SOEP The German Socio-Economic Panel study

418

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP — The German Socio-Economic Panel study at DIW Berlin 2017

SOEP-Core v32 – Documentation on Biography and Life History Data

Jan Goebel (ed.)

DIW SOEP

Running since 1984, the German Socio-Economic Panel study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing. The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at http://www.diw.de/soepsurveypapers

Editors:
Dr. Jan Goebel, DIW Berlin
Prof. Dr. Martin Kroh, DIW Berlin and Humboldt Universität Berlin
Prof. Dr. Carsten Schröder, DIW Berlin and Freie Universität Berlin
Prof. Dr. Jürgen Schupp, DIW Berlin and Freie Universität Berlin

Please cite this paper as follows:

Jan Goebel (ed.). 2017. SOEP-Core v32 – Documentation on Biography and Life History Data. SOEP Survey Papers 418: Series D. Berlin: DIW/SOEP

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
© 2017 by SOEP
ISSN: 2193-5580 (online)
DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany
soeppapers@diw.de

SOEP-Core v32 – Documentation on Biography and Life History Data

Jan Goebel (ed.)

The files described in this documentation are part of a collection, which is released with doi:10.5684/soep.v32

1 General Introduction and Overview of available datasets 6

2 Biographical Information in the Meta File PPFAD (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989) 19

2.1 The Month of Birth in the data set PPFAD 19
2.2 Construction of variables 20
2.3 Year of birth 23
2.4 Year of death 24
2.5 Immigration information 26
2.6 Living in East or West Germany in 1989 35

3 MIGSPELL: The Migration-Biography (Samples M1 and M2) 37

3.1 Introduction to the new release of MIGSPELL 37
3.2 Summary description of MIGSPELL 38
3.3 Overview: the structure of the migration biography questions 38
3.4 Description of the variables in MIGSPELL 41
3.4.1 Summary description of the changes in MIGSPELL 41
3.4.2 Synopsis of the variables in MIGSPELL (systematic order) 42
3.4.3 Levels and value labels of the categorical variables (alphabetical order) 43
3.4.4 The date-variables in MIGSPELL 48
3.5 The imputation of missing date values 49
3.5.1 General remarks 49
3.5.2 Procedure for the imputation of the missing date values 50
3.6 The integration of the migration biographies of the three waves 2013-2015 (bd, be, bf) 53
3.6.1 Structure Tables: $$p_mig-variables to MIGSPELL-variables for waves bd to bf 55
3.6.2 Synopsis: Mapping of the migbiography variables of waves 2013-2015 to the MIGSPELL variables 61

4 Activity Biography in the Files PBIOSPE and ARTKALEN 65

5 BIOJOB: Detailed Information on First and Last Job 74

5.1 Overview 74
5.2 Structure and Contents of BIOJOB 75
5.3 Steps of Coding 94

6 The couple history files BIOCOUPLM and BIOCOUPLY, and marital history files BIOMARSM and BIOMARSY 97

6.1 Sources of the couple and marital history 98
6.2 Construction of marital histories 99
6.3 Construction of couple histories 101
6.4 BIOCOUPLM: A monthly couple biography 105
6.5 BIOCOUPLY: An annual couple biography 108
6.6 BIOMARSM: A monthly marital history 111

6.7 BIOMARSY: A annual marital biography 113

7 BIOBIRTH: A Data Set on the Birth Biography of Female Respondents 116

7.1 Population and purpose of the data set BIOBIRTH 116
7.2 Structure of the data set 117
7.3 Information basis of the birth biography 118
7.4 A new source of biographical information – the youth questionnaire 120
7.5 The fertility histories of male respondents in BIOBIRTH 120
7.6 Integration of "Familien in Deutschland" – FiD 121
7.7 Identification process of the children in the SOEP data base 122
7.8 Identification of the children of parents with completed fertility histories 124
7.9 Identification of the children for women who have no biography data/ not completed the biography questionnaire 124
7.10 Updating BIOBIRTH 124

8 BIOTWIN: TWINS in the SOEP 128

8.1 Population and contents of the data set BIOTWIN 128
8.2 The twin survey of 2006 129
8.3 Construction of variables in the data set BIOTWIN 129

9 BIOSIB: Information on siblings in the SOEP 132

9.1 General description of the data set 132
9.2 Sources of information on siblings in the SOEP 132
9.3 Overview on the number of siblings in BIOSIB 133
9.4 Organization of the data in BIOSIB 133

10 BIOAGE01, BIOAGE03, BIOAGE06, BIOAGE08, BIOAGE10, BIOAGE12: Generated variables from the "Mother & Child", "Parent", and "Pupils" questionnaires 140

10.1 Introduction 140
10.2 Respondents in the 'Bioagel' Data Set 141
10.3 Topics and Variables 144
10.4 Generated Variables 145

11 BIOAGE17: The Youth Questionnaire 149

11.1 Genesis and Target Population of the Youth Questionnaire 149
11.2 Contents and Structure of the Data Set BIOAGE17 151
11.3 Special Features of Some Questions and Variables 152

12 BIOSOC: Retrospective Data on Youth and Socialization 161

12.1 Structure of the Data Set BIOSOC 162
12.2 Special Features of Some Questions and Variables 162

13 BIOPAREN: Biography Information for the Parents of SOEP-Respondents 170
13.1 Short summary 170
13.2 How biography information has been collected in the SOEP 170
13.3 How is BIOPAREN generated? 173
13.4 What's new in version v32? 177
13.5 Complete list of variables in BIOPAREN 178
14 BIOIMMIG: Generated and Status Variables from SOEP for Foreigners and Migrants 213
14.1 Content 213
14.2 Status Variables and Carrying Forth of Information 213
14.3 Updating of Time-Dependent Information 215
14.4 Using this File 216
14.5 Using BIOIMMIG as a Cross-Section 216
14.6 Documentation of the Variables 217
15 BIORESID: Variables on Occupancy and Second Residence 269
15.1 Sources of Variables 270
15.2 Population of Interest 270
15.3 Variable List of the Data Set BIORESID 273
15.4 Recent Changes in the Data Set 273
16 BIOEDU: Data on educational participation and transitions 274
17 LIFESPELL: Information on the Pre- and Post-Survey History of SOEP-Respondents 277

1 General Introduction and Overview of available datasets

by Jan Goebel

By compiling a comprehensive set of questions on the individual life history into user-friendly variables, the SOEP database provides users with a representative collection of biographical information for the entire German population.[1] This covers information on the individual career path since the age of 15, on marital status and childhood biography, the first job, social background and migration history. The function of these data is, on the one hand, to make important background information available for analyses (e.g. information on fertility as an explanatory variable when analyzing labor market supply of women), and, on the other hand, to support self-contained analyses (e.g. on occupational careers or intergenerational transmission of education).

In general, each respondent of the SOEP questionnaire (surveying age starts in the calendar year a person turns 17 years) will answer the biographical questions only once (retrospectively). In the beginning of the SOEP, this occurred within the framework of the first three waves (1984 to 1986). Due to the inevitable 'mortality rate' of the panel (refusal to participate, death, relocation abroad), this process unfortunately leads to missing biographical entries for persons who did not participate in all three waves. Because of this, since 1988 all biographical information (occupation, marital status, family, first job and social background) is, in principle, collected during the first interview for new respondents in existing sample households. It should be noted that - due to the costs involved and the increased response burden—the main objective of surveying the biographical information in the course of the very first interview is not applied to the first wave of new subsamples. For example, in sample C (East Germany, field work started in 1990) the biographical questionnaire was first collected in 1992. Consequently, the surveyed persons in sample C who left SOEP before 1992 or who refused to complete the biography questionnaire in 1992 have no biographical information included in the SOEP data. Starting with Sample J in 2011 we now enable member of new subsamples to fill in an integrated questionnaire, combining individual and biographical questions. This procedure allows us to collect biographical information during the first wave without increased response burden, however at the expense of a slightly different individual questionnaire when compared to "old" samples. In such a case, the effected variables will be set to "-5 Not included in this version of the questionnaire" for the entire subsample.

Summing up, in principle most of the biographical information in the SOEP is collected by means of the so-called 'Lebenslauf' (`life history') questionnaire. Although naming

[1] A general introduction into the SOEP database can be found in our Desktop Companion (DTC) at **http://about.paneldata.org/soep/dtc/**

conventions, positioning of questions and the scope of this questionnaire have been changed and revised several times (see below), it has been addressed once at each respondent throughout the SOEP. Since 2000, a separate youth questionnaire exists which contains youth-specific questions.[2] A whole new series of age-triggered instruments for collecting biographical data was implemented in 2003. The target of the *first* of these questionnaires is to collect information about *newborn* children. It is aimed at their mothers of children aged up to 15 months. As a result, the SOEP has started to survey the development of children from the very beginning of their life and will provide users with a completely new type of data. In 2005, a follow-up questionnaire targeted at children aged 2 to 3 years was implemented. Again, the information was collected from the mothers. It contained questions on the child's individual development and the mother's specific experiences during this formative period of raising the child. There will be follow-up interviews to collect data about these children at specific ages which are typically associated with decisions relevant to their individual development. The respective questionnaire targeted at children aged 5 to 6 was implemented in 2008. A questionnaire targeted at children aged 7 to 8 years is used for the first time in 2010. This questionnaire will be answered by mothers and fathers (in contrast to earlier age-triggered instruments which were answered by mothers, only) and will therefore delivered in two separate files. A questionnaire targeted at children aged 9 to10 years is used for the first time in 2012. In 2014 we introduced a specific questionnaire for pupils aged 11 to 12 years.

A chronological listing of the various changes related to the survey of biographically relevant information for the time period 1984 to 2015 can be found below. The differences in gathering information among and between the various sub-samples are reported with respect to 'Timing' (*when* respondents were asked), 'Coverage' (*which* parts of the biographical topics and single indicators were asked), and 'Positioning' of the biographical questions in the diverse survey instruments.

1984 The focus of the survey from samples A and B was the occupational biography. This information was collected (retrospectively) with the help of a 'life-course calendar' and covered the time period from the age of 15 up to the current age (or up until and including the maximum age of 65). The 'calendar' takes the form of a matrix with one column for every year of age and up to nine specifications of occupational activities (school, apprenticeship/training, military and community service, employed full time, employed part time, unemployed, househusband/wife, retired and other; question 62 in the standard Individual Questionnaire in Wave 1).

1985 The focus for samples A and B was on collecting marital and family status information in retrospect (questions 81-88 in the standard Individual

[2] Due to survey-related reasons biographical information was not asked of first-time respondents aged 16 or 17 years until 1999. For this group of persons, much of the biographical information (i.e. on marital status and family information, occupation history since the age of 16 and social background) can generally be reconstructed using variables collected by means of the Individual Questionnaire, i.e. from the yearly ongoing survey.

Questionnaire in Wave 2). The number of children born up to that point in time is collected in detail as well as the eventual date at which the children moved out of the parents' home (only female participants were asked those questions). In addition, residency during childhood, the date a person moved out of the parents' home, as well as the start, end and potential reasons for the termination of up to three marriages are asked of each surveyed person.

1986 The focus of the life history data was on social background and entry into the workforce (questions 10-13 and 80-87 in the standard Individual Questionnaire, Wave 3). Information on every surveyed person's parents is included, i.e. their year of birth and, where appropriate, the year of death, their level of education and vocational training as well as their working status at the time the respondent was 15 years of age. Furthermore, the father's type of gainful employment was also asked. With respect to entering the workforce, information is available on the age of the surveyed person when he/she first started to work and the type of employment he/she had. When appropriate, the age at each job change was also asked.

1987 No biographical information was collected in this year.

1988 The complete collection of biographical questions was included in the blue Individual Questionnaire for first time participants. For those persons who were new additions to the SOEP since 1985 and who had missed portions or all of the biographical questions, this missing information was collected in 1988. For this reason, it is only since 1988 that complete biographical information has been available on all three biographical areas for all persons surveyed up to this point (as long as they were still included in the SOEP population). Young adults up to age 17 were excluded from this retrospective collection of biographical information due to reasons of content (it did not make sense to collect the biographical information here). However, some technical problems arose when determining the exact minimum cut-off age of the persons included in this retrospective survey.

1989 to 90 During this time period, the form of the survey on biographical data from 1988 remained unchanged.

1990 The SOEP random sample was expanded. For the first time individuals and households in East Germany (sample C) were surveyed. However, biographical data for sample C were collected later on.

Since 1991 New respondents of sample A (West Germany) and B (Foreigners / 'Guest Workers') answered the biographical questions in an independent Biography Questionnaire.

1992 Due to the differing occupational titles, educational degrees, and biographies between East and West Germany, an additional biography questionnaire was developed for sample C (East Germany) which was first used in 1992 in order to cover all respondents in this sample. This additional questionnaire is identical with the western version in its structure and the format of its questions, with just a few answer categories being modified and speech delimitations were effected (for example, on occupational position or the description of the successfully completed apprenticeship). This extensive group of questions was also applied to everyone who had been a new addition to the survey in East Germany since 1993.

1994 An updated version of the biography questionnaire version called 'Lebenslauf' ('life history') was introduced for all the four samples A, B, C, and D1/D2.[3] The formats of some of the questions were slightly changed, and new questions were added, although some questions were included for only one of the samples (i.e. questions relevant to immigration were only directed towards sample D).

1996 The Biography Questionnaire 'Lebenslauf' ('life history') was fully integrated for all samples, for example, using appropriate filter questions the immigration relevant information was also asked from persons in samples A to C in an identical form.

1998 Introduction of the supplementary sample E.

1999 The 1996 form of the Biography Questionnaire 'Lebenslauf' ('life history') was given to members of sample E for the first time.

2000 The 1996 version of the Biography Questionnaire 'Lebenslauf' ('life history') was changed slightly. For example, information on having own children is collected for men as well, as is the information on the respondent's mother's occupation at the time that the respondent was 15 years old.
A preliminary-version of the Youth Questionnaire was designed and given to 17 year old youths (only samples A to E). Data on social background were collected from young adults with single or no parents in the household.
In the year 2000, a new supplementary sample F with over 6000 surveyed households was established.

2001 The Biography Questionnaire 'Lebenslauf" ('life history') was further expanded and now also includes more questions on school, i.e. marks, and activities during childhood.

[3] (A) 'West Germany', (B) 'Guest Workers / Foreigners', (C) 'East Germany', (D1/D2) 'Immigrants since 1984', persons from D2 were first surveyed in 1995.

Biographical data are collected for the first time for all persons belonging to sample F using this updated Biography Questionnaire.
The revised Youth Questionnaire, the standard version for the forthcoming years, is used in the field for all 17 year old teenagers in addition to the Individual Questionnaire.

2002 A new sample G is drawn, which is only targeted at high-income households, i.e. households with a monthly net household income of more than 7,500 DM (≈ 3,850 €). This sample was also asked retrospective information on inheritances, which was collected in 2001 for samples A through F.

2003 Persons from sample G answered the Biography Questionnaire for the 'first' time.
The new questionnaire 'Mother and Child' was given to mothers of newborns (all samples).

2004 The Biography Questionnaire was slightly expanded with questions concerning the 'numbers of brothers and sisters' and the 'location a person lived at before reunification (East Germany, West Germany, abroad)'. The question on siblings is also asked in the Youth Questionnaire.

2005 The new questionnaire 'Mother and Child II' ("Infants") targeted at children aged 2 to 3 years was implemented (for all samples).

2006 Introduction of the supplementary sample H with valid interview information for about 2.600 individuals. As a standard procedure, these new respondents do not fill in the biography questionnaire in order to reduce response burden in wave 1.

Starting in 2006, the age for first-time respondents has been changed to be the calendar year in which the person turns 18 years of age. Those aged 17 in 2006 are asked to fill in the extended "Youth Questionnaire" (data is stored in the file $PAGE17) instead of the "Individual Questionnaire" (data stored in $P). These extended questions cover indicators on subjective well-being, health (including body measures), labor force participation and education.

2007 Members of Sample H have answered the biographical background questionnaire for the very first time.

2008 The new questionnaire 'Mother and Child III' ("Pre-School") targeted at children aged 5 to 6 years was implemented for the first time (for all samples).

2009 Introduction of the new subsample I with valid interview data on about 2.500 adults. As a standard procedure, these new respondents do not fill in the biography questionnaire in order to reduce response burden in wave 1.

2010 Members of Sample I have answered the biographical background questionnaire for the very first time. Members of Samples L1 and L2 are answered only the first part of the biographical questionnaire. The new questionnaire 'Parents I' targeted at children aged 7 to 8 years was implemented for the first time.

2011 Members of the new subsample J are asked to fill in an integrated version of the individual questionnaire and the biographical background questionnaire. Members of subsamples L1 and L2 are asked to answer the second part of the questionnaire. Members of sample L3 are asked for the first part of the questionnaire.

2012 Members of the new subsample K are asked to fill in an integrated version of the individual and biographical questionnaire. Members of subsample L3 are asked to answer the second part of the biographical background questionnaire. The new questionnaire targeted at children aged 9 to 10 years was implemented for the first time.

2013 Members of the new subsample M1 are asked to fill in an integrated version of the questionnaire. Because the sample is targeted on persons with migrational background questions on the detailed migration biography are included (see chapter MIGSPELL).

The new pupil questionnaire for the 11 – 12 years old was implemented (for all samples).

2015 Members of the new subsample M2 are asked to fill in an integrated version of the questionnaire. Because the sample consists of immigrants to Germany who have arrived since 2010 questions on the detailed migration biography are included (see chapter MIGSPELL).

A series of problems may emerge when combining biographical information and storing data collections spanning multiple waves. This is due to the fact that the biographical information over time both within and between the sub-samples of SOEP is not always consistent with regards to

- *Positioning* (this includes differences among the various surveying instruments, i.e. the Individual Questionnaire and the single Biography Questionnaire 'Lebenslauf' ('life history'), as well as differences in the position of several indicators in the various versions of the questionnaires),
- *Coverage* (this includes both the changes in the targeted population and the partitions of the survey asked of each person and the corresponding indicators used), and
    - *Timing* (this refers to the point in time when the biographical information was collected for a person in relation to the very first survey).

The biography data sets can always be divided into time invariant (e.g. first year of immigration to Germany, first job, place a person grew up) and time dependent (e.g. marital status, number of children, occupational biography) variables. Whereas time invariant information is by definition valid at every point in time after it has been collected, the time dependent information originally collected needs to be updated whenever a change has occurred. Alternatively, the information that is still valid must be included over the entire analysis period under investigation. In other words, since for the most part identical biographical information for different individuals is collected in SOEP at various points in time, all information regarding an eventual status change or an expansion of the original information must be accounted for over the entire time period of the analysis.

A yearly update of the biographical data therefore involves the following tasks:

- *Time dependent information must be*

  - collected for persons answering the survey questions for the first time and

  - carried forward or changed for persons repeating the SOEP interview.

- *Time invariant* information must be integrated into existing data sets for persons answering the survey questions for the first time.

The goal is for all biography relevant information provided to be up-to-date, without any loss of information with respect to the original variables, and in a user friendly form within the framework of the yearly data set updates. The time dependent variables will correspond then to the status of the most recently realized personal interview. The individual steps of the complex revision of the data sets are described in the corresponding documentation.

Additional Information:

- Unless otherwise indicated,[4] the symbol '$' in a variable name or a file name stands for a wave specific prefix or suffix: for example, the variable $KMUTTI from the file $KIND indicates the vector of the variables AKMUTTI, …, ZKMUTTI, BAKMUTTI up to BFKMUTTI from the file AKIND, …, ZKIND to BFKIND. '$$' indicates the survey year (2 digits) and is used as a suffix: for example, NATION$$ stands for NATION84 to NATION15 from the files APGEN to BFPGEN.

- The file BIOLELA is mentioned frequently within the framework of the following documentation of the individual steps needed for generating biographical variables. This file is not a component of the standard updates of the SOEP data sets, but encompasses all of the biographical entries collected until 1996 (in the Individual Questionnaire and the

[4] The SOEP data set released in 2011 (SOEP v27) and later will include a two-letter rather than a single-letter wave prefix. Since we came to the end of the Latin alphabet with the letter Z in data release v26, we decided to use the wave prefix BA for the cross-sectional data format.

Biographical Questionnaire) from the SOEP respondents. This file is rather complex due to the differences in the surveying procedures mentioned above and is therefore one central input for nearly all of the following variables on individuals who entered the survey prior to 1996. However, BIOLELA does not contain information necessary for updates (e.g. giving birth after having answered the Biographical Questionnaire). Furthermore, identical information is distributed over a multitude of single variables. The information in BIOLELA is only suitable for very restricted analyses without additional tests and supplements. Beginning with 1997, there are wave-specific $LELA files containing the biography information as collected in the respective year. These files (i.e. BIOLELA and $LELA) can be made available on request to interested users of the SOEP data, and a cumulative long format of these original files is available in our new distribution format SOEPlong (dataset BIOL).

The following table displays in a general overview the full set of biographical information as surveyed in the Biography Questionnaire 'Lebenslauf' ('life history') in 2006 and the current version of the user-friendly edition of this information. The designated numbers in the Biography Questionnaire 'Lebenslauf' ('life history') refer to the 2006 version with all samples fully integrated; due to the multitude of differences in the data collection process (as mentioned above), this does not imply that all of the following named variables were collected from all respondents nor that all information is available accordingly in the final biographical files.

Table 1: Biographical data in SOEP

Files in the SOEP Database	Biography Sub-area	Number of Question in the 'Lebenslauf' Questionnaire (2006)	Comparable Questions in the Youth Question-naire (2006)	SOEP Target Population	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave BE)
PPFAD	Country of birth	2, 3	*61, 62*	All persons surveyed	Individual	No	Available
PPFAD	Year of immigration	4	*63*	For persons not born in Germany	Individual	No	Available
MIGSPELL	Migration biography	*migration modul in CAPI version*		Sample M only	SPELL	No	Available
BIOIMMIG	Immigration biography	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 15a	*64, 65, 66, 67, 68, 69, 70, 71*	For persons not born in Germany	Individual	No	Available
PPFAD	Living in East or West Germany in 1989	16	-	All persons surveyed	Individual	No	Available
BIOPAREN	Place of childhood; Life at childhood residence; grew up with parents, Living together with parents	17,17a, 19, 20	*72, 73, 75, 76*	All persons surveyed	Individual	No	Available
BIOSIB	Number of brothers and sisters	18	*74*	All persons surveyed	Individual	Yes	Available
BIOPAREN	Parents living region, year of birth, year of death, nationality, country of birth	21, 22, 23, 23a	*77, 78, 79*	All persons surveyed	Individual	Partly (year of death from PPFAD)	Available

Files in the SOEP Database	Biography Sub-area	Number of Question in the 'Lebenslauf' Questionnaire (2006)	Comparable Questions in the Youth Question-naire (2006)	SOEP Target Population	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave BE)
BIOPAREN	Religious affiliation of parents	28	*85*	All persons surveyed	Individual	No	Available
BIOSOC	Parents took care about efforts at school	29	*41*	All persons surveyed	Individual	No	Available
BIOSOC	Respondent's last school marks	30	*37*	All persons surveyed	Individual	No	Available
BIOSOC	Relationship to parents during youth	31	*13*	All persons surveyed	Individual	No	Available
BIOSOC	Sport and activities during youth	32, 33, 34, 35	*16, 21,22, 25*	All persons surveyed	Individual	No	Available
PBIOSPE	Occupational biography	36	-	All persons surveyed	Spell	Yes ($P, $PKAL)	Available
BIOSOC	Year and place of acquiring a school degree	37, 38, 41	*27*	All persons surveyed	Individual	No (although possible using $P)	Available
$PGEN	Level of school degree	39, 40, 42	*28*	All persons surveyed	Individual	Yes ($P)	Available
BIOSOC	Number of foreign classmates in last attended school class	43	*45*	All persons surveyed	Individual	No	Available
BIOSOC	Target school degree	44, 45	*29, 30*	All persons surveyed	Individual	No	Available

Files in the SOEP Database	Biography Sub-area	Number of Question in the 'Lebenslauf' Questionnaire (2006)	Comparable Questions in the Youth Question-naire (2006)	SOEP Target Population	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave BE)
BIOSOC	Target vocational degree	53, 54	*48, 49*	All persons surveyed	Individual	No	Available
BIOJOB	First job (age, occupational position, public sector, industry)	55, 56, 57, 58, 59, 60a, 60b	-	All persons surveyed	Individual	Yes, if person previously did not work ($P)	Available
BIOJOB	Occupational changes	61	-	All persons surveyed	Individual	Yes	Available
BIOJOB	Last job (year, scope, public sector branch, occupational position)	62, 63, 64, 65, 66, 67	-	All persons surveyed	Individual	Yes	Available
BIORESID	Year since living personally in current apartment; second residence	68, 69	-	All persons surveyed	Individual	No	Available
BIOBIRTH BIOBRTHM	Births	70	-	All women surveyed; since 2000 men, too	Individual	Yes ($P, $PBRUTTO, $KIND)	Available
BIOMARSY	Family status (marriage biography)	71, 72	-	All persons surveyed	Spell	Yes ($P, $PBRUTTO)	Available
BIOCUPLY	Partnership biography	*partnership modul in CAPI version*		All persons surveyes	SPELL	Yes ($P, $BRUTTO)	Available
MIGSPELL	Migration biography	*migration modul in CAPI version*		Sample M only	SPELL	No	Available

Files in the SOEP Database	Biography Sub-area	Number of Question in the 'Lebenslauf' Questionnaire (2006)	Comparable Questions in the Youth Question-naire (2006)	SOEP Target Population	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave BE)
BIOAGE17	Youth	Youth Questionnaire		16 and 17 year old respondents	Individual	No	Available
BIOAGEL	Newborns	Mother & Child Questionnaire		Mothers of newborns	Individual	No	Available
BIOAGEL	Infants	Questionnaire on children aged 2 to 3 years		Mothers	Individual	No	Available
BIOAGEL	Preschooler	Questionnaire on children aged 5 to 6 years		Mothers	Individual	No	Available
BIOAGEL	Elementary school	Questionnaire on children between the age of seven and eight		Parents	Individual	No	Available
BIOAGEL	Elementary school	Questionnaire on children aged 9 to 10 years		Mothers	Individual	No	Available
BIOEDU	Educational history			All persons surveyed	Individual	Yes	Available
Lifespell	Pre- and Post-Survey history	Drop- out studies		All persons	Spell	$BRUTTO	Available

2 Biographical Information in the Meta File PPFAD (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989)[1]

by Elisabeth Liebau, Christian Schmitt, and Diana Schacht

The file PPFAD includes, among other more survey related variables like responding status, some most important demographical information for each person who has ever participated in SOEP in at least one wave. These are, on the one hand, longitudinally checked data on sex (variable SEX) and the date of birth (year of birth in variable GEBJAHR in 4-digits and month of birth in variable GEBMONAT), and, on the other hand, generated demographic variables on the year of death (TODJAHR and TODINFO), on the country of origin (GERMBORN and CORIGIN), on the year of the first immigration to Germany (IMMIYEAR), on the migration background (MIGBACK and MIGINFO), as well as on the geographic area a person lived in prior to German unification (LOC1989). In the following section, the construction of these generated variables will be explained briefly.

