Der Einfluß gefälschter Interviews auf Survey-Ergebnisse

Schnell, Rainer

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zeitschrift für Soziologie, Jg. 20, Heft 1, Februar 1991, S. 25–35

Der Einfluß gefälschter Interviews auf Survey-Ergebnisse

Rainer Schnell
Institut für Angewandte Sozialforschung, Universität zu Köln, Greinstr. 2, D-5000 Köln 41

Zusammenfassung: Zu den vielen möglichen Kritikpunkten an Umfrageergebnissen gehören Verzerrungen durch gefälschte Interviews. Diese Bedenken werden analytisch und empirisch untersucht. Interviewfälschungen sind ein Spezialfall von „Missing-Data-Problemen" und können daher mit denselben Formeln abgeschätzt werden. Die entsprechenden analytischen Ergebnisse legen für einfache Statistiken nur kleine Verzerrungen nahe. Da solche analytischen Abschätzungen für multivariate Statistiken kaum möglich sind, werden mögliche Verzerrungseffekte mit verschiedenen Methoden empirisch untersucht. Die Ergebnisse einer Untersuchung zur „Qualität" gefälschter Interviews mit 22 „Interviewern", die je 10 „Interviews" durchführten, zeigt zwar u. a. eine größere Konsistenz gefälschter Interviews im Vergleich zu echten Interviews, aber keine größeren Unterschiede zu den echten Daten. Diese Fälschungen hätten sich weder auf die Berechnung univariater Statistiken noch auf multivariate Analysen ausgewirkt, wenn sie 5%-Bestandteil eines Datensatzes gewesen wären. Schließlich wird mit einigen Simulationen die Robustheit eines Regressionsmodells selbst gegenüber höheren Anteilen von Interviewfälschungen demonstriert. Falls die Abschätzung möglicher Effekte von Interviewfälschungen notwendig erscheint, müssen ähnliche Simulationen in jedem Einzelfall durchgeführt werden.

Zu den vielen möglichen Kritikpunkten an den Ergebnissen empirischer Sozialforschung gehört die Verzerrung der Ergebnisse durch gefälschte Interviews. Die Publikumswirksamkeit dieses Argumentes steht – wie so oft – in umgekehrtem Verhältnis zu dem Ausmaß vorhandener empirischer Daten: Es ist weder allgemein der Anteil gefälschter Interviews, noch die „Qualität" der Fälschungen, noch die mögliche Verzerrung der Ergebnisse durch die Fälschungen bekannt. Die Folklore der empirischen Sozialforschung ist voller Horrorgeschichten über Studien, bei denen zumindest ein Teil der Datenerhebung durch Interviewer allein in deren Wohnungen stattfand. Da sich – aus guten evolutionären Gründen – die Aufmerksamkeit bei Menschen eher auf ungewöhnliche denn auf reguläre Ereignisse richtet, könnte die Wahrnehmung weniger Unregelmäßigkeiten (Fälschungen) zu einer subjektiv weit größeren Gefährdung der Ergebnisse führen, als sie objektiv möglich ist. Die maximal mögliche Verzerrung von Survey-Ergebnissen durch Fälschungen läßt sich zumindest teilweise quantifizieren. Um den Effekt gefälschter Interviews auf Survey-Ergebnisse allgemein[1] zu bestimmen, muß zunächst gezeigt werden, daß die möglichen Verzerrungen eine Funktion des Ausmaßes der Fälschungen und der „Güte" der Fälschungen sind.[2]

Analytische Abschätzung der möglichen Verzerrungen durch Fälschungen

Falls überhaupt Abschätzungen der möglichen Effekte von gefälschten Interviews auf die Verzerrung von statistischen Schätzern erfolgten, wurden diese anscheinend bisher nicht veröffentlicht.[3]

Die Möglichkeit der Abschätzung der Effekte wird durch die Überlegung ermöglicht, daß ein Datensatz mit Fälschungen einem Datensatz, bei dem fehlende Daten durch Ersetzungen geschätzt wurden („Imputations") entspricht. Die Ersetzung fehlender Werte in Datensätzen durch „Experten-

[1] Der Vorteil einer analytischen Lösung liegt wie stets in ihrer Allgemeinheit: Sind die Parameter bekannt, so sind die Effekte berechenbar. Die Parameter können aus verschiedenen Quellen geschätzt werden, folglich erlauben analytische Lösungen die Berechnung der Effekte. Keine (erst recht: keine qualitative) Erhebung tatsächlichen Fälschungsverhaltens (z. B. durch Befragungen von Fälschern) erlaubt solche quantitativen Abschätzungen.

[2] Als Güte der Fälschung wird hier die Differenz zwischen dem „wahren Wert" des eigentlich zu Befragenden und der gefälschten Angabe des Interviewers bezeichnet.

[3] Dies mag zum Teil durch die offensichtliche Besorgnis vieler Erhebungsorganisationen begründet sein, irgendwelche Probleme bei ihren Datenerhebungen einzugestehen. Insbesondere in der Bundesrepublik werden von den kommerziellen Instituten kaum Daten zu Fälschungen, Ausschöpfungen usw. veröffentlicht. Das statistische Bundesamt ist in dieser Hinsicht führend: Weder das Ausmaß der Probleme noch die offensichtlichen Korrekturen an erhobenem Material werden öffentlich dokumentiert.

ratings“[4] unterscheidet sich nur durch die quantifizierbare Güte der „Expertenschätzung“ gegenüber der Interviewfälschung. Schließlich ersetzt bei einer Fälschung der Interviewer als Experte fehlende Daten durch seine Schätzung.[5] Die Verzerrung der Schätzungen durch Fälschungen und die Verzerrung durch Ersetzung fehlender Werte ist daher formal identisch. Fälschungen stellen also so betrachtet lediglich eine Variante eines speziellen „Missing-Data-Problems“ dar (vgl. Schnell 1986). Da die Auswirkungen gefälschter Interviews daher den Auswirkungen von Nonresponse ähneln, können die Formeln zur Berechnung des Nonresponsebias auf dieses Problem angewendet werden.[6] Die folgenden Formeln sind lediglich einfache Adaptionen der Formeln für den Nonresponsebias bei Kalton (1983: 6–10).

Der einfachste Fall betrifft die statistische Schätzung von Anteilswerten einer Variablen. Die mögliche Verzerrung der Schätzung der Anteilswerte kann nicht größer sein als der Anteil der Fälschungen insgesamt. Der sich ergebende Anteilswert (P_g) ist eine Funktion der Differenz zwischen dem Anteil in den echten Interviews (P_t) und dem Anteil in den gefälschten Interviews (P_f) gewichtet mit dem Anteil der Fälschungen an allen Fällen (A_f):

$$P_g = P_t - A_f (P_t - P_f)$$

Bei 5% Fälschungen kann sich also maximal eine Differenz von 5% gegenüber dem tatsächlichen Anteilswert ergeben. Sobald die Interviewer nur minimal bessere Schätzungen abgeben als durch Würfeln zu erreichen wäre, werden die Verzerrungen kleiner.

Für die Mittelwerte ergeben sich analog zu den Anteilswerten die Schätzungen:

$$M_g = M_t - A_f (M_t - M_f)$$

Die meisten Variablen der empirischen Sozialforschung besitzen sehr kleine Wertebereiche, z. B. 1 bis 7 oder 1–10. Nur in seltenen Fällen wird der Wertebereich 0–100 überschritten. Bei 5% Fälschungen bedeutet dies also bei den 0–100-Skalen eine maximale Verzerrung von ± 5, bei den 7-stufigen Skalen um ± 0.3.

