

Wagner, Gert G.; Schupp, Jürgen; Rindtl, Ulrich

Working Paper — Digitized Version

The Socio-Economic Panel (SOEP) for Germany: Methods of production and management of longitudinal data

DIW Discussion Papers, No. 31a

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Wagner, Gert G.; Schupp, Jürgen; Rindtl, Ulrich (1991) : The Socio-Economic Panel (SOEP) for Germany: Methods of production and management of longitudinal data, DIW Discussion Papers, No. 31a, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/95772>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Deutsches Institut für Wirtschaftsforschung

Discussion Paper No. 31a

The Socio-Economic Panel (SOEP) for Germany - Methods of Production and Management of Longitudinal Data

Gert G. Wagner, J. Schupp und U. Rendtel

Berlin, August 1991

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 1000 Berlin 33
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

The Socio-Economic Panel (SOEP) for Germany -

Methods of Production and Management of Longitudinal Data

Gert G. Wagner, Juergen Schupp and Ulrich Rendtel¹

I Aims

The Socio-Economic Panel (SOEP) is a survey started in 1984 that can be placed under the heading of population and income statistic. The SOEP is very similar to the well known Panel Study of Income Dynamics (PSID) for the United States of America but the questionnaire of the SOEP is broader. The SOEP data amplifies official statistics, both regarding additional variables that arise as well as the longitudinal character of the data, i.e. the repeated surveying of the same respondents. This kind of data is highly conducive to describing and analyzing changes such as those triggered by the developments in the GDR. Thus the SOEP was expanded already in June 1990 to include the territory of the GDR, respectively East Germany ("Neue Bundeslaender").

1.1 Well-Being and Micro-Economic Approach

For researching structural transformation and the ensuing change on the macro-level of economy and society the empirical analysis of micro-units has stepped into the foreground since the 1970s. After initially analyzing extensive cross-sectional data sets of private households and individuals it didn't take long to realize that - just as in the traditional time series analysis of aggregate data - micro units need to be temporally followed in order to test empirically hypotheses about transformation processes within a society.

With the SOEP, panel data should be supplied for two theoretical approaches: the choice of the economic variables on micro-economic and micro-econometric

¹ The authors are grateful to their colleagues within the SOEP-group at the German Institute for Economic Research (Deutsches Institut für Wirtschaftsforschung, Berlin): Joachim Frick, Elke Holst, Peter Krause, Rainer Pischner, Marlis Riebschläger und Johannes Schwarze. The German Institute for Economic Research is a cosponsor of the SOEP which is a project of the German National Science Foundation (Deutsche Forschungsgemeinschaft, Bonn). Since 1990 the SOEP has been financed by the Federal/State Commission for Education Planning and Promotion of Research (Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung). The SOEP was started in 1984 in cooperation with the German Institute of Economic Research and a special research unit of the universities of Frankfurt and Mannheim (Sonderforschungsbereich 3 "Mikroanalytische Grundlagen der Gesellschaftspolitik"). This research unit was funded by the National Science Foundation.

approaches, by which the former can be placed under human capital theories in the deepest sense, and the latter concern theoretical approaches to labor market segmentation and poverty research. The sociological (and some for political science) variables are determined by the ideas of social indicator movement.

Along with the longitudinal aspect and the theory induced selection of variables, the SOEP with its household-based survey design incorporates yet another major advance in sociological as well as micro-economic science. By interviewing every adult person in the household central information from the "household" primary network is made available. In particular more recent theoretically innovative approaches of family economics as well as all the literature on network analyses stress the necessity for household data bases of this sort.

The opening of the inner-German borders and the merger of the GDR with the FRG were naturally a challenge for a wide-ranging household sample. A repeated survey is of course ideally suited to an empirical recording of upheaval in the GDR and its repercussions in the FRG.

1.2 Program of Questionnaire

Both the analytical design and the survey of the SOEP are therefore complex in a number of ways: This can be said for the organization of the variable selection in a finely-meshed economic and sociological approach; the placement of the individual perspective in a household context; the representative cross-sectional and longitudinal analyses as well as the overproportional inclusion of a separate migrant worker sample. Purely cohort-oriented surveys (such as the NLS in the USA) work from a narrower theoretical basis.

The analytical possibilities extend through the following eight topics²:

- Demography and Population
- Labor Market and Unemployment
- Income, Taxation and Social Security
- Housing
- Health
- Education and Training

² The documentation for all questionnaires fills at present three volumes. Even a condensed outline would be impossible in the space of one book. See therefore the SOEP "User's Manual".

- Economic Output of Private Households
- Basic Orientation and Values.

In the majority of the topics in the SOEP information on objective living conditions as well as subjective perceptions are gathered. In addition to the thematic mashing, another exceptional feature of the SOEP is the mashing of very disparate time references. For many indicators the time reference of the SOEP questions ranges over the respective period in which the survey was taken to deep in the past. Thus, for example, information - particularly concerning economic conditions - is also gathered by retrospective questioning (e.g. about the previous year's income). Questions about one's well-being, among others, also have a prospective character (e.g. "How content do you think you'll be in five year's time?"). Such questions, together with the purely time-related factual and employment questions, make up the standard survey program for investigating stability and intra-individual change of respondents. In addition to this standard part of the longitudinal indicators there are wave-specific questions themes of the questionnaire.

Overview I shows the focal questioning points ("topical moduls") for the first ten years of the survey.

Overview 1: Topical moduls of the Socio-Economic Panel

Year	
1984	Working Life History
1985	Marital and Family History
1986	Social Origins, Occupational Starts, Residential Environment
1987	Social Security, Early Retirement, Need to be Cared For and Care of Children
1988	Balance of Assets
1989	Continued Education and Qualification
1990	Time Expenditure and Preferences
1991	Family and Financial Support
1992	Social Services and Social Security (repeated from Wave 4)
1993	Labor Market

The questions in the SOEP are standardized and "closed", meaning that the response possibilities are clearly limited. There are, however, a series of "open questions", the answers to which are noted in uncoded text in the questionnaire. The SOEP thus gives general opportunity to do a certain amount of "qualitative social research", since in the course of time a wide range of uncoded text information on all household members is contained in the data base. For data protection reasons, however, this information is not distributed to external data users (cf. section 6.3 below).

In the SOEP Study it is also standard procedure that the occupation and the economic sector are transcribed in uncoded text and is recorded later. This procedure proved extraordinarily useful in avoiding classification problems with the survey in the GDR (cf. Geiss and Hoffmeyer-Zlotnik 1991). Since for the old GDR economy and the transitional processes there are no comparable and reliable international classification schemes, registering the respondents' information in uncoded text allows for any sort of re-coding according to new classification schemes.

1.3 Representativeness in longitudinal and cross-sectional analyses

In general, when we speak of the representativeness of a random sample we don't mean a perfect microcosmos of the entire target population. What is merely meant is that for each question a variety of representative terms and standards should be usefully formulated because, in the practice, a perfect sample in the sense of a "microcosmos" is not to found³.

Roughly speaking, a sample survey can serve two purposes: Either its aim is description, i.e. the target population with all its aggregates and distributions is depicted as accurate as possible⁴, which is important if the results are to affect political decision-making. Or - and this is paramount in the sciences - with the aid of a random sample, theories about causal hypotheses can be tested. Although it may seem surprising at first, the standards for achieving representativeness are lower for testing theories than they are for purely descriptive purposes.

An initial problem with representativeness is isolating (by definition) the relevant target population. Most random samples in the Federal Republic of Germany are not designed to represent the total population because foreigners and the institutionalized population aren't surveyed. This was attempted in the SOEP, although admittedly difficulties of a practical nature in surveying and contacting have shown that the institutional population can not be adequately represented. A subsample was designed for the most important foreign group, the Southern Europeans from the former guestworker-recruiting countries.

The SOEP target population is therefore a broader one than in other German surveys, although the SOEP hasn't succeeded in portraying the entire residential population either. Special problems dealing with immigration will be discussed in a later chapter shortly. Extending the relevant target population to include the territory of the GDR posed no technical problems for the sample. Operations in this area will be discussed in the following section.

Usually, when one speaks of representative problems after the target population has been established, this is either a matter of the structure of the intended target population being reflected correctly in the sample or, with disproportionately-designed samples, whether an appropriate procedure for "correcting" the sample can be found. To be kept in mind with panel surveys is the third stage of representativeness: whether as a result of repeated surveying the sample members will begin to demonstrate behavioral changes (the "panel effect").

³ Cf. on this subject the three-part synopsis by Kuskal and Mosteller (1979).

⁴ Strictly speaking, "good" is meant here in regard to specific research questions. The valuation will differ for each problem and question.

Considering the thematic content of the SOEP surveys and the fact that the interviewing takes place only once a year, serious panel effects are not to be reckoned with. To be sure, problems could crop up if "target households" refuse to participate. In Sections 4.1 and 5 this will be gone into more concretely. For now we'll just let the general comment stand.

1.4 Description of Contents

In Section 2 which follows, the surveying and the field work for the 1st wave of the SOEP will be described. Panel-related problems are not touched upon here, but the 1st-wave surveying was crucial to the later representativeness of the longitudinal section. In Section 3 we begin with panel-related material. First the so-called follow-up concept, which of course is of central importance for a longitudinal survey, will be presented. In addition the "panel-specific" construction of the survey instruments will be discussed and documented. In Section 4 the development of the SOEP samples to date will be dealt with, and in Section 5 necessary evaluations and projections of the sample results will be gone into. In Section 6 processing the data, which is far more complicated than with simple cross-sectional surveys, will be discussed and illustrated.

2 Surveying and Field Work in the 1st Wave

In this section the random samples, the surveying and the field work in the 1st SOEP wave will be clearly defined. In the old Federal Republic of Germany the panel began in 1984 with samples A and B (for Germans and foreigners respectively). In June 1990, with sample C (Germans in the GDR), it was expanded to include the territory of what was still the GDR at that date.

The surveying is carried out by "Infratest Social Research" (Munich). Infratest is responsible for the field work, for collecting and processing the data and for documenting the survey before the anonymous micro-data are turned over to the Panel Project Group.

2.1 Target Population and Respondents

The target population to be represented by the SOEP is defined firstly as the German residential population of the FRG in 1984 including (West) Berlin, secondly as the German residential population in the GDR in June 1990. In the FRG, in addition to German persons in private households, households of selected foreign groups were included in the study.

Respondents are all household members who are 16 years of age and over; information on the younger household members is obtained from the "chief respondent" (head of household). By means of these "proxy information", analyses for children (e.g. on their educational history) are also possible. This information can be linked to the personal information obtained later when the children have reached respondent age. Since all adult persons in a household are to be interviewed, the telephone interview as a surveying method is eliminated, although in the case of an interviewer change the telephone plays an important part as a contact medium between respondent household and interviewer/survey institute.

For technical reasons the original FRG sample in 1984 was carried out separately for two populations:

Sample A "Germans in the FRG" covers persons in private households with head who is not Turkish, Greek, Yugoslavian, Spanish or Italian.

Sample B "Immigrant Sample" covers persons in private households with a Turkish, Greek, Yugoslavian, Spanish or Italian head.

Inmates of institutions in the true sense of the word (hospitals/sanatorien; rest homes; military installations) were not included in the first sample; later, however, persons from the initial households were included who had taken up residence temporarily or permanently in institutions of this kind (cf. Section 3.1 below) and who were still capable of taking part in the survey.

Sample C "Germans in the GDR" covers persons in private households where the household head (chief wage-earner in the household) is a GDR citizen. This meant that ca. 1.7% of the residential population in the GDR in June 1990 was excluded from the sample as foreigners. Because they are predominantly institutionalized residents, and institutional households aren't included in the survey, they would have already been excluded when the target population was established.

The size of the samples was pragmatically determined. For method-related reasons, it is optimal for the sample to be as big as possible; however, available financial means are notoriously tight.

2.1.1 Sample A "Germans in the FRG"

Sample A was intended to encompass a net amount of 4,500 households; in the end the completed net sample contained 4,554 households. The ADM master tape from 1982⁵ served as a basis for collecting sample A. 584 sample points were randomly selected from it by means of a multi-stage stratified sampling procedure. The interviewer selected the households within the selected constituency according to the random-route procedure. Working from a given start address the interviewer had to write down 84 addresses. Every seventh household was a "target household" and thus to be recruited for the survey.

2.1.2 Sample B "Immigrants in the FRG"

Strictly speaking, sample B consists in turn of five autonomous samples for the five numerically largest foreign nationality groups living as immigrants in the FRG in 1984. To facilitate detailed evaluations, the projected net case number for the five nationalities (the Southern European recruiting countries for so called guestworkers) was set higher than it would be in proportion to the percentage of the population they constitute (disproportionate approach). In setting the case numbers for the nationalities - a total of 1,400 foreign households were to be surveyed - the anticipated mobility/re-migration rates were taken into account.

