



4.

## Datenprüfung und übergebener Datenbestand

Die Verfahren der Datenerfassung und Datenprüfung sind für die Datenqualität von großer Bedeutung. Im SOEP wird dieser Teil der Erhebungsarbeiten mit großem Aufwand, großer Sorgfalt und voller Transparenz gehandhabt. Die Verantwortung liegt bei Infratest Burke. Regeln und Kriterien der Datenprüfung werden mit der SOEP-Gruppe im DIW abgesprochen. Darüber hinaus erhält das DIW die Daten jeweils in zweifacher Form, nämlich den bereinigten und den unbereinigten Datenbestand. Eingriffe in die Daten, die im Zuge der Datenbereinigung vorgenommen wurden, können damit bei Bedarf festgestellt und überprüft werden.

In einer Panelbefragung ist auch die *Kontinuität* der Verfahren und Kriterien von Datenbereinigungen wichtig, um bearbeitungsbedingte Effekte im Längsschnitt zu vermeiden. Es ist daher erfreulich, daß das System der Datenerfassung, Datenprüfung und Datenbereinigung im SOEP der Sache nach weitgehend unverändert seit Beginn der Erhebung im Jahr 1984 beibehalten werden konnte. Die technischen Verfahren und Programme sind seither allerdings schrittweise immer wieder modernisiert und verbessert worden. Inhaltlich werden die Prüfprogramme im Rahmen einer gleichbleibenden Grundstruktur jedes Jahr aktualisiert, d.h. an die Veränderungen der Fragebogen angepaßt.

Wegen der bestehenden Kontinuität konnte auch darauf verzichtet werden, diesen Teil der Erhebungsarbeiten von Infratest im *Methodenbericht* jedes Jahr erneut darzustellen. Zuletzt wurde das Vorgehen in den Methodenberichten zum SOEP Ost/Welle 1-2 und im Methodenbericht für das Jahr 1992 (West/Welle 9 / Ost/Welle 3) beschrieben. Der jetzige Methodenbericht zum SOEP '97 soll diese Darstellung aktualisieren und mit erweiterter Dokumentation vertiefen (vgl. *Anhang 2*).

4.1

### Ablauf und Verfahren der Datenprüfung

Datenerfassung und Datenprüfung umfassen im SOEP '97 zehn aufeinander folgende Arbeitsschritte. Diese werden im folgenden kurz dargestellt.

Die Arbeit beginnt mit der Überarbeitung der Erfassungs- und Prüfprogramme und dem Test dieser Programme zu Anfang des Bearbeitungsjahres. Die eigentliche Erfassungs- und Prüfarbeit beginnt Ende Februar und wird parallel zur Feldarbeit durchgeführt, wobei der Rücklauf an Fragebogen in „Wochenportionen“ abgearbeitet wird. Der geprüfte Datenbestand für die Ost-Stichprobe wird bis Ende September, derjenige für die West-Stichproben Mitte Dezember an das DIW übergeben.

Die folgende Beschreibung des Arbeitsablaufs setzt an dem Punkt an, an dem der Rücklauf an Fragebogen registriert und haushaltsweise in der Datenbank („Paneldatei“) erfaßt ist.

### (1) Manuelle Vorprüfung („Editing“)

Prüfung der Vollständigkeit der ausgefüllten Fragebogen pro Haushalt. Vorbereitung der Fragebogen für die Datenerfassung. Dabei geht es insbesondere um Fehler oder handschriftliche Vermerke, die bei der Erfassung Probleme bereiten würden bzw. in der späteren maschinellen Datenprüfung nicht oder nur schwer erkennbar sind (z.B. schlecht lesbare Texte, Eintragungen außerhalb der vorgesehenen Antwortfelder, unzulässige Mehrfachnennungen usw.). Dabei können eventuell vorhandene handschriftliche Anmerkungen berücksichtigt werden. Textantworten auf offene Fragen bzw. „Sonstiges“-Kategorien werden signiert oder auf Strichlisten erfaßt.

Der Editingplan im einzelnen ist als **Anhang 2.1** beigefügt.

### (2) Datenerfassung (Scanner)

Die Datenerfassung erfolgte bis 1994 mit manueller Dateneingabe. Seit 1995 ist sie auf ein modernes Scannersystem umgestellt. Die Umstellung wurde durch Tests und Kontrollen sorgfältig vorbereitet. Ein Vergleich von 120.000 erfaßten Angaben mit den Eintragungen im jeweiligen Originalfragebogen ergab eine Fehlerquote der Datenerfassung von 0,08%, das sind 8 Fehler auf 10.000 Angaben. Die Erfassungsqualität ist damit deutlich besser als bei herkömmlicher Dateneingabe.