Die Verzerrung (B) von Subgruppenmittelwertdifferenzen (MD) ist folglich:

$$B_{MD} = A_{fa} (M_{Ta} - M_{Fa}) - A_{fb} (M_{Tb} - M_{Fb})$$

wobei A_{fa} und A_{fb} die Anteile der Fälschungen in den Subgruppen a und b, M_{Ta} und M_{Tb} die Mittelwerte der echten Interviews in den Subgruppen und M_{Fa} und M_{Fb} die Mittelwerte der gefälschten Interviews in den Subgruppen sind.

Da die gesamte Varianz mit

$$S^2 = (1 - A_f) S_T^2 + A_f S_f^2 + A_f (1 - A_f) (M_T - M_F)^2$$

geschätzt werden kann, ergibt sich die Verzerrung der Varianz als

$$B_s^2 = A_f (S_T^2 - S_F^2) - A_f (1 - A_f) (M_T - M_F)^2.$$

Bei angenommenen 5% Fälschungen, einer 10% geringeren Varianz der Fälschungen und fast maximalen Differenzen der Mittelwerte von standardnormalverteilten Variablen (6.0) wäre bereits eine Überschätzung der Varianz um den Faktor 2.7 möglich. Geht man hingegen von realistischeren (fast) identischen Mittelwerten in beiden Gruppen aus, so wird bei 5% Fälschungen und standardnormalverteilten Variablen für eine 5%-Unterschätzung der Varianz die Annahme konstanter Werte für die Fälschungen (Varianz=0) erforderlich. Falls die fälschenden Interviewer nicht allzu unrealistische Mittelwerte produzieren, ist also bei 5% Fälschungen auch bei starker Homogenität der Fälschungen nur mit einer minimalen Verzerrung der Varianz zu rechnen.

Die sich ergebende Kovarianz läßt sich mit

$$S_{xy} = (1 - A_f) S_{Txy} + A_f S_{Fxy} + A_f (1 - A_f) (M_{Tx} - M_{Fx}) (M_{Ty} - M_{Fy})$$

berechnen[7], wobei S_{xy} die geschätzte Kovarianz der Variablen x und y, S_{Txy} und S_{Fxy} deren Kovarianz für die echten, bzw. gefälschten Daten und M_{Tx} und M_{Fx} deren Mittelwerte sind. Die Verzerrung der Kovarianz ist dann

$$BS_{xy} = A_f (S_{Txy} - S_{Fxy}) - A_f (1 - A_f) (M_{Tx} - M_{Fx}) (M_{Ty} - M_{Fy}).$$

---

[4] Zu solchen Expertenratings vgl. Rummel (1970: 262–263); zur empirischen Kritik der Leistungsfähigkeit von Expertenurteilen allgemein vgl. Dawes (1988: 201–227).

[5] Dies gilt nicht nur für Teil- und Totalfälschungen, sondern auch für die Fälschung durch bewußte Verletzung der Auswahlregeln: Ob die Ersetzung der Zielperson durch den Interviewer durch eine Fälschung oder den Statistiker durch ein Korrekturverfahren (z. B. „Doppeln“ oder Gewichten) vorgenommen wird, ist für die maximal mögliche Verzerrung weitgehend bedeutungslos (lediglich durch die Tatsache der Verdopplung entstehen einige mathematisch unangenehme Verbindungen zwischen den sonst als unabhängig betrachteten Stichprobenelementen, dies führt vor allem zu veränderten Schätzungen der Varianz der Schätzer, vgl. hierzu Platek/Gray 1983: 270–274).

[6] Hinweise für kompliziertere Statistiken lassen sich der Arbeit von Santos (1981) entnehmen, die sich aber ausschließlich mit den Effekten von Ersetzungsverfahren auf Schätzungen unter Annahme verschiedener Ausfallmodelle beschäftigt.

[7] Diese Formel gilt in dieser Form natürlich nur bei dem gleichen Ausmaß von Fälschungen in beiden Variablen.

Bei Annahme identischer Mittelwerte ist der Bias eine einfache Funktion des Anteils der Fälschungen. Bei angenommenen 5% Fälschungen und einer Kovarianz von null bei den Fälschungen wird die Kovarianz folglich nur um 5% unterschätzt. Bei Annahme identischer Mittelwerte, 5% Fälschungen und einer nur im Vorzeichen unterschiedlichen Kovarianz bei den Fälschungen (die Interviewer würden hierbei von einer impliziten Theorie mit falschem Vorzeichen ausgehen) ergäbe sich also eine Unterschätzung der Kovarianz um 10%. Bei realistischen Mittelwerten der Interviewerschätzungen sind also auch bei Kovarianzen kaum größere Verzerrungen zu erwarten.

Die analytischen Ergebnisse lassen somit für kleine Anteile von Fälschungen bei einfachen univariaten Statistiken nur kleine Veränderungen durch die Fälschungen erwarten. Geht man von der (wie noch zu zeigen sein wird: realistischen) Annahme nicht allzu großer Differenzen der Mittelwerte der gefälschten Variablen von den Mittelwerten der echten Variablen aus, so ist auch für Varianzen und Kovarianzen nicht mit großen Verzerrungen zu rechnen.

Für einfache Statistiken wie Mittelwerte, Varianzen und Kovarianzen lassen sich die möglichen Verzerrungen analytisch abschätzen. Für komplexere Statistiken, z. B. Regressionskoeffizienten, ist die Herleitung des Bias hingegen schwierig, in vielen Fällen kaum möglich. Für die praktische Abschätzung der möglichen Effekte bei komplexeren Statistiken muß daher auf einfache Simulationen zurückgegriffen werden.[8] Zentral für solche Simulationen sind natürlich wiederum das (vermutete) Ausmaß der Fälschungen und die Güte der Fälschungen. Die Güte der Fälschung hängt ihrerseits von der Art der Fälschung ab. Daher sollen die wenigen veröffentlichten Ergebnisse zu diesen Aspekten kurz referiert werden.

Das Ausmaß gefälschter Interviews

Durch die übliche Art von Interviewerkontrollen scheinen meist weit weniger als 1% der Interviewer aufzufallen.[9] Der Anteil der gefälschten Interviews dürfte wesentlich höher liegen, da meist nur höchstens 25% aller Interviews überprüft werden und die verwendeten Kontrolltechniken (Versendung von Kontrollpostkarten an vermutlich Befragte mit der Bitte um Rücksendung, telefonische Kontrollen) nicht als zuverlässig gelten können (vgl. Hauck 1969). Bei den wenigen veröffentlichten Studien, die intensive Interviewerkontrollen durchführten, liegen die Anteile dann auch stets höher.

Z. B. berichten Biemer/Stokes (1989: 25) die Ergebnisse eines zwischen 1982 und 1985 durchgeführten Projekts der amerikanischen Zensusbehörde zu Interviewerfälschungen. Hierbei konnten 3–5% aller Interviewer eine Fälschung nachgewiesen werden. Case (1971: 42) berichtet von 13 Studien, bei denen zusammen 2449 Befragte für eine Kontrolle der Interviewer ausgewählt wurden. Hiervon konnten 89% telefonisch erreicht werden. 4,1% der Interviews wurden als Fälschung erkannt, bei weiteren 22,7% gab es Durchführungsprobleme.