For surveying population B, a random selection, separately for each nationality, was first made of counties and metropolitan areas. Using the immigrant registration records there, the respondents were then selected according to random procedure. The household of the respondent selected in this manner then came into the sample, provided that the household head had the same citizenship as the selected respondent. In a number of counties and metropolitan areas - particularly in Baden-Württemberg - it wasn't possible to draw from the immigrant registration lists; the alternative solution here was to randomly select counties and then use the local residents' registration lists.

The number of sample points was set at 80 for the (strongly overly-represented) Turks and 40 for each of the remaining nationalities. 20 addresses per point were then drawn from the registers, some of which were used as reserve addresses.

The number of addresses used per sample point in the immigrant sample shows a stronger mean variation than in sample A, the reason being the substantially higher proportion of quality-neutral attrition caused chiefly by false or no-longer-current addresses, in which case a reserve address was to be used.

2.1.3 Sample C "Germans in the GDR"

Although in the Spring of 1990 German unification was already in sight, it made sense to set the size of the East-sample C in such a way that independent analyses for the GDR respectively the new "Bundeslaender" would be possible. Therefore, compared to sample A, a greater sampling rate was chosen and a goal of at least 2,000 households was striven for; 2,179 were ultimately surveyed.

Because access was granted to addresses from the central residents' file of the GDR, a different - and better - sample method than in samples A and B was possible. Serving as a basis for the selection was the master sample (issue date: March 14, 1990) issued for the GDR by our survey institute Infratest. This is, in contrast to the ADM master sample in the old FRG⁶ a random selection of private addresses drawn from the central residents' data base (cf. Pietzke 1991).

To design sample C of SOEP the Infratest master sample was used in the following ways:

6 The ADM group first brought out a drawing volume for the new Bundeslaender using the election districts as a basis after the first national election in both Germanys in Dec.1990.

- First a household-proportional allocation was calculated for 360 sample points which followed the stratification of the master sample according to county and community size.
- For each stratum the sample points which corresponded to this household-proportional allocation were then taken from the master sample by systematic random selection.
- Finally, for each of the available addresses for these sample points a person 16 years of age or older was selected as a "start address". In order to produce a representative household sample (and to spare costs and travelling time by lumping together the respondent addresses) the random-route method was chosen. Commencing with this start address, each interviewer was to list the households on a formally described and clearly defined random route, whereby the start address itself was not to be surveyed.
- Ten private households were to be listed and recruited for panel participation⁷ unless it turned out while making contact that one of the listed addresses didn't belong to the target population (or that the residence was vacant). In that case up to two substitute addresses could be listed and contacted. Every third household was a "target household" and thus to be recruited for the survey⁸.

Insiders from the GDR social research organizations raised objections against taking addresses out of the central register, claiming that often at these addresses other people lived there than the ones who were registered there. Although this assertion is difficult to prove or disprove, in view of the housing shortage in the GDR it is plausible and probable. The quality of the SOEP, however, is in no way affected by this because the address sample was merely used to ensure a random and representative regional distribution of the respondent households only. Whether or not the residence is vacant and who may live there is in fact irrelevant for the random-route method.

2.2 Gross Amount of Addresses and Quality-Neutral Attrition in the 1st Wave

2.2.1 Samples A and B

According to the sample plan the original gross approach in sample A encompassed 7,008 addresses. Of the 1,168 reserve addresses included therein, 158 were not used because in each respective sample point a maximum response rate (9 or 10 households with completed interviews) had already been attained or seemed to be within reach. But in the sample points with weak response rates the sample was boosted with 1,129 addresses. So altogether 7,979 addresses were used.

⁷ A separate, preliminary address collection was not possible due to lack of time.

⁸ This three-address interval is routinely used in the random-route procedure by the ADM institutes. With the West-samples A and B a wider interval was constructed (7 households; cf. Section 2.1.1 *ibid*) in order to attain a higher degree of independence for the households. This proved impossible in the GDR because the interviewers were unfamiliar with the random-route procedure to begin with, so it made no sense to burden them additionally with more distance to cover.

In order to maintain the mathematical sample gross, households that don't belong to the target population "Private Households Excluding the Separately-Interviewed Households in Sample B" must be subtracted from the total amount of start addresses. These addresses are defined as "quality-neutral drop-outs" and the result as "edited gross amount".

As Table 1 shows (left block), there was 5.8% quality-neutral attrition in sample A. The edited gross amount encompassed 7,519 addresses. The quality-neutral attrition is a result of the address procedure, namely the interviewer's notation of house numbers along the pre-determined route. With some addresses it doesn't become clear until contact is made at a later date that the household doesn't belong in the target population (because the household members are in sample B). This was the case in 2.8% of the addresses on the lists. With other addresses it was discovered upon closer inspection that they were business addresses (0.5%) or vacant dwellings (1.9%). And then 1.0% were either false addresses or couldn't be found.

Sample B (the immigrant sample) gives us a completely different picture because here addresses supplied by the registration offices were worked with. The extent of the quality-neutral attrition due to false or no-longer-current addresses is very much greater here than in sample A. The average rate of attrition for the five immigrant samples is 22% of the utilized addresses.

Table 1: Gross and Net Samples in the 1st Wave of the SOEP

	Sample								
	Cases	A	%	Cases	B	%	Cases	C	%
Adresses Used	7979	-	-	2659	-	-	3616	-	-
Quality-neutral drop-outs ¹⁾	460	-	-	579	-	-	502 ²⁾	-	-
Edited gross amount total	7519	100		2080	100		3144	100	
Attrition due to									
- refusal	2403		32	452		22	496		16
- failure to contact household	243		3	148		7	184		6
- other reasons	319		4	65		3	255		8
Total number of drop-outs	2965		39	665		32	935		30
Households queried									
- before editing	4554		61	1415		68	2179		70
- after editing ³⁾	4528		60	1393		67	-		-

- 1) Cases with false adresses, uninhabited house/apartment, not a private household or wrong nationality of household head.
- 2) Also included here are 29 total drop-outs from sample points (i.e. from 290 addresses) which were not dealt with at all during the field period.
- 3) Errors of falsifications that first came to light during the 2nd-wave field work and data checking.

The next question to arise concerns the sample response rate. In sample A, 4,554 addresses could be taken into the net sample after concluding all of the field phases and the processing work. With an edited gross amount of 7,519 addresses this means for the time being a sample response rate of 60.6%.

Better results were shown with the foreign households from sample B. The response rates ranged from 64.7% for the Italians to 70.0% for the Turks.

In sample A refusal was the main cause of attrition. Because of the long field period the share of non-contacted households could be reduced to 3.2% (a percentage not attained in normal cross-sectional surveys). The other grounds for attrition were provided by households that couldn't be surveyed due to linguistic difficulties (0.2%); the foreigners in question didn't belong to any of the five nationalities from sample B and therefore "rode along" with sample A. Lastly there is the 0.8% of the addresses for which no survey information exists. As a rule these are reserve addresses which the interviewer didn't realize were supposed to be contacted.

The refusal rates for the foreign households from sample B are visibly lower than for the German households. However, the share of non-contacted households is higher.

The quality of a sample (cross-sectional representativeness) is measured by the conformity of characteristic distribution in the realized sample (net sample) with the characteristic distribution in the target population. The distribution in the target population is estimated using external statistics, posing of course the problem that they on their part could be biased again (particularly the income and consumer samples). Beyond that, the possibility was used for the SOEP of undertaking an "internal validation", which results from comparing basis information on households for which it wasn't possible to conduct complete interviews with information on households willing to participate.

The internal validation is a matter of whether those households and persons who were willing to participate in the survey systematically distinguish themselves from those who were not prepared to participate. The practical difficulty with this type of validation is that structural characteristics of households which aren't prepared to participate in a survey are normally unknown. Because of the significance of the sample validation for the SOEP, however, the attempt was made here to gather as much information on the drop-outs as possible. This is a difficult enough undertaking, one that is - rightfully so - hindered further by strict observance of data privacy laws.

This information is available for 31% of the drop-outs in sample A. In sample B this area of the analysis had to be omitted due to lack of information.

Table 2: Internal Validation of the SOEP Samples A and B

	Sample A		Sample B	
	SOEP	Attrition	SOEP	Attrition
Total number of cases	4 554	2 906	1 415	652
<u>Size of household</u>				
Basis (case number)	4 554	1 728	1 415	261
	%	%	%	%
1 person	26	30	15	15
2 persons	32	36	17	18
3 persons	19	16	20	26
4 or more persons	24	18	48	42
<u>Sex of household's head¹⁾</u>				
Basis (case number)	4 554	1 799	1 415	269
	%	%	%	%
male	76	67	92	91
female	24	33	8	9
<u>Age of household's head¹⁾</u>				
	4 554	1 523	1 415	213
	%	%	%	%
under 29	14	9	12	9
30 - 39	18	12	29	28
40 - 49	22	18	34	34
50 - 59	17	17	22	23
60 - 69	13	17	3	6
70 - 79	12	20	0	-
80 or older	4	7	0	-
<u>Vocational status²⁾</u>				
<u>of household's head¹⁾</u>				
Basis (case number)	4 494	894	-	-
Blue-collar worker	39	33	-	-
White-collar worker	36	38	-	-
Civil servant	11	10	-	-
Self-employed	11	13	-	-
Trainee	1	1	-	-
Has never been employed	4	6	-	-
<u>Status-group rating</u>				
lowest ³⁾	25	23	-	-
highest ⁴⁾	15	11	-	-

1) Household head/respondent.

2) For unemployed: last position applies.

3) Unskilled worker, low-ranking office worker.

4) Highly-qualified managerial employee, high-level official, self-employed person, self-employed person with 10 or more employees.

Table 3: External validation of the SOEP samples A and B as well as C

	A		B		C	
	Sample-	Dev. ¹⁾	Sample	Dev. ¹⁾	Sample	Dev. ¹⁾
					-	-
<u>Persons</u>						
<u>Bundesland</u>						
(West) Berlin	3,9	-0,6	3,2	-0,6	-	-
Hamburg	2,8	-0,6	2,8	-0,3	-	-
Bremen	1,3	-0,1	1,1	0,0	-	-
Schleswig-Holstein	4,0	-0,2	1,2	-0,7	-	-
Lower Saxony	11,2	0,0	5,7	-0,6	-	-
North Rhine-Westphalia	26,8	-0,5	28,5	-2,7	-	-
Hesse	8,8	-0,3	13,6	2,6	-	-
Rhineland-Palatine/ Saarland	7,5	+0,3	4,2	+0,1	-	-
Baden-Württemberg	15,5	+0,9	26,9	+4,2	-	-
Bavaria	18,3	+1,2	12,8	-1,9	-	-
Mecklenburg- West Pomerania	-	-	-	-	11,7	+0,1
Brandenburg	-	-	-	-	16,2	+0,3
Saxony-Anhalt	-	-	-	-	18,5	+0,3
Thuringia	-	-	-	-	16,7	+0,4
Saxony	-	-	-	-	30,3	+0,1
(East) Berlin	-	-	-	-	6,7	-1,1
<u>District population</u>						
less than 5.000 inhabitants	12,0	+1,5	2,8	-0,2	-	-
5.000 to less than 20.000	14,0	+0,2	13,7	-0,3	-	-
20.000 to less than 100.000	10,1	-0,2	7,7	+0,4	-	-
100.000 to less than 500.000	16,1	+0,3	17,5	+2,4	-	-
500.000 or more						
- central districts	32,4	-1,6	38,1	-3,0	-	-
- outlying districts	15,4	-0,2	20,1	+0,6	-	-
less than 2.000	-	-	-	-	26,6	+2,2
2.000 to less than 10.000	-	-	-	-	17,2	-2,3
10.000 to less than 50.000	-	-	-	-	23,9	+0,4
50.000 to less than 100.000	-	-	-	-	5,8	-1,2
100.000 or more	63,9	-1,5	75,7	0,0	26,5	+0,8

Table 3 continued:

<u>Sex</u>						
male	47,8	-1,5	-	-	47,5	+0,5
female	52,2	+1,5	-	-	52,5	-0,5
<u>Age group</u>						
16-19 years old	9,0	+0,5	-	-	7,2	+1,0
20-29	19,6	+1,3	-	-	20,8	+0,7
30-39	16,6	+1,6	-	-	23,8	+4,7
40-49	19,6	+1,6	-	-	17,4	+2,2
50-59	14,9	+0,2	-	-	16,0	-0,4
60-69	10,4	1,1	-	-	9,0	-2,8
70 years and more	10,0	-4,1	-	-	5,7	-5,5

1) Deviation of the non-weighted sample from the official population statistics in percentage points.

In regard to the regional distribution characteristics and the household structure, sample A reflects the familiar shortcomings of survey research. The population in the conurbations - and here particularly in the central zones - is more difficult to recruit for survey participation than the population in the medium and small-sized towns and communities. Elderly persons are under-represented due to difficulties in interviewing. In comparing with the population statistics, however, it should be pointed out that those data also contain institutionalized residents who are not included in the SOEP sample.

The case number of drop-outs for who information is available on the occupational status of the household head is with $N = 894$ quite small. The first result which comes to the fore is that the share of blue-collar workers among the drop-outs is not as large as in the realized sample. The white-collar workers and self-employed persons are, in contrast, slightly over-proportionately represented among the drop-outs; with the civil servants no deviation is noticeable. This result is oddly inconsistent with the current theory that survey research is "middle-class biased" and it speaks for the quality of surveys which are carefully designed, and carried out.

