# Germany 1984: Survey Information

## Summary table

| | |
|---|---|
| **Generic information** | |
| Name of survey | German Socio-Economic Panel (GSOEP) / *Das Sozio-oekonomische Panel(SOEP)* or *Leben in Deutschland* |
| Institution responsible | Deutsches Institut für Wirschaftsforschung (DIW) |
| Frequency | Annual |
| Survey year / Wave | 1985 - Wave B (#2) |
| Collection period | 8 months in 1985 |
| Survey structure | Cross-sectional and longitudinal |
| Coverage | Private households in the whole territory |
| Geographic information | Province *(Bundesland)* |
| Files delivered | Several files at different levels and with different samples |
| **Sample size** | |
| Households | 5,322 interviewed households (LIS dropped about 100 hhlds because of individual members non-participation or zero weight) |
| Individuals | 16,363 individuals in interviewed households, of which 11,090 where interviewed (i.e. individuals aged 17 and over who participated) |
| **Sampling** | |
| Sampling design | *Initial samples:* there are 6 different samples, all multi-stage random samples, which are regionally clustered (around federal states, administrative districts and type of community). The respondents (households) are selected by random-walk. *Follow-up concept:* old households with old and new persons (births and moved in) are followed up as well as new households with old (moved out) and new persons (births and split-offs) |
| Sampling frame | 1982 ADM master tape for sample A, immigrant registration records and local residents' registration lists for sample B, central residents' file of the GDR for sample C |
| **Questionnaires** | The GSOEP survey instruments include: household head schedule, schedule for each individual over 16, schedule for foreigners and address record (in order to follow people through the panel) |
| **Standard classifications** | |
| Education | 8 different national categories plus two for Eastern German diplomas and one residual category |
| Occupation | 4-digit ISCO-88 standard |
| Industry | 2-digit NACE-standard |
| **Income** | |
| Reference period | Calendar year 1984 |
| Unit of collection | Mostly individual, some income sources at the household level |
| Period of collection | Mostly monthly income with number of months in 1984, some yearly. |
| Gross/net | Variables are recorded gross of taxes and contributions, which are also imputed separately (CNEF) |
| **Data editing / processing** | |
| Consistency checks | Extensive |
| Weighting | Different longitudinal and cross-sectional weights (both at the individual and household level) to correct for the different sampling probabilities of the subsamples, for non-response (non-willingness to participate in the first wave) and attrition in the subsequent waves |
| Imputation | The Cross-National Equivalent File (CNEF) includes completely simulated taxes and contributions (on the basis of a microsimulation model – the Schwarze routine) and fully imputed missing income information due to non-response. |

This document draws extensively upon the following document: "DTC Companion to the German Socio-Economic Panel Study (GSOEP)", edited by John P. Haisken-DeNew and Joachim R. Frick, DIW Berlin, Version 6.0 - December 2002, Updated toWave18 (R).

**Table of contents:**



## A. General characteristics

Official name of the survey/data source:
German Socio-Economic Panel (GSOEP) / *Das Sozio-Oekonomische Panel (SOEP)*

Administrative Unit responsible for the survey:


GSOEP DIW-Berlin
Mailing address: D-14191 Berlin Germany
Street Address: Koenigin-Luise-Strasse 5, D-14195 Berlin, Germany
General Information:
Fax:+49 /30 /89789-109 (Attn:GSOEP Secretary)
Fax:+49 /30 /89789-200 (Attn:GSOEP Secretary)
WWW: http://www.diw.de/gsoep/
Hotline: soepmail@diw.de


The GSOEP was started in 1984 as a longitudinal survey of private households and persons in the Federal Republic of Germany. The central aim of this panel study is to collect representative micro-data on persons, households and families in order to measure stability and change in living conditions by following principally a micro-economic approach enriched with sociology and political science variables, mainly determined by the Social Indicator movement.

A rather stable set of core questions is asked every year covering the most essential areas of interest of the study: population and demography, education, training, and qualification, labor market and occupational dynamics, earnings, income and social security, housing, health, household production, basic orientation (preferences, values, etc.) and satisfaction with life in general and certain aspects of life. Additionally, as a yearly topical module, the basic information in one of these areas is enlarged by detailed questions.

In order to measure change and stability across time, the GSOEP-questions are targeted at different dimensions of time (past, present and future) using also different measurements

of time (information at a given point of time, periodical information, calendar information, life history information).

## B. Population, sampling size and sampling methods

Coverage

In order to start the survey in early 1984, the original GSOEP sample was drawn in 1983. The target population to be represented by the GSOEP is defined as the residential population of the FRG in 1984 including West Berlin.

In the FRG, selected foreigner groups were oversampled in the study. The sampling probability for the eastern sample is also larger than the probability for the main sample in West Germany. Those different sampling probabilities were chosen to make sure that the number of cases in the sample are large enough for analyses of the three samples on their own. The institutionalized population, in the true sense of the word (hospitals, nursing homes, military installations) was not representatively included in the 1st wave. Later, however, persons from the initial households who had taken up residence temporarily or permanently in institutions of this kind were followed.

