Die Homogenität sozialer Kategorien als Voraussetzung für "Repräsentativität" und Gewichtungsverfahren

Schnell, Rainer

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
SSG Sozialwissenschaften, USB Köln

Empfohlene Zitierung / Suggested Citation:
Schnell, R. (1993). Die Homogenität sozialer Kategorien als Voraussetzung für "Repräsentativität" und Gewichtungsverfahren. *Zeitschrift für Soziologie*, *22*(1), 16-32. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-121791

Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität" und Gewichtungsverfahren[1]

Rainer Schnell

Fakultät für Soziologie, Universität Mannheim, Gebäude A 5, D-6800 Mannheim 1

Zusammenfassung: Die meisten Datensätze der empirischen Sozialforschung basieren auf Surveyinterviews. Das größte methodische Problem bei Surveyerhebungen sind Ausfälle (Nonresponse), insbesondere durch Verweigerungen. Zur Korrektur von Ausfällen werden in der Praxis Gewichtungsverfahren eingesetzt. Allen Gewichtungsverfahren liegen inhaltliche, soziologisch relevante Annahmen zugrunde. Diese Annahmen werden jedoch fast nie expliziert oder überprüft. Zentral für Gewichtungsverfahren ist die Annahme, daß die Gewichtungsklassen homogen sind. Auf ähnlichen Annahmen basieren auch Quota-Verfahren und sogenannte „Repräsentanzbeweise". Anhand empirischer Daten wird gezeigt, daß diese zentrale Annahme falsch ist. Weiterhin wird mit einer Simulation systematischer Ausfälle auf der Basis der Daten des ALLBUS 1980 demonstriert, daß Standardgewichtungsverfahren nicht in der Lage sind, systematische Ausfälle zu kompensieren.

1. Einleitung

Die meisten Untersuchungen der empirischen Sozialforschung basieren auf standardisierten Interviews im Rahmen von Zufallsstichproben. Zu den methodisch bedeutsamsten Problemen bei Zufallsstichproben aus der allgemeinen Bevölkerung gehört das Nonresponseproblem: Ein Teil der ausgewählten Stichprobe fällt durch Abwesenheit, Krankheit oder Verweigerung aus der Befragung aus.

Das Ausmaß des Nonresponse scheint in den letzten Jahren in den meisten Industrieländern zugenommen zu haben. In der BRD dürften bei den meisten Zufallsstichproben der kommerziellen Institute die Ausschöpfungsquoten derzeit eher zwischen 50 und 60 Prozent liegen als bei 70 Prozent (Hansen 1988: 399). Das Ausmaß des Nonresponse scheint viele Rezipienten empirischer Sozialforschung zu beunruhigen: So scheint es Wyss (1990: 71) notwendig, darauf hinzuweisen, daß eine 70%-Ausschöpfung als Gütekriterium „wissenschaftlich" nicht begründet werden kann. Kromrey (1990: 202) glaubt (fälschlicherweise), daß bei „Rücklaufquoten um 70 Prozent" „auch mit den ausgefeiltesten Konzepten der Inferenzstatistik keinerlei Aussage über die Genauigkeit der Ergebnisse mehr möglich" sei.[2] Dies hat einige Autoren dazu gebracht, die „Repräsentativitätsproblematik" qualitativer Methoden mit denen von Zufallsstichproben zu vergleichen (Lamnek 1989: 92) oder gar für Quota-Samples zu plädieren (Lamnek 1978: 568–569).[3]

Im allgemeinen Umgang mit dem Nonresponseproblem wird vor allem auf einfache Konzepte zurückgegriffen: „Repräsentativität" zeigt sich in der Übereinstimmung von Stichprobenhäufigkeiten mit den entsprechenden bekannten Anteilen in der Grundgesamtheit.[4] Wenn sich diese nicht ohnehin

[1] Den Anstoß für diese Arbeit gab eine Diskussion mit Bernhard von Rosenbladt anläßlich der Tagung der Methodensektion der DGS in Berlin im April 1991. Für kritische Anmerkungen danke ich Stefan Bender, Elke Esser, Frank Kalter, Thomas Klein und Johannes Kopp.

[2] Nur um Mißverständnisse zu vermeiden: Gerade bei solchen Ausschöpfungsquoten gibt es nur noch *eine* Möglichkeit einer begründbaren Aussage: „Multiple Imputation" (Rubin 1987), siehe dazu weiter unten.

[3] Dabei wird übersehen, daß Quota-Stichproben ebenfalls unter Nonresponse leiden: Bausch (1990: 77) schreibt explizit, daß der Vorteil des Quotenverfahrens darin besteht, daß es kein Nonresponseproblem besäße. Dies ist falsch. Das Nonresponseproblem bei Quota-Stichproben wird nur verdeckt: Da das Ausmaß des Nonresponse nicht ersichtlich ist, kann bei solchen „Stichproben" keinerlei „Korrektur" oder auch nur der geringste Ansatz einer Abschätzung der Verzerrung erfolgen: Man kann nur noch an die Unverzerrtheit der Ergebnisse glauben. Dies gehört nicht zu den üblicherweise akzeptierten Voraussetzungen der Anwendung einer Methode.

[4] Entgegen weitverbreiteten Ansichten ist „Repräsentativität" kein Begriff, der in der technischen und mathematischen Literatur zu Auswahlverfahren verwendet wird: In diesem Sinn handelt es sich um keinen „wissenschaftlichen" Begriff, vgl. Kruskall/Mosteller (1979a, 1979b, 1979c, 1980), zusammenfassend Schnell et al. (1992: 314–315).

zeigen läßt, dann werden die verzerrten Stichproben „gewichtet". Da die gewichteten Stichproben in einigen bekannten Merkmalen mit der Grundgesamtheit übereinstimmen, wird daraus auf die Unverzerrtheit anderer Merkmale geschlossen („Repräsentanzbeweise").[5] Ziel dieser Arbeit ist es, zu zeigen, auf welchen Annahmen Gewichtungsverfahren (und auch Quotenstichproben) basieren und empirisch zu demonstrieren, daß diese Annahmen falsch sind. Schließlich soll anhand simulierter Ausfälle gezeigt werden, daß Standardgewichtungsverfahren weder Ausfälle aufgrund von Schwererreichbarkeit noch aufgrund inhaltlicher Prozesse korrigieren können.

2. Zur Anwendung von Gewichtungsverfahren

Die geschilderten Ursachen führen dazu, daß die meisten verfügbaren Umfrage-Ergebnisse auf gewichteten Datensätzen basieren. Es wird – meist ohne explizite Begründung – gewohnheitsmäßig gewichtet:

Der von ZUMA in gemeinsamer Verantwortung mit GFM-GETAS dreimal jährlich durchgeführte „Sozialwissenschaften-Bus" wird standardmäßig „iterativ gewichtet" ausgeliefert, für den ALLBUS wurde bis einschließlich 1988 ein „ZUMA-Gewicht" berechnet.[6] Selbst das Statistische Bundesamt akzeptiert die Gewichtung von in seinem Auftrag erstellten Studien (Forsa 1989). Die größte regelmäßig in der BRD durchgeführte nichtamtliche Erhebung, die Media-Analyse, wird nach Alter, Geschlecht, Haushaltsgröße, Bundesland, Gemeindegrößenklasse und Befragungstag gewichtet.[7] Einige kommerzielle Studien werden der Media-Analyse durch Gewichtung angeglichen, wobei die durch die bereits gewichtete Media-Analyse gewonnene Mediennutzungswahrscheinlichkeit nochmals zur Gewichtung verwendet wird (Beispiel: Spiegel 1991: 189). Auch bei explizit für analytische Fragestellungen konzipierten Erhebungen wird gewohnheitsmäßig gewichtet, so z. B. bei der Deutschen Herz-Kreislauf-Präventionsstudie nach Geschlecht, Alter, Bundesland und Gemeindegrößenklasse (DHP 1989: 5–7).

In den Methodenberichten vieler Erhebungen werden Gewichtungsverfahren höchstens mit einem Satz abgehandelt. Die dabei verwendete Terminologie ist überaus uneinheitlich, identische Verfahren werden mit einer Vielzahl von Begriffen belegt.[8] Durch Benennung mit einem eindrucksvollen Label wird ein Konsens in Hinsicht auf die Anwendung von Gewichtungen unterstellt und jegliche Diskussion abgebrochen. Diese fehlende Thematisierung des Problems wird dadurch erleichtert, daß Gewichtungsverfahren bei vielen Sozialwissenschaftlern Unsicherheiten auslösen, was nicht zuletzt daran liegen dürfte, daß dieses Thema in kaum einem einführenden Lehrbuch der empirischen Sozialforschung oder Statistik thematisiert wird.[9]

Damit geschieht – wissenssoziologisch höchst interessant – etwas Eigentümliches: Die mathematische Literatur zu Gewichtungsverfahren verliert sich in technischen Details (z. B. Graef/Blien 1989, Gabler 1991), wobei die zentrale inhaltliche Annahme *aller* Gewichtungsverfahren[10] nicht mehr thematisiert wird; die meisten Sozialwissenschaftler akzeptieren aber die mechanische Anwendung von Gewichtungsverfahren als Teil der schwarzen

---

5 Dies ist ein Standardargument von Verfechtern der Quotenstichprobe: Zeigt sich die Übereinstimmung zwischen Stichprobe und Grundgesamtheit auch bei nichtquotierten Merkmalen, dann sei dies ein Beweis für die „Repräsentativität" der Stichprobe, vgl. z. B. Kaplitza (1982: 168) und Noelle-Neumann/Piel (1983: 223–231).

6 Technische Details einschließlich des verwendeten GAUSS-Programms finden sich bei Rothe/Wiedenbeck (1988), vgl. auch Rothe (1989).

7 Eine Rekonstruktion der verwendeten Gewichtung geben Rothe/Wiedenbeck (1987: 52–56).

8 Neben „iterativer Gewichtung" findet sich „Raking", „Iterated Proportional Fitting" (auch: „Deming-Stephan algorithm"), „Minimax-Gewichtung", „Entrop-Gewichtung", „Redressment", „Strukturangleichung", „Post-Stratification", „weighting-class correction", „Cell-Weighting", „Linear-Weighting", „Rim-Weighting", „Furness-Method", „Cross-Fratar procedure", „Mostellerising" usw. (die zuletzt genannten finden sich bei Upton 1987: 363). Zwar gibt es technische Unterschiede in der Durchführung (und den Zielkriterien) zwischen einzelnen Verfahren, die Logik der Verfahren ist aber ähnlich.

9 Sharot (1986: 269) beginnt seine elementare Einführung in die Gewichtungspraxis mit dem Hinweis, daß es keine „(. . .) single, reasonably comprehensive, introductury explanation of the process of weighting (. . .)" gibt. Dies änderte sich erst durch die Veröffentlichung von Elliot (1991).