Das Scannersystem hat einen zusätzlichen Vorteil, der sozusagen als Nebeneffekt genutzt wird. Das Verfahren erfaßt zunächst optisch komplette Fragebogenseiten, um im zweiten Schritt die markierten Angaben zu „lesen“. Durch ein Zusatzprogramm werden die als Bild erfaßten Fragebogenseiten abgespeichert und damit archiviert. Die ausgefüllten Fragebogen sind dann zunächst auch noch in Papierform vorhanden; sie werden jedoch nicht mehr benötigt, da der Inhalt auf CD-Rom verfügbar ist (rd. 40 CD-Rom pro Befragungswelle). Bei Bedarf kann das gespeicherte Originalbild des ausgefüllten Fragebogens am Bildschirm aufgerufen werden.

### (3) Basisprüfung (Prüfung A)

DV-gestützte Prüfung der Vollständigkeit und Konsistenz der verschiedenen Datensätze pro Haushalt. Damit wird insbesondere auch sichergestellt, daß die Identität einer Person, der ein Datensatz zugeordnet wird, zweifelsfrei feststeht. Dies ist Grundvoraussetzung konsistenter Längsschnittdaten.

Eine genauere Beschreibung enthält **Anhang 2.2**.

*(4) Prüfung Haushaltsfragebogen (Prüfung B)*

In einem DV-Programm werden mögliche Fehler und Lücken in den Daten definiert. Bei fehlenden Angaben, die als zulässig definiert sind, wird der KA-Code „-1“ gesetzt. Bei einer Reihe von Variablen werden fehlende Angaben jedoch nicht akzeptiert, sondern lösen eine Einzelfallprüfung aus, die nach Möglichkeit zur Vervollständigung der Angaben führt. Dieser Teil des Prüfprogramms wird als Programm B1 bezeichnet.

Das Programm B2 enthält die Prüfungen auf inhaltliche Fehler. Tritt ein Fehler auf, wird ein Fehlerprotokoll auf Papier ausgedruckt. In diesem Protokoll können ggf. Datenbereinigungen und Erläuterungen vermerkt werden. Die eigentliche Datenkorrektur erfolgt online mit Hilfe eines Maskenprogramms, das pro Frage den Wortlaut und die Antwortvorgaben sowie den eingetragenen Antwortcode zeigt.

Nach Eingabe eventueller Datenkorrekturen wird der Prüflauf über das Prüfprogramm wiederholt, um die Fehlerfreiheit der Daten bestätigen zu lassen bzw. in einem weiteren Durchgang herzustellen.

Die Regeln für die Datenbereinigung im Fall aufgetretener Fehler oder fehlender Angaben sind in sogenannten „Bearbeitungsregeln“ schriftlich dokumentiert. Die Bearbeitungsregeln für den Haushaltsfragebogen des SOEP '97 sind in **Anhang 2.3** wiedergegeben.

*(5) Prüfung Personenfragebogen (Prüfung C)*

Das Vorgehen entspricht demjenigen beim Haushaltsfragebogen. Das Prüfprogramm ist wiederum geteilt in ein Programm C1, mit dem bestimmte Werte und KA-Codes gesetzt werden, und ein Programm C2, das zum Ausdruck von Fehlerprotokollen und manuellen Prüfungen im Einzelfall führt.

Bestimmte Fehlerarten können und müssen in jedem Fall korrigiert werden. Das gilt für

- unzulässige Mehrfachnennungen (MFN)
- sowie Filterfehler.

Bei anderen Fehlertypen ist eine Korrektur nicht in jedem Fall möglich. Das gilt für

- Wertebereiche (Wert liegt außerhalb eines als plausibel definierten Bereichs)
- Summen (Aufaddierung mehrerer Werte ergibt unplausiblen Summenwert)
- Nicht plausibel: Angabe steht im Widerspruch zu Angaben an anderer Stelle des Fragebogens.

Ist eine Angabe eindeutig (!) unplausibel und kann nicht korrigiert werden, wird der Wert gelöscht und durch den Code „-3“ ersetzt.

Die Bearbeitungsregeln für Prüfprogramm C (Personenfragebogen) sind in **Anhang 2.4** wiedergegeben.

## (6) Längsschnittprüfungen

Die Längsschnittprüfungen beziehen sich ebenfalls auf den Personenfragebogen, sind jedoch in einem eigenen Prüfprogramm zusammengefaßt. Dabei geht es im wesentlichen

- um die „Kalenderprüfung“, d.h. die Prüfung des fortgeschriebenen Tätigkeitskalenders
- und die Prüfung von Statusänderungen, und zwar zum Familienstand und zur familiären Situation (Frage 103), zum Erwerbsstatus und zum Bildungsabschluß.