GSOEP samples

*Sample A "Residents in the FRG"* - Covers persons in private households with a household head who does not belong to the main foreigner groups of Guestworkers, i.e. household heads who are Turkish, Greek, Yugoslavian, Spanish or Italian). Because only a few foreigners are in Sample A it is often called the West German Sample of SOEP. In 1984 it covered 4528 households (4298 in the 95% Scientific Use Version) with a sampling probability of about 0.0002.

*Sample B "Foreigners in the FRG"* - Covers persons in private households with a Turkish, Greek, Yugoslavian, Spanish or Italian household head. Compared to Sample A the population of Sample B is oversampled and started with 1393 households (1326 in the 95% Scientific Use Version). The sampling probability was about 0.0008.

Sample size

The table below shows the starting samples sizes of the samples A and B for the first survey year (1984). Already in the second wave, the number of successful interviews has reduced. The reduction in the population size for both samples is mainly the result of person-level drop-outs, refusals, moving abroad, etc. However, due to new persons moving into already existing households, and children reaching the minimum respondent's age of 16, and thereby increasing the sample size, this negative development is onset somewhat.

Starting sampling size

| Sample | Year | Households (net) | Persons (gross) | Respondents (net) | Children (net) |
|---|---|---|---|---|---|
| | | 100% sample | | | |
| A and B | 1984 | 5921 | 16205 | 12245 | 3915 |
| | | 95% sample | | | |
| A and B | 1984 | 5624 | 15397 | 11610 | 3711 |

However, this cross-sectional view is insufficient when examining the longitudinal development of the sample, which is in?uenced by demographic and ?eld-work related factors as outlined below.

Determinants of the Sample Development:

- *Demographic factors:* persons exit by death and moving abroad, persons enter by birth, moving into a GSOEP household from somewhere else in Germany or from abroad, reaching age of 16 years (minimum respondents age), new households and persons from a split of at least one old person from an old household.
- *Field-work related factors* (2 stages): making a successful contact to a given household, realizing a successful interview, social groups which are hardest to contact (single person households, residentially mobile households and persons, young adults leaving parental home).

However, in order to improve response rates, the GSOEP has implemented a respondent-incentive program such that small "bonuses" or gifts are given, and every effort is made to maintain the personal contact between respondents and the survey:

- for each successful interview, each respondent receives a small gift related to the yearly topical module and takes part in a monthly nationwide lottery;
- addresses are kept up to date by the ?eld work agency throughout the entire year in order to be informed about residential mobility; for example by sending them a brochure containing some results based on last years data;
- the interview situation (face-to-face) ensures a personal relationship, which makes it harder to withdraw from the survey; thus, the stability of the interviewer over time is very crucial.

Initial sampling design

All samples of GSOEP are multi-stage random samples, which are regionally clustered. The respondents (households) are selected by random-walk. The strata from which the primary sample units (PSU's, similar to voting districts) were selected are given by regional strata, defined by *Bundesland* (federal state), *Regierungsbezirk* (administrative district) and *Gemeindetyp* (type of community). Note that these units may change over time. For subsample B, strata are defined by the nationality of the head of the household.

*Sample A "West German Residents" ("German Sample")*

Sample A was intended to net 4,500 households. In the end the completed net sample contained 4,554 households. The ADM ("Arbeitsgemeinschaft Deutscher Marktforschungsinstitute" - Working Group of the German Marketing Research Institutes) master tape from 1982 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 random start address the interviewer had to select every seventh household as a target household. Households whose household-head belonged to the definition of sample B were discarded.

*Sample B "Foreigners in West Germany"*
Sample B consists of five autonomous samples for the five numerically largest foreign nationality groups living as immigrants in the FRG in 1984. To facilitate detailed analyses, a sample of 1400 net cases was projected. Thus the sampling rate for this sample exceeds the rate for sample A. Anticipated out-migration rates were taken into account in setting a sampling rate which gave a high probability that after several waves a considerable number of 1000 cases are still in the sample. Population B was selected from primary sampling units (PSUs) of counties and metropolitan areas. A random selection of PSUs was independently drawn for each nationality. Using immigrant registration records in each PSU, the respondents were then selected by probability sampling, i.e. systematic sampling with random start address. 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-Wurttemberg - it was not 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. Some 80 PSUs were drawn for the (strongly overly-represented) Turks and 40 PSUs for each of the remaining nationalities. Some 20 addresses were then drawn from the registers for each PSU, some of which were used as reserve addresses. The number of addresses used per sample point in the sample B show a stronger variation than in sample A. Substantially more addresses were false or contained no-longer eligible respondents, in which case a reserve address was to be used.