10 Zwar kann man bei einigen Verfahren inhaltliche Theorien über den Ausfallprozeß berücksichtigen, dies wird aber weder demonstriert noch existieren „Standardmodelle" des Ausfallprozesses: Die sozialwissenschaftliche Literatur zur Erklärung von Ausfällen besteht (mit wenigen Ausnahmen, z. B. Esser 1986) fast ausschließlich aus der Auflistung demographischer Korrelate von Nonresponse.

Magie der Statistiker. Anders formuliert: Mathematiker interessieren sich nicht für die inhaltlichen Annahmen, Rezipienten empirischer Sozialforschung bleiben diese inhaltlichen Annahmen verborgen oder sie betrachten diese Annahmen fälschlich als korrekt.

3. Das zentrale Problem: Homogenität der Gewichtungsklassen

Nimmt man an, daß eine Person eine (für die gegebenen Surveybedingungen) konstante Wahrscheinlichkeit der Teilnahme besitzt, dann könnte man – falls man diese Wahrscheinlichkeit kennt – auf der Basis jeder Stichprobe einen interessierenden Parameter unverzerrt schätzen. Konstante Teilnahmewahrscheinlichkeiten sind aber unrealistisch. Man kann die Annahme konstanter Wahrscheinlichkeiten für eine Person abschwächen zur Annahme homogener Antwortwahrscheinlichkeiten in Subgruppen. Oh/Scheuren (1983: 157) bezeichnen die Bestimmung homogener Gruppen in Hinsicht auf die Responsewahrscheinlichkeiten als das zentrale Problem der Gewichtungsverfahren.[11] Diese sind aber ohne zusätzliche Annahmen nicht meß- oder schätzbar.[12] Als robustes Verfahren betrachten Oh/Scheuren die Verwendung von Gruppen, bei denen die Binnengruppenvarianz der Nonrespondenten gering und die Mittelwertdifferenzen zwischen den Gruppen bei den beobachteten Variablen groß sind. Leider sind nur die Binnengruppenvarianzen der Respondenten beobachtbar. Damit müssen folgende Annahmen gemacht werden:

Entweder

1. Es gibt eine beobachtbare Homogenität der Respondenten bei den beobachteten Variablen
2. Diese Homogenität gilt auch für die Nonrespondenten
3. Die Mittelwerte von Respondenten und Nonrespondenten unterscheiden sich

oder

4. Die Responsewahrscheinlichkeiten sind in den Gruppen homogen
5. Die Responsewahrscheinlichkeiten für alle Gruppen sind größer Null
6. Die Responsewahrscheinlichkeiten unterscheiden sich.

Die Annahme beobachtbarer Homogenität kann direkt empirisch getestet werden, die anderen Annahmen sind nur unter den methodischen Problemen von Nonresponse-Studien im allgemeinen indirekt prüfbar.[13]

Die empirische Überprüfung der Annahme beobachtbarer Homogenität wird in Abschnitt 4 erfolgen. Dabei wird sich zeigen, daß die üblichen Gewichtungsklassen in Hinsicht auf sozialwissenschaftliche Variablen nicht homogen sind. Damit ist eine Legitimation der Gewichtungsverfahren über die Homogenität der Gewichtungsklassen nicht möglich.[14]

Damit verbliebe noch das Argument der Homogenität der Responsewahrscheinlichkeiten innerhalb der Gewichtungsklassen. Da die Gewichtungsklassen stets durch die Kreuztabellierung demo-

---

[11] Eine detaillierte Darstellung und Kritik der Annahmen finden sich bei Schnell (1986: 97–106), eine Zusammenfassung bei Esser et al. (1989: 154–158).

[12] Dalenius (1983: 412) lehnt daher die Verwendung des Konzepts „response probabilities" allgemein ab.

[13] Nonresponse-Untersuchungen basieren
- auf dem Vergleich von Aggregatstatistiken mit Stichprobenergebnissen (dann sind nur wenige Merkmale überprüfbar, die Aggregatergebnisse müssen als unverzerrt angesehen werden)
- auf der Grundlage von ehemaligen Teilnehmern, die bei Paneluntersuchungen ausfallen (damit werden ehemalige Teilnehmer und keine Nonrespondenten untersucht)
- auf der Extrapolation auf der Grundlage der „Schwierigkeit" des Interviews (somit wird ein „Widerstandskontinuum" angenommen: eine mehrfach als falsch nachgewiesene irrige Annahme)
- auf Interviewerschätzungen (diese sind meist selbst unvollständig und sehr wenig zuverlässig, zudem sind nur wenige Merkmale verfügbar)
- auf den Angaben „konvertierter" Verweigerer (solange nicht *alle* konvertieren – und das geschieht nie –, werden nur leichter Befragbare untersucht).

Keine dieser Methoden zur Untersuchung von Nonresponsebias ist ohne Probleme, vgl. auch Smith (1983: 401). Daneben gibt es weitere Probleme (Datenschutz, Außendarstellung der Erhebungsinstitute, Organisation und Kontrolle der Feldarbeit etc.), die die praktische Durchführung von Nonresponsestudien beeinträchtigen.

[14] Da die (demographischen) Gewichtungsklassen nicht homogen in bezug auf die Responsewahrscheinlichkeiten sind und empirisch keine demographisch homogene Nonrespondentenkategorie nachgewiesen werden kann (vgl. die Diskussion der Annahmen 4 und 5) ist auch die Annahme 2 sehr unwahrscheinlich. Die Annahme 3 ist zwar gelegentlich empirisch korrekt (es gibt zumindest gelegentlich beobachtbare Unterschiede in Hinsicht auf einige Variablen zwischen Respondenten und Nichtrespondenten), allerdings variieren diese Differenzen kaum zwischen den Gewichtungsklassen. Damit sind 2 der drei notwendigen Annahmen fast sicher falsch, die dritte Annahme meistens.

graphischer Variablen gebildet werden, entspricht dies der Annahme, daß es einen starken Zusammenhang zwischen demographischen Variablen und Teilnahmeverhalten gibt. Dies ist aber nicht der Fall. Demographische Merkmale zeigen nur unregelmäßige Zusammenhänge mit dem Teilnahmeverhalten und insbesondere mit Verweigerungsverhalten.[15] Alle bisherigen empirischen Ergebnisse zum Zusammenhang von demographischen Variablen und Nonresponse belegen, daß von einer homogenen Kategorie der „Nonrespondenten“ oder auch einer „Gruppe der Verweigerer“ nicht gesprochen werden kann.[16] Weiterhin gibt es im Gegensatz zu der bei Statistikern sehr beliebten Annahme eines „harten Kerns“ der Nonrespondenten, der sich jedem Befragungsversuch entzieht, in der gesamten sozialwissenschaftlichen Literatur zu Nonresponse keine empirischen Hinweise auf die Existenz einer solchen Gruppe. Da sich sowohl Einflüsse der Interviewer, der Erhebungsorganisation und des Erhebungsgegenstandes zeigen lassen, kann es sich beim Teilnahmeverhalten nicht um ein unveränderliches „Persönlichkeitsmerkmal“ handeln, sondern offensichtlich um Entscheidungshandeln, das sowohl von den Situationsmerkmalen wie auch durch individuelle Präferenzen des Handelnden bedingt wird (vgl. Esser 1986). Die Nichtexistenz des „harten Kerns“ läßt die Annahme, daß in keiner Gruppe die Responsewahrscheinlichkeit Null ist, als vermutlich korrekt erscheinen. Anders gesagt: Irgendjemand aus einer Gruppe läßt sich immer befragen. Der problematische Teil liegt in der Annahme, daß die Responsewahrscheinlichkeiten in den demographischen Subgruppen (den Gewichtungsklassen) homogen sind. Da kaum vermutet werden kann, daß die Handlungskalküle der Befragten so stark typisiert sind, daß sich diese innerhalb demographischer Subgruppen nicht mehr unterscheiden, muß die Annahme homogener Responsewahrscheinlichkeiten zweifelhaft erscheinen. Die schwachen Korrelationen zwischen Teilnahmeverhalten und demographischen Variablen stützen tendenziell die Annahme unterschiedlicher Responsewahrscheinlichkeiten. Die Schwäche der Zusammenhänge unterminiert jedoch die Annahme homogener Responsewahrscheinlichkeiten. Zusammenfassend: Die notwendigen Homogenitätsannahmen für Gewichtungsverfahren sind weder in Hinsicht auf „inhaltliche“ Variablen noch in Hinsicht auf Responsewahrscheinlichkeiten erfüllt. Theoretisch ist damit eine Anwendung der Gewichtungsverfahren bei sozialwissenschaftlichen Fragestellungen nicht zu begründen.[17]

4. Die empirische Überprüfung der Homogenität von Gewichtungsklassen

Der (eher triviale) empirische Nachweis der Inhomogenität der Gewichtungsklassen kann mit jedem beliebigen Datensatz der empirischen Sozialforschung erfolgen. Für alle Analysen dieser Arbeit liefert der ALLBUS 1980 die Datengrundlage. Die Entscheidung für den ALLBUS 1980 basiert auf der außerordentlich exakten Dokumentation der Feldoperationen, die in diesem Umfang auch von der sogenannten „ALLBUS Nonresponse Studie 1986“ nicht mehr erreicht wurde. Der ALLBUS 1980 eignet sich daher für methodische Analysen in besonderem Maße. So sind z.B. die Kontaktversuche der Interviewer in der zusätzlichen Methodendatei zum ALLBUS 1980 enthalten.

Weiterhin enthält der Bericht von Kirschner (1984: 163) alle notwendigen Details der Stichprobengewichtung, einschließlich der Gewichtungstabelle, die vom erhebenden Institut (GETAS) verwendet wurde. Für die Gewichtung des ALLBUS 1980 wurden von GETAS (wie von vielen Erhebungsinstituten bei verschiedenen Studien) die Variablen Geschlecht, Alter und Bundesland als Gewichtungsvariablen verwendet. Die Logik der Gewichtung erfordert Homogenität der durch die Gewichtungsvariablen gebildeten Gewichtungsklassen in Hinsicht auf die interessierenden Variablen. Wählt man als Homogenitätsmaß Eta^2, so kann eine Abschätzung der Homogenität der Gewichtungsklassen durch die „Varianzerklärung“ der abhängigen

[15] Das gilt selbst für den häufig von Praktikern der empirischen Sozialforschung erwähnten Zusammenhang mit „Alter“, vgl. hierzu Schnell (1991: 120–121).

[16] Vgl. zusammenfassend Goyder (1987: 117), Groves (1989: 208), Schnell et al. (1992: 322).