Obwohl insgesamt nur sehr wenige Daten hierzu veröffentlicht werden[10], scheint daher eine Schätzung des Anteils gefälschter Interviews mit ca. 5% aller Interviews realistisch.[11]

Formen der Fälschung

Vollständige Fälschungen sind für Interviewer schwierig herzustellen. Wesentlich einfacher als vollständige Fälschungen sind Teilfälschungen, bei denen einige Basisinformationen tatsächlich erfragt werden (z. B. telefonisch, bei Nachbarn oder bei einem anderen Haushaltsmitglied). Schließlich gibt es für die Interviewer noch die Möglichkeit der Befragung der falschen Zielperson.

Biemer/Stokes (1989: 25) berichten, daß 72% aller Fälschungen Totalfälschungen waren, weitere 17% der Fälschungen bestanden aus der falschen Angabe, daß eine Wohnung unbewohnt sei. Im National Crime Survey (NCS) bestanden 20 der 26 bestätigten Fälschungen aus der Befragung der falschen Person („Proxy-Interviews"). Es ist daher kaum erstaunlich, daß in dieser Studie fast ¾

[8] Diese entsprechen den „multiple imputations" für Nonresponse von Rubin (1987).

[9] Dies ist eine vorläufige Schätzung, die auf der Durchsicht aller im Zentralarchiv für empirische Sozialforschung in Köln vorhandenen Feldberichte für bundesweite Studien mit echten Zufallsstichproben basiert. Dem Zentralarchiv bin ich für die freundlicherweise gewährte Zugangsmöglichkeit zu den Feldberichten zu Dank verpflichtet. Eine umfangreiche quantitative Analyse der Feldberichte befindet sich in Vorbereitung.

[10] Nur wenige Feldberichte enthalten entsprechende Angaben. Auch in der Methodenliteratur werden solche Schätzungen kaum publiziert. Reuband (1990) berichtet z. B. nur die Zahl gefälschter Interviews des Methodenberichts des ALLBUS 1984.

[11] In Übereinstimmung mit dieser Schätzung geben Kirschhofer-Bozenhardt/Kaplitza (1982: 133) ohne jeden Beleg einen „internationalen Erfahrungswert" von 5–6% an.

aller Fälschungen nur durch eine Wiederholungsbefragung entdeckt wurden.

Die Wahrscheinlichkeit, daß Interviews gefälscht werden, variiert zwischen den Interviewern. Es gibt Hinweise darauf, daß sich die Fälschungen bei wenigen Interviewern konzentrieren: Case (1971: 42) berichtet, daß bei den beteiligten 632 Interviewern seiner 13 Studien mehr als 45 % der Fälschungen und Fehler auf ca. 18 % der Interviewer entfielen. Fast der Hälfte der Interviewer konnten keinerlei Fehler oder Fälschungen nachgewiesen werden, bei ca. 35 % gab es gelegentliche Durchführungsfehler. In der Studie von Biemer/Stokes (1989: 25) schien der Anteil der Fälschungen mit der Dauer der Tätigkeit als Interviewer zu sinken. Biemer/Stokes (1989: 25) erwähnen selbst, daß dies auch bedeuten kann, daß erfahrene Interviewer besser fälschen.[12]

Betrachtet man nur die nachgewiesenen Fälschungen, so fälschten erfahrene Interviewer einen kleineren Anteil ihrer Interviews als weniger erfahrene Interviewer (19 % der Interviews vs. 30 %). Die erfahreneren Interviewer begingen auch weniger Totalfälschungen (13 %) als unerfahrene Interviewern (ca. 50 %). Erfahrene Interviewer modifizieren eher die Auswahlregeln für die Befragten in ihrem Sinne (Schreiner et al. 1988: 492), indem z. B. ein leichter erreichbares Haushaltsmitglied anstelle der eigentlichen, schwer erreichbaren Zielperson im Haushalt befragt wird.

Die Qualität gefälschter Interviews

Keine einzige Studie scheint tatsächliche erkannte Fälschungen mit den „wahren Werten" der Zielpersonen zu vergleichen. Zur „Qualität" gefälschter Interviews scheint es neben einer (von Reuband 1990 zitierten) unveröffentlichten Studie von Jean Converse (1968) nur ein ebenfalls unveröffentlichtes Papier von Hippler (1979) zu geben.[13] Reuband (1990) legt die bisher umfassendste Studie zum Thema vor.[14] Sein wichtigstes Ergebnis besteht in dem Nachweis, daß zumindest Studenten in der Lage sind, solche Antwortmuster in fiktiven Interviews zu produzieren, die sich nicht von echten Antwortmustern – auch nicht in ihren Randverteilungen – unterscheiden lassen. Insgesamt sind die Differenzen zwischen echten und gefälschten Interviews in Reubands Untersuchung eher gering, lediglich die Konsistenz der Angaben in den gefälschten Interviews ist etwas größer als in echten Interviews.

In Hinsicht auf die hier interessierenden Aspekte weist die Studie von Reuband aber einige Lücken auf.[15] Die für die Nutzung von Umfragen zentrale Frage: „Wie robust sind die Ergebnisse gegenüber Fälschungen?" bleibt in der Literatur bisher unbeantwortet. Um der Beantwortung dieser Frage etwas näher zu kommen, wurde eine eigene Studie durchgeführt.

Theoretische Grundlage der empirischen Erhebung

Interviewern stehen drei Möglichkeiten der Fälschung zur Verfügung: Totalfälschung, Teilfälschung und Befragung der falschen Zielperson. Am einfachsten für den Interviewer und am schwierigsten nachzuweisen ist die Befragung der falschen Zielperson. Die komplizierteste Aufgabe für Interviewer sind Totalfälschungen, diese sind auch am ehesten zu entdecken. Interviewer, die ihre Bögen rein zufällig ausfüllen, werden vermutlich nur kurz in ihrem Beruf tätig sein.

Die Konsequenzen der drei Fälschungsstrategien sind unterschiedlich: Da die Befragung der falschen Zielperson ein echtes Antwortmuster erbringt, kann aus diesem nicht auf die Fälschung

---

[12] Wobei eine bessere Fälschung nur bedeutet, daß hier eine geringere Entdeckungswahrscheinlichkeit besteht. Es könnte sein, daß erfahrene Interviewer die Befragten zu stark typisieren und daher inhaltlich „schlechtere" Fälschungen produzieren und trotzdem geringere Entdeckungswahrscheinlichkeiten besitzen.

[13] Falls kommerzielle Unternehmen mit dem ihnen vermutlich reichlich zur Verfügung stehenden Material an erkannten Fälschungen systematische Studien zur Qualität der Fälschungen unternommen haben sollten, so sind diese anscheinend unveröffentlicht geblieben.

[14] Die Studie von Reuband basiert auf zwei Experimenten mit 39 bzw. 57 Studenten, die insgesamt 495 bzw. 464 Interviews „fälschten".