## The Follow-Up Concept

One of the most crucial features of a longitudinal survey to cope with problems of representativeness is the concept, according to which respondents are traced across time. Since in the GSOEP all household members are to be interviewed individually once they reach the age of 16, the next generation is automatically taken into account. In principle, all persons who took part in the very ?rst wave of the survey as well as their children whenever born, are to be surveyed in the following years. In case of residential mobility, the person is to be followed within the survey territory (Federal Republic of Germany). Third persons moving into an existing GSOEP household are to be surveyed, or "followed-up" even in case of subsequently leaving that household. The weighting scheme takes into account this "follow-up" of everybody. Temporary drop-outs or persons and households which could not be successfully interviewed in a given year are

followed until there are two consecutive temporary drop-outs of all household members or a ?nal refusal. In the case of a successful interview after a drop-out, there is also a small questionnaire including questions on central information which is missing for the year of the drop-out (e.g. employment status). "New "persons become part of the GSOEP population due to birth, or residential mobility. Those persons living in GSOEP households, who then move out or "split-off" into new households, are still followed, but under a new household identi?er. See the table below.

The Emergence of New Households

| | | Households | |
|---|---|---|---|
| | | Old | New |
| Persons | Old | • "classic case ": without change of address<br>• entire household moves | • Move-out |
| | New | • Birth<br>• Move-in | • Birth<br>• caused by split-offs of old persons |

As a result of the follow-up concept, already in the second wave, several "new" households were added to the GSOEP population.

**B. Data collection and acquisition**

Data collection period

The SOEP ?eld work takes 8 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.

Interview methodology

The interview methodology of the GSOEP is based on a set of pre-tested questionnaires for households and individuals. Principally an interviewer tries to obtain face-to-face interviews with all members of a given survey household aged 16 years and over. Thus, there are no proxy interviews for adult household members. Additionally one person (head of household) is asked to answer a household related questionnaire covering information on housing, housing costs, and different sources of income (e.g. social transfers like social assistance or housing allowances). This covers also some questions on children in the household up to 16 years of age, mainly concerning attendance at institutions (kindergarten, elementary school, etc.).

There are different versions of the questionnaires. First, the questionnaires for the foreigner's sample (B) cover additional measures of integration or information on re-migration behaviour. The questionnaires are not uniform for all samples (e.g. there is a separate schedule for foreigners and ensuing foreigner speci?c data ?les APAUSL and

BPAUSL). Secondly, there is a need to differentiate between ?rst time respondents and those with a repeated interview, since some information does not have to be asked every year, unless a change occurred. Additionally each respondent is asked to ?ll out a biography questionnaire covering information on the life course (e.g. marital history, social background, employment biography etc.).

Additional information can be obtained from the so-called "address log". This is filled in by the interviewer even in case of non-response, thus providing very valuable information for attrition analyses. For researchers interested in methodological issues this data also contains information on the process of the ?eld work, e.g. the number of contacts, reason for eventual drop-outs, or the interview method. For successfully contacted households, the address log covers the size of the household, some regional information, survey status etc., while the individual data for all household members includes the relation to the household head, survey status of the individual and some demographic information.

Survey instruments

The SOEP survey instruments include:

- Household head schedule
- Schedule for each individual over 16
- Schedule for foreigners
- Address record (in order to follow people through the panel).

The address protocol records a vast amount of information on the composition of and change within the household. That information is essential for attrition analyses, the weighting of the sample and for longitudinal analyses. The questionnaires are lavishly designed in comparison to commercial questionnaires because they have to be flexibly used by the interviewer:

- in all subsamples the questionnaire may be administered as a personal interview or as a booklet that the respondent completes himself and returns;
- in sample B the interviewer respectively the respondent has the option of conducting the interview in German, in the native language of the respondent, or in a mixed mode.

The SOEP field procedures are as follows:

- personally conducted oral interviews are conducted whenever possible;
- 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;
- 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 assistance.

These rules are obviously soft. Only one rule is 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 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 the potential amount of persons who can be contacted and are willing to do an interview can be hold on a high level. The information on the interview method applied in individual cases is available in the data. So here too systematic analyses of method-related in?uences can be made. To date there has been no indication that the interview method has a strong in?uence on the results.

Maintenance of Motivation

*Respondents' motivation* - The following methods of motivating respondents are employed (see Pol 1989, 42-45):

- giving the study a catchy name; all sample respondents know the SOEP by the name of "Life in Germany";
- respondents are given an illustrated informative brochure on the aims of the study (in sample B the brochure is translated into the respondents' native language);
- providing an information sheet on data privacy;
- following each interview with a letter of thanks after completion of the ?eld work for each wave;
- providing each respondent with a ticket for a well-known TV lottery.

*Interviewers' motivation* - Motivation of the interviewer is certainly an important in?uential factor for the respondents' willingness to participate. Good training, sufficient information about the project, a clear structuring of the survey instruments and information on research results furnish a foundation for 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. In some years the interviewers get a book with description results of the SOEP on request.

*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 the extreme case number of interviewers deployed would correspond to the number of sample points (these are the clusters which are administered by one interviewer);
- concentrating the survey work on a minimum number of highly-quali?ed interviewers.