The validation for sample B is far more difficult. Whereas for the German residential population relatively reliable and diversified distribution information is available from official sources, available structural data on the immigrant residential population is scanty. Thus possibilities for an external sample validation using official data are practically non-existent, making the internal validation that much more important (cf. Table 2).

In regard to the socio-demographic structure of the sample - household sizes as well as the age and sex of the household head - the result is extremely satisfactory. In regard to the regional distribution, the drop-out structure shows the same result as for sample A, namely intensified attrition in the core zones of the conurbations.

2.2.2 *Sample C*

As shown in Table 1 (right block), with sample C in the GDR a response rate of 70% of the "edited gross addresses" was attained. This is a field result that is practically impossible to achieve with similar studies in the FRG. Not even in 1984 in the 1st wave of the SOEP in the FRG could similar results be achieved, despite great effort and month-long field work. The response rate at that time was noticeably lower with 60.6% of the German and 68.0% of the foreign households (left block in Table 1).

From an edited gross sample with 3,114 GDR household addresses it was possible with a field duration of a bit more than one month to successfully recruit a total of 2,179

households⁹. A total of 4,453 persons 16 years or older were queried at these households.

A special problem, about which similar reports were to be heard from other survey institutes active in the GDR in 1990, was the total attrition of sample points. Total attrition means that the addresses of a sample point don't get processed at all because for one reason or another the interviewer responsible for these addresses could not or would not carry out this assignment, and a replacement was unavailable or could not be deployed in time (it could even be that the field organization was not alerted in time to the fact that the addresses remained uncontacted).

In the SOEP basis survey 29 out of 360 sample points were total drop-outs. These, however, are distributed randomly throughout the entire GDR and thus have no noticeable effect on the sample. The neutrality of these drop-outs is also verified by the regional validation of the sample.

The total attrition is, on the one hand, a consequence of the inadequate telecommunication facilities in East Germany, and secondly an indication of the problems involved in setting up or restructuring an interview staff. Our Survey institute Infratest had opted for taking over an already existing interviewer network from the former GDR and reorganizing it to meet with the new standards. Thus, we made an analysis of possible interviewer effects in the survey data. As in West Germany, these proved to be inconsequential; there is, in particular, no effect of the entry-date of interviewers into the staff (cf. Riebschlaeger and Wagner 1991).

The completed sample was examined immediately after the survey was taken with - as long as comparative statistics were available - official statistical data from the former GDR (cf. Table 3; right block).

- Regionally there is practically no sample deviation in the distribution among the Länder.
- Although in the community-size classes there are some deviations from the target population in the completed sample, they don't show the general under-representation of metropolitan population as is normally observed in the West. It is noticeable, however, that the persons/households in Berlin were slightly under-represented.
- In the distribution of sexes there are no recognizable deviations worth mentioning.

⁹ Three percent of the households didn't get fully surveyed until the first week of July. Since the interview date is recorded in the data set, it's possible to determine during the evaluation process whether the data was collected before or after the currency union (July 1).

- As in sample A, the elderly age groups - especially the 70-years-and-older group - are clearly under-represented in the net sample. The comparative figures from the population statistics, however, are somewhat biased because the institutional residents - who are not surveyed in the SOEP - couldn't be edited out. Nevertheless: the age structure bias in the SOEP is a consequence of an under-representation of one-person households. Their exact representation, though, is difficult to establish due to a lack of appropriate basis data from official sources. A validation of the socio-economic structures is even more difficult, since the data basis of the official GDR statistics in this field is even scantier (cf. Wagner and Schupp 1991); cf. Pischner (1991) concerning projections.

2.3 The Interviewer Sample

A scientific panel survey such as the SOEP provides the opportunity of designing an "interviewer panel" as an additional sample. In order to examine potential interviewer effects (cf. Hoag 1981) a number of interviewer characteristics are supplied by the survey institute along with the actual survey data. In the SOEP this interviewer information can be applied without time limitations and can be linked to the surveyed households. Thus an additional data record that can be used for important methodical analyses is made available (cf. on this subject Rendtel 1990, Riebschlaeger and Wagner 1991).

The interviewer sample for samples A and B from 1984 encompasses 631 persons. Sample C was carried out in 1990 by 308 interviewers.

3 Panel-Related Construction of Survey Instruments and How to Ensure Continued Respondent Participation

For a panel survey the concept of following the sample members and ensuring their continued participation are the central parameters of the study. The procedure developed for this purpose will be explained in following, whereby it will also be seen that due to learning effects at several points in the course of time there have been procedural changes made.

3.2 The Follow-up Concept

For a serial survey which includes all household members the criterion for following persons must be set so as to maintain the representativeness of the selected target population. When one disregards immigration with self-establishment of a household, it is possible with the right follow-up concept to reflect the natural population dynamic of the target population in the panel sample. It is necessary first of all that young household members maturing into respondent age (meaning completion of their 16th year) also get interviewed; secondly, persons who leave an "initial household" must be

followed according to a specific pattern. This means that persons who moved in while the panel was in progress and moved out again afterwards don't necessarily have to be followed for the sake of cross-sectional representativeness. However, children who have moved directly from a foreign country into a household (since they, in a manner of speaking, virtually belong to the target population of the 1st wave) must be followed, as well as children who move into initial households while the panel is in progress (cf. on this subject Galler 1986). This concept of following initial persons was first chosen for the SOEP.

For longitudinal analyses, especially of demographic events, it is more practical to follow not only the initial persons but all persons who have ever come in contact with the SOEP. This procedure increases the number of events which are of particular research interest. But this leads potentially to a snowball-like growth effect in the sample because, hypothetically, with enough mobility and a correspondingly lengthy panel duration the entire target population will take root in the sample in the end. However, this is only theoretically the case, because in practice the respondents' successive refusal to participate leads to a reduction in the sample instead. As was to be expected, little willingness to participate is shown by persons who move into an established panel household. Following non-initial persons will therefore create no problems of size in the sample. Only the number of analyzable cases sinks less quickly than with a straight initial-person concept. Since 1990 (West-wave 7) all non-initial persons have been followed because of the special attraction of following new persons, since these persons generally constitute unusually mobile groups, which are interesting for event-oriented analyses¹⁰.

In cross-sectional evaluations the weights for households with non-initial persons are somewhat lower than, for instance, other households (cf. also Section 5). The simplified follow-up rules do not affect the validity of the initial person concept for longitudinal studies in the least. Here, as opposed to cross sections, only the initial persons have a positive longitudinal-projection factor.

Another major problem with the new follow-up concept is presented by the immigrants who establish new households in the FRG. Because of the fact that at the beginning of the 80s the immigration of migrant workers or their family members was the only factor that played a major role, the decision was made in the conceptual stages of the SOEP not to do a successive surveying of immigrant households. Even with the pioneering American study PSID no immigrants were surveyed in the course of 25 waves despite the high rate of immigration to the USA. This can be justified in that

10 Beckett et al. (1989) also recommend this for the PSID sample. In adhering to the original SOEP concept it came to light anyway that the concept of following initial persons was too complicated for the actual field work, i.e. persons were

the primary concern is longitudinal analysis, in which immigrants - by definition - don't play such a big role¹¹.

The dramatic increase in the immigration rate in the old Federal Republic of Germany since 1988 (Eastern-Bloc citizens of German descent, Germans from the GDR and seekers of political asylum) meanwhile necessitates a supplementing of the panel sample if information on the residential population is to be continued to be given because this immigrant population became very important quantitatively (according to our estimates, this population now in 1991 constitutes 3% of the private households in the FRG).

3.2.2 Re-Surveying of Longitudinal Characteristics with Insufficient Information

In the SOEP a distinction is made between "final" and "temporary" drop-outs. A drop-out is considered final in the case that a household/person can no longer be found despite intensive effort to do so or if a final refusal is given to further participation in the survey. In all other cases the attrition is assessed at first as "temporary". This means that in the succeeding wave a renewed attempt will be made to contact the household and to persuade them to participate further in the survey. If this in turn isn't successful the "temporary" becomes a "final" drop-out - this household will no longer be included in the following panel waves. If, however, the household can be re-motivated to take part, a gap in the data caused by the attrition in the preceding wave will remain.

also interviewed who shouldn't have been according to the rules of the initial-person concept. Moreover, the expanded concept greatly simplifies the weighting procedure, since the sampling probabilities are now easier to calculate.

11 It must be admitted that the neglect of immigrants by a panel study with the poverty problem as a focal point of its research is argumentatively not at all easy to justify. In 1990 an - externally funded - immigrant sample of so-called Hispanics will be included in the PSID for the first time. The second, equally important immigrant population, the Asiatics, are still missing in the PSID.

Table 4: Temporary Attrition - Utilizable Cross-Sectional Data Records from Samples A and B with "Gaps"

	Persons	Households
Wave 2	202	91
Wave 3	176	79
Wave 4	83	33
Wave 5	126	55
Wave 6	147	66

In addition to these household-related gaps, other individual-person gaps appear when a household couldn't be completely surveyed in one wave but in the following waves all household members are interviewed. There were, for example, 123 households with incomplete interviews in the 2nd West-wave, 53 of which participated fully in the succeeding 3rd wave. Accordingly, there are 250 persons to be reckoned with who have gap-ridden event histories. Although this case number is small in relation to the whole sample, it should be taken into consideration that the gaps in each wave appear repeatedly for other households or persons. A procedure was therefore developed whereby at least the most important longitudinal characteristics for the missing year are able to be reconstructed.

In the 3rd year of SOEP the decision was made to try to close these gaps. The results were unexpectedly good. Of 192 persons from 134 households to be queried, longitudinal information could be reconstructed for 176 of them. In the majority of cases this was done by telephone. Owing to the good results the decision was then made to reconstruct the gaps in wave 2 as well. The long period of time between the confirmation of the gaps in wave 2 (summer 1986) and the reconstruction (Autumn 1987) proved to be a handicap; not more than roughly 75% of the persons concerned could be included in the field work because a relatively high number of target persons had dropped out of the panel in the meantime. Since 1987 (West-wave 4) re-surveying is done routinely.

3.2.3 Reconstruction of Biographical Information

The central thematic content of the first typical moduls of the three waves for samples A and B (West) was biographical information which only had to be recorded once in the course of the panel: Job history in yearly stages of the individual life course, beginning with the age of 15 (wave 1); marital history, information on childhood and the move away from the parental home, information from women about their children (wave 2) as well as social background, occupational starts (wave 3).

This information is missing completely or partially for persons who weren't included in the panel until later and for persons who were temporary drop-outs in the 2nd or in the 3rd wave. The retrospective questions were therefore put into a supplementary biographical questionnaire. In the fall of 1987 the new persons who had entered the SOEP in the 2nd, 3rd or 4th wave were sent the questionnaire along with a small gift, a pocket calendar¹². The reconstruction of the biographical information on the temporary drop-outs from wave 2 was done primarily by telephone - along with the

¹² Not included in the supplement questionnaire, though, were the persons who were new to the survey because they had just turned 16, since most of the questions didn't apply to them or the information on the parental home was already available.

reconstruction of the longitudinal characteristics in uncompleted operations (cf. the results in the preceding section).

Since 1988 (West-wave 5) a routine gap reconstruction is done by means of a special biographical questionnaire. For the East-sample C the biographical reconstruction for all respondents will take place in 1992 as a part of the regular fieldwork of wave 3.

3.3 The Method Mix in Surveying

The SOEP survey instruments are

- Cover sheets (address protocol)
- Household questionnaires
- Individual questionnaires

With regard to the layout and the filter questions, the questionnaires are compared to commercial questionnaires lavishly designed because they have to be flexibly used by the interviewer:

- In samples A and B, along with the basic form of the oral interview it's also possible for the respondent to fill out the questionnaire himself.
- In sample B one has the option of conducting the interview in German or in the respective foreign language (or in a mixed form too).

In the cover sheets (the so called address protocol) a vast amount of information on the composition of and change within the household is also recorded, information which is essential for making complicated longitudinal analyses.

The SOEP surveying procedures constitute a method mix:

- The basic form of the survey is the personally conducted oral interview.
- The respondent, however, is permitted to fill out the questionnaire, which is handed to and explained to him by the interviewer.
- In the event of a refusal to participate or non-appearance of target persons a new interview date will be agreed upon in writing or by telephone assistance.
- If the respondent wishes, the (new) interview date can be cancelled and, as an exception, the interview will be conducted in writing (i.e. by mail) or by telephone.

One drawback is admittedly very strictly implemented: information on a respondent can only be obtained from the respondent him/herself. Proxy interviews which are common, for instance, in the American SIPP study and are necessary in the PSID for all household members other than the head of the household, are principally not allowed. There are only a handful of cases where exceptions are made, for example when an immigrant household member gives permission to another household member to fill out his personal questionnaire.

With this multi-method approach alone the potential amount of persons who can be contacted and are willing to do an interview can be permanently guaranteed and maximally utilized (cf. Table 5).