[15] Reuband arbeitete ausschließlich mit Studenten aus Einführungsveranstaltungen der empirischen Sozialforschung als fiktiven Interviewern und kann daher nur wenig über Unterschiede zwischen den Interviewern aussagen. Weiterhin gibt er zu vielen einzelnen Aspekten keine quantitativen Angaben, so z. B. über den Prozentsatz korrekter Schätzungen (dies ist aufgrund des Designs der Reubandstudie auch nicht möglich). Schließlich gibt es bei Reuband zwar einen Vergleich der Aggregatergebnisse der Fälschungen mit Umfragedaten, er unternimmt aber keinen Versuch abzuschätzen, wie sich die Umfragedaten durch die Fälschungen verändert hätten.

geschlossen werden. Die mögliche Verzerrung durch diese Art der Fälschung ist identisch mit dem Fall der (methodisch unzulässigen) Ersetzung eines Befragten (z. B. bei Nonresponse) durch einen anderen Befragten (vgl. hierzu Chapman 1983). Die resultierende Verzerrung ist eine Funktion der Differenzen zwischen der Zielperson und der befragten Person: Bei vollständiger Homogenität der Befragtenpopulation ist trivialerweise keine Verzerrung möglich, ansonsten steigt die Verzerrung mit der Heterogenität an.[16] Analytisch scheint dieser Fall kaum realistisch modelliert werden zu können, hier kann aber sehr leicht eine Abschätzung durch Simulation erfolgen.

Bei Totalfälschungen werden hohe Anforderungen an das Vorstellungsvermögen der Fälscher gestellt: Hierbei muß der Fälscher mit einem impliziten Modell des Antwortverhaltens („Laientheorien") arbeiten. Das gilt ebenso für Teilfälschungen. Den schlimmst möglichen Fall stellt hierbei keineswegs ein „zufälliges Ankreuzen" dar (dieses führt nur zur Erhöhung der unsystematischen Meßfehler) sondern falsche Laientheorien. Sowohl analytisch als auch in der Simulation ist das „zufällige Ankreuzen" unproblematisch für die Abschätzung der Verzerrung, dies ist bei der Verwendung falscher Modelle des Antwortverhaltens durch die Fälscher anders: Die Verzerrung hierdurch kann größer sein als bei rein zufälligem Ankreuzen. Dies wird insbesondere bei populären Laientheorien, z. B. über Einkommen und Wahlverhalten, der Fall sein, da entsprechende (falsche) Modelle des Antwortverhaltens von vielen Fälschern verwendet werden. Damit wird für die Abschätzung der möglichen Verzerrung durch Fälschungen die Frage nach der Güte der Laientheorien zentral. Eine Möglichkeit, die Güte der Umsetzung der Laientheorien zu überprüfen, besteht in der Untersuchung der Fähigkeit von Interviewern, die tatsächlichen Angaben der Befragten aus wenigen Schlüsselmerkmalen, meist demographischen Variablen, schätzen zu können.

Durchführung der Erhebung

Um den Vergleich geschätzter Angaben mit echten Angaben auf individueller Ebene durchführen zu können, wurden den Interviewern dieser Studie demographische Variablen von tatsächlich im Rahmen des ALLBUS 1988 Befragten als Basis der Schätzung anderer Variablen vorgeben. Da das tatsächliche Antwortverhalten[17] der Befragten bekannt ist, können die Schätzungen der Interviewer mit den tatsächlichen Angaben der Befragten direkt verglichen werden. Dieses Design erlaubt somit den individuellen Vergleich der Genauigkeit der Schätzung.[18]

Aus dem Datensatz des ALLBUS 1988 (ZA-Nr. 1670, n = 3052) wurden zunächst die Berliner Befragten ausgeschlossen, um die geplante Auswertung der Fragen zur Wahlabsicht bei der Bundestagswahl zu erleichtern. Aus der resultierenden Datei (n = 2915) wurde maschinell eine Zufallsauswahl (n = 300) gezogen. Diese Datei enthielt neben der ID-Nummer des Befragten unter anderem die Daten von 11 Variablen, von denen angenommen wurde, daß sie für Interviewer leicht erkennbar bzw. durch eine Befragung anderer als der Zielperson, vor allem von Nachbarn, leicht erfragbar wären: Land, Gemeindegrößenklasse, Telefon im Haushalt, Geschlecht, Alter, Familienstand, Zahl der Kinder unter 3 Jahren, Zahl der Personen im Haushalt, Typ der Wohnung, Berufstätigkeit und Stellung im Beruf. Daneben enthielt die Datei die Daten von 18 weiteren Variablen, die später von den Interviewern geschätzt werden sollten. Diese Variablen umfaßten ein Ethnozenstrismus-Item, politische Items, die Links-Rechts-Skala, die subjektive Schichteinstufung, die Wahlentscheidung bei der letzten Bundestagswahl, den allgemeinen Schulabschluß, die Frage nach Geschwistern, nach der Zugehörigkeit zu einer Religionsgemeinschaft, der Kirchgangshäufigkeit, dem Haushaltsnettoeinkommen, der Wahlabsicht, sowie eine Oben-Unten-Skala der gesellschaftlichen Selbsteinstufung. Weiterhin wurde die Frage nach

[16] Genau diese vollständige Homogenität innerhalb einer durch die Quotenvorgaben gebildeten Zelle wäre die einzige mögliche Legitimation für Quotenstichproben.

[17] Zwar besteht die entfernte Möglichkeit, daß auch die im ALLBUS 1988 vorhandenen Daten dieser Befragten ebenfalls Fälschungen sind. Der Anteil von Fälschungen dürfte beim ALLBUS allerdings niedriger liegen als bei anderen Befragungen. Sollte der ALLBUS 88 immer noch 5% Fälschungen enthalten, so wäre in dieser Studie mit ca. 7 Fällen zu rechnen, bei denen Fälschungen mit Fälschungen verglichen werden. Die wesentlichen Schlußfolgerungen dieser Studie könnten durch diese Fälle vermutlich kaum verändert werden.

[18] Ein solches Design verwendete auch Hippler (1979), der sich ohne Quellenangaben auf „mehrere Experimente in den USA" (Hippler 1979: 2) bezieht (hierbei handelt es sich vermutlich um die Arbeiten von Jean Converse). Das andere Design der Reuband-Studie erlaubt diesen Vergleich dagegen nicht.

der Anwesenheit Dritter beim Interview und die Frage nach der Dauer des Interviews aufgenommen.[19] Für jeden Fall dieser Datei wurde ein „Fragebogen" gedruckt, der die 11 Basisangaben und die Fragen zu den 18 Schätzvariablen[20] enthielt.

Da für die Untersuchung nur 22 „Interviewer" zur Verfügung standen, wurden 220 der 300 Fragebögen zufällig ausgewählt und jeweils 10 Interviews an die Interviewer verteilt. Jeder der Interviewer sollte auf Grund der 11 Angaben die Werte für die 18 anderen Variablen schätzen. Weiterhin sollten die Interviewer einen kurzen Interviewerfragebogen beantworten. Erhoben wurde Alter, Geschlecht, Semesterzahl, Zahl tatsächlich durchgeführter Interviews, Erfahrung in der Datenbereinigung bzw. Datenanalyse und benötigte Zeit für die Durchführung der Fälschungen. Die Datenerhebung für diese Studie erfolgte im April/Mai 1990. Die geschätzten Angaben der Interviewer wurden dann mit den tatsächlichen Daten zusammengeführt und bilden zusammen mit den Daten des Interviewerfragebogens die Datei, auf der ein Teil der folgenden Analysen basiert.