A maximum deployment of interviewers resulted in few clusters of interviewer effects as possible. This strategy was chosen starting in 1984. It has become noticeable in the ?eld work in several waves, however, that some interviewers are much better-suited than others to realize the high response rate that the SOEP requires. Moreover, a change of interviewers was determined to decrease the probability of respondents' refusal. 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 that a great many households refuse to participate. Thus it has proved necessary to seek a balance between as many or as few interviewers as possible.

## C. Definition of the survey units

Household

Only private households are considered.

Every group of persons, who live together and economically spend and earn together, where meals are shared. Those living alone, and earning (or responsible for) their own money, constitute single-person-households.

Other private households include:

- so-called "private households in institutions": persons who live in institutions, but are responsible for earning their own living, e.g. a gatekeeper husband and wife at a hospital, or a superintendent family in a home for children.
- most persons in a residence (e.g. convent), as long as they run their own household, and are not cared for and fed by the institution.

NON-private households include those persons who live in "institutions" and are mainly fed and cared for by the institutes communal facilities.

All persons who normally live in the household, but who are at the time of interview at the hospital, on vacation, doing military or civilian service, are indeed considered to be part of the household.

Head of household

The head of the household is defined as the person who knows best about the general conditions under which the household acts and is supposed to answer this questionnaire in each given year. This reduces the risk of longitudinal inconsistencies.

## D. Contents

The GSOEP was started in 1984 as a longitudinal survey of private households and persons in the Federal Republic of Germany. The central aim of this panel study is to collect representative micro-data on persons, households and families in order to measure stability and change in living conditions by following principally a micro-economic approach enriched with sociology and political science variables, mainly determined by the "Social Indicator" movement.

A rather stable set of core questions is asked every year covering the most essential areas of interest of the study:

- population and demography;
- education, training, and quali?cation;
- labour market and occupational dynamics;
- earnings, income and social security;
- housing;
- health;
- household production;
- basic orientation (preferences, values, etc.) and satisfaction with life in general and certain aspects of life.

Additionally, as a yearly topical module, the basic information in one of these areas is enlarged by detailed questions as seen in the following table.

Special Topics Modules

| Year | Wave | Sample | Topic |
|---|---|---|---|
| 1984 | A/1 | A B | Employment biography since age 15 (Bio) |
| 1985 | B/2 | A B | Marriage and family biography (Bio) |

Notes:
W: West German Sample: A, B

In order to measure change and stability across time, the GSOEP-questions are targeted at different dimensions of time (past, present and future) using also different measurements of time (information at a given point of time, periodical information, calendar information, life history information):

- questions about a point of time (present) e.g. current employment status or current levels of satisfaction;
- single retrospective questions on certain events in the past (in the past) e.g. how often did you change your job during the last ten years?;
- retrospective life event history since the age of 15 (in the past) e.g. employment or marital history;
- monthly calendar on income and labor market related issues (in the past) e.g. employment status January through December last year;
- questions concerning a period of time (in the past) e.g. demographic changes since the last interview like marriage or death of spouse;
- questions concerning future prospects (future) e.g. satisfaction with life ?ve years from now, or job expectations.

## E. Quality of data

Results of Sampling in the 1st Waves

*Quality-neutral attrition rates* - According to the sample plan, the original gross number of households 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. In order to calculate the drop-out rate of the ?rst wave, households that did not belong to the target population "Private Households Excluding the Separately-Interviewed Households in Sample B" had been subtracted from the total amount of start addresses. These addresses are defined as "quality-neutral drop-outs" and the result as "edited gross amount". 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 does not become clear until contact is made at a later date that the household does not belong to the target population (because the household members belong to 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 could not be found. Sample B (the foreigner sample) gives a different picture because addresses were supplied by the registration offices. The extent of the quality-neutral attrition due to false or no-longer-current addresses is greater here than in sample A. The average rate of attrition for the ?ve immigrant samples is 22% of the utilized addresses.

*Sample response rates* - The sample response rate is varied by stratum. In sample A, 4,554 addresses could be taken into the net sample after concluding all of the ?eld phases and the processing work. A sample response rate of 60.6%is implied. It is common that initial responses in panel surveys are signi?cantly lower than the response rates in subsequent waves. See Duncan and Kalton (1987) p.109 and Duncan and Hill (1989). Sample B yields better results. The response rates range from 64.7% for the Italians to 70.0% for the Turks. In sample A, the main cause of attrition was refusal. Due to a long period of ?eld work, non-contacted households could be reduced to 3.2% (a percentage not attained in normal cross-sectional surveys). Only 0.2% of households could not be surveyed due to linguistic difficulties. 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 did not 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.