Nähere Einzelheiten sind in **Anhang 2.5** dargestellt.

## (7) Prüfung des Haushaltseinkommens

Dieser Arbeitsschritt ist im SOEP '97 neu in die Datenprüfung aufgenommen worden. Er wird daher gesondert im folgenden Kapitel 4.2 dargestellt.

## (8) Prüfung Lebenslauf (Prüfung D)

Diese Prüfung bezieht sich auf den „Personenfragebogen 2: Lebenslauf“, der von erstmals befragten Personen zusätzlich zu beantworten ist. Das Vorgehen entspricht dem oben beim Haushalts- und Personenfragebogen dargestellten Verfahren. Die Bearbeitungsregeln sind in **Anhang 2.6** wiedergegeben.

## (9) Vercodung von Ausbildungsabschluß, Beruf und Branche

Im Personenfragebogen werden Ausbildungsabschluß, Beruf und Branche an verschiedenen Stellen offen erfragt. Die Textangaben werden im Wortlaut erfaßt und anschließend nach vereinbarten Klassifikationen vercodet.

Fragen 59-60 fragen nach einem eventuellen *Ausbildungsabschluß*, der im letzten Jahr abgeschlossen wurde. Sofern es sich um einen Hochschulabschluß handelt, wird dieser nach dem Fachrichtungsverzeichnis des Statistischen Bundesamtes vercodet. Sofern es sich um einen beruflichen Ausbildungsabschluß handelt, wird er nach dem erweiterten Verzeichnis der Ausbildungsberufe vercodet. Die Vercodung wird durch die Datenprüfer bei Infratest vorgenommen.

Die berufliche Tätigkeit wird nach der *International Standard Classification of Occupations (ISCO 88)* vercodet, die *Branche* nach einer SOEP-spezifischen, an NACE angelehnten Klassifikation. Die Verschlüsselung erfolgt durch ZUMA, Mannheim. Infratest erstellt einen Auszugsfile mit den entsprechenden alphanumerischen Daten und nimmt eine erste Durchsicht der Texte auf korrekte Schreibweise vor. Die von ZUMA zugeordneten Codes werden unmittelbar an das DIW übermittelt.

Im SOEP '97 war dabei die folgende Menge an Texten für die Vercodung aufzubereiten:

	Anzahl
Personenfragebogen:	
derzeitiger Beruf (Frage 28)	7.598
derzeitige Branche (Frage 31)	7.301
berufliche Nebentätigkeit (Frage 56)	1.115
Lebenslauf-Fragebogen:	
Beruf des Vaters (Frage 33)	174
erste eigene Berufstätigkeit (Frage 47)	216
sofern nicht mehr berufstätig:	
Branche der letzten beruflichen Tätigkeit (Frage 52)	66

#### (10) Abschlußprüfung

Der kumulierte geprüfte Datenbestand (getrennt nach Haushaltsfragebogen, Personenfragebogen, Lebenslauf-Fragebogen und Bruttoband) wird abschließend anhand einer Grundauszählung auf Vollständigkeit und eventuelle Auffälligkeiten überprüft.

## 4.2 Prüfung des Haushaltseinkommens

Im Haushaltsfragebogen wird nach dem derzeitigen monatlichen Gesamteinkommen des Haushalts gefragt (Frage 50). Zusätzlich kann man ein rechnerisches Gesamteinkommen des Haushalts bilden, indem die einzelnen Einkommensarten - die sowohl im Haushaltsfragebogen als auch für die einzelnen Haushaltsmitglieder im jeweiligen Personenfragebogen erfragt werden - aufsummiert werden. Bis 1994 bestand dabei die Schwierigkeit, daß diese Einzelangaben sich überwiegend auf das „letzte Kalenderjahr“ bezogen. Seit 1995 werden sie auch für den Zeitpunkt „heute“ erfragt. Damit bietet sich nun erstmals die Möglichkeit, die zwei verschiedenen Messungen des Haushaltseinkommens einander gegenüberzustellen und daraus ggf. Hinweise auf den „richtigen“ Wert oder auf mögliche Korrekturen und Ergänzungen von Einkommensangaben abzuleiten.<sup>5</sup>

Infratest entwickelte im Frühjahr 1997 einen Vorschlag für ein mögliches Vorgehen. Der Vorschlag beruhte auf der genauen Analyse von 100 Testdatensätzen, d.h. 100 befragten Haushalten. Das Ergebnis ist in **Anhang 2.7** wiedergegeben.