Die 22 Interviewer entstammen dem Umfeld dreier soziologischer Forschungsinstitute. Diese Art der Interviewerrekrutierung ist für nicht an Marktforschungsinstitute[21] delegierte Projekte typisch.[22] Die einzige wesentliche Abweichung gegenüber der üblichen Praxis besteht darin, daß neben 14 Studenten und 6 (zum Teil ehemaligen) wissenschaftlichen Mitarbeitern auch ein Soziologie-Professor „Interviews" durchführte. Dadurch konnte das Merkmal „sozialwissenschaftliche Kenntnisse" stärker variiert werden, als bei einer rein studentischen Stichprobe.

Das Alter der Interviewer lag zwischen 23 und 46, der Median bei 30 Jahren. Der „Interviewerstab" bestand aus 8 Männern und 14 Frauen. Die Studenten waren im 3. bis 24. Fachsemester, der Median lag bei 12 Semestern. Die Zahl der von den Interviewern durchgeführten echten Interviews lag zwischen 0 (6 Fälle) bis 250 (1 Fall), der Median lag bei 25. Nur 5 Interviewer hatten noch nie selbständig eine Datenanalyse durchgeführt, bei immerhin 4 Interviewern gehörte dies zu den ständigen Aufgaben. Die Interviewer benötigten zwischen 15 und 90 Minuten für das Ausfüllen aller Fragebögen, im Mittel 33 Minuten.

Ergebnisse zur „Qualität" gefälschter Interviews

Eine naheliegende Hypothese über das Fälschungsverhalten geht von der Unterschätzung der Varianz metrischer Variablen durch die Interviewer aus, da diese die Befragten zu stereotyp wahrnehmen und beschreiben würden. Betrachtet man die entsprechenden Variablen in dieser Untersuchung, so kann dies nicht bestätigt werden (vgl. Tabelle 1): Von 9 (bzw. 10) metrischen Variablen zeigt sich bei 3 (bzw. 4) Variablen eine Unterschätzung der Varianz, bei 5 Variablen eine Überschätzung der Varianz.[23] Betrachtet man die wahren und geschätzten Werte als wiederholte Messungen, so wären 4 der Differenzen signifikant (Links-Rechts-Skala, subjektive Schichteinstufung, Oben-Unten-Skala, subjektive Kompetenz für aktive Rolle in einer politischen Gruppe).

Bei den 4 bis 10 Kategorien umfassenden Skalen liegt die mittlere Abweichung zwischen –.47 und .28, im allgemeinen steigt die Abweichung mit der Zahl der Kategorien. Die Korrelation zwischen der Zahl der Kategorien und der mittleren absoluten Abweichung in diesen 9 Fällen liegt bei .82. Bei den metrischen Variablen können die Abweichungen der geschätzten von den tatsächlichen Werten im Aggregat als klein bezeichnet werden. Die Varianz der von den Interviewern geschätzten Daten unterscheidet sich zwar bei einigen Varia-

[19] Es handelt sich um die Variablen V12, V63–V66, V101, V106, V110, V154, V425, V431, V432, V435, V436, V507, V511, V519 und V527 des Datensatzes des ZA.

[20] Für die letzte Schätzvariable (Dauer des Interviews) waren die Anweisungen für die Interviewer dieser Studie offensichtlich mißverständlich, so daß die Ergebnisse hierzu nicht interpretierbar sind. Diese Variable wurde daher aus allen Analysen ausgeschlossen.

[21] Methodisch ist natürlich eine entsprechende Studie mit einem Interviewerstab eines kommerziellen Institutes in der BRD höchst wünschenswert. Da aber die kommerziellen Institute in der BRD fast nie Befragungen ihrer Interviewer ermöglichen, scheint eine solche Studie z. Z. in der BRD für die akademische Sozialforschung kaum realisierbar.

[22] Bei den von Buchhofer (1979) untersuchten 143 Interviewprojekten der empirischen Sozialforschung wurde bei 51% der Projekte ein eigener Interviewerstab ins Leben gerufen (Buchhofer 1979: 87), hierbei waren in 87% der Fälle ausschließlich Studenten als Interviewer tätig, in den verbleibenden 13% ein hoher Anteil (Buchhofer 1979: 172).

[23] Die unterschiedlichen Angaben basieren auf der Berücksichtigung bzw. Nichtberücksichtigung der Einkommensschätzung mit nur 41 gültigen Fällen.

Tabelle 1 Mittelwerte und Standardabweichungen echter (T) und gefälschter (F) ALLBUS-Interviews, sowie t-Werte der Differenz.

Variable		Mean T	Mean F	S T	S F	t
Gastarbeiter keine politische Tätigkeit	V12	3.93	4.23	2.30	1.90	−1.56
Politiker kümmern sich nicht	V63	2.15	2.05	.89	.91	1.23
Eigene aktive politische Rolle	V64	3.00	2.72	.99	.98	3.24
Keinen Einfluß auf Regierung	V65	2.33	2.17	.97	.98	1.85
Politik zu kompliziert	V66	2.72	2.60	1.03	1.00	1.27
Links-Rechts-Skala	V101	5.38	5,75	1.67	1.91	−2.19
Subjektive Schicht	V106	2.70	2.85	.63	.66	−2.57
Kirchgangshäufigkeit	V432	4.16	4.08	1.30	1.45	.71
Haushaltsnettoeinkommen	V435	2508	2542	1251	1190	−.19
Oben-Unten-Skala	V511	5.05	5.52	1.61	1.67	−3.34

blen ein wenig von den tatsächlichen Daten, aber keineswegs systematisch: Weder werden die Varianzen systematisch überschätzt oder unterschätzt noch variiert die Varianz der Schätzungen mit einem Interviewermerkmal.

Der mittlere Anteil vollständig korrekter Schätzungen der Interviewer lag je nach Interviewer zwischen 31% und 46%, im Mittel bei 37%.[24] Insgesamt wurden zwischen 0% und 71% der Werte korrekt geschätzt, der Median lag bei 35%. Die Varianz des Anteils korrekter Schätzungen schwankt zwischen den Interviewern erheblich (Standardabweichungen zwischen 8.9 und 19.3). Die Zahl korrekter Schätzungen variierte mit keinem der erhobenen Interviewermerkmale.

Obwohl die Randverteilungen und deskriptiven Statistiken für die meisten Variablen eine erstaunlich hohe Übereinstimmung zwischen echten und gefälschten Daten zeigt, ergaben sich bei einigen wenigen Variablen bei einzelnen Ausprägungen größere Differenzen.[25] Gaben z. B. 20 Befragte an, bei der Bundestagswahl nicht gewählt zu haben, so vermuteten die Interviewer dies nur bei 5 Personen. Bei der Wahlabsicht wurde von den Fälschern der CDU-Anteil um ca. 10% überschätzt, ebenso wurde hier die explizite Verweigerung der Auskunft unterschätzt (1,8% gegen 10,5% tatsächlich).

Die Interviewer überschätzten leicht den Anteil nicht substantieller Antworten. Die Befragten erreichten im Mittel 1.59 fehlende Angaben, die Interviewer hingegen 1.79. Insbesondere unterschätzten die Interviewer den Anteil vollständiger Angaben. Bei 25,9% der Befragten war jede Frage beantwortet, bei den Interviewern hingegen nur 0,5%. Die mittlere Anzahl unvollständiger Angaben und die Varianz der Anzahl fehlender Angaben hing ebenfalls mit keinem Interviewermerkmal zusammen. Lediglich bei den Interviews, die von Personen, die regelmäßig Datenanalysen durchführen, gefälscht wurden, ergaben sich tendenziell unvollständigere Antworten (Überschätzung: .38 gegenüber .17).