Table 5: Development of the Interview Methods for Personal Questionnaires in the German Households from Sample A¹⁾

	Year of Interview						
	1984	1985	1986	1987	1988	1989	1990
	Column percentage						
Oral Interview	61	58	59	56	55	53	50
Questionnaire filled out by respondent	-	22	23	23	26	26	27
Mixed form	2	6	5	8	9	10	12
Proxy interview	0	0	0	0	0	0	0
Sum total of interviews completed under face-to-face interviewer guidance	98	95	94	92	91	90	89
Questionnaire filled out by respondent under telephone guidance	2	5	6	8	9	10	11
Sum total of all completed interviews	100	100	100	100	100	100	100

1) Based on the personal questionnaires

The information on the interview method applied in individual cases is available in the data. So here too systematic analyses of method-related influences can be made. To date there has been no indication that the interview method has a strong influence on the results.

3.4 Continued Motivation of Panel Respondents

3.4.1 Instruments

With the SOEP there are two important areas of ensuring sample continuance:

- The contacting and contact maintenance for recruiting the respondents for initial and repeated participation (motivation).
- The tracing of households that move to a new address between two surveys (updating the address register).

The following methods of motivating respondents are employed.

- Giving the study a catchy name. All sample respondents know the SOEP by the name of "Life in Germany".
- An illustrated informative brochure on the aims of the study (including in sample B a translation into the respective native language).
- An information sheet on data privacy¹³
- A letter of thanks after completion of the field work for each wave.
- Each respondent receives a ticket for a well-known TV lottery.
- Since 1987 (4th West-wave) all panel households receive a small gift ("loyalty bonus") worth 5-10 DM.

In addition to these measures, the motivation of the interviewer is certainly an important influential factor for the respondents' willingness to participate. A good training, sufficient information about the project, a clear structuring of the survey instruments and information on research results furnish the basics of successful interviewing. For years now, all of the interviewers involved in the surveys receive a thank-you card from the client (the DIW) at the end of the year in order to underscore the relevance of their engagement.

¹³ For legal reasons the data privacy information is of special importance for recurring surveys but on the whole is probably not a means of winning people's trust because, as test examinations show, the impression is made that the questions are of a more sensitive nature than is in fact the case (cf. Hippler, etc. 1990).

With the 1st wave in the then GDR, it was an advantage for the SOEP in regard to the respondents that the project bore the title "Life in Germany". In form letters and in the accompanying GDR survey brochure it was pointed out that in the old FRG the survey had been running since 1984, and its lofty intentions of documenting life in Germany could now be fully realized.

Unlike the 1st SOEP wave in 1984 in the FRG, the planned interview date for the households could not be pre-announced by mail. Due to the badly-functioning telephone system, meetings were hardly able to be arranged by phone in the routine way this has always been done by the SOEP. This led to no problems worth mentioning, however, because in the GDR unannounced visits were a part of daily life.

3.4.2 Household - Interviewer Continuity

A panel survey represents a panel not only for the respondents but for the interviewers themselves.

The SOEP had two interviewer-deployment strategies to choose from:

- Assigning the survey work to as many interviewers as possible, meaning that in a borderline case the number of interviewers deployed would correspond to the number of sample points.
- Concentrating the survey work on a minimum number of highly-qualified interviewers.

Starting point in 1984 was the first strategy, i.e. a maximum deployment of interviewers in order to provoke as few clusters of interviewer effects as possible. It has become noticeable in the field work in several waves, however, that for the high response rate which the SOEP requires some interviewers are better-suited than others. Moreover, a change of interviewers is an important determinant for respondents' refusal¹⁴. Thus it proved necessary to seek an optimum between as many or as few interviewers as possible, whereby it had to be kept in mind that the loss of a single interviewer, who interviewed a lot of households very effectively, increases the danger of a great many households refusing to participate.

3.5 Panel Files and Following Up Addresses

A "panel file" which was built up within the survey institute Infratest is very important for the success of the field work. This file contains the addresses, telephone numbers, interview method, and so on for every household.

¹⁴ It could also be shown that prior length of participation in the panel - with the exception of the newly-arrived persons - has no statistically significant influence on the willingness to participate (cf. Rendtel 1990).

During the whole year the whereabouts of each surveyed household and person no longer at the same address as the previous year is checked up on. Information on this follow-up work gets stored in the panel files.

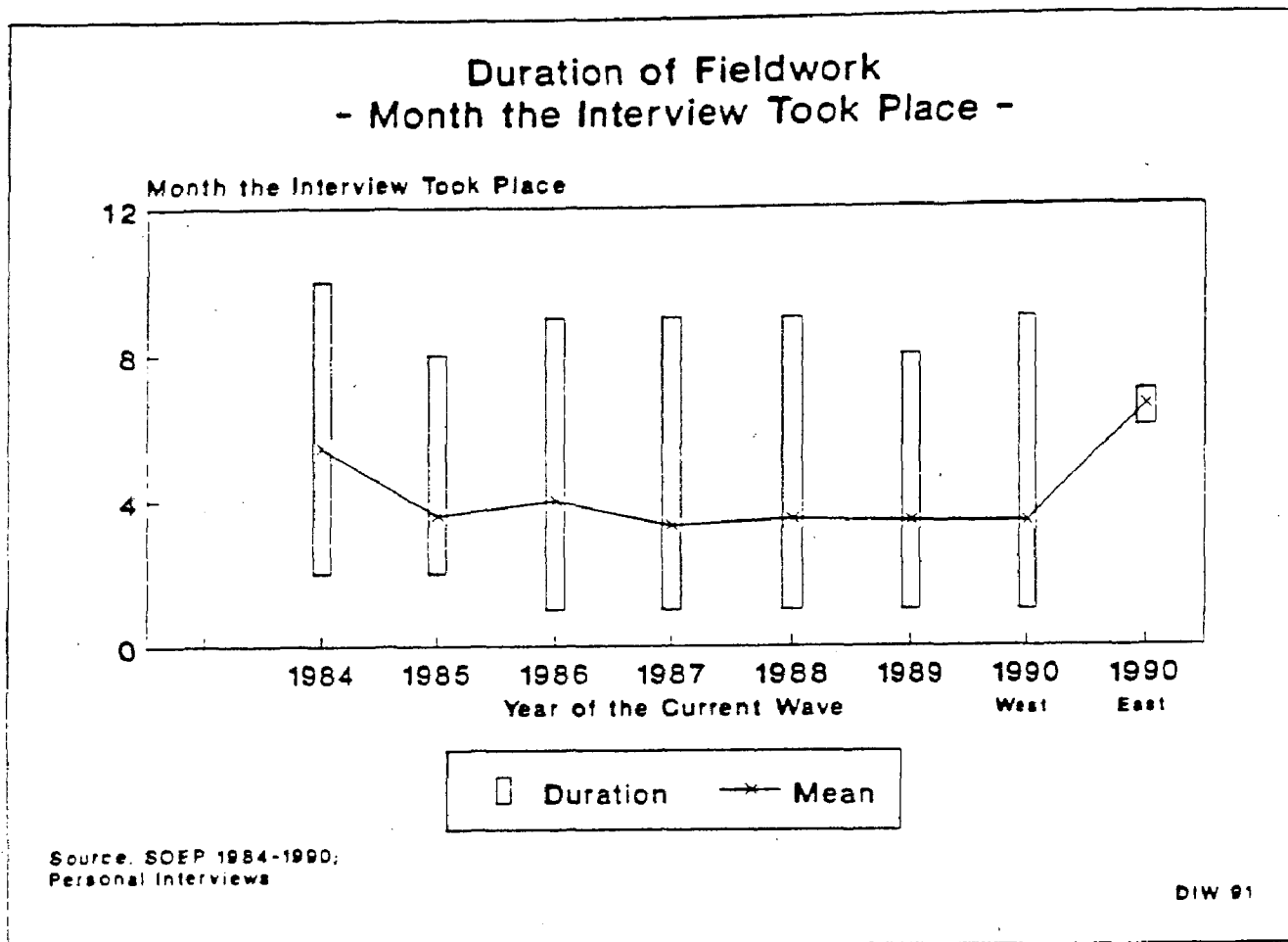
Each year approximately 7% of all the households from sample A are to be found at a new address, whereby this share is in decline again since 1988 after an initial increase. Moreover, there are what could be termed "partial moves" by persons who have left households and established new ones at a new address for the SOEP follow-up concept. This is the case each year for about 4% of all households from sample A.

The address research for the survey institute can be given a very good rating because it's had over 90% success since the beginning of the project. The source for the new addresses is in far more than half of the cases the interviewers themselves, who relay the new addresses to the field office. The remaining addresses are obtained half of the time by inquiring at the postal department and the remainder by inquiring at the residents' registration office.

The field-work sequence was altered after wave 3. Now the cases that are considered difficult are contacted first in order to ensure enough time for getting the job done without jeopardizing the field schedule.

The SOEP field work takes months to complete because a more than 90% response rate has to be attained in the follow ups. In order to make reporting-date-based evaluations, the date of the interview is retained in the analyzable data record, too.

Figure 1



4 Development of Samples A and B including C

The following section will deal with the development in the size and structure of the sample.

4.1 Samples A and B

Determinants of the Development

The sample development in the SOEP is in one respect a result of demographic change (departures by death or a move abroad, entries when household members reach respondent age or establish new households). On the other hand the state of the sample is fundamentally determined by the success of the field work. The degree of this success is measured by two aspects:

- Was it possible to contact the households from the previous wave again?
- Was permission to interview given in the contacted households?

Table 6 shows that the attrition resulting from inaccessibility plays a minor role in general. The drop-out rate of 1.9% in wave 2 could also be reduced to 0.9% in wave 7. However, there are varying degrees of difficulty in contacting households from the foregoing wave. This is verified in Table 6 by the clearly disparate rates of attrition in the respective household groups. The rate of attrition in the group of the one-person households who have moved is thus substantially higher than it is in households with no change of address. The field work for both household groups could be considerably improved; even more difficult to track down are "split-off households". Here too there's no discernable trend towards improvement in the field work.

Table 6: Attrition Rates in % due to Inaccessibility

	Wave 2		Wave 3		Wave 4		Wave 5		Wave 6		Wave 7	
	N	%	N	%	N	%	N	%	N	%	N	%
Total	6051	1,9	5814	1,4	5465	1,0	5342	0,9	5156	0,9	5044	0,9
Households with no change of address:	5413	0,8	5039	0,4	4808	0,1	4683	0,1	4545	0,2	4472	0,0
One-person households that have moved:	119	21,0	180	14,4	142	7,7	143	5,6	126	4,7	122	5,7
Splitt-off households:	221	11,7	295	8,4	242	10,4	242	7,4	246	11,8	263	12,9

N = Total number of cases (old households and new households - households dissolved on account of death - households that have moved abroad)

% = Percentage of households not contacted

The success of the next phase of the field work is in proportion to the number of interviews obtained from the re-contacted households. These households are divided into three sub-groups: participant from the previous year, temporary drop-out from the previous year, and newly established households. The attrition rates shown in Table 7 demonstrate the varying degrees of success with each single group. As anticipated, the best results were obtained with households who had also participated in the foregoing panel. Cause for gratification here is that it was possible to boost the success of the field work. From wave 2 to wave 4 the rate of attrition was halved from 9.7% to 4.7%. This positive trend, though, couldn't be continued in wave 5, in which the rate of attrition increased by 1.7% up to 6.4%. Here it should be taken into account, however, that in wave 5 with the main survey topic, "Assets & Liabilities", a very delicate subject matter was dealt with. In the 6th and especially in the 7th wave this negative effect is visibly diminished.

More difficult to survey, however, are the new households (see middle row in Table 7) which are often established by young people after moving out of the parental home, although here too the rate of attrition could be substantially reduced from 29.2% (wave 2) down to 17.6% (wave 4). Then in wave 5 a re-increase of 6.1 percentage points up to 23.7% takes place. So the negative effect of the main survey topic "Assets & Liabilities" was much more considerable in the new-household group than with the members of repeatedly-surveyed households.

Table 7: Rate of Attrition in % of the Contacted Households after Participation in the Previous Year

	Wave 2		Wave 3		Wave 4		Wave 5		Wave 6		Wave 7	
	N	%	N	%	N	%	N	%	N	%	N	%
Participant from previous year	5 742	9,7	5 235	8,3	5 019	4,7	4 937	6,4	4743	5,8	4630	4,8
New household	195	29,2	238	21,4	182	17,6	194	23,7	183	16,9	198	23,2
Temporary drop-out in previous year	-	-	259	59,5	197	52,8	154	71,4	169	57,4	154	49,4

N = Total number of cases (contacted households)

% = Percentage of households without completed interview

The positive trend in the development of the rates of response for the new households in the 7th wave couldn't be continued, however. Here the rate of attrition climbed again up to 23.2 vH. Newly-established households - and also "old" households that have moved, by the way - present a subgroup with certain participation-motivation problems, cf. Rendtel (1990). These respondents don't always grasp the reasons for continuing to participate in the survey in a new household context or at a new residence.