*Comparison with external sources* - The quality of a sample (cross-sectional representativeness) can be inferred from the conformity of a distribution of characteristics

in the sample with the distribution in the population. The distribution in the target population is estimated using external statistics, posing the problem that they themselves could be biased as well (particularly the income and consumer samples). It should be pointed out that the official statistics also contain institutionalised residents who are not included in the GSOEP sample. With regard to the regional distribution characteristics and the household structure, sample A re?ects familiar shortcomings of survey research. The population in the metropolitan areas, 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. The socio-demographic structure of sample B, household sizes as well as the age and sex of the household head, appears satisfactory. In regard to the regional distribution, the drop-out structure shows the same result as for sample A, namely intensi?ed attrition in the core zones of the metropolitan areas.

Attrition in the Course of Time (Wave 2)

For details of the level and the structure of attrition in the GSOEP see Pannenberg (2002a). Death and moving abroad are natural causes for dropping out and are not a problem for analysis. But dropping out due to refusal of respondents and in some cases due to problems of ?nding a household again ("unsuccessful follow-up") may cause problems when the dropouts are not random. The following characteristics have found to be of signi?cant importance in the GSOEP:

- Unsuccessful follow-up: household moved, split-off, large city, single household.
- Refusal of Respondents: resident of East Berlin, age of head of household, female head, household moved, split off, separation/divorce of partner, change of interviewer, number of interviews with the same interviewer, low household income, item non-response on income, expected loss of job, migration from East to West Germany.

Weighting

The goal of any sample is to draw conclusions from the sample and apply them to the "recorded" target population. Due to different sampling probabilities, non-response in the ?rst wave and attrition in the course of time, a weighting ("projection") of the sample cases is required in order to be able to infer the case numbers of the target population. For SOEP we must distinguish three steps of weighting:

1. Cross-sectional weighting of wave 1
2. Weighting of longitudinal populations
3. Cross-sectional weighting of waves 2 and thereafter

*1. Cross-sectional Weighting of Wave 1* - The selection probabilities (and thus the weighting factors) for the ?rst wave of a panel are of special importance, because these values are used as the starting point for deriving all other weighting factors. The survey design of the ?rst wave of the German Socio-Economic Panel contained a two stage selection procedure. The primary units (sample points) are polling districts in sample A

and counties in sample B. The secondary units (households of the ?rst wave) are drawn from the polling districts using a random route procedure, and in sample B the persons drawn from that county 's foreigner register.

The primary units have been strati?ed. For sample A, 148 regional cells, i.e. strata, were formed from the characteristics state, county, and local district. The strata sizes $n_h$ were chosen to be proportional to the number of households $N_h$ in those regional cells, i.e. $n_h$ is proportional to $N_h$, where h = 1 ,...,148. Then the primary units were chosen in each regional cell using systematic sampling proportional to size. The size proportions are related to the number of households of the primary units. The ordering of the primary units is important when using systematic sampling with random start and ?xed intervals. The sequencing of primary units was carried out with the characteristics district, community, city section, and polling district number. Within each foreigner group in sample B the districts, i.e. primary units, were sequenced by state and county. Again the primary units for each foreigner group were selected by systematic sampling proportional to size. The size of primary units was the number of foreigners of that nationalities within that area.

The second stage of the sample's selection was the performance of a random route procedure of sample points (sample A) and for sample B the selection of person from the foreigner registers. The selection of a person from the registers used again a systematic selection with random start number and ?xed intervals. In some ways the random route procedure for the sample points can also be interpreted as such a systematic selection. There the procedure 's inspection rules, provide a sequencing of the households. Every seventh household was chosen. Through the sequencing of the primary units and the systematic selection procedure, a kind of net is generated which covers the survey area uniformly.

One must recognize that the proportional selection probabilities at the primary unit level are such that in each primary unit (the objective was $n_s$ equals eight households) a ?xed number $n_s$ of secondary units are drawn. If $p_k$ is the selection probability for the k-th primary unit and $p_{i/k}$ the selection probability for the i-th unit in the k-th primary unit with size $N_k$, then the selection probability $P(C_i = 1)$ for the i-th unit is $p_i$:

$$p_i = p_i / {}_k p_k = n_s / N_k * \text{const} * N_k = \text{const} * n_s.$$

The designs of selection of households in sample A and persons in sample B are approximately identical. However, in sample B households in which several persons are of the same nationality, the selection probabilities of households equal to the sum of selection probabilities of all household members. Hence the selection probabilities of households in sample B are proportional to the number of its household members who are 16 years and older. Because all household members were included in the survey, the selection of the households and the persons living in them is identical. Therefore the selection probabilities of a given household and of its members are the same, except for persons who have secondary residences. These persons have doubled the selection probability, i.e. an implicit assumption was made that these persons have the same selection probability at both residences.

On the basis of household or foreigner aggregates in the primary units the design probabilities $P(D_i =1)$ can be determined for the start wave. Because respondents were told at the beginning of the survey that this would be an annual survey, the respondents knew that more time and effort would be required than for a usual cross-sectional survey. Hence a sampling rate of about 65% (average of sample A and B) in the ?rst wave can be considered a success.