Die Hypothese der Stereotypisierung der Befragten durch die Interviewer wurde durch den Vergleich der Übereinstimmung trivialer Modelle des Antwortverhaltens mit dem tatsächlichen Antwortverhalten überprüft. Bei Stereotypisierung ist für ein solches triviales Modell, z. B. für ein Maß interner Konsistenz einer Likert-Skala, mit einem besseren Modellfit zu rechnen.

Diese Hypothese kann als bestätigt angesehen werden: Mit den 4 Items zur subjektiven politischen Kompetenz ergibt sich mit den echten Daten eine Likert-Skala mit einem Alpha von .64 bei Item-Scale-Korrelationen zwischen .33 und .52. Mit den geschätzten Daten ergibt sich ein Alpha von .87, die Item-Scale-Korrelationen liegen zwischen .65 und .83. Die Interviewer überschätzen

[24] In einer in dieser Hinsicht vergleichbaren Untersuchung von Hippler (1979: 11) ergaben sich für 21 ZUMA-Interviewer Anteile korrekter Schätzungen zwischen 35% und 50%. Obwohl die Variablen sehr ähnlich waren, erlaubt die unterschiedliche Zahl von Kategorien lediglich die Aussage, daß die Größenordnung der korrekt geschätzten Werte vergleichbar ist.

[25] Eine tabellarische Darstellung aller Differenzen der Randverteilungen würde den zur Verfügung stehenden Platz weit überschreiten. Eine Randauszählung und der Datensatz können beim Autor angefordert werden.

offensichtlich die Konsistenz der Angaben bei den „Einstellungsfragen" zu diesem Konstrukt.

Ebenso wird die Korrelation zwischen subjektiver Schichteinschätzung und der Oben-Unten-Skala überschätzt: Bei den Fälschungen ergibt sich r = –.69, bei den wahren Werten r = –.47. Entstammten die Korrelationen unabhängigen Stichproben, so ergäbe sich hier eine hochsignifikante Differenz.

Vergleicht man die Unterschiede in Hinsicht auf die Selbsteinschätzung auf der Links-Rechts-Skala zwischen den Angaben der Wahlabsicht, so zeigt sich die Stereotypisierung deutlich: Ergibt sich mit den tatsächlichen Daten ein Eta von .51 (Eta^2 = .26), so zeigt sich bei den gefälschten Daten ein Eta von .66 (Eta^2 = .44).

Interessant ist der Effekt der Fälschungen auf die multivariaten Statistiken. Hierzu wurde eine Likert-Skala aus den politischen Items gebildet und eine multiple Regression mit subjektiver Schichteinschätzung, Oben-Unten-Skala und Nettoeinkommen gerechnet. Mit den echten Daten ergaben sich mit subjektiver Schicht als einzigem erklärungskräftigen Prädiktor (beta = .41, b = 1.71) eine erklärte Varianz von 16,1%. Bei den gefälschten Daten besaß die subjektive Schichteinschätzung dagegen einen etwas geringeren Einfluß (mit beta = .25, b = 1.21). Als stärkster Prädiktor ergab sich hier die Oben-Unten-Skala (beta = –.46, b = .90). Insgesamt erklären die Variablen 40.2% der Varianz.

Eine Faktorenanalyse (Hauptkomponentenmethode, Oblimin-Rotation, pairwise) mit einem Item zur Einstellung gegenüber Gastarbeitern, 4 Items zur subjektiven politischen Kompetenz, der Links-Rechts-Skala und der Kirchgangshäufigkeit erbringt mit den echten Daten zwei nahezu orthogonale Faktoren, die zusammen 50% der Varianz erklären. Mit den gefälschten Daten ergeben sich ebenfalls zwei Faktoren, die aber zusammen 68% der Varianz erklären. Bei den 12 Faktorenladungen stimmen 10 im Vorzeichen überein, die maximale Differenz beträgt .30. Die Kommunalitäten liegen bei den gefälschten Daten immer höher, wobei die Differenz zwischen .05 und .29 liegt. Die Faktorenkorrelation beträgt tatsächlich –.07, bei den gefälschten Daten hingegen –.21.

Obwohl sich im wesentlichen bei den deskriptiven Statistiken keine großen Unterschiede zwischen gefälschten und echten Daten zeigen lassen, ist die Stereotypisierung nachweisbar: Die Interviewer arbeiten mit zumindest impliziten Modellen des Antwortverhaltens. Die Stereotypisierung ist aber weder stark genug noch führt sie zu so ungewöhnlichen und vorhersagbaren Kovarianzstrukturen, als daß sie für die Identifikation gefälschter Interviews verwendet werden könnte.

Ergebnisse zum Effekt gefälschter Interviews

Da sich insbesondere bei multivariaten Analysen Unterschiede zwischen den Antworten bei gefälschten und echten Interviews ergeben, stellt sich die Frage, ob diese Differenzen Parameterschätzungen auf der Basis eines Datensatzes, der vermutlich zu 95% aus echten Interviews besteht, verzerren würden. Um die möglichen Verzerrungen durch Fälschungen abzuschätzen, wurden zunächst von den 220 Fällen, für die Fälschungen vorlagen, 147 maschinell zufällig ausgewählt. Von jedem Interviewer wurden daher zwischen 4 und 10 Interviews verwendet. Danach wurden die Daten des ALLBUS 1988 derjenigen Fälle, für die zufällig ausgewählte Fälschungen vorlagen, durch die Fälschungen ersetzt. Dieser modifizierte Allbus enthielt also 4.8% bekannte Fälschungen. Mit diesem Datensatz wurden einige Analysen gerechnet, deren Ergebnisse mit den ursprünglichen verglichen wurden.

Beispielsweise hatten sich (wie oben erwähnt wurde) zwischen den Schätzungen der Interviewer und dem tatsächlichen Antwortverhalten bei der Wahlabsicht Unterschiede bis zu ca. 10% ergeben. Im modifizierten Allbus (mit Fälschungen) gab es erwartungsgemäß kaum noch Differenzen, so veränderte sich der CDU-Anteil von 24.6% auf 25.1%, der SPD-Anteil von 29.7% auf 29.3%. Der Anteil nichtsubstantieller Antworten veränderte sich von 31.1% auf 29.7%. Alle diese Differenzen liegen innerhalb der Stichprobenschwankungen und zeigen durch die Fälschung keine signifikante Veränderung der Randverteilung.

Eine multiple Regression mit der subjektiven politischen Kompetenz als abhängiger Variablen (Summe der Items V63–V66) und subjektiver Schichteinstufung, Oben-Unten-Skala und Nettoeinkommen als Pädiktoren zeigt für die echten Allbusdaten 11.6% erklärte Varianz, wobei nur zwei Prädiktoren signifikant sind. Die gefälschten Daten ergeben 12.6% erklärte Varianz, dabei sind alle Prädiktoren signifikant. Lediglich bei dem zusätzlich signifikanten Prädiktor ergibt sich eine etwas größere Veränderung des Regressionskoeffizienten. Insgesamt ist das Ergebnis dieser Regression gegenüber den Fälschungen sehr robust (vgl.