Nachdem das DIW den Vorschlag grundsätzlich begrüßt und gutgeheißen hatte, wurde die Einkommensrechnung pro Haushalt, wie sie in Anhang 2.7 dargestellt ist, für alle 6.810 befragten Haushalte im SOEP '97 durchgeführt. Beim Vergleich zwischen aufsummierten Einzeleinkommen und pauschal erfragtem Haushaltseinkommen (HHFr50) ergibt sich folgendes Bild:

	Anzahl	%
Keine oder geringfügige Abweichung beider Werte (mit einem Toleranzbereich von +/-20%)	4.735	69,5
HHFr50 ist kleiner ( $\leq 80\%$ )	623	9,2
HHFr50 ist größer ( $\geq 120\%$ )	517	7,6
Keine Rechnung, da KA bei Einzeleinkommen	935	13,7
Summe Haushalte	6.810	100,0

Für die mittleren beiden Teilgruppen, also für 1.140 Haushalte, wurden Fehlerprotokolle bzw. Übersichten über alle möglichen Einkommenskomponenten ausgedruckt und Fall für

<sup>5</sup> In dieser Einkommensrechnung bleiben jährliche Sonderzahlungen sowie generell alle Kapitaleinkünfte außer Betracht. Diese werden nicht als laufende Einkommen zum Zeitpunkt „heute“ erfragt, sondern nur rückwirkend für das letzte Kalenderjahr. Im Regelfall wird man annehmen können, daß auch das pauschal erfragte derzeitige monatliche Haushaltseinkommen diese Einkommenskomponenten nicht einschließt.

Fall gesichtet. Bei Bedarf wurden die Originalfragebogen herangezogen, da diese teilweise handschriftliche Erläuterungen enthalten.

Das Ergebnis dieser aufwendigen Prüfung ist folgendes:

- (1) Eine Korrektur oder Ergänzung einzelner Einkommensangaben war in 53 Fällen möglich. Das sind weniger als 1%. Meist kann man zwar plausible Hypothesen über die Gründe der Divergenzen aufstellen - aber dies ist kein ausreichend sicheres Fundament, um Änderungen der Daten vorzunehmen.

Es bliebe höchstens die Möglichkeit, durch eine telefonische Nachbefragung zu versuchen herauszufinden, welche der Einkommensangaben im Fragebogen fehlerhaft oder lückenhaft waren. Zu befürchten ist allerdings, daß dies zu negativen Reaktionen auf seiten der Befragten führen und damit die weitere Teilnahmemotivation beeinträchtigen könnte. Dieser Weg soll daher nicht beschritten werden.

- (2) Die Einzelfallprüfungen bestätigten den bereits im Test gewonnenen Eindruck, daß in der Regel der höhere der beiden Gesamteinkommenswerte die höhere Plausibilität aufweist. Der jeweils niedrigere Wert ist der Tendenz nach als „underreporting“ - mit verschiedenen Gründen im Einzelfall - zu sehen. Den geringsten Fehler dürften Einkommensanalysen im SOEP dann machen, wenn sie den jeweils höheren der beiden Gesamteinkommenswerte zugrunde legen.

Im Datensatz, der an das DIW übergeben wurde, sind beide Werte als Variable vorhanden. Zusätzlich wurde ein Kontrollcode generiert, der angibt, ob für den jeweiligen Haushalt Personenfragebogen für alle Haushaltsmitglieder ab 16 Jahren vorliegen. (Ist dies nicht der Fall, führt das häufig zu einem zu niedrigen Wert für das errechnete Haushaltseinkommen.)

Der Effekt der empfohlenen Faustregel („höherer Wert des Gesamteinkommens sticht“) läßt sich mit folgender Auswertung zeigen:

Durchschnittliches Haushaltseinkommen	DM	Index
Pauschal erfragt (HH Fr. 50)	3.852	100
Aus Einzelangaben errechnet	3.968	103
Der jeweils höhere Wert	4.187	109

Das im SOEP erfaßte monatliche Haushaltseinkommen erhöht sich bei Anwendung des vorgeschlagenen Verfahrens um 9%.

### 4.3 Übergebener Datenbestand

Der vollständige geprüfte Datenbestand des SOEP '97 wurde dem DIW Mitte Dezember 1997 auf CD Rom übergeben. Ebenfalls enthalten war der ungeprüfte Datenbestand.

Der Datenbestand umfaßt folgende Teile:

	Satzlänge	Fallzahl
Nettodaten Haushaltsfragebogen	600	6.810
Nettodaten Personenfragebogen	2.000	13.283
Nettodaten Lückenfragebogen	2.000	162
Nettodaten Lebenslauf-Fragebogen	1.400	494
Bruttodaten Haushalte (Paneldatei)	63	7.423
Bruttodaten Personen <sup>6</sup> (Paneldatei)	77	19.031

---

<sup>6</sup> einschl. Kindern unter 16 Jahren.