[17] In der Regel interessieren sich Sozialwissenschaftler eher für Variablenzusammenhänge als für Punktschätzungen. Bemerkenswerterweise existiert in der statistischen Fachliteratur ein weitgehender Konsens, daß bei korrekter Modellierung eines interessierenden Zusammenhangs eine Gewichtung nicht notwendig ist (Hoem 1989). Diese korrekte Modellierung ist allerdings für kaum eine sozialwissenschaftliche Fragestellung zu gewährleisten (Alexander 1987: 188). Eine Gewichtung bei einer analytischen Fragestellung muß daher das explizite Ziel der Nonresponsebias-Reduktion durch die Gewichtung trotz einer Fehlspezifikation des interessierenden Zusammenhangs verfolgen.

Tabelle 1 Erklärte Varianz (eta²) aller metrischen Variablen des ALLBUS 1980 nach Gewichtungsvariable und Variablengruppe

Gewichtungsvariable	Variablengruppe	Minimum	Maximum	Mittelwert	Median
SEX	Einstellungen	0.000	0.074	0.005	0.002
	Demographie	0.000	0.214	0.038	0.012
	Fakten	0.000	0.058	0.008	0.002
	Interview	0.000	0.039	0.006	0.001
AGE	Einstellungen	0.000	0.134	0.027	0.013
	Demographie	0.003	0.800	0.160	0.071
	Fakten	0.001	0.106	0.022	0.013
	Interview	0.003	0.043	0.020	0.014
LAND	Einstellungen	0.003	0.073	0.012	0.010
	Demographie	0.003	0.054	0.015	0.011
	Fakten	0.002	0.055	0.011	0.008
	Interview	0.001	0.041	0.013	0.004
SEX* AGE* LAND	Einstellungen	0.032	0.194	0.068	0.054
	Demographie	0.039	0.813	0.257	0.243
	Fakten	0.030	0.160	0.075	0.062
	Interview	0.028	0.148	0.072	0.064

Variablen durch die Gewichtungsvariablen erfolgen.[18]

Man kann von den 374 Variablen des Datensatzes[19] des ALLBUS 1980 ca. 151 Variablen als „metrisch" betrachten. Hiervon kann man 70 Variablen als Einstellungsmessungen, 30 als demographische Variablen, 42 als Faktenfragen und 9 als erhebungstechnische Variablen („Interview") auffassen.[20] Für alle „metrischen" Variablen des ALLBUS 1980 wurde Eta² getrennt und gemeinsam für alle Gewichtungsvariablen berechnet.

Die Tabelle 1 zeigt das nach Variablengruppen getrennte Ergebnis dieser Berechnungen. Während die Variable „Bundesland" im Mittel weniger als 2% einer beliebigen Variablen erklärt, erklärt die Variable „Geschlecht" im Mittel zwar weniger als 1% der Varianz, allerdings bei den demographischen Variablen fast 4%. Weder „Geschlecht" noch „Bundesland" sind damit in irgendeinem Sinn erklärungskräftige Variablen – wie sollten sie auch?

Beide Variablen sind in Hinsicht auf die zumeist erhobenen „inhaltlichen" Variablen nur durch viele „indirekte Effekte" mit tatsächlichen handlungsrelevanten Anfangsbedingungen für das Handeln verknüpft. „Bundesland" und „Geschlecht" sind so schlechte Messungen handlungsrelevanter Variablen, daß sie als „erklärende Variable" vernachlässigbar sind: In sehr wenigen Ausnahmen werden maximal 7% Varianz erklärt.[21]

[18] Zu Eta² vgl. ausführlich Benninghaus (1990: 344–367).

[19] Die Zahl bezieht sich auf den PC-Datensatz des Zentralarchivs für empirische Sozialforschung/Köln mit der Nummer 1000. Die zugehörige „Methodendatei" enthält insgesamt 502 Variablen

[20] Als Faktenfrage wurden betrachtet: V62 bis V77, V105 bis V108, V150, V221, V315 bis V333 und V334; als demographische Variable: V4, V5, V7, V132, V134, V138, V142, V143, V146, V152, V154, V158, V163, V165, V180, V206, V209, V214, V217, V222, V223, V227, V230, V231, V235, V282, V290, V291, V293 und V295; als Einstellungsvariable V9 bis V25, V27, V28, V29 bis V37, V78 bis V83, V88, V89, V90 bis V104, V109, V114, V115 bis V119, V121 bis V131 und V161; als befragungstechnische Variable: V336 bis V344. Selbstverständlich sind die Kategorien unscharf und damit diskutabel.

[21] Selbstverständlich gibt es (vor allem bei demographischen Variablen) einige triviale Ausnahmen, z. B. die Zahl selbstgeborener Kinder etc.

Interessanter ist die Variable „Alter“. Zwar spielt „Alter“ in keiner soziologischen Theorie eine Rolle („age of the bones“ ist kein soziologisches Konstrukt), „Alter“ wird aber häufig als Proxy-Variable verwendet, so z.B. für „Lebenserfahrung“ oder „kognitive Inflexibilität“ sowie als Indikator für die Stellung im Lebenszyklus. Die Heterogenität der Konstrukte impliziert, daß auch hier keine perfekten Zusammenhänge erwartbar sind. Die Variable „Alter“ als Indikator für die Stellung in einem vergleichsweise stark typisierten Lebenszyklus sowie die damit verbundene Altershomophilie (Ältere sind zumeist mit Älteren verheiratet, daher besteht auch eine Korrelation zwischen Alter des Ehegatten und Ehedauer usw.) läßt die beobachtbaren relativ starken Zusammenhänge bei demographischen Variablen als trivial erscheinen. Aber selbst hier zeigt sich, daß in der überwiegenden Zahl der demographischen Variablen der größere Teil der Varianz nicht erklärt werden kann. Sobald man zu Faktenfragen oder gar Einstellungsfragen übergeht, verliert auch „Alter“ seine prädiktive Kraft: Der Median der erklärten Varianz liegt unter 1,5%.

Bei gleichzeitiger Berücksichtigung aller Gewichtungsvariablen lassen sich zwar in Einzelfällen bis zu 81% der Varianz erklären, dies sind aber tatsächliche „Ausreißer“, die durch die Altershomophilie und ähnliches erklärt werden. Der Median aller erklärten Varianzen liegt selbst bei den demographischen Variablen unter 25%, bei den Einstellungs- und Faktenfragen sogar unter ca. 6%. Darüber hinaus bleiben auch bei den am besten „erklärten“ Variablen (abgesehen von der Demographie) mehr als 80% der Varianz unerklärt: Damit kann von der Homogenität der Gewichtungsklassen in Hinsicht auf die erhobenen Variablen nicht gesprochen werden.

Zusammenfassend muß festgestellt werden, daß die „Erklärungskraft“ der demographischen Variablen Alter, Geschlecht und Bundesland für die meisten in sozialwissenschaftlichen Surveys erhobenen Variablen sehr gering ist. Nichts deutet darauf hin, daß dies für andere demographische Variablen als Gewichtungsvariablen anders wäre. Die Annahme der Homogenität von Gewichtungsklassen in Hinsicht auf interessierende Variablen ist somit falsch. Dies gilt dann auch für die Verwendung dieser Variablen als Quotierungsmerkmale. Trotz der Simplizität des Nachweises, daß die erklärten Varianzen gering sind, schreibt Kaplitza (1982: 161) über Quotierungsmerkmale: „Für die drei Kontrollmerkmale Geschlecht, Alter und soziale Schicht ist dieser Zusammenhang [zum Untersuchungsgegenstand, R.S.] fast immer gegeben, denn sie bewirken für die Antworten auf sehr viele Fragen den größten Anteil der Variation.“ Wie für Alter und Geschlecht gezeigt wurde, ist dies schlicht unzutreffend: Der größte Anteil der Variation bleibt ungeklärt.

5. Die empirische Leistungsfähigkeit von Gewichtungsverfahren

Wie gezeigt wurde, sind die zentralen Homogenitätsannahmen für Gewichtungsverfahren entweder empirisch falsch (soweit sie direkt prüfbar sind) oder zumindest vermutlich falsch (soweit sie indirekt prüfbar sind). Gerade aufgrund der nur indirekten Prüfbarkeit könnte noch ein Rest von Unsicherheit über einen vermeintlichen Nutzen von Gewichtungsverfahren zur Korrektur von Ausfällen verbleiben. Daher erschien eine empirische Untersuchung der Leistungsfähigkeit der Gewichtungsverfahren bei systematischen Ausfällen in Hinsicht auf die Korrektur von Variablenzusammenhängen notwendig.

5.1 Verwendete Daten und inhaltliche Modelle

Wie bereits erwähnt, wurde der ALLBUS 1980 als Datengrundlage ausgewählt. Der ALLBUS 1980 enthält die Daten von 2955 Fällen, bei denen in drei Fällen das Alter des Befragten nicht erhoben wurde. Da die Variable „Alter“ bei der Gewichtung verwendet wurde, mußten diese drei Fälle aus der Analyse ausgeschlossen werden. Die Analysedatei enthält somit 2952 Fälle. Für die Analyse wurden drei Themen ausgewählt:

Feick/Mayntz (1982: 421) untersuchen u.a. den Zusammenhang zwischen Alter und der wahrgenommenen Beschwerdemöglichkeit gegenüber Ämtern und Behörden.[22] Sie finden anhand der ALLBUS-Daten einen Effekt des Alters in Abhängigkeit des Bildungsniveaus: Bei den weniger Gebildeten spielt Alter keine Rolle in Hinsicht auf die wahrgenommenen Beschwerdemöglichkeiten, wohl aber bei den höher Gebildeten: Je jünger die Hochgebildeten sind, desto weniger werden Beschwerdemöglichkeiten gesehen. Feick und

[22] V80 (Frage 17C): Frage: „Der Bürger hat viele Möglichkeiten, sich gegen Entscheidungen von Ämtern und Behörden zu wehren“; Zustimmung oder Ablehnung mit 7stufiger Skala von „stimme überhaupt nicht zu“ bis „stimme voll und ganz zu“.

Mayntz erklären dies mit unterschiedlichen Erwartungshorizonten und Wertemustern, vor allem mit den „postmaterialistischen Zielen" dieser Gruppe (Feick/Mayntz 1982: 422). Feick/Mayntz stützen ihre Analyse auf die Ergebnisse einer dreidimensionalen Kreuztabellierung (BILDUNG*ALTER*BESCHWERDE), wobei sie für jedes Bildungsniveau den Gamma-Koeffizienten für ALTER*BESCHWERDE berechnen. Sie berichten lediglich den Koeffizienten für Abiturienten (.40) und die Tatsache, daß die Gamma-Koeffizienten mit abnehmender Bildung kleiner werden.

Als zweites Thema wurde ein „Standardmodell" der empirischen Sozialforschung (wie es schon bei Feick/Mayntz anklingt) gewählt: „Postmaterialismus" in Abhängigkeit von Alter und Bildung. Hierzu wurde aus den Antworten auf die Frage nach den wichtigsten politischen Zielen (V110–V112) nach den Angaben des ZUMA-Skalenhandbuchs (Allmendiger et al. 1983: B02) der Postmaterialismusindex berechnet. Dieser wurde – wie stets mit mäßigem Erfolg – in einer multiplen Regression durch „Alter" und „Bildung" „erklärt".