2.1 The Month of Birth in the data set PPFAD

Introduction of variables

From wave T onwards (2003) the data set PPFAD contains – in addition to the year of birth – the month of birth (GEBMONAT). This new variable is accompanied by the supporting variable GEBMOVAL which indicates the data source for the month of birth.

GEBMONAT and GEBMOVAL can take the following characteristics:

- GEBMONAT: Month of birth;

  1 (January) to 12 (December)

- GEBMOVAL: Month of birth— data-source

  1 Generated

  2 Info as stored in PPFAD

  3 Info derived from data set $KIND

  4 Info derived from data set SP (own response)

  5 derived from data set $LELA (own response)

  6 derived from BIOAGE01 (mother-child-questionnaire) (NEW with Wave W / Survey year 2006)

[1] Based on earlier work of Joachim R. Frick, Olaf Groh-Samberg, and Florian Henkel.

7 derived from Youth Questionnaire (own response)
(NEW with Wave Z / Survey year 2009)

The month of birth was asked in wave S individual questionnaire (SP). Furthermore, the month of birth was asked in the biography data set, starting with wave T ($LELA, file not available with the SOEP data distribution). Additionally the month of birth is recorded for all children within the file $KIND (starting with wave T). With wave W, an further source of information was introduced with a number of biographical questionnaires, including the mother-child questionnaire (filled in by mothers of newborns), and a number of additional biographical questionnaires, where parents report on their children around ages three, six and eight (BIOAGE01, BIOAGE03, BIOAGE06, BIOAGE08a, BIOAGE08b). All these biographical questionnaires record the year and month of the child. Information from the Youth Questionnaire (self-response, age 17) is also considered.

All these sources of information are used to derive the month of birth, where valid information is used to replace missings in one or more of the mentioned sources. This procedure provides the relevant information for most of the current panel members. The information remains missing for persons who lack any of the above information, including temporary dropouts or people who exited in a previous wave, and before providing data in any of the sources mentioned above. For some of those persons, the month of birth could be reconstructed (this refers primarily to newborns for whom the month of moving into the household is considered as a proxy in case no other reliable information is available). This reconstruction remains an approximation and might differ from the true month of birth in individual cases.

The variable GEBMOVAL displays an ordinal scaling of the level of reliability, where individual response on one's own date of birth is given preference over derived information, and parent response is considered more reliable at younger ages of the child.

2.2 Construction of variables

The month of birth is constructed in an hierarchical order from the files:

- Generated (basis: $P, $PBRUTTO $KIND)
- $KIND
- SP
- $LELA
- BIOAGE01, BIOAGE03, BIOAGE06, BIOAGE08a, BIOAGE08b
- Youth Questionnaire ($PAGE17)

whereas the latter information overrides the former.

This means the generated information will only be utilized if no further, questionnaire based information for the month of birth is available.

The generated month of birth could only be constructed for people who were born while their parents were members of the SOEP. The information was derived from two sources:

- For newborn children the month of moving into the household was used as an approximation of the real month of birth (relevant file $PBRUTTO).
- For parents who reported a birth in a certain month, a link to the child was established, assigning the month of birth to the child (relevant file $P).

Several adjustments and tests of the generated data have been done which showed that – in the cases in which the generated data was also collected by SP, $LELA, $KIND, and $BIOAGE[n] – the data generation is almost always congruent with the collected data and therefore has proven to be reliable.

Frequencies: Month of Birth and Month of Birth: Data Source (File: PPFAD / up to Wave BF)

Table 1: GEBMONAT Month of Birth

		Frequency	Percent	Cumulative Percent
Valid	-5 Not Present in Version of Questionnaire	17,243	15.15	15.15
	-3 Answer improbable	1	0	15.15
	-1 No Answer	15,719	13.81	28.96
	1 January	7,268	6.,38	35.34
	2 February	6,648	5.84	41.18
	3 March	7,363	6.47	47.65
	4 April	6,624	5.82	53.47
	5 May	6,941	6.1	59.56
	6 June	6,553	5.76	65.32
	7 July	6,971	6.12	71.44
	8 August	6,772	5.95	77.39
	9 September	6,857	6.02	83.42
	10 October	6,563	5.77	89.18
	11 November	6,045	5.31	94.49
	12 December	6,272	5.51	100.00
	Total	113,840	100.00	

Source: SOEP v32, doi: 10.5684/soep.v32

Table 2: GEBMOVAL Month Of Birth, Data Source

		Frequency	Percent	Cumulative Percent
Valid	-5 Not Present in Version of Questionnaire	17,243	15.15	15.15
	-3 Not Valid	1	0	15.15
	-1 No Answer	15,719	13.81	28.96
	1 Generated from gebmonth (parents)	1,524	1.34	30.29
	3 $KIND, Info from mother	11,437	10.05	40.34
	4 Info From SP	26,769	23.51	63.86
	5 Info From $LELA	28,471	25.01	88.87
	6 Info From bioage[n]	7,294	6.41	95.27
	7 Info from $PAGE17	5,382	4.73	100.00
	Total	113,840	100.00	

Source: SOEP v32, doi: 10.5684/soep.v32

2.3 Year of birth

(not generated)

2.4 Year of death

Variable TODJAHR Year of death - 4 digits –

The Variable TODJAHR contains the four digit year entered as the year of death.

Codes

$$$$ effective year entered for persons whose year of death could be determined

(1) from the drop-out file PBR_EXIT8, that is, the outcome of the yearly field work

(2) within the scope of the Infratest-Verbleibstudie (Study conducted by Infratest to follow-up on drop-outs) carried out in 1992

(3) within the scope of the Infratest-Verbleibstudie (Study conducted by Infratest to follow-up on drop-outs) carried out in 2001

(4) within the scope of the Infratest-Verbleibstudie (Study conducted by Infratest to follow-up on drop-outs) carried out in 2007

(5) within the scope of the Infratest-Verbleibstudie (Study conducted by Infratest to follow-up on drop-outs) carried out in 2008

Missing codes

(-2) Persons currently living or no longer existing in the sample

Essentially, the deaths of SOEP respondents are reported in the course of the yearly household interview during which the status of the currently living members of the household, as well as the changes due to births and deaths since the last year are surveyed. Furthermore, within the framework of, up to now, three subsequent address investigations of SOEP drop-outs ("Infratest-Verbleibstudie"), demographical drop outs due to mortality or move abroad have been identified. The mortality information is used in generating the variable TODJAHR.

In the first "Verbleibstudie" conducted from April to June in 1992 a total of 53 persons could be identified as deceased. In incorporating this information into the variable TODJAHR attention was given to the fact that an exact year of death could be determined for only 35 of

[8] Help for old friends: The file PBR_EXIT includes all observations that exited from survey households since the previous wave for demographic reasons (death, emigration). Together with the file PBR_HHCH (covering observations who changed household from one wave to the next) these two files replace the file YPBRUTTO used in former releases of SOEP data.

these persons. An exact date was missing for 16 persons, that is, only the qualitative information on their death was available. As a substitute for these cases, the year of the Wave in which the person dropped out of SOEP was used. For 2 persons implausible entries were corrected.

Within the scope of the second Infratest-"Verbleibstudie" conducted in 2001, over 700 persons were identified as deceased. Included in this number are multiple identifications, i.e., persons who were already determined to be deceased through the standard follow up process or in course of the first "Verbleibstudie 1992" mentioned above. This displays essentially a very high correspondence of results from the standard follow up and the ex-post determination of the time of death. For 10 persons the missing information on the year of death was imputed with the help of the year in which they dropped out of the SOEP sample.

In the few cases in which there were conflicting information between the first two follow-up studies and the information from PBR_EXIT (formerly YPBRUTTO), in principle the information from the "Verbleibstudie" was used.

In the third of those studies, another 21 individuals were identified as deceased between 2001 and 2005. For 18 of those persons a valid year of death could be investigated, the remaining three observations are set to the standard missing code "-1".

Again, some of these deaths have also been registered in the most recent of those follow-up studies which was carried out in 2008. In this study a total of 982 individuals were identified as deceased some of which date back to the late 1980s.

Variable TODINFO Year of death – source of information

Codes

1 'from continued surveying (PBR_EXIT / YPBRUTTO)'

2 'Infratest-Verbleibstudie (Follow-up Study) 1992'

3 'Infratest-Verbleibstudie (Follow-up Study) 2001'

4 'Infratest-Verbleibstudie (Follow-up Study) 2007'

5 'Infratest-Verbleibstudie (Follow-up Study) 2008'

For all of the persons who could be identified as deceased, the variable TODINFO contains the corresponding source of information.

2.5 Immigration information

Introduction of variables

The SOEP data comprises a sizeable number of immigrants to Germany and their descendants. Several user-friendly variables identify these groups (GERMBORN, CORIGIN, IMMIYEAR, MIGBACK, MIGINFO) and thus give information on the migration background of all persons who have ever been a part of a SOEP household (i.e., the population from PPFAD). In detail, GERMBORN and CORIGIN give information on the country of birth with the exception of persons who immigrated to Germany before 1950 who are considered being born in Germany (1949 was the founding year of the Federal Republic of Germany). IMMIYEAR specifies the last year of immigration to the Federal Republic of Germany for all persons considered being born in Germany and MIGBACK is useful to identify immigrant descendants by combining information on respondents and their parents. Last, MIGINFO indicates the quality of information given in MIGBACK.

All SOEP samples include immigrants to Germany and their descendants. The shares vary however across samples depending on the target population covered in these samples. Naturally, samples covering the entire German population (Sample A, E, F, H, I, J, K, L1, L2 and L3) or specific groups such as persons from the former GDR (Sample C) comprise a lesser number of immigrants and their descendants than the samples covering foreigners and migrants in particular (Sample B, D, M1 and M2). Furthermore, it has to be noted that the share of missing values within sub-sample G is particularly high, since most participants within the high income sub-sample (Sample G) were not asked to state their country of birth when the sample was launched (around 900 cases).[9]

Table 3 illustrates the share of persons who were born in Germany (or immigrated before 1950) and those who immigrated to Germany since 1950 differentiating between SOEP sub-samples for those members of SOEP households who gave at least one interview ($netto).

[9] For more information see: Liebau E. and Tucci I. (2015) Migrations- und Integrationsforschung mit dem SOEP von 1984 bis 2012: Erhebung, Indikatoren und Potenziale. DIW Berlin: SOEP Survey Papers 270.

Table 3: GERMBORN distribution across SOEP samples (A to M2)

	Born in Germany or immigrated before 1950 (1)	Immigrated to Germany since 1950 (2)	Missing values (-1/-2)	Total
[1] A Initial Sample (West, 1984)	13,955	457	163	14,575
%	*95.8*	*3.1*	*1.1*	*100.0*
[2] B Migration	1,518	3,871	98	5,487
%	*27.7*	*70.6*	*1.8*	*100.0*
[3] C Original Sample (East)	6,865	94	11	6,970
%	*98.5*	*1.4*	*0.2*	*100.0*
[4] D Migration 1994/5 (1984-92/94 West)	589	1,001	7	1,597
%	*36.9*	*62.7*	*0.4*	*100.0*
[5] E Refreshment 1998	2,204	200	21	2,425
%	*90.9*	*8.3*	*0.9*	*100.0*
[6] F Refreshment 2000	12,113	1,467	14	13,594
%	*89.1*	*10.8*	*0.1*	*100.0*
[7] G High-Income 2002	2,422	127	710	3,259
%	*74.3*	*3.9*	*21.8*	*100.0*
[8] H Refreshment 2006	2,684	214	76	2,974
%	*90.3*	*7.2*	*2.6*	*100.0*
[9] I Innovation (Incentives) 2009	2,346	269	77	2,692
%	*87.2*	*10*	*2.9*	*100.0*
[10] J Refreshment 2011	4,987	703	6	5,696
%	*87.6*	*12.3*	*0.1*	*100.0*
[11] K Refreshment 2012	2,381	306	4	2,691
%	*88.5*	*11.4*	*0.2*	*100.0*
[12] L1Birth Cohorts 2010 (2007-2010)	3,086	979	23	4,088
%	*75.5*	*24*	*0.6*	*100.0*
[13] L2 Family Types 2010	4,836	597	41	5,474
%	*88.3*	*10.9*	*0.8*	*100.0*
[14] L3 Family Types 2011	1,797	124	7	1,928
%	*93.2*	*6.4*	*0.4*	*100.0*
[15] M1 Migration 2013 (1995-2010)	1,473	4,003	0	5,476
%	*26.9*	*73.1*	*0*	*100.0*
[16] M2 Migration 2015 (2009-2013)	130	1,581	0	1,711
%	*7.6*	*92.4*	*0*	*100.0*
Total No.	63,386	15,993	1,258	80,637
%	*78.6*	*19.8*	*1.6*	*100.0*

Source: All survey participants with at least one SOEP interview from 1984 to 2015 (n=80,637); SOEP v32, doi: 10.5684/soep.v32.

Variables GERMBORN, CORIGIN and IMMIYEAR

Information for GERMBORN, CORIGIN and IMMIYEAR is collected primarily from the wave-specific individual questionnaires ($P or $PAUSL) or the variations of the "Biography/Life history" questionnaires (integrated biographical data files for Waves A to L in BIOLELA or life course information on first-time respondents since Wave M in $LELA) and from the additional 16-17-year-old questionnaire in use since 2000 ($JUGEND *Table 4* lists information used for generating GERMBORN, CORIGIN and IMMIYEAR.

Table 4: Information used for GERMBORN, CORIGIN and IMMIYEAR

File	Information used
BIOLELA / $LELA	Country of birth
BIOLELA / $LELA	Year of immigration to Germany
BIOLELA	Area of origin
BIOLELA	Has always lived in Germany since immigration
$LELA	Born in Germany
$LELA	Immigration status
$LELA	Flight and expulsion until 1950 (yes/no)
$LELA	Mother: Born In Germany
$LELA	Country of birth of mother
$JUGEND	Born in Germany
$JUGEND	Country of birth
$JUGEND	Year of immigration to Germany
$JUGEND	Immigration status
$JUGEND	Citizenship
$JUGEND	Second citizenship
$JUGEND	Previous citizenship
$JUGEND	Mother: Born In Germany
$JUGEND	Country of birth of mother
$PAUSL / $P	Born in Germany
$PAUSL / $P	Country of birth
$PAUSL / $P	Year of immigration to Germany
$P	Area of origin
$P	Citizenship
$P	Second citizenship
MP	Emigrant of German descent from Eastern Europe
MIGSPELL	Last year of immigration to Germany
$P_MIG	Immigration status
$P_MIG	Previous citizenship
$PGEN	Citizenship (*NATION$$*)
$PBRUTTO	Citizenship *(PNAT$)*
$PBRUTTO	Member of household *($PZUG)*
LPBRUTTO	Country of birth
LPBRUTTO	Immigration group

Source: SOEP v32, doi: 10.5684/soep.v32.

In the following sections, the variables GERMBORN, CORIGIN and IMMIYEAR are described in detail. Special attention is given to the filtering function of GERMBORN for CORIGIN and IMMIYEAR.

Variable GERMBORN "Born in Germany"

Value	Value Label	Frequency	Percent	Cumulative Percent
-2	does not apply	6,803	6.0	6.0
-1	no answer / don't know	2,390	2.1	8.1
1	Born in Germany or immigrated before	87,446	76.8	84.9
2	Immigrated to Germany since 1950	17,201	15.1	100.0
Total		113,840	100.0	

Source: SOEP v32, doi: 10.5684/soep.v32.

GERMBORN specifies whether a person was born in Germany or in another country. Persons who immigrated to Germany before 1950 are considered as being born in Germany (1949 was the founding year of the Federal Republic of Germany; see also IMMIYEAR).

Although most of the respondents gave information on their country of birth, for some cases, additional indicators were used to minimize the portion of missing values. These indicators were used in the following order:

- Mothers' country of birth, their immigration history, and their place of residency at the time of the respondents' birth were taken into account to approximate the respondents' most probable place of birth. For instance, when a respondent was born after his/her mother immigrated to Germany, the respondent is considered to have been born in Germany.
- The respondents' immigration year (IMMIYEAR), country of origin (CORIGIN), and citizenship (NATION$$ from PGEN, $P or $JUGEND, or $PNAT from $BRUTTO) were further taken into account to identify a respondents' country of birth. For instance, an immigration year implies another country of birth than Germany.
- Minor children living in a household consisting only of persons born in Germany were considered to also have been born in Germany.
- When sample D was launched, the country of birth was obtained from the address protocol (variable LPHERKFT). This information was used to fill in the missing values in GERMBORN.

If the country of birth is still missing after this procedure, GERMBORN is coded "-1". Permanent non-respondents are assigned the missing value "-2". A high share of these are sample G cases that were not asked to state their country of birth when the sample was launched (around 900 cases).

For persons who according to GERMBORN were not born in Germany, the variables CORIGIN and IMMIYEAR designate the country of origin and the year of immigration to Germany, respectively.

Variable CORIGIN "Country of origin"

Value	Value Label	Frequency	Percent	Cumulative Percent
-2	does not apply	6,803	6.0	6.0
-1	no answer / don't know	2,494	2.2	8.2
1	Germany	87,446	76.8	85.0
2	Turkey	2,625	2.3	87.3
…				
183	Niger	1	0.0	99.8
222	Unspecified Eastern European country	155	0.1	100.0
333	Other unspecified foreign country	50	0.0	100.0
Total		113,840	100.0	

Source: SOEP v32, doi: 10.5684/soep.v32.

CORIGIN contains information on the country of birth for all persons who have ever been a part of a SOEP household (i.e., the population from PPFAD). Respondents who were born in Germany were assigned the code "1" (see GERMBORN). Persons who were not born in Germany were assigned another country of birth than Germany depending on the information given in the wave-specific individual questionnaires ($P or $PAUSL) or the variations of the "Biography/Life history" questionnaires (integrated biographical data files for Waves A to L in BIOLELA or life course information on first-time respondents since Wave M in $LELA), and from the additional questionnaire for 16-17-year-olds in use since 2000 ($JUGEND).

For those respondents who were not born in Germany and whose country of birth could not be determined, additional indicators were used to minimize the number of missing values. These indicators were used in the following order:

- Respondents' country of origin was considered to be the same country as that of their first, second, or previous citizenship if this was not German (NATION$$ from PGEN, $P or $JUGEND, or $PNAT from $BRUTTO).
- Since 1996, first-time respondents' have been asked to state whether they are a member of a broader group of immigrants such as Ethnic Germans from Eastern Europe or whether they are EEA or EU citizens. Respondents citing Eastern Europe as their country of origin were coded "222 - Unspecified Eastern European country," whereas respondents who stated that they were from an EU country or a precursor country were coded "444 – Unspecified country within EU".
- Respondents who stated being from another region than Germany (including East and West Germany before 1989) are coded "333" (see also variables GP10803 to JP108B03).
- Mothers' country of birth was considered to be the respondents' most probable place of birth if the respondent was born before the mother immigrated to Germany.

- When sample D was launched, the country of birth was obtained from the address protocol (variable LPHERKFT). This information was used to fill in further missing values in CORIGIN.

If the country of birth was still missing after this procedure, CORIGIN was coded "-1". CORIGIN includes a few more missing values than GERMBORN due to cases in which it was not possible to determine a country of birth other than Germany. Permanent non-respondents were assigned the missing value "-2". A high share of these are sample G cases that were not asked to state their country of birth when the sample was launched (around 900 cases).

For persons who were born in another country than Germany, IMMIYEAR designates the year of immigration to Germany.

Variable IMMIYEAR "Year of immigration to Germany "

Value	Value Label	Frequency	Percent	Cumulative Percent
-2	does not apply	94,249	82.8	82.8
-1	no answer / don't know	3,364	3.0	85.8
1950		19	0.0	85.8
…				
2015		24	0.0	100.0
Total		113,840	100.0	

Source: SOEP v32, doi: 10.5684/soep.v32.

IMMIYEAR contains information on the year of immigration to Germany for all persons who have ever been a part of a SOEP household (i.e., the population from PPFAD) and who were not born in Germany (see GERMBORN). The information on this variable was collected from the wave-specific individual questionnaires ($P or $PAUSL) or the variations of the "Biography/Life history" questionnaires (integrated biographical data files for Waves A to L in BIOLELA or life course information on first-time respondents since Wave M in $LELA), and from the additional questionnaire for 16-17-year-olds in use since 2000 ($JUGEND). Since sample M, information on all of a respondent's stays in Germany is collected (up to 15 moves between countries, see MIGSPELL). For all cases in which a respondent had more than one stay in Germany, IMMIYEAR contains the respondent's last year of immigration to Germany.

For those respondents who were not born in Germany and whose year of immigration could not be determined, additional indicators were used to minimize the portion of missing values. These indicators were used in the following order:

- When a respondent entered the SOEP for the first time because he/she had just moved into the household from abroad (see $PZUG from $PBRUTTO), the household entry year was considered to be the same as the immigration year.
- Mother's year of immigration was used as a proxy for the respondent when the respondent was born before the mother immigrated to Germany.

If the year of immigration was still missing after this procedure, IMMIYEAR was coded "-1". Persons born in Germany were coded "-2" (see GERMBORN) as were permanent non-respondents. A high share of the latter are sample G cases that were not asked to state their country of birth when the sample was launched (around 900 cases).

Variables MIGBACK and MIGINFO

The SOEP data comprises a sizeable number of immigrants to Germany and their descendants. The variable MIGBACK is useful in identifying the latter. It combines information on respondents' country of birth (see GERMBORN) and parental information such as their place of birth and their citizenship. The information for this variable comes predominantly from PPFAD (GERMBORN), PGEN (NATION$$), and the relevant biographical data sets (BIOPAREN, BIOIMMIG). The variables were also updated using information from the wave-specific individual questionnaires ($P or $PAUSL), the variations of the "Biography/Life history" questionnaires (integrated biographical data files for Waves A to L in BIOLELA or life course information on first-time respondents since Wave M in $LELA), and the additional questionnaire for 16-17-year-olds in use since 2000 ($JUGEND). *Table 5* lists information used for generating MIGBACK and MIGINFO.

Table 5: Information used for MIGBACK and MIGINFO

File	Information/variable used
PPFAD	Born in Germany (*GERMBORN*)
PGEN	Citizenship (*NATION$$*)
$P	Acquisition of German citizenship (at birth/after)
$P	Dual citizenship
$P	Both parents born in Germany (Yes/No)
$JUGEND	Acquisition of German citizenship (at birth/after)
$JUGEND	Mother: Born in Germany
$JUGEND	Mother: Country of birth
$JUGEND	Father: Born in Germany
$JUGEND	Father: Country of birth
BIOIMMIG	Status group of migrant (*BIIMGRP*)
BIOPAREN	Mother: Country of origin (*MORIGIN*)
BIOPAREN	Father: Country of origin (*VORIGIN*)
BIOPAREN	Mother: Citizenship (*MNAT*)
BIOPAREN	Father: Citizenship (*VNAT*)
$LELA	Mother: Born in Germany
$LELA	Father: Born in Germany
$LELA	Mother: Country of birth

File	Information/variable used
$LELA $PBRUTTO AKIND/EKIND	Father: Country of birth PNAT AK07A and EK03A

Source: SOEP v32, doi: 10.5684/soep.v32.

In the following sections, the variables MIGBACK and MIGINFO are described in detail. Special attention is given to the filtering function of GERMBORN for MIGBACK.

Variable MIGBACK "Migration Background"

Value	Value Label	Frequency	Percent	Cumulative Percent
-1	no answer / don't know	6,781	6.0	6.0
1	No migration background	72,619	63.8	69.8
2	Direct migration background	17,201	15.1	84.9
3	Indirect migration background	14,827	13.0	97.9
4	Migration background, not differentiated	2,412	2.1	100.0
	Total	113,840	100.0	

Source: SOEP v32, doi: 10.5684/soep.v32.

Respondents were assigned to the MIGBACK categories based on country of birth (see GERMBORN): Being born in another country than Germany indicates, by definition, a direct migration background (2), while respondents born in Germany may have either no (1) or an indirect (3) migration background. Respondents whose parents had no migration background were assigned the code "1- no migration background", while respondents whose father or mother had a migration background were assigned the code "3 - indirect migration background". However, parental information is not available for all respondents. Whenever this information was missing (see MIGINFO) but the respondent was born in Germany and when further indicators also suggested that there was no migration background (e.g. NATION$$), the respondent was considered having "1 - no migration background" (see MIGINFO). Since some of these respondents may be the descendants of immigrants, MIGBACK may slightly underestimate the number of persons having a "3 - indirect migration background".

Whenever information on a respondent's country of birth was missing (see GERMBORN), the respondent's first, second, and previous citizenships were taken into account. Having had non-German citizenship at some point in time was considered an indicator of a migration background. In a similar manner, BIIMGRP from BIOIMMIG was taken into account. If the migration background could not be differentiated further, however, respondents were assigned the code "4 – Migration background, not differentiable". Given that citizenship is also

available for household members who never participated in an interview (e.g., $PNAT, AK07A, EK03A), MIGBACK contains fewer missing entries than, for example, GERMBORN.

If the migration background is still missing after this procedure, MIGBACK is coded "-1". Note that any updates in related variables may also lead to an update of the MIGBACK variable. For instance, a respondent who never stated his or her citizenship but later states having German citizenship will be classified as having a migration background of some form. This retrospective perspective may lead to updates of the migration background variable with every new wave.

To provide the highest level of transparency possible, we include a variable for the sources used to create the migration background variable: MIGINFO.

Variable MIGINFO "Information source of MIGBACK"

Value	Value Label	Frequency	Percent	Cumulative Percent
-1	no answer / don't know	6,781	6.0	6.0
1	Direct information without parental information	27,728	24.4	30.3
2	Proxy information without parental information	1,294	1.1	31.5
3	Direct information with parental information	51,994	45.7	77.1
4	Proxy information with parental information	26,043	22.9	100.0
	Total	113,840	100.0	

SOEP v32, doi: 10.5684/soep.v32.

MIGINFO can indicate the quality of information given in MIGBACK. MIGINFO provides information about the usage of proxy information in the generation process of MIGBACK due to missing values in respondents' and their parents' migration histories in the SOEP. Overall, MIGINFO can take on four different codes: either direct or proxy information is available on respondents, and either parental information is available or not.

With proxy information, we are referring to information on the respondent reported either by the parents (AK07A or EK03A) or by the interviewer ($PNAT from $PBRUTTO or the respective Infratest information), whereas we consider information on the country of birth (GERMBORN) or information on a respondents' citizenship (NATION$$ from PGEN, $P or $JUGEND) to be direct respondent information. Please note that information on GERMBORN has partially been derived from other indicators (see GERMBORN). Whenever this is the case, MIGINFO reports proxy and not direct information.

The parental information refers to any information on the migration background of the respondents' mother or father or both. This includes information on the country of birth (see GERMBORN), country of citizenship (NATION$$ from PGEN, $P or $JUGEND), or proxy information on the migration history ($PNAT from $PBRUTTO or the respective Infratest information).

MIGBACK information is considered to be particularly reliable for cases coded "3" on MIGINFO, in contrast to the other cases of missing parental information (1 and 2 in MIGINFO). The cases coded "4" in MIGINFO refer primarily to children in SOEP households whose information was reported by their parents.

2.6 Living in East or West Germany in 1989

The variable LOC1989 in the meta-file PPFAD provides information about the geographic area a person lived in *prior to* the German reunification, differentiating "East Germany (DDR incl. East Berlin", "West Germany (Bundesrepublik Deutschland incl. West Berlin)", and "abroad (Ausland)". This information has been generated for all individuals in SOEP.

Variable LOC1989 "Where did you live in 1989?"

Codes

1 East Germany (German Democratic Republic [DDR] including East Berlin)
2 West Germany (Federal Republic of Germany [BRD] including West Berlin)
3 Abroad (Ausland)

Missing Codes

-2 does not apply; born after 1989
-1 not available

After asking this information from all respondents in 2003 (variable TP121 in file TP), a corresponding question has been included in the biography questionnaire since wave U (2004) [Question 16 / variable UB16 in file ULELA, UJ58 in file UJUGEND] which will collect this time-independent information from all future first time respondents. For all respondents interviewed up until 2003, the following information was used as input to generate LOC1989:

- Information on place and date of last school attendance [variables BSSCHEND and BSSCHWO in file BIOSOC / variables $B38 and $B3701 in file $LELA with $ starting in wave U, 2004],

- Sample affiliation [variable PSAMPLE in file PPFAD],
- year moved in at current address [variable BRMOVEIN in file BIORESID / variable $B68 in file $LELA with $ starting in wave U, 2004],
- sample region [variables $SAMPREG in file PPFAD],
- year of first immigration to Germany [variable IMMIYEAR in file PPFAD]
- In case of inconsistent information from these various sources, the data collected in 2003 via variable TP121 and the information from the biography questionnaire collected since 2004 is considered superior. Persons without any individual information and aged less than 18 years in 1989 were assigned parental information, if available.
- The variable LOC1989 is completed for SOEP samples A through M2.

Variable LOC1989

Where did you live in 1989?	Freq.	Percent	Cum.
[-2] trifft nicht zu	30,756	27.02	27.02
[-1] keine Angabe	12,471	10.95	37.97
[1] East Germany (DDR) incl. East Berlin	14,422	12.67	50.64
[2] West Germany (FRG) incl. West Berlin	47,839	42.02	92.66
[3] Abroad (Ausland)	8,352	7.34	100.00
Total	113,840	100.00	

Source: POPULATION of PPFAD SOEP v32, doi: 10.5684/soep.v32

3 MIGSPELL: The Migration-Biography (Samples M1 and M2)

Integrated Version: Waves bd to bf

by Klaudia Erhardt[1]

3.1 Introduction to the new release of MIGSPELL

The previous release, v31, of MIGSPELL included the waves bd and be of the migrant survey within the SOEP. At that time, although the questions were changed in a significant way between 2013 and 2014, the MIGSPELL-generation changes were minor, because it was known that other changes and the inclusion of the new sample M2 would be incorporated in the next wave. Therefore, we postponed the decision on the ultimate structure of MIGSPELL until wave bf was ready to be included.

With this release, v32 of MIGSPELL, the integration of waves bd to bf has been carried out. The integration entailed major modifications in the number and coding of the MIGSPELL variables, due to different operationalizations of the status at entry. But these amendments of the question design apply only to the migration biographies of respondents who were born abroad and their migrations to Germany, but not their recorded moves back to their birth country or to other countries. Questions related to German-born respondents as well as to stays abroad of foreign-born respondents remained unchanged over the waves.

In addition, in the new release, an improved imputation procedure for the replacement of missing dates has been implemented, meaning that fewer spells will drop from analyses due to a missing start date.

We also eliminated some bugs that had been noticed since the last release. So the end of the last spell of a migration biography is the interview month, as was always intended, but in the last release was a constant that served as a place holder in programming that had been forgotten to be replaced.

In the following sections, the variables of MIGSPELL and their generation are described in detail.

[1] Contact: kerhardt@diw.de

3.2 Summary description of MIGSPELL

MIGSPELL is derived from the migration biographies, which are collected from each new respondent of the IAB-SOEP migration samples M1 and M2.[2] The moves of the respondents were captured through a loop structure of the questionnaire (see Figure 1 and Figure 2 on pp. 3 - 4), with the number of loops limited to 15. The $$p_mig files of the SOEP-distribution hold these data in "wide" format: A proper set of variables for each potential loop has been laid out. For MIGSPELL, the original data has been transformed into spell format. Each migration that actually took place is represented by a spell. Additionally, a spell has been generated for the period from birth to first move (if there has been any), or from birth to interview date (if the respondent has not moved). This has the advantage that every respondent from the IAB-SOEP migration samples is represented in the MIGSPELL dataset, not only those who had moved to another country. Because two moves could be captured within one loop cycle, the maximum number of spells a person can have in MIGSPELL is 31. However, this is extremely rare.[3]

MIGSPELL contains data on the moves of foreign-born migrants as well as on the stays abroad of German-born respondents. The requirement was to capture only periods of at least three months, which was generally observed, but not always.

3.3 Overview: the structure of the migration biography questions

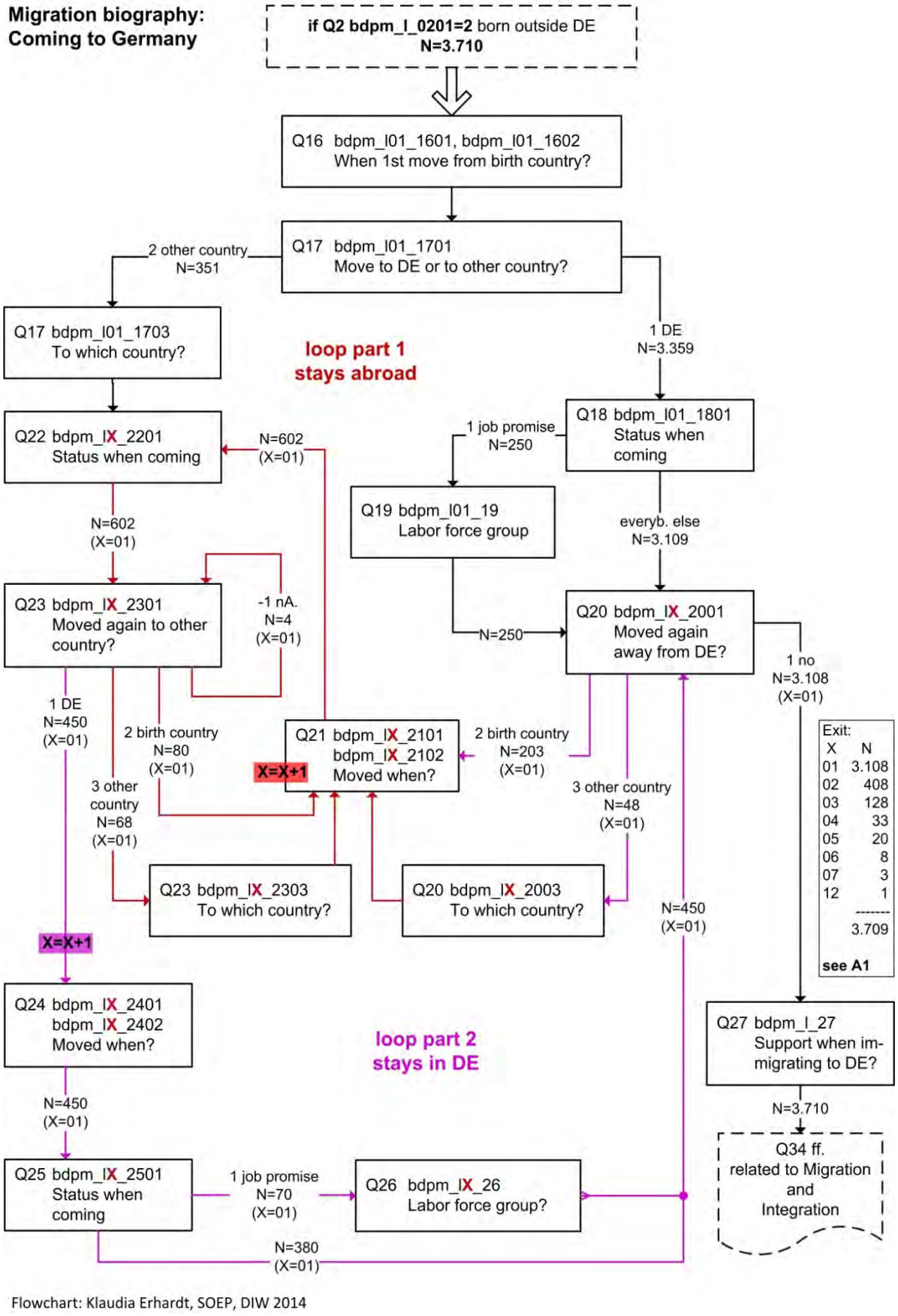
Figure 1 and Figure 2 show two flow charts of the migration biography questions of the IAB-SOEP migration sample. Because compiling such flow charts is very labor-consuming, we provide them only for the first wave of sample M1. As mentioned, the questions to survey the stays abroad of foreign-born and German-born respondents remained unchanged for samples M1 and M2 over all waves. Concerning the moves to Germany of foreign-born respondents, the changes affected mainly the response alternatives (and subsequently, minor filter paths), but not the whole structure as such. For this, the flow charts are still helpful for all the waves to date.

As you see from the "Exit"-lists in the charts, few respondents undergo more than two loop cycles, i.e. undertook more than 4 moves between countries.

[2] See **http://www.diw.de/en/diw_01.c.485464.en/iab_soep.html** for a description and additional documenation of the IAB-SOEP migration sample M1 for waves bd and be (survey years 2013 and 2014). For sample M2 and the survey year 2015, documentation will follow.

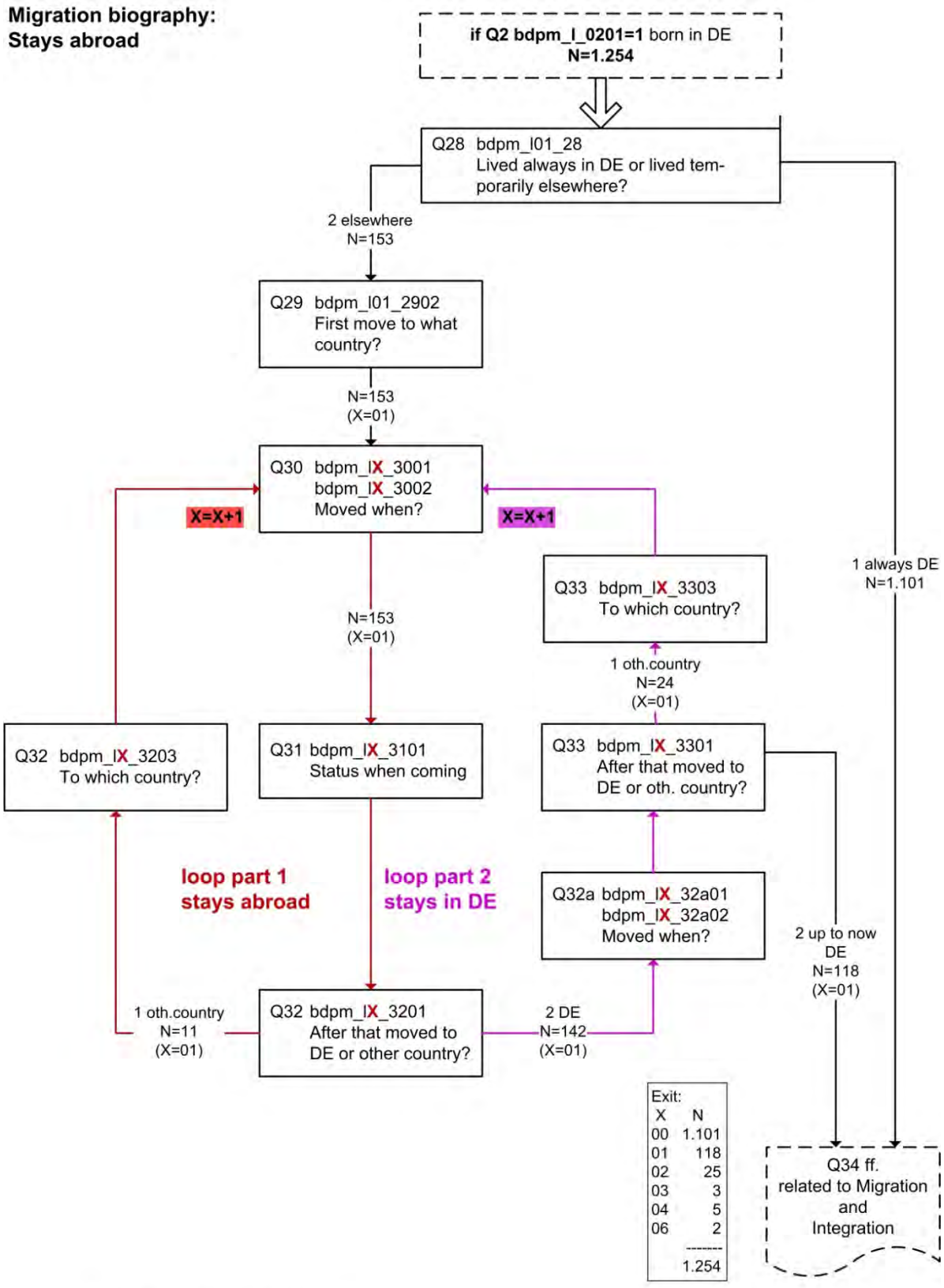
[3] For details on the transformation from wide data to spell data, see: Klaudia Erhardt. 2014. How to Generate Spell Data from Data in “Wide” Format. Based on the migration biographies of the IAB-SOEP Migration Sample. SOEP Survey Papers 228: Series G. Berlin: DIW/SOEP

Figure 1: Flow chart of the "Coming to Germany"-part of the migration biography



Flowchart: Klaudia Erhardt, SOEP, DIW 2014

Figure 2: Flow chart of the "Stays abroad"-part of the migration biography



Flowchart: Klaudia Erhardt, SOEP, DIW 2014

3.4 Description of the variables in MIGSPELL

3.4.1 Summary description of the changes in MIGSPELL

As some of the migration biography questions have changed from wave to wave, it was necessary to map each wave separately to the target variables in MIGSPELL. As a result, the variables in this release of MIGSPELL describing the conditions under which the migration took place have also changed.

This applies mainly to the former MIGSPELL variable **status**. It had to be differentiated into **status1** and **status2**. In the first wave, the status when moving to Germany did not mirror all possible legal conditions of immigration. In the second wave, the question design changed to resolve this, but the adaptation of MIGSPELL was postponed until the current release, in view of further changes of the status-related questions in wave bf.

In addition, the former variables **statusde** and **statusoc**, which only differentiated status according to the type of the stay (a stay in Germany or a stay abroad), have been abandoned in this release. If needed, these variables can easily be constructed by using the **stype** variable.

There have also been changes in the open answers to the status-questions. If respondents choose the category "other", they were asked to specify. In wave bd, open answers were captured not only for moves abroad, but also for moves to Germany. In following waves, this was only the case for moves abroad. In this, v32, release of MIGSPELL the new variable **ostatus** replaces the former variables **ostatus**, **ostatusde** and **ostatusoc**, the two latter variables being redundant to ostatus, because, like statusde and statusoc, they only differentiated according to the type of the stay. Please note: A redesign of the construction and coding of ostatus has taken place (see section 1.4.3).

Variable **lfgroup** has received additional levels and the coding has been changed.

A new variable **jobpr** has been included. Also, several flag variables were added to indicate imputed date values.

The former variable **syear** has been renamend **intyear**, for consistency reasons: In SOEP panel data and in SOEP long, **syear** is the wave indicator, running through years in which a respondent participated in the SOEP, or to which the data relates, respectively. In MIGSPELL, **intyear** is a constant for each respondent, because in the IAB-SOEP-migration samples, the migration biography is surveyed only once, usually in the first wave of participation.

Variable **support** is no longer part of MIGSPELL, because it is not a loop variable, but is related only to the last move to Germany of foreign-born respondents. If needed, this information can be merged to MIGSPELL from the $$p_mig data files (variables bdpm_l_27 for

intyear 2013, bepm_l_21 for intyear 2014, and bfpm_l_2601, bfpm_l_2602, bfpm_l_2603 for intyear 2015).

3.4.2 Synopsis of the variables in MIGSPELL (systematic order)

Table 1: Variables in MIGSPELL

Variable	Variable label	Description
Identificators		
hhnr	Original Household Number	Identifier throughout the SOEP
persnr	Never Changing Person ID	Identifier throughout the SOEP
mignr	Running no. of move	MIGPELL-specific identifier. The combination of persnr and mignr identifies a spell. Runs from 0 because first spell is not a move but episode from birth to first move or interview date, respectively.
Time-invariant Characteristics		
intyear	Interview year	Year when migration biography was surveyed
migfilter	German-born/abroad-born	
bcountry	Country of birth	
lastmig	Year of last move to Germany	= starty of the last spell of a person (only foreign-borns)
lastmig_imp	Year of last move to Germany (imputed version)	= starty_imp of the last spell of a person (only foreign-borns)
Time-variant Characteristics		
starty	Start year	Start year of an episode. Birth year in first spell of a respondent.
startmo	Start month	Start month of an episode. Birth month in first spell of a respondent.
starty_imp	Start year imputed	Same as starty, with imputed missing values and birth year updated with birth year from ppfad
tartmo_imp	Start month imputed	Same as startmo, with imputed missing values and birth month updated with birth month from ppfad
start	Start (months from jan 1900)	generated from starty_imp and startmo_imp
end	End (months from jan 1900)	= start - 1
stype	Spell type	Move to Germany / move abroad
country	Country of the stay	

Variable	Variable label	Description
status1	Legal background of entry	see variable concordance table (Table 8, page 1)
status2	Status at entry	see variable concordance table (Table 8, page 1)
ostatus	Status at entry: unstandard. answers	derived from open answers of level "other", see variable concordance table (Table 8, page1)
jobpr	Job agreement at entry	see variable concordance table (Table 8, page1)
lfgroup	Labor force group	see variable concordance table (Table 8, page 1)
move	Type of move	to Germany / back to birth country / to another country
nmtype	Type of next move	to Germany / back to birth country / to another country
tcountry	Target country of next move	
staytime	Duration of stay (months)	end - start + 1 (generated from imputed date variables)
Technical Variables		
censor	Censor	Censor Variable generated from original dates
censor_imp	Censor imput. version	Censor Variable generated from imputed dates
nspells	Number of spells	Number of moves of a respondent + 1 (= mignr(max) + 1)
f_birth-date_corr	Flag: birth-date updated with birth-date from ppfad	= 1 if starty_imp and/or startmo_imp of first spell of a person is updated with ppfad if a later interview gave different birth-date information
f_starty_imp	Flag: starty imputed	= 1 if missing starty has been imputed
f_startmo_imp	Flag: startmo imputed	= 1 if missing startmo has been imputed

3.4.3 Levels and value labels of the categorical variables (alphabetical order)

bcountry Birth country

Coding according to SOEP-convention.

Source variables see section 1.6.1, Table 2 to Table 7

censor Censor Marker

Generated variable, based on the original start variables

0 Uncensored
1 Right censored
2 Right miss. censored
3 Left censored
4 Left and right censored
5 Left and right miss. censored
6 Left miss. censored
7 Left miss. and right censored
8 Left miss. and right miss. censored

censor_imp Censor Marker, imputed version.

Generated variable, based on the imputed start variables. Same levels as censor

country Country of the stay.

Coding according to SOEP-convention, = tcountry of preceding spell

end End date of a spell as months from jan 1900

Generated variable = start of the preceding spell + 1

In last spell, end is generated from interview date (welle in $$p_mig, and bdpm_pmonin, bepm_monin, or bfpm_monin)

f_birth-date_corr Flag variable

= 1 if starty and/or startmo of first spell was updated with birth-date from ppfad

f_startmo_imp Flag variable

= 1 if missing value in startmo was replaced by imputed value

f_starty_imp Flag variable

= 1 if missing value in starty was replaced by imputed value

intyear Year when the migration biography was surveyed

= variable welle in the $$p_mig files

jobpr Job agreement at entry

Generated Variable. Source variables see section 1.6.2, Table 8

1 Yes (undiff.: only stays outside Germany, and stays in Germany before bf)
2 Prospective job (since bf)
3 Employment contract (since bf)

4 Job as self-employed (since bf)
5 No (since be)
6 Did not look for job (since bf)
7 Does not apply, was a child (since bf)

lastmig starty of last move

= -2 for German-born respondents

lastmig_imp starty_imp of last move

= -2 for German-born respondents

lfgroup Labor force group

Generated Variable. Source variables see section 1.6.2, Table 8

1 Seasonal worker, contract for work and labor
2 Highly qualified and experts with special entry conditions
3 Qualified labor force with priority check by the Fed. Work Agency (not: bd, be)
4 Other labor force with priority check by the Fed. Work Agency (not: bd, be)
5 Trainee, au pair (not bd)
6 Self-employed, entrepreneur
7 Other
8 Relocated to Germany by employer (only bd, be)
9 Sent to Germany by company (only bd, be)

migfilter Marker for German-born/abroad-born respondents

Generated variable, based on bdpm_l_0201, bepm_l_0301, or bfpm_l_0301.

1 Born in Germany
2 Born abroad

move Type of move

= nmtype of preceding spell

nmtype Type of next move

Source variables see section 1.6.1, Table 2 to Table 7

1 To Germany
2 Back to country of birth
3 To another country

nspells Number of spells of a person.

Generated variable

ostatus Status at entry ("other" open answers)

Generated from open answers to status-variables. Replaces variables ostatus, ostatusde and ostatusoc of former releases of migspell.

Open answers are captured for the status variables related to stays abroad. For stays in Germany, open answers have been captured only in the first wave (bd).

In former releases of migspell, coded open answers have been assigned directly to the variable status (and statusde, statusoc), if they fitted a level.

In contrast, in this release of migspell, ostatus is designed to mirror if a level fits to a level of status1 or status2, but no direct assignment to status1 or status2 has been made. The first digit of the two-digit codes in ostatus indicates if the level fits variables status1 or status2, the second digit indicates which level of status1 or status2 a certain level of ostatus fits. Example: open answers with the meaning "Spouse, child, or family member" were coded 23 in ostatus, because the standardized answers with the same meaning are assigned to code 3 of status2.

One-digit codes do not correspond to a level of status1 or status2

1 Visit of family or friends
2 Love attachment
3 Au pair, gap year spent on voluntary social work
4 Intern, trainee
5 Military service, soldier
11 German migrant from Eastern Europe
21 Labor force (not bd)
22 Labor force with job agreement at entry
23 Spouse, child, family member
24 Asylum seeker, refugee
25 Student, apprentice
26 Seeking for job
27 Tourist
28 With tourist visa
29 None of these / other

start Start date of a spell in months from jan 1900

Generated variable from starty_imp and startmo_imp

startmo Start month of an episode (original values)

Source variables see section 1.6.1, Table 2 to Table 7

startmo_imp Start month of an episode (imputed values)

Missing values in startmo are replaced with imputed values, where possible (see section 1.5)

starty Start year of an episode. (original values)

Source variables see section 1.6.1, Table 2 to Table 7

starty_imp Start year of an episode (imputed values)

Missing values in starty are replaced with imputed values, where possible (see section 1.5)

status1 Legal background of entry

Generated Variable. Mapping of source variables to levels of status1 see section 1.6.2, Table 8.

status1 and status2 replace variables status, statusde, statusoc of former releases of migspell

1 German migrant from Eastern Europe

2 German citizen, grown-up outside Germany (since be)

3 EU-citizen (only be)

4 EU- or EEZ-citizen with right to free movement (since bf)

5 EU- or EEZ-citizen without right to free movement (since bf)

6 Other citizens

status2 Status at entry

Generated Variable. Mapping of source variables to levels of status2 see section 1.6.2, Table 8

status1 and status2 replace variables status, statusde, statusoc of former releases of migspell

1 Labor force (not bd)

2 Labor force with job agreement at entry

3 Spouse, child, family member

4 Asylum seeker, refugee

5 Student, trainee

6 Seeking for job

7 Tourist

8 With tourist visa

9 None of these / other

staytime Duration of stay (months)

Generated from imputed start and end variables: = end - start + 1

stype Spell type (marker for type of stay)

Generated from variable move

1 Stay in Germany

2 Stay abroad

tcountry Target country of the next move

Coding according to SOEP-convention

Source variables see section 1.6.1, Table 2 to Table 7

3.4.4 The date-variables in MIGSPELL

With spell data, the information on start date and end date of each episode is crucial. The migration biographies constitute a special kind of spell data, without parallelities and gaps: Because the questionnaire did not allow for reporting more than one place of habitation at the same time, the spells of a person can not be parallel to each other. Further - as one necessarily has to stay at one place or another - the migration biographies have no true gaps, the stays are successive to each other.

The participants were only asked at what time they moved to a certain country, but not at what time they left. Because of the successive character of the data at hand, the end of a spell could be derived from the start of the next spell.

However, sometimes respondents were not able to recollect the exact dates of their moves, or they named dates that contradicted the time sequence of moves, which led to missing values in the start date variables. The respondents were not bullied into stating a date and time contradictions were not immediately clarified during the interview, so these kinds of missings were to be expected. Overall, 369 spells (=2.4%) have a missing value in the original startyear and/or the startmonth.

In order to make those spells accessible for data analyses, the missing values were replaced by imputed values to the extent possible, resulting in only 9 spells with missing values in the imputed version of the startyear and/or startmonth.

From the imputed versions of startyear and startmonth the **start** and **end** variables have been generated. They contain integers counting the month that have passed since January 1, 1900, (i.e. including January). This is in contrast to the SOEP standard, where the counting of time begins January 1, 1983, and also to the Stata standard, where the counting begins January 1, 1960.

We had to advance the zero-point of the time scale, because otherwise the SOEP-convention on missing codes (which are integers < 0 and > -10) would have conflicted: The migration biographies may begin earlier than January 1983 or 1960, which would result in negative values if we had used the SOEP or Stata standard. In certain cases, missing and valid codes would then have become indistinguishable.

If you are using Stata for data analyses and want to benefit from the inbuilt Stata time and date functions, you can transform the **start** and **end** variables into the Stata standard by subtracting 271. But before doing so, you have to replace the SOEP missing values by values that are distinguishable from valid values, such as Stata missing codes:

```
gen start_stata = start
gen end_stata = end
recode start_stata end_stata (-1 = .a) (-3 = .b)
replace start_stata = start_stata - 721
replace end_stata = end_stata - 721
format start_stata end_stata %tmCCYY_mon
```

3.5 The imputation of missing date values

3.5.1 General remarks

As already mentioned, the **start** and **end** variables have been generated from the *imputed* startyear and startmonth information. The original variables **starty** and **startmo** have not been touched, so that they can be used for analyses instead of the imputed versions. All replacements of values have been made in the variables **starty_imp** and **startmo_imp**.

In this section we explain how the imputation was performed.

Before presenting the applied rules and algorithms for the replacement of missing dates, the treatment of the special case "birth-date" is described.

The birth-date is represented in starty and startmo of the first spell of a respondent, which relates to the episode from birth to the first move to another country. If there had been missing values in the birth-date (which was not the case), they would not have been subject to imputation. In the first spell, there is no floor for possible values, and therefore a span for the estimation is not to be determined. However, the birth-date was compared with the birth-date from ppfad. The birth-date is surveyed in each wave anew, and ppfad holds the newest answer of the respondent. If there was a difference between the birth-date from starty and startmo with the one from ppfad, starty_imp and starty_mon was updated with the newer information, and the flag variables f_startmo_imp and f_starty_imp, respectively, were set to 1.

In one case, the difference was more than 28 years, which was regarded as an improbable value, and the update of starty_imp and startmo_imp was undone.

3.5.2 Procedure for the imputation of the missing date values

In the next section, the imputation of startyear and startmonth for the missing dates is explained. The corresponding Stata syntax is too complex to be part of this documentation. In case of questions please contact the author via kerhardt@diw.de

3.5.2.1 The imputation of missing values in the startyear

In a first step, the consistency of the startyears' sequence is tested and set to -3 if it is inconsistent, i.e. the startyear is earlier than the startyear of the previous spell. After that, all spells with a missing value (-1 or -3) are flagged. Then a marker is generated for spells with directly succeeding missing startyears.

For series of more than one spell with a missing startyear, rule 1 applies:

Rule 1: If there are more than one missing startyear in directly succeeding spells, they will not be imputed. Instead, the startmonth is also set to missing, namely, set to the value of the startyear (which is either -1 or -3).

Single spells with a missing value in startyear are treated as follows:

Rule 2: The missing startyear is replaced with the startyear of the next spell minus 1. If the next spell is the last spell of a case, the missing startyear is replaced by the interview year minus 1.

If the such imputed startyear is smaller than the startyear of the preceding spell, or if it is equal to the startyear of the preceding spell but the startmonth is smaller than the startmonth of the preceding spell (that is to say, if the imputed start time is earlier than the start time of the preceding spell), then the subtraction of 1 from the startyear of the next spell is undone, the imputed startyear equals the startyear of the next spell.

With this procedure, in a rare constellation, the spell with a missing startyear conflicts with the time sequence of either the preceding or the following spell. For example: The spell with a missing startyear has a startmonth "May", the preceding spell starts June, 2011, the following spell starts February, 2012. To handle this kind of constellation, a second test of the sequence consistency is performed, resulting in a reset of the originally missing, now imputed startyear to missing. In practice, mostly the startmonth is also missing if the startyear is missing, so that the chance to meet this constellation is very small.

Following rule 2, a spell with an imputed startyear lasts one year and 11 months maximally.

3.5.2.2 The imputation of missing values in the startmonth

The imputation of missing startmonths is only done if the imputed startyear is not missing. Otherwise, there is no sense in keeping or imputing the startmonth information, and startmo_imp is set to the same missing code than starty_imp (either -1 or -3). So the frame for the imputation of missing startmonths is always the startyear of the spell in question.

For the imputation of missing startmonths, rule 1 does not apply. More than one succeding missing startmonth may be imputed if the startyear is known, such that a lower and upper limit for the missing data can be established. Multiple succeeding spells with missing start-month information are referred to as a "series" of spells with a missing startmonth.

The principle of the imputation of a series of spells with missing startmonth is to portion the time between the last and the next established startmonth of the same year onto the spells with a missing startmonth that lay in between.

It has to be distinguished between cases where a) all spells in a year have missing startmonths and b) at least one spell in a year has a valid startmonth.

Considerations for case a) all spells of a year have missing startmonths:

Rule 3: 12 months are divided between the number of succeeding spells with a missing startmonth in the same year. The algorithms are:

a1) startmo_imp = 1 + round(c * span)

a2) span = 12 / ctot + 1

with:
c = running number of the spell with a missing startmonth (the succeeding spells with missing starmonth within a year numbered from 1 to n)
ctot = total of succeeding spells with missing startmonth within a year

EXAMPLE:
startmonth before imputation: -3, -1, -1
span = 12 / 3+1 = 3
startmonth after imputation: 4, 7, 10
(1 + 1 * 3 / 1 + 2*3 / 1 + 3*3)

NOTE: the algorithm means that only 11 spells with a missing startmonth can be placed within one year, although there is enough "place" for 12 spells, if each lasts 1 month. The reason for that is the "+ 1" in formula a1). Only by this, a single spell in a year with missing startmonth is assigned the desired result 7 instead of 6.

As the provisions for surveying the migration biographies say to capture only stays that lasted at least 3 months (which was not always met, however), the constellation of 12 spells with a missing startmonth in a year does not occur empirically.

Considerations for case b) there is at least one spell with a valid startmonth in the same year:

Rule 4 The adjacent spells with a valid startmonth in the same year form the lower and upper limits of the time span that can be divided between the number of succeeding spells with a missing startmonth in the same year. If a series of spells with a missing startmonth begin in the first or end in the last month in a year, the lower limit is 1, the upper limit is 12, respectively. Otherwise, the lower limit is the startmonth of the last spell before the series with missing startmonth, the upper limit is the startmonth of the next spell after the series with missing startmonth.

The algorithms are:

b1) startmo_imp = lol + round(c * span)

b2) span = upl - lol + 1 / ctot + 1

with:
lol = lower limit = startmonth of the last spell with valid startmonth before the series of spells with missing startmonth, or 1 respectively
upl = upper limit startmonth of the next spell with valid startmonth before the series of spells with missing startmonth, or 12 respectively
c and **ctot** as above.

We now understand, that formulas a1) and a2) are only special cases of formulas b1) and b2).

NOTE: The minimum permissible value for span is 1. In other words, between lower and upper limit there must be a gap of at least as many free time units (months) as there are spells with a missing startmonth in the series.

EXAMPLE:
startmonth before imputation: 2, -1, 3, 10
span = 10-2 / 2+1 = 8/3 = 2,66
imputed startmonths: 2 + round(1 * 2,66) = 5
2 + round(2 * 2,66) = 7
startmonth after imputation: 2, 5, 7, 10

After the imputation of missing values in starty and startmo, the variables starty_imp and startmo_imp are either both missing or both non-missing, because we decided a) to impute all missing startmonths if starty was valid and b) not to keep a nonmissing startmonth if starty was missing.

From the imputed variables starty_imp and startmo_imp the variable **start** has been generated by:

```
gen start = cond(starty_imp > 0 & startmo_imp > 0, ///
((starty_imp-1900)* 12) + startmo_imp, starty_imp)
```

meaning: if starty_imp and startmo_imp are not missing, start is calculated as:

```
starty_imp - 1900 * 12) + startmo_imp
```

otherwise it is equal to starty_imp (which is a missing code).

The **end** of a spell is generated by:

```
by persnr: gen end = cond(start[_n+1] > 0, start[_n+1]-1,
start[_n+1])
```

meaning: end is calculated as the start of the next spell minus 1, provided that the start of the next spell is not missing and that the next spell belongs to the same case.

3.6 The integration of the migration biographies of the three waves 2013-2015 (bd, be, bf)

The questionnaire for the survey of the migration biographies was different from wave to wave. The version for wave 2015 is expected to be a final version, but because legal regulations are involved, which may change in future, this is not set in stone. Regardless, the questionnaire for the new respondents of the M1 and M2 IAB-Soep-Migration samples has remained the same for the forthcoming wave bg, whereas the questionnaire for the M3 IAB-Soep-Migration sample (i.e. refugees) is very different, and it is neither sensible nor feasable to include the migration biographies of the M3 sample into MIGSPELL.

An integration of different structures necessarily has to compromise. Consequently, some core variables of MIGSPELL - while mirroring the variables of the bdp_mig-file perfectly in the first wave - now have a more complex relationship to the $$p_mig files. For analyses and their correct interpretation you should be aware of the information that goes - and does not go - into the different categories of the MIGSPELL variables. E.g., some categories systematically receive data only from a single wave, or from a certain group in wave bd, but from another group in wave bf.

To allow for keeping track of the relationship between the variables of the $$p_mig files and the MIGSPELL file, we provide the single structure tables for German and abroad-born respondents of each wave (Table 2 to Table 7) and a synopsis of the levels of each variable of the migration biographies as a source for the core MIGSPELL variables (Table 8) in the following sections.

3.6.1 Structure Tables: $$p_mig-variables to MIGSPELL-variables for waves bd to bf

Table 2: bdp integrated Synopsis 1: Stays abroad (for German-born migrants)

Var##	country	starty	startm	status1	status2	jobpr	lfgroup	nmtype	tcountry	coverage
01	bdpm_l_0203 -2 tnz: DE integr.	bdpm_l_0103	bdpm_l_0102	---			---	bdpm_l01_28	bdpm_l01_2902	all cases
from here: repetition for each value of Var##										
01		bdpm_l01_3001	bdpm_l01_3002	---	bdpm_l01_3101	bdpm_l01_3101	---	bdpm_l01_3201	bdpm_l01_3203	stay abroad
01		bdpm_l01_32a01	bdpm_l01_32a02	---	---	---	---	bdpm_l01_3301	bdpm_l01_3303	stay in DE
02		bdpm_l02_3001	bdpm_l02_3002	---	bdpm_l02_3101	bdpm_l02_3101	---	bdpm_l02_3201	bdpm_l02_3203	stay abroad
02		bdpm_l02_32a01	bdpm_l02_32a02	---	---	---	---	bdpm_l02_3301	bdpm_l02_3303	stay in DE
etc.										
15		bdpm_l15_3001	bdpm_l15_3002	---	bdpm_l15_3101	bdpm_l15_3101	---	bdpm_l15_3201	bdpm_l15_3203	stay abroad
15		bdpm_l15_32a01	bdpm_l15_32a02	---	---	---	---	bdpm_l15_3301	bdpm_l15_3303	stay in DE (up to int.date)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 3: bdp integrated Synopsis 2: Coming to Germany (for migrants not born in Germany)

XX index minus 1 because of yellow marked variables in loop transition 1

Var##	country	starty	startm	status1	status2	jobpr	lfgroup	nmtype	tcountry	coverage
01	bdpm_l_0203 -2 tnz: DE integr.	bdpm_l_0103	bdpm_l_0102	---			---	bdpm_l01_1701	bdpm_l01_1703	all cases
01		bdpm_l01_1601	bdpm_l01_1602	---	bdpm_l01_2201	bdpm_l01_2201	---	bdpm_l01_2301	bdpm_l01_2303	stay abroad (birth country → not DE)
01		bdpm_l01_1601	bdpm_l01_1602	bdpm_l01_1801	bdpm_l01_1801	bdpm_l01_1801	bdpm_l01_19	bdpm_l01_2001	bdpm_l01_2003	stay in DE
01		bdpm_l01_2101	bdpm_l01_2102	---	bdpm_l01_2201	bdpm_l01_2201	---	bdpm_l01_2301	bdpm_l01_2303	stay abroad (birth country → DE)
from here: repetition for each value of Var##										
02		bdpm_l01_2401	bdpm_l01_2402	bdpm_l01_2501	bdpm_l01_2501	bdpm_l01_2501	bdpm_l01_26	bdpm_l02_2001	bdpm_l02_2003	stay in DE
02		bdpm_l02_2101	bdpm_l02_2102	---	bdpm_l02_2201	bdpm_l02_2201	---	bdpm_l02_2301	bdpm_l02_2303	stay abroad
03		bdpm_l02_2401	bdpm_l02_2402	bdpm_l02_2501	bdpm_l02_2501	bdpm_l02_2501	bdpm_l02_26	bdpm_l03_2001	bdpm_l03_2003	stay in DE
03		bdpm_l03_2101	bdpm_l03_2102	---	bdpm_l03_2201	bdpm_l03_2201	---	bdpm_l03_2301	bdpm_l03_2303	stay abroad
etc.										
15		bdpm_l14_2401	bdpm_l14_2402	bdpm_l14_2501	bdpm_l14_2501	bdpm_l14_2501	bdpm_l14_26	bdpm_l15_2001	bdpm_l15_2003	stay in DE
15		bdpm_l15_2101	bdpm_l15_2102	---	bdpm_l15_2201	bdpm_l15_2201	---	bdpm_l15_2301	bdpm_l15_2303	stay abroad
16		bdpm_l15_2401	bdpm_l15_2402	bdpm_l15_2501	bdpm_l15_2501	bdpm_l15_2501	bdpm_l15_26	bdpm_l02_2001	bdpm_l02_2003	stay in DE (up to int.date)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 4: bep integrated Synopsis 1: Stays abroad (for German-born migrants)

Var##	country	starty	startm	status1	status2	jobpr	lfgroup	nmtype	tcountry	coverage
01	bepm_l_0303 -2 tnz: DE integr.	bepm_l_0103	bepm_l_0102	---			---	bepm_l01_22	bepm_l01_2302	all cases
from here: repetition for each value of Var##										
01		bepm_l01_2401	bepm_l01_2402		bepm_l01_2501	bepm_l01_2501	---	bepm_l01_2601	bepm_l01_2603	stay abroad
01		bepm_l01_2701	bepm_l01_2702		---	---	---	bepm_l01_2801	bepm_l01_2803	stay in DE
02		bepm_l02_2401	bepm_l02_2402		bepm_l02_2501	bepm_l02_2501	---	bepm_l02_2601	bepm_l02_2603	stay abroad
02		bepm_l02_2701	bepm_l02_2702		---	---	---	bepm_l02_2801	bepm_l02_2803	stay in DE
etc.										
15		bepm_l15_2401	bepm_l15_2402		bepm_l15_2501	bepm_l15_2501	---	bepm_l15_2601	bepm_l15_2603	stay abroad
15		bepm_l15_2701	bepm_l15_2702		---	---	---	bepm_l15_2801	bepm_l15_2803	stay in DE (up to int.date)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 5: bep integrated Synopsis 2: Coming to Germany (for migrants not born in Germany)

XX index minus 1 because of yellow marked variables in loop transition 1

Var##	country	starty	startm	status1	status2	jobpr	lfgroup	nmtype	tcountry	coverage
01	bepm_l_0303 -2 tnz: DE intgr.	bepm_l_0103	bepm_l_0102	---	---	---	---	bepm_l01_0701	bepm_l01_0703	all cases
01		bepm_l_0601	bepm_l_0602	---	bepm_l01_1401	bepm_l01_1401	---	bepm_l01_1501	bepm_l01_1503	stay abroad (birth country → not DE)
01		bepm_l_0601	bepm_l_0602	bepm_l_08	bepm_l_08 bepm_l_0901 bepm_l_10 bepm_l01_11	bepm_l_0902	bepm_l01_11	bepm_l01_1201	bepm_l01_1203	stay in DE
01		bepm_l01_1301	bepm_l01_1302	---	bepm_l01_1401	bepm_l01_1401		bepm_l01_1501	bepm_l01_1503	stay abroad (birth country → DE → not DE)
from here: repetition for each value of Var##										
02		bepm_l01_1601	bepm_l01_1602	bepm_l01_17	bepm_l01_17 bepm_l01_1801 bepm_l01_19 bepm_l01_20	bepm_l01_1802	bepm_l01_20	bepm_l02_1201	bepm_l02_1203	stay in DE
02		bepm_l02_1301	bepm_l02_1302	---	bepm_l02_1401	bepm_l02_1401	---	bepm_l02_1501	bepm_l02_1503	stay abroad
etc.										
15		bepm_l14_1601	bepm_l14_1602	bepm_l14_17	bepm_l14_17 bepm_l14_1801 bepm_l14_19 bepm_l14_20	bepm_l14_1802	bepm_l14_20	bepm_l15_1201	bepm_l15_1203	stay in DE
15		bepm_l15_1301	bepm_l15_1302	---	bepm_l15_1401	bepm_l15_1401	---	bepm_l15_1501	bepm_l15_1503	stay abroad
16		bepm_l15_1601	bepm_l15_1602	bepm_l15_17	bepm_l15_17 bepm_l15_1801 bepm_l15_19 bepm_l15_20	bepm_l15_1802	bepm_l15_20	---	---	stay in DE (up to int.date)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 6: bfp integrated Synopsis 1: Stays abroad (for German-born migrants)

Var##	country	starty	startm	status1	status2	jobpr	lfgroup	nmtype	tcountry	coverage
01	bfpm_l_0303 -2 tnz: DE integr.	bfpm_l_0103	bfpm_l_0102	---			---	bfpm_l01_34	bfpm_l01_3502	all cases
from here: repetition for each value of Var##										
01		bfpm_l01_3601	bfpm_l01_3602		bfpm_l01_3701	bfpm_l01_3701	---	bfpm_l01_3801	bfpm_l01_3803	stay abroad
01		bfpm_l01_3901	bfpm_l01_3902		---	---	---	bfpm_l01_4001	bfpm_l01_4003	stay in DE
02		bfpm_l02_3601	bfpm_l02_3602		bfpm_l02_3701	bfpm_l02_3701	---	bfpm_l02_3801	bfpm_l02_3803	stay abroad
02		bfpm_l02_3901	bfpm_l02_3902		---	---	---	bfpm_l02_4001	bfpm_l02_4003	stay in DE
etc.										
15		bfpm_l15_3601	bfpm_l15_3602		bfpm_l15_3701	bfpm_l15_3701	---	bfpm_l15_3801	bfpm_l15_3803	stay abroad
15		bfpm_l15_3901	bfpm_l15_3902		---	---	---	bfpm_l15_4001	bfpm_l15_4003	stay in DE (up to int.date)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 7: bfp integrated Synopsis 2: Coming to Germany (for migrants not born in Germany)

XX index minus 1 because of yellow marked variables in loop transition 1

Var##	country	starty	startm	status1	status2	jobpr	lfgroup	nmtype	tcountry	coverage
01	bfpm_l_0303 -2 tnz: DE intgr.	bfpm_l_0103	bfpm_l_0102	---	---	---	---	bfpm_l01_0701	bfpm_l01_0703	all cases
01		bfpm_l_0601	bfpm_l_0602	---	bfpm_l01_1601	bfpm_l01_1601	---	bfpm_l01_1701	bfpm_l01_1703	stay abroad (birth country → not DE)
01		bfpm_l_0601	bfpm_l_0602	bfpm_l_08	bfpm_l_09 bfpm_l_11	bfpm_l_13	bfpm_l_10 bfpm_l_12	bfpm_l01_1401	bfpm_l01_1403	stay in DE
01		bfpm_l01_1501	bfpm_l01_1502	---	bfpm_l01_1601	bfpm_l01_1601		bfpm_l01_1701	bfpm_l01_1703	stay abroad (birth country → DE → not DE)

from here: repetition for each value of Var##

02		bfpm_l01_1801	bepm_l01_1802	bfpm_l01_19	bfpm_l01_20 bfpm_l01_22	bfpm_l01_24	bfpm_l01_21 bfpm_l01_23	bfpm_l02_1401	bfpm_l02_1403	stay in DE
02		bfpm_l02_1501	bfpm_l02_1502	---	bfpm_l02_1601	bfpm_l02_1601	---	bfpm_l02_1701	bfpm_l02_1703	stay abroad
03		bfpm_l02_1801	bfpm_l02_1802	bfpm_l02_19	bfpm_l02_20 bfpm_l02_22	bfpm_l02_24	bfpm_l02_21 bfpm_l02_23	bfpm_l03_1401	bfpm_l03_1403	stay in DE
03		bfpm_l03_1501	bfpm_l03_1502	---	bfpm_l03_1601	bfpm_l03_1601	---	bfpm_l03_1701	bfpm_l03_1703	stay abroad

etc.

15		bfpm_l14_1801	bfpm_l14_1802	bfpm_l14_19	bfpm_l14_20 bfpm_l14_22	bfpm_l14_24	bfpm_l14_21 bfpm_l14_23	bfpm_l15_1401	bfpm_l15_1403	stay in DE
15		bfpm_l15_1501	bfpm_l15_1502	---	bfpm_l15_1601	bfpm_l15_1601	---	bfpm_l15_1701	bfpm_l15_1703	stay abroad
16		bfpm_l15_1801	bfpm_l15_1802	bfpm_l15_19	bfpm_l15_20 bfpm_l15_22	-bfpm_l15_24	bfpm_l15_21 bfpm_l15_23	---	---	stay in DE (up to int.date)

Source: SOEP v32, doi: 10.5684/soep.v32

3.6.2 Synopsis: Mapping of the migbiography variables of waves 2013-2015 to the MIGSPELL variables

Table 8, which follows, shows the mapping of the variables of the single waves to the variables of the new release of MIGSPELL in full detail, level by level. Thus you can see at a glance, for instance, that level 3 of the MIGSPELL variable status1 receives data only from variables of wave be.

Annotation on the representation of the variables in Table 8:

- The numerals (in case of four-digit numerals: the first two digits of the numerals) after the last low line in the variable names indicate the numbers of the related question in the wave-specific questionnaire. The number signs (##) are wild-cards for the loop numerator in the variable name.
- The numbers in parenthesis indicate the levels of the source variables that entered in the respective levels of the MIGSPELL variables.
- The different colors of the variable names refer to different filter paths for different respondent groups and types of moves: black: Born outside Germany / a move to Germany, blue: Born outside Germany / a move abroad, red: Born in Germany / a move abroad.

Table 8: Synopsis of the migration biography variables as source variables for the MIGSPELL variables

bdp_mig	bep_mig	bfp_mig	migspell	signification
status1				
			status1 : -6	
bdpm_l01_1801 (all levels) bdpm_l##_2501 (all levels) bdpm_l01_2201 (all levels) bdpm_l##_3101 (all levels)	bepm_l_08 (4,6,7,8,9) bepm_l##_17 (4,6,7,8,9) bepm_l##_2501 (all levels)		status1 : -5	
			status1 : -3	
	bepm_l_08 (-2) bepm_l##_17 (-2)	bfpm_l_08 (-2) bfpm_l##_19 (-2) bfpm_l##_3701 (all levels)	status1 : -2	
	bepm_l_08 (-1) bepm_l##_17 (-1)	bfpm_l_08 (-1) bfpm_l##_19 (-1)	status1 : -1	
bdpm_l01_1801 (2) bdpm_l##_2501 (2)	bepm_l_08 (1) bepm_l##_17 (1)	bfpm_l_08 (1) bfpm_l##_19 (1)	status1 : 1	German migrant from Eastern Europe

bdp_mig	bep_mig	bfp_mig	migspell	signification
	bepm_l_08 (2) bepm_l##_17 (2)	bfpm_l_08 (2) bfpm_l##_19 (2)	status1 : 2	German citizen, grown-up outside Germany
	bepm_l_08 (3) bepm_l##_17 (3)		status1 : 3	EU-citizen
		bfpm_l_08 (3) bfpm_l##_19 (3)	status1 : 4	EU- or EEZ-citizen with right to free movement
		bfpm_l_08 (4) bfpm_l##_19 (4)	status1 : 5	EU- or EEZ-citizen without right to free movement
		bfpm_l_08 (5) bfpm_l##_19 (5)	status1 : 6	Other citizens
status2				
			status2 : -6	
			status2 : -5	
			status2 : -3	
			status2 : -2	
	bepm_l_08 (-1)		status2 : -1	
	bepm_l_0901 (2) bepm_l_10 (1) bepm_l##_1801 (2) bepm_l##_19 (1)	bfpm_l_09 (1) bfpm_l_11 (1) bfpm_l##_20 (1) bfpm_l##_22 (1)	status2 : 1	Labor force
bdpm_l01_1801 (1) bdpm_l##_2501 (1) bdpm_l##_2201 (1) bdpm_l##_3101 (1)	bepm_l##_1401 (1) bepm_l##_2501 (1)	bfpm_l##_1601 (1) bfpm_l##_3701 (1)	status2 : 2	Labor force with job agreement at entry
bdpm_l01_1801 (3) bdpm_l##_2501 (3) bdpm_l##_2201 (2) bdpm_l##_3101 (2)	bepm_l_08 (6) bepm_l_0901 (4) bepm_l##_17 (6) bepm_l##_1801 (4) bepm_l##_1401 (2) bepm_l##_2501 (2)	bfpm_l_09 (4) bfpm_l_11 (4) bfpm_l##_20 (4) bfpm_l##_22 (4) bfpm_l##_1601 (2) bfpm_l##_3701 (2)	status2 : 3	Spouse, child, family member
bdpm_l01_1801 (4) bdpm_l##_2501 (4) bdpm_l##_2201 (3)	bepm_l_08 (4) bepm_l##_17 (4) bepm_l##_1401 (3)	bfpm_l_11 (5) bfpm_l##_22 (5) bfpm_l##_1601 (3)	status2 : 4	Asylum seeker, refugee
bdpm_l01_1801 (5) bdpm_l##_2501 (5) bdpm_l##_2201 (4) bdpm_l##_3101 (3)	bepm_l_08 (7) bepm_l_0901 (3) bepm_l##_17 (7) bepm_l##_1801 (3) bepm_l##_1401 (4) bepm_l##_2501 (3)	bfpm_l_09 (3) bfpm_l_11 (3) bfpm_l##_20 (3) bfpm_l##_22 (3) bfpm_l##_1601 (4) bfpm_l##_3701 (3))	status2 : 5	Student, trainee

bdp_mig	bep_mig	bfp_mig	migspell	signification
bdpm_l01_1801 (6) bdpm_l##_2501 (6) bdpm_l##_2201 (5) bdpm_l##_3101 (4)	bepm_l_0901 (1) bepm_l01_11 (8) bepm_l##_1801 (1) bepm_l##_20 (8) bepm_l##_1401 (5) bepm_l##_2501 (4)	bfpm_l_09 (2) bfpm_l_11 (2) bfpm_l##_20 (2) bfpm_l##_22 (2) bfpm_l##_1601 (5) bfpm_l##_3701 (4)	status2 : 6	Seeking for job
		bfpm_l_09 (5) bfpm_l_11 (6) bfpm_l##_20 (5) bfpm_l##_22 (6)	status2 : 7	Tourist
	bepm_l_08 (8) bepm_l##_17 (8)		status2 : 8	with tourist visum
bdpm_l01_1801 (7) bdpm_l##_2501 (7) bdpm_l##_2201 (6) bdpm_l##_3101 (5)	bepm_l_08 (9) bepm_l_0901 (5) bepm_l##_17 (9) bepm_l##_1801 (5) bepm_l##_1401 (6) bepm_l##_2501 (5)	bfpm_l_09 (6) bfpm_l_11 (7) bfpm_l##_20 (6) bfpm_l##_22 (7) bfpm_l##_1601 (6) bfpm_l##_3701 (5)	status2 : 9	None of these / other
jobpr				
			jobpr : -6	
bdpm_l01_1801 (1,2,3,4,5,6,7 od. all levels?) bdpm_l##_2501 (1,2,3,4,5,6,7 od. all levels?) bdpm_l01_2201 (all levels?)	bepm_l_0902 (-2,-1,1,2) bepm_l##_1802 (-2,-1,1,2)		jobpr : -5	
			jobpr : -3	
		bfpm_l_13 (-2) bfpm_l##_24 (-2)	jobpr : -2	
		bfpm_l_13 (-1) bfpm_l##_24 (-1)	jobpr : -1	
bdpm_l01_1801 (1) bdpm_l##_2501 (1) bdpm_l01_2201 (1) bdpm_l##_3101 (1)	bepm_l_0902 (1) bepm_l##_1802 (1) bepm_l##_1401 (1) bepm_l##_2501 (1)	bfpm_l##_1601 (1) bfpm_l##_3701 (1)	jobpr : 1	Yes (undiff.)
		bfpm_l_13 (1) bfpm_l##_24 (1)	jobpr : 2	Prospective job
		bfpm_l_13 (2) bfpm_l##_24 (2)	jobpr : 3	Employment contract
		bfpm_l_13 (3) bfpm_l##_24 (3)	jobpr : 4	Job as self-employed
	bepm_l_0902 (2) bepm_l##_1802 (2)	bfpm_l_13 (4) bfpm_l##_24 (4)	jobpr : 5	No

bdp_mig	bep_mig	bfp_mig	migspell	signification
		bfpm_l_13 (5) bfpm_l##_24 (5)	jobpr : 6	Did not look for job
		bfpm_l_13 (6) bfpm_l##_24 (6)	jobpr : 7	Does not apply, was a child
lfgroup				
			lfgroup : -6	
			lfgroup : -5	
			lfgroup : -3	
			lfgroup : -2	
			lfgroup : -1	
bdpm_l01_19 (2) bdpm_l##_26 (2)	bepm_l01_11 (2) bepm_l##_20 (2)	bfpm_l_10 (1) bfpm_l_12 (1) bfpm_l##_21 (1) bfpm_l##_23 (1)	lfgroup : 1	Seasonal worker, contract for work and labor
bdpm_l01_19 (5) bdpm_l##_26 (5)	bepm_l01_11 (5) bepm_l##_20 (5)	bfpm_l_12 (2) bfpm_l##_23 (2)	lfgroup : 2	Highly qualified and experts with special entry conditions
		bfpm_l_12 (3) bfpm_l##_23 (3)	lfgroup : 3	Qualified labor force with priority check by the Fed. Work Agency
		bfpm_l_12 (4) bfpm_l##_23 (4)	lfgroup : 4	Other labor force with priority check by the Fed. Work Agency
	bepm_l01_11 (6,7) bepm_l##_20 (6,7)	bfpm_l_10 (3) bfpm_l_12 (5) bfpm_l##_21 (3) bfpm_l##_23 (5)	lfgroup : 5	Trainee, au pair
bdpm_l01_19 (1) bdpm_l##_26 (1)	bepm_l01_11 (1) bepm_l##_20 (1)	bfpm_l_10 (4) bfpm_l_12 (6) bfpm_l##_21 (4) bfpm_l##_23 (6)	lfgroup : 6	Self-employed, entrepreneur
bdpm_l01_19 (6) bdpm_l##_26 (6)	bepm_l01_11 (9) bepm_l##_20 (9)	bfpm_l_10 (5) bfpm_l_12 (7) bfpm_l##_21 (5) bfpm_l##_23 (7)	lfgroup : 7	Other
bdpm_l01_19 (3) bdpm_l##_26 (3)	bepm_l01_11 (3) bepm_l##_20 (3)		lfgroup : 8	Relocated to Germany by employer
bdpm_l01_19 (4) bdpm_l##_26 (4)	bepm_l01_11 (4) bepm_l##_20 (4)		lfgroup : 9	Sent to Germany by company

Source: SOEP v32, doi: 10.5684/soep.v32

4 Activity Biography in the Files PBIOSPE and ARTKALEN

by Paul Schmelzer and Maik Hamjediers[1]

The spell file PBIOSPE is based on the information on activity status over the life course, which is collected as a matrix from every respondent answering the Biography Questionnaire (Question 45 in 2015).[2] The observations start at the age of 15 and end at the current age (up to age 65). This information on activity status covers only the period up to the time the biography is collected. To update the ongoing occupational career in PBIOSPE, information from the yearly Individual Questionnaire is also used. In this questionnaire, respondents are always asked their occupational status for every month of the previous year (Question 118 in 2015).[3] Therefore, the information on activity status collected on a monthly basis in the yearly personal questionnaire and stored in the file ARTKALEN in spell format is aggregated into yearly values and combined with the information gathered from the Biography Questionnaire.[4]

In the following, the method of combining the data is described. There have been **no changes** how the data is generated since the previous version, distributed in 2015.[5] But if you have been working with older versions of the dataset (versions distributed in 2008 and earlier) you should check the section at the end of the chapter, where you will find information on previous changes. But before we move on to the details, we provide a brief overview of the contents of PBIOSPE. Table 2 contains a list of all the variables in the dataset. The variables BEGIN and END indicate the beginning and the end of a spell. These variables are age entries. There are also variables that refer to calendar years: BEGINY and ENDY (Y stands for Year). The variable SPELLTYP contains information on the activity status during the spell, e.g., employed full-time or unemployed. The SPELLNR is a serial identifier of spells of each activity status of a given person. Missing information on the beginning or end of a spell causes what is known as censoring problems. There are two types of missing data. First, data can be missing on periods outside the observation window (before the age of 15 and after the age of 65). Second, data can be missing on years within the observation window due to item non-response in particular years or due to temporary drop-outs (the latter applies to calendar

[1] Based on earlier work by Rainer Pischner, Henning Lohmann, Marco Giesselmann and Mila Staneva.
[2] See Chapter 1 for general information on the collection of biography information.
[3] For persons who were temporarily unavailable for interviewing, it is sometimes possible to fill in the gaps in their occupational status. If these persons fill out the additional questionnaire for temporary drop-outs later on, we can use the information collected there (see files $PLUECKE).
[4] For more information, see Haisken-DeNew, John and Joachim R. Frick (2005): *DTC - Desktop Companion to the German Socio-Economic Panel Study (SOEP)*, Chapter 3.
[5] The only exception is the lack of information on short work hours (spelltyp '2' in ARTKALEN), due to omitting this question in the questionnaire of 2015.

information only). In this case, we speak of “gaps.” There are nine different patterns (see Table 1).

Table 1: Coding of the variable ZENSOR

Right: / Left:	not censored	censored missing	censored before gap
not censored	1	2	3
censored missing	4	5	6
censored after gap	7	8	9

Note: ‘(99) Gap’ spells are all markes as (-2)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 2: Contents of PBIOSPE (variables)

Variable	Description
HHNR	Original Household Number
PERSNR	Never Changing Person ID
SPELLNR	Serial Number Of The Spell Per Person
SPELLTYP	Type Of Spell
BEGIN	Age Spell Begins
END	Age Spell Ends
BEGINY	Year Spell Begins
ENDY	Year Spell Ends
ZENSOR	Censor Variable
SPELLINF	Spell Construction Information
ERHEBJ	Survey Year Biography Data
KALYEAR	First Observation Year Calendar
BEGINB1	Age Spell Begins, 1st Initial Biography Spell
ENDB1	Age Spell Ends, 1st Initial Biography Spell
BEGINK1	Age Spell Begins, 1st Initial Calendar Spell
ENDK1	Age Spell Ends, 1st Initial Calendar Spell
BEGINYB1	Year Spell Begins, 1st Initial Biography Spell
ENDYB1	Year Spell Ends, 1st Initial Biography Spell
BEGINYK1	Year Spell Begins, 1st Initial Calendar Spell
ENDYK1	Year Spell Ends, 1st Initial Calendar Spell
BEGINB2	Age Spell Begins, 2nd Initial Biography Spell
ENDB2	Age Spell Ends, 2nd Initial Biography Spell
BEGINK2	Age Spell Begins, 2nd Initial Calendar Spell
ENDK2	Age Spell Ends, 2nd Initial Calendar Spell
BEGINYB2	Year Spell Begins, 2nd Initial Biography Spell
ENDYB2	Year Spell Ends, 2nd Initial Biography Spell
BEGINYK2	Year Spell Begins, 2nd Initial Calendar Spell
ENDYK2	Year Spell Ends, 2nd Initial Calendar Spell
BEGINB3	Age Spell Begins, 3rd Initial Biography Spell
ENDB3	Age Spell Ends, 3rd Initial Biography Spell
BEGINK3	Age Spell Begins, 3rd Initial Calendar Spell
ENDK3	Age Spell Ends, 3rd Initial Calendar Spell
BEGINYB3	Year Spell Begins, 3rd Initial Biography Spell
ENDYB3	Year Spell Ends, 3rd Initial Biography Spell
BEGINYK3	Year Spell Begins, 3rd Initial Calendar Spell
ENDYK3	Year Spell Ends, 3rd Initial Calendar Spell
BEGINK4	Year Spell Begins, 4th Initial Biography Spell
ENDK4	Year Spell Ends, 4th Initial Biography Spell
BEGINYK4	Year Spell Begins, 4th Initial Calendar Spell
ENDYK4	Year Spell Ends, 4th Initial Calendar Spell

As mentioned above, PBIOSPE combines information collected in the biography questionnaire and the calendar matrix of the individual questionnaire. The two types of information are merged into PBIOSPE following a number of rules. First of all, it is important to acknowledge that the Biography Questionnaire Matrix as well as the Individual Questionnaire Matrix allow for multiple activity statuses for a given year or month. No concept of main activity is used. A common combination is, for instance, “housewife/-husband” and “working part-time”. There are a number of other plausible combinations, but also combinations that are less plausible. However, a list of valid combinations of activity

statuses defined according to legal or similar constructs would need to be based on very strong assumptions. In addition—in particular in case of the yearly matrix in the Biography Questionnaire—activities are reported that took place in a calendar year in consecutive months, which makes it impossible to exclude combinations of activities. Therefore, no data cleaning is performed at this stage. As a consequence, the data may contain information on more than one activity for a given point in time.

This also defines the rules for aggregating the monthly ARTKALEN data into yearly values. Take, for example, a person who was in full-time employment from January to November 2007, and unemployed in December 2007. The exact months are recorded in the dataset ARTKALEN. In the aggregated data, which is merged with the yearly data from the Biography Questionnaire, you find the information that the person worked full-time and was also unemployed in the year 2007. There is a second level of aggregation of ARTKALEN information as the data on type of activity, which is recorded in the variable SPELLTYP is more detailed than in PBIOSPE. The respective information is aggregated as described in Table 3.

Table 3: Aggregation of ARTKALEN spell information into PBIOSPE

	PBIOSPE	ARTKALEN
1	School/University	School, College (1)
2	Apprenticeship/Training	Vocational Training (4), First Job Training, Apprenticeship (13), Continuing Education, Retraining (14)
3	Military/Civilian service	Military, Community Service (9)
4	Full-time employed	Full-Time Employment (1), Short Work Hrs (2)
5	Part-time employed	Part-Time Employment (3), Second Job (11), Mini-job (up to 400 euros) (15)
6	Unemployed	Unemployed (5)
7	House-Husband/Wife	Housewife, Husband (10)
8	Retired	Retired (6)
9	Other	Maternity Leave (7), Other (12)
99	Gap	Information on gaps in ARTKALEN is not used. Gaps are calculated on the basis of the merged dataset.

Source: SOEP v32, doi: 10.5684/soep.v32

As stated above, the calendar information is used to update the biography information. However, there is also a certain overlap of the periods covered by the two types of data. This is shown in Table 4. It indicates, for persons included in PBIOSPE, the year in which the biography information was collected (variable ERHEBJ). This year is usually also the last year for which biography information is available.[1] The table also shows the first year recorded in the calendar data (variable KALYEAR).

[1] Please note that some biographies were collected in 2011 although they are part of Wave 27. This results from the fact that some members of Sample I were interviewed in early 2011 instead of 2010.

Table 4: Overlap between biography and calendar information

	First observation in ARTKALEN (compared to erhebj*)					
		earlier				
erhebj*	same year or later %	1 year %	2 years %	3 years %	4+ years %	Total n
1984	0.1	100.0	0.0	0.0	0.0	11,001
1987	0.0	36.4	33.5	30.1	0.0	505
1988	0.0	100.0	0.0	0.0	0.0	164
1989	0.5	99.5	0.0	0.0	0.0	193
1990	0.0	100.0	0.0	0.0	0.0	180
1991	0.0	100.0	0.0	0.0	0.0	157
1992	0.0	8.4	3.6	88.0	0.0	3,930
1993	0.0	76.6	0.3	2.3	20.7	304
1994	0.2	98.3	0.3	0.2	1.0	918
1995	0.2	99.1	0.0	0.1	0.6	1,037
1996	0.2	97.9	0.0	0.0	1.9	480
1997	0.0	98.5	0.0	0.0	1.5	478
1998	0.7	98.1	0.0	0.2	1.0	415
1999	0.1	26.6	72.8	0.0	0.5	1,821
2000	0.0	90.2	0.9	7.7	1.3	235
2001	0.0	6.3	93.6	0.0	0.0	7,529
2002	0.2	48.1	0.4	39.0	12.4	526
2003	0.1	16.9	81.3	0.1	1.6	2,193
2004	0.0	68.8	4.2	20.1	6.9	432
2005	0.0	89.0	3.4	0.7	6.9	292
2006	0.0	92.2	4.2	0.0	3.7	217
2007	0.0	16.2	83.4	0.1	0.3	1,858
2008	0.0	68.9	2.9	26.9	1.3	309
2009	0.0	89.5	2.1	0.5	7.9	190
2010	4.1	79.2	16.7	0.0	0.0	7,887
2011	2.5	91.7	5.7	0.0	0.1	5,670
2012	3.9	95.6	0.3	0.2	0.1	2,205
2013	92.6	7.1	0.2	0.1	0.0	3,860
2014	38.1	61.0	0.7	0.0	0.2	423
2015	0.0	79.3	12.0	7.5	1.1	266
Total	7.7	59.8	24.7	7.3	0.6	55,675

Notes: *) Year of biography data collection (variable ERHEBJ). Source: SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

In the majority of cases (59.8 percent), the earliest calendar information is available for the year before the biography interview. This is the case for persons who answered the Biography Questionnaire in their first year as survey respondents. The calendar in the Individual

Questionnaire refers to the year before the survey. There are, however, changes over time. In 1998, it was decided that first-time respondents from new samples would not be given the Biography Questionnaire in the first wave but in the second in order to reduce the entry threshold for these new respondents. Consequently, for the majority of persons in years after new samples were integrated (1999, 2001, 2003, 2007, 2010/11 – Samples E to I), the earliest calendar information is available two years before the biography information was collected. This was once again changed in 2011 and in 2013: respondents from samples J and K were given both questionnaires in their first year in SOEP and 2013 the sample M was introduced via a special Biography Questionnaire about individual migration histories without surveying the calendar data. The same applies to first-time respondents who are members of an old sample (e.g., persons who moved into a panel household) - they answer the Biography Questionnaire at the time of their first interview. The pattern is quite stable for most years before 1999. Notable exceptions is the year 1992. This is explained by the integration of East Germany into the SOEP in 1990 (Sample C). The majority of the respondents in these samples answered just the Biography Questionnaire at the entrance into the SOEP. Another exception is the year 1987. In the years 1985 to 1987, the life course matrix was not part of any of the questionnaires. Therefore the respective biography information was only available for persons who were interviewed in 1984. In 1988, biographic information was also collected for persons who became respondents in 1985, 1986, and 1987 (for all years ERHEBJ=1987). In addition There are even some cases (0.6 percent = 307 cases) where the biography information was collected a long time after the person started to respond to the Individual Questionnaire (up to 31 years). These are respondents who failed to answer to the Biography Questionnaire at a given time and therefore the biography information was collected later. In these—albeit very rare—cases, there is substantial overlap between the periods covered by the calendar and biography information.

Table 5: Sources of PBIOSPE spells

	n	%	% cum.
biography only	211,450	51.9	51.9
calendar only	130,600	32.0	84.0
1 biography, 1 calendar spell	63,394	15.6	99.6
2+ biography, 1 calendar spell(s)	582	0.1	99.7
1 biography, 2+ calendar spell(s)	1,062	0.3	100.0
2+ biography, 2+ calendar spell(s)	35	0.0	100.0
Total	407,123	100.0	

Source: SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

After merging the information from the Biography Questionnaire and ARTKALEN, the data is transformed into spells, whereby each spell is defined by the duration of a given status. A question that arises when merging the data is how to handle overlapping pieces of information. The basic principle is to assign a value of a given status in a given year if the status is recorded in the calendar or in the biography information or both. An example might help to illustrate this: the calendar records full-time employment for the years 2005 and 2007 while the biography records full-time employment for the period from 2000 up to 2006. The merged data from PBIOSPE contains a spell that begins in 2000 and ends in 2007. However, the initial information is restored by including additional variables, which allows for alternative ways of merging the data (see below). The variables SPELLINF, ERHEBJ, and KALYEAR contain general information on the sources of the information captured in a given spell. Table 5 shows that the majority of spells are based on biography information only (51.9 percent). Slightly less than one-third of all spells (31.7 percent) are not observed in the Biography Questionnaire but only in the calendar data. The remainder of spells contain information from biography as well as calendar data. Usually these spells combine one period observed in the Biography Questionnaire with a period observed in the calendar. Only 0.4 percent of the spells combine more than one period in any of the two sources (SPELLINF=4, 5 or 6).

The variables BEGINB1-ENDYK4 document the initial information from the two different sources and are probably not of interest to the majority of users. However, on the basis of these variables, users are able to fully separate the Biography data from the aggregated ARTKALEN data. This is advisable if you want to use the more detailed ARTKALEN information and combine it with the yearly information from PBIOSPE for earlier years only. The variable names indicate the “source” of the original information utilized (B: Biography - Questionnaire or K: calendar information from the yearly survey). As an example, we discuss one of the spells that combines information on more than one period from any of the two sources. The spell number 4 of person 9205 starts in 1983 and ends in 1994 (SPELLTYP=4: full-time employment). As the variable SPELLINF (=5) shows, this a spell that combines one period from the biography data with two periods from the calendar data. According to the biography data, the person worked full-time from 1983 (BEGINYB1) until 1992 (ENDYB1). There is overlapping information from the calendar data available from 1986 onwards (KALYEAR). According to these data, the person worked full-time from 1986 (BEGINYK1) to 1990 (ENDYK1) and from 1993 (BEGINYK2) to 1994 (ENDYK2). During the years 1991 and 1992, no full-time employment is recorded in the calendar data, which contradicts the information from the biography data.

Table 6: Example of combined spell

persnr	spellnr	spelltyp	beginy	endy	spellinf	erhebj	kalyear	beginyb1	endyb1	beginyk1	endyk1	beginyk2	endyk2
9205	4	4	1983	1994	5	1998	1986	1983	1992	1986	1990	1993	1994

Source: SOEP v31 (PBIOSPE), doi: 10.5684/soep.v32

In PBIOSPE, no attempt is made to "resolve" such contradictions, as this would require rather strong assumptions. More important, such assumptions would differ according to the research question, which makes it even more difficult to provide a standard solution. Therefore, in such cases, we generate spells in the same manner as in less difficult cases, namely by combining the information from the calendar and the biography data. In the given example, this results in a full-time employment spell that starts in 1983 and ends in 1994. As mentioned above, there are very few spells that combine information on two or more periods (SPELLINF=4, 5, 6, less than 0.4 percent of all spells). There are even fewer such spells where the period of overlap is as long as in this example, where the biography data was collected many years after the persons joined the survey (ERHEBJ=1998, KALYEAR=1986). However, users who are interested in combining biography and calendar data in a different manner can use the variables BEGINB1-ENDYK4 to fully separate the two types of data and to recombine the data on the basis of different rules of aggregation.

Changes in the previous version of PBIOSPE (release 2009):

The description in this chapter refers to the version of PBIOSPE released in 2014 (waves 1-30). There have been no changes how the data is generated since the previous version, distributed since 2009. But users who are only familiar with older versions of PBIOSPE (releases 2008 and earlier) will observe some differences. In 2009, the data generation has been updated completely, but without changing the basic principles. Therefore, there are only a few barely discernible deviations in the main variables (due to slight changes in the consistency checks of the data). But there are a number of visible changes in the form of additional variables or additional values in already existing variables:

- documentation of censoring:
    - gaps in the data are recorded as spells (SPELLTYP=99)
    - the variable ZENSOR is more detailed and informs about the type of censoring (end of observation window, gap due to missing data)
- documentation of set-up of single spells:
    - new variable KALYEAR: contains the first year for which calendar information is available

- o new variables BEGINB1-3, ENDB1-3, BEGINYB1-3, ENDYB1-3, BEGINK1-4, ENDK1-4, BEGINYK1-4, ENDYK1-4 (these variables replace BEGINBIO, ENDBIO, BEGINYB, ENDYB, BEGINKAL, ENDKAL, BEGINYK, ENDYK): Like the replaced variables, these variables document the original calendar and biography data. The new variables have been added to have a full documentation also for spells in which three or more initial spells are merged (spells with SPELLINF>=4). For the large majority of spells (SPELLINF<=3) only the first of each set of variables is filled. The new variables can be used to separate biography and calendar data, e.g., if you want to combine on your own biography data with data from ARTKALEN.
- o additional value in variable SPELLINF: the value 6 indicates that a spell has been constructed out of 2 or more biography and 2 or more calendar spells

- additional changes:
  - o variable ERHEBJ: value -2 if no biography information for a person is available (old version: value 0)
  - o The variable FEHLCODE is no longer provided, as its values appeared to be more confusing than helpful. It contained information on data problems in the biographies collected in 1984 only. Information on gaps and overlaps is now documented for all years but not in a single variable.
  - o variable SPELLTYP: value 3 indicates part-time and marginal employment until 2004. In 2005 a separate category for marginal employment (also "mini-job" or "400-euro job") is introduced (value 15) and value 3 is restricted only to part-time employment.

5 BIOJOB: Detailed Information on First and Last Job

by Paul Schmelzer and Tobias Wolfram[2]

5.1 Overview

Biographical data in the GSOEP stem from various sources. All information for the waves 1984 to 1995 is compiled in the BIOLELA-file of the SIR-GSOEP-database. Since 1996 a standardised version for all samples has been provided, and new biographical data is stored in

[2] Based on work of Tanja Schmidt, Anita Kottwitz, Daniel Wachtlin, Mathis Schroeder, Thorsten Schneider; and Hansjoerg Haas (Update waves X/Y/Z/BA/BB/BC/BD (2007-2013)).

wave-specific files ($LELA). To have a general phrasing, all biographical files are referred to as LELA-files. (LELA stems from the German 'LEbensLAuf', curriculum vitae.)

The LELA-data relevant for BIOJOB consists of

- the age at and year of entry into the working force given by different theoretical considerations
- the type of occupation at entry (blue/white collar worker, self-employed, civil servant)
- detailed occupational information at entry
- changes of occupation
- intended educational degree or vocational/professional training
- the year of the last employment
- the type of occupation in the last job.

Since 2000 a new questionnaire (in the following referred to as Youth Questionnaire) has been provided for respondents who are 16 or 17 years old. The youth respondents answer the Youth Questionnaire instead of the biographical one. The Youth Questionnaire provides less detailed information about the job biography because respondents usually have not entered the labour market at the age of 16 or 17.

In 2001 members of the F sample became part of the biojob population. They had to answer the biography questionnaire if their year of birth was prior to 1982. Members of the F sample with a birth year in the range from 1982 to 1984 answered the Youth Questionnaire.

Members of sample G (2002) answered the biography questionnaire in 2003, Persons who where born between 1986 and 1987 answered the Youth Questionnaire.

Members of sample H (2006) answered the biography questionnaire first time in 2007 and therefore are part of the BIOJOB population.

Sample I has been moved to the SOEP-Innovation study and, since 2011 (wave 28/BB), are no longer part of the core SOEP population. Members of sample I are still part of of the BIOJOB population until 2010.

Since 2006 respondents who are 16 or 17 years old filled in a youth questionnaire instead of the standard Individual Questionnaire, which provides less detailed information about the current job.

Recently two new projects were integrated into BIOJOB: The IAB-SOEP migration sample M (2013) and the "Familien in Deutschland" (L1-L3) data which was incorporated into the GSOEP in 2015 and is for the first time available with the distribution of wave BE.

The purpose of BIOJOB is to provide a file, that offers the user convenient access to biographical information on past job activities. Up to now all but two variables of BIOJOB

are time-invariant. Information on occupational changes and on the age at the most recent change of occupation refer to the date of the respondent's biography interview.

5.2 Structure and Contents of BIOJOB

BIOJOB consists of generated variables as well as plain questionnaire information. In this section the generated variables are explained and their coding is illustrated.

Concerning different sources of information, the following priority scheme is applied: First the plain information stemming directly from questions on the relevant topic in the latest valid LELA-file is used. In case of inconsistencies, which will be explained later on, the latest valid information stemming from the PBIOSPE file is also used. The PBIOSPE file consists of spell data concerning the retrospective question 'what did you do since the age of 15' in the Biography Questionnaire as well as the question on activities in the last year in the Individual Questionnaire (for detailed information see chapter 3).

Contents of BIOJOB

Population: All persons with an entry in any LELA-/YOUTH-file up to 2015, even if information on employment is missing.

number of cases: 80,379 *waves*: A(1984) - BF(2015) *samples*: A, B, C, D, E, F, G, H, I, J,K,L,M

variables:

HHNR	original household identifier
PERSNR	unique individual identifier
BIOYEAR	year of biography / youth interview
AGEFJOB	age at first job
AGEINFO	information source AGEFJOB
EINSTIEG_ARTK	Year of first job (different generation process involving ARTKALEN)
EINSTIEG_ARTK_INFO	information source EINSTIEG_ARTK
EINSTIEG_PBIO	Year of first job (different generation process involving PBIOSPE)
EINSTIEG_PBIO_INFO	information source EINSTIEG_ARTK
NOJOB	never worked before the time of the interview
STILLFJ	still employed in first job
OCCFJOB	occupational position first job

FULLTIME	first job was a full-time or part-time job
FJBLUE	first job blue collar worker
FJSELFE	first job self-employed
FJSEFSIZ	number of employees FJSELFE
FJWHITE	first job white collar worker
FJCIVS	first job civil servant
ISCO88	International Standard Classification of Occupation 1988, first job
STBA	classification of career according to the Federal Statistical Office, Germany, (Statistisches Bundesamt), version 1992, first job
EGP	Erikson and Goldthorpe's Class Category (EGP), first job
ISEI	International Socio-Economic Index of Occupational Status after Ganzeboom (ISEI), first job
MPS	Magnitude Prestige Scale after Wegener, first job
SIOPS	Treiman Standard Int. Occ. Prestige Scale, first job
REQEDUC	required education for first job
CIVILSFJ	first job was in civil service
NACEFJ	NACE branch code first job
OCCMOVE	number of occupational changes
AGEATMV	age at most recent occupational change
INTEDUC1 to INTEDUC4	intended educational degree
CURREMPL	employed at time of biography interview
YEARLAST	year of last employment
SCOPELJ	last job was a full-time or part-time job
CIVILSLJ	last job was in civil service
NACELJ	NACE branch code last job
OCCLJOB	occupational position last job
LJBLUE	last job blue collar worker
LJSELFE	last job self-employed
LJSEFSIZ	number of employees LJSELFE
LJWHITE	last job white collar worker
LJCIVS	last job civil servant

If data are missing, we use the SOEP missing value definition:

-1	no answer / don't know: item nonresponse

-2	does not apply
-3	after intensive checks a given value was found to be implausible and was finally deleted (to be interpreted like -1)

Description of variables

AGEFJOB/AGEINFO

The variable AGEFJOB provides the age at entry into the working force. AGEINFO is a pointer variable indicating the source of the age information.

In the Biography Questionnaire people either have to give information on their age at entry into the working force or have to state that they have never worked before the time of the interview. The latter information is used in the variable NOJOB.

In the Youth Questionnaire people have to answer whether they are currently working in a regular occupation. They are not asked about the age at their first occupation, but since people answering the Youth Questionnaire are normally at the age of 16 or 17, in most cases we can assume that a full-time job at this age is their first regular employment.

Information on the coding procedure of AGEFJOB is provided in the following subsections where (a) to (i) refer to LELA respondents, (j) to (p) to youth respondents respectively.

LELA-respondents

a) For people who are or have ever been employed at the time of answering the biographical questions their age at the time of entry into the working force is taken from the LELA-files.

b) When we observe, that the person has not been in the working force at the time of responding, but starts to work later on, data of the PBIOSPE-file is used. Using the spell information in PBIOSPE, we are able to collect the age at the first job.

c) A replacement of the LELA-data takes place, when respondents state that they have worked before the age of fifteen, but have a spell entry later than the age of fifteen. This rule is not applied when the spell starts at the age of fifteen, since this is the minimum value for spell data in the questionnaires.

d) The same procedure is applied, when people answer, that they have never worked at the time of the interview, but have a spell which starts before the first interview.

e) In some cases the AGEFJOB value is higher than the start of the corresponding working spell in PBIOSPE. In general, the AGEFJOB value is maintained. Only when the value is

greater than 27, is it replaced by the PBIOSPE data. (95% of these cases have an AGEFJOB below 27.)

f) If we observe item non response concerning AGEFJOB and NOJOB, but spell information is available, the missing value is replaced by the corresponding PBIOSPE spell data.

g) If even the 'What did you do since you were 15' question had not been answered, there still was a chance to extract similar information out of the PBIOSPE-file by considering the question 'What did you do every month last year'.

h) If we still had no valid information, the value of AGEFJOB was left out of the dataset.

i) Due to the fact that PBIOSPE information are collected only until the end of the year preceding the actual wave, for respondents without first job information from both the biography questionnaire and PBIOSPE we further look for a first job using information from the current wave individual questionnaire.

YOUTH-respondents

j) For respondents who are regularly employed, information is taken from the Youth Questionnaire; AGEFJOB is coded as year of questioning minus year of birth minus one (only if the respondent does not state that he/she is still in school, etc.).

k) If we additionally observe a spell starting before the respondent answers the Youth Questionnaire, information from PBIOSPE is used if the respondent does not state in the current questionnaire that he/she is still in school, etc.

l) If respondents answer that they have no regular employment but provide an employment spell starting after the time of the first interview, information from $P (for details see m) is taken if available (only if the respondent does not state that he/she is still in school, etc.).

m) For respondents with inconsistent first job information (simultaneous employment and school attendance/apprenticeship, differing job info in Youth Questionnaire and PBIOSPE) the question 'Are you currently engaged in paid employment?' asked in the Individual Questionnaire turned out to be the most reliable source of information. If a respondent states to be full- or part-time employed in a wave subsequent to the youth interview, AGEFJOB info is derived from the latest information of that kind.

n) If people do not answer at least one of the questions 'Do you currently earn money?" and 'Do you earn money as an apprentice, full-time worker or part-time-worker?" but have an employment spell, like in m) the earliest $P information is taken if available (only if the respondent does not state that he/she is still in school, etc.).

o) If information from the Youth and the Individual Questionnaire (including PBIOSPE) are inconsistent concerning AGEFJOB, then the variable is set to missing.

p) Due to the fact that PBIOSPE information are collected only until the end of the year preceding the actual wave, for respondents without first job information from both the Youth Questionnaire and PBIOSPE we further look for a first job using information from the current Individual Questionnaire.

The pointer variable AGEINFO provides the coding information described above. Value labels of AGEINFO indicating the source of information are:

(1) LELA-files (case (a) above)

(2) PBIOSPE if AGEFJOB<15, but spell begin > 15 (c)

(3) PBIOSPE if 'not worked' at interview but later spell begin (b)

(4) PBIOSPE if 'not worked' at interview but earlier spell begin (d)

(5) PBIOSPE if AGEFJOB>27 and earlier spell begin (e)

(6) implausible information therefore set missing (h)

(7) PBIOSPE if 'not worked'-question and AGEFJOB not answered, but 'what done at 15'-question answered (f)

(8) PBIOSPE if 'not worked'-question, AGEFJOB and 'what done at 15'-question not answered, but 'what done last year'-question answered (g)

(9) completely missing

(10) SP if no info from bio interview and PBIOSPE but employment in current Individual Questionnaire (i)

(11) info drawn from Youth Questionnaire(j)

(12) info drawn from PBIOSPE for persons who state in the Youth Questionnaire to be regularly employed and additionally have an employment spell starting earlier (k)

(13) info drawn from $P for persons who state in the Youth Questionnaire not to earn money relating to an employment/job or to earn money but relating to a part-time job or a practical training, and have a subsequent employment spell (l)

(14) info drawn from $P for persons with inconsistent first job information from the Youth Questionnaire or PBIOSPE, but valid employment information from an Individual Questionnaire subsequent to the biography interview (m)

(15) info drawn from $P for persons with item non response in one of the questions 'Do you already earn money from jobs?' or 'Do you earn that money as a trainee, full-time or part-time employee?' and with info in PBIOSPE (n)

(16) completely missing
(17) set to missing because of inconsistent information (o)
(18) info drawn out of UP, the last wave of the SOEP (p)

For more than 50% of the cases with AGEINFO = 3, 7, or 8 (AGEINFO=7 or 8 only if information collected after biography interview) it is possible to extract information from the regular questionnaires.

For respondents with AGEINFO=10 or 11, information referring to the variables OCCFJOB, FJBLUE, FJWHITE, FJSELFE, FJSEFSIZ, FJCIVS, REQEDUC and CIVILSFJ are taken from the Individual Questionnaire (same year as of youth interview). While for respondents having AGEINFO=10 this approach is intuitive, for the persons having AGEINFO=11 we act on the assumption that the job declared in the respective Individual Questionnaire is still the first job of that person. This assumption seems plausible due to the low age of all persons responding to the YOUTH Questionnaire.

In the YOUTH Questionnaire there is no question on the first job. But we can follow up their professional career by the statements given in the activity calendar in the subsequent waves. This can lead to problems if these youths report student jobs. For that reason we decided to take information from the question "Are you currently engaged in paid employment?" asked in the Individual Questionnaires of subsequent waves as the relevant source of information for this group of respondents. The earliest information of that kind determines the variable AGEFJOB.

Some respondents have very low values with respect to AGEFJOB. Most of these jobs turn out to be low-skilled and starting before 1970. The respective persons are either blue collar workers (mostly unskilled) or self-employed (mostly helping in family business). We think these characteristics suggest that these specifications are valid.

EINSTIEG_ARTK/EINSTIEG_ARTK_INFO

The variable EINSTIEG_ARTK (by Marco Giesselmann and Mila Staneva) provides the year of available survey information related to the entry into the working force. It is primarily based on information found in the spell dataset ARTKALEN and generally founded on a different conceptualization of job entry than AGEFJOB. EINSTIEG_ARTK_INFO is a pointer variable indicating the source of the information. There are three main reasons why two seemingly redundant variables like AGEFJOB and EINSTIEG_ARTK are both included in BIOJOB.

a) EINSTIEG is based on a more clear and consistent definition of what a first labor market entry is. Here the first labor market entry is conceptualized as the entry in the first job after the completion of (secondary and tertiary) education and apprenticeship.

AGEFJOB, though, captures labor market entries at very different stages of the educational and employment biography. One reason for this is that it largely relies on a self-assessment of what a labor market entry is. In the Biography Interview all respondents are asked when they first started to work and this leads to very diverse self-reported labor market entries ranging from the first side-job in high school to the first full-time stable employment matching the own professional field.

b) As described above for people who have never been employed at the time point of the Biography Interview the very first observed labor market entry from the spell-data PBIOSPE is used for the generation of the "age at first job"-variable. This also leads to inconsistent labor market entries since this generating strategy often captures student side-jobs.

c) Additionally, EINSTIEG refers to the earliest yearly measurement after the transition. Although this year is, due to panel structure and interview date, not necessarily to the year of entry, only this strategy allows a clear assignment of covariates from the yearly measurement to the labor market entry.

Among the several plausible concepts and operationalization of labor market entries with SOEP Data (among them AGEFOB), we consequently hold this indicator particularly suited for scholars who want to study the impact of labor market institutions on early career outcomes. By not being focused on first full-time or standard employment, but also regarding shifts into atypical employment as labour market entry, employment biographies spanned from this entry point capture the uncertainties and instabilities associated with the early career phase. At the same time, side-jobs or apprenticeships are explicitly assigned to the educational phase and excluded from the concept of labor market entry. A detailed description of the generation process of EINSTIEG_ARTK is given in the respective documentation file *Introduction to the Variable EINSTIEG.*

EINSTIEG_PBIO/EINSTIEG_PBIO_INFO

The variable EINSTIEG_PBIO provides the year of the entry into the working force. It is primarily based on information found in the spell dataset PBIOSPE and founded on the same concept as EINSTIEG_ARTK. EINSTIEG_PBIO_INFO is a pointer variable indicating the source of the information. The motivation behind another operationalization of job entry is straight-forward: Using the algorithm of EINSTIEG_ARTK only the job market entry-years of less than 7500 respondents can be reconstructed. This low number is explained by the fact that to be identified by our algorithm the beginning of the job (and the end of the educational) biography of a SOEP-participant has to be part of the ARTKALEN-dataset. This is only the case if one became part of the survey in late youth or early adolescence and did not leave the

sample before a first employment could be observed, so only for a fraction of first jobs as defined by EINSTIEG_ARTK dates can be estimated.

To offer a compromise between the problems of AGEFJOB and the few observations reconstructed by EINSTIEG_ARTK a third variable was created: EINSTIEG_PBIO, which instead of using information from ARTKALEN employs the dataset PBIOSPE, which includes spell-data gathered from the retrospective activity calendar which is part of the biography questionnaire. For understandable practical reasons though this data is just available on a yearly basis and not a monthly one like it is the case with ARTKALEN. The implied loss of granularity induces a potential higher risk of misclassifications compared to EINSTIEG_ARTK while enabling us to reconstruct job entries for a vastly higher amount of respondents, namely almost everyone who ever filled out the biography questionnaire. Still the potential use of EINSTIEG_PBIO compared to EINSTIEG_ARTK is much more restricted as again for most identified first jobs which fall in the time frame before the person became part of the sample and whose information deviates from agefjob there is just no further information available at all. A more detailed description of the generation process of EINSTIEG_PBIO is given in the respective documentation file Introduction to the Variable EINSTIEG.

NOJOB

The underlying question for the variable NOJOB is ‘I have never been employed up to this date’. This variable has the label ‘never been employed until the date of the interview” (1).

If NOJOB has a missing value, in general there should exist AGEFJOB information, for special cases, see above. Due to the lack of a comparable question in the Youth Questionnaire, respondents of this questionnaire are given the value (1) as long as no consistent AGEFJOB information is available.

STILLFJ

This variable is based on the question ‘Are you still employed in the same job and at the same place?’. It applies only to LELA respondents who do not state ‘I have never been gainfully employed’ and whose biography interview was after 2000.

Value labels:

(1) Yes
(2) No

FULLTIME

The FULLTIME-variable is used to indicate, whether the first job of a person was a full-time or a part-time job. The value labels are

(0) part-time job or marginal employment

(1) full-time job.

This variable is generated out of the file PBIOSPE for all respondents. For persons with first job information stemming from the Biography Questionnaires, FULLTIME possibly does not refer to the declared first job if PBIOSPE does not contain the respective job spell (i.e. due to item non response or incomplete answering of the activity biography within the Biography Questionnaire).

OCCFJOB

The variable OCCFJOB provides information on the occupational position at the first job. Due to different versions of the questionnaires in the GSOEP´s different samples we face some difficulties. Table 1 gives an overview.

Table 1: Number of Possible Values for Occupational Classifications in the First Job

	Farmers (not self-employed)	Blue Collar Workers	Self-employed	White Collar Workers	Civil Servants
Sample A, B (84-95)	-	5	5	5	4
Sample C (90-95)	4	5	5	4	4
Sample D (94/95)	4	5	5	4	4
Sample A,B,C,D (96)	-	3	4	3	4
Sample A,B,C,D (97-99), E (99)	-	3	4	4	4
Sample A,B,C,D,E (00)	-	3	6	4	4
Sample A,B,C,D,E,F (01)	-	3	10	4	4
Sample A,B,C,D, E,F (02)	-	5	10	6	4
Sample A,B,C,D, E,F,G(06),H(06), I(10),J(11),K(12)	-	5	10	6	4

Source: SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

Facing these differences we decided to standardise the occupational classification. Only four types of occupational status were taken into account: blue collar workers, white collar workers, civil servants, and self-employed. The group 'Farmers' is included in the blue collar worker group.

The potential value labels for OCCFJOB are:

(1) blue collar worker
(2) self-employed
(3) white collar worker
(4) civil servant

Further details are provided by the variables FJBLUE (for blue collar workers), FJSELFE (self-employed), FJWHITE (white collar workers), and FJCIVS (civil servants). Table 2 shows the number of possible values.

Table 2: Number of Possible Values for the subcategories of the variable OCCFJOB

	FJBLUE	FJSELFE	FJWHITE	FJCIVS
Sample A,B,C,D, E,F,G,H,I,J,K(12)	9	4	7	4

Source: SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

Due to the fact that the PBIOSPE-file is used for the coding of AGEFJOB in certain cases (see above) there is less information on OCCFJOB than on AGEFJOB.

FJBLUE

The FJBLUE variable provides detailed information on the first occupational status if the person was a blue collar worker. Certain value labels are only given for certain samples, because of the already mentioned differences in the questionnaires.

The following value labels are assigned:

(10) un- and semiskilled farmers (sample C/D)
(11) unskilled worker
(12) semiskilled worker
(20) skilled worker
(30) farmers (sample C/D) being foreman or master craftsman
(31) foreman (sample A/B)
(32) foreman (sample C/D)
(40) master craftsman
(41) farmers (sample C/D) in middle and higher management

FJSELFE/FJSEFSIZ

The FJSELFE variable provides detailed information on the first occupational status if the person was self-employed. FJSEFSIZ gives the number of employees in the respondent's firm. Again there are differences due to the different versions of questionnaires.

The following value labels are assigned:

(10) independent farmer
(20) free lances, self employed academics
(30) other self employed workers
(40) helping within family business

FJSEFSIZ has the following value labels:

(10) number of employees ≤ 9 (all subsamples (see exceptions for samples C/D), up until wave M)
(11) no co-workers (all subsamples, from wave R on)

(12) number of co-workers 1-9 (all subsamples, from wave N on)
(20) number of employees > 9 (all subsamples (see exceptions for samples C/D))
(30) number of employees ≤ 10 (sample C (waves I to L) / D (waves K to L), only if info drawn from biography questionnaire)
(40) number of employees > 10 (sample C (waves I to L) / D (waves K to L) , only if info drawn from biography questionnaire)

FJWHITE

FJWHITE gives detailed information on persons, who were first employed as white collar workers. The subvalues of unskilled labour without degree (21), or with degree (22) are, due to uncomparable values in the LELA-files, only drawn from the $P-Files. (Beginning with BIOJOB 2004).

Potential value labels:

(10) industrial foreman
(20) employee / unskilled labour
(21) same as (20), but without degree
(22) same as (20), but with degree
(30) employee / skilled labour
(40) employee / professional labour
(50) employee / managerial labour

FJCIVS

FJCIVS provides detailed information on first employment as a public servant.

The following value labels occur:

(10) low level civil servant
(20) middle level civil servant
(30) high level civil servant
(40) executive civil servant

ISCO88, STBA EGP, ISEI, MPS, SIOPS

These variables – job classifications and different prestige scores – concerning in each case the first job but are not generated within this file and therefore they are not described within this documentation.

REQEDUC

REQEDUC provides information about the required education for the first job. This information has been asked in the Biography Questionnaire for the first time in the year 2001, but comparable information are gathered by the Individual Questionnaire in all waves.

For all respondents having their first job subsequent to their biography interview, information is drawn out of the generated file $PGEN. Neither respective variables in $P nor those in $PGEN provide full information for all waves. In both data sources no differentiation is made between vocational college degree and university degree. As $PGEN info is equally coded in all waves, it is preferred to $P info.

Potential value labels:

(10) no training
(20) completed vocational training
(30) vocational college or university degree
(31) vocational college degree
(32) university degree

CIVILSFJ

CIVILSFJ indicates if the first job was assigned to the civil service or not. This information has been asked in the 2001 Biography Questionnaire for the first time

For respondents having their first job subsequent to their biography interview, information is drawn out of the generated file $PGEN where this information is provided since the first wave in 1984.

The following value labels occur:

(1) Yes
(2) No

NACEFJ

NACEFJ provides information about the industrial sector of the first job according to the branch classification NACE. This variable is not generated within this file. The description of its value labels is therefore not part of this documentation.

OCCMOVE

The variable OCCMOVE is based on the question 'Did you change your occupation and if you did, more than once?'. Information stems from the year of the biography interview. For respondents of the Youth Questionnaire as well as persons having their first job after the biography interview no information is available.

Labels of **OCCMOVE:**

(1) never changed occupation
(2) changed once
(3) changed more than once

AGEATMV

This variable is based on the question 'If you changed your occupation, how old were you at the most recent change?'. Information stems from the year of the biography interview. For respondents of the Youth Questionnaire as well as persons having their first job after the biography interview no information is available.

CURREMPL

This variable is based on the question ' Are you gainfully employed at the current time?'. The question applies only to LELA respondents who do not state 'I have never been gainfully employed' or 'Still employed in the first job'. This question has been asked in 1994 for the first time.

Value labels:

(1) Yes
(2) No

YEARLAST

This variable is based on the question 'When was the last time you were gainfully employed?'. The question applies only to LELA respondents who do not make at least one of the following statements in their biography interview:

'I have never been gainfully employed.'

'Still employed in the first job'

'Gainfully employed at the current time'.

This question has been asked in 1994 for the first time.

SCOPELJ

SCOPELJ indicates if the last job was a full time or part time job.

Information is only provided for respondents who answer the respective question within the Biography Questionnaires. The respective question applies only to respondents who do not make at least one of the following statements:

'I have never been gainfully employed.'

'Still employed in the first job'

'Gainfully employed at the current time'.

This question has been asked in 1994 for the first time. For youth respondents no information is available.

Value labels:

(1) full-time employed
(2) part-time employment
(3) marginal / irregular employment

CIVILSLJ

CIVILSLJ indicates if the last job was assigned to the civil service or not.

Information is only provided for respondents who answer the respective question within the Biography Questionnaires. The respective question applies only to respondents who do not make at least one of the following statements:

'I have never been gainfully employed.'

'Still employed in the first job'

'Gainfully employed at the current time'.

This question has been asked in 1994 for the first time. For youth respondents no information is available.

The following value labels occur:

(1) Yes
(2) No

NACELJ

NACELJ provides information about the industrial sector of the last job according to the branch classification NACE. The respective question applies only to respondents who do not make at least one of the following statements in their biography interview:

'I have never been gainfully employed.'

'Still employed in the first job'

'Gainfully employed at the current time'.

This question has been asked in 1994 for the first time.

This variable is not generated within this file. The description of its value labels is therefore not part of this documentation.

OCCLJOB

The variable OCCLJOB provides information on the occupational position at the last job. The respective question applies only to respondents who do not make at least one of the following statements in their biography interview:

'I have never been gainfully employed.'

'Still employed in the first job'

'Gainfully employed at the current time'.

This question has been asked in 1994 for the first time.

Due to different versions of the questionnaires in the GSOEP´s different samples we face some difficulties. Table 3 gives an overview:

Table 3: Number of Possible Values for Occupational Classifications in the Last Job

	Farmers	Blue Collar	Self-employed	White Collar	Civil Servants

	(not self-employed)	Workers		Workers	
Sample A,B (94/95)	-	5	5	5	4
Sample C,D (94/95)	4	5	5	4	4
Sample A,B,C,D (96-99), E (99)	-	5	5	6	4
Sample A,B,C,D,E (00)	-	5	6	6	4
Sample A,B,C,D, E,F (01/02)	-	5	10	6	4
Sample A,B,C,D, E,F,G(06),H(06),I(10), J(11)	-	5	10	6	4

Source: SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

Facing these differences we decided to standardise the occupational classification. Only four types of occupational status were taken into account: blue collar workers, white collar workers, civil servants, and self-employed. The group 'Farmers' is included in the blue collar worker group.

The potential value labels for OCCLJOB are:

(1) blue collar worker
(2) self-employed
(3) white collar worker
(4) civil servant

Further details are provided by the variables LJBLUE (for blue collar workers), LJSELFE (self-employed), LJWHITE (white collar workers), and LJCIVS (civil servants). Table 4 shows the number of possible values.

Table 4: Number of possible values for the subcategories of the variable OCCLJOB

	LJBLUE	LJSELFE	LJWHITE	LJCIVS
Sample A,B,C,D, E,F,G,H,I,J(84-11),K(12)	9	4	7	4

Source: SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

LJBLUE

The LJBLUE variable provides detailed information on the last occupational status if the person was a blue collar worker. Certain value labels are only given for certain samples, because of already mentioned differences in the questionnaires.

The following value labels are assigned:

(10) un- and semiskilled farmers (sample C/D)
(11) unskilled worker
(12) semiskilled worker
(20) skilled worker
(30) farmers (sample C/D) being foreman or master craftsman
(31) foreman (sample A/B)
(32) foreman (sample C/D)
(40) master craftsman
(41) farmers (sample C/D) in middle and higher management

LJSELFE/LJSEFSIZ

The LJSELFE variable provides detailed information on the last occupational status if the person was self-employed. LJSEFSIZ gives the number of employees in the respondent's firm. Again there are differences due to different versions of questionnaires.

The following value labels are assigned:

(10) independent farmer
(20) free lances, self employed academics
(30) other self employed workers
(40) helping within family business

LJSEFSIZ has the following value labels:

(10) number of employees ≤ 9 (all subsamples (see exceptions for samples C/D), until wave M)
(11) number of co-workers = 0 (all subsamples, from wave N on)
(12) number of co-workers 1-9 (all subsamples, from wave N on)
(20) number of employees > 9 (all subsamples (see exceptions for samples C/D))
(30) number of employees ≤ 10 (sample C (waves I to L) / D (waves K to L), only if info drawn from biography questionnaire)
(40) number of employees > 10 (sample C (waves I to L) / D (waves K to L) , only if info drawn from biography questionnaire)

LJWHITE

LJWHITE gives detailed information on persons, who were last employed as white collar workers. The values (21) and (22) are drawn from the BIOLELA-File and from the $P-files.

Potential value labels:

(10) industrial foreman
(20) employee / unskilled labour
(21) same as (20), but without degree
(22) same as (20), but with degree
(30) employee / skilled labour
(40) employee / professional labour
(50) employee / managerial labour

LJCIVS

LJCIVS provides detailed information on last employment as a public servant.

The following value labels occur:

(10) low level civil servant
(20) middle level civil servant
(30) high level civil servant
(40) executive civil servant

INTEDUC1 to INTEDUC4

The variables INTEDUC1, INTEDUC2, INTEDUC3, and INTEDUC4 provide information on the educational degree or the vocational/professional training a respondent intends to complete in the future, asked at the time of the biography interview. We create these four variables since multiple answers are explicitly allowed in the questionnaire. The intended education is stored with respect to the hierarchy given by the questionnaire, i.e., the highest degree is placed in INTEDUC1. For example, a person intending to finish an apprenticeship (1) and university (7) would have INTEDUC1 = 7 and INTEDUC2 = 1. Since this question has been asked for the first time in 1996, we do observe a large number of missing values for INTEDUC1 to INTEDUC4.

(1) apprenticeship
(2) full-time vocational school
(3) technical school
(4) education as a civil servant

(5) accredited professional school
(6) technical or professional college
(7) university

General remark:

Some persons answered more than once the Biography Questionnaire (but this occurs very rarely). The data-set BIOJOB contains only information from one Biography Questionnaire, in most cases the earlier one.

5.3 Steps of Coding

1. Creating a dataset using the data concerning all aspects of the job biography (working force entry, position, etc.) drawn from BIOLELA, MLELA, NLELA, OLELA, PLELA, QLELA, RLELA, SLELA, TLELA, ULELA, VLELA, WLELA, XLELA, YLELA, ZLELA, BALELA, BBLELA, BCLELA (internal DIW files with biographical information up to wave BB), QJUGEND, RJUGEND, SJUGEND, TJUGEND, UJUGEND, VJUGEND, WJUGEND, XJUGEND, YJUGEND, ZJUGEND, BAJUGEND, BBJUGEND, BCJUGEND (internal DIW youth biography files), QP, RP, SP, TP, UP, VP, WP, XP, YP, ZP, BAP, BBP, BCP (needed for consistency checks with respect to the youth biography files).
2. Using the PBIOSPE-data to retrieve spell information during the first occupation.
3. Using PPFAD for personal data (year of birth, sex, sample).
4. Using several files containing generated information about job classification (ISCO), prestige scores and industry sector classification (NACE) concerning the first job.
5. Combining all data concerning the employment biography into a new data file BIOJOB, where priority is set as mentioned above.
6. Coding of AGEFJOB. (for details, see above)
7. Setting the pointer variable AGEINFO indicating the source of the information of AGEFJOB. (for details, see above)
8. Excluding one value for respondents, who stated to have two occupational positions in their first job. Exclusion based on consistency checks.
9. Assignment of the variable OCCFJOB, with respect to the different versions of the questionnaire. Possible value labels: FJBLUE, FJSELFE, FJWHITE, FJCIVS.

10. Definition and assignment of new value-labels for the sub-category FJBLUE, nine labels possible, for details see above.
11. Definition and assignment of new value-labels for the sub-category FJSELFE, four labels possible, for details see above.
12. Definition of the variable FJSEFSIZ, indicating the numbers of employees.
13. Definition and assignment of new value-labels for the sub-category FWHITE, seven labels possible, for details see above.
14. Definition and assignment of new value-labels for the sub-category FJCIVS, four labels possible, for details see above.
15. Coding of the variables REQEDUC and CIVILSFJ.
16. Coding of the variables INTEDUC1 to INTEDUC4.
17. Computing the age at the most recent change of occupation if necessary.
18. Check of consistency: Does information about the age at the most recent change of occupation make sense? If inconsistencies appear, the value is set to a missing value.
19. Assignment of value labels for the variables specifying the last job:
20. Definition and assignment of value labels of the variable CURREMPL indicating if a respondent is gainfully employed at the time of the biography interview.
21. Specification of the year of last employment (YEARLAST).
22. Coding of the variables SCOPELJ and CIVILSLJ.
23. Excluding one value for respondents, who stated to have two occupational positions in their last job. Exclusion based on consistency checks.
24. Assignment of the variable OCCLJOB, with respect to the different versions of the questionnaire. Possible value labels: LJBLUE, LJSELFE, LJWHITE, LJCIVS.
25. Definition and assignment of new value-labels for the sub-category LJBLUE, nine labels possible, for details see above.
26. Definition and assignment of new value-labels for the sub-category LJSELFE, four labels possible, for details see above.
27. Definition of the variable LJSEFSIZ, indicating the numbers of employees.
28. Definition and assignment of new value-labels for the sub-category LJWHITE, seven labels possible, for details see above.
29. Definition and assignment of new value-labels for the sub-category LJCIVS, four labels possible, for details see above.

30. Collecting of job information for people with AGEINFO = 3, 7 or 8, if possible.
31. Collecting of job information for people with AGEINFO = 12, 14 or 16, if possible.
32. Coding of the variable FULLTIME.
33. Definition of missing values for all variables.
34. Hand-editing of inconsistencies between different variables.
35. Final listing
36. Definition and assignment of new value-labels for the sub-category LJCIVS, four labels possible, for details see above.
37. Collecting of job information for people with AGEINFO = 3, 7 or 8, if possible.
38. Collecting of job information for people with AGEINFO = 12, 14 or 16, if possible.
39. Coding of the variable FULLTIME.
40. Coding of the EINSTIEG variables
41. Definition of missing values for all variables.
42. Hand-editing of inconsistencies between different variables.
43. Final listing

6 The couple history files BIOCOUPLM and BIOCOUPLY, and marital history files BIOMARSM and BIOMARSY

by Maik Hamjediers and Paul Schmelzer[1]

With the BIOMARSM/Y and BIOCOUPLM/Y the SOEP provides consistent and continuous marital and partnership histories for nearly all adult respondents. Whereas BIOMARSM and BIOCOUPLM just build on the prospective information at the time of each interview, BIOMARSY and BIOCOUPLY – containing also retrospective data – provide complete marital histories of respondents, starting at the year of their birth. But since until wave 27 no questions on a respondents' couple history were asked, BIOCOUPLY includes only those respondents who have answered the biography questionnaire in wave 28 or later (and those who have been observed since the age of 17). Thus, all four datasets contain different information and therefore Table 1 gives an overview on the most important differences.

Table 1: Overview of the datasets

	Provided type of history	*Time unit*	*Starting time*	*Sample*
BIOMARSM	Marital Histories	Month	Entry into SOEP	All adult SOEP-participants
BIOMARSY	Marital Histories	Annual	Year of birth	All adult SOEP-participants
BIOCOUPLM	Relationship Histories	Month	Entry into SOEP	All adult SOEP-participants
BIOCOUPLY	Relationship Histories	Annual	Year of birth	Adult respondents answered the biography questionnaire after wave 27 or which were observed since the age of 17

Source: SOEP v32, doi: 10.5684/soep.v32

Note that the marital status in the $PGEN data files, stored as $FAMSTD, is derived from BIOCOUPLM and BIOMARSM at the time of the interview. Although the partner indicator PARTZ$ supplied in the $PGEN data files is considered in the generating process of BIOCOUPLM and BIOMARSM, due to different generating processes, it might not entirely

[1] This documentation is a new version of previous SOEP documentations for the same files and has benefited from the work made by previous generators. For readability reasons, we do not specifically cite and specify text that has been used directly from the older SOEP documents.

match with $FAMSTD. Furthermore, consistency checks between waves were done as well. That way, changes in $FAMSTD between data distributions but also in the couple and marital history datasets are possible for former waves.

This documentation proceeds with a short description of the sources for the generating process and of the editing process of constructing logical and consistent marital and couple histories. These steps are important to understand the description of the four data files following in the next parts.

6.1 Sources of the couple and marital history

For the construction of individual marital and relationship histories we gathered information

1. on current marital *and couple* status conducted in the personal questionnaire $P (and $PAUSL and $PLUECKE);
2. on monthly information on events, that may have occurred since the last personal interview, which are also stored in $P (and $PLUECKE);
3. on the generated partner pointer PARTZ$$ of the generated dataset $PGEN, which links current partners living in the same household;
4. on the *marital and relationship biography* from the biographical questionnaire $LELA (and BIOLELA).

The personal questionnaire comprises a question on the marital status at the month of interview (Question 147 in 2015), whereby information on registered partnerships is included only since 2011. Furthermore it is asked for the current couple status, which entails whether someone has a partner and if so whether they are living together (Question 148 and 149 in 2015). For immigrants we also used information on the marital status derived from the foreigner questionnaire stored until 1995 in $PAUSL and for temporary drop outs we replaced missing information with data from a short version of the personal questionnaire, which information is stored in $PLUECKE. Moreover at least once the majority of SOEP participants answers the biography instead of the personal questionnaire, and for those waves the current marital and couple status is stored in $LELA. This hereby out of the four sources gathered information on the current marital couple status is used to generate spells, beginning at the time of the interview and reaching to the next interview in which a change in the status is reported.

For those interviewed again it is also asked for any changes that occurred during the last year (Question 173 in 2015).[2] This monthly information on the events ‘moved in together’, ‘marriage’, ‘divorce’, ‘separation’, ‘death of partner’ and since 2011 ‘starting a new

[2] Due to the fact that events were collected retrospectively from first of January of the last calendar year until the month of interview, events in the beginning of a year could have been reported twice in two consecutive waves. Taking variations up to twelve months for the same event into account, the timing of the events were corrected.

relationship' that may have occurred since the last personal interview supplements the previous generated marital and couple status spells In addition, we added the generated partner pointer, which was just used as another source of. These three kind of information – events, the current status and the partner indicator – were used to generate the monthly datasets BIOMARSM and BIOCOUPLM.

For the annual files, we additionally resorted on the retrospective information of the biographical questionnaire, which nearly all respondents just answer one time while participating in the SOEP. Since until 2010 it was just asked for the last three marriages (Question 72 in 2010), for all respondents having used the old version it was only possible to generate marital and no couple biographies from their year of birth on. In 2011, design of the retrospective questionnaire collecting information on couples changed notably. It now contains information on up to three previous marriages, registered same-sex partnerships or long-term relationships, defined as lasting for at least six months (Question 83 in 2015). Hence, up to four relationships, are possible to mention.[3] Therefore for all respondents who have used the biography questionnaire after 2010 we generated the annual data BIOCOUPLY, whereas all responses to any biographical questionnaire were used to generate BIOMARSY. Naturally, the information drawn from the biographical questionnaires are proceeded by the responses to the personal questionnaire for the following waves. This is done by extending the monthly spelldata from BIOCOUPLM and BIOMARSY to an annual format; thus for this time, also shorter relationships were possible to mention, for example via the monthly information on the events mentioned above.

6.2 Construction of marital histories

Information on marital history mainly stems from respondents' retrospective reports on their own history. Thus, no other benchmark on the substance of these reports exists. This led to inconsistencies – for example, overlapping of reported marriages or impossible changes between legal states. In contrary to the relationship histories, where odd patterns of the reported histories occurred but were also possible for the most part, rules for logical histories of the legal marital status had to be laid down in the generation process:

1. Every individual marital history has to start with the state 'unmarried'. We did not allow a person to be married before age 16.
2. From 'unmarried', one can only change to 'married' or into 'living in a registered same-sex partnership'.

[3] Please note, that also between 2011 and 2014 the questions of the biography questionnaire were changed. From 2011 to 2013 it was also possible to mention a fourth previous relationship, thus the questions for that relationship was limited compared to the question about the other relationships. Furthermore, the biography questionnaire of the National Survey of Families (FiD-Sample) left out the option of reg. same-sex partnerships.

3. There is no possible return to ‘unmarried’ once a person was ever ‘married’.[4] The only possible change from ‘married’ is to ‘divorced’, ‘widowed’ or ‘divorced or widowed’.
4. The only possible change from ‘divorced’ or ‘widowed’ is to ‘married’.
5. Reported separations of married spouses are taken into account as well. This led to the new category ‘married, separated’ which is coherent with the BIOCOUPL datasets and contains the time from a separation until a divorce or death of the respondent’s spouse. Yet, this rule was not applied for move-outs and later returns into the household of the same partner. If another marriage with a new partner is reported without a divorce or a death of the previous partner, after the ‘married, separated’-spell an episode of the category ‘divorced or widowed’ is also added and the beginning and end of those spells as well as the censoring were adjusted (Figure 1).

Figure 1: Example for treatment of separated marriages in BIOMARSM/Y

Respondents information

persnr	spellnr	spelltyp	begin	end	divorce
1	1	married	10	50	0
1	2	married	100	150	0

Added spells

persnr	spellnr	spelltyp	begin	end	divorce
1	1	married	10	50	0
1	2	married, separated	50	-1	-1
1	3	divorced or widowed	-1	100	0
1	4	married	100	150	0

Algorithms rearrange and correct original data to produce logically consistent marital histories following these five rules. Thus, we inserted an obligatory first ‘unmarried’ spell starting with birth and ending with the first marriage. Possible contradictions in the data were checked and edited as well, i.e.: a) contradictions between responses given in the same year in the different, previously described sources; b) illogical sequences between years.

While most of the original data was used, we edited some spells, e.g. because of the notion of “romantic wedding”. This idea assumes that respondents often define their marital status in subjective or even affective terms rather than referring to the legal marital status. For instance, we frequently observe unmarried couples both reporting to be married for one year, and then returning to define themselves as singles (this phenomenon we call “romantic marriage”). Another rather emotional defined marital status are changes between the reported state of being divorced and being married, but not living together. This might stand for an insecure period after a break-up of a marriage and the changes between the states are corrected until

[4] Except for the annulations of a marriage. However, given that this event is extremely rare – in particular compared to the many returns to ‘unmarried’ that we find in the data – we did not consider the possibility of annulations.

any clear indication for an actual divorce is reported. A third frequent pattern refers to divorced respondents who report to be unmarried after they started a new relationship, since they may want to avoid the new partner to come to know about the former marriage. Since this kind of change of the legal marital status is impossible, the status of being divorced is assigned till the next reported marriage.

If responses on marital status alternate between 'divorced' and 'widowed' without reporting another marriage in between, the information reported most often was used to correct these contradictions. Sequences alternating short-termed between 'divorced' / 'unmarried' and 'married' are replaced by 'married', as long as a new partner, a longer consistent sequence or the report of an event does not confirm this possibly new marriage. If the very end of a reported sequence indicates an 'unmarried' spell after being married and without further information, the marital history ends with a 'divorced or widowed', indicating that the reason for the end of a marriage is not known. Likewise, reported weddings during still existing marriages are ignored mostly.

In cases where information of both partners on their joint marriage is available, contradictions in dates were not dissolved and the original information is not replaced. You can identify these spells by sorting the couples in the BIOCOUPLM dataset using the variable COUPID and decide which information you find more reliable.

Completeness is a further criterion for construction of marital histories in the sense that the spell system is a closed system of spells starting from birth or the entrance into the SOEP going to the last year of sample membership. Due to item as well as partial unit non-response (i.e., a person of a SOEP households refuses to give a personal interview) and due to inconsistent information, 'gap' spells are introduced as another category of SPELLTYP on its own. Gap spells can occur at any place in the spell system, i.e., there are no restriction rules like the ones above. If information on marital status is missing for more than two years we inserted gap spells indicating that we have no knowledge of what happened during these periods. Likewise, missing retrospective biography information is indicated by an inserted gap spell. Also, if three finished marriages were reported in the biography questionnaire and a fourth current marriage via the next marital status, another gap is inserted. Missing information due to item non-response in the life course questionnaire may also affect the dates of the beginning or end of a spell.

6.3 Construction of couple histories

As stated above, most of reported odd patterns in the couple histories may be possible and hence, no verified decision between measurement error and uncommon reality can be made. For that reason, whenever possible, couple histories are left as they were stated. Only in rare cases, restrictions, corrections on orderings of events or changes of declared years are

conducted (read the following paragraphs for details). Further corrections to smooth out irregularities are thus left to the user.

Firstly, the following default rules were obeyed to obtain logical and consistent histories if no other information forced to do otherwise:

1. Every individual couple history starts at the year of their birth with the state ‘single’. Because legal marriages are not possible before the age of 16, we restricted the age of marriage to be at least 16. Thus, we did not restrict age within a relationship, that is, unmarried relationships are allowed to start anytime.
2. Every spell set for a certain couple starts with the state ‘coupled, partner not in household’. One exception exists: if respondents report a year of moving together that lies before the start of their relationship, this specific couple history starts with ‘coupled, partner in household’. Note that in this case the information when this couple moved together is not available in BIOCOUPLM/Y anymore. Analogical the date of moving out is also lost if it was reported to come after the end of a relationship. Both dates you can still look it up in the original source.
3. If there is no evidence to the contrary, it is assumed that married couples live together and moved together before marriage. That is: for married couples their specific couple history starts with ‘coupled, partner not in household’ and is followed by a spell ‘coupled, partner in household’ before their marriage spell ‘married, spouse in household’ starts. Thus, if a couple moved together in the same year they married, a spell ‘coupled, partner in household’ is included anyway. The information whether spells needed to be inserted in order to this rule is stored in the variable REMARK (see Figure 2). Note that dates of becoming a couple and moving together or whether they moved together at all are not known in that case and were set equal to the time of marriage.

4. Since it was possible to mention another relationship in the questionnaire, it is assumed that periods between those relationships mentioned were 'single' states and

Figure 2: Example for additional spells before marriages

Respondents information

persnr	spellnr	spelltyp	begin	end	remark
1	1	single	10	50	1
1	2	married, living together	50	100	1

Added spells

persnr	spellnr	spelltyp	begin	end	remark
1	1	single	10	50	1
1	2	coupled, not living together	50	50	2
1	3	coupled, living together	50	50	2
1	4	married, living together	50	100	1

thus assigned accordingly. Note that this relates to long-term relationships (longer than six months) reported via the biography questionnaire only. In contrast to those, relationships derived from reported changes in the familiar situation in the personal questionnaire may also be shorter than six month.

5. Any (formerly) married couple that is not an active relationship anymore, i.e. married couple which is separated but not yet divorced, ends with a spell 'married-separated' (see Figure 3). As long as they are not divorced (or the partner hasn't died) yet, the end date of those separated spells is the same as their last interview year. Contrary to this, the separation spell is not added if marriage clearly ended with a divorce or the death of the respondent's partner and no previous break-up was reported. Be aware, that because the question for a separation does not differentiate between a move-out and a clear ending of a relationship, there is some uncertainty whether a marriage is still active, but not in a joint household or the relationship ended and the spouses are n

Figure 3: Example for treatment of separated marriages in BIOCOUPLM/Y

Respondents information

persnr	spellnr	spelltyp	begin	end	divorce
1	1	married, living together	10	50	0
1	2	coupled, not living together	100	150	0

Added spells

persnr	spellnr	spelltyp	begin	end	divorce
1	1	married, living together	10	50	0
1	2	married, separated	50	150	-1
1	3	single	50	100	0
1	4	coupled, not living together	100	150	0

yet. By using the information of following interviews we identified the spelltyp as clear as possible. Note that 'married, separated' is a redundant spell in terms of couple history: It always and fully overlaps with other spells that contain the actual couple status(es) over the entire separation episode.

6. Because BIOCOUPLM/Y documents couple statuses and not marital statuses, it is possible to become single after marriage (in BIOMARSM/Y the only possible change from 'married is to 'divorced' or 'widowed').

7. Information on the partner's death or a divorce is stored in the variables PDEATH and DIVORCE. It is always just assigned to the last spell of a relationship; if e.g. a marriage ends in divorce, just the last spell containing also the definite end of the relationship is marked via DIVORCE and all previous spells for the same marriage (e.g. 'married, not living together' before a move-in) are marked as 'spell does not end with a divorce' (see also Figure 3). If not applicable, that is if a person is currently single, PDEATH and DIVORCE are set to '(-2) does not apply'. If a respective couple is not married, DIVORCE is set to '(-2), does not apply', as well. (Be careful to change missings into system missing values if you plan to sum up these indicators column-wise.)

Checks were applied for three kind of inconsistencies: a) information from the different sources (e.g. biography and personal questionnaire) for the same years may conflict, b) dates within a personal couple history might appear plausible or not and c) contradictions between information given by a respondent and her relationship to the head of household.[5] In general it was tried to leave as much of the original information as it is given by the respondent as possible to avoid making assumption which users wouldn't follow and to allow for analyses of uncommon patterns as well.

Concerning the reported status(es) (a) and dates (b), it is possible that a person has multiple, overlapping relationships in a certain period. That is, relationships (in opposition to marriages) do not need to be consecutive. The variable SPELLNR provides the sorting order, whereby the most recent spell comes last, which represents the relationship that started most recently. If no clear dates were available, the sorting in SPELLNR reflects the order of reporting by the respondent. Thus, some overlaps and inconsistencies might still appear in the dataset, which you may want to remove.

Contradictions in both partners' answers (c) are left as they were, even if both were interviewed in the current wave. E.g. information on starting date or the date of marriage of their joint relationship often differs. For very few cases, there were even contradictions about the actual partnership between two individuals in the sample. E.g., if one part of a couple or the position to the head of household suggested a couple in a household, but one or both

[5] See $PSTELL in the dataset $PBRUTTP. This information was given by the head of household in the Household Questionnaire and processed in partner pointer PARTZ$$.

answered to be single or coupled with someone else (outside the household). Because it may depend on a research question which information is needed, we left irregularities within couples as reported by each spouse; neither date nor the relationship status is edited or overwritten between both partners of a current couple. To check for contradictions you can link them via couple identifier COUPID or PARTNR$$.

Some respondents refused to answer some or all questions in the relationship part. To some extent, missing values and contradiction are similar in the generating sense. Hence, some of the following rules applied to missing values, may seem familiar:

1. If a respondent stated to be married or in a relationship, but it is not known whether the couple lived together as well, these relationship spells were set as 'coupled, partner not living in the household' and the respective REMARK was coded as 'edited'.
2. Furthermore, if we had clear information about whether a respondent was married or not, but there was contradicting information about whether a respondent was in a relationship or not, we assigned these spells as 'unknown' (98). This applies also to the counterpart, that means the relationship status was definite, but we lacked information whether the couple was married or not.
3. In some cases it is possible that a respondent might have had earlier relationships, but they could not be named anymore due to the above mentioned restrictions of the biography questionnaire. Thus, it is not known whether there were other partners before the latest mentioned. In these cases, gaps were introduced, that is a spell 'unit nonresponse' (99) was inserted.
4. The relationship history of respondents, who haven't filled out the biography questionnaire, were also supplemented with additional spells: from age 0 they are stated 'single'; between this first single spell until the first mentioned relationship gaps are included as well.
5. For many non-interviewed persons' information on the current relationship can be reconstructed, because the respective partner is available from his or her personal interview. Thus, those non-interviewed adults remain in the BIOCOUPLM dataset, even though information is not directly retrieved from them. If information on their current couple status was given by their partner, it is fully copied to these non-respondents as well; the variable REMARK is set as 'added' for these spells.

6.4 BIOCOUPLM: A monthly couple biography

The spell data in the file BIOCOUPLM contains information on marital biographies starting with responses to the first personal interview. The data file comprises fourteen variables: the original household, couple and individual identifiers HHNR, COUPID and PERSNR as well

as eleven specific variables. Please note for the couple identifier COUPID that the BEGIN and END the identified partner spells are not necessarily equal: Differences may occur because we give a preference to respondents' own history. To achieve the same dates for BEGIN and END one might give a preference to one of the partners (i.e. by using only the information reported by the women in the relationships).

Variable SPELLNR is a chronological index number for each individual's spells during the observation period. Variables BEGIN and END indicate the month in which a marital spell starts and ends. Monthly histories start with a value of 1 in January 1983 and ranges until the current margin in month 384, i.e., December 2014. In principle, the month in which a spell starts is the very same month the previous spell ends in. Please take into account that the months of BEGIN and END are not imputed when the exact date of the change in the familiar situation is not reported. Instead, the time of the interview is left as the BEGIN or END date when a change is observed in the data. We introduced therefore a variable called EVENTS, containing the information whether the exact month of the BEGIN or END of a spell is reported via the monthly information on events or not.

Variable SPELLTYP documents the partnership status with the possible categories listed below. Note that it is only asked for same-sex partnership since wave 28, thus, this information cannot be reconstructed for previous relationships. The state 'single' is not equal to the legal state 'never married'; it refers only to a person's current state of having a partner or not. 'Married, separated' spells refer only to couples who are still married, i.e. who are separated but not yet divorced. The spell 'Married, separated' always overlaps (in terms of starting and ending time) with other spells that contain the actual couple status(es) over the entire separation episode and is therefore redundant in terms of completeness of couple histories (e.g. spell "Coupled, partner in household" might overlap with spell 'Married, separated'). The additionally assigned codes 'implausible', 'unknown' and 'unit nonresponse' indicate implausible or a lack of information for the respective period.

Variables of BIOCOUPLM

HHNR	Identifier of original sample household
PERSNR	Personal identifier
COUPID	Couple identifier
SPELLNR	Consecutive spell number (chronological order)
SPELLTYP	Partnership status (-3) implausible (1) Married, spouse not in household (2) Married, spouse not in household (3) Coupled, partner in household (4) Coupled, partner not in household (5) Single (6) Married, separated (7) Registered same-sex partnership, partner in household (8) Registered same-sex partnership, partner not in household (98) Unknown (99) Unit nonresponse
BEGIN	Month when spell begins [1=Jan 1983 to 396=Dec 2015]
END	Month when spell ends [1=Jan 1983 to 396=Dec 2015]
BEGINY	Year spell begins [1983 to 2015; -3=implausible; -1=missing]
ENDY	Year spell ends [1983 to 2015; -3=implausible; -1=missing]
PDEATH	Death indicator: spell ends with death of partner?
DIVORCE	Divorce indicator: spell ends with divorce?
CENSOR	Censoring information [0 to 14] (see explanation below)
EVENTS	Month information: exact month of spell begin or end known?
REMARK	Further spell information (1) Original spell (2) Edited spell (3) Added spell (4) Gap Spell (5) First Spell

In addition, the indicator variables PDEATH and DIVORCE are provided. PDEATH indicates whether a respective spells ends with the death of a person's partner. Please note that this is not restricted to married persons, thus does not only refer to widowhood. DIVORCE is similar and indicates whether the marriage spell ends in divorce. Hence if it did not end in divorce, it is coded as missing when a new marriage has been reported or as a still ongoing marriage. In the second case END is updated by the month of the last interview. Both indicator variable PDEATH and DIVORCE are assigned with '(-2) does not apply' for single and '(-1) missing' for gap spells.

It is important to note that a 'first spell' in BIOCOUPLM is not the very first spell of a person, but the first observed partnership status since the person is taking part in the SOEP. Accordingly, the first spell in BIOCOUPLM is almost always left-censored, that means the beginning of the spell is not known. Variable CENSOR informs whether a spell is left- or right-censored (in fact most spells in BIOCOUPLM are censored). Table 2 gives an overview for the coding of the reasons responsible for censoring. With regards to left censoring we

distinguished the reasons ‘missing’ (which also applies for most of the first spells), ‘after gap’; for right censoring the reasons are categorized in ‘missing’, ‘before gap’, ‘last spell’, and ‘death of the respondent’, whereas ‘death’ of course is always the ‘last spell’ and therefore the censor is overwritten with the former category. Certainly spells can also be coded as left- and right-censored combined.

Table 2: Coding of the variable CENSOR

Right: / Left:	not censored	censored missing	censored before gap	censored last spell	censored death
not censored	0	3	4	5	6
censored missing	1	7	8	9	10
censored after gap	2	11	12	13	14

Note: ‘(99) Gap’, ‘(98) unknown’ and ‘(-3) implausible’ spells are all uncensored

Source: SOEP v32, doi: 10.5684/soep.v32

Variable REMARK provides information on whether we had to edit or supplement original information provided by respondents in order to construct consistent couple biographies. Spells in BIOCOUPLM are marked as ‘edited’ if the editing process involved substitution of reported information because of inconsistence with responses in previous or following interviews or information reported by a partner. Furthermore we have added spells, e.g. between two marriages because two marriages with separate persons at the same time are not legal (see former section for more details on editing); these spells are marked as ‘added spells’. Finally ‘first spells’ and ‘gap spells’ are marked separately, whereas also ‘added’ or ‘edited’ first spells are marked as ‘first spells’.

6.5 BIOCOUPLY: An annual couple biography

The spells in the data file BIOCOUPLY contain retrospectively and prospectively collected information on couple history since a respondent’s year of birth on an annual basis. Since until wave 27 no questions on a respondent’s couple history were asked retrospectively, BIOCOUPLY includes only those respondents who have answered the biography questionnaire in wave 28. So except those, who were observed in the SOEP annually at least since the age of 17, people who answered the biography questionnaire before wave 28 are excluded from BIOCOUPLY. Thus, BIOCOUPLY comprises a much smaller sample than BIOCOUPLM and the BIOMARSM/Y datasets. The newly developed retrospective part of the person questionnaire covers up to four relationships in addition to the current status.

The data files contain thirteen variables. In contrast to BIOCOUPLM, BIOCOUPLY just contains the original household and individual identifier, due to the fact that no information about the identity of each partner was surveyed in the biography questionnaire and therefore a partner cannot be identified for the relationships before the entrance into the SOEP. The following table provides an overview of the variables in BIOCOUPLY, which contains beside the identifier eleven further variables.

Variables of BIOCOUPLY

HHNR	Identifier of original sample household
PERSNR	Personal identifier
SPELLNR	Consecutive spell number (chronological order)
SPELLTYP	Partnership status (-3) Implausible (1) Married, spouse in household (2) Married, spouse not in household (3) Coupled, partner in household (4) Coupled, partner not in household (5) Single (6) Married, separated (7) Registered same-sex partnership, partner in household (8) Registered same-sex partnership, partner not in household (98) Unknown (99) Unit nonresponse
BEGIN	Age of respondent when spell begins [-3=implausible; -1=missing]
END	Age of respondent when spell ends [-3=implausible; -1=missing]
BEGINY	Year when spell begins [-3=implausible; -1=missing]
ENDY	Year when spell ends [-3=implausible; -1=missing]
PDEATH	Death indicator: spell ends with death of partner?
DIVORCE	Divorce indicator: spell ends with divorce?
CENSOR	Censoring information [0 to 14] (see explanation above)
SOURCE	Source of information (1) derived only from biography questionnaire (2) derived from biography and personal questionnaire (3) derived only from personal questionnaire
REMARK	Further spell information (1) Original spell (2) Edited spell (3) Added spell (4) Gap Spell (5) First Spell

Variable SPELLTYP documents the partnership status with the possible categories listed above. Note that it was only asked for same-sex partnership since wave 28, thus, this information cannot retrospectively reconstructed for responds of the personal questionnaire. The state ‘single’ is not equal to the legal state ‘never married’; it refers only to a person’s current state of having a partner or not. ‘Married, separated’ spells apply only to former couples who are still married, i.e. who are separated but not yet divorced. It always overlaps

with other spells that contain the actual couple status(es) over the entire separation episode and is therefore redundant in terms of completeness of couple histories. The additionally assigned codes ‘implausible’, ‘unknown’ and ‘unit nonresponse’ indicate implausibility or a lack of information for the respective period.

Variable SPELLNR is a chronological index number for each individual’s spell during the observation period. Due to the fact that the spells’ duration is measured in months via the personal questionnaire, it is important to note that an individual may encounter several events in the same year. In this case the variable SPELLNR allows the user to order spells with respect to the respondent’s life course.The variables BEGINY and ENDY provide the years in which a spell begins and ends, whereas the variables BEGIN and END indicate respondent’s age for users’ convenience. In BIOCOUPLY, spell systems for each individual always start with the respondent’s birth. The SPELLTYP of the first spell per definition is ‘single’.

As SPELLTYP does contain missing values, so do BEGIN(Y) and END(Y), indicating that the exact year of change in the couple status is not known. Missing dates indicate that the year was either not reported (-1) (or especially for ‘married, separated’-spells the event of divorce is not reported) or that the reported year of change is implausible (-3), i.e. contradictory to other information. In order to differentiate the reasons for missing information the user can utilize the variables REMARK and CENSOR.

Consistent with BIOCOUPLM the data file BIOCOUPLY also provides the indicator variables PDEATH and DIVORCE. PDEATH indicates whether a respective spell ends with the death of a person’s partner, but is not restricted to married persons, thus does not only refer to widowhood. The states of widowhood can easily be derived from BIOCOUPLY by using the spells of marriage in SPELLTYP and checking for the death of the respective spouse in PDEATH (also BIOMARSM/Y contains the marital status and therefore information about widowhood). PDEATH of single spells are assigned a ‘(-2) does not apply’. DIVORCE works in a similar fashion, indicating whether the last marriage spell, that is the separated spell, ended in divorce. Hence, if it did not end in divorce it is coded as a still ongoing marriage. In this case END is updated by the year of the last interview.

Variable REMARK provides information on whether we had to edit or supplement original information given by respondents in order to construct consistent couple biographies. Spells in BIOCOUPLY are marked as ‘edited’ (in contrast to ‘original’) if the editing process involved substitution of reported information because of inconsistence with responses in previous or following interviews or information reported by a partner. Furthermore we have added spells, e.g. between two marriages because two marriages with separate persons at the same time are not legal (see later section for more details on editing); these spells are marked as ‘added spells’. Furthermore, ‘first spells’ and ‘gap spells’ are marked separately, whereas ‘added’ or ‘edited’ first spells are marked as ‘first spells’. For BIOCOUPLY we also generated a variable SOURCE, which contains information about whether the respective spell

was generated on basis of answers to the biography questionnaire or / and of responses to the personal questionnaire.

The variable CENSOR indicates whether a spell is left-censored, right- censored or censored on both sides (see Table 2 above within the explanations on BIOCOUPLM). The coding furthermore provides information on the reasons of censoring. In principle, spells might be censored if they precede or follow a gap spell or if BEGIN or END is missing. The last spell for each person is marked as right-censored if a person is still in the SOEP and the current relationship status is open (‘last spell’).

6.6 BIOMARSM: A monthly marital history

Spells in data file BIOMARSM contain prospectively collected information on marital biographies starting with the information reported in the first personal interview. The data file comprises eleven variables: the case and individual identifiers HHNR and PERSNR as well as nine spell specific variables. Variable SPELLTYP documents marital status with the possible categories ‘unmarried’, ‘married’, ‘divorced’, ‘widowed’ and ‘divorced or widowed’. Once married, a later spell ‘not married’ is not assigned anymore. Note that we renamed the known SOEP code ‘(1) single’ to ‘not married’. This is necessary to indicate that it is possible the respondent might have a partner anyway. If you are interested in this information, we recommend using BIOCOUPLM instead of BIOMARSM. Take also into account that we do not differentiate between marriages and registered same-sex partnerships for the SPELLTYP categories (3) to (6). SPELLTYP has one additional category ‘divorced or widowed’ which indicates that a marriage definitely ended though we do not know whether via divorce or death of the spouse. This may be due to missing information from the biographical questionnaires or due to a respondent’s frequent shifts between both categories without ever reporting the death of the partner or divorce as an event. A sixth state is ‘gap’ indicating a lack of reliable data for this period.

Variable SPELLNR is a chronological index number for each individual's spells during the observation period. Variables BEGIN and END indicate the month in which a marital

Variables of BIOMARSM

HHNR	Identifier of original sample household
PERSNR	Personal identifier
SPELLNR	Consecutive spell number (chronological order)
SPELLTYP	Marital status (1) Unmarried (2) Married (3) Divorced / reg. same-sex partnership annulled (4) Widowed / reg. same-sex partnership deceased (5) Divorced or widowed / reg. same-sex partnership annulled or deceased (6) Married, separated / reg. same-sex partnership, separated (7) Living in reg. same-sex partnership (9) Gap
BEGIN	Month when spell begins [1=Jan 1983 to 396=Dec 2015]
END	Month when spell ends [1=Jan 1983 to 384=Dec 2014]
BEGINY	Year spell begins [1983 to 2015; -3=implausible; -1=missing]
ENDY	Year spell begins [1983 to 2015; -3=implausible; -1=missing]
CENSOR	Censoring information [0 to 14] (see explanation above, Table 2)
EVENTS	Month information: exact month of spell begin or end known?
REMARK	Further spell information (1) Original spell (2) Edited spell (3) Added spell (4) Gap Spell (5) First Spell

spell starts and ends. Monthly histories start with a value of 1 in January 1983 and ranges until the current margin in month 384, i.e., December 2014. In principle, the month in which a spell starts is the very same month the previous spell ends in. Please take into account that the months of BEGIN and END are not imputed when the exact month of the change in the familiar situation is not reported. Instead, in those cases the time of the interview is left as the BEGIN or END date when a change is observed in the data. To distinguish between these two cases we introduced a new variable called EVENTS, containing the information whether the exact month of the BEGIN or END of a spell is reported or not.

It is important to note that a 'first spell' in BIOMARSM is not the very first spell of a person, but the first observed marital status since the person is taking part in the SOEP. Accordingly, the first spell in BIOMARSM is almost always left-censored. Variable CENSOR informs about whether a spell is left- or right-censored and if so, why. Most spells in BIOMARSM are in fact censored. In order to provide the user with detailed information on the nature of censorship we distinguished 'left', 'right', and combined 'left- and right-censored spells' with respect to the reason for censoring: 'first spell' or 'last spell', 'spell ends with death', spell

‘precedes’ or ‘succeeds a gap’ (see Table 2 above within the explanations on BIOCOUPLM). Of course, ‘death’ and ‘last spell’ are not mutually exclusive, thus we overwrite the latter with the former reason for being right-censored if the last interview is in the year of death or precedes it.

Variable REMARK provides information on whether we had to edit or supplement original information provided by respondents in order to construct consistent couple biographies. Spells in BIOMARSM are marked as ‘edited’ if the editing process involved substitution of reported information because of inconsistence with responses in previous or following interviews or information reported by a partner. Furthermore we have added spells, e.g. between two marriages because two marriages with separate persons at the same time are not legal (see later section for more details on editing); these spells are marked as ‘added spells’. Finally ‘first spells’ and ‘gap spells’ are marked separately, whereas also ‘added’ or ‘edited’ first spells are marked as ‘first spells’.

6.7 BIOMARSY: A annual marital biography

Data file BIOMARSY supplements BIOMARSM with retrospectively collected information on the marital history since a respondent’s year of birth. Whereas the marital history in BIOMARSM is measured in months, BIOMARSY depicts the marital biography on an annual basis. In contrast to BIOCOUPLY the BIOMARSY data set contains also the respondents to the biography questionnaire before wave 28. Please note, that until wave 28 the biography questionnaire just asked for three previous marriages and therefore the number of reported marriages in BIOMARSY is limited.

The BIOMARSY file comprises eleven variables. The individual and household identifiers HHNR and PERSNR as well as SPELLTYP are basically the same in all data sets. Once married, a later spell ‘not married’ is not assigned anymore. Note again that we renamed the known SOEP code ‘(1) single’ to ‘not married’ and take into account that we do not differentiate between marriages and reg. same-sex partnerships for the SPELLTYP categories (3) to (6). This is to indicate that it is possible the respondent might have a partner anyway. Like BIOMARSM, it is important to notice that SPELLTYP has one additional category ‘divorced or widowed’ which indicates that a marriage definitely ended, though we do not know whether via divorce or death of the spouse. This may be due to missing information from the biographical questionnaires or due to a respondent’s frequent shifts between both categories without ever reporting the death of the partner or divorce as an event.

Variables of BIOMARSY

HHNR	Identifier of original sample household
PERSNR	Personal identifier
SPELLNR	Consecutive spell number (chronological order)
SPELLTYP	Marital status (1) Unmarried (2) Married (3) Divorced / reg. same-sex partnership annulled (4) Widowed / reg. same-sex partnership deceased (5) Divorced or widowed / reg. same-sex partnership annulled or deceased (6) Married, separated / reg. same-sex partnership, separated (7) Living in reg. same-sex partnership (9) Gap
BEGIN	Age of respondent when spell begins [-3=implausible; -1=missing]
END	Age of respondent when spell ends [-3=implausible; -1=missing]
BEGINY	Year when spell begins
ENDY	Year when spell ends
CENSOR	Censoring information [0 to 14] (see explanation above, Table 2)
SOURCE	Source of information (1) derived only from biography questionnaire (2) derived from biography and personal questionnaire (3) derived only from personal questionnaire
REMARK	Further spell information (1) Original spell (2) Edited spell (3) Added spell (4) Gap Spell (5) First Spell

Regarding the fact that duration of spells is measured in years it is important to notice that an individual may encounter several events in the same year. In this case the variable SPELLNR allows the user to order the spells with respect to a respondent's life course. The variables BEGINY and ