Bei allen in diesem Rahmen getesteten Hypothesen wird die Verteilung der abhängigen Variablen beachtet und entsprechende Optionen des Befehls angewendet. Außerdem findet eine Kontrolle für den Modus der Befragung statt.

5.2 Moduseffekte

Moduseffekte kann man generell so interpretieren, dass verschiedene Untersuchungsmodi unterschiedliche Ergebnisse liefern.
Die Unterschiede können u.a. durch die im theoretischen Teil bereits angesprochenen *modusabhängigen Einflüsse des Interviewers* oder Abwesenheit vs. Vorhandensein der Interviewereinflüsse entstehen. Daher können die Moduseffekte zuerst in Form von signifikant unterschiedlichen Intraklassenkorrelationskoeffizienten, z.B. für Face-to-Face- gegenüber CATI-Befragungen modelliert werden. Die Berechnung des Koeffizienten erfolgt auf die gleiche Weise wie für die Interviewereffekte. Auf diese Weise wird die zweite Hypothese zu den Moduseffekten

[11] Mehr zu den Einflussvariablen siehe im Kapitel 6.1 „Testen der Interviewereffekte am DEFECT-Datensatz“.
[12] φ^2 entspricht s_p^2, δ^2 entspricht s_a^2 und τ_0^2 entspricht s_b^2 in der Gleichung 5.2.

am DEFECT-Datensatz getestet[13]. Da es unmöglich ist, für die Einflüsse des variierenden Fragebogens in der Frauenstudie zu kontrollieren, wird auf diese Modellierung der Moduseffekte für die Studie verzichtet.

Des Weiteren kann man den jeweiligen Modus als gruppierende Variable betrachten und die Unterschiede zwischen den Antwortverteilungen mit einer einfaktoriellen ANOVA feststellen. Hier können ebenfalls die Intraklassenkorrelationskoeffizienten zum Vergleich der Effekte verwendet werden. Der Unterschied zu der Berechnung der Interviewereffekte besteht darin, dass hier die Beobachtungen nicht nach dem Interviewer, sondern nach dem Modus gruppiert werden und der Intraklassenkorrelationskoeffizient daher die Ähnlichkeit der Beobachtungen innerhalb eines Modus misst. Auch dieser Ansatz kann aus den bereits dargestellten Gründen nur für Hypothesentests am DEFECT-Datensatz angewendet werden.

Drittens kann man die Modi im Hinblick auf die Datenqualität untersuchen. So lassen sich die Ergebnisse z.B. in Bezug auf den Item- oder Unit-Nonresponse, die Tendenz zur Wahl der extremen oder letzten Antwortkategorien etc. untersuchen. So werden hier u.a. die Recency-Effekte am DEFECT-Datensatz untersucht ($H3$). Der Test basiert auf dem Vergleich der Anteile der gewählten Antwortkategorien im CATI- und im Mail-Survey.
Die Moduseffekte in der Frauenstudie können nicht als Vergleich der Intraklassenkorrelationskoeffizienten modelliert werden. Daher werden alle Hypothesen als Vergleich verschiedener Anteilswerte modelliert und geprüft. Also werden die Anteile des Item-Nonresponse im mündlichen und schriftlichen Interview für neutrale ($H1a$) und sensitive ($H1b$) Items verglichen und auf Signifikanz der Differenzen geprüft. Nachfolgend werden in der $H2$ die Anteile der beantworteten offenen Fragen in Viktimisierungssurveys in beiden Befragungen getestet.

Fazit

Die Interviewer- sowie Moduseffekte können mit Hilfe verschiedener mathematischer Modelle, wie z.B. der multivariaten Regressionsmodelle, Strukturgleichungsmodelle oder Varianzanalysen dargestellt werden. Im vorangegangenen Kapitel

[13] Diese Hypothese lautet: „In Face-to-Face-Interviews sind die Interviewereffekte größer als in telefonischen Befragungen“.

wurden die für diese Arbeit relevanten Konzepte beschrieben. Es handelte sich dabei um den Intraklassenkorrelationskoeffizienten und die Varianalyse sowie hierarchischen Modelle, die in STATA mit -gllamm- geschätzt werden. Weiterhin wurden Besonderheiten der Modellierung der Moduseffekte anhand der beiden Datensätze angesprochen.
Die genannten Modelle werden in den nachfolgenden Kapiteln zur Formulierung und Prüfung der Hypothesen verwendet.

6 Formulierung und Prüfung der Hypothesen zu Interviewereffekten

6.1 Testen der Interviewereffekte anhand des DEFECT-Datensatzes

6.1.1 Modelle

Als eine der wichtigsten Verzerrungsursachen der Befragungsergebnisse gilt der Interviewer. Dabei können die Ergebnisse von zwei Faktorengruppen beeinflusst werden. Zum einen spielen Verhalten, Gebaren und Einstellungen des Interviewers zum Interviewgegenstand eine wichtige Rolle, zum anderen sind es seine sichtbaren Eigenschaften, wie z.B. Geschlecht, Alter, Rasse etc., die sich auf die Antworten der Befragten verzerrend auswirken können.
Der Einfluss der ersten Faktorengruppe ist nicht ohne ein spezielles Design zu schätzen. Bei den in den Sozialwissenschaften verbreiteten Sekundäranalysen entsteht oft das Problem, dass die Datensätze die dafür benötigten Angaben zu den Interviewern und Befragten nicht enthalten. Allerdings geht man davon aus, dass ein intensives Interviewertraining und eine gute Interviewerkontrolle mögliche Verzerrungen dieser Art weitgehend ausgleichen können (vgl. dazu Hansen *et al.* (1951), Groves & Magilavy (1986) und Groves *et al.* (2004)).
Ein breiteres und wohl interessanteres Forschungsfeld bieten die Verzerrungen durch die Faktoren der zweiten Gruppe. Diese werden in meiner Arbeit als folgende Hypothesen formuliert und getestet:

- $H1$: Interviewer haben einen verzerrenden Einfluss auf die Befragungser-

gebnisse. Dieser Einfluss ist auf sensitive Fragen zur Viktimisierung sowie auf emotional geladene, schwierige, offene oder Einstellungsfragen größer.

- $H2$: Männliche Interviewer führen bei Viktimisierungsbefragungen zu größeren Interviewereffekten als weibliche Interviewer. Diese Effekte sind bei „anfälligen“ Items größer als bei „weniger anfälligen“[1].
- $H3$: In gleichgeschlechtlichen Dyaden sind Interviewereffekte geringer als in verschiedengeschlechtlichen Dyaden.
- $H4$: Männliche Interviewer haben bei Viktimisierungsbefragungen einen größeren Einfluss auf weibliche Befragte als weibliche Interviewer.
- $H5$: Weibliche Interviewer haben bei Viktimisierungsbefragungen einen größeren Einfluss auf männliche Befragte als männliche Interviewer.
- $H6$: Der Interviewereffekt wird mit steigender Altersdifferenz zwischen dem Interviewer und dem Befragten größer.
- $H7$: Interviewereffekte sind in der Gruppe der negativen Altersdifferenzen größer als in der Gruppe der positiven Altersdifferenzen.

6.1.2 Ergebnisse

Hypothese 1: Interviewer haben einen verzerrenden Einfluss auf die Befragungsergebnisse. Dieser Einfluss ist auf sensitive Fragen zur Viktimisierung sowie auf emotional geladene, schwierige, offene oder Einstellungsfragen größer

In der ersten Hypothese wird der allgemeine Einfluss des Interviewers getestet. Wenn es diesen Einfluss nicht gäbe, bliebe noch die Möglichkeit, nach den Effekten einzelner Interviewereigenschaften, wie z.B. des Geschlechts, Alters oder Bildungsstandes in den Subgruppen zu suchen.
Wie bereits im Kapitel 5 erklärt, wird für die Analyse der Effekte die Random-Effects-ANOVA (für das einfaktorielle Modell als -loneway- oder -oneway- in STATA implementiert) und der von Kish (1962) vorgeschlagene Intraklassenkorrelationskoeffizient *rho* verwendet. Eine weitere Bezeichnung für diesen Koeffizienten

[1] Die „anfälligen“ Items wurden im Kapitel 4.1 bereits als Typ1-Items, die „weniger anfälligen“ als Typ2-Items definiert.

- p_{int} - findet sich u.a. bei Groves *et al.* (2004: 276) und wird in dieser Arbeit verwendet. Die erste Hypothese wird - wie auch die anderen - ahhand von 16 Fragen aus dem DEFECT-Fragebogen getestet.

Ergebnisse

Für die Durchführung der Tests werden die Beobachtungen nach den jeweiligen Interviewern gruppiert. Danach wird der Intraklassenkorrelationskoeffizient für jedes Item berechnet. Der Datensatz enthält die Intervieweridentifikationsnummern für die drei Face-to-Face-Befragungen sowie für die telefonische Befragung. Somit kann man die Interviewereffekte mit und ohne Kontrolle für den Modus der Befragung ausrechnen.
Ohne Kontrolle für den Modus können bei 14 der 16 getesteten Items signifikante Ergebnisse verzeichnet werden. Als nicht signifikant erweist sich der Interviewereffekt für das Item f33a „Von einem Fremden geschlagen oder verletzt innerhalb der letzten 12 Monate“ und f43 „Körpergröße in cm“. Der Intraklassenkorrelationskoeffizient nimmt für die signifikanten Items Werte zwischen 0.018 und 0.158 an. Im Durchschnitt beträgt der Koeffizient für die untersuchten Items (inklusive der nicht signifikanten) 0.094.
Diese Werte werden in der Literatur als bedeutend angesehen, da sie in Kombination mit der Interviewerauslastung zu starken Designeffekten führen können: „[...] even small p_{int}'s when combined with large workloads can produce major increases in variances of sample statistics“ (Groves *et al.* (2004): 277).
Tabelle 6.1 liefert einen Überblick über die Ergebnisse der Koeffizientenwerte ohne Kontrolle für den Befragungsmodus[2].
Durch die Kontrolle des Modus steigen die Werte für Face-to-Face-Befragungen und sinken für das CATI-Survey, was von stärkeren Effekten in ersten und schwächeren im zweiten Fall zeugt. Während in persönlichen Interviews 15 Items signifikante Intraklassenkorrelationskoeffizienten zwischen 0.016 und 0.218 mit einem Durchschnitt von 0.157 verzeichnen, sind in der CATI-Befragung nur sieben Koeffizienten signifikant und liegen zwischen 0.009 und 0.070. Da die Moduseffekte in einem eigenen Kapitel ausführlich behandelt werden, soll hier nicht weiter auf

[2]Für alle Tests in dieser Arbeit wird ein Signifikanzniveau von 95% festgelegt. Die Abkürzung „n.s.“ weist auf ein nicht signifikantes Ergebnis auf diesem Niveau hin. * hinter einem Ergebnis bedeutet, dass dieses bei Null trunkiert wurde.

die Unterschiede eingegangen werden.

Tabelle 6.1: p_{int} für Typ1- vs. Typ2-Items

Typ1-Items	p_{int}	Typ2-Items	p_{int}
f10	0.137	f1	0.148
f13_2	0.122	f31	0.146
f13_4	0.158	f38_4	0.060
f14_2	0.118	f43	0.007 (n.s.)
f14_3	0.127		
f25	0.124		
f33a	0.000* (n.s.)		
f34	0.089		
f35	0.119		
f35a	0.063		
f42	0.071		
lf44	0.018		
$\overline{p_{int}}$	0.096	$\overline{p_{int}}$	0.090

Die aufgeführten Ergebnisse erlauben es, die Hypothese für beide Itemtypen als akzeptiert zu betrachten. Andererseits lassen sich die vermuteten Unterschiede zwischen den für die Effekte „anfälligen“ und „weniger anfälligen“ Items nicht bestätigen, da der Vergleich der durchschnittlichen Intraklassenkorrelationskoeffizienten in beiden Gruppen nur einen sehr geringen Unterschied zeigt.
Die Ergebnisse des Bartlett-Tests bestätigen die Verletzung der Annahme der Varianzenhomogenität bei vielen Items. Da man bei ANOVA aber von einem gegenüber der Verletzung dieser Annahme robustem Verfahren ausgeht[3], können die Ergebnisse der Analysen zum Testen der Hypothesen verwendet werden.

Die Varianzanalyse stellt kein optimales Modell für diese Daten dar. Erstens werden die Verteilungsannahmen der abhängigen Variablen für die meisten Items nicht erfüllt. Zweitens entspricht ihr Skalenniveau nicht dem für das Modell angenommenen Intervallskalenniveau. Drittens kann die Varianz des Interviewers nicht von der des Sampling-Points getrennt werden. Daher wird diese Hypothese anschließend mit Hilfe von -gllamm- erneut geprüft.

[3]Siehe dazu Kapitel 5.

Das neue Modell beinhaltet drei Ebenen, die die hierarchische Struktur der Daten wiedergibt - Ebene der Befragten (unterste Ebene), die der Interviewer (mittlere Ebene) und der Sampling-Points (die oberste Ebene). Da es sich hier um die Interviewereffekte handelt, sollte die durch den Interviewer erzeugte Varianz in den Antworten von der Sampling-Point-Varianz getrennt werden.
Das Modell ermöglicht durch zusätzliche Befehlsoptionen[4] der dichotomen oder ordinalen Verteilung der abhängigen Variablen Rechnung zu tragen. Also sollten die neu berechneten Intraklassenkorrelationskoeffizienten bessere Schätzungen der Interviewereffekte liefern als die Ergebnisse der ANOVA.
Sowohl die Interviewer als auch die Sampling-Points werden hier als Zufallsauswahl aus den Populationen aller Interviewer und aller Sampling-Points betrachtet. Daher wird das Random-Intercept-Modell berechnet. „The random intercept can be interpreted as the combined effect of all unobserved subject-specific covariates, often referred to as *unobserved heterogeneity*“ (Rabe-Hesketh & Everitt (2004))[5]. Auch in diesem Fall wird für den Befragungsmodus kontrolliert, so dass die Ergebnisse auch zum Testen der Moduseffekte verwendet werden können.

Ergebnisse von Gllamm

Die sogenannte „condition number“ des Modells hat für alle Items Werte, die auf ein gut spezifiziertes Modell hinweisen (zwischen 1.65 und 571) (vgl. Rabe-Hesketh & Everitt (2004)).
Zuerst werden die Intraklassenkorrelationskoeffizienten für die Ebene der Interviewer und die Ebene der Sampling-Points getrennt berechnet[6]. Der Einfluss der Interviewerebene ist insbesondere in der persönlichen Befragung deutlich stärker als der Einfluss der Sampling-Point-Ebene. Dies ergibt der Vergleich von Koeffizienten, die den Anteil der erklärten Varianz durch die jeweilige Ebene darstellen. Für 15 Items sind die Koeffizienten der Intervierebene höher, in einem Fall sind diese für beide Ebenen nicht signifikant. Ein gepaarter T-Test bestätigt die Si-

[4] Folgende Optionen werden verwendet: family (binomial), link (logit) oder link (ologit).
[5] Rabe-Hesketh & Everitt (2004) bezieht sich in diesem Satz auf Zeitreihenanalysen, bei denen die erste Ebene die Beobachtungen an einem Individuum zu einem gegebenen Zeitpunkt und die zweite Ebene das Individuum bilden. Somit kann hier die Individuumebene der in meinem Modell vorhandenen Interviewerebene gleichgesetzt werden.
[6] Alle hier ausgerechneten Koeffizienten finden sich in den Tabellen A.1 und A.2 im Anhang A.

gnifikanz der Unterschiede zwischen den Koeffizienten. Dieser Test zeigt, dass die Annahme stärkerer Interviewereffekte im Vergleich zu Sampling-Point-Effekten besonders für Face-to-Face-Befragungen zutrifft. Ein Vergleich der Ergebnisse zwischen den persönlichen und telefonischen Befragungen findet beim Testen der Moduseffekte anhand des DEFECT-Datensatzes im nächsten Kapitel statt.
Eine weitere Vergleichsmöglichkeit ergibt sich für die Koeffizientenpaare, die mit Hilfe von ANOVA und -gllamm- berechnet werden[7]. In der Face-to-Face-Befragung liefert -gllamm- in neun Fällen höhere Intraklassenkorrelationskoeffizienten für die ausgewählten 16 Items. Bei sechs weiteren ist der von ANOVA berechnete Koeffizient höher und in einem Fall gab es keinen Unterschied. Für die CATI-Befragung gibt es bei fünf Itempaaren keinen Unterschied, ebenfalls bei fünf einen höheren Koeffizienten in der ANOVA-Gruppe und bei sechs weiteren Itempaaren den höheren Koeffizientenwert in der Gllamm-Gruppe. Somit kommen die Interviewereffekte, die mit -gllamm- berechnet werden, etwas stärker zur Geltung als die Ergebnisse der Varianzanalyse.
Bezüglich des Skalenniveaus der abhängigen Variablen werden im letzten Schritt die ANOVA-Koeffizienten mit Gllamm-Koeffizienten getrennt für Variablen verschiedener Skalleniveaus verglichen. Ein signifikanter Unterschied ergibt sich für dichotome Items. Für sechs der sieben Items sind die Gllamm-Koeffizienten signifikant höher, in einem Fall gibt es keinen Unterschied zwischen den beiden Berechnungsarten. Dies zeigt, dass die verletzte Annahme der intervallskalierten abhängigen Variable zur Unterschätzung der Interviewereffekte führt. Dieses Problem kann mit Hilfe von -gllamm- gelöst werden.

Die entsprechende Diagnostik wird an das Skalenniveau angepasst. Dies betrifft vor allem die Überprüfung der Residuen verschiedener Ebenen. Während bei kontinierlichen abhängigen Variablen die Residuen aller drei Ebenen normalverteilt sein sollen (Rabe-Hesketh & Everitt (2004)), gilt diese Annahme für Residuen der dichotomen oder ordinalen Variablen nur für die Residuen der zweiten und dritten Ebene. Bei dichotomen Variablen folgen die Residuen der ersten Ebene (die sogenannten Pearson's Residuen) der logistischen Verteilung. Die Normalverteilungsannahme der Residuen wird graphisch mit Hilfe von Kerndichteschätzern mit eingezeichneten Normalverteilungskurve sowie Normal-Probability-Plots überprüft.