Although at the beginning of the SOEP it was only possible to survey institutionalized residents in the immigrant sample - because the addresses were taken from personal registers - it was hoped at first that the panel sample - via follow-ups - would, so to speak, grow into the institutional area in the succeeding waves. It turned out, though, that in the panel survey the crossover into the institutional area (e.g. homes for the elderly) was practically non-existent. In the German sample the number of institutionalized residents stagnated at approximately 30 persons, while in the immigrant sample this number drops steadily from 47 persons (wave 1) down to 19 persons (wave 7). This is chiefly due to the anticipated decrease in the number of foreigners living in workers' dormitories. So although there are hardly any persons with interviews in the institutional area in the sample, the point of change can be exactly determined and can therefore be analyzed as well.

A renewed surveying of (temporary) drop-outs is not a common procedure. All the same, through these efforts approximately 50% of the temporary drop-outs could meanwhile be re-recruited for the survey. These households would have otherwise left the study for good. So in each wave a reduction of the sample by about 80 households is avoided.

4.1.2 Results of the Sample Development

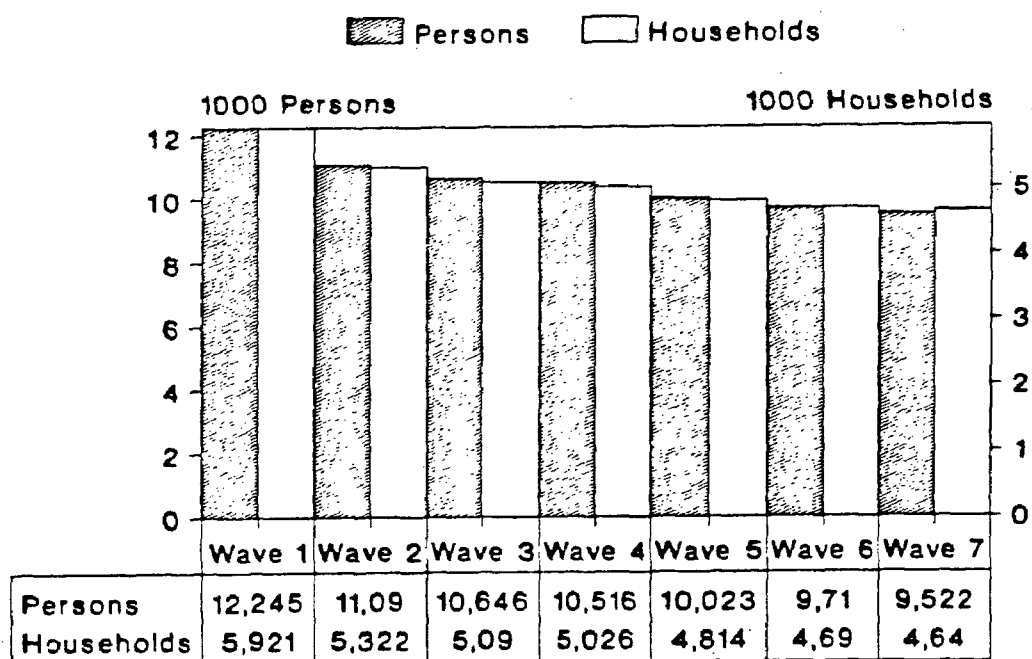
Figure 2 demonstrates how the case numbers for interviewed individuals and households have developed. On the whole, a stabilization of the panel case numbers is clearly discernable. In wave 7 interviews were completed for 4,640 households and 9,522 respondents. This corresponds to 78.3% or 77.8% of the initial volume of wave 1.

When observing the persons in the SOEP households, a distinction should be made between all of the household members and the personally-interviewed persons (respondents). To be sure, a great deal of information is only available on the respondents, and of course subjective indicators can only be taken from directly surveyed persons. Nevertheless, with the aid of the household questionnaires central (longitudinal) indicators for non-interviewed children are also obtained. Information on

a total of 16,205 persons from the SOEP households in wave 1 is available. In wave 7, with 12,253 persons, 75.6% of the initial volume still remains.

Figure 2

Number of Interviews in the SOEP



Source : German SOEP , Wave 1 to Wave 7

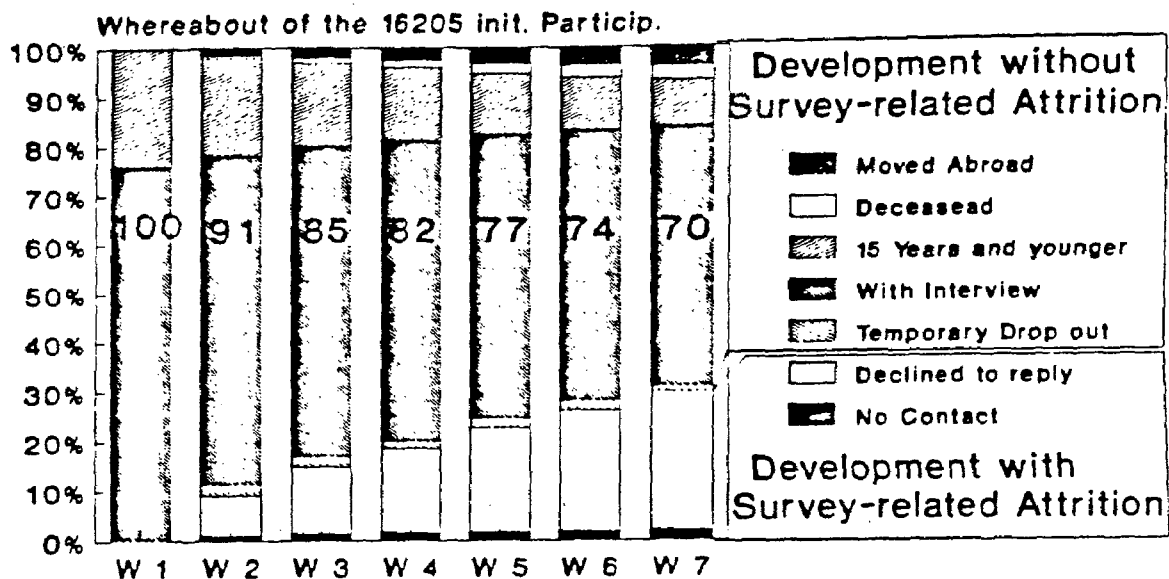
Looking now at the respondents alone, Figure 2 shows a nearly parallel development for the household and the individual blocks. It would at first have been expected that the individual graph would develop more unfavorably than the household one. This effect appears when household interviews can be realized for a considerable number of households but not all respondents are willing to be interviewed. It can be demonstrated, however, that the willingness to supply information is to a great extent identical on the household and the individual levels, cf. Rendtel (1990).

In the development of the SOEP samples A and B attention should be drawn to the fact that with sample B significant population mobility is recorded which in its net effect implies a reduction in the target sample because since 1984 more foreigners, who are represented by sample B, have left the FRG than have moved there. 565 persons in the panel sample have left the FRG while 210 persons who have come from abroad to join their families have entered the sample. To date only 122 of these persons have been queried, since half of them are less than 16 years of age.

For longitudinal evaluations the development of cross-sectional sample-sizes is less important than the percentage of persons who are in the sample stock during the panel running-time. It is natural that especially the percentage of those who've been surveyed in all of the waves decreases through demographic movement (death, moves abroad). In figure 3 the development of the stock comprised of the 16,205 initial persons from wave 1 is shown. For 70% of the initial stock, complete longitudinal information is available on the first seven waves. Up to wave 7 the losses due to demographic reasons comprise 6.9% of the initial stock. The losses due to survey-related attrition are four times as high (29.8% of the initial stock) as the losses due to demographic change. This development is, with regard to the "analyzability" of the SOEP, in no way cause for concern; neither in regard to the representativeness nor in regard to the analyzable absolute case numbers. The representativeness of the results attained by the SOEP is achieved by weighting the data (cf. Section 5). The case numbers themselves are still sufficient in the cross section as well as in the longitudinal section. If there are small sample problems for certain sub-groups, which are expressed in a lack of significant results, then these are problems which because of the total case number were already inherent in the 1st wave.

Figure 3

Development of Sample Size Basis : All Initial Participants from the 1st Wave of the SOEP



Source : German SOEP , Wave 1 to Wave 7

In appraising causal-analytical longitudinal models it should be kept in mind that more and more methods are developed which evaluate the "incomplete cases" (complete information analysis) in addition to evaluating those cases for which information from all waves is available (complete case analysis). One also speaks of evaluating "unbalanced panels". For an econometric application the estimation of an earning-equation is taken as an example, see Löwenbein and Rendtel (1991) as well as Licht and Steiner (1991). Appraising longitudinal models with "unbalanced panel" methods becomes more urgent as the panel continues because the data stock increasingly loses its rectangular file shape. Figure 4 underscores this development. In the top section the gradual merging of the volume of the 12,245 respondents from the first wave is demonstrated. In doing this all of the drop outs are included together ("eliminated" = demographic plus survey-related drop-outs). However, temporary drop-outs are not accounted for (cf. figure 3 for their volume).

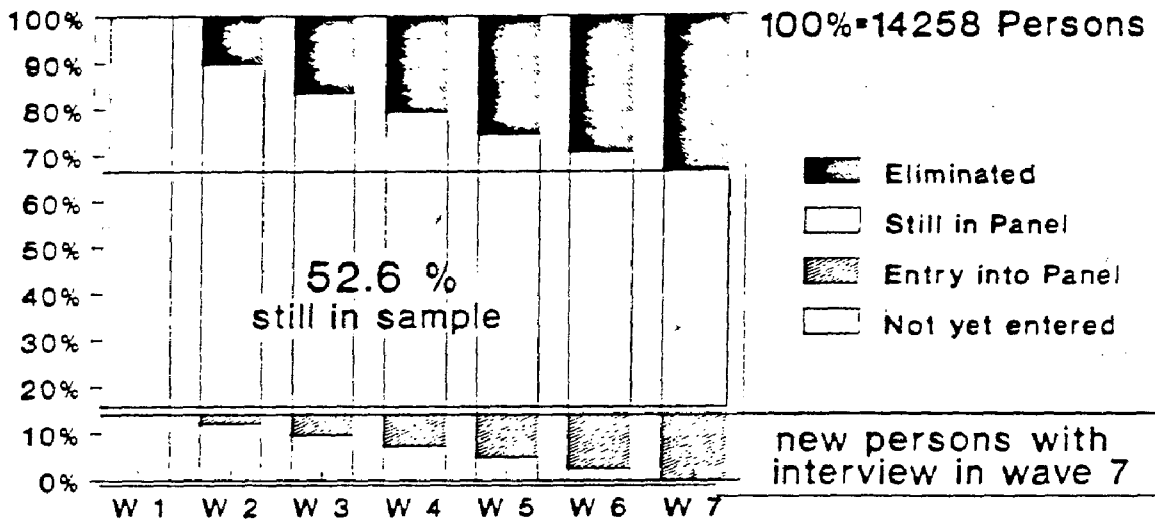
Of the 12,245 respondents from the 1st wave, a total of 61.2% are still in the sample by wave 7. In the figure this group comprises exactly 52.6% of the surveyed total of 14,256 persons. Queried simultaneously in the 7th wave were 2,013 persons who were not yet of respondent age in the 1st wave or who weren't taken into the panel until later. In the bottom section of figure 4 we can see how long these persons have been answering questionnaires. The data set of the panel now begins to assume a rhombus-like form. By limiting the evaluation to persons who have taken part in all of the surveys, the records from only 7,506 or 14,256 (= 12,245 + 2,013) persons are analyzed. This corresponds to the rate of 52.6% as shown in figure 4. The relationship becomes even more unfavorable if firstly the 611 persons who entered the panel sample between waves 2 and 6 and then left again before the 7th wave are included, and secondly if only persons with no temporary drop-out are accounted for in the framework of the complete case analysis. This reduces the number of persons by another 415. The share of persons analyzed in this fashion amounts to no more than 47.7%.

The best-known method for analyzing incomplete histories is the event-analysis method (cf. Hujer & Schneider and Ott & Poetter as well as numerous other articles by Hujer et al. 1991). For event-based analyses the SOEP offers, despite sinking case numbers, a steadily increasing number of analyzable events for each wave. This development is clearly shown in figure 5. The purely linear ascent of all of the events is a trend of particular interest. This development also makes prognoses for the case numbers of future waves possible.

Although a great number of episodes aren't classic cases of "independent events", because they're occurring with one and the same person, the steadily increasing number of spells does reinforce the "power" of the panel data.