The response probabilities $P(R_{1i} =1 \mid D_i =1)$ for the ?rst wave were estimated in two steps: the ?rst step used only regional characteristics of all households (participating and non-participating); the second step compares sample information of participating households with the corresponding information from other surveys (micro-census and European Communities labor survey).

The regional characteristics of all households (participating as well as non-participating) are known. The response probabilities can be estimated in each of the 148 regional cells by using the ratio of participating households to the total number of attempted households. The reciprocal of the products $P(D_i =1) * P(R_{1i} \mid D_i =1)$ is available for the user as design and regional weights in the GSOEP-database under the label AHDESREG (in the ?le HHRF which contains the weighting variables for households) or APDESREG (in the ?le PHRF with personal weighting variables). These calculations of the selection probabilities can be enhanced by linking the estimated number of household and person characteristics with those from other corresponding surveys. This is done because it is assumed that:

- the population estimate in the other surveys are more exact than those estimated from the panel;
- this adaptation also makes the calculation of other characteristics more precise.

The higher precision of the Micro Census is derived from its 50 times larger sample size and the respondents' obligation to provide information. The second argument is more difficult to substantiate. The modi?ed selection probabilities are not uniquely de?ned in that the estimation results agree with the J<n restrictions from the Micro Census. A specific solution is achieved if the modi?ed weights and the original weights have minimal information distance, see Ireland and Kullback (1968). This speci?c solution is characterized by

$$\tilde{P}(R_{1i} =1 \mid D_1 =1) = P(R_{1i} =1 \mid D_1 =1) * \exp\left(\Sigma_{j=1,N} ?_j m_j(i)\right)$$

for the modi?ed response probabilities $\tilde{P}(R_{1i} = \mid D_i =1)$. See (Rendtel 1987).

When the i-th unit has the j-th characteristic, then m j(i) takes on the value one, which is controlled by the distribution of the micro-census. The unknown l coefficients can be determined by iterative application of the "should/is " adaptation, i.e. the "iterative proportional ?tting" algorithm. Alternatively the l's can be ascertained by a direct minimization of the information distance with the Newton-Raphson algorithm (see Merz (1983)).

The adaptation of estimation results to certain marginal distributions is therefore equivalent to the assumption that the response probabilities are in accordance with a main effects model, where the main effects are generated by the variables for the marginal distribution. This supports the original assertion that the adaptation of additional data increases the estimation precision of other characteristics. Under the main effects model the uncorrected estimation of population totals is biased. However the validity of this model cannot be checked empirically. To do this it would be necessary to know the model variables for missing households, which is generally not the case. The relationship of the response probabilities may be even more complex. For instance, if signi?cant interaction effects appear alongside the main effects, then the adaptation procedure can lead to increased distortions of population estimates.

The three-step-weighting process of wave 1 of samples A and B takes into account all information which are available for the calculations of sampling probabilities. Thus the ?nal step of adjusting the marginal distributions of the sample and external statistics is changing the weighting factors slightly only. But the three-step-procedure make the variance of the weighting factors bigger than a one-step-procedure might do.

*2. Longitudinal Weighting* - The weighting of longitudinal populations is straightforward. To get an estimate for time t+x it is only necessary to know for certain subgroups in the population how big the drop-out rate is. The inverses of the drop-out rate give the weighting factor. The calculation of dropout rates can be done by cross-tabulations or - much better - by LOGIT-regression analysis. For details see the background paper "Documentation of Sample Sizes and Panel Attrition in the GSOEP" by Pannenberg (2002a). To determine the reasons for attrition in wave t+1 the characteristics of a household in wave t =0 can be used. Additionally characteristics of the ?eldwork for wave t+1 can be used, for example the information that a household moved or the interviewer changed. This analysis is done for each wave. The longitudinal weighting factors adjust from one wave (beginning with wave 2) to another wave. To arrive at the correct weight for longitudinal analyses in the course of multiple waves, the longitudinal factors need only be multiplied by each other.

*3. Cross-sectional Weighting of Wave 2* - For cross-sectional weighting not only "old friends" must be weighted but new members of the sample too. 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. Thus it is not necessary, as the PSID is doing, to assign zero-weights to persons who join old households. But 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 without (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 members (i.e. non-initial persons too), this lower weight compensates for the increase in case numbers caused by the new arrivals.

Cross-National Equivalent File (CNEF)

As a by-product of the GSOEP, a German Cross-National Equivalent File (CNEF) is also created. The Cross-National Equivalent File is created by Cornell University, in close cooperation with DIW-Berlin, ISER-Essex and StatsCan-Ottawa, consisting of variables from the German GSOEP, American PSID, Canadian SLID and British BHPS, based on common de?nitions. The income variables are all annualized, meaning that the typical German SOEP variables asking about monthly income components have been transformed. The Equivalent File variable names are identical across datasets, adding to ease of use. The reader is referred to the standard Equivalent File documentation in Burkhauser, Butrica, Daly, and Lillard (2001) for further information (all used original variables names from the data sets are included with the algorithms). The codebooks are available at http://www.human.cornell.edu/pam/gsoep/equiv ?l.cfm.