Als drittes Thema wurde eine Arbeit von Diekmann (1984) zur Einkommensdiskriminierung von Frauen im Angestelltenverhältnis ausgewählt. Diekmanns Modell basiert auf dem bekannten Blau/Duncan-Modell, das modifiziert für eine Teilmenge der Befragten des ALLBUS80 herangezogen wird.[23] Diekmann erklärt das Einkommen bei Angestellten für Männer und Frauen getrennt durch Bildung, Berufsprestige, Berufserfahrung (= Alter), Bildung des Vaters, Berufsprestige des Vaters und vier Leistungskategorien (einfache Tätigkeiten, schwierige selbständige Tätigkeit nach allgemeiner Anleitung, selbständige Leistung in verantwortungsvoller Tätigkeit, Führungsaufgaben), die als Dummy-Variablen in eine multiple Regression eingehen.[24]

5.2 Verwendete Ausfallmechanismen und Untersuchungsdatensätze

Um den Einfluß von Ausfällen und Gewichtungen zu simulieren, wurden die den drei inhaltlichen Modellen (Feick/Mayntz, Diekmann, „Postmaterialismus") zugrundeliegenden Analysen mit durch die Ausfallmechanismen veränderten Datensätzen gerechnet, wobei die Ergebnisse der drei inhaltlichen Modelle jeweils mit den Ergebnissen des Ausgangsdatensatzes (ALLBUS 1980) verglichen wurden.

Damit kann sich eine Überlagerung zweier Effekte ergeben: ein Gewichtungseffekt, der sich schon im Ausgangsdatensatz zeigt (ohne zusätzliche Ausfälle) und ein Gewichtungseffekt in den anderen Datensätzen, der durch die zusätzlichen Ausfälle hervorgerufen wird. In fast allen Fällen sind die Unterschiede zwischen gewichteten und ungewichteten Ergebnissen im Ausgangsdatensatz im Vergleich zu den Effekten zusätzlicher Ausfälle sehr gering; dies kann im Einzelfall jeweils den Plots entnommen werden. Die Effekte der Ausfälle auf die Gewichtung lassen sich durch eine Analyse der Veränderung der gewichteten Ergebnisse in Abhängigkeit von der Höhe der Ausfälle im Vergleich zum gewichteten ALLBUS ohne Ausfälle bestimmen.

Es wurden zwei verschiedene Ausfallmechanismen verwendet: Ein Mechanismus simuliert die variierende Erreichbarkeit der Befragten, der andere Mechanismus die möglicherweise systematisch variierende Kooperationsbereitschaft der Befragten.

Die variierende Erreichbarkeit wurde dadurch simuliert, daß aus dem Ausgangsdatensatz (n = 2952) sukzessive diejenigen Befragten gelöscht wurden, die erst beim sechsten, fünften, vierten, dritten und zweiten Kontaktversuch des Interviewers erreicht wurden.[25] Die letzte Stufe dieser Variante besteht also nur aus denjenigen Befragten, die beim ersten Kontaktversuch des Interviewers erreicht wurden und sich befragen lie-

[23] Diekmann (1984: 349) schließt diejenigen Personen aus, die nach Ansicht des Interviewers unzuverlässige Antworten gaben (V343 = 3). Weiterhin werden nur ganztags Beschäftigte berücksichtigt (V136 = 1). Die von Diekmann genannten Fallzahlen konnten bei den Männern nicht exakt reproduziert werden, die Abweichung beträgt aber nur 3 Fälle (Diekmann rechnet mit 284 bzw. 161 Fällen, diese Analysen basieren demgegenüber auf 287 Männern und ebenfalls 161 Frauen). Durch „Listwise"-Löschung der Fälle mit fehlenden Werten sind dies bereits 73%-Stichproben der inhaltlich definierten Subgruppe.

[24] Trotz der z.T. detaillierten Angaben bei Diekmann konnten nicht alle Variablen fehlerfrei reproduziert werden. Insbesondere gilt dies für die beiden Bildungsvariablen, die bei Diekmann offensichtlich in Jahren codiert in die Regressionen eingehen. Da Diekmann aber nicht angibt, wie er die ausschließlich erfragten Bildungsabschlüsse in Jahre umcodiert, wurde mit den nicht umcodierten Abschlüssen gerechnet. Dies wirkt sich zwar selbstverständlich auf die Regressionskoeffizienten aus, allerdings auf die standardisierten Regressionskoeffizienten in kaum merklichem Ausmaß.

[25] Die Erreichbarkeit wurde aus den Variablen der Kontaktprotokolle (V440–V469) konstruiert.

ßen. Da man als erste Annäherung an die tatsächliche Ziehung von Quota-Stichproben von der Hypothese ausgehen kann, daß Quota-Stichproben vermutlich eher aus leicht erreichbaren Zielgruppen rekrutiert werden,[26] können Effekte variierender Erreichbarkeit bei unverzerrter Randverteilung demographischer Variablen vermutlich auch auf Quota-Stichproben verallgemeinert werden.

Obwohl Erreichbarkeit z. B. mit demographischen Variablen zusammenhängt (vgl. z. B. Hoag 1981: 11), werden Ausfälle aufgrund von Schwererreichbarkeit zumindest in der Praxis echter Zufallsstichproben häufig als „unsystematische Ausfälle" behandelt.[27] Neben diesem „unsystematischen" Ausfall über Erreichbarkeit wurde noch ein weiterer, explizit systematischer Ausfallmechanismus simuliert.

Von „systematischen Ausfällen" spricht man dann, wenn zwischen den Ursachen für den Ausfall und den interessierenden inhaltlichen Variablen ein Zusammenhang besteht. Da die systematische Variation der Kooperationsbereitschaft immer vom Untersuchungsthema und der Zielperson abhängen kann, muß für eine Untersuchung systematischer Ausfälle stets ein jeweils spezieller Ausfallmechanismus konstruiert werden.[28]

Da „Postmaterialismus" bei zwei inhaltlichen Modellen eine Rolle spielt, wurde eine mit Postmaterialismus korrelierte Variable gesucht, die für einen systematischen Ausfallmechanismus verwendbar war.

Der ALLBUS enthält u.a. die 11-stufigen-„Thermometer"-Fragen zur Parteisympathie, wobei auch nach der Sympathie gegenüber den „Grünen" (V127) gefragt wurde. Da Postmaterialismus im ALLBUS80 mit Sympathie gegenüber den Grünen mit $r = .22$ korreliert ist, wurde die Sympathie gegenüber den Grünen als Indikator für eine simulierte Teilnahmebereitschaft gewählt. Falls also die Kooperationsbereitschaft bei einer Untersuchung von der Sympathie mit den Grünen abhängen würde, dann wäre der simulierte Ausfallmechanismus realistisch: Je näher man den Grünen steht, desto eher verweigert man die Teilnahme an sozialwissenschaftlichen Untersuchungen.

Der zweite Ausfallmechanismus basiert auf der Annahme dieser Hypothese. Dies wurde dadurch umgesetzt, daß aus den Analysen sukzessive diejenigen Befragten ausgeschlossen wurden, die zunehmend Sympathie gegenüber den Grünen bekundeten: Auf der letzten Stufe dieser Variante besteht der Datensatz nur noch aus denjenigen Befragten, die den Grünen völlig bis leicht ablehnend[29] gegenüberstehen.

Die beiden Ausfallmechanismen sind weitgehend unabhängig voneinander: Die Zahl der Kontakte bis zum Interview korreliert im ALLBUS80 mit Sympathie für die Grünen zwar „hochsignifikant" ($p < 0.001$) mit $r = .07$, dies entspricht aber eben nur ca. 0,5% gemeinsamer Varianz.

Mit jedem dieser Datensätze wurde jedes der drei inhaltlichen Modelle gerechnet, wobei jeder Datensatz einmal ungewichtet und zum anderen nach Alter, Geschlecht und Bundesland kombiniert gewichtet gerechnet wurde.[30] Damit sind in jedem der gewichteten Datensätze die (kombinierten) Variablen Alter, Geschlecht und Bundesland in den tatsächlichen Bevölkerungsanteilen der jeweiligen Subgruppen vorhanden. Alle gewichteten

[26] Zusätzlich können sich Interviewer in Quota-Stichproben an ihren Netzwerkbeziehungen orientieren und so indirekt ein Schneeball-Sample realisieren, vgl. hierzu Hoag (1986).

[27] In diesem Sinne gibt Landgrebe (1992: 22) die Empfehlung, daß die Ausschöpfung nur noch gesteigert werden solle, solange durch verstärkte Bemühungen während der Feldzeit eine Veränderung inhaltlicher Merkmale festgestellt werden kann. Diese Empfehlung basiert auf der nicht explizierten Annahme, daß ein monotoner Zusammenhang zwischen inhaltlichen Merkmalen und Erreichbarkeit besteht. Da Mischverteilungen und Schwellenwertprozesse nicht ausgeschlossen werden können, ist das vorgeschlagene Vorgehen nicht als allgemeines Verfahren tauglich.

[28] Dies engt zwangsläufig die Interpretation der Ergebnisse ein: Weder folgt aus der Demonstration einer Verzerrung durch systematische Ausfälle, daß Ergebnisse immer durch diesen Mechanismus verzerrt werden noch folgt aus der Demonstration der Unverzerrtheit die generelle Unverwundbarkeit der Ergebnisse: Bei jeder Untersuchung können andere systematische Mechanismen wirken.

[29] –5 bis –1 bei den Antwortvorgaben; dies entspricht im Datensatz des ALLBUS80 den Ausprägungen 1–5 der Variablen V127.

[30] Die Gewichte wurden als einfaches Soll/Ist-Gewicht auf Personenebene für jeden Datensatz neu berechnet, wobei die GETAS-Soll-Tabelle des ALLBUS 80 (Kirschner 1984: 163) als Soll verwendet wurde. Es wurde also *nicht* lediglich eine bereits im Datensatz vorhandene Gewichtungsvariable verwendet, sondern jede simulierte Stichprobe völlig neu gewichtet. Hierzu wurden die Gewichtungsvariablen für jede Stichprobe ausgezählt, die Angaben der Gewichtungstabelle durch die Resultate dividiert und das Ergebnis als neues Gewicht dem Datensatz zugeschrieben.

Tabelle 2 Datensatzbeschreibung.