Tabelle 2 Vergleich der Ergebnisse einer multiplen Regression zwischen dem ALLBUS 1988 und dem ALLBUS 1988 mit Fälschungen.

		ALLBUS mit Fälschungen			ALLBUS		
Variable		B	Beta	T	B	Beta	T
Nettoeinkommen	V435	2.00E-4	.10	3.58	2.07E-4	.11	3.74
Subj. Schicht	V106	1.10	.27	8.64	1.08	.27	8.51
Oben-Unten	V511	–.12	–.07	–2.14	–.07	–.04	–1.33
(Constant)		6.22		11.07	6.05		10.90
R^2		12.6%			11.6%		

Tabelle 2). Um Mißverständnisse zu vermeiden, sei darauf hingewiesen, daß dieses Ergebnis selbstverständlich keinerlei allgemeine Geltung beanspruchen kann.

Simulationsergebnisse zum Effekt gefälschter Interviews

Um den Einfluß abweichender Ergebnisse in kleinen Teilgruppen für das Gesamtergebnis abschätzen zu können, wurden eine Reihe von Simulationen durchgeführt. Hierdurch können dann sowohl unterschiedliche Anteile von Fälschungen, als auch der Effekt verschiedener Fälschungsformen (z. B. „korrekte" Laientheorie, falsche Laientheorie, Zufallsmuster) auch in ihrem gemeinsamen Effekt auf multivariate Statistiken abgeschätzt werden, was analytisch nicht möglich ist.

Eine Variante bestand in der Simulation zufälliger Antwortmuster der Interviewer. Hierbei wurde für jede Variable des Regressionsmodells eine normalverteilte Variable mit gleichem Mittelwert und gleicher Standardabweichung wie die Originalvariable erzeugt.[26] Für die 147 Fälschungen wurden die Werte dieser Variablen in das Modell eingegeben (in dieser Subgruppe lag die erklärte Varianz bei nicht signifikant von 0 % verschiedenen 1.2 %). Für den so modifizierten ALLBUS mit Fälschungen lag die erklärte Varianz bei 9.6 %, gegenüber dem ALLBUS verändern sich die Regressionskoeffizienten fast nicht. Die größte Differenz ist die Veränderung des nichtstandardisierten Regressionskoeffizienten für V106 von 1.08 auf 0.94. Allerdings wären mit diesen Daten alle drei Prädiktoren als signifikant bezeichnet worden, im ALLBUS hingegen nur zwei.

In einem anderen Modell wurde davon ausgegangen, daß die Interviewer lediglich ihre subjektive Schichteinschätzung und die Oben-Unten-Skala als Prädiktoren für die Summe der Items der Skala verwendet hätten.[27] Damit ergibt sich für die simulierten Fälschungen ein $R^2 = 97.7\%$, mit beta = –.89 bzw. –.44. Werden diese simulierten Daten als Fälschungen in einem modifizierten ALLBUS berücksichtigt, so fällt R^2 von 11.6 % auf 10.1 %, alle Prädiktoren werden signifikant; die Koeffizienten verändern sich hingegen nur unwesentlich (der Regressionskoeffizient von V511 wächst von –.07 auf–.18, der Regressionskoeffizient für V106 sinkt von 1.08 auf .86, die standardisierten Koeffizienten verändern sich von –.04 auf –.10 bzw. von .27 auf .21. Obwohl sich ein Koeffizient mehr als verdoppelt, würde dies im allgemeinen kaum als inhaltliche Differenz gedeutet werden. Statistisch ist die Differenz der Koeffizienten für V511 aber signifikant. Trotzdem ist das Modell bei 5 % Fälschungen bemerkenswert stabil.

Um den Effekt eines größeren Anteils von Fälschungen beurteilen zu können, wurde eine weitere Simulation gerechnet, bei der die „gefälschten" Interviews der zweiten Simulation zusammen mit 853 zufällig ausgewählten echten Fällen des ALLBUS einen Datensatz mit nur noch 1000 Fällen bildeten, wobei 14.7 % Fälschungen waren. Hier ergeben sich deutlichere Veränderungen: 8.6 % erklärte Varianz, nur noch ein signifikanter Prädiktor (V511, b = –.48), die beiden anderen Koeffizienten werden fast null. Verwendet man den

[26] Die dabei unter dem empirischen Minimum bzw. über dem empirischen Maximum liegenden Werte wurden auf das Minimum bzw. Maximum recodiert.

[27] Die simulierte Skala ergab sich durch die Gleichung: SCALE = 18 – 0.9 * V106 – 0.8 * V511 + 0.2 * NORMAL (1).

Tabelle 3 Vergleich der Ergebnisse einer multiplen Regression zwischen einer Stichprobe des ALLBUS 1988 und der Stichprobe mit Fälschungen.

		n = 1000, modifizierter ALLBUS mit 14,7% Fälschungen			n = 1000, ALLBUS echte Daten		
Variable		B	Beta	T	B	Beta	T
Nettoeinkommen	V435	9.27E-05	.05	.93	9.63E-05	.05	.98
Subj. Schicht	V106	.84	.21	3.20	.82	.22	3.21
Oben-Unten	V511	–.30	–.18	–2.67	–.16	–.10	–1.45
(Constant)		8.15		6.72	7.48		6.46
R^2		12.3%			8.2%		

gleichen Datensatz mit 1000 Fällen und 14.7% Fälschungen, hierbei aber die tatsächlichen Fälschungen anstelle der simulierten, so ergeben sich neben einer erklärten Varianz von 12.3% zwei signifikante Prädiktoren (vgl. Tabelle 3).

Wie man sieht, verändert sich das R^2 etwas stärker in Richtung höherer erklärter Varianz, der Koeffizient für die Oben-Unten-Skala (V511) verdoppelt sich fast: Dadurch wird er auch hier zum signifikanten Prädiktor. Die beiden anderen Koeffizienten sind bemerkenswert stabil. Vergleicht man die Originaldaten des Subsets mit den Originaldaten des gesamten Datensatzes, so fällt auf, daß die Veränderungen durch die Subsetbildung mindestens genau so groß sind. Die Schätzungen innerhalb des Subsets unterscheiden sich nicht signifikant voneinander, die Subsetschätzungen unterscheiden sich aber von den Schätzungen des ALLBUS. So wird z. B. der Koeffizient für Nettoeinkommen (V435) halbiert, der damit im Subset auch nicht mehr signifikant von null verschieden ist.

Um dem naheliegenden Einwand zu begegnen, diese Ergebnisse wären lediglich auf den ohnehin sehr schlechten Fit des überprüften Modells zurückzuführen, wurde mit vollständig simulierten Daten ein weiteres 3-Variablen-Modell mit einer wesentlich höheren erklärten Varianz (91%) überprüft. Für die 4.8% „gefälschten" Daten wurden zwei Versionen mit je einer anderen Korrelationsstruktur berechnet: Ein reines Zufalls-Modell, bei dem alle Variablen unabhängig normalverteilt waren, und ein Zwei-Populationsmodell, bei dem in der zweiten Population ein anderes Modell galt. In beiden Fällen ergaben sich zwar hochsignifikante, „inhaltlich" aber eher bedeutungslose Differenzen: R^2 sank auf minimal 85%, die größte Veränderung eines Regressionskoeffizienten lag in der Reduktion von .90 auf .85. Der Fit des Modells war zwar deutlich schlechter, die „strukturellen" Koeffizienten veränderten sich aber kaum. Eine Wiederholung dieses Experiments mit einer Subgruppe von 20.2% Personen mit veränderter Kovarianzstruktur führte zu einer Reduktion auf minimal 67% erklärter Varianz, die größte Veränderung eines Regressionskoeffizienten lag in der Verminderung von .90 auf .72. Trotz einer Subgruppe von 20% mit einer anderen Kovarianzstruktur sind die Ergebnisse recht stabil.