With respect to the normal GSOEP variables, the CNEF also includes completely simulated taxes and social contributions, as well as a full imputation of all missing income information due to item non-response. For a description of these procedures, see below.

Simulation of taxes and social contributions

The GSOEP does not currently provide information on the annual tax payments of its respondents, but they are completely simulated using a simulation package that uses a methodology to compute taxes and social contributions for GSOEP respondents that is conceptually similar to the one used in the Panel Study of Income Dynamics for United States households (cf. Johannes Schwarze, 1995: "Simulating German income and social security tax payments using the GSOEP", Cross-National Studies in Aging Program Project Paper No. 19). A brief overview of the simulation package follows.

The German tax system is complex. Thus, it would be difficult to incorporate each of the regulations described into a simple, easily updated program. Therefore, in general the simulation programs are based on a set of simplifying assumptions:

- all married persons file jointly
- all filing units take the standard deductions
- no filing unit itemizes
- when no standard deduction exists the allowance is ignored
- average national insurance contribution rates for old age pensions, health insurance,
- and unemployment insurance apply to all employees.

*a) Simulating Income Tax Burdens* - The first step in computing the tax base is to compute potentially taxable income. To do this, assumptions about income related expenses for each of the seven income sources have to be made. One key assumption concerns how respondents report income from different sources. It is not clear whether respondents report their potentially taxable income (income minus expenses) or their gross income. Moreover, respondents may report gross income in some cases and income net of expenses in others. For example, individuals may report their gross income from

labor earnings but their net income (profits) from self-employment. In this tax simulation package income from sources 1 through 3—self-employment income—and income from rentals or leasing is assumed to be net of expenses. Income from all other sources—4, 5, 6, and 7—is assumed to be gross income from which the standard deductions are subtracted.

Additional adjustments to income from employment and from social security and employer pensions for workers other than civil servants must be made. Income from employment must be adjusted for individuals who received short-time or bad-weather allowances. These allowances are tax-free and should be subtracted from taxable wage and salary income. Since the survey question on this topic only asks for the number of weeks respondents received such benefits, the simulation program simply assumes that these benefits are a fixed percentage of reported weekly gross income. Benefits for non-civil service retirees have to be divided into the taxable profit share portion and the non-taxed contribution portion. In principle, this division could be done separately for each person based on actual retirement age. However, for simplicity retirement is assumed to occur at age 60 for everyone, producing a constant profit share portion of benefits equal to 28 percent.

Additional deductions, such as the ones for old age and special expenses in Steps 2 and 3 of the income tax calculation, are treated uniformly for all filers. Persons over age 65 are assumed to deduct the lesser of 40 percent of their income or 3,720 DM. Deductions for self-employed individuals are set equal to the upper limit allowed because contributions to private old age and health insurance are not known. The simulation package also takes into account special regulations for joint filers and pensioners. Deductions for other special expenses are set equal to the standard deduction of 108 DM (216 DM for joint filers). Extraordinary expenses and loss deductions are ignored within the tax simulation package.

The tax simulation package computes the child allowance exactly as the German income tax laws require and applies this deduction to the taxable income base. The computation of the housekeeping deduction is only included in the tax simulation program for individual filers (M-TAX- I.SAS), because the deduction can only be taken by single persons. This allowance is a fixed amount which is deducted whenever a child is present in a single adult household.

*b) Simulating Social Security Contributions* - Most employees are compulsory members of the statutory social security system. As a first step, contributions to old-age insurance, health-care insurance, and unemployment insurance are calculated for every individual with positive income from employment. The income base is income from employment up to a certain limit (limits as well as contributions rates for every year can be seen in the programming code).

It is assumed that all private sector workers are charged the same contribution rate. The rate applied in the tax simulation program is the average rate of the statutory health insurance agencies. This assumption is made despite the fact that there is detailed

information about health care insurance in some waves of the GSOEP. The same assumption is made in the case of civil servants. These individuals can be identified in the GSOEP data by their employment status and are excluded from compulsory social insurance. However, civil servants are partially funded by their employer if they purchase private health insurance. As a result, almost every civil servant is covered in part by private health insurance.

It is difficult to identify all marginal employees even with detailed survey data. The present version of the tax simulation program does not consider hours worked in determining such employment. Rather, marginal employment is approximated based solely on yearly income from employment. If yearly gross labor earnings are below the ceiling, it is assumed that the worker is in marginal employment and no contributions to the social security system are computed.