[7]Die Koeffizienten finden sich in den Tabellen A.1 und A.2 im Anhang A.

Die Residuen bei kontinuierlichen Variablen sind auf allen drei Ebenen annähernd normalverteilt. Diese Annahme wird am seltensten bei der CATI-Befragung verletzt. In der Face-to-Face-Befragung ist sie für die Hälfte der kontinuierlichen Variablen (Items f14_2, f14_3 und f38_4) für Residuen aller drei Ebenen verletzt. Für vier von sieben binär kodierten abhängigen Variablen der Face-to-Face-Befragung werden annähernd normalverteilte Residuen auf der Ebene des Interviewers und des Sampling-Points festgestellt. Bei einem Item wird die Annahme auf der Interviewerebene (f34), bei zwei weiteren auf Ebene zwei und drei verletzt (f33a und f35a). Darüber hinaus können beim Item f35a auf jeder Ebene Ausreißer festgestellt werden[8], die insbesondere auf der Interviewerebene auf Grund ihrer Anzahl eine Rolle spielen können. Die Ausreißer spielen für die restlichen Items dagegen keine oder eine sehr geringe Rolle. In der CATI-Befragung wird die Annahme bei allen Items für die Residuen der Interviewerebene verletzt. Im Falle der Residuen der Sampling-Point-Ebene können vier von sieben Items dieser Annahme genügen. Eine Ausreißerkontrolle verzeichnet beim Item f35a (Interviewerebene) und f42 (Sampling-Point-Ebene) eine relativ hohe Anzahl von Ausreißern.
Bei ordinalskalierten Variablen wird bei der Prüfung der Normalverteilung von Residuen der zweiten und dritten Ebene nur in Ausnahmefällen eine Verletzung festgestellt. In Bezug auf die Ausreißer zeigen sich die Ergebnisse für Face-to-Face-Befragung weitgehend als stabil. Die Ergebnisse der CATI-Befragung dagegen weisen für alle drei Items einige Ausreißer auf der Interviewerebene auf. Da sich die Schlußfolgerungen über die getesteten Interviewereffekte hauptsächlich auf die Ergebnisse der Face-to-Face-Befragung stützen, können diese Ausreißer als unbedeutend eingestuft werden.

Hypothese 2: Männliche Interviewer führen bei Viktimisierungsbefragungen zu größeren Interviewereffekten als weibliche Interviewer. Diese Effekte sind bei „anfälligen“ Items größer als bei „weniger anfälligen“

Für die Prüfung dieser Hypothese wird auf die STATA-Befehle -loneway- und -anova- zurückgegriffen. Der Einfluss des Interviewergeschlechts kann nur für persönliche Interviews untersucht werden, da in der CATI-Befragung nur Frauen als Interviewer eingesetzt wurden und somit die Varianz auf der unabhängigen

[8]Dafür werden die Boxplots der standardisierten Residuen entsprechender Ebenen untersucht.

Variable fehlte. Die berechneten Koeffizienten sind daher frei von den zusätzlichen Moduseffekten.
Der Befehl -loneway- mit dem Interviewergeschlecht als gruppierender Variable liefert dabei die Intraklassenkorrelationskoeffizienten für das jeweilige Item und kann als erster Hinweis auf den vorhandenen Einfluss des Interviewergeschlechts gelten. Bei neun von 16 Items wird ein signifikanter Einfluss des Interviewergeschlechts festgestellt. Die signifikanten Werte variieren zwischen 0.002 und 0.035. Im Vergleich zum allgemeinen Interviewereinfluss ($H1$) scheint das Interviewergeschlecht eine kleinere Rolle zu spielen, da der Anteil der Varianz dieses Faktors an der totalen Varianz ist relativ gering. Das gleiche gilt für r^2, das im Vergleich zu den Ergebnissen der ersten Hypothese deutlich kleinere Werte aufweist[9]. Eine der Erklärungen dafür wäre die Verletzung der Annahmen der Varianzanalyse[10].
Eine zweifaktorielle ANOVA mit partiellen Quadratsummen und Einflussvariablen „Geschlecht des Befragten“, „Geschlecht des Interviewers“ und „Interaktionseffekt der Geschlechter“ dient ebenfalls als Hinweis auf die Effekte des Interviewergeschlechts sowie auf die Interaktionseffekte der Geschlechter. Diese verzeichnet für zehn Items einen signifikanten Interviewereinfluss. Außerdem sind die adjustierten r^2-Werte der zweifaktoriellen ANOVA für viele Items höher als die r^2-Werte der einfaktoriellen Analyse (liegen zwischen 0.003 und 0.183).
Die Annahmen der ANOVA sind, wie bei der ersten Hypothese, verletzt. Entsprechend den Ergebnissen des Bartlett-Tests wird die Hypothese der gleichen Varianzen in den Gruppen nur für sechs Items bestätigt. Dennoch erlaubt die Robustheit der ANOVA gegenüber der Verletzung von Varianzhomogenitätsannahme eine sinnvolle Interpretation der Ergebnisse.

Die eigentliche Prüfung der Hypothese erfordert einen direkten Vergleich der p_{int}-Werte der von männlichen mit den von weiblichen Interviewern befragten Personen. Tabelle 6.2 faßt die Ergebnisse des Hypothesentests zusammen.
Im Anschluß an die Berechnung der Koeffizienten wird ein einseitiger T-test für unabhängige Stichproben durchgeführt. Die Annahmen des T-Tests sind weitgehend erfüllt: die Beobachtungen stammen aus einer Zufallsstichprobe, die Koeffi-

[9] r^2 liegt bei signifikanten Ergebnissen zwischen 0.001 und 0.018.
[10] keine kontinuierliche abhängige Variable, Varianzheterogenität in den Gruppen. Dazu ausführlich Kapitel 5.

Tabelle 6.2: Einfluss des Interviewergeschlechts

Item	p_{int} männl. Interviewer	p_{int} weibl. Interv.
f10	0.277	0.140
f13_2	0.218	0.147
f13_4	0.246	0.170
f14_2	0.220	0.214
f14_3	0.233	0.163
f25	0.177	0.137
f33a	0.000* (n.s.)	0.000* (n.s.)
f34	0.135	0.075
f35	0.207	0.129
f35a	0.211	0.008* (n.s.)
f42	0.071	0.084
lf44	0.023	0.056
$\overline{p_{int}}$ Typ1-Items	0.168	0.110
f1	0.246	0.150
f31	0.212	0.179
f38_4	0.184	0.108
f43	0.032	0.023 (n.s.)
$\overline{p_{int}}$ Typ2-Items	0.169	0.115
$\overline{p_{int}}$ gesamt	0.168	0.111

zientenwerte sind annähernd normalverteilt[11], die Variablen sind intervallskaliert und die Varianzen in den Gruppen sind gleich[12]. Die Ergebnisse des T-Tests bestätigen die aufgestellte Hypothese, dass männliche Interviewer zu stärkeren Effekten führen als weibliche.
Ein Unterschied in der Ausprägung der Effekte für verschiedene Itemtypen kann hingegen nicht festgestellt werden.

[11] geprüft mit einem Skewness-Kurtosis-Test
[12] geprüft mit einem -sdtest-.

Hypothese 3: In gleichgeschlechtlichen Dyaden sind Interviewereffekte geringer als in verschiedengeschlechtlichen Dyaden

Diese Hypothese wird mit der STATA-Routine -loneway- auf zweifache Weise getestet.

Im ersten Fall wird der jeweilige Interviewer als gruppierende Variable definiert und die Intraklasskorrelationskoeffizienten in gleichgeschlechtlichen Dyaden[13] (Mann-Mann und Frau-Frau) mit den in verschiedengeschlechtlichen Dyaden (Mann-Frau und Frau-Mann) verglichen.

Dabei werden folgende Werte ausgerechnet: p_{int} für signifikante Items liegt bei 0.075-0.421 mit einem Durchschnitt über 16 getestete Items von 0.193 (GG-Dyaden) und 0.227 (VG-Dyaden). In acht von 16 Fällen sind die Unterschiede in den Koeffizienten der beiden Gruppen in die postulierte Richtung signifikant: der p_{int} in verschiedengeschlechtlichen Dyaden ist höher als in gleichgeschlechtlichen[14]. Da die Annahmen für die Durchführung eines T-Tests erfüllt sind, wird ein einseitiger T-Test durchgeführt. Dieser führt jedoch nicht zur vorläufigen Akzeptierung der Forschungshypothese[15].

Im zweiten Fall wird der Einfluss des Interviewersgeschlechts (Interviewergeschlecht als gruppenbildende Variable) für gleichgeschlechtliche vs. verschiedengeschlechtliche Dyaden als p_{int} geschätzt und dessen Werte miteinander verglichen. Auch hier erfolgt eine Kontrolle für den Befragungsmodus, da die Koeffizienten nur für persönliche Interviews berechnet werden.

Fünf der Typ1-Fragen werden nur an Frauen gestellt, somit fehlt für diese Items die Varianz auf der Einflussvariable. Die Gruppierung sowie die Berechnung der Koeffizienten ist für diese Fälle nicht möglich. Dadurch reduziert sich die Anzahl der zu vergleichenden Koeffizienten auf elf Paare. Die p_{int}-Werte für signifikante Items liegen zwischen 0.008-0.648 mit einem Durchschnitt für elf getestete Items von 0.138 (GG-Dyaden) und 0.145 (VG-Dyaden)[16].

Insgesamt liefert dieser Vergleich weniger verlässliche Ergebnisse. Obwohl der Vergleich der Durchschnittswerte der Koeffizienten immer in Richtung der For-

[13] Gleichgeschlechtliche Dyaden werden als GG-Dyaden, verschiedengeschlechtliche als VG-Dyaden abgekürzt.

[14] Die entsprechende Tabelle A.3 findet sich im Anhang A.

[15] Siehe dazu den Datensatz ttestsintervdefect.dta. Die Intraklassenkorrelationskoeffizienten für gleichgeschlechtliche Dyaden sind in der Variable GG1, für verschiedengeschlechtliche Dyaden in der Variable VG1 gespeichert.

[16] Siehe Tabelle A.4 im Anhang A.

schungshypothese ausfällt, wird beim jeweiligen Paarvergleich deutlich, dass eine definitive Ablehnung der $H0$ auf Grund dieser Daten nicht erfolgen kann. Nur bei drei Vergleichspaaren ist der p_{int}-Wert in den verschiedengeschlechtlichen Dyaden höher. Bei acht weiteren Vergleichspaaren ist der Koeffizient dagegen in den gleichgeschlechtlichen Dyaden höher.
Ein anschließender T-Test der durchschnittlichen Koeffizienten über alle elf Items ergibt keinen signifikanten Unterschied zwischen den Gruppen[17].

Hypothese 4: Männliche Interviewer haben bei Viktimisierungsbefragungen einen größeren Einfluss auf weibliche Befragte als weibliche Interviewer

Diese Hypothese geht von einem Interaktionseffekt zwischen dem Geschlecht des Interviewers und des Befragten aus. Einen ersten Eindruck über diesen Effekt kann eine zweifaktorielle ANOVA (erster Faktor „Geschlecht des Befragten“, zweiter Faktor „Geschlecht des Interviewers“) mit einem Interaktionseffekt geben. Diese ergibt in zehn von 104 Fällen einen signifikanten Interaktionseffekt[18]. Somit beträgt der Anteil von signifikanten Effekten an allen möglichen 10/104=0.10.
Nachfolgend ist ein direkter Vergleich der Intraklassenkorrelationskoeffizienten erforderlich, um die Richtung des Interaktionseffekts feststellen zu können. Dafür wird eine einfaktorielle ANOVA mit Koeffizienten für folgende zwei Gruppen berechnet: a) von Männern befragte Frauen und b) von Frauen befragte Frauen[19]. Der p_{int}-Mittelwert für die erste Gruppe beträgt 0.192, für die zweite Gruppe 0.082. Ein einseitiger T-Test für unabhängige Stichproben zeigt einen signifikanten Unterschied zwischen den Gruppen in die postulierte Richtung. Auch der direkte Vergleich der Koeffizienten für die beantworteten Items in beiden Gruppen zeigt für alle Items deutlich höhere Werte für die Gruppe der von Männern befragten Frauen. Die Nullhypothese wird daher abgelehnt.

[17] Die Intraklassenkorrelationskoeffizienten finden sich im Datensatz ttestsintervdefect.dta. Für gleichgeschlechtliche Dyaden sind diese in der Variable GG2, für verschiedengeschlechtliche Dyaden in der Variable VG2 gespeichert.
[18] Dabei werden die Fragen, die nur an Frauen gestellt wurden, nicht berücksichtigt.
[19] Die entsprechenden Koeffizienten finden sich in der Tabelle im Anhang A.

Hypothese 5: Weibliche Interviewer haben bei Viktimisierungsbefragungen einen größeren Einfluss auf männliche Befragte als männliche Interviewer

Das Testen dieser Hypothese erfolgt auf die gleiche Weise wie das der Hypothese vier. Diese Aussage stellt, wie auch die $H4$ eine Verfeinerung der dritten Hypothese über die stärkeren Effekte in verschiedengeschlechtlichen Dyaden dar. Obwohl diese Aussage für männliche Interviewer in der $H4$ zu trifft, können die Ergebnisse des Hypothesentests für die $H5$ nicht zur Annahme der Forschungshypothese führen. Die Intraklassenkorrelationskoeffizienten sind beim Paarvergleich in der Gruppe der von Männern befragten Männer höher als in der Gruppe der von Frauen befragten Männern, was der Hypothese widerspricht. Der nachfolgende einseitige T-Test bestätigt die Signifikanz der Koeffizientenunterschiede in die der Hypothese entgegengesetzte Richtung. Die H_0 wird beibehalten.

Hypothese 6: Der Interviewereffekt wird mit steigender Altersdifferenz zwischen dem Interviewer und dem Befragten größer

Da die CATI-Befragung keine Angaben zum Alter des Interviewers enthält, werden die Hypothesen über den Einfluss des Interviewералters ($H6$ und $H7$) nur für Face-to-Face-Befragung getestet.
Als erstes wird der allgemeine Einfluss des Interviewералters untersucht. Die einfaktorielle ANOVA liefert für zwölf von 16 Items signifikante p_{int}-Werte zwischen 0.012 und 0.061 mit sehr geringen r^2-Werten (zwischen 0.03 und 0.08). Dies bestätigt, dass sich das Alter des Interviewers auf das Befragtenverhalten verzerrend auswirkt. Da dieser Test nur auf die Unterschiede hinweist, deren Richtung jedoch nicht bestimmen kann, soll ein passenderes Modell gefunden werden. Unter Berücksichtigung dessen wird im zweiten Schritt der Einfluss der absoluten und nicht absoluten Altersdifferenzen[20] untersucht. Beide Effekte sind jeweils in fünf Fällen signifikant. Für nicht absolute Altersdifferenzen können nur sehr geringe Werte von p_{int} (0.013-0.046) und r^2 festgestellt werden. Die p_{int}-Werte bei absoluten Vergleichen fallen dagegen deutlich höher aus. So liegt der Intraklassenkorrelationskoeffizient zwischen 0.023 und 0.149 mit einem Durchschnitt für signifikante Ergebnisse von 0.067. In beiden Fällen gehören vier der fünf si-

[20] Die Variable „Absolute Altersdifferenz“ wird wie folgt berechnet: Absoluter Betrag von „Alter des Interviewers“ minus „Alter des Befragten“; die Variable „nicht absolute Altersdifferenz“ als „Alter des Interviewers“ minus „Alter des Befragten“.

gnifikanten Items dem Typ1 („anfällig“ für Interviewereffekte) an.

Die eigentliche Hypothesenprüfung zielt auf die Bestimmung der Richtung der Unterschiede in Koeffizienten in verschiedenen Gruppen. Dies erfordert einen direkten Vergleich der Koeffizienten in folgenden Altersdifferenzgruppen: 1) 0-10 Jahre Differenz, 2) 11-20 Jahre, 3) 21-30 Jahre, 4) 31-40 Jahre und 5) 41 und mehr Jahre. Die Annahme hierbei ist, dass je größer die Altersdifferenz zwischen dem Interviewer und dem Befragten, desto stärker ist die verzerrende Wirkung des Interviewers, die auch in diesem Fall als Intraklassenkorrelationskoeffizient definiert wird.
Die Koeffizienten werden für jede Gruppe und jedes der 16 Items ausgerechnet und ein Mittelwert des Koeffizienten für jede Gruppe gebildet. Aus der Berechnung des Gruppenmittelwertes werden fünf Items ausgeschlossen, die in mehr als zwei Gruppen nicht signifikante Ergebnisse aufweisen[21]. Folgende Mittelwerte werden für die entsprechenden Altersdifferenzgruppen ausgerechnet: 1) 0.174,
2) 0.248, 3) 0.245, 4) 0.259 und 5) 0.272[22].
Im Einklang mit der Hypothese finden sich die geringsten Werte in der Gruppe mit der kleinsten Altersdifferenz und die größten in der Gruppe mit der größten Differenz. Außerdem steigen die Koeffizientenwerte, wie ebenfalls angenommen, mit der steigenden Altersdifferenz. Zwar sind die Unterschiede zwischen den Gruppen nicht allzu stark ausgeprägt, die Forschungshypothese kann jedoch vorläufig akzeptiert werden.