Nummer	Sympathie Grüne	Kontakt	% des Samples
1	11	6	100
2		5	99
3		4	98
4		3	92
5		2	75
6		1	40
7	10		95
8	9		93
9	8		89
10	7		82
11	6		73
12	5		51
13	4		45
14	3		38

Nummer: Nummer des Datensatzes
Sympathie Grüne: Maximale Sympathie für Grüne im Datensatz
Kontakt: Maximale Zahl der Kontakte bis zum Interview im Datensatz
% des Samples: Fallzahl im Datensatz als Prozent des Ausgangsdatensatzes

Stichproben sind im oben erläuterten naiven Sinn damit ausnahmslos „repräsentativ". Neben dem vollständigen Datensatz gibt es entsprechend 5 Datensätze mit unterschiedlicher „Erreichbarkeit" und 8 Datensätze mit unterschiedlicher Kooperationsbereitschaft; also 14 verschiedene Datensätze (vgl. Tabelle 2). Da jeder dieser Datensätze sowohl gewichtet als auch nichtgewichtet analysiert wurde, liegen den Analysen insgesamt 2*14 Datensätze zugrunde.

5.3 Univariate Ergebnisse

Zunächst sollen die univariaten Ergebnisse in Hinsicht auf einige für die inhaltlichen Modelle zentrale Variablen kurz vorgestellt werden.

Die Abbildung 1 zeigt den Anteil der Personen, die keine Beschwerdemöglichkeiten gegenüber Ämtern und Behörden sehen in Abhängigkeit vom Ausmaß der Ausfälle. Die Ausfälle erfolgen hier systematisch, d.h. je näher die Befragten den Grünen stehen, um so eher fällt die Zielperson aus der Stichprobe heraus.

Der Anteil der Personen, die keine Beschwerdemöglichkeit sehen, sinkt mit dem Anwachsen der Ausfälle: Fällt z. B. dasjenige Viertel der Befragten aus der Stichprobe, das den Grünen am nächsten steht, so ist die Abweichung der resultierenden Stichprobe von der Ausgangsstichprobe nach den üblichen Kriterien schon „signifikant" (die gestrichelte horizontale Linie stellt die untere Grenze des 95% Konfidenzintervalls dar). Bei der Verringerung der Stichprobe auf die Hälfte erreichen sowohl die gewichteten als auch die ungewichteten Schätzungen Abweichungen von mehr als 3% vom wahren Wert: Beide Abweichungen würden als „hochsignifikant" interpretiert werden. Die Gewichtung korrigiert also den systematischen Ausfall nicht.

Bei dem entsprechenden Ausfallmechanismus über die Kontaktzahl verändert sich die Schätzung des Prozentsatzes „keine Beschwerdemöglichkeit" hingegen kaum.

Die Abbildung 2 zeigt die Veränderung der Schätzung des Anteils ganztägig Berufstätiger durch einen systematischen Ausfallmechanismus. Die ungewichtete Schätzung weicht bei einem Ausfall von ca. 1/4 der Stichprobe „signifikant" vom „wahren" Wert nach unten ab. Die gewichtete Schätzung bleibt vergleichsweise stabil.

Auch hier verändert sich bei dem entsprechenden Ausfallmechanismus auf der Basis der Erreichbarkeit die Schätzung des Prozentsatzes „ganztags berufstätig" kaum.

Die Abbildung 3 zeigt die Veränderung des Postmaterialismusindex in Abhängigkeit von steigenden systematischen Ausfällen. Schon bei nur 11% systematischen Ausfällen (aufgrund von hoher Sympathie für die Grünen) weichen die ungewichteten Schätzungen „signifikant" vom „wahren Wert" ab, bei ca. 1/4 Ausfall der Gesamtstichprobe weichen auch die gewichteten Schätzungen ab. Auch hier kann die Gewichtung die Verzerrung nicht korrigieren.

Wie die Abbildung 4 zeigt, gilt dies in diesem Beispiel selbst für den angenommenen Ausfallmechanismus über die Schwererreichbarkeit. Hätte man sich beim ALLBUS80 auf diejenigen Fälle beschränkt, die durch *einen* Kontaktversuch hätten erreicht werden können, so wären sowohl die gewichtete als auch die ungewichtete Schätzung des Postmaterialismusindex signifikant vom „wahren Wert" verschieden. Da eine Stichprobe, die lediglich die leicht erreichbaren Zielpersonen umfaßt,

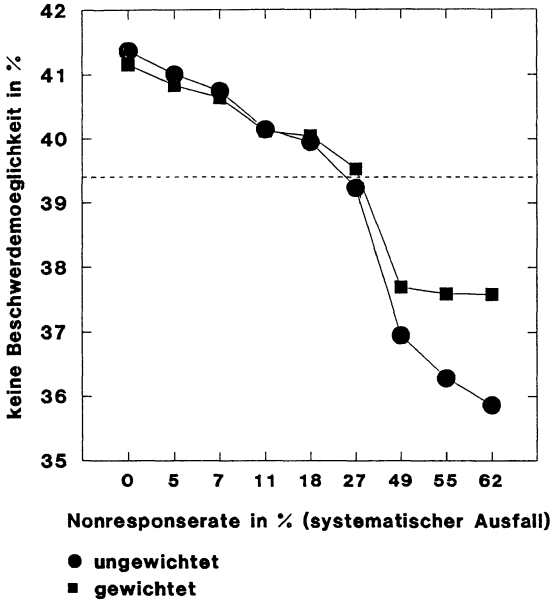

Abbildung 1 Anteil der Personen, die keine Beschwerdemöglichkeiten gegenüber Ämtern und Behörden sehen in Abhängigkeit vom Ausmaß der Ausfälle.

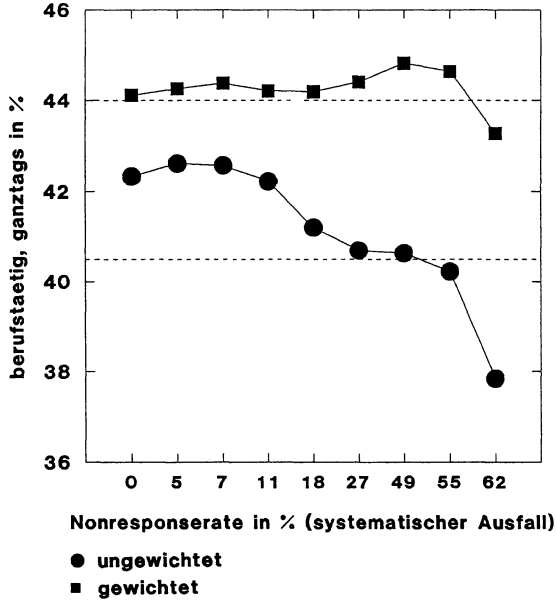

Abbildung 2 Veränderung des Anteils ganztägig Berufstätiger durch einen systematischen Ausfallmechanismus.

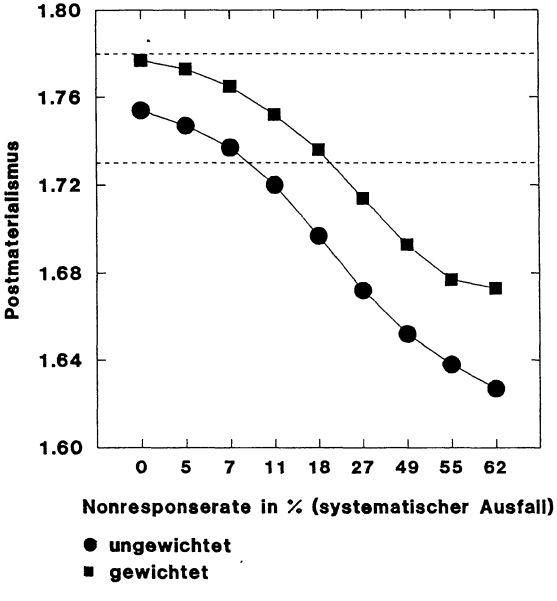

Abbildung 3 Veränderung des Postmaterialusmusindex in Abhängigkeit von steigenden systematischen Ausfällen.

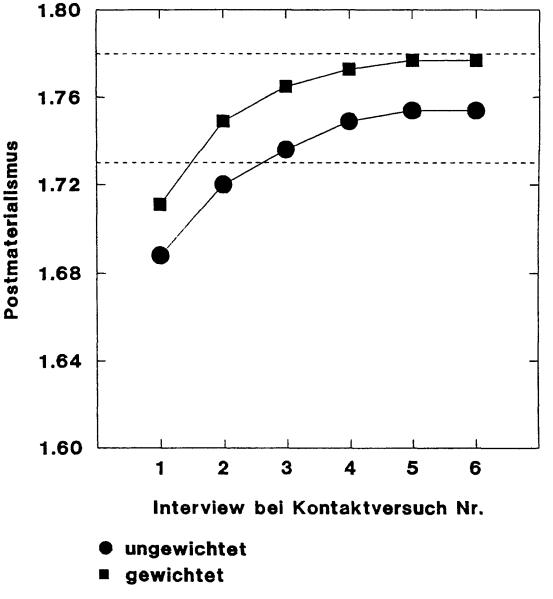

Abbildung 4 Veränderung des Postmaterialismusindex in Abhängigkeit von abnehmender Erreichbarkeit.

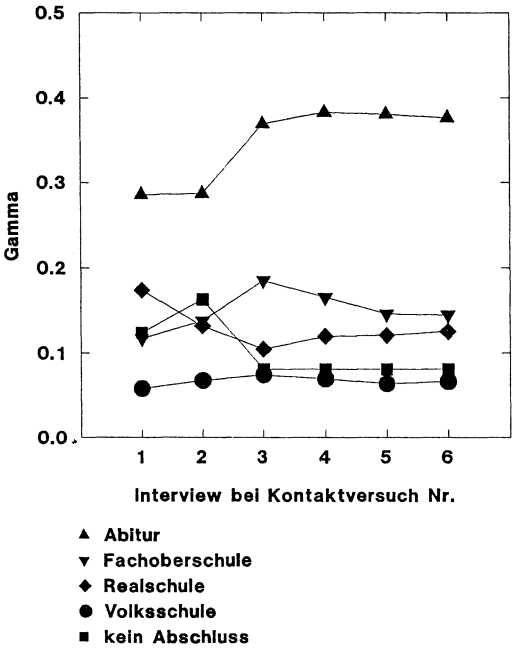

Abbildung 5 Veränderung der Gamma-Koeffizienten des Zusammenhangs zwischen der Einschätzung von Beschwerdemöglichkeiten gegenüber Ämtern und der Altersgruppe des Befragten, getrennt nach Bildungsgruppen (ungewichtet).