Figure 4

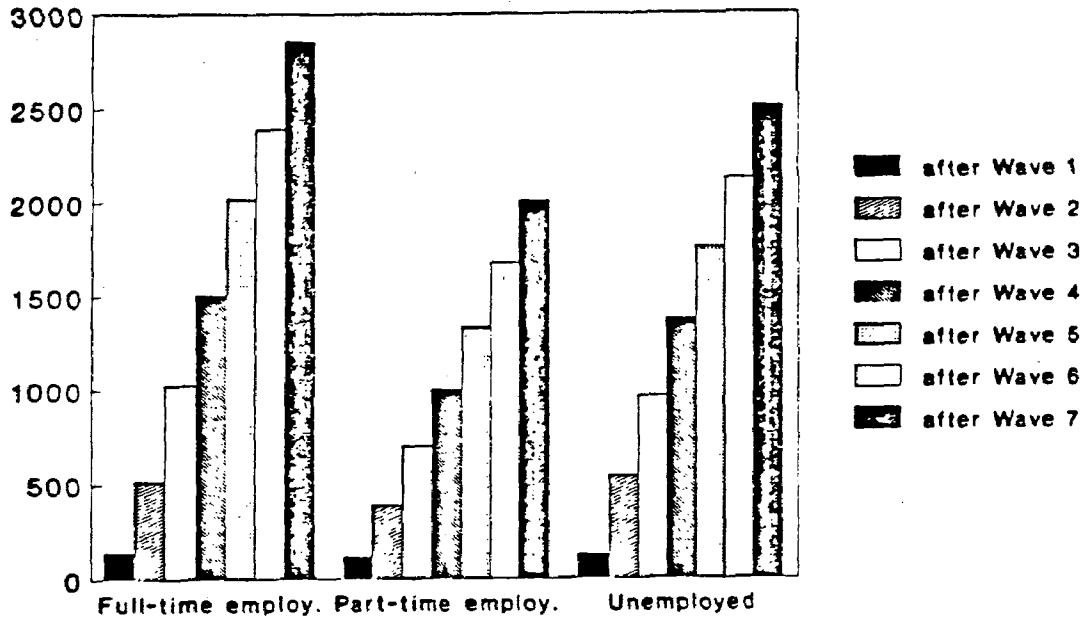
Development of the Sample Comparison: 1st Wave Respondents with New Respondents up to the 7th Wave



Source : SOEP Wave 1 to Wave 7

Figure 5

Number of uncensored Spells in SOEP Development up to Wave 7



Source : German SOEP , Wave 1 to Wave 7

4.2 *Sample C*

To date only preliminary information from wave 2, which was carried out in the first half of 1991, is available for sample C. The picture presented is a very gratifying one:

After the very good sample response rate (70%) in the 1st wave (East), the prospective situation in March in the new Bundeslaender for carrying out empirical surveys had changed. The downswing in the public mood which had been reported by a number of institutes dampened the optimism in regard to participation in the SOEP in 1991. The field work is not yet completed at this date (June 15, 1991); final results can not yet be given at the present time. However, the completed household data, most of which has already been delivered, are now in the recording and editing stage.

The interviewers' impressions as well as the written reactions of the respondents confirm the current difficulties in carrying out empirical studies on socio-economic themes. On the whole, though, the change in the peoples' mood doesn't appear to have a fundamentally negative effect on the sample utilization. For the "old" households at the old addresses the current projections from Infratest are working from a mortality rate of maximum 10%. This would even be a mild improvement on the results from the 2nd wave (West) in 1985.

What is proving to be more difficult in the new Bundeslaender is surveying new households, whereby address research is usually necessary. The surveying time in these cases is considerably longer than in the old Bundesländer. Here presumably the western results - a 63% response rate for "new" households - will not be attained, at least not in the scheduled field time, which ends at the beginning of July. At present extra efforts are being made to make this difficult collaborative process run more smoothly, and if necessary the surveying will be continued after the official termination date for the field work.

All in all, though, the field results for sample C must be given an exceptionally positive rating because it's highly probable that the sample response rate for the 2nd wave in the East will be at least as good, if not slightly better than it was for sample A. This means that after two waves the cumulative willingness to participate in sample C is 10% higher than was the case in sample A.

4.3 *Interviewer Sample*

With regard to the development of the interviewer samples, a report can only be given on the interviewers from samples A and B; longitudinal results for the interviewer staff from sample C are not yet available.

The interviewer-deployment procedure described in Section 3.4.2 has led to a drastic reduction in the number of interviewers in the course of the SOEP operations.

A small number of new interviewers are incorporated into the project per wave. These new interviewers' rates of success are generally somewhat lower than those of the interviewers who've already been with the project for some time.

631 interviewers were originally deployed. In wave 7 only 280 remain. The average number of household interviews climbed from 9 up to 17 per interviewer.

The majority of the households are still tended personally by the interviewers, and the predominant surveying method is still the oral interview. The amount of alternative interview methods, however, grows consistently, though marginally, from wave to wave (cf. Table 5 above).

The monitoring of the interviewers which occurs when a household is surveyed again has also enabled us to estimate the minimum percentage for cases of deception by professional interviewer staffers. The results of the 2nd wave brought to light that 9 household interviews in the 1st wave obviously must have been forged. Four interviewers were involved, i.e. 0.6% of all of the interviewers. It's presumed that in ordinary cross-sectional surveys the share of forged interviews is greater, since the interviewers are less qualified and also don't have to fear being controlled by means of 100% re-surveying.

5 *Weighting Procedures*

The goal of any sample is to draw conclusions from the sample and apply them to the "recorded" target population. A projection of the sample cases is required in order to be able to infer the case numbers of the target population. If it's not a simple (non-stratified) random sample, weighting the sample data will be necessary in order to diagram the structures of the target population. This is also necessary for the SOEP.

Chosen for the marginal adjustment of the 1st wave samples A and B were the characteristic combinations shown in the left block of Overview 1, which affix a total of 316 restrictions to the projection results. For the frame adjustment for sample C only person-related and regional restrictions could be given because for 1990 no reliable household-structural data were available from the GDR. With the projection of the 115 characteristic combinations shown in the right block in Overview 1 a plausible household structure was arrived at (cf. Pischner 1991).

Overview 1: SOEP Projection Frame

Sample A and B (1984)	Sample C (1990)
<u>Private households</u>	
Sex of household's head	-
Age of household's head	-
Household size	-
Nationality of household's head	-
<u>Residential population in private households</u>	
Sex	Sex
Age	Age
Marital status	Marital status
Nationality	-
<u>Schoolchildren total</u>	
Sex	-
Type of school	-
Nationality of household's head	-
<u>Employed persons in private households, Sample A:</u>	
Self-employed persons and their assistants according to:	
Sex	-
Age	-
Agricultural profession (ISCO=6)/other	-
Self employed persons:	
Sex	-
Age	-
Occupation (main group ISCO)	-
<u>Employed persons in private households, Sample B:</u>	
Nationality	-
Sex	-
Age	-
<u>Regional Distribution</u>	
-	GDR districts

The major difference between a panel survey and a series of independent cross-sectional surveys is that the panel's preceding waves provide the initial stock for the following survey wave.

It is quite clear that special survey features of this sort must be taken into account in weighting the procedures (cf. also Ernst 1989). This means that the single waves of a panel survey should not be treated like a series of independent cross sections. Neither would an approach of this sort supply an answer to the problems involved in weighting longitudinal sections.

Rendtel (1991b) provides details on this approach and on empirical experience with it.

Generally speaking, the goal of a projection can be described as estimating the incidence of interesting characteristic combinations in the target population from the sample. The estimation on the inverse selection probabilities is based on the randomizing approach. In this approach the characteristic features Y_i of the individual units of the population are regarded as non-random. The only random factor is the selection of the individuals. The random variable C_i indicates here whether the unit i belongs to the sample ($C_i = 1$) or not ($C_i = 0$).

If \hat{Y} is determined by a linear estimating function of the form

$$\hat{Y} = \sum_{i=1}^N \alpha_i C_i Y_i$$

then unbiasedness of \hat{Y} requires:

$$\alpha_i = 1/P(C_i = 1)$$

This gives:

$$\hat{Y} = \sum_{i=1}^N \frac{1}{P(C_i = 1)} C_i Y_i = \sum_{i=1}^n \frac{1}{P(C_i = 1)} Y_i$$

The procedure of sampling in a panel survey can be described as a multi-phase process. The sample for a longitudinal section over T panel waves is considered to be a selection process with $2T$ steps, which can be described as follows (the index i for the sample units has been omitted in order to simplify the notation):

- Step 1: Design sample (setting up the sample) $P(D = 1)$
- Step 2: Response in the first wave $P(R_1 = 1 | D)$
- Step 3: Contact successfully established in the second wave
 $P(K_2 = 1 | D, R_1)$
- Step 4: Response given in the second wave
 $P(R_2 = 1 | D, R_1, K_2)$

- ...
- ...
- ...
- Step $2T$: Response given in the T^{th} wave
 $P(R_T = 1 | D, R_1, K_2, \dots, R_{(T-1)}, K_T)$

The probability of selection $P(C = 1)$ for the whole sampling process over all the individual steps is given by the product of the individual probabilities:

$$\begin{aligned} P(C = 1) &= P(D = 1, R_1 = 1, K_2 = 1, \dots, R_T = 1) \\ &= P(D = 1) \cdot P(R_1 = 1 | D) \\ &\quad \cdot P(K_2 = 1 | D, R_1) \\ &\quad \cdot P(R_2 = 1 | D, R_1, K_2) \\ &\quad \cdot \dots \\ &\quad \cdot P(R_T = 1 | D, R_1, K_2, \dots, R_{T-1}, K_T) \end{aligned}$$

So the problem of weighting longitudinal sections is thus reduced to calculating the start probabilities of taking part in the 1st wave of the panel and determining the probabilities of remaining, i.e. probabilities in each respective selective stage of remaining in the panel.

For cross-sectional evaluations the inclusion of the non-initial persons is no problem as long as the sample probabilities for households in the year in question are known or can be estimated. Nor does this pose problems for the (since 1990) simplified follow-up concept by which non-initial persons are also continually drawn into the panel (cf. Huang 1984). On the contrary: the PSID concept of only following initial persons and omitting non-initial persons from the weighting (i.e. these persons are assigned the weight 0) leads to the problem that the weighting results between the household plane and the individual plane are no longer consistent with each other. At this point it becomes clear that the demographically-oriented concept for the initial persons over the course of time is not identical with the concept of the population in private households. This identity is, strictly speaking, only valid for the starting wave of the panel.

Households with newly-arrived members have a higher chance of being selected; this is also valid for the original follow-up concept. Because the selection probabilities of households are arrived at solely by the selection probabilities of its members at the start of the panel as well as the follow-up rules, households with new arrivals have higher chances of selection than households with none (because there were at least two paths by which they could be reached). As a consequence households with new arrivals have to be assigned a lower weight. If one applies the household weight to all household persons (i.e. non-initial persons too), this lower weight compensates for the increase in case numbers caused by the new arrivals. The following-up adhering to the initial-person concept implies - in contrast to the simplified follow-up rules - a wealth of very confusing exclusion rules for the individual "paths" of the persons between the households and these are very difficult to implement for the weighting procedure. Finally, a following-up adhering to the initial-person concept removes entirely from the panel sample persons with a high rate of mobility who actually don't have to be assigned more than a low weight as immobile persons.

If one regards the entire weighting scheme with the practical eye of someone who wishes to evaluate the SOEP data, this means that per wave for each household in the data set a cross-sectional and a longitudinal weighting factor are made available. The cross-sectional factor furnishes for the entire sample the valid values for the survey year in the target population. The longitudinal factor indicates how the panel dynamic

changes the selection probability of each household. To arrive at the correct weight for longitudinal analyses in the course of multiple waves, the longitudinal factors should only be multiplied by each other.

6 Data Structures

6.1 The Logical Concept of Data Processing

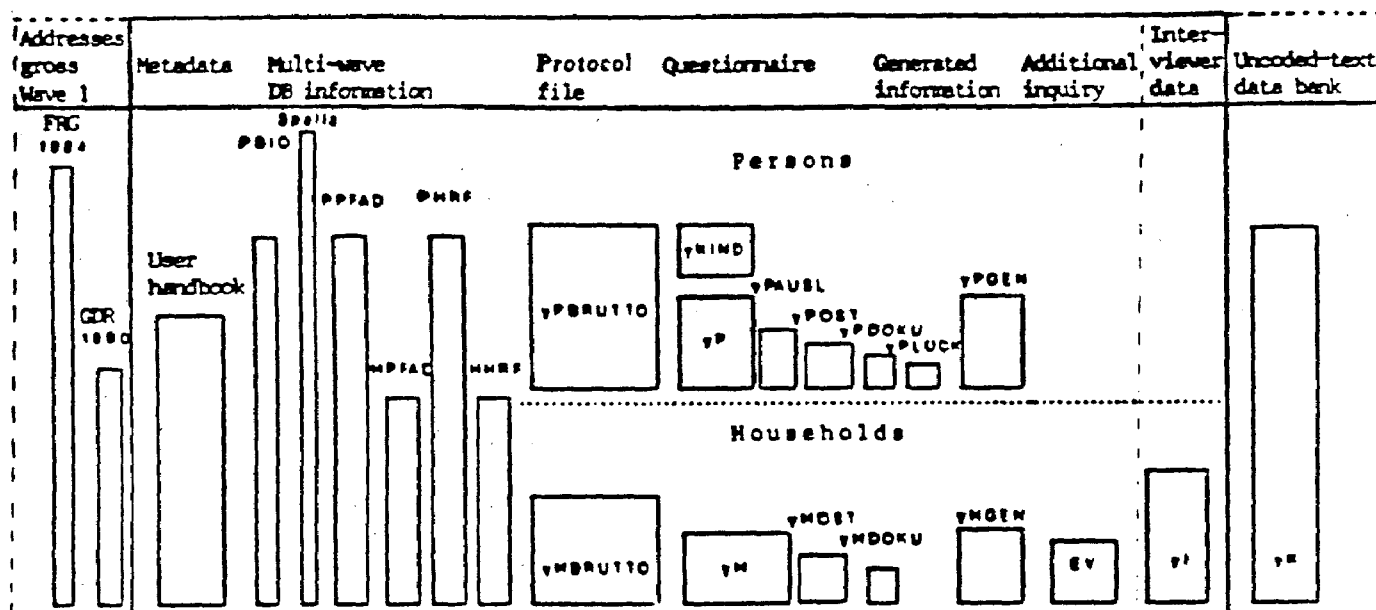
In the course of a panel study like the SOEP a large amount of data is accumulated that must be assigned to various hierarchical planes. Thus the data is stored within a data base system. Figure 6 shows all machine-readable data types that exist for the SOEP. The data that are distributed are contained in the rectangle inside the illustration. The data types outside are in the truest sense of the word of peripheral importance. Additionally the uncoded texts are extremely sensitive data in regard to the data privacy laws and can only be evaluated within the DIW in co-operation with the Panel Group.