Hypothese 7: Die Interviewereffekte sind in der Gruppe der negativen Altersdifferenzen größer als in der Gruppe der positiven Altersdifferenzen

In diesem Zusammenhang bedeuten positive Altersdifferenzen, dass in der Gruppe die Interviewer älter als die Befragten sind. Außerdem enthält diese Gruppe alle Paare mit Nulldifferenzen. Die Gruppe der negativen Altersdifferenzen besteht dagegen aus den Paaren, bei denen die Interviewer jünger als die Befragten sind.
Die einfaktorielle ANOVA mit der Gruppierungsvariable „Gruppierte nicht-absolute Altersdifferenz Interviewer-Befragte“ verzeichnet für zehn Items signifikan-

[21] Es geht dabei um folgende Items: f33a, f38_4, f35a, f43, lf44.
[22] Siehe Tabelle A.6 im Anhang A.

te Intraklassenkorrelationskoeffizienten zwischen 0.007 und 0.176. Dies bedeutet, dass die Zugehörigkeit zu einer bestimmten Gruppe die Werte der abhängigen Variable beeinträchtigt.
Zum selben Ergebnis führen auch die Berechnungen mit -gllamm-. Die Zugehörigkeit zur Gruppe der positiven Altersdifferenzen resultiert bei allen Items in höheren Werten der abhängigen Variable. Zudem ergibt eine Analyse der Intraklassenkoeffizienten, die für jede Ebene berechnet werden, einen deutlich höheren Anteil der durch die Interviewerebene erklärten Varianz im Vergleich zum durch die Sampling-Point-Ebene erklärten Varianzanteil[23].
Zudem können für beide Gruppen die jeweiligen p_{int}-Werte bestimmt werden. Diese Werte werden mit dem STATA-Befehl -loneway- ausgerechnet, wobei der jeweilige Interviewer als gruppierende Variable fungiert und eine Sortierung der Beobachtungen nach den Altersdifferenzgruppen erfolgt.
Auf der Basis der in der Tabelle 6.3 aufgeführten Ergebnisse kann man von einem starken Interviewereinfluss in beiden Gruppen ausgehen. Im Einklang mit der aufgestellten Hypothese ist der Effekt in der Gruppe der negativen Altersdifferenzen höher, jedoch nur für die Typ2-Items. In der Gruppe der Typ1-Items ist die entgegengesetzte Tendenz zu beobachten. Obwohl der Durchschnitt des Intraklassenkorrelationskoeffizienten für alle 16 Items in der Gruppe der positiven Differenzen größer ist, ist dieser Unterschied laut dem durchgeführten T-Test nicht signifikant. Die Nullhypothese kann nicht abgelehnt werden.

[23]Die mit -gllamm- berechneten Varianzen und daraus errechneten Intraklassenkorrelationskoeffizienten finden sich im Datensatz ttestsdefectintervgllamm.dta, entsprechende t-tests der Koeffizienten im Do-File ttestsdefectintervgllamm.do.

Tabelle 6.3: Interviewereffekte für negative vs. positive Altersdifferenzgruppen

Typ 1-Items	p_{int} negative D.	p_{int} positive D.
f10	0.224	0.254
f13_2	0.163	0.266
f13_4	0.195	0.248
f14_2	0.258	0.290
f14_3	0.178	0.279
f25	0.152	0.209
f33a	0.065 (n.s.)	0.000* (n.s.)
f34	0.161	0.166
f35	0.219	0.203
f35a	0.000* (n.s.)	0.180
f42	0.098	0.067
lf44	0.050	0.000* (n.s.)
$\overline{p_{int}}$	0.147	0.180
Typ 2-Items		
f1	0.238	0.207
f31	0.211	0.219
f38_4	0.235	0.140
f43	0.037	0.021 (n.s.)
$\overline{p_{int}}$	0.180	0.147
Total $\overline{p_{int}}$	0.155	0.172

Fazit

Die Analysen bestätigen, dass der Interviewer einen wichtigen Einflussfaktor im Befragungsprozess darstellt ($H1$).
Zudem wird festgestellt, dass seine verzerrende Wirkung geschlechtsspezifisch ist - sie fällt bei männlichen Interviewern stärker aus als bei weiblichen ($H2$). Die Ausprägung der Effekte ist bei beiden Hypothesen ($H1$ und $H2$) unabhängig vom definierten Itemtyp (Typ1- oder Typ2-Items).
Weiterhin können keine stärkeren Interviewereffekte in verschiedengeschlechtlichen im Vergleich zu gleichgeschlechtlichen Dyaden bestätigt werden ($H3$). Bei dieser Hypothese bestehen verschiedengeschlechtliche Dyaden aus folgenden Paaren: 1) männlicher Interviewer - weiblicher Befragte und 2) weiblicher Interview-

er - männlicher Befragte. Eine Ausdifferenzierung der Hypothese für diese Paare ergibt, dass die erste verschiedengeschlechtliche Dyade im Vergleich zur gleichgeschlechtlichen „weiblicher Interviewer - weiblicher Befragte“ stärkere Interviewereffekte verzeichnet ($H4$). Bei der zweiten verschiedengeschlechtlichen Dyade gegenüber der gleichgeschlechtlichen „männlicher Interviewer - männlicher Befragte“ ist dies nicht der Fall ($H5$).
Bei der Untersuchung der Alterseffekte ergibt sich, dass der Effekt mit steigender Altersdifferenz zwischen dem Interviewer und dem Befragten stärker wird ($H6$). Zuletzt vermutete ich, dass die Altersdifferenz keine absolute Bedeutung hat. Im Falle, dass der Interviewer jünger als der Befragte ist, sollte die Altersdifferenz größere Bedeutung haben als wenn der Interviewer älter als der Befragte ist. Diese Hypothese ($H7$) bleibt unbestätigt.

6.2 Testen der Interviewereffekte anhand des Frauendatensatzes

Im folgenden Kapitel werden die Interviewereffekte anhand der Frauenstudie getestet.
Im ersten Abschnitt des Kapitels werden Besonderheiten des Studiendesigns erläutert, die zur Reduzierung der Anzahl der Forschungshypothesen geführt hatten. Danach werden die Hypothesen formuliert und getestet. Ein anschließendes Fazit bietet einen Überblick über die Ergebnisse.

6.2.1 Modelle

Wie bereits angesprochen, basiert der Datensatz der Frauenstudie auf einem anderen Design als der DEFECT-Datensatz.
In diesem Zusammenhang ist als erstes zu erwähnen, dass bei der Frauenstudie kein interpenetriertes Design verwendet wurde, so dass eine Trennung der Interviewer- von den Sampling-Point-Effekten nicht möglich ist. Deswegen ist bei den berechneten Interviewereffekten zu beachten, dass sie einen gewissen (nicht eindeutig zu bestimmenden) Anteil der Sampling-Point-Effekte beinhalten (Diehl (1977)).
Des Weiteren erforderte die Spezifik der Untersuchung laut den Studienverantwortlichen den Einsatz ausschließlich weiblicher Interviewer. Diese Entscheidung stützte sich auf die Annahme, „dass sich weibliche Zielpersonen erfahrungsgemäß gegenüber Frauen als Interviewerinnen bei sehr persönlichen Themen - insbesondere bei Fragen zu sexueller Gewalt - leichter und vertrauensvoller öffnen können“ (Fredebeul *et al.* (2004): 8).
Da außerdem nur Frauen befragt wurden, können Hypothesen zum Einfluss des Interviewergeschlechts und zu den Interaktionseffekten der Geschlechter ($H2$, $H3$, $H4$ und $H5$ im Kapitel 6.1) nicht behandelt werden.

Es werden somit folgende Hypothesen zu den Interviewereffekten untersucht:

- $H1$: Interviewer haben einen verzerrenden Einfluss auf die Befragungsergebnisse der Viktimisierungssurveys. Dieser Einfluss ist bei sensitiven Fragen zur Viktimisierung sowie auf emotional geladene, schwierige, offene oder

Einstellungsfragen größer[24].

- $H2$: Der Interviewereffekt bei Viktimisierungssurveys wird mit steigender Altersdifferenz zwischen dem Interviewer und Befragten größer. Dieser weist für „anfällige“ Items größere Werte auf[25].

- $H3$: Interviewereffekte sind in der Gruppe der negativen Altersdifferenzen höher als in der Gruppe der positiven Altersdifferenzen[26].

Entsprechend der im Kapitel 4 beschriebenen Datenaufbereitung ergibt sich eine weitere Besonderheit für das Testen der Hypothesen anhand der Frauenstudie. Die Intraklassenkorrelationskoeffizienten weisen in Abhängigkeit von der jeweiligen Kodierung der Antworten erhebliche Unterschiede auf.
Dabei geht es vor allem um den Einfluss von „weiß nicht“-, „verweigert“- und „keine Angabe“-Kategorien. Der Intraklassenkorrelationskoeffizient variiert für dasselbe Item in ein und demselben Modell abhängig davon, ob die genannten Kategorien berücksichtigt werden oder nicht sowie abhängig von der Kodierung (als letzte Kategorie, z.B. mit einem Wert von 7, 8 oder 9 oder als erste Kategorie mit einem Wert von 0). Da die abhängige Variable in ANOVA als metrisch behandelt wird, hängt der Einfluss der Kategorien auf den Koeffizienten vom Kodierungswert ab. Diese Tatsache wird bei den jeweiligen Berechnungen berücksichtigt.
Nach einigen Überlegungen werden die ausgesuchten Items auf zweifache Weise umkodiert. Im ersten Fall werden alle gültigen Antwortkategorien beibehalten und die „weiß nicht“-, „verweigert“- und „keine Angabe“-Kategorien zu Missings umgewandelt. Im zweiten Fall erfolgt eine binäre Umkodierung der Items. Dabei bilden alle gültigen Antwortkategorien die erste und alle anderen die zweite Kategorie. Alle Hypothesen werden für beide Kodierungsarten geprüft. Auf die sich daraus ergebenden Unterschiede bzw. Gleichheit der Ergebnisse wird bei der Prüfung der jeweiligen Hypothesen eingegangen.
In Anlehnung an die Tests im DEFECT-Datensatz werden die ausgewählten Items auf dieselbe Weise kategorisiert. Eine ausführliche Beschreibung der Kategorisierung wurde bereits im Kapitel 5 vorgenommen. Die Items werden dement-

[24]Diese Hypothese entspricht der anhand des DEFECT-Datensatzes getesteten $H1$ (Interviewereffekte).
[25]Entspricht der anhand des DEFECT-Datensatzes getesteten $H6$ (Interviewereffekte).
[26]Entspricht der anhand des DEFECT-Datensatzes getesteten H_7 (Interviewereffekte).

sprechend in „Items mit erhöhter Wahrscheinlichkeit der Interviewereffekte“ und „Items mit geringerer Wahrscheinlichkeit der Interviewereffekte“ aufgeteilt (auch als Typ1- bzw. Typ2-Items bezeichnet).
Um eine bessere Vergleichbarkeit der Ergebnisse für verschiedene Itemtypen untereinander zu gewährleisten, wird - wenn möglich - die gleiche Anzahl an Items pro Typ ausgewählt.
Alle Hypothesen zu den Interviewereffekten werden nur für die Items der mündlichen Befragung untersucht, da ich nur für diese Befragungsform bedeutende Interviewereffekte vermute[27]. Zwar wird die schriftliche Befragung nicht als frei von diesen Effekten eingeschätzt, diese werden jedoch in ihrer Intensität als geringfügig angenommen[28].

6.2.2 Ergebnisse

Hypothese 1: Interviewer haben einen verzerrenden Einfluss auf die Befragungsergebnisse der Viktimisierungssurveys. Dieser Einfluss ist bei sensitiven Fragen zur Viktimisierung sowie auf emotional geladene, schwierige, offene oder Einstellungsfragen größer

Zuerst werden die Intraklassenkorrelationskoeffizienten für beide Itemtypen („anfällige“ vs. „weniger anfällige“) für nicht binär kodierte Items berechnet. Das Vorhandensein von Interviewereffekten wird für alle 35 Items außer für das Item nf732[29] „Zufriedenheit mit ärztlicher Hilfe“ bestätigt. Die Koeffizienten weisen signifikante Werte zwischen 0.010 und 0.083 mit einem Durchschnitt von 0.044 für Typ1-Items und zwischen 0.013 und 0.170 mit einem Durchschnitt von 0.045 für Typ2-Items auf[30]. Die Forschungshypothese über den verzerrenden Einfluss der Interviewer kann vorläufig akzeptiert werden. Die spezifische Ausprägung der Effekte für verschiedene Itemtypen lässt sich jedoch nicht bestätigen.
An dieser Stelle soll noch einmal darauf hingewiesen werden, dass der im je-

[27] Für die Entstehung der Interviewereffekte ist die Anwesenheit des Interviewers ausschlaggebend. Dies ist bei einer persönlichen Befragung gewährleistet. Bei einer klassischen schriftlichen Befragung fehlt der Interviewer, somit können keine Interviewereffekte auftreten.
[28] Beim Ausfüllen des schriftlichen Drop-Offs der Frauenstudie waren die Interviewer anwesend. Das Design der Studie erlaubte jedoch eine Reduzierung dieser Effekte.
[29] Im Originalfragebogen handelt es sich dabei um das Item f732. Die Bezeichnung nf732 ist durch die erwähnte Umkodierung der Itemkategorien entstanden. Siehe dazu den Do-File „cranalysefrauen.do“.
[30] Alle Koeffizienten sind in der Tabelle A.7 im Anhang A aufgeführt.

weiligen Intraklassenkorrelationskoeffizienten enthaltene Anteil der Varianz der Interviewer an der totalen Varianz einen zusätzlichen Anteil der Sampling-Point Varianz enthält.

Für binär kodierte Items ergibt sich ein etwas anderes Bild. Nach der binären Umkodierung der ausgewählten 35 Items haben 13 Items sehr schiefe Verteilungen (mehr als 90 % der Befragten wählten eine der beiden Kategorien). Diese Items werden aus der Analyse ausgeschlossen[31]. Von den verbleibenden 12 Items gehört je die Hälfte von ihnen einem der beiden Typen an.
Die berechneten Koeffizienten sind für alle zwölf Items signifikant und liegen zwischen 0.022 und 0.131 mit einem Durchschnitt von 0.113 für Typ1-Items und von 0.066 für Typ2-Items. Die Ergebnisse des Mann-Whitney-Tests weisen auf Signifikanz der Unterschiede zwischen den Itemtypen hin. Die Forschungshypothese kann für die binär kodierten Items in ihrem vollen Umfang vorläufig akzeptiert werden.
Abschließend kann man die Hypothese über den Einfluss der Interviewer in Viktimisierungsurveys als bestätigt betrachten, da die Werte des Intraklassenkorrelationskoeffizienten überzufällig größer als Null sind. Ein Unterschied zwischen „anfälligen“ und „weniger anfälligen“ Items kann jedoch nur für dichotome Items bestätigt werden.
Bezüglich der p_{int}-Werte für die gleichen Items, die unterschiedlich kodiert werden, zeigen sich deutlich höhere Werte für die binär kodierten Items. Daraus folgt, dass je stärker die Annahme der normalverteilten abhängigen Variable verletzt wird, desto empfindlicher sind die Ergebnisse der Varianzanalyse. Für binär kodierte Items, bei denen diese Annahme am stärksten verletzt wird, erfolgt eine Überschätzung der Interviewereffekte.

[31] Es geht dabei um folgende Items: f110_3, f110_4, f111_3, f111_4, f203_1, f203_3, f300, f501_a1, f602, f605, f800, f802, f100, f101, f112_a, f112_c, f211, f213, f408, f600, f732, f908, f914. Siehe dazu das Do-File „cranalysefrauen.do“.

Hypothese 2: Der Interviewereffekt bei Viktimisierungssurveys wird mit steigender Altersdifferenz zwischen dem Interviewer und dem Befragten größer. Dieser weist für „anfällige“ Items größere Werte auf

Bevor der Einfluss der Altersdifferenz berechnet wird, soll geprüft werden, ob es in diesem Zusammenhang überhaupt einen Effekt des Alters der Interviewer auf die Befragungsergebnisse gibt. Dies geschieht mit einer einfaktoriellen ANOVA, mit der für die ausgesuchten Items Intraklassenkorrelationskoeffizienten berechnet werden.
Für nicht binär kodierte Items verzeichnet der Koeffizient für Typ1-Items in jeweils 16 von 18 Fällen signifikante Werte zwischen 0.004 und 0.040 mit einem Durchschnitt für signifikante Koeffizienten von 0.014. 15 von 16 Koeffizienten der Typ2-Items sind signifikant mit den Werten zwischen 0.004 und 0.022 und einem Durchschnitt von 0.012. Wie auch im Falle der Prüfung dieser Hypothese für den DEFECT-Datensatz sind die Werte von r^2 sehr gering bis gering (0.007-0.088), was unter anderem auf die Verletzungen der Normalverteilungsannahme der abhängigen Variable zurückgeführt werden kann.
Ein Unterschied der Koeffizientenwerte abhängig vom Itemtyp kann nicht festgestellt werden.

Dichotom kodierte Items haben auch bei diesem Test höhere Werte als nicht dichotome Items. Außerdem sind die Werte der „anfälligen“ Items wie angenommen höher als die der „weniger anfälligen“. Die jeweiligen Durchschnitte betragen 0.032 (Typ1) und 0.018 (Typ2). Die Ergebnisse des Mann-Whitney-Test bestätigen die Signifikanz der Unterschiede.
Da die Intraklassenkorrelationskoeffizienten signifikante Werte für die überwiegende Anzahl der Items bei den beiden Tests aufweisen, gilt der allgemeine Alterseffekt als bestätigt. Des Weiteren gilt, dass dieser Effekt für die dichotomen Items eine stärkere Bedeutung hat als für nicht dichotome.