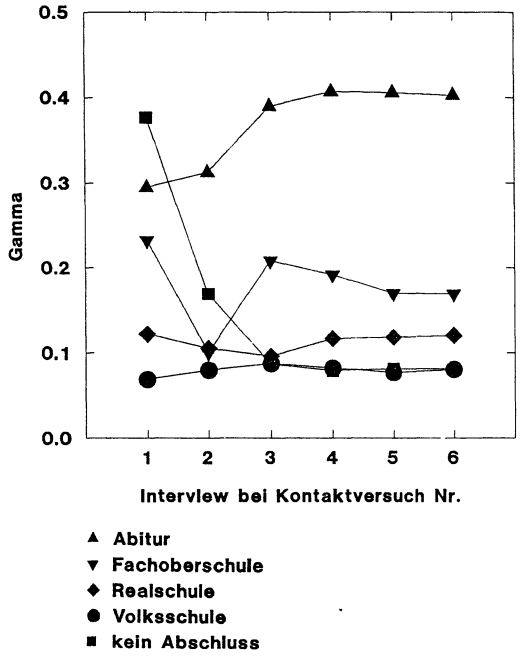

Abbildung 6 Veränderung der Gamma-Koeffizienten des Zusammenhangs zwischen der Einschätzung von Beschwerdemöglichkeiten gegenüber Ämtern und der Altersgruppe des Befragten, getrennt nach Bildungsgruppen (gewichtet).

vermutlich einer Quota-Stichprobe entspricht, ist dies ein weiteres Beispiel für die Unzulänglichkeiten von Quota-Stichproben: Selbst wenn die soziodemographischen Merkmale (entweder über den Quotenplan oder wie hier über Gewichtung) unverzerrt sein sollten, so folgt daraus nichts über die Unverzerrtheit anderer Merkmale (solange die Gewichtungsklassen nicht vollständig homogen sind). In Abschnitt 4 wurde empirisch gezeigt, daß diese Homogenitätsannahme für fast alle sozialwissenschaftlich interessanten Variablen nicht erfüllt ist. Wie dieses Ergebnis zeigt, gilt dies nachweislich z. B. für den Postmaterialismusindex.[31]

5.4 Multivariate Ergebnisse

Die Abbildung 5 zeigt die Veränderung der Gamma-Koeffizienten des Zusammenhangs zwischen der Einschätzung von Beschwerdemöglichkeiten gegenüber Ämtern und der Altersgruppe des Befragten. Feick/Mayntz (1982: 421–422) erwähnen, daß Altersunterschiede in den unteren Bildungsschichten in der hier interessierenden Hinsicht von nur geringer Bedeutung seien, hingegen bei Real- und Fachoberschulabschluß deutlich und bei Abiturienten besonders wirksam hervortreten. Dies wird von den Autoren durch Bezug auf postmaterialistische Ziele erklärt. Abbildung 5 zeigt, daß für die ungewichteten Daten diese berichtete Tendenz unabhängig von der Kontaktzahl stets nachweisbar ist, wobei allerdings die niedrigste Bildungsgruppe bei nur einem oder zwei Kontaktversuchen diese Regelmäßigkeit durchbricht.[32]

Die Abbildung 6 zeigt die gleichen Zusammenhänge wie Abbildung 5, nur diesmal anhand der gewichteten Daten. Die Gewichtung verändert die Abfolge der Bildungsgruppen jenseits von 2 Kontaktversuchen nicht mehr, wohl aber führt die Ge-

[31] Auf das Problem der Verzerrung durch Quota-Stichproben bei der Untersuchung von Postmaterialismus haben in anderem Zusammenhang schon Böltken/Gehring (1984) hingewiesen.

[32] Die asymptotischen Standardschätzfehler der Gamma-Koeffizienten sind in den meisten der hier berichteten Fälle so groß, daß im Gegensatz zu Feick/Mayntz die Unterschiede zwischen den Gruppen nicht inhaltlich interpretierbar erscheinen.

wichtung zu einer deutlicheren „Bestätigung" der Hypothese von Feick/Mayntz. Bei weniger als zwei Kontaktversuchen führt die Gewichtung zu einer drastischen Veränderung der Gamma-Koeffizienten in der niedrigsten Bildungsgruppe: Gamma wird von .12 (ungewichtet) auf .38 „hochgewichtet": Damit wird dies zum stärksten „beobachteten" Zusammenhang. Die Hypothese von Feick/Mayntz wäre bei Beschränkung auf einen Kontaktversuch und Gewichtung nicht mehr akzeptiert worden.

Das Regressionsmodell zur Vorhersage des Postmaterialismus erklärt im Ausgangsdatensatz insgesamt 16,5% der Varianz, wobei die Prädiktoren Bildung (β: .26) und Alter (β: –.26) jeweils „hochsignifikante" Prädiktoren darstellen. Die erklärte Varianz verändert sich durch die Reduzierung der Anzahl der Kontaktversuche weder in den gewichteten noch in den ungewichteten Datensätzen. Bei der Annahme eines systematischen Ausfalls reduzieren sich die erklärten Varianzen (mit zunehmenden Ausfällen fast linear) geringfügig auf ca. 10%. Zwischen gewichteten und ungewichteten Datensätzen besteht hier kein Unterschied. Durch die Reduzierung der Kontaktversuche verändern sich die Regressionskoeffizienten für Bildung und Alter praktisch kaum, auch hier spielt die Gewichtung keine Rolle. Bei Annahme eines systematischen Ausfalls verändert sich der Regressionskoeffizient für Bildung geringfügig (maximal von b = .18 auf b = .16), die Schätzungen weichen aber in keinem Fall signifikant vom Ausgangsdatensatz ab. Betrachtet man hingegen den Regressionskoeffizienten für Alter, so zeigt die Abbildung 7 eine deutliche Veränderung: Mit steigenden Ausfällen sinkt der Koeffizient von $\beta = -.26$ auf maximal –.19. Aufgrund des Standardfehlers für den Regressionskoeffizienten für Alter (0.00123 bei b = 0.0101 im Ausgangsdatensatz) weichen alle Schätzungen mit mehr als 18% systematischen Ausfällen signifikant vom Ausgangsdatensatz ab.[33] Die Bedeutung der Variablen „Alter" wäre bei systematischen Ausfällen „signifikant" als geringer (und damit auch geringer als die der Bildungsvariablen) betrachtet worden.

Die erklärte Varianz im Diekmann-Modell verändert sich bei der Reduzierung der Kontaktzahl kaum wesentlich, lediglich bei den Frauen ist eine kleine Steigerung der erklärten Varianz von .46 auf .56 feststellbar (sowohl mit als auch ohne Gewichtung).

Bei Annahme systematischer Ausfälle zeigt sich bei den Männern eine Abnahme der erklärten Varianz, wobei maximal eine Reduktion von 46% auf 36% erfolgt (vgl. Abbildung 8). Bei den Frauen ist der Einfluß systematischer Ausfälle auf die erklärte Varianz des Modells stärker: Fällt mehr als ein Viertel der Stichprobe systematisch aus, so weichen die Schätzungen der gewichteten Stichproben deutlich von der Ausgangsstichprobe ab: Die erklärten Varianzen steigen von .46 auf bis zu .66 an. Bemerkenswert ist die geringere Überschätzung der erklärten Varianz bei den ungewichteten Stichproben der Frauen bei extremen Ausfallraten.

Das interessanteste Ergebnis ist allerdings die offensichtliche Produktion eines Artefakts durch Gewichtung: Ohne Ausfälle erklärt das Regressionsmodell bei Männern und Frauen fast exakt den gleichen Anteil an Varianz. Durch systematische Ausfälle wächst die Differenz in der erklärten Varianz auf maximal .16 bei den ungewichteten Daten[34] (27% Ausfälle) und 0.27 bei den gewichteten Daten (62% Ausfälle).[35] Zumeist werden aber schon Unterschiede von mehr als 10% erklärter Varianz inhaltlich interpretiert. Die Unterschiede in der Leistung des Modells wären aber allein durch die Gewichtung produziert worden.

Betrachtet man die inhaltlichen Aussagen des Diekmann-Modells, so ist vor allem der unterschiedliche Einfluß der Bildung auf das Einkommen bei Männern und Frauen von zentraler Bedeutung (Diekmann 1984: 342). Untersucht man den Einfluß der Reduzierung der Zahl der Kontakte auf dieses Ergebnis, so muß die außerordentliche Stabilität dieses Resultats festgestellt werden. Lediglich bei nur einem Kontaktversuch verändern sich die Koeffizienten deutlich (im Sinne eines stärkeren Unterschieds zwischen Männern und Frauen), allerdings sind die Effekte nicht signifikant.

Die Abbildung 9 zeigt die Veränderung des Regressionskoeffizienten für Bildung in Hinsicht auf Einkommen im Diekmann-Modell, getrennt für Männer und Frauen, in Abhängigkeit eines systematischen Ausfallmechanismus, wobei gewichtete und nichtgewichtete Ergebnisse getrennt ausge-

[33] Ebenfalls ab 18% systematischer Ausfälle weichen alle geschätzten Kovarianzmatrizen der drei Variablen signifikant von der Ausgangskovarianzmatrix ab.

[34] Die Differenz liegt in diesem Fall bei den gewichteten Daten bei 0.17.

[35] Die Differenz liegt in diesem Fall bei den ungewichteten Daten bei 0.08.

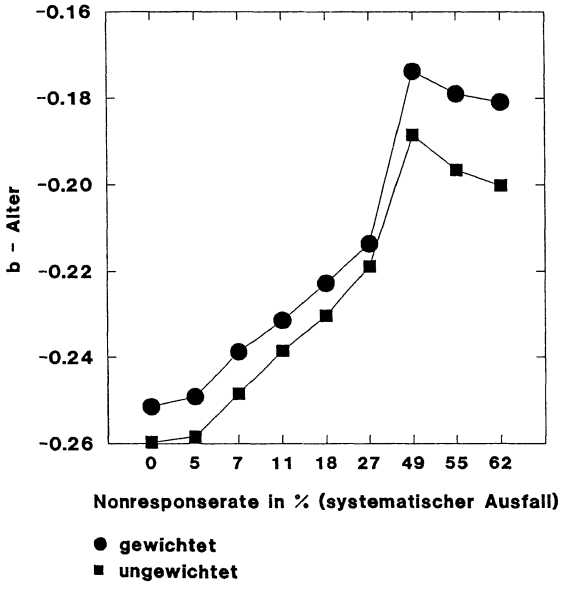

Abbildung 7 Veränderung des Regressionskoeffizienten für Alter im Postmaterialismusmodell.