The data are processed with a data base system that simplifies the combining of data belonging to different hierarchical planes as well as the combining in the course of time (cf. Frick et al. 1991)¹⁵. In the social sciences there is some dissension about retaining data in data base systems since in this case the data analysis requires an additional human capital investment of the researcher in order to learn how to operate the corresponding system. But the dissemination of very large rectangular files containing only those interviews gathered in the ongoing wave seems to the user only at first glance to be the simplest method. They don't have to learn any of the - still complicated - data base languages and can extract the interesting cases and variables from the "master super-file" without undue mental strain (cf. Solenberger et al. 1989 on this concept). Some "hidden" costs turn up, however:

- The analyzing of and forming prognoses for panel mortality is possible only with the help of additional files; but they are in fact usually omitted. Thus it is not possible to model the attrition process with the rectangular file which could bias the results.

¹⁵ For alternative data base concepts see Engel (1991), Brecht (1990) and David (1989), as well as Schmaus (1990) for a SPSS solution which admittedly is sub-optimal for large panels like the SOEP.

Figure 6 SOEP Data Base Tables



*Sample survey C (GDR or East Germany) has been fully integrated into the data bank. Up to 1989 the cases under this sample survey are marked in the multi-wave DB information (PPFAD and HPFAD) as "not yet called". The uncoded-text data bank has not been integrated into the actual SOEP data bank for reasons of data protection. Evaluation of uncoded-text data is possible only in the framework of special agreements with DIW.

Legend:

- y: Running index for wave A (=1984) to up to, currently, G (=1990)
- PBIO: Person-related biographical information
- Spells: Monthly activity and earnings calendars
- PPFAD, HPFAD: Person- respectively household-related participant information
- PHRF, HHRF: Person- respectively household-related weighting factors
- yPLUECKE: reconstructed person information in case of participant gaps
- yPBRUTTO, yHBRUTTO: Person- and household-related information from the address protocol
- yH: Household-related standard information from the household questionnaire
- yP, yKIND: Person-related standard information from the person questionnaire (yP) respectively from the household questionnaire (yKIND)
- yPAUSL: Person-related additional information for foreigners from sample B
- yPOST: Person-related additional information for the East-sample C
- yPDOKU, yHDOKU: Person- respectively household-related documentation information on corrected variables
- yPGEN, yHGEN: Person- respectively household-related generated variables
- GHOST: Household information for the Eastern sample C (y=G, for the year 1990)
- yEV: Asset balance of the households in wave 5 (y=E, for the year 1988)
- yI: Interviewer information
- yK: Person-related uncoded text from the open questions of the questionnaire

- Since, despite all of progress made in data-processing, the extraction of the research population from the large rectangular file is time-consuming. New runs are avoided since it can be somehow justified. Therefore some additional variables that would have enhanced the analysis were excluded in many studies. They are admittedly without being noticed but obviously these are hidden costs.
- The analysis is reduced as a rule to a matter of individual observation, although assigning context information from household members would be possible and for many research questions most valuable. This reduction is the source of another kind of hidden costs.

We therefore hold the decision to keep the SOEP data within a data base management system and transmit the data in data base format to be correct and forward-looking¹⁶. Flexible evaluations that employ the full information content of a household panel are only possible with a data base management system.

6.2 Logical Structure of the Variables

During the processing of the survey data itself the decisive problem crops up of whether the questionnaire should be displayed in a one-to-one relation in the data base or if a logical structure that follows the typical user's wishes should be selected instead of the structure geared to the respondents. The data SOEP's bank was originally conceived as a one-to-one copy of the questionnaire; the concept was then modified, however, because of diverse analyzing experiences:

A basic problem with handling panel data is that the survey instruments follow psychological considerations that enable the interview to run as smoothly as possible. These determine the questioning sequence and certain questions from repeated surveys don't have to be asked every year. If, for example, the amount of education completed has been established, it needs only be asked about again if a filter question when there has been a change in the previous year is answered affirmatively. A one-to-one copy of the questionnaire then means that for many persons the amount of education completed is not contained in the data from the wave in progress. The educational level (of these persons) must first be reconstructed from the data from preceding waves. For most data users this is clearly sub-optimal. In order to attain an user friendly data file it is therefore necessary to supply generated variables with processed information. In particular so-called status variables which, for example, would contain the completed

16 The data from the PSID study were originally disseminated as a purely rectangular file of the cases realized in the last wave. This makes it impossible to analyze attrition processes and to correct interview-related distortions. Without a data base, the non-respondent file which has been distributed additionally for some years can be assigned to the central file only with a great deal of effort. This is why Beckett et al. (1988) stress the fact that analyzing the attrition process of the PSID was a "Herculean task".

amount of education in every wave although it would not be asked for with every respondent.

The logic of the data base design for SOEP thus follows cross-sectional rules to make life as easy as possible for users. Of course for longitudinally-based analyses a vast array of variables are made available to the user in order to simplify individual and household combinations in the course of several waves. For this purpose "meta data" were designed (the tables are called PPFAD and HPFAD) which provide for each person or household that has ever made an appearance in the course of the study a data set whose attribute (variables) feeds information on the accompanying logical keys into the data base. With the aid of this data the fate of each person and each household can be easily traced over the course of time.

Experience with the SOEP has shown that the generation of "status variables" and "meta data" is more important for the user than increasingly meticulous "error clean-ups" that can never cease anyway. With complex data sets one will detect inconsistencies repeatedly if a new formulation of a question is introduced into the data. In the end the user has to do the necessary clean-up work himself anyway¹⁷.

6.3 Data Distribution and Data Protection

The information content of data as complex as the SOEP data can not be realized completely by one research group alone. Therefore dissemination of the data to all independent scientists at universities and research institutes is necessary to exploit the full richness of the data. This means also that some research question will be worked by several researchers. But this is not a waste of resources, because mistakes can only be uncovered and errors of serious consequence avoided through multiple research.

In general statistical numbers don't present a "one-to-one" image of reality. Instead, with the aid of theoretical constructions operational indicators are defined and measured. After completion of the survey theoretical constructions can be reconstructed from the empirical indicators. This process is a constant source of dissension and the empirical calculations are error-ridden, too. We know from the time-series analysis what difficult problems of operationalisation exist and how many errors have already been made. This scientific discussion process of time-series-analyses is possible because anyone can gain access to the time-series data from the economic national account. Complete transparency and intrasubjective verifiability are the rule. Transparency to such a degree can only be attained with panel data if the data are available in

17 It would also be naive from the standpoint of scientific theory to assume that every data user would arrive at the exactly the same answers when working with a similar type of research question. There are basically too many operational problems in detail and in the cleaning options, ruling out the possibility that different users could present identical results taken from identical micro data if the research questions are sufficiently complex.

anonymous form to any and all interested scientists¹⁸. In order to promote further evaluation, the SOEP data are disseminated free of charge for scientific purposes.

The interest in SOEP data has grown steadily in past years. As of July 1991 over 100 data-dissemination contracts had already been closed, whereby one contract often serves for a number of scientists who work with the data. It deserves to be mentioned that a strong interest in the German SOEP data is taken abroad. Since many countries have data privacy laws less stringent than those in the Federal Republic of Germany, an especially anonymous version of the data set was developed which can be disseminated to researchers in every country in the world¹⁹.

In the "public use file" of the SOEP a small number of variables were deleted which were of no great importance for most of the analyses anyway. The variables in question were the precise nationality of the foreigners in the sample (so now one merely knows that they're non-Germans), the regional affiliation of the sample household and detailed information about assets (general information is of course available).

Moreover, only a 95% random sample²⁰ from the SOEP will be disseminated for public use. This stems from the consideration that it is usually easy to reveal a respondent's identity if one knows that he is contained in a random sample. If, however, some of the cases are removed from the sample one can no longer be sure that the person in question is in this subsample, and the degree of anonymity thus becomes greater²¹.

At the present time the data are distributed as SIR export files or as raw data on magnetic tape, whose files are constructed like the data base charts. A distribution on floppy disc is also possible. Distribution with CD-ROMs is in preparation.

7 Utilization

The data of the SOEP are now being analysed by over 100 user groups (as of July 1991). Following the definition of the "public use files", the data were distributed to eleven foreign groups within six months. Outside of the Federal Republic of Germany

-
- 18 New problems come up with the cleaning of the data. Within a data base system cleaned values can be marked as such without difficulty (yDOKU table from the SOEP data base). This transparency is not to be found - for example - with data from the economic national account. The collecting and evaluation problems don't come to light until the latest "revision" of the economic national account statistics is published by the statistical bureaus.
- 19 For this purpose it was necessary to expand the "Notice on Data Privacy", which is handed out in every surveyed household to include a section on data dissemination to foreign countries.
- 20 A panel subsample should not represent an only 95% random sample of the records for each wave. In this case the longitudinal information would be destroyed. The public use sample is therefore determined by a selection of households from the first wave. The information from these particular households (and all split-offs that originated there) is disseminated wave by wave as the public use file.
- 21 The number of cases removed must be substantial. Otherwise one would be able to reconstruct the characteristics of the deleted persons by a comparison of the distribution of the entire sample and of the subsample.

the SOEP data is available in the following countries up to now: Australia, Canada, Great Britain, Luxembourg, the Netherlands and the USA.

The focal points are poverty research, the microeconomic analyses of employment opportunities and income dynamics, the development of subjective life satisfaction, and - subject to explosive growth - comparisons between West and East Germany in many areas. An increase in international comparisons is making itself felt.

For an overview of the literature, see in addition to this paper's selected bibliography particularly Krupp and Schupp (1988), Wagner (1990), Rendtel and Wagner (1991) as well as Hujer et al. (1991). For the initial results of the Eastern Sample, see Projektgruppe Panel (1991).

There are currently over 300 single publications based on data from the SOEP²². They are contained in a literature data base which is available on floppy disc. This data base can be read on any computer with an MS-DOS operating system.

22 Results are available for the following topics: Social Structure and the Quality of Life, Subjective Well-Being in East and West Germany, Life Satisfaction in West and East Germany, Situation of Lone-Parent Families, Marital Behaviour and Divorce Risks, The Welfare Position of Divorced and Widowed Women in the Federal Republic and the USA, Education Expansion and Decreasing Birth Rates, Remarriage after Divorce, Changes in Educational Opportunities, Educational Expansion and Changes in Womens' Entry into Marriage and Motherhood, Structural Residential Mobility, Formation and Dissolution of One-Person Households in the United States and the FRG, Changes of the Youth Phase, Demands on Dentists, Social Situation of Home-cared Persons, Behaviour of Health-insured Persons, Social Differences in Life Expectancy, Health Condition and Health Care, Stability and Change in the Political Parties, Time Expenditure, Distribution of Leisure Sports in East and West Germany, Prognosis for the Development of Leisure Sports to the Year 2000, Assimilation of Foreigners, Language Skills of Foreigners, Re-migration of Migrant Workers, Transition from School to Vocational Training, Further Vocational Training, Description of the Income Distribution, Employment Tenure and Earnings, Temporal Aspects of Social Inequality, Description of the Cycles of Individual Work Income and from Household Income, Poverty in Cross Section and in Longitudinal Section, Determinants of Change in Household Income, Discrimination and the Labor Market, Labor Market Participation of Women in West and East Germany, An Optimal Wage Structure in East Germany, Development of Household Income and Living Costs in East Germany, Part-time Work in East and West Germany, Marital Behaviour in East and West Germany, Non-Standard Employment, Microeconomic Analyses of Female Labor Supply, Effects of the Tax Reform in West Germany, Distributional Consequences of the Introduction of Income Tax in East Germany, Determinants of Unemployment and the Length of Unemployment, Effects of Unemployment on Wage Development, Comparison of Household and Work Income in East and West Germany, Efficiency of the Family Financial Compensation Program, Regional Differences in Work Income, Regional Mobility, Relation of Company Size to Employees' Wages, Productivity and Competitive Potential of the GDR Economy, Pathways to New Jobs, Labor Market Expectations in West and East Germany, Profit Shares and Development of Individual Wages, Re-Examination of the Linearity of the Human Capital Theory, The Situation of Home-Care Patients in the FRG, Sectoral Wage Patterns, Are There Compensating Wage Differentials?, Occupational Illness and Work Income, Determining Factors for Union Membership, Determinants of Self-Employment, Structure and Consistency of Working Time Preferences, The Influence of Inheritances on the Distribution of Wealth, Intersectoral Wage Differentials, Labor Market Segmentation in the FRG, The Earnings Function under Test, Pension Reform and Income Distribution, Preferences for a Pension Reform, Transition into Retirement, Raising Children and Old Age Pensions, Housing Patterns and Mobility of the Aged, Retirement in East and West Germany.