Um einen Interaktionseffekt des Alters, wie in der $H2$ formuliert, näher untersuchen zu können, wird eine neue Variable „Altersdifferenz zwischen dem Interviewer und Befragten“ generiert. Diese Variable kann zum Einen in ihrer jeweiligen absoluten oder nicht absoluten Ausprägung sowie für beide Fälle gruppiert oder ungruppiert betrachtet werden. Im Falle der absoluten Altersdifferenz spielt es

keine Rolle, ob der Interviewer oder der Befragte älter ist. So finden sich in derselben Altersdifferenzgruppe - z.B. „fünf Jahre Altersdifferenz“ - Dyaden mit dem um fünf Jahre älteren Interviewer sowie dem um fünf Jahre älteren Befragten. Im Falle der nicht absoluten Altersdifferenzen bilden diese Dyaden unterschiedliche Gruppen. Die entsprechenden Differenzen lassen sich außerdem gruppieren, wodurch aus einem Faktor mit sehr vielen Ausprägungen (alle möglichen Differenzen) ein Faktor mit wenigen Ausprägungen ensteht.
Daher wird im nächsten Schritt der Einfluss der absoluten ungruppierten Altersdifferenz, danach der Einfluss der nicht absoluten ungruppierten und anschließend der Einfluss der absoluten gruppierten Altersdifferenz untersucht. Der Effekt der nicht absoluten gruppierten Altersdifferenz wird gesondert in $H3$ getestet.

Einfluss der absoluten ungruppierten Altersdifferenz

Diese Variable[32] hat in elf von 18 (Typ1-Items) bzw. acht von 17 (Typ2-Items) Fällen bei nicht binär kodierten Items signifikante Unterschiede der Koeffizienten als Ergebnis. Die zum Teil sehr niedrigen Werte der Koeffizienten können mit der besonderen Gruppierung der Beobachtungen zusammenhängen. Möglicherweise heben sich die Effekte der unterschiedlich gerichteten Altersdifferenzen, die durch absolute Werte in einer Gruppe zusammengefasst werden, gegenseitig auf. Deswegen soll auch der Effekt der nicht absoluten Altersdifferenz untersucht werden.
Es ist kein Unterschied zwischen den Effekten bei verschiedenen Itemtypen zu beobachten. Die Koeffizienten in den beiden Gruppen sind weitgehend identisch.

Bei binär kodierten Items ist der Effekt für fünf (Typ1) bzw. sechs (Typ2) von jeweils sechs Items signifikant. Obwohl die entsprechenden Werte der Koeffizienten gering sind, weisen diese auf einen vorhandenen Effekt hin und bestätigen wie auch im Falle der nicht dichotomen Items die Forschungshypothese. In Bezug auf die Unterscheidung des Effekts für verschiedene Itemtypen kann die postulierte Differenz in beiden Fällen nicht als erwiesen gelten.

[32] Zur Generierung dieser Variable wird der absolute Betrag der Differenz zwischen den Variablen „Alter des Interviewers“ und „Alter des Befragten“ gebildet.

Einfluss der nicht absoluten ungruppierten Altersdifferenz

Diese Variable ist ein Resultat der Substrahierung des Befragtenalters vom Intervieweralter. Der Effekt ist für elf von 18 Items des ersten Typs und zehn von 17 Items des zweiten Typs überzufällig von Null verschieden mit einem höheren Durchschnitt für Typ2-Items (0.064 entgegen 0.017 für den ersten Itemtyp). Bei binär kodierten Items sind die Koeffizientenunterschiede in sechs (Typ1) bzw. fünf (Typ2) von jeweils sechs Fällen signifikant mit einem ebenfalls höheren Durchschnitt für Typ2-Items (0.093 gegen 0.027).
Die Signifikanz der Unterschiede zwischen den Itemtypen wird für die nicht dichotomen Items mit einem einseitigen T-Test für ungleiche Varianzen getestet. Diese Option wird gewählt, da der Test auf Gleichheit der Varianzen (-sdtest- in STATA) zur Ablehnung der H_0 führt. Der T-Test ist signifikant und deutet auf stärkere Effekte für Typ2-Items hin.
Die Ergebnisse bestätigen den vermuteten Effekt der Altersdifferenz zwischen dem Interviewer und Befragten. Eine Differenzierung der Ergebnisse nach Itemtypen führt zur Schlussfolgerung, dass diese im Gegensatz zur vermuteten Konstellation für beide Kodierungsarten für den zweiten Itemtyp stärker ausfallen.

Einfluss der absoluten gruppierten Altersdifferenz

Dieser Test ist die eigentliche Hypothesenprüfung, die die Richtung der Unterschiede in Koeffizienten in verschiedenen Altersdifferenzgruppen bestimmen soll. Wie auch für den DEFECT-Datensatz erfolgt hier ein direkter Vergleich der Koeffizienten in folgenden Altersdifferenzgruppen: 1) 0-10 Jahre, 2) 11-20 Jahre, 3) 21-30 Jahre, 4) 31-40 Jahre und 5) 41 und mehr Jahre Differenz. Mit der wachsenden Altersdifferenz zwischen dem Interviewer und dem Befragten soll die verzerrende Wirkung des Interviewers, die als Intraklassenkorrelationskoeffizient definiert wird, stärker ausfallen.
Die Koeffizienten werden für jede Gruppe und jedes Item ausgerechnet und ein Mittelwert des Koeffizienten für die Gruppen gebildet. Aus der Berechnung des Gruppenmittelwertes werden vier der Typ1-Items sowie drei der Typ2-Items ausgeschlossen, die pro Item in mehr als zwei Gruppen nicht signifikante Ergebnisse aufweisen[33]. Folgende Mittelwerte sind für die entsprechenden Altersdifferenz-

[33] Folgende Items werden ausgeschlossen: nf309, nf310, nf604_1, nf605 (Typ1) sowie nf207, nf732, nf920_1 (Typ2).

gruppen ausgerechnet: 1) 0.069, 2) 0.060, 3) 0.070, 4) 0.132 und 5) 0.138[34].
Im Einklang mit der Hypothese sind die geringsten Werte in der Gruppe mit der kleinsten Altersdifferenz und die größten in der Gruppe mit der größten Differenz zu beobachten. Außerdem sind die Koeffizientenwerte in den ersten drei Gruppen für beide Itemtypen geringer als für die letzten zwei Gruppen mit Altersdifferenzen von 31 Jahren und mehr. Obwohl die Steigerung der Unterschiede zwischen den Gruppen nicht kontinuierlich geschieht, so kann die Forschungshypothese jedoch vorläufig akzeptiert werden, da die Effekte für die Dyaden mit den größten Altersdifferenzen größer sind. Die zweite Aussage der Forschungshypothese über die stärkeren Effekte für die „anfälligen“ Items wird nicht bestätigt.
Die sehr hohen r^2-Werte (bis zu 0.97) erlauben es, das Modell als passend zu betrachten.

Eine weitere Untersuchung bezieht sich auf die binär kodierten Items und wird auf dieselbe Weise durchgeführt. Der Vergleich der Gruppenmittelwerte des Intraklassenkorrelationskoeffizienten bietet ähnliche Erkenntnisse wie für nicht dichotome Items: die Koeffizienten in den letzten zwei Gruppen mit den höchsten Altersdifferenzen sind höher als in den ersten beiden. Dies trifft für beide Itemtypen zu[35].
Der Effekt der absoluten gruppierten Altersdifferenz kann somit auch für binär kodierte Items als bestätigt gelten. Eine weitere Ausdifferenzierung des Effekts für unterschiedliche Itemtypen (der Effekt für Typ1-Items soll stärker als für Typ2-Items sein) kann nicht als erwiesen betrachtet werden.

Hypothese 3: Interviewereffekte sind in der Gruppe der negativen Altersdifferenzen höher als in der Gruppe der positiven Altersdifferenzen

Die Gruppe der positiven Altersdifferenzen besteht aus den Interviewer-Befragten-Dyaden, in denen Interviewer älter als Befragte oder beide gleichen Alters sind. Die Gruppe der negativen Altersdifferenzen besteht dagegen aus Paaren, bei denen Interviewer jünger als Befragte sind.
Eine einfaktorielle Varianzanalyse verzeichnet für 14 von 18 (Typ1) bzw. elf von 17 (Typ2) der nicht binär kodierten Items signifikante Intraklassenkorrelationskoeffizienten. Damit wird der Unterschied zwischen den Effekten in den zwei Al-

[34]Siehe dazu Tabelle A.8 im Anhang A.
[35]Siehe dazu Tabelle A.9 im Anhang A.

tersdifferenzgruppen bestätigt. Um die genauere Richtung der Unterschiede feststellen zu können, wird ein direkter Vergleich der für jedes Item (jeweils für die Gruppe der negativen und positiven Differenzen) berechneten Intraklassenkkorrelationskoeffizienten durchgeführt. Die entsprechenden Werte lassen sich mit dem STATA-Befehl -loneway- ausrechnen. Dabei fungiert der jeweilige Interviewer als Grundlage für die Gruppierung und die Beobachtungen werden nach Altersdifferenzgruppen sortiert.
Tabelle 6.4 liefert einen Überblick über die Ergebnisse für nicht binär kodierte Items.

Tabelle 6.4: p_{int} für nicht binär kodierte Items

Typ 1-Items			**Typ 2-Items**		
Item	p_{int} negat. D.	p_{int} posit. D.	Item	p_{int} negat. D.	p_{int} posit. D.
nf110_3	0.064	0.021	nf100	0.044	0.032
nf110_4	0.033	0.057	nf101	0.035	0.018
nf111_3	0.040	0.046	nf102	0.042	0.032
nf111_4	0.024	0.023	nf112_a	0.042	0.033
nf203_1	0.088	0.097	nf112_c	0.075	0.040
nf203_3	0.037	0.023	nf205	0.113	0.075
nf204	0.082	0.071	nf207	0.040	0.058
nf300	0.074	0.081	nf208	0.100	0.110
nf303	0.071	0.054	nf211	0.034	0.018
nf309	0.023	0.041	nf213	0.016	0.003 (n.s.)
nf310	0.000 (n.s.)	0.011 (n.s.)	nf408	0.075	0.032
nf501_a1	0.050	0.076	nf600	0.045	0.049
nf602	0.083	0.067	nf732	0.131	0.022 (n.s.)
nf604_1	0.028	0.042	nf908	0.093	0.094
nf605	0.055 (n.s.)	0.033 (n.s.)	nf914	0.038	0.119
nf800	0.047	0.037	nf920_1	0.008 (n.s.)	0.019
nf802	0.023	0.008 (n.s.)	nf928	0.236	0.150
nf913	0.046	0.077			
$\overline{p_{int}}$	0.051	0.058	$\overline{p_{int}}$	0.072	0.059

Insgesamt wird für jede getestete Gruppe eine hohe Anzahl signifikanter Koeffizienten berechnet. Die Koeffizienten für nicht binär kodierte Items liegen zwischen 0.016 und 0.236 mit einem Durchschnitt von 0.062 (negative Altersdifferenzen)

bzw. 0.018 und 0.150 mit einem Durchschnitt von 0.059 (positive Altersdifferenzen). Bei 17 Vergleichspaaren (p_{int}-Werte für beide Gruppen für jedes Item) sind die Koeffizienten in der Gruppe der negativen, bei zwölf Paaren in der Gruppe der positiven Altersdifferenzen höher und bei fünf Paaren können keine Unterschiede in den Koeffizienten festgestellt werden. Auch der Vergleich der durchschnittlichen Intraklassenkorrelationskoeffizienten liefert keine Grundlage zur Ablehnung der $H0$.
Im Falle der binär kodierten Items sind die Koeffizienten insgesamt größer (0.030-0.144 mit einem Durchschnitt von 0.109 für die erste und 0.041-0.153 mit einem Durchschnitt von 0.103 für die zweite Gruppe). Aber auch in diesem Fall kann die Nullhypothese nicht abgelehnt werden. Der Interviewereffekt fällt abhängig vom jeweiligen Item in der Gruppe der positiven oder negativen Altersdifferenzen höher aus.

Fazit

Insgesamt betrachtet, lässt sich die Mehrzahl der Hypothesen zu Interviewereffekten für diesen Datensatz bestätigen. So werden die verzerrende Wirkung des Interviewers ($H1$), der Einfluss der Altersdifferenz zwischen dem Interviewer und dem Befragten ($H2$) für beide Itemtypen und beide Kodierungsarten (dichotom vs. nicht dichotom) festgestellt. Die in der $H1$ sowie $H2$ angenommene stärkere Ausprägung dieser Effekte für Typ1-Items ist nur für binär kodierte Items bei der ersten Hypothese zu beobachten. Für nicht dichotome Items ($H1$) sowie für beide Kodierungsarten ($H2$) kann diese verstärkte Wirkung für Typ1-Items nicht bestätigt werden.
Die dritte Hypothese lässt sich anhand der Daten ebenfalls nicht bestätigen.

7 Formulierung und Prüfung der Hypothesen zu Moduseffekten

In dieser Arbeit werden die drei klassischen Modi betrachtet - die persönliche, telefonische und schriftliche Befragung. Die für die Enstehung der Moduseffekte relevanten Charakteristiken wurden im Kapitel 2 ausführlich beschrieben. Dieses Kapitel befasst sich mit der Formulierung und Prüfung entsprechender Hypothesen für jeden Datensatz.

7.1 Testen der Moduseffekte anhand des DEFECT-Datensatzes

7.1.1 Modelle

Beim Befragungsmodus sind drei Faktoren ausschlaggebend: a) Charakteristiken des Erhebungsmodus (Kontrollmöglichkeiten beim Befragten oder beim Interviewer, Zugang zum Befragungsinstrument, Zeitdruck), b) Einfluss der An- bzw. Abwesenheit des Interviewers und c) Informationsübertragungskanäle (sprachlich, visuell oder beides) und zusätzliche Kommunikationsmöglichkeiten (Gestik etc.) (De Leeuw (1992)).
Die vorliegende Arbeit testet in diesem Zusammenhang folgende Hypothesen anhand des DEFECT-Datensatzes:

- $H1$: Der Befragungsmodus hat einen verzerrenden Einfluss auf die Befragungsergebnisse der Viktimisierungssurveys.
- $H2$: In Face-to-Face-Interviews sind die Interviewereffekte größer als in telefonischen Befragungen.

- $H3$: Telefoninterviews weisen Recency-Effekte auf: der Anteil der gewählten letzten Antwortvorgaben ist in einem Telefoninterview höher als in einem Mail-Survey.

7.1.2 Ergebnisse

Hypothese 1: Der Befragungsmodus hat einen verzerrenden Einfluss auf die Befragungsergebnisse der Viktimisierungssurveys

Diese Hypothese untersucht den allgemeinen Einfluss des Befragungsmodus auf die Befragten. Bei der Beschreibung des Analysemodells in Kapitel 5 wurden die unterschiedlichen Möglichkeiten der Modellierung von Moduseffekten erläutert. Für die Prüfung dieser Hypothese tritt der Modus als gruppierende Variable auf[1]. Danach werden die Unterschiede zwischen den Antwortverteilungen in den so gebildeten Gruppen betrachtet. Der mit Hilfe einer einfaktoriellen ANOVA berechnete Intraklassenkorrelationskoeffizient bildet die Basis für die Inferenz. Die Moduseffekte gelten dann als vorhanden, wenn der Intraklassenkorrelationskoeffizient signifikante Werte aufweist[2].
Die p_{int}-Werte verzeichnen in 13 von 16 Fällen signifikante Ergebnisse zwischen 0.002-0.053 mit einem Durchschnitt von 0.020. Die Werte von r^2 sind ebenfalls in einem sehr niedrigen Bereich zwischen 0.001 und 0.033. Die Ergebnisse deuten auf einen relativ geringen Anteil der durch die Befragungsart erklärten Varianz der abhängigen Variable. Allerdings ist auch ein geringer Anteil ein Hinweis auf systematische Verzerrungen, die durch die abhängige Variable - hier den Befragungsmodus - entstanden sind.
Beim Berechnen der ANOVA für alle Items[3] ist der Koeffizient für 93 Items signifikant. Auf Grund der sehr hohen Anzahl von signifikanten Koeffizienten kann die Anwesenheit der Moduseffekte bestätigt werden.

[1] Dies ist die zweite Art der Modellierung von Moduseffekten im Kapitel 5.2.
[2] Hierbei handelt es sich um folgende Modi: Face-to-Face-Interview, telefonische und schfriftliche Befragung.
[3] Für diese Analyse bleiben 111 von insgesamt 135 Fragen, da sehr schiefe Items entsprechend den im Kapitel 4 beschriebenen Datenaufbereitungsprozess ausgeschlossen wurden.

Hypothese 2: In Face-to-Face-Interviews sind die Interviewereffekte größer als in telefonischen Befragungen

In einem persönlichen Interview hat der Interviewer viel mehr Möglichkeiten, die Befragten bewusst oder unbewusst zu beeinflussen. Seine sichtbaren Merkmale kommen bei einer Face-to-Face-Befragung besser zur Geltung, was eine weitere Ableitung der Einstellungen und Erwartungen des Interviewers (nicht sichtbarer Merkmale) durch den Befragten zur Folge haben und den Interviewereinfluss verstärken kann.

Diese Hypothese modelliert Moduseffekte als Unterschiede in den Interviewereinflüssen bei verschiedenen Modi, die durch bestimmte Modusmerkmale ausgelöst werden[4]. Es erfolgt demnach der Vergleich von Intraklassenkorrelationskoeffizienten für jedes Item getrennt nach dem Modus. Dafür werden die Beobachtungen nach Modus sortiert und der jeweilige Interviewer als Gruppenvariable definiert. Tabelle 7.1 ermöglicht den direkten Vergleich der Intraklassenkorrelationskoeffizienten.

Die Ergebnisse der Analyse sind eindeutig: für die Face-to-Face-Befragung werden für alle Items außer f33a („Innerhalb der letzten 12 Monate von einem Fremden geschlagen oder verletzt“) signifikante p_{int}-Werte ausgerechnet. Diese sind sowohl für Typ1-Items als auch für Typ2-Items sehr hoch. Der Durchschnitt des Koeffizienten für die Face-to-Face-Befragung beträgt 0.157. Der Intraklassenkorrelationskoeffizient weist hier auf modusspezifische Interviewereffekte hin, die im persönlichen Interview sehr stark ausgeprägt sind.