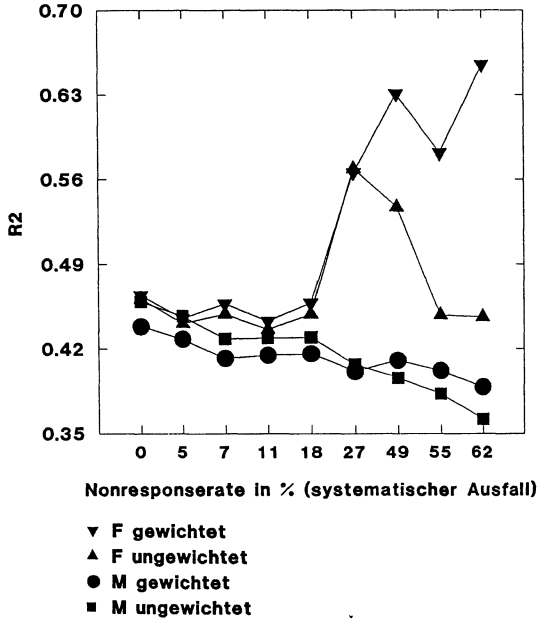

Abbildung 8 Veränderung der erklärten Varianz im Diekmann-Modell.

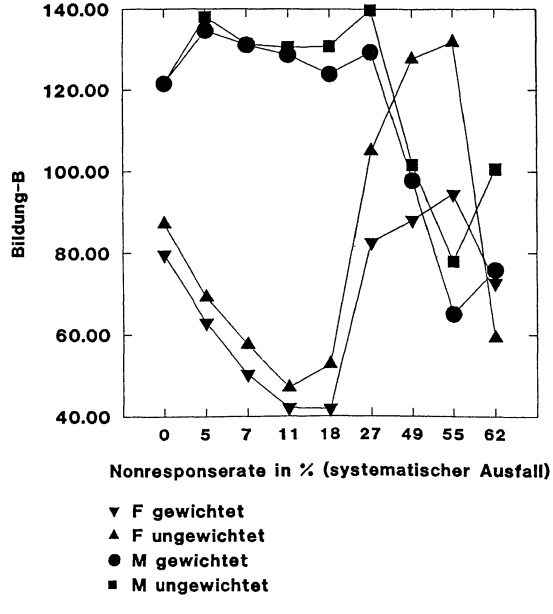

Abbildung 9 Veränderung des Regressionskoeffizienten für Bildung in Hinsicht auf Einkommen im Diekmann-Modell.

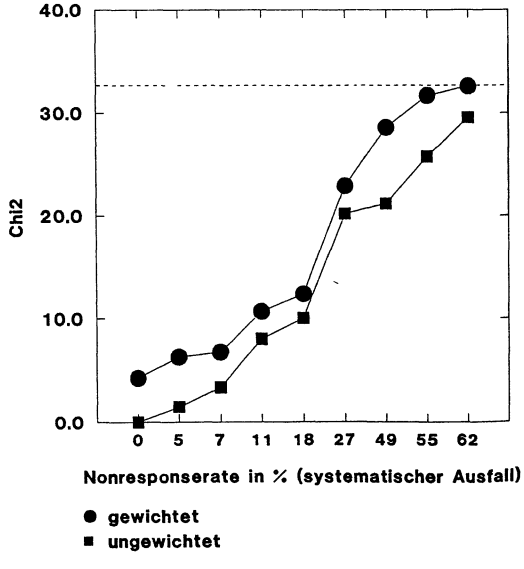

Abbildung 10 Chi² der Abweichung der Kovarianzmatrix im Diekmann-Modell (nur Frauen).

wiesen werden. Eine vergleichsweise geringe Nonresponserate von 11% führt zu einem deutlicheren Unterschied zwischen den Geschlechtern, der auch durch die Gewichtung nicht korrigiert wird. Allerdings ist diese Veränderung noch innerhalb der Konfidenzintervalle (trotzdem würde dies vermutlich – in Übereinstimmung mit Diekmann 1982: 342 – inhaltlich interpretiert werden). Bei sehr starken systematischen Ausfällen (ca. 50% und mehr) kehren sich die relativen Größen der Koeffizienten in den Subgruppen teilweise sogar um: Im ungewichteten Datensatz ist der Einfluß der Bildung auf das Einkommen in drei Fällen bei den Frauen stärker als bei den Männern, im gewichteten Datensatz einmal. In diesen Fällen wären möglicherweise andere Schlußfolgerungen gezogen worden als bei Diekmann.

Die Effekte der Ausfallmechanismen und der Gewichtung auf die anderen Variablen des Modells sind nicht einheitlich. Um die Darstellung etwas zu verkürzen, soll abschließend daher nur noch ein globaler Test auf die Gleichheit der dem Diekmann-Modell zugrundeliegenden Kovarianzmatrizen erläutert werden.

Der Test der Kovarianzmatrizen der Modelle (jeweils getrennt für Männer und Frauen)[36] bei Reduzierung der Kontaktversuche zeigt lediglich bei Reduzierung auf 1 oder 2 Kontaktversuche deutliche Abweichungen der Kovarianzmatrizen, die aber nur in einem Fall (bei Männern, nur ein Kontaktversuch, gewichtete Stichprobe) auf dem 5%-Niveau signifikant von der Ausgangsstichprobe verschieden ist: Diese Bedingung entspräche aber der Realisierung einer Quota-Stichprobe: unverzerrte Randverteilung (Alter, Geschlecht, Bundesland), trotzdem signifikant verschiedene Kovarianzmatrix durch Auswahl der leichter Erreichbaren.

Bei Annahme systematischer Ausfälle steigen sowohl bei Männern als auch bei Frauen die Abweichungen der geschätzten Kovarianzmatrizen mit der Nonresponserate an (vgl. Abbildung 10, nur Frauen). Nur eine der Stichproben mit immerhin 62% Ausfällen weicht auf dem 5%-Niveau (der kritische Chi^2-Wert ist in der Abbildung gestrichelt eingetragen) signifikant von der Kovarianzmatrix der Ausgangsstichprobe ab.

6. Schlußfolgerungen

Gewichtungsverfahren und Quotenstichproben basieren beide auf einer fast nie erwähnten Annahme: Die durch die Gewichtungsvariablen bzw. Quotenmerkmale gebildeten Klassen müssen entweder in Hinsicht auf alle interessierenden Variablen homogen sein oder innerhalb einer solchen Klasse dürfen sich die Antwortwahrscheinlichkeiten von Respondenten und Nichtrespondenten nicht unterscheiden. Anhand der Daten des ALLBUS 1980 wurde gezeigt, daß die Annahme der Homogenität der Gewichtungsklassen mit Sicherheit falsch ist. Die Annahme identischer Responsewahrscheinlichkeiten ist mit hoher Sicherheit ebenfalls falsch, allerdings läßt sich dies – per Definition von „Nonrespondent“ – nicht beweisen.

Da die Voraussetzungen für Gewichtungsverfahren damit nicht erfüllt sind, können diese auch keineswegs zur Korrektur von Ausfällen verwendet werden. In vielen (und vor allem: ohne explizite Theorie nicht vorhersagbaren) Fällen werden Schätzungen auf der Basis gewichteter Datensätze noch weiter verzerrt.[37] Die Verzerrungen umfassen selbst Veränderungen der Vorzeichen von Variablenbeziehungen und Subgruppenabfolgen. Verzerrungen konnten sowohl für „unsystematische“ als auch systematische Ausfälle anhand simulierter Ausfälle auf der Basis des ALLBUS 1980 gezeigt werden. Das gleiche Argument gilt für Quotenstichproben: Die Verzerrungen durch Beschränkung auf die am leichtesten Erreichbaren und Kooperationswilligsten ist auch durch die Übereinstimmung in den demographischen Variablen prinzipiell nicht auszuschließen, in den meisten untersuchten Fällen sind die Verzerrungen unter der „Quota“-ähnlichen Bedingung am größten.

Weiterhin wurde damit demonstriert, daß aus der Unverzerrtheit „demographischer Variablen“ in den Stichproben *nichts* über die Unverzerrtheit anderer Variablen folgt: Die u.a. in der Marktforschung gängigen „Repräsentanzbeweise“ sind vollständig unsinnig.[38]

Quota-Stichproben sind daher auch bei einer Stabilisierung der Ausschöpfung in *echten* Zufallsstichproben aus der „allgemeinen Bevölkerung“ von weniger als 70% keine ernstzunehmende Al-

[36] Die Dummy-Variablen wurden aus diesem Test ausgeschlossen, so daß nur die Kovarianzmatrizen der Variablen Alter, Berufsprestige, Bildung, Bildung des Vaters, Berufsprestige des Vaters und Einkommen getestet wurden.

[37] Ohne explizite theoretische Begründung gelangten auch Brög/Meyburg (1984: 188) sowie vor allem Hoag (1981: 16) zu dieser Schlußfolgerung.

[38] Diese „Repräsentanzbeweise“ werden auch als „Strukturkontrollen“ bzw. „externe Stichprobenvalidierungen“ bezeichnet.

ternative zu sorgfältig durchgeführten und kontrollierten (und damit teuren) Zufallsstichproben. Die Begründung hierfür liegt darin, daß ausschließlich echte Zufallsstichproben die prinzipielle Berechnung der Auswahlwahrscheinlichkeit für ein Element der Grundgesamtheit erlauben. Bei einer 100%-Ausschöpfung ist dies trivialerweise ausschließlich über das Design der Stichprobe möglich; in der Praxis müssen (und können) die Auswahlwahrscheinlichkeiten auf der Basis inhaltlicher Theorien geschätzt werden. Wie gezeigt wurde, scheitert die implizite Schätzung der Auswahlwahrscheinlichkeit über ein Gewichtungsverfahren an der nachweisbaren Inhomogenität der Gewichtungsklassen.

7. Über den Umgang mit Nonresponse

Die Feststellung, daß man am besten kein Nonresponse-Problem hat, ist ebenso alt wie korrekt. Dem wird aber keineswegs immer in der Praxis Rechnung getragen. Bei vielen Amateur-Projekten der empirischen Sozialforschung gibt es kaum eine ausreichende Zahl von Kontaktversuchen, einen Wechsel der Erhebungstechnik oder des Interviewers zur Reduktion von Nonresponse. Bei schriftlichen Erhebungen wird regelmäßig „aus Zeitgründen" auf Mahnschreiben verzichtet. Bei kommerziellen Instituten fallen bis zu 10% der Stichprobenelemente aus, weil die Interviewer keinen einzigen Kontaktversuch unternahmen. Berichtete Verweigerungen werden kaum je validiert. Schließlich werden Interviewer für ihre aufreibende Tätigkeit ungewöhnlich schlecht bezahlt.