Imputation of item non-response on income questions

The imputation of item-non-response related missing income data in the SOEP follows a two step procedure (cf. Grabka and Frick 2003). The general principle is to apply the row and column imputation technique (hereafter L & S) whenever longitudinal income data is available, and to run purely cross-sectional imputation techniques otherwise. As a matter of fact, the empirical implementation of L & S in the case of SOEP fails in all those cases where a given income component is not observed in any other wave of data considered in the imputation process. This includes not only first time respondents, but also those observations for whom a given income variable has been surveyed for the very first time. In all of those cases there is a need for an alternative imputation procedure which is based on cross-sectional data only, i.e., on data observed from other units (individuals or household, respectively) in the very same wave. The different techniques applied for the various SOEP income variables are briefly overviewed below:

- Following *logical imputation*, institutional or external information is used to impute missing amounts of those income components which are perfectly related to otherwise observed information, e.g. child benefit which is fixed per child, direct housing support for owner occupiers which is related to the number of children and the construction year of the building, as well as nursing care insurance which is fixed to the observed needs.

- *Median Substitution* takes place for income components which are of minor relevance in terms of the number of affected cases ($n < 10$) as well as with respect to the level (e.g. military service pay, maternity benefit). Median Substitution for Subgroups is performed for e.g. housing benefit for owner occupiers by household size.

- *Median Share Substitution* is chosen if a link between two income variables can be established, e.g. the median share of the monthly labor earnings and the Christmas bonus in the private sector in Germany is about 35%. Any observation with a missing Christmas bonus in the private sector is assigned an imputed value

given by the individually observed labor income times the (median) share of 35%. This allows for more variation of the imputed income values than single median substitution would do.

- *Regression-based substitution* is used for more complex income constructs e.g. "interest and dividends" or "individual labor income from first job"; in the latter case Mincer-type wage regressions are applied for imputation purposes (cf. Grabka and Frick 2003).

## F. Uses of the survey

Publications


Autorengemeinschaft Panel (Eds.); 1990: Das Sozio-ökonomische Panel für die Bundesrepublik Deutschland nach fünf Wellen. in: Vierteljahreshefte zur Wirtschaftsforschung (Berlin), Heft 2.

Berntsen, R.; 1989: Einkommensanalyse mit den Daten des Sozio-ökonomischen Panels unter Verwendung von generierten Einkommensdaten, SfB 3 Arbeitspapier Nr. 291.

Berntsen, R; W. Dobroschke-Kohn; 1990: Update 1984/85, Dataset Description Federal Republic of Germany, SfB 3 Documentation.

Frick, J.R. and M. Grabka (2003): "Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income Distribution", *DIW Discussion Paper 376*, October 2004.

Grabka, M. and J.R. Frick (2003): "Imputation of Item non-Response on Income Questions in the SOEP – 1984-2002", *DIW Research Notes 29*, October 2003.

Hanefeld, U.; 1987: Das Sozio-ökonomische Panel, Grundlagen und Konzeptionen, Frankfurt/ New York.

Hanefeld, U.; 1984: The German Socio-Economic Panel, American Statistical Association, 1984 Proceedings of the Social Statistics Section, Washington D.C.

Kassella, T.; U. Hochmuth; 1989: Ein synthetisches Mikrodatenfile des Haushaltssektors für steuerpolitische Simulationen, SfB 3 Arbeitspapier Nr. 299.

Rendtel, U.; 1987: Methodische Konzepte für die Hochrechnung von Paneldaten. in: Vierteljahreshefte zur Wirtschaftsforschung, Heft 4-1987.

Rendtel, U.; 1988: Panelmortalität. in: Vierteljahreshefte zur Wirtschaftsforschung, Heft 1/2-1988.

Rendtel, U.; 1989: Über den Einfluß der Panelselektivität auf Längsschnittanalysen. in: Vierteljahreshefte zur Wirtschaftsforschung, Heft 1-1989

Rendtel, U.; 1990: Hochrechnung und Stichprobenfehler in Panelerhebungen, SfB 3 Arbeitspapier Nr. 321.

Rendtel, U.; 1988: Population Estimates and Representative Character of the Data, english translation of an article in: Krupp/Schupp (Eds.); 1987: Lebenslagen im Wandel: Daten 1987, Frankfurt.

Rendtel, U.; 1990: Teilnahmebereitschaft in Panelstudien: Zwischen Beeinflussung, Vertrauen und sozialer Selektion. in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Nr.2/1990, pp. 280-299.

Sonderforschungsbereich 3 (SfB 3), Deutsches Institut für Wirtschaftsforschung (DIW)(ed.): Benutzerhandbuch (user handbook), Frankfurt/M./Mannheim, Berlin. Source: Das Sozio-ökonomische Panel Deutsches Institut für Wirtschaftsforschung.

Schwarze, J. (1995): "Simulating German Income and Social Security Tax Payments Using the GSOEP", Cross-National Studies in Aging, Program Project Paper No. 19, Maxwell School of Citizenship and Public Affairs, Syracuse University, New York.

Witte, James C.; 1990: The Potential for Comparative Panel Research using data from the U.S. Survey of Income and Programme Participation (SIPP) and the German Socio-Economic Panel (SOEP), DIW working paper, Berlin.

Poverty and income distribution.