In der telefonischen Befragung dagegen sind die p_{int}-Werte lediglich für fünf von 16 Items signifikant mit einem Durchschnitt von 0.035.

In allen Vergleichspaaren, Face-to-Face- vs. CATI-Koeffizienten, weist der erstere höhere Werte auf. Für zehn Vergleichspaare ist der erste Koeffizient signifikant und der zweite nicht signifikant.

Da die Varianzen in den beiden Gruppen ungleich sind, wird ein einseitiger T-Tests für unabhängige Stichproben mit ungleichen Varianzen durchgeführt. Dieser erlaubt eine vorläufige Akzeptierung der Forschungshypothese.

Eine weitere Bestätigung der Hypothese liefern die Ergebnisse von -gllamm-[5].

[4] Die erste im Kapitel 5.2 beschriebene Modellierungsart der Moduseffekte.

[5] Siehe dazu Tabellen A.1 und A.2 im Anhang A.

Tabelle 7.1: p_{int} für Face-to-Face vs. CATI-Befragung

Typ 1-Items	p_{int} Face-to-Face	p_{int} CATI
f10	0.218	0.010
f13_2	0.187	0.040
f13_4	0.203	0.070
f14_2	0.217	0.027
f14_3	0.195	0.003 (n.s.)
f25	0.175	0.006 (n.s.)
f33a	0.000 (n.s.)	0.025 (n.s.)
f34	0.131	0.000 (n.s.)
f35	0.178	0.005 (n.s.)
f35a	0.175	0.000 (n.s.)
f42	0.074	0.028
lf44	0.035	0.000 (n.s.)
$\overline{p_{int}}$	0.163	0.035
Typ 2-Items		
f1	0.218	0.000 (n.s.)
f31	0.190	0.008 (n.s.)
f38_4	0.140	0.012 (n.s.)
f43	0.016	0.007 (n.s.)
$\overline{p_{int}}$	0.141	0.007 (n.s.)
$\overline{p_{int}}$ total	0.157	0.035

Ein T-Test der Intraklassenkorrelationskoeffizienten bestätigt signifikant höhere Interviewereffekte für die persönliche Befragung.
Außerdem werden die Varianzen bzw. Intraklassenkorrelationskoeffizienten beider Ebenen (Interviewer vs. Sampling-Point) für jeden Modus miteinander verglichen. In der Face-to-Face-Befragung ist die durch den Interviewer erklärte Varianz[6] in 15 von 16 Fällen deutlich größer als die durch den Sampling-Point erklärte Varianz.
In der CATI-Befragung gibt es keinen Hinweis auf die Dominanz der Interviewereffekte gegenüber den Sampling-Point-Effekten. In fünf Fällen sind die Interviewereffekte größer, in sechs Fällen dagegen die Sampling-Point-Effekte und in

[6]Diese kann man auch als Homogenität der Beobachtungen eines Interviewers bezeichnen.

weiteren fünf Fällen kann der Unterschied in den Koeffizienten vernachlässigt werden. Dieser Vergleich kann als eine zusätzliche Bestätigung der Forschungshypothese betrachtet werden.

Hypothese 3: Telefoninterviews weisen Recency-Effekte auf: der Anteil der gewählten letzten Antwortvorgaben ist in einem Telefoninterview höher als in einem Mail-Survey

Die Entstehung der Recency-Effekte wird durch die Besonderheiten der Stimuluspräsentation[7] bei verschiedenen Modi begründet. Demnach sind die Möglichkeiten bei telefonischen Befragungen im Vergleich zu persönlichen oder schriftlichen Interviews auf die Audiopräsentation beschränkt. Bei mehreren Antwortkategorien fällt es den Befragten schwer, sich diese zu merken. Dies betrifft vor allem die ersten Kategorien und kann zusammen mit dem zeitlichen Aspekt der Befragung (erhöhter Zeitdruck bei telefonischen Interviews) zur Überrepräsentation der letzten Antwortvorgaben führen.
Diese Hypothese stützt sich auf die Definition von Moduseffekten, die diese als Aspekte der Datenqualität beschreibt und untersucht[8]. Anstelle der Intraklassenkorrelationskoeffizienten werden für 13 ausgewählte Items, die mehr als vier Antwortkategorien beinhalten[9] getrennt nach Modi die Anteile der letzten Antwortkategorien an der Gesamtzahl der gültigen Antworten ausgerechnet. Die Analyse ergibt für zehn Paarvergleiche (CATI vs. Mail) höhere Werte für Anteile im CATI-Survey und in zwei Fällen höhere Werte für den Mail-Survey. Der Durchschnitt der Anteile über alle 13 Items beträgt in der CATI-Befragung 0.306 und in der Mail-Befragung 0.248 mit den dazugehörigen Standardabweichungen von 0.204 und 0.183.
Ein einseitiger T-Test für unabhängige Stichproben zeigt jedoch keinen signifikanten Unterschied zwischen den Anteilen im CATI- und Face-to-Face-Interview. Dies kann an der geringen Testpower liegen.
Da jedoch die jeweiligen paarweisen Vergleiche für die aufgestellte Hypothese sprechen, wird sie vorläufig akzeptiert.

[7] Siehe dazu Kapitel 2.1.3, Punkt 2.
[8] Siehe dazu Kapitel 5.2.
[9] Der Vorgang ist im Do-File „defectmodus.do“ zu finden.

Fazit

Die DEFECT-Daten erlauben eine Modellierung und Prüfung der Moduseffekte auf unterschiedliche Weise: mit Hilfe des Intraklassenkorrelationskoeffizienten oder als Vergleich der Datenqualität verschiedener Befragungen. Die Hypothesentests ermöglichen zuerst eine Feststellung der allgemeinen Moduseffekte ($H1$). Nachfolgend werden stärkere Effekte in persönlichen als in telefonischen Interviews bestätigt ($H2$). Für diese Hypothese erfolgt anschließend eine Trennung der Varianzen nach der Interviewer- und Sampling-Point-Varianz mit Hilfe des Stata-Tools -gllamm-. Als Letztes werden die Recency-Effekte untersucht und ihre Relevanz für telefonische Befragungen bestätigt. Somit können alle drei Forschungshypothesen vorläufig akzeptiert werden.

7.2 Testen der Moduseffekte anhand des Frauendatensatzes

Analog zum Testen der Interviewereffekte für die Frauenstudie werden hier zuerst die Besonderheiten des Studiendesigns beschrieben, die ihre Auswirkungen auf die Hypothesentests haben. Danach werden die Forschungshypothesen formuliert und untersucht und die Untersuchungsergebnisse in einem Fazit zusammengefasst.

7.2.1 Modelle

Wie bereits angesprochen, zeichnet sich das Design der Studie durch einige Besonderheiten aus. Für die Untersuchung der Moduseffekte ist vor allem der Ablauf beider Befragungen ausschlaggebend. Dieser soll hier kurz beschrieben werden.
Die Befragten nahmen zuerst an einem Face-to-Face-Interview teil. Danach wurde ihnen ein schriftlicher Selbstausfüller (Drop-Off) vorgelegt, dessen Ausfüllen im Schnitt 15-20 Minuten dauerte. 94 % aller Frauen, die am Face-to-Face-Interview teilgenommen hatten, füllten den Drop-Off ebenfalls aus. Der im Beisein der Interviewerin ausgefüllte Fragebogen wurde dieser in einem geschlossenen Umschlag übergeben. In Ausnahmefällen bestand die Möglichkeit, den Fragebogen im Freiumschlag per Post zurückzuschicken oder ihn von der Interviewerin zu einem späteren Zeitpunkt abholen zu lassen (Müller & Schröttle (2004)).
Das beschriebene Design führt zu einigen Einschränkungen bei der Formulierung sowie beim Testen der Hypothesen. Zusätzlich zum nicht interpenetrierten Design und zur fehlenden Varianz des Interviewer- und Befragtengeschlechts wirkt sich die Varianz des Fragebogens auf das Testen der Moduseffekte aus.
Im Gegensatz zum DEFECT-Fragebogen, der bei allen Befragungsmodi konstant bleibt, sind die Fragebögen der Frauenstudie verschieden. Ein sehr umfangreicher mündlicher Fragebogen enthält unter anderem Fragen zur allgemeinen Lebenssituation, Gesundheit und Sicherheit der Frauen sowie Fragen, die auf die Erfassung unterschiedlicher Gewalterfahrungen der Frauen ausgerichtet sind.
Der viel kürzere schriftliche Fragebogen konzentriert sich auf die Gewalterfahrungen in der aktuellen und den früheren Partnerschaften. Obwohl mehrere Items identisch oder fast identisch formuliert sind und gleiche (oder fast gleiche) Antwortkategorien beinhalten, darf die Varianz des Fragebogens beim Testen der

Moduseffekte nicht vernachlässigt werden.
Durch die erwähnte Varianz entstehen zusätzliche Effekte, für die nicht kontrolliert werden kann: Fragenreihenfolgeeffekte (question order effects), Kontexteffekte, Effekte durch unterschiedliche Formulierung der Items etc..
Bei den Hypothesentests werden dementsprechend zwei Stategien angewendet, die die Untersuchung der Effekte im beschriebenen eingeschränkten Umfang ermöglichen: zum einen wird versucht, identische oder fast identische Items in beiden Fragebögen zu finden und zu untersuchen, zum anderen werden die zu vergleichenden Items als Zufallsauswahl aller möglichen Items betrachtet und bestimmte Anteile (z. B. Anteile beantworteter offener Fragen) in beiden Modi miteinander verglichen.

In Bezug auf die Moduseffekte werden für die Frauenstudie folgende Hypothesen untersucht:

- $H1a$: Der Item-Nonresponse bei Viktimisierungsbefragungen ist in einem Face-to-Face Interview geringer als in einem Selbstausfüller (einer schriftlichen Befragung).
- $H1b$: Der Item-Nonresponse ist bei *sensitiven Fragen* (insbesondere zur Viktimisierung) in einem Face-to-Face Interview stärker ausgeprägt als in einem Selbstausfüller (einer schriftlichen Befragung).
- $H2$: Der Anteil der beantworteten offenen Fragen in Viktimisierungssurveys ist in einem Selbstausfüller (einer schriftlichen Befragung) geringer als in einer Face-to-Face-Befragung.

7.2.2 Ergebnisse

Hypothese 1a: Der Item Item-Nonresponse bei Viktimisierungsbefragungen ist in einem Face-to-Face Interview geringer als in einem Selbstausfüller (einer schriftlichen Befragung)

Wenn man den Item-Nonresponse generell für alle möglichen Fragen betrachtet, geht man davon aus, dass die Anwesenheit des Interviewers die Antwortbereitschaft erhöht und somit zur geringeren Verweigerungsquote führt (De Leeuw (1992)). Dieser Zusammenhang wird nachfolgend untersucht.

Ein Vergleich beider Fragebögen hat ergeben, dass einige Itempaare gebildet werden können, die aus identisch oder fast identisch formulierten Items bestehen. So lautet z.B. die Frage f45 im schriftlichen Selbstausfüller „Haben Sie oder andere in solchen Situationen mit Ihrem Partner jemals die Polizei eingeschaltet?“ Diese enthält folgende Antwortmöglichkeiten: „1) Ja, die Polizei wurde von mir *selbst* eingeschaltet; 2) Ja, die Polizei wurde *von anderen* eingeschaltet und 3) Nein, die Polizei wurde *nicht* eingeschaltet.“
Eine ähnliche Frage - f744 - enthält der mündliche Fragebogen. Sie ist wie folgt formuliert: „Haben Sie oder andere infolge der Situation die Polizei eingeschaltet?“ Die vorgegebenen Antwortkategorien lauten: 1) Ja, Polizei wurde *von mir selbst* eingeschaltet; 2) Ja, Polizei wurde *durch andere* eingeschaltet und 3) Nein, Polizei wurde nicht eingeschaltet.“ Bei dieser Frage geht es um eine Situation, die aus der Sicht der Befragten von allen vorher erwähnten am schlimmsten oder am belastendsten war.
Offensichtlich variiert die Formulierung, der inhaltliche Bezug und der Kontext der Fragen. Da aber nur die aggregierten Werte (Anteile von „verweigert“- und „keine Angabe“-Antworten) in beiden Modi verglichen werden, kann die Hypothese auf Grund dieser Daten untersucht werden.

Die Itemauswahl beschränkt sich nicht auf sensitive Fragen zu Gewalterfahrungen. Vielmehr wird Wert auf Ähnlichkeit der Items in verschiedenen Modi gelegt. Zuerst werden die Antwortkategorien „verweigert“ und „keine Angabe“ einheitlich kodiert. Dann werden die Anteile dieser Antwortkategorien an allen gültigen Antworten für jedes Item ausgerechnet. Als letzter Schritt wird der Durchschnitt der Anteile über alle 38 Items pro Modus ausgerechnet.
Der durchschnittliche Anteil der interessierenden Kategorien beträgt für den schriftlichen Drop-Off 0.126 und für den mündlichen Fragebogen 0.014. Somit ist der Anteil im schriftlichen Selbstausfüller deutlich höher als im Face-to-Face-Interview, was im Einklang mit der aufgestellten Hypothese ist und diese vorläufig bestätigt. Wie vermutet, ist der Item-Nonresponse im Face-to-Face-Interview geringer, was vor allem auf die höheren Einflussmöglichkeiten seitens der Interviewer und die damit verbundene verbesserte Motivation der Befragten zurückgeführt werden kann.

Hypothese 1b: Der Item-Nonresponse ist bei sensitiven Fragen (insbesondere zur Viktimisierung) in einem Face-to-Face Interview stärker ausgeprägt als in einem Selbstausfüller (einer schriftlichen Befragung)

Im Gegensatz zu neutralen Fragen wird bei sensitiven Fragen (hier insbesondere Fragen zur sexuellen, psychischen und physischen Gewalt) erwartet, dass der Item-Nonresponse in schriftlichen Befragungen geringer als bei telefonischen oder persönlichen Interviews ausfällt. Wie De Leeuw (1992) in seiner Untersuchung zu Moduseffekten betont, wenn „very sensitive questions [...] are asked, [...] mail surveys can even show less item nonresponse [...]“ (De Leeuw (1992): 32).
Für die Hypothese werden daher nur Fragen ausgewählt, die als sensitiv eingestuft werden. Die Mehrheit der ausgewählten Items bezieht sich auf detaillierte Angaben zur erlebten Gewalt. So lautet z.B. die Frage f13 „[...] Wie häufig haben Sie persönlich erlebt, dass einer *Ihrer früheren Partner* Sie körperlich angegriffen hat, Sie zum Beispiel geschlagen, geohrfeigt, an den Haaren gezogen, getreten oder mit einer Waffe oder einem Gegenstand bedroht hat?“
Pro Modus werden 81 sensitive Items ausgesucht, die als eine Zufallsauswahl der möglichen sensitiven Items betrachtet werden. Dadurch wird der beabsichtigte Hypothesentest möglich.
Das Vorgehen beim Testen entspricht dem der Hypothese 1a. Für beide Modi werden die jeweiligen durchschnittlichen Anteile von „verweigert“- und „keine Angabe“-Antworten an allen gültigen Antworten ausgerechnet und verglichen. Dieser Anteil beträgt für die schriftliche Befragung 0.056 und für die mündliche 0.025. Der Vergleich der Anteile widerspricht dem in der Hypothese 1b postulierten Zusammenhang.
Dafür gibt es jedoch eine plausible Erklärung. Erstens geht es beim schriftlichen Selbstausfüller um keinen klassischen Mail Survey, für den die Hypothese von De Leeuw (1992) bestätigt wurde. Die Anwesenheit der Interviewerin beim Ausfüllen sowie die Tatsache, dass sie im Großteil der Fälle den Fragebogen persönlich bekam, führt dazu, dass im Unterschied zu einem Mail Survey die Interviewereffekte beim schriftlichen Drop-Off nicht ausgeschlossen werden können. Dadurch ließe sich der höhere Item-Nonresponse bei der schiftlichen Befragung erklären.
Zweitens soll an dieser Stelle die fehlende Varianz im Interviewer- und Befragtengeschlecht als Ursache für die Ablehnung der Hypothese aufs Neue betont werden. Im Falle der vorhandenen Geschlechterdyaden hätte man vermutlich stärkere In-

terviewereffekte im Face-to-Face-Interview und fehlende Effekte im Mail Survey beobachten können. Dies hätte dann zum geringeren Item-Nonresponse bei sensitiven Fragen in der schriftlichen Befragung geführt. Da dies jedoch nicht der Fall ist, kann der erwartete Zusammenhang nicht festgestellt werden.

Hypothese 2: Der Anteil der beantworteten offenen Fragen in Viktimisierungssurveys ist in einem Selbstausfüller (einer schriftlichen Befragung) geringer als in einer Face-to-Face-Befragung

Diese Hypothese untersucht einen bestimmten Aspekt der Datenqualität. Die Annahme ist, dass verschiedene Modi unterschiedliche Datenqualität aufweisen und Moduseffekte daher als Vergleich der Datenqualität konzipiert werden können.
Wenn es sich bei dieser Hypothese ausschließlich um sehr sensitive Fragen handeln würde und man von starken Interviewereffekten beim mündlichen Interview (vor allem von Geschlechtereffekten) ausgehen würde, wäre ein umgekehrter Zusammenhang denkbar. Da diese beiden Bedingungen nicht zutreffen, erwarte ich eine höhere Antwortbereitschaft auf offene Fragen in einem Face-to-Face-Interview, bei dem die Motivation der Befragten durch die Interviewer positiv beeinflusst werden kann, der Zeitdruck eine geringere Rolle spielt und die Kontrollmöglichkeiten[10] beim Interviewer liegen.
Für die Analyse werden 15 offene Fragen aus dem mündlichen und 19 aus dem schriftlichen Fragebogen ausgewählt. Es werden dann die Anteile der gegebenen Antworten pro Item und der durchschnittliche Anteil pro Modus über ausgewählte Items begerechnet. Bei der mündlichen Befragung werden im Schnitt 86 % aller offenen Fragen beantwortet. Im schriftlichen Fragebogen sind es nur 79 %. Der Vergleich der Durchschnitte fällt zugunsten der Forschungshypothese aus, die somit akzeptiert werden kann.