Ein großer Teil der Ausfälle ist aber praktisch kaum zu vermeiden. Der einzig methodisch korrekte Umgang mit Nonresponse bei der Datenanalyse liegt in der expliziten inhaltlichen theoretischen Modellierung des Ausfallprozesses. Dann und *nur* dann kann das Ausmaß eventueller Verzerrungen durch Ausfälle abgeschätzt werden (Rubin 1987). Die Ausschöpfungsquote allein besagt überhaupt nichts. Für eine Abschätzung der Effekte des Ausfalls benötigt man die Responsequote *und* starke explizite, empirisch bewährte Hypothesen über die Unterschiede zwischen Respondenten und Nichtrespondenten.[39] Am einfachsten läßt sich dies für die Verzerrung des Mittelwertes zeigen: Die Verzerrung des Mittelwerts B ist gleich dem Anteil der Nonrespondenten M multipliziert mit der Differenz der Mittelwerte der Respondenten $\bar{Y}_R$ und dem Mittelwert der Nonrespondenten $\bar{Y}_N$: $B = M^*(\bar{Y}_R - \bar{Y}_N)$. Man benötigt also beide Angaben. Da die Parameter der Nonrespondenten nicht bekannt sind, müssen diese über mehrere unterschiedliche inhaltliche Hypothesen über den Ausfallprozeß geschätzt werden: Dies ist „multiple Imputation". Diese Vorgehensweise ist vergleichsweise schwierig: Neben technischen Kenntnissen wird vor allem eine inhaltliche (und das heißt bei den meisten soziologischen Fragestellungen: eine handlungstheoretische) *Erklärung* des Ausfallprozesses notwendig. Diese muß aber bei jeder veränderten Problemstellung von jedem Datenanalytiker jeweils neu entwickelt und möglichst empirisch geprüft werden. Die nicht zu überwindende Bindung an inhaltliche Modelle bedingt auch das nahezu vollständige Desinteresse mathematischer Statistiker an der Entwicklung solcher Ausfallmechanismen: Es gibt derzeit nur *ein* Standardmodell für systematische Ausfälle.[40] Daß solche Abschätzungen daher praktisch nie durchgeführt werden, ist kaum verwunderlich. Für die Entwicklung solcher Modelle benötigt man detaillierte Informationen über das Zustandekommen eines Ausfalls: Exakte Kontaktprotokolle sind daher unverzichtbare Voraussetzung für jeden Versuch einer Modellbildung. Selbst Methodenstudien enthalten diese Informationen in der Regel nicht. Weiterhin existieren fast keine Detail-Studien zum Ablauf einer Verweigerung; über den Prozeß ist nur wenig bekannt. Handlungstheoretisch ist zu erwarten, daß die meisten Verweigerungen situationelle Verweigerungen sind. Sollte diese Hypothese korrekt sein, dann wären die tatsächlichen Konsequenzen größerer Ausfallquoten durch Verweigerungen eher gering. Dies ist aber bisher empirisch nicht belegt.

Die Ausführungen sollten gezeigt haben, daß es zwar einfache praktische und etwas kompliziertere theoretische Techniken für den Umgang mit Nonresponse gibt, diese aber in keinem Fall kostenneutral sind. Standardgewichtungsverfahren und Quotastichproben sind zwar billig, bieten aber weder theoretisch noch empirisch eine Alternative. Es stellt sich abschließend die Frage, ob nicht wenige, inhaltlich theoretisch fundierte, teure Zufallsstichproben langfristig der Vielzahl von

[39] Vgl. z.B. Kalton (1983: 6–10).

[40] Bei diesen Modellen handelt es sich um Verallgemeinerungen des sogenannten Heckmann-Modells, vgl. z.B. Little/Rubin (1987: 223–230), Schnell (1986: 71–79). Dieses Modell besitzt eine Reihe eigener statistischer Probleme, die praktischen Anwendungen einen eher experimentellen Charakter verleihen.

schlecht vorbereiteten Erhebungen von „Einstellungen", die die Praxis der empirischen Sozialforschung dominieren, vorzuziehen sind.

Literatur

Alexander, C.H., 1987: A model-based justification for survey weights. American Statistical Association Proceedings of the Section on Survey Research Methods, 183–188.

Allmendinger, J./Schmidt, P./Wegener, B., 1983: ZUMA-Handbuch Sozialwissenschaftlicher Skalen, Bonn.

Bausch, T., 1990: Stichprobenverfahren in der Marktforschung. München: Vahlen.

Benninghaus, H., 1990: Einführung in die sozialwissenschaftliche Datenanalyse. München: Oldenbourg.

Böltken, F./Gehring, A., 1984: Zur Empirie des Postmaterialismus. Quota und Random, Äpfel und Birnen, Kraut und Rüben. ZA-Information 15: 38–52.

Brög, W./Meyburg, A.H., 1984: Non-Response-Effekte in großangelegten Mobilitätsuntersuchungen, Teil 1. Planung und Analyse 4: 168–172, 188.

Dalenius, T., 1983: Some Reflections on the Problem of Missing Data., in: Madow, W.G./Olkin, I. (eds.): Incomplete Data in Sample Surveys, Vol. 3. New York, S. 411–413.

DHP 1989: Gesundheitssurvey der Deutschen Herz-Kreislauf-Präventionsstudie (DHP), Dokumentation des Datensatzes für die 1. Stufe. Bonn, April 1989.

Diekmann, A., 1984: Einkommensdiskriminierung von Frauen – Messung, Analyseverfahren und empirische Anwendungen auf Angestellteneinkommen in der Bundesrepublik; S. 315–351 in: Mayer, K.U./ Schmidt, P. (Hrsg.): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Frankfurt: Campus.

Elliot, D., 1991: Weighting for non-response. A survey researchers guide. London: Office of Population Censuses and Surveys.

Feick, J./Mayntz, R., 1982: Bürger im Bürokratischen Staat: Repräsentative Beurteilungen und Handlungseinschätzungen. Die Verwaltung 15, 4: 409–434.

Forsa, 1989: Statistik in Deutschland: Akzeptanz, Erfahrungen und Meinungen (Studie im Auftrag des Statistischen Bundesamtes), unveröffentlichter Bericht, Dortmund.

Gabler, S., 1991: Eine allgemeine Formel zur Anpassung an Randtabellen. ZUMA-Nachrichten 29: 29–43.

Graef, F./Blien, U., 1989: Ein allgemein einsetzbares Verfahren zur Gewichtung von Stichproben, zur Disaggregierung von Daten und zur Ermittlung von Tabellen aus heterogenen Informationen. Allgemeines Statistisches Archiv 73: 122–142.

Esser, H., 1986: Über die Teilnahme an Befragungen. ZUMA Nachrichten 18: 38–47.

Esser, H./Grohmann, H./Müller, W./Schäffer, K.A., 1989: Mikrozensus im Wandel. Stuttgart: Metzler-Poeschel.

Goyder, J., 1987: The Silent Minority: Nonrespondents on Sample Surveys. Cambridge: Polity Press.

Groves, R.M., 1989: Survey Errors and Survey Costs. New York: Wiley.

Hansen, J., 1988: 70 Prozent? Ein Beitrag zur Ausschöpfung von Random-Stichproben. Planung und Analyse 15, 10: 398–401.

Hoag, W., 1981: Realisierte Stichproben bei Panels: Eine vergleichende Analyse. ZUMA Nachrichten 9: 6–18.

Hoag, W.J., 1986: Der Bekanntenkreis als Universum: Das Quotenverfahren der Shell-Studie. Kölner Zeitschrift für Soziologie und Sozialpsychologie 38: 123–132.

Hoem, J.M., 1989: The Issue of Weights in Panel Surveys of Individual Behavior; S. 539–565 in: Kasprzyk, D. et al. (eds.): Panel Surveys. New York: Wiley.

Kaplitza, G., 1982: Die Stichprobe; S. 136–186 in: Holm, K. (Hrsg.): Die Befragung, Bd. 1. München, 2. Aufl.

Kirschner, H.P., 1984: ALLBUS 1980: Stichprobenplan und Gewichtung; S. 114–182 in: Mayer, K.-U./ Schmidt, P. (Hrsg.): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Frankfurt.

Kromrey, H., 1990: Buchbesprechung: Methoden der empirischen Sozialforschung. Soziologische Revue 13: 201–202.

Kruskall, W./Mosteller, F., 1979a: Representative Sampling, I: Non-scientific Literature. International Statistical Review 47: 13–24.

Kruskall, W./Mosteller, F., 1979b: Representative Sampling, II: Scientific Literature, Excluding Statistics. International Statistical Review 47: 111–127.

Kruskall, W./Mosteller, F., 1979c: Representative Sampling, III: the Current Statistical Literature. International Statistical Review 47: 245–265.

Kruskall, W./Mosteller, F., 1980: Representative Sampling, IV: the History of the Concept in Statistics, 1895–1939. International Statistical Review 48: 169–195.

Landgrebe, K.P., 1992: Ausschöpfungen. Planung und Analyse 2: 19–22

Lamnek, S., 1978: Zugang zu und Ausschöpfung von Umfragepopulationen. Interview und Analyse 10/11/ 12/1978: 510–515, 566–570.

Lamnek, S., 1989: Qualitative Sozialforschung, Band 2: Methoden und Techniken, München.

Little, R.J.A./Rubin, D.B., 1987: Statistical Analysis With Missing Data. New York: Wiley.

Noelle-Neumann, E./Piel, E., 1983: Eine Generation später: Bundesrepublik Deutschland 1953–1979. München: Saur.

Oh, H.L./Scheuren, F.J., 1983: Weighting Adjustment for Unit Nonresponse; S. 143–184 in: Madow, W.G./ Olkin, I./Rubin, D.B. (Hrsg.): Incomplete Data in Sample Surveys. New York: Wiley, Vol. 2.

Rothe, G./Wiedenbeck, M., 1987: Stichprobengewichtung: Ist Repräsentativität machbar? ZUMA Nachrichten 21: 43–58.

Rothe, G./Wiedenbeck, M., 1988: Beschreibung der Prozeduren zur Berechnung des ZUMA-Gewichts

beim ALLBUS 1988. ZUMA Technischer Bericht Nr. 88/12.
Rothe, G., 1989: Gewichtungen zur Anpassung an Statusvariablen – Eine Untersuchung am ALLBUS 1984, 6. ZUMA-Arbeitsbericht 89/21.
Rubin, D.B., 1987: Multiple Imputations for Nonresponse in Surveys. New York: Wiley.
Schnell, R., 1986: Missing-Data-Probleme in der empirischen Sozialforschung. Dissertation, Bochum.
Schnell, R., 1991: Wer ist das Volk? Zur faktischen Grundgesamtheit bei allgemeinen Bevölkerungsumfragen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 43: 106–137.
Schnell, R./Hill, P.B./Esser, E., 1992: Methoden der empirischen Sozialforschung, 3. Auflage. München: Oldenbourg.
Sharot, T., 1986: Weighting survey results. Journal of the Market Research Society 28: 269–284.
Smith, T.W., 1983: The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey. Public Opinion Quarterly 47: 386–404.
Spiegel-Dokumentation, 1991: Prozente 5, Hamburg.
Upton, G.J.G., 1987: On the use of rim weighting. Journal of the Market Research Society 29: 363–366.
Wyss, W., 1990: Darf eine heilige Kuh geschlachtet werden? Die 70-Prozent-Marke bei der Ausschöpfung von Random-Umfragen. Planung und Analyse 17: 69–71.