Selected Bibliography (mainly of english written papers)

- Becker, Irene 1989: Die Datenbestände des Sfb 3, Sfb 3 Arbeitspapier Nr. 317, Frankfurt, Mannheim
- Beckett, Sean et al. 1988: The Panel Study of Income Dynamics after Fourteen Years - An Evaluation, in: Journal of Labor Economics, 6(4), pp. 472-492
- Bird, Edward 1991: Income Variation Among West German Households, in: Rendtel and Wagner, pp. 409-436
- Brecht, Beatrix 1990: Aufbau, Struktur und Anwendung des Sozio-oekonomischen Panels in INGRES, Diskussionsbeiträge des Sfb 178, Serie II-Nr. 120, Konstanz
- Burkhauser, Richard et al. 1990: Economic Burdens of Marital Disruptions- A Comparison of the United States and the Federal Republic of Germany, in: Review of Income and Wealth, 36(4), pp. 319-333
- Burkhauser, Richard et al. 1991: Wife or Frau, Women Do Worse - A Comparison of Men and Women in the United States and Germany After Marital Dissolution, in: Demography, 28(3), pp. 353-360
- David, Martin 1989: Managing Panel Data for Scientific Analysis - The Role of Relational Data Base Management Systems, in: D. Kasprzyk, pp. 226-241
- Ernst, Lawrence R. 1989: Weighting Issues for Longitudinal Household and Family Estimates, in: D. Kasprzyk, pp. 139-159
- Esser, Hartmut et al. 1989: Mikrozensus im Wandel. Band 11 der Schriftenreihe Forum der Bundesstatistik herausgegeben vom Statistischen Bundesamt, Stuttgart
- Frick, Joachim 1990: Die SIR-Datenbank des sozio-oekonomischen Panels - Ein Tutorial zu Aufbau, Syntax und problemorientierten Anwendungen, Berlin
- Frick, Joachim et al. 1991: SIR and the German SOEP, ESF Working Paper "Household Panel Studies", Essex
- Galler, Heinz P. 1986: Eine Gewichtungskonzept für das Sozio-ökonomische Panel, Sfb 3 Arbeitspapier Nr. 204, Frankfurt, Mannheim
- Geiss, Alfons J. und Jürgen H.P. Hoffmeyer-Zlotnik 1991: Zur Vercodung von Beruf, Branche und Prestige in der DDR, in: Projektgruppe Panel (Hg.): Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in Ostdeutschland, Frankfurt, New York, pp. 139-147

- Gerlach, Knut and Ulrich Schasse 1990: On-The-Job Training Differences by Sex and Firm Size, in: *Zeitschrift fuer Wirtschafts- und Sozialwissenschaften*, 110(2), pp. 261-272
- Grohmann, Heinz 1985: Vom theoretischen Konstrukt zum statistischen Begriff - Das Adäquationsproblem, in: *Allgemeines Statistisches Archiv*, 69 (1), pp. 1-15.
- Hanefeld, Ute 1984: The German Socio-Economic Panel- in: *American Statistical Association, Proceedings of the Social Statistics Section*, Washington, D.C., pp.117-124
- Hart, Robert A. and Olaf Huebler 1991: Are Profit Shares and Wages Substitutes or Complementary Forms of Compensation ?, in: *Kyklos*, 44(2), pp. 221-231
- Hauser, Richard 1992: Einleitung, in: R. Hauser et al. (eds.), *Mikroanalytische Grundlagen der Gesellschaftspolitik - Ausgewählte Probleme und Lösungsansätze*, Weinheim (in press)
- Headey, Bruce et al. 1990: The Duration and Extent of Poverty - Is Germany a Two-Third-Society ?, WZB Working Paper No. P 90-103, Berlin: mimeo.
- Hippler, Hans-J. et al. 1990: Der Einfluß von Datenschutzzusagen auf die Teilnahmebereitschaft an Umfragen, in: *ZUMA-Nachrichten*, 27, pp. 54-67
- Huang, H. 1984: Obtaining Cross-Sectional Estimates from a Longitudinal Survey. Experiences of the Income Survey Development Programm, *Proceedings of the Section of Survey Research Methods*, American Statistical Association, pp. 676-681
- Huebler, Olaf 1989: Individual Overtime Functions With Double Correction for Selectivity Bias, in: *Economic Letters*, (29), pp. 87-90
- Hujer, Reinhardt and Hilmar Schneider 1989: The Analysis of Labor Market Mobility Using Panel Data, in: *European Economic Review*, 33, pp. 530-536
- Hujer, Reinhard et al. (eds.) 1991: *Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel*, Frankfurt, New York
- Kalton, G. 1989: Modelling Considerations: Discussion from a Survey Sampling Perspective, in: D. Kasprzyk et al., pp. 575-585
- Kasprzyk, Daniel et al. (eds.) 1989: *Panel Surveys*, New York u.a.
- Kirschner, H.P. 1984: Allbus 1980 - Stichprobenplan und Gewichtung, in: K.U. Mayer und P. Schmidt (Hg.), *Allgemeine Bevölkerungsumfragen der Sozialwissenschaften - Beiträge zu methodischen Problemen des Allbus 1980*, Frankfurt, pp. 114-182

- Krause, Peter and Gert Wagner 1991: Datenhaltung bei sozialwissenschaftlichen Panel-Studien, in: B. Engel et al., Datenbankorganisatorische Probleme und Grundlagen des NIFA-Panels - Ergebnisse eines Workshops, Arbeitspapier des Sfb 187 "Neue Informationstechnologie und flexible Arbeitssysteme", Bochum
- Krupp, Hans-Juergen and Ute Hanefeld 1987 (eds.): Lebenslagen im Wandel - Analysen 1987, Frankfurt, New York
- Krupp, Hans-Juergen and Juergen Schupp 1988 (eds.): Lebenslagen im Wandel - Daten 1988, Frankfurt, New York
- Kruskal, William und Frederik Mosteller 1979: Representative Sampling, I-III, in: International Statistical Review, 47, pp. 13-24, pp. 111-127 und pp. 245-265
- Loewenbein, Oded und Ulrich Rendtel 1991: Selektivitaet und Panelanalyse, in: U. Rendtel und G. Wagner, pp. 156-187
- Mayer, Karl-Ulrich 1990: Lebensverläufe und sozialer Wandel, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 31, pp. 7-21
- Ott, Notburg 1991: Intrafamily Bargaining and Household Decisions, Berlin u.a.
- Pietzke, Iwan-Rainer 1991: Anwendung der Informations- und Kommunikationstechnik in der Kommunalverwaltung - heute und in der Zukunft, in: Verwaltungsorganisation, 25 (1), pp. 15-20.
- Pischner, Rainer 1991: Die Gewichtung der Ost-Stichproben des Sozio-ökonomischen Panels und der Pilotstichprobe des Wohlfahrtssurveys, in: Projektgruppe Panel (Hg.): Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in Ostdeutschland, Frankfurt, New York, pp. 97-112
- Pol, Frank van de 1989: Issues of Design and Analysis of Panels, Amsterdam
- Projektgruppe Panel 1990: Das Sozio-oekonomische Panel fuer die Bundesrepublik Deutschland nach fuenf Wellen, in: Vierteljahrshefte zur Wirtschaftsforschung, No. 2, pp. 141-151.
- Projektgruppe Panel (ed.) 1991a: Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in den Neuen Bundesländern, Frankfurt, New York
- Projektgruppe Panel (ed.) 1991b: Das Sozio-oekonomische Panel - Benutzerhandbuch, Version 5/91, Berlin

- Rendtel, Ulrich 1991a: Über die Behandlung des Selektivitätsproblems bei der Auswertung von Paneldaten - dargestellt an zwei Fallbeispielen aus dem Sozio-ökonomischen Panel, in: Beiträge zur Arbeitsmarkt und Berufsforschung (144), pp. 89
- Rendtel, Ulrich 1991b: Die Schätzung von Populations-Werten in Panel-Erhebungen, in: Allgemeines Statistisches Archiv, 75 (4).
- Rendtel, Ulrich 1991c: Participation-Rates in Panel-Surveys - Influence, Confidence and Social Selection - The Development of Participation-rates in the German Socio-economic Panel, in: The German Journal of Psychology - Abstracts and Reviews, in press
- Rendtel, Ulrich 1991d: Weighting Procedures and Sampling Variance in Household Panels, Working Paper ESF Panel Working Group, Essex: mimeo
- Rendtel, Ulrich and Gert Wagner (eds.) 1991: Lebenslagen im Wandel - Die Entwicklung der Einkommen seit 1984, Frankfurt, New York
- Riebschläger, Marlis und Gert Wagner 1991: Interviewerstab und Interviewereffekte in der DDR-Basisbefragung des Sozio-ökonomischen Panels, in: Projektgruppe Panel (ed.): Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in den Neuen Bundesländern, Frankfurt, New York, pp. 127-138.
- Schmaus, Guenther 1990: Organisation der Daten des Luxemburger Haushaltspanels, Document PSELL No. 17, Walferdange/Luxemburg
- Schnell, Rainer 1991: Wer ist das Volk?, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43(1)
- Schnell, Rainer, Paul B. Hill und Elke Esser 1988: Methoden der empirischen Sozialforschung, München, Wien
- Schupp, Juergen 1985: Erfahrungen mit dem Aufbau einer SIR/DBMS-Datenbank beim Sozio-ökonomischen Panel, in: Status GmbH (Hg.), Datenbank SIR/DBMS Benutzerkonferenz, Berlin, pp. 117-142.
- Schupp, Jürgen 1991: Zur Durchführung des Sozio-ökonomischen Panels, in: Herget, Hermann (Hrsg.), Chancen von Panelerhebungen und zeitbezogener Analyse für die Berufsbildungsforschung, Berichte zur beruflichen Bildung, Heft 124, Berlin und Bonn, pp. 111-127
- Schupp, Juergen and Gert Wagner 1990: Die DDR-Stichprobe des SOEP, in: Vierteljahrshefte zur Wirtschaftsforschung, No. 2, pp. 152-159.

- Schupp, Juergen and Gert Wagner 1991: Die Ost-Stichprobe des Sozio-oekonomischen Panels - Konzept und Durchfuehrung der "SOEP-Basiserhebung 1990" in der DDR, in: Projektgruppe Panel, pp. 25-41.
- Solenberger, Peter et al. 1989: Data Base Management Approaches to Household Panel Studies, in: D. Kasprzyk, pp. 190-225
- Stahl, Konrad 1987: Housing Patterns and Mobility of the Aged- The United States and West Germany, in: D. Wise (ed.), The Economics of Aging, Chicago.
- Wagner, Gert 1990: Das Sozio-ökonomische Panel - Ein Instrument zur Dauerbeobachtung privater Haushalte, in: H. Rapin (Hg.) Der private Haushalt im Spiegel sozialempirischer Erhebungen, Frankfurt, New York, pp. 93-113
- Wagner, Gert 1991a: Die Erhebung von Einkommensdaten im Sozio-ökonomischen Panel (SOEP), in: U. Rendtel und G. Wagner (Hg.): Lebenslagen im Wandel - Zur Einkommensdynamik in Deutschland seit 1984, Frankfurt, New York , pp. 26-33
- Wagner, Gert 1991b: Indicators to Characterizing the Stability of Panel Studies - The Case of the German SOEP, ESF Working Paper "Household Panel Studies", Essex
- Wagner, Gert und Jürgen Schupp 1990: Das Sozio-ökonomische Panel im sich einenden Deutschland, Sfb 3 Arbeitspapier Nr. 326, Frankfurt, Mannheim
- Witte, James C. und Herbert Lahmann 1988: Formation and Dissolution of One-Person Households in the United States and Waest Germany, in: Sociology and Social Research, (1), pp. 31-41
- Witte, James C. 1990: The Potential of Comparative research Using Data from the U.S. Survey of Income and Program Participation (SIPP) and the German Socio-Economic Panel (SOEP), DIW Discussion Paper No. 14, Berlin: mimeo