Fazit

Es können also zwei der postulierten Hypothesen zu Moduseffekten anhand der Frauenstudie bestätigt werden. So wird zuerst der höhere Item-Nonresponse bei Viktimisierungsbefragungen bei Selbstausfüller im Unterschied zu persönlichen Befragungen bestätigt ($H1$). Des Weiteren können die Unterschiede im Befrag-

[10] z.B. wie lange ein Fragebogen ausgefüllt bzw. beantwortet wird.

tenverhalten zwischen den Modi bezüglich der Beantwortung offener Fragen ebenfalls bestätigt werden. Demnach sind die Befragten beim persönlichen Interview eher bereit, die zeitaufwendigeren offenen Fragen zu beantworten als wenn sie diese schriftlich beantworten sollten.

8 Fazit

Die vorliegende Arbeit beschäftigte sich mit den Verzerrungen der Ergebnisse sozialwissenschaftlicher Befragungen. Der Schwerpunkt der Arbeit lag dabei auf der Untersuchung der Non-Sampling Errors.
Die Grundlage für die Untersuchung bildete die Varianz der Datenqualität, die durch die Charakteristiken des Untersuchungsgegenstandes, Interviewermerkmale oder Charakteristiken des Befragten herbeigeführt werden kann. Hier wurde insbesondere auf die Einflüsse des Interviewers, die Interaktion der Interviewer- und Befragtenmerkmale sowie auf die Einflüsse des Befragungsmodus (Moduseffekte) eingegangen.
Die Ausprägung der Interviewer- und Moduseffekte wurde speziell für Viktimisierungsbefragungen untersucht. Dabei ging ich davon aus, dass sowohl die Interviewer- als auch Moduseffekte in Viktimisierungssurveys eine größere Bedeutung haben sollten. Diese Vermutung stützte sich auf die Ergebnisse früherer Untersuchungen zu Interviewereffekten bei sensitiven Fragen (vgl. Tourangeau (1996)).
Weiterhin wurde versucht, die hierarchische Struktur der Daten bei den Analysen zu beachten und, wenn möglich, Sampling-Point- von Interviewereffekten zu trennen. Für die Analysen wurden zwei aktuelle deutsche Datensätze ausgewählt - der DEFECT-Datensatz und die Frauenstudie[1].

Theoretisches Modell

Das im Kapitel 2 vorgestellte Erklärungsmodell des Befragtenverhaltens von Esser (1985) diente hier in erster Linie zur Erklärung der im Befragungsprozess entstehenden Interviewereffekte. Weiterhin konnte die Entstehung der Moduseffekte ebenfalls mit diesem Modell begründet werden, wenn die Moduseffekte als unterschiedliche Ausprägungen der Interviewereffekte bei verschiedenen Modi

[1] Zur Beschreibung der Datensätze siehe Kapitel 3.

modelliert wurden. Das Modell von Esser (1985) führte zur Formulierung der Hypothesen zu Interviewer- und Moduseffekten, die anschließend geprüft wurden. Ein entscheidender Vorteil des Modells bestand in der Angabe der Mechanismen der Entstehung der untersuchten Effekte sowie der Rahmenbedingungen, unter denen die Effekte vorkommen können. Außerdem war das Modell auch für die Erklärung des Nichtentstehens der interessierenden Effekte geeignet, da es explizit die Faktoren benennt, die dafür verantwortlich sein können.
Das Modell wurde demnach für Untersuchungen, wie die hier vorgenommenen, als geeignet betrachtet.

Analytisches Modell

Das zentrale Konzept für die Analysen in dieser Arbeit war der von Kish (1962) vorgeschlagene Intraklassenkorrelationskoeffizient. Er ermöglichte einen schnellen, von der jeweiligen Stichprobengröße unabhängigen Vergleich der Ergebnisse. Dieser war hier zwar nicht notwendig, wäre jedoch insbesondere beim Design neuer Studien von Relevanz, um deren Qualität erhöhen zu können. Der Intraklassenkorrelationskoeffizient stellte in den Analysen eine überschaubare und in ihrer Bedeutung eine gut nachvollziehbare Größe dar.
Der Weg zum Koeffizienten wurde auf zweifache Weise bestritten: über eine ein- bzw. zweifaktorielle ANOVA oder ANCOVA sowie über die Berechnungen mit Hilfe von -gllamm-. Beim Testen der Hypothesen zu Interviewereffekten im Kapitel 6 wurde bereits darauf hingewiesen, dass die Varianzanalyse kein optimales Modell für die hier verwendeten Daten darstellte, da eine Ähnlichkeit der Beobachtungen eines Klumpens (hier der Befragten in einem Sampling-Point) nicht modelliert werden konnte. Außerdem war das Skalenniveau einer Vielzahl von Variablen nicht ausreichend, um eine genaue Schätzung gewährleisten zu können. Für einige Hypothesen wurde daher eine Schätzung der hierarchischen Modelle mit -gllamm- vorgenommen.
Dabei wurde festgestellt, dass die Sampling-Point-Ebene einen sehr geringen Anteil an der totalen Varianz der abhängigen Variable erklärte. Somit waren die ANOVA-Ergebnisse in diesen Fällen gegenüber der erwähnten Klumpeneffekte auf der Sampling-Point-Ebene robust. Ein Problem stellte jedoch das Skalenniveau der Variablen dar. Für dichotome Items wurden die mit ANOVA berechneten Interviewereffekte unterschätzt, für einige andere Items wiederum

überschätzt. Insgesamt betrachtet stellte ANOVA kein optimales, aber ein durchaus akzeptables Modell dar, das das Erreichen des wichtigsten Ziels - des Auffindens von Interviewer- und Moduseffekten - ermöglichte.
Bei den Untersuchungen der Moduseffekte wurden außerdem verschiedene Anteilswerte mit Hilfe von T-Tests verglichen.

Interviewereffekte

Die Analysen des DEFECT-Datensatzes bestätigten eine starke verzerrende Wirkung des Interviewers im Befragungsprozess. Es wurde auch der Frage nach den Ursachen dieser Wirkung nachgegangen und verschiedene Hypothesen zum Effekt des Alters und des Geschlechts untersucht. Demnach fielen die Interviewereffekte bei männlichen Interviewern stärker als bei weiblichen aus. In der Ausprägung der Effekte für verschiedene Itemtypen konnte jedoch kein Unterschied festgestellt werden[2].
Weiterhin wurden keine stärkeren Interviewereffekte in verschiedengeschlechtlichen im Vergleich zu gleichgeschlechtlichen Dyaden bestätigt. Beim näheren Vergleich wurde festgestellt, dass männliche Interviewer bei Viktimisierungsbefragungen einen größeren Einfluss auf weibliche Befragte hatten als weibliche Interviewer. Eine ähnliche Wirkung bei der zweiten verschiedengeschlechtlichen Dyade „weiblicher Interviewer - männlicher Befragte“ vs. gleichgeschlechtliche „männlicher Interviewer - männlicher Befragte“ ließ sich nicht auffinden. Die Untersuchung der Alterseffekte ergab, dass der Effekt mit steigender Differenz zwischen dem Interviewer und dem Befragten stärker wurde. Die Hypothese über eine größere Bedeutung der Altersdifferenz in den Fällen, in denen der Interviewer älter als der Befragte war, blieb unbestätigt.
Für die Frauenstudie ließ sich die Mehrzahl der Hypothesen zu Interviewereffekten bestätigen. So wurden die allgemein verzerrende Wirkung des Interviewers und der Einfluss der Altersdifferenz zwischen dem Interviewer und dem Befragten für beide Itemtypen festgestellt. Die verstärkte Wirkung der Effekte für Typ1-Items wurde nur für binär kodierte Items[3] bei der ersten Hypothese beobachtet. Analog zum DEFECT-Datensatz wurde die Hypothese über stärkere Effekte in der Gruppe der negativen Altersdifferenzen im Vergleich zur Gruppe der positiven

[2]Zur Klassifizierung der Items siehe Kapitel 4.
[3]Zur Kodierung siehe Kapitel 4.

Altersdifferenzen nicht bestätigt.

Moduseffekte

Für die DEFECT-Daten wurden die Moduseffekte mit Hilfe des Intraklassenkorrelationskoeffizienten oder als Vergleich der Datenqualität verschiedener Befragungen modelliert. Die Hypothesentests führten zur Bestätigung der allgemeinen Moduseffekte. Weiterhin wurden stärkere Effekte in persönlichen als in telefonischen Interviews festgestellt. Für diese Hypothese wurde außerdem eine Trennung der Varianzen nach der Interviewer- und Sampling-Point-Varianz vorgenommen. Als Letztes wurden im Einklang mit früheren Untersuchungen die Anwesenheit der Recency-Effekte in telefonischen Befragungen bestätigt (vgl. Schwarz *et al.* (1989)). Somit konnten alle drei Forschungshypothesen vorläufig akzeptiert werden.
Anhand der Frauenstudie konnten zwei der postulierten Hypothesen zu Moduseffekten bestätigt werden. So wurde der höhere Item-Nonresponse bei Viktimisierungsbefragungen bei schriftlichen Interviews im Unterschied zu persönlichen Befragungen festgestellt. Nachfolgend konnten die Unterschiede im Befragtenverhalten zwischen den Modi bezüglich der Beantwortung offener Fragen ebenfalls bestätigt werden. Somit zeichneten sich die persönlichen Interviews im Unterschied zu den schriftlichen Interviews durch eine höhere Bereitschaft aus, die zeitaufwendigeren offenen Fragen zu beantworten.

Ausblick

Die festgestellten Interviewer- und Moduseffekte in beiden Studien führen zur erneuten Betonung der Wichtigkeit der sozialen Interaktionsprozesse während einer Befragung. Die Interaktion zwischen dem Interviewer und dem Befragten sowie bestimmte Merkmale eines Befragungsmodus sind oft für entstehende Verzerrungen in Antworten der Befragten verantwortlich. Diese Tatsache stellt erhöhte Anforderungen an die zukünftigen Studiendesigns, die mit diesem Wissen die Verzerrungen wenn möglich ausschließen oder zumindest kontrollieren sollten. Durch eine Kontrolle, wie sie im DEFECT-Datensatz vorgenommen wurde, können die entstandenen Effekte berechnet und die auf Grund der Daten berechneten Schätzungen der Populationsparameter - z.B. Standardfehler - korrigiert und somit präzisere Schätzungen erreicht werden.

Diese Arbeit konzentrierte sich auf einige ausgewählte Einflussfaktoren, wie z.B. Intervieweralter, Interviewergeschlecht oder ausgewählte Effektarten, wie z.B. die Recency-Effekte. Die Palette der möglichen Interviewer- und Moduseffekte ist jedoch vielfältiger. So könnten in weiteren Untersuchungen der Einfluss der Anwesenheit Dritter (differenziert nach Kindern/Erwachsenen oder Verwandte/nicht Verwandte) für Viktimisierungsbefragungen getestet werden, da dieser Einfluss für solche Studien von großer Bedeutung sein kann.
In Bezug auf die Analysemodelle können weiterhin Random-Slope-Modelle genutzt werden, um den Einfluss bestimmter Einflussfaktoren als gruppenspezifisch zu modellieren. Die durch die variierende Steigung der Geraden ausgedrückte gruppenspezifische Wirkung des ausgewählten Einflussfaktors kann neben dem Intraklassenkorrelationskoeffizienten als Hinweis auf die vorhandenen Interviewer- oder Moduseffekte gelten.

Literaturverzeichnis

Atteslander, P., & Kneubühler, H.-U. 1975. *Verzerrungen im Interview: zu einer Fehlertheorie der Befragung*. Studien zur Sozialwissenschaft, vol. 32. Opladen: Westdeutscher Verlag.

Bailar, B. 1983. *Encyclopedia of Statistical Sciences*. Vol. 4. New York. Chap. Interpenetrating Subsamples, pages 197–201.

Becker, R., & Günther, R. 2004. Selektives Antwortverhalten bei Fragen zum delinquenten Handeln. Eine empirische Studie über die Wirksamkeit der 'sealed envelope technique' bei selbst berichteter Delinquenz mit Daten des ALLBUS 2000. *ZUMA-Nachrichten*, **54**(Jg. 28), 39–59.

Behrens, K., & Löffler, U. 1999. *ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.* Opladen: Leske und Budrich. Chap. Aufbau des ADM-Stichproben-Systems, pages 69–91.

Bishop, G., Hippler, H.-J., Schwarz, N., & Strack, F. 1987. *A Comparison of Response Effects in Self-administered and Telephone Surveys*. ZUMA-Arbeitsbericht 11. Zentrum für Umfragen, Methoden und Analysen e. V., Mannheim.

Bradburn, N.M. 1983. *Handbook of Survey Research*. Quantitative Studies in Social Relations. Academic Press, Inc., New York. Chap. Response Effects, pages 289–328.

Christian, L.M., Dillmann, D.A., & Smyth, J.D. 2005 (December). *The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys*. TSM 2 Second Draft, unpublished.

Collins, M., & Butcher, B. 1982. Interviewer and Clustering Effects in an Attitude Survey. *Journal of the Market Research Society*, **25**, 39–58.

Davis, P, & Scott, A. 1995. The Effect of Interviewer Variance on Domain Comparisons. *Survey Methodology*, **21**, 99–106.

De Leeuw, E. D. 1992. *Data Quality in Mail, Telephone, and Face to Face surveys.* Proefschrift Vrije Universiteit Amsterdam. TT-Publikaties, Amsterdam.

De Vaus, D.A. 1991. *Surveys in Social Research.* 3rd. edn. Allen & Unwin Pty Ltd.

DeMaio, T.J. 1984. *Surveying Subjective Phenomena.* Vol. 2. New York: Russell Sage. Chap. Social Desirability and Survey Measurement: A Review., pages 257–282.

Dennis, J.M., Chatt, C., Li, R., Motta-Stanko, A., & Pulliam, P. 2005 (January). *Data Collection Mode Effects Controlling for Sample Origins in a Panel Survey: Telephone versus Internet.* Presented at the 2005 Annual Meeting for the American Association of Public Opinion Research.

DFG. 1999. *Qualitätskriterien der Umfrageforschung: Denkschrift.* Akademie Verlag, Berlin.

Diehl, J. M. 1977. *Varianzanalyse.* Methoden in der Psychologie, vol. 3. Fachbuchhandlung für Psychologie, Verlagsabteilung.

Esser, H. 1985. *Befragtenverhalten als „rationales Handeln" - Zur Erklärung von Antwortverzerrungen in Interviews.* ZUMA-Arbeitsbericht 1. ZUMA, Mannheim.

Esser, H. 1986. *Können Befragte lügen? Zum Konzept des „wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung.* ZUMA-Arbeitsbericht 2. ZUMA, Mannheim.

Fredebeul, C., Gilberg, R., Hess, D., Marwinski, G. Kästnerand K., & Prussog-Wagner, A. 2004 (Mai). *Methodenbericht „Lebenssituation, Sicherheit und Gesundheit von Frauen in Deutschland".* infas - Institut für angewandte Sozialwissenschaft GmbH im Auftrag des Bundesministeriums für Familie, Senioren, Frauen und Jugend.

Freeman, J., & Butler, E.W. 1976. Some Sources of Interviewer Variance in Surveys. *The Public Opinion Quarterly*, **40**(1), 79–91.

Groves, R. M., & Magilavy, L.J. 1986. Measuring an Explaining Interviewer Effects in Centralized Telephone Surveys. *The Public Opinion Quarterly*, **50**(2), 251–266.

Groves, R.M. 1989. *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

Groves, R.M., & Couper, M.P. 1998. *Nonresponse in Household Interview Surveys.* Wiley Series in Probability and Statistics. Survey Methodology Section. Wiley Interscience publication.

Groves, R.M., & Fultz, N.H. 1985. Gender Effects among Telephone Interviewers in a Survey of Economic Attitudes. *Sociological Methods Research*, **14**, 31–52.

Groves, R.M., Jr., F.J. Fowler, Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. 2004. *Survey Methodology.* Wiley Series in Survey Methodology. A John Wiley & Sons, Inc. Publication.

Hansen, M.H., Hurwitz, W.N., Marks, E.S., & Mauldin, W. P. 1951. Response Errors in Surveys. *Journal of the American Statistical Association*, **46**(254), 147–190. Bureau of the Census.

Hanson, M.H., & Marks, E.S. 1958. Influence of the Interviewer on the Accuracy of Survey Results. *Journal of the American Statistical Association*, **53**, 635–655.

Hochstim, J.R. 1967. A Critical Comparison of Three Strategies of Collecting Data from Households. *Journal of the American Statistical Association*, **62**, 976–988.

Holm, K. 1974. Theorie der Frage. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, **26**, 91–114.

Huddy, L., Billig, J., Bracciodetta, J., Hoeffler, L., Moynihan, P.J., & Pugliani, P. 1997. The Effect of Interviewer Gender on the Survey Response. *Political Behaviour*, **19**(3), 197–220.

Iversen, G.R., & Norpoth, H. 1976. *Analysis of Variance*. Quantitative Applications in the Social Sciences, nos. 07–001. Sage University Papers.

Johnson, T.P., & Parsons, J. A. 1993 (May). Measuring Interviewer Effects on Self-Reports from Homeless Persons. *In: Proceedings of the Survey Research Methods Section, American Statistical Association*, vol. 1. American Association for Public Opinion Research, St. Charles, IL.