Zusammenfassend: Bei einem größeren Anteil Fälschungen zeigt sich auch bei multivariaten Statistiken eine Zunahme der Verzerrungen. Obwohl sich dadurch hierbei unterschiedliche inhaltliche Interpretationen ergeben würden, können die Ergebnisse – gemessen an der relativen Unpräzision sozialwissenschaftlicher Theorien und Messungen (die sich zum Beispiel darin zeigt, daß die korrekte Vorhersage des Vorzeichens schon als Bestätigung der Theorie interpretiert wird) – als erstaunlich robust bezeichnet werden. Die größeren Differenzen mit den simulierten Daten legen allerdings den Schluß nahe, daß bei Variablen, für die falsche Laientheorien verwendet werden und einem hohen Anteil von Fälschungen die Resultate weitgehend unbrauchbar werden.

Schlußfolgerungen

Obwohl der Anteil der Fälschungen bei Surveyinterviews vermutlich klein ist, kann das Problem dennoch nicht ignoriert werden. So unbedeutend Interviewerfälschungen für univariate Statistiken wie Anteile, Mittelwerte oder Streuungen sein mögen, so fatal können sich statistisch selbst einzelne Fälschungen (als „Ausreißer") auf multivariate Analysen auswirken. Dies gilt insbesondere für

kleine Stichproben bzw. für die häufig durchgeführten Analysen sehr kleiner Subsets größerer Datensätze. Für univariate Statistiken mag die Hoffnung einiger Praktiker, daß eine Stichprobe von 2000 Befragten statistisch zu robust sei, „(...) um von einem Zwanzigstel unkorrekter Antworten verbogen werden zu können" (Kirschhofer-Bozenhardt/Kaplitza 1982: 133), noch einigermaßen begründet sein, wie dies auch die vorliegende Untersuchung zeigt. Bei anspruchsvolleren Datenanalysetechniken, die gerade für eine theoretisch orientierte Sozialwissenschaft unverzichtbar sind, gilt dies allerdings nicht mehr *mit mathematischer Sicherheit.* Bei multivariaten Analysen, die auf dem allgemeinen linearen Modell basieren (wie z. B. Faktoren- und Varianzanalysen, multiple Regressionen, Pfadanalysen), reichen prinzipiell wenige Fälle zur grundlegenden Veränderung der Ergebnisse vollständig aus. Die Verwendung multivariater Techniken setzt daher bei kleinen Stichproben und dem Verdacht von Interviewfälschungen vor der Analyse eine zusätzliche, ungewöhnlich umfangreiche und auf die Betrachtung einzelner abweichender Fälle orientierte Datenprüfung und Bereinigung („Data Screening") durch den Datenanalytiker voraus. Die Berücksichtigung mindestens der Interviewernummer und einiger Interviewermerkmale im Datensatz ist daher unverzichtbar.[28]

Die analytischen Ergebnisse zeigen für einfache Statistiken bei großen Fallzahlen und kleinen Anteilen von Fälschungen die Robustheit der Ergebnisse gegenüber Fälschungen. Die empirischen Ergebnisse der Erhebung dieser Studie und die Simulationsergebnisse legen dies auch für multivariate Statistiken nahe. Einzelne Belege, daß sich Fälschungen auch nicht auf die Ergebnisse multivariater Analysen auswirken, sind aber leider nicht in der Lage, das Argument der Verzerrung durch wenige Fälschungen vollständig zu entkräften. Prinzipiell sind solche Verzerrungen möglich, das Ausmaß hängt von den Gegebenheiten des speziellen Sachverhalts ab. Das Ausmaß der Verzerrung komplexer Statistiken durch Fälschungen kann – wie bei fehlenden Werten allgemein – nicht analytisch abgeschätzt werden, sondern muß im Einzelfall über eine Art „multiple Imputation" (Rubin 1987) beurteilt werden.

Literatur

Biemer, P. P./Stokes, S. L., 1989: The Optimal Design of Quality Control Samples to Detect Interviewer Cheating. Journal of Official Statistics, 5, 1: 23–39.

Buchhofer, B., 1979: Projekt und Interview. Hamburg: Beltz.

Case, P. B., 1971: How to Catch Interviewer Errors. Journal of Advertising Research, 11, 2: 39–43.

Chapman, D. W., 1983: The Impact of Substitution on Survey Estimates; in: Madow, W. G./Olkin, I./Rubin, D. B. (Hrsg.): Incomplete Data in Sample Surveys, Vol. 2, S. 45–61, New York.

Dawes, R. M., 1988: Rational Choice in an Uncertain World, San Diego: Harcourt Brace Jovanovich.

Hauck, M., 1969: Is Survey Postcard Verification Effective? Public Opinion Quarterly, 23: 117–120.

Hippler, H.-J., 1979: Untersuchung zur „Qualität" von absichtlich gefälschten Interviews, ZUMA-Arbeitspapier, Februar 1979.

Kalton, G., 1983: Compensating for Missing Survey Data, Ann Arbor: Institute for Social Research.

Kirschhofer-Bozenhardt, A. v./Kaplitza, G., 1982: Das Interviewernetz. S. 127–135 in: K. Holm (Hrsg.): Die Befragung, Band 1, 2. Auflage, München: Francke.

Platek, R./Gray, G. B., 1983: Imputation Methodology; in: Madow, W. G./Olkin, I./Rubin, D. B. (Hrsg.): Incomplete Data in Sample Surveys, Vol. 2, S. 255–333, New York.

Reuband, K.-H., 1990: Interviews, die keine sind – „Erfolge" und „Mißerfolge" beim Fälschen von Interviews. KZfSS 42: 706–733.

Rubin, D. B., 1987: Multiple Imputations for Nonresponse in Surveys, New York: Wiley.

Rummel, R. J., 1970: Applied Factor Analysis. Evanston: Northwestern University Press.

Santos, R. L., 1981: Effects of Imputation on Complex Statistics. Technical Report, Survey Research Center, University of Michigan.

Schnell, R., 1986: Missing-Data-Probleme in der empirischen Sozialforschung. Dissertation, Ruhr-Universität Bochum.

Schreiner, I./Pennie, K./Newbrough, J., 1988: Interviewer Falsification in Census Bureau Surveys. American Statistical Association Proceedings of the Section on Survey Research Methods, 491–496.

[28] Dies ist in der BRD unverantwortlicherweise leider ebenso wenig gängige Praxis wie die Erstellung eines umfangreichen Feldberichtes für jede Auftragsstudie. Die Ursache hierfür liegt u. a. auch in dem offensichtlichen und nicht zu rechtfertigendem Desinteresse der meisten Auftraggeber gegenüber der Herstellung ihrer „Ergebnisse".