Johnson, W.T., & Delameter, J.D. 1976. Response Effects in Sex Surveys. *Public Opinion Quarterly*, **40**, 165–181.

Kahn, R., & Cannell, Ch. F. 1968. *International Encyclopedia of the Social Sciences*. Vol. 8. Chap. Interviewing, pages 149–161.

Kish, L. 1962. Studies of Interviewer Variance for Attitudinal Variables. *Journal of the American Statistical Association*, **57**(297), 92–115.

Kohler, U., & Kreuter, F. 2001. *Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. R. Oldenbourg Verlag München.

Kreiselmaier, J., & Porst, R. 1989. *Methodische Probleme bei der Durchführung telefonischer Befragungen: Stichprobenziehung und Ermittlung von Zielpersonen, Ausschöpfung und Non-Response, Qualität der Daten.* ZUMA-Arbeitsbericht 12. Zentrum für Umfragen, Methoden und Analysen e. V., Mannheim.

Maccoby, E.E., & Maccoby, N. 1972. *Das Interview: Formen, Technik, Auswertung*. 7 edn. Praktische Sozialforschung. Köln: Kiepenheuer & Witsch. Chap. Das Interview: Ein Werkzeug der Sozialforschung, pages 37–85.

Mangione, T.W., Fowler, F.J., & Louis, T.A. 1992. Question Characteristics and Interviewer Effects. *Journal of Official Statistics*, **8**, 293–307.

Marczyk, G.R., De Matteo, D., & Festinger, D. 1964. *Essentials of Research Design and Methodology*. John Wiley & Sons, Inc.

Maxwell, S.E., & Delaney, H.D. 2004. *Designing Experiments and Analysing Data: a Model Comparison Perspective*. 2nd edn. Mahwah, New Yersey: Lawrence Erlbaum Associates, Inc.

McGraw, K.O., & Wong, S.P. 1996. Forming Inferences About Some Intraclass Correlations Coefficients. *Psychological Methods*, **1**(1), 30–46.

Müller, U., & Schröttle, M. 2004 (September). *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.

Newton, R.R., & Rudestam, K.E. 1999. *Your Statistical Consultant: Answers to Your Data Analysis Questions*. Sage Publications, Inc.

O'Muircheartaigh, C., & Campanelli, P. 1998. The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **161**(1), 63–77.

Phillips, D.L. 1971. *Knowledge from What: Theories and Methods in Social Research*. Chicago: Rand McNally and Co.

Rabe-Hesketh, S., & Everitt, B. 2004. *A Handbook of Statistical Analyses Using Stata*. 3rd edn. Chapmann & Hall/CRC.

Rabe-Hesketh, S., & Skrondal, A. 2005. *Multilelevel and Longitudinal Modeling Using Stata*. College Station, Texas: A Stata Press Publication.

Rabe-Hesketh, S., Skrondal, A., & A.Pickles. 2004. *GLLAMM Manual*. Paper 160 edn. University of California, Berkeley. U.C. Berkeley Division of Biostatistics Working Paper Series.

Reinecke, J. 1991. *Interviewer- und Befragtenverhalten: Theoretische Ansätze und methodische Konzepte*. Studien zur Sozialwissenschaft, vol. 106. Westdeutscher Verlag GmbH.

Robins, L.N. 1974 (May). *The Vietnam drug user returns*. Series A 2. Special Action Office for Drug Abuse Prevention, Special Action Office Monograph, US Government Printing Office, Washington D.C.

Schnell, R. 1994. *Graphisch gestützte Datenanalyse*. München: R. Oldenbourg Verlag.

Schnell, R., & Kreuter, F. 2000. Das DEFECT-Projekt: Sampling-Errors und Nonsampling-Errors in komplexen Bevölkerungsstichproben. *ZUMA-Nachrichten*, **47**(Jg. 24), 89–102.

Schnell, R., & Kreuter, F. 2005. Separating Interviewer and Sampling-Point Effects. *Journal of Official Statistics*, **21**(3), 389–410.

Schnell, R., Hill, P.B, & Esser, E. 2005. *Methoden der empirischen Sozialforschung.* 7 edn. R. Oldenbourg Verlag München.

Schuman, H., & Presser, S. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* Quantitative Studies in Social Relations. New York: Academic Press.

Schwarz, N., Bishop, G., Hippler, H.-J., & Strack, F. 1989. *Psychological Sources of Response Effects in Self-administered and Telephone Surveys.* ZUMA-Arbeitsbericht 1. Zentrum für Umfragen, Methoden und Analysen e. V., Mannheim.

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. 2002. *The Impact of Administration Mode on Response Effects in Survey Measurement.* ZUMA-Arbeitsbericht 90. Zentrum für Umfragen, Methoden und Analysen e. V., Mannheim.

Sczesny, S., & Stahlberg, D. 1999. Sexuelle Belästigung am Telefon. Definition, Prävalenz, Formen und Verarbeitung. *Zeitschrift für Sozialpsychologie*, **30**(2/3), 151–164.

Snijders, T.A.B., & Bosker, R.J. 1999. *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage Publications Ltd.

StataCorporation. 2003. *Stata User's Guide. Release 8.* A Stata Press Publication.

Stock, J.S., & Hochstim, J.R. 1951. A Method of Measuring Interviewer Variability. *Public Opinion Quarterly*, **15**, 322–324.

Sudman, S., & Bradburn, N.M. 1974. *Response Effects in Surveys: a Review and Synthesis.* Monographs in social research, vol. 16. Chicago: Aldine.

Sudman, S., & Bradburn, N.M. 1982. *Asking Questions.* Jossey-Bass Publishers.

Tucker, C. 1983. Interviewer Effects in Telephone Surveys. *The Public Opinion Quarterly*, **47**(1), 84–95.

Williams, J.A. Jr. 1964. Interviewer-Respondent Interaction: A Study of Bias in the Information Interview. *Sociometry*, **27**(3), 338–352.

A Zusätzliche Tabellen

Tabelle A.1: Prüfung der $H1$ (Interviewereffekte) am DEFECT-Datensatz. Koeffizienten (p_{int}), berechnet mit Anova und -gllamm- für die Face-to-Face-Befragung

Item	Niveau[1]	p_{int}-Anova	p_{int}-Interv.	p_{int}-SP
f10	ordinal	0.218	0.219	0.053
f31	binär	0.190	0.214	0.053
f43	interval	0.016	0.014	0.002
f13_2	ordinal	0.187	0.231	0.015
f1	interval	0.218	0.194	0.020
f14_3	interval	0.195	0.199	0.002
f34	binär	0.131	0.156	0.072
f25	binär	0.175	0.221	0.035
f35a	binär	0.175	0.192	0.162
f38_4	interval	0.142	0.149	0.000*
f35	binär	0.178	0.220	0.030
f13_4	ordinal	0.203	0.243	0.031
f14_2	interval	0.024	0.186	0.030
f42	binär	0.070	0.088	0.010
f33a	binär	0.000*	0.000*	0.000*
f44	interval	0.035	0.026	0.010

[1]Unter „Niveau“ ist hier die Verteilung der abhängigen Variable zu verstehen. Unter „p_{int}-Anova“ finden sich die Intraklassenkorrelationskoeffizienten, die für die Face-to-Face-Befragung mit Hilfe der ANOVA berechnet wurden. p_{int}-SP bzw. p_{int}-Interv. sind die Koeffizienten, die mit Hilfe von -gllamm- für die Sampling-Point-Ebene bzw. die Interviewerebene berechnet wurden. Die entsprechenden Werte für die telefonische Befragung sind in der Tabelle A.2 aufgeführt.

Tabelle A.2: Prüfung der $H1$ (Interviewereffekte) am DEFECT-Datensatz. Koeffizienten (p_{int}), berechnet mit Anova und -gllamm- für die CATI-Befragung

Item	Niveau[2]	p_{int}-Anova	p_{int}-Interv.	p_{int}-SP
f43	interval	0.001	0.005	0.000*
f42	binär	0.217	0.121	0.000*
f13_2	ordinal	0.042	0.067	0.027
f14_3	interval	0.007	0.019	0.031
f10	ordinal	0.009	0.021	0.088
f33a	binär	0.000*	0.000*	0.000*
f44	interval	0.000*	0.000*	0.000*
f34	binär	0.000*	0.056	0.000*
f1	interval	0.000*	0.012	0.005
f14_2	interval	0.027	0.007	0.010
f38_4	interval	0.016	0.000*	0.013
f35a	binär	0.000*	0.000*	0.000*
f35	binär	0.004	0.000*	0.027
f13_4	ordinal	0.070	0.037	0.042
f31	binär	0.012	0.000*	0.000*
f25	binär	0.008	0.000*	0.000*

[2]Unter „Niveau“ ist hier die Verteilung der abhängigen Variable zu verstehen. Unter „p_{int}-Anova“ finden sich die Intraklassenkorrelationskoeffizienten, die für die Face-to-Face-Befragung mit Hilfe der ANOVA berechnet wurden. p_{int}-SP bzw. p_{int}-Interv. sind die Koeffizienten, die mit Hilfe von -gllamm- für die Sampling-Point-Ebene bzw. die Interviewerebene berechnet wurden. Die entsprechenden Werte für die persönliche Befragung sind in der Tabelle 6.1 aufgeführt.

Tabelle A.3: Interviewereffekte in gleichgeschlechtlichen vs. verschiedengeschlechtlichen Dyaden. Prüfung der $H3$ am DEFECT-Datensatz

Typ 1-Items	p_{int}GG-Dyaden	p_{int}VG-Dyaden
f10	0.242	0.300
f13_2	0.147	0.226
f13_4	0.254	0.197
f14_2	0.214	0.243
f14_3	0.241	0.184
f25	0.187	0.186
f33a	0.084 (n.s.)	0.000* (n.s.)
f34	0.075	0.154
f35a	0.008 (n.s.)	0.293
f35	0.129	0.215
f42	0.206	0.257
lf44	0.294	0.296
$\overline{p_{int}}$	0.173	0.213
Typ 2-Items	p_{int}GG-Dyaden	p_{int}VG-Dyaden
f1	0.272	0.221
f31	0.217	0.181
f38_4	0.100	0.268
f43	0.421	0.418
$\overline{p_{int}}$	0.253	0.272
Total $\overline{p_{int}}$	0.193	0.227

Tabelle A.4: Zweiter Test der Interviewereffekte für GG- vs. VG-Dyaden. Interviewereffekte am DEFECT-Datensatz ($H3$)

Typ 1-Items	p_{int}GG-Dyaden	p_{int}VG-Dyaden
f10	0.109	0.054
f13_4	0.029	0.015
f14_3	0.034	0.000* (n.s.)
f25	0.028	0.000* (n.s.)
f33a	0.000* (n.s.)	0.000* (n.s.)
f42	0.250	0.358
lf44	0.433	0.506
$\overline{p_{int}}$	0.126	0.133
Typ 2-Items	p_{int}GG-Dyaden	p_{int}VG-Dyaden
f1	0.030	0.015
f31	0.000* (n.s.)	0.000* (n.s.)
f38_4	0.008	0.000* (n.s.)
f43	0.593	0.648
$\overline{p_{int}}$	0.158	0.166
Total $\overline{p_{int}}$	0.138	0.145

Tabelle A.5: $H4$, Interviewereffekte anhand des DEFECT-Datensatzes (Männliche Interviewer haben bei Viktimisierungsbefragungen einen größeren Einfluss auf weibliche Befragte als weibliche Interviewer)

Typ 1-Items	p_{int}Frau befragt vom Mann	p_{int} Frau befragt von Frau
f10	0.369	0.069
f13_2	0.218	0.082
f13_4	0.255	0.151
f14_2	0.220	0.089
f14_3	0.254	0.085
f25	0.169	0.057
f33a	n.a.[3]	0.363
f34	0.135	0.034
f35	0.207	0.060
f35a	0.211	0.000* (n.s.)
f42	0.165	0.103
lf44	0.082	0.011 (n.s.)
Typ 2-Items		
f1	0.228	0.064
f31	0.226	0.115
f38_4	0.256	0.013 (n.s.)
f43	0.080	0.016 (n.s.)
Total $\overline{p_{int}}$	0.192	0.082

[3]n.a.: der Koeffizient kann auf Grund der fehlenden Varianz „innerhalb“ der Interviewer nicht ausgerechnet werden.

Tabelle A.6: Einfluss der absoluten gruppierten Altersdifferenz. Interviewereffekte am DEFECT-Datensatz ($H6$)

Item	p_{int} 0-10	p_{int} 11-20	21-30	31-40	41 u. mehr
f1	0.175	0.304	0.275	0.327	0.339
f10	0.175	0.242	0.391	0.317	0.154 (n.s.)
f13_4	0.223	0.249	0.256	0.127	0.338
f14_3	0.167	0.271	0.253	0.000* (n.s.)	0.458
f25	0.180	0.199	0.133	0.133	0.000* (n.s.)
f31	0.207	0.207	0.274	0.253	0.188
f42	0.065	0.093	0.095 (n.s.)	0.405	0.280
f13_2	0.291	0.233	0.080 (n.s.)	0.431	0.247
f14_2	0.063 (n.s.)	0.498	0.545	0.208 (n.s.)	0.571
f34	0.121	0.241	0.179 (n.s.)	0.466	0.087 (n.s.)
f35	0.251	0.185	0.209	0.186 (n.s.)	0.333
$\overline{p_{int}}$	0.174	0.248	0.245	0.259	0.272

Tabelle A.7: $H1$, Interviewereffekte am Frauendatensatz für nicht binär kodierte Items

Typ1-Items	p_{int}	Typ2-Items	p_{int}
f110_3	0.040	f100	0.036
f110_4	0.034	f101	0.023
f111_3	0.038	f102	0.031
f111_4	0.019	f112_a	0.026
f203_1	0.080	f112_c	0.045
f203_3	0.019	f205	0.085
f204	0.077	f207	0.047
f300	0.066	f208	0.098
f303	0.050	f211	0.023
f309	0.023	f213	0.013
f310	0.010	f408	0.016
f501_a1	0.057	f600	0.025
f602	0.072	f732	0.032 (n.s.)
f604_1	0.023	f908	0.045
f605	0.083	f914	0.030
f800	0.040	f920_1	0.012
f802	0.017	f928	0.170
f913	0.042		
$\overline{p_{int}}$	0.044		0.045

Tabelle A.8: Einfluss der absoluten gruppierten Altersdifferenz. Interviewereffekte am Frauendatensatz ($H2$) für nicht binär kodierte Items

Typ1-Items	p_{int} 0-10	p_{int} 11-20	21-30	31-40	41 u. mehr
nf110_3[4]	0.027	0.038	0.029	0.071	0.215
nf110_4	0.029	0.042	0.035	0.073	0.132
nf111_3	0.059	0.052	0.035	0.042	0.076
nf111_4	0.023	0.029	0.002 (n.s.)	0.010 (n.s.)	0.049
nf203_1	0.080	0.081	0.096	0.170	0.140
nf203_3	0.033	0.034	0.061	0.086	0.188
nf204	0.074	0.122	0.051	0.157	0.158
nf300	0.094	0.057	0.080	0.135	0.100
nf303	0.065	0.036	0.061	0.185	0.084
nf501_a1	0.066	0.060	0.076	0.126	0.103
nf602	0.079	0.058	0.066	0.133	0.103
nf800	0.043	0.042	0.048	0.030 (n.s.)	0.043 (n.s.)
nf802	0.020	0.024	0.030	0.002 (n.s.)	0.065
nf913	0.083	0.037	0.072	0.123	0.079
$\overline{p_{int}}$	0.055	0.051	0.053	0.096	0.110
Typ2-Items					
nf100	0.043	0.026	0.060	0.045	0.102
nf101	0.023	0.019	0.025	0.044	0.111
nf102	0.035	0.030	0.077	0.010 (n.s.)	0.036 (n.s.)
nf112_a	0.037	0.037	0.045	0.107	0.087
nf112_c	0.059	0.059	0.071	0.213	0.190
nf205	0.115	0.115	0.093	0.141	0.219
nf208	0.091	0.120	0.143	0.134	0.031 (n.s.)
nf211	0.031	0.023	0.009 (n.s.)	0.073	0.114
nf213	0.018	0.018	0.012 (n.s.)	0.056	0.017 (n.s.)
nf408	0.097	0.041	0.202	0.293	0.279
nf600	0.064	0.052	0.018 (n.s.)	0.185	0.206
nf908	0.166	0.108	0.121	0.293	0.251
nf914	0.193	0.064	0.220	0.432	0.139
nf928	0.188	0.253	0.130	0.322	0.549
$\overline{p_{int}}$	0.083	0.069	0.088	0.168	0.167
Total $\overline{p_{int}}$	0.069	0.060	0.070	0.132	0.138

Tabelle A.9: Einfluss der absoluten gruppierten Altersdifferenz. Interviewereffekte am Frauendatensatz ($H2$) für binär kodierte Items

Typ1-Items	p_{int} 0-10	p_{int} 11-20	21-30	31-40	41 u. mehr
nnf204[5]	0.151	0.116	0.134	0.212	0.198
nnf303	0.160	0.143	0.157	0.177	0.109
nnf309	0.158	0.144	0.163	0.170	0.109
nnf310	0.160	0.144	0.161	0.174	0.106
nnf604_1	0.065	0.081	0.062	0.083	0.161
nnf913	0.122	0.094	0.108	0.161	0.153
$\overline{p_{int}}$	0.136	0.120	0.131	0.163	0.139
Typ2-Items					
nnf102	0.199	0.115	0.173	0.483	0.378
nnf205	0.122	0.095	0.107	0.214	0.241
nnf207	0.155	0.123	0.130	0.181	0.078
nnf208	0.142	0.103	0.117	0.166	0.063
nnf920_1	0.065	0.011 (n.s.)	0.068	0.131	0.045 (n.s.)
nnf928	0.060	0.024	0.038	0.028 (n.s.)	0.046 (n.s.)
$\overline{p_{int}}$	0.124	0.079	0.106	0.201	0.142
Total $\overline{p_{int}}$	0.130	0.099	0.118	0.182	0.141

[4]Die Items des analytischen Datensatzes wurden kopiert und umbenannt, um Transformationen an den ursprünglichen Variablen zu vermeiden. Dabei bekam jedes Item ein „n“ davor. Dadurch wurde z.B. das Item f204 zum nf204 usw.

[5]Die Items des analytischen Datensatzes wurden hier ebenfalls kopiert und umbenannt. Dabei bekam jedes Item ein „nn“ davor. Dadurch wurde z.B. das Item f204 zum nnf204 usw.

B Übersicht der ausgewählten Items

B.1 Items aus dem DEFECT-Datensatz

Typ1-Items

- **f10**: Denken Sie einmal nur an Ihre Wohngegend, also an alles, was Sie in 5 Gehminuten erreichen können. Wie sicher fühlen Sie sich, oder würden Sie sich fühlen, wenn Sie hier in dieser Gegend nachts draußen alleine sind? Fühlen Sie sich...
  sehr sicher - ziemlich sicher - ziemlich unsicher - sehr unsicher - weiß nicht.

- **f13_2**: Haben Sie manchmal Angst davor, daß Ihnen hier in dieser Wohngegend eine der folgenden Straftaten passieren könnte?
  Haben Sie immer, oft, gelegentlich, selten oder nie Angst davor,
  (Diese Teilfrage richtet sich nur an Frauen:) ...daß jemand Sie in Ihrer Wohngegend in sexueller Absicht tätlich angreift oder bedroht?

- **f13_4**: Haben Sie manchmal Angst davor, daß Ihnen hier in dieser Wohngegend eine der folgenden Straftaten passieren könnte?
  Haben Sie immer, oft, gelegentlich, selten oder nie Angst davor,
  ...daß ein Fremder Sie in Ihrer Wohnung aus geringem Anlaß schlägt oder körperlich verletzt? *(richtet sich an alle Befragten.)*

- **f14_2**: Bitte denken Sie nur an Ihre Wohngegend und an das, was Ihnen persönlich dort innerhalb der nächsten 12 Monate passieren könnte. Bitte geben Sie für jede der folgenden Straftaten an, für wie wahrscheinlich Sie

es halten, daß Ihnen persönlich hier so etwas in den nächsten 12 Monaten passiert. Geben Sie bitte einen Wert zwischen 100 % und 0 % an. Dabei bedeutet 100 % „es passiert mir ganz sicher“ und 0 % „ich halte es für ausgeschlossen“. Dazwischen können Sie jeden beliebigen Wert wählen.
Für wie wahrscheinlich halten Sie es,
(Diese Teilfrage richtet sich nur an Frauen): ...daß jemand Sie in Ihrer Wohngegend in sexueller Absicht tätlich angreift oder bedroht?

- **f14_2**: *(richtet sich an alle Befragten):* ... daß jemand Ihnen in Ihrer Wohngegend Gewalt androht, um an Ihr Geld ider Ihre Wertgegenstände zu kommen.

- **f25**: Was meinen Sie: Würde ein Einbrecher denken, daß es sich lohnt, in Ihre Wohnung einzubrechen?
  ja - nein - weiß nicht.

- **f33a**: Ist das innerhalb der letzten 12 Monate passiert? (f33: Wurden Sie selbst schon einmal von einem Fremden geschlagen oder verletzt?)
  ja - nein

- **f34**: *(nur Frauen:)* Wurden Sie am Telefon schon einmal sexuell belästigt?
  ja - nein - hatte noch nie Telefon

- **f35**: *(nur Frauen:)* Einmal abgesehen von sexueller Belästigung am Telefon: Haben Sie schon einmal daran gedacht, daß jemand Sie sexuell bedrohen oder in sexueller Absicht tätlich angreifen könnte?
  ja - nein

- **f35a**: *(nur Frauen:)* Haben Sie auch innerhalb der letzten 14 Tage daran gedacht? (sieh Frage oben.)
  ja - nein

- **f42**: Glauben Sie, daß Sie sich gegen einen gewaltsamen Angriff von einem unbewaffneten jungen Mann wehren könnten?
  ja - nein - weiß nicht

- **f44**: Wieviel wiegen Sie? Bitte geben Sie Ihr Körpergewicht in kg an.

Typ2-Items

- **f1**: Wie zufrieden sind Sie - alles in allem - mit der öffentlichen Sicherheit und der Bekämpfung der Kriminalität? Bitte antworten Sie auf einer Skala von 0 bis 10 - wobei 0 bedeutet, daß Sie ganz und gar unzufrieden sind - und 10 bedeutet, daß Sie ganz und gar zufrieden sind. Mit den Werten dazwischen können Sie Ihre Meinung abstufen.

- **f31**: Haben Sie schon einmal daran gedacht, daß ein Fremder Sie aus geringem Anlaß schlagen oder verletzen könnte?
  ja - nein

- **f38_4**: Die Straftaten, über die wir bisher gesprochen haben, sind nicht alle gleich wahrscheinlich. Niemand weiß genau, welche Ereignisse wie häufig eintreten. Denken Sie bitte einmal an 1000 Erwachsene aus der Gegend, in der Sie wohnen. Was würden Sie sagen: Wie viele davon werden innerhalb der nächsten 12 Monate in dieser Gegend in sexueller Absicht tätlich angegriffen oder bedroht? Bitte Anzahl Erwachsener eintragen.

- **f43**: Wie groß sind Sie? Bitte geben Sie Ihre Körpergröße in cm an.

B.2 Items aus dem Datensatz der Frauenstudie

Typ1-Items

- **f110_3**: Im folgenden geht es um Ihre Einschätzung von sich selbst und Ihrem Leben. Wir möchten Sie bitten, sich anhand der folgenden Aussagen selbst einzuschätzen. Geben Sie bitte für jede Aussage an, inwieweit diese auf Sie zutrifft. „Ich finde auch dann noch Wege, ein Problem zu lösen, wenn andere schon entmutigt sind.“
  trifft genau zu - trifft eher zu - trifft eher nicht zu - trifft gar nicht zu[1].

[1] Für die Beantwortung fast aller Fragen dieser Befragung wurden Listen mit Antwortvorgaben vorgelegt. Oft wurden die Fragen sehr umfangreich formuliert. Hier finden sich zum Teil verkürzte Versionen der Items (Ohne ausführliche Erklärung der Begriffe, die während der Befragung statt gefunden hatte.)

- **f110_4**: [...] „Ich fühle mich von Zeit zu Zeit richtig nutzlos.“ (Siehe oben.)
  trifft genau zu - trifft eher zu - trifft eher nicht zu - trifft gar nicht zu.

- **f111_3**: Nun geht es um die Beziehungen zu anderen Menschen. Bitte geben Sie an, inwieweit die Aussagen auf Sie zutreffen. Wenn Sie nicht ganz sicher sind, dann wählen Sie die Antwort, die Ihrer Meinung am nächsten kommt. „Es gibt genug Menschen, die mir helfen würden, wenn ich Probleme habe“.
  trifft genau zu - trifft eher zu - trifft eher nicht zu - trifft gar nicht zu.

- **f111_4**: (...)„Mir fehlt eine richtig gute Freundin bzw. ein richtig guter Freund“.
  trifft genau zu - trifft eher zu - trifft eher nicht zu - trifft gar nicht zu.

- **f203_1**: Haben Sie häufig, gelegentlich, selten oder nie Angst, dass ein Fremder Sie körperlich oder sexuell angreifen oder verletzen könnte?
  häufig - gelegentlich - selten - nie

- **f203_3**: (...) dass jemand aus Ihrer Familie oder Ihr Partner Sie körperlich oder sexuell angreifen oder verletzen könnte?

- **f204**: Wenn Sie abends alleine auf ein öffentliches Verkehrsmittel wie Bus, U- bzw. S-Bahn, Straßenbahn oder Zug warten oder damit fahren, wie sicher fühlen Sie sich dann? Nennen Sie mir bitte die entsprechende Kennziffer von dieser Liste. 1 bedeutet dabei „sehr sicher“, 6 bedeutet „überhaupt nicht sicher“. Mit den Werten dazwischen können Sie Ihr Urteil abstufen.
  sehr sicher bis überhaupt nicht sicher - Fahre nicht mit öffentlichen Verkehrsmitteln - Es gibt keine öffentlichen Verkehrsmittel bei uns

- **f300**: Viele Frauen fühlen sich in ihrem Alltag manchmal durch Bemerkungen, Berührungen oder Gesten sexuell bedrängt oder belästigt. Das kann auf der Straße oder an öffentlichen Orten sein, aber auch am Arbeitsplatz, in Ausbildung oder Studium sowie im Freundes-, Bekannten- und Familienkreis. Wie häufig haben Sie sich persönlich schon sexuell bedrängt oder belästigt gefühlt? Würden Sie sagen ...
  häufig - gelegentlich - selten - nie

- **f303**: Wenn Sie jetzt alle Situationen zusammennehmen, in denen Sie sexuell belästigt oder bedrängt wurden: Wie häufig haben Sie persönlich solche

Situationen durch wenig oder gar nicht bekannte Personen an öffentlichen Orten, Straßen, Plätzen bisher erlebt?
häufig - gelegentlich - selten - nur einmal - nie

- **f309**: Wie häufig haben Sie persönlich solche Situationen durch einen Partner oder Ehepartner erlebt? Würden Sie sagen...
  häufig - gelegentlich - selten - nur einmal - nie - trifft nicht zu, hatte nie einen Partner

- **f310**: Wie häufig haben Sie persönlich solche Situationen durch andere Familienangehörige oder Verwandte erlebt? Würden Sie sagen...
  häufig - gelegentlich - selten - nur einmal - nie

- **f501_a1**: Haben Sie persönlich diese Situation schon einmal erlebt? (ja - nein) Wenn ja, war das auch in den letzten 12 Monaten? (ja - nein): „Habe schon erlebt, dass man mich schwer beleidigt, eingeschüchtert oder aggressiv angeschrien hat.“

- **f602**: Haben Sie sich schon einmal eine Zeit lang große Sorgen darüber gemacht, wieviel Sie essen oder darüber, zu dick zu sein, zuzunehmen oder zu dick zu werden?
  ja - nein

- **f604_1**: Hat Ihnen schon einmal jemand, wie z.B. Verwandte oder Freunde, gesagt, dass Sie viel zu dünn seien oder dass Sie wie ein Skelett aussehen würden? Wenn ja, in welchem Jahr war das?
  ja - nein (evtl. Jahr)

- **f605**: Hatten Sie in dieser Zeit, als Ihr Gewicht am niedrigsten war, große Angst, wieder zuzunehmen?
  ja - nein

- **f800**: Wie häufig haben Sie seit dem Alter von 16 Jahren ungewollte sexuelle Handlungen erlebt, zu denen Sie gedrängt oder psychisch oder moralisch unter Druck gesetzt wurden? Würden Sie sagen...
  häufig - gelegentlich - selten - nur einmal - nie

- **f802**: Wie häufig haben Sie seit dem Alter von 16 Jahren solche erzwungenen sexuellen Handlungen erlebt? War das...
  häufig - gelegentlich - selten - nur einmal - nie

- **f913**: Bitte schätzen Sie, wie hoch in etwa der prozentuale Anteil ist, den Sie zum Haushaltseinkommen beitragen.
  circa ... Euro - habe kein eigenes Einkommen

Typ2-Items

- **f100**: Wie zufrieden sind Sie alles in allem mit Ihrer derzeitigen Lebenssituation? Bitte sagen Sie es anhand der Werte zwischen 1 und 6, wobei 1 bedeutet „sehr zufrieden“, und 6 „sehr unzufrieden“. Mit den Werten dazwischen können Sie Ihr Urteil abstufen. (Eine Liste mit Antwortvorgaben wird vorgelegt.)

- **f101**: Wie zufrieden sind Sie mit Ihrer derzeitigen Lebenssituation, was den Bereich Freunde und Familie betrifft? Nennen Sie mir bitte wieder einen Wert zwischen 1 und 6.

- **f102**: Wie zufrieden sind Sie mit Ihrer derzeitigen Lebenssituation, was den Bereich Ausbildung und Berufsleben betrifft? Nennen Sie mir bitte wieder einen Wert zwischen 1 und 6.

- **f112_a**: Auf dieser Liste sind verschiedene Freizeitaktivitäten aufgeführt. Bitte sagen Sie mir für jede Aktivität, wie häufig Sie diese zur Zeit machen. Freunde, Verwandte oder Bekannte besuchen.
  häufig - gelegentlich - selten - nie

- **f112_c**: (...)Besuch von Kino, Theater oder sonstigen Kulturveranstaltungen. (Siehe oben.)

- **f205**: Wie oft benutzen Sie abends alleine öffentliche Verkehrsmittel? Was von dieser Liste trifft zu?
  Täglich - mehrmals pro Woche - einmal pro Woche - 1-3mal im Monat - mehrmals im Jahr - seltener - nie

- **f207**: Wenn Sie alleine mit Ihrem Auto in eine Parkgarage fahren, wie sicher fühlen Sie sich dann? Nennen Sie mir bitte die entsprechende Kennziffer

von dieser Liste. 1 bedeutet dabei „sehr sicher“, 6 bedeutet „überhaupt nicht sicher“. Mit den Werten dazwischen können Sie Ihr Urteil abstufen. (zusätzliche Antwortmöglichkeiten: Benutze keine Parkgaragen - Trifft nicht zu, habe bzw. fahre kein Auto, es gibt hier keine Parkgaragen.)

- **f208**: Wie oft benutzen Sie Parkgaragen alleine?
  Täglich - mehrmals pro Woche - einmal pro Woche - 1-3mal im Monat - mehrmals im Jahr - seltener - nie

- **f211**: Wenn Sie abends oder nachts alleine in Ihrer Wohnung sind, wie sicher fühlen Sie sich dann? Nennen Sie mir bitte wieder die entsprechende Kennziffer von dieser Liste (siehe f207). Zusätzliche Antwortmöglichkeit: „Bin abends nie allein zu Hause“.

- **f213**: Wurden Sie schon einmal beraubt oder waren Sie Opfer eines Überfalls, wo Ihnen jemand Ihre Tasche, Geldbörse, Schmuck etc. gewaltsam entrissen hat?
  ja - nein

- **f408**: Sind Sie zur Zeit erwerbstätig? Unter Erwerbstätigkeit wird jede bezahlte bzw. mit einem Einkommen verbundene Tätigkeit verstanden, egal welchen zeitlichen Umfang sie hat. Was von dieser Liste trifft auf Sie zu?
  Ich bin zur Zeit: Vollzeit erwerbstätig (35 Stunden pro Woche und mehr) - Teilzeit erwerbstätig (15 bis unter 35 Stunden pro Woche) - Geringfügig oder unregelmäßig erwerbstätig (unter 15 Stunden pro Woche) - Vorübergehend freigestellt (z.B. Mutterschafts-/Erziehungsurlaub, sonstige Beurlaubung) - Auszubildende / Lehrling / Umschülerin / Studentin - nicht erwerbstätig (einschließlich: Schülerin, Arbeitslose, Vorruhestand, Rentnerin)

- **f600**: Wie würden Sie Ihren aktuellen Gesundheitszustand beschreiben? Bitte sagen Sie es anhand von Werten zwischen 1 und 6, wobei 1 bedeutet „sehr gut“, und 6 „sehr schlecht“. Mit den Werten dazwischen können Sie Ihr Urteil abstufen.

- **f732**: Wie zufrieden waren Sie mit der medizinischen Hilfe? Nennen Sie mir bitte einen Wert von dieser Liste. 1 bedeutet „sehr zufrieden“, 6 „sehr

unzufrieden“. Mit den Werten dazwischen können Sie Ihr Urteil abstufen.

- **f908**: Wie viele Personen leben insgesamt in Ihrem Haushalt, Kinder und Sie selbst mit eingeschlossen? Zählen Sie bitte dazu auch Kleinkinder bzw. Personen, die normalerweise hier wohnen, aber zur Zeit abwesend sind, z.B. im Krankenhaus oder im Urlaub sind.
  lebe allein - insgesamt ... Personen

- **f914**: Welchen Familienstand haben Sie heute? Sind Sie ...
  verheiratet und leben mit Ihrem Ehepartner zusammen - verheiratet und vom Ehepartner getrennt lebend - sind Sie geschieden - verwitwet - oder sind Sie ledig

- **f920_1**: Ist Ihr heutiger Partner etwa gleich alt wie Sie, älter oder jünger als Sie? Etwa gleich alt.
  ja - nein

- **f928**: Bitte geben Sie mir den ungefähren Nettoverdienst Ihres Partners / Ihrer Partnerin an. (Eine Liste wird vorgelegt.)

C Abkürzungsverzeichnis

ADM	Arbeitskreis deutscher Marktforschungsinstitute
ANOVA	Analysis of Variance
ANCOVA	Analysis of Covariance
CATI	Computer Assisted Telephone Interviewing
DFG	Deutsche Forschungsgemeinschaft
GG-Dyaden	Gleichgeschlechtliche Dyaden Interviewer-Befragte
gllamm	Generalized Linear Latent and Mixed Models (STATA-Befehl)
H	Hypothese
IFF	Interdisziplinäres Frauenforschungszentrum (Universität Bielefeld)
infas	Institut für angewandte Sozialwissenschaft GmbH (Bonn)
n.a.	kann nicht ausgerechnet werden (Intraklassenkorrelationskoeffizient)
n.s.	nicht signifikant
p_{int}	Intraklassenkorrelationskoeffizient
$\overline{p_{int}}$	Durchschnittlicher Intraklassenkorrelationskoeffizient
RLD	Randomized Last Digit
VG-Dyaden	Verschiedengeschlechtliche Dyaden Interviewer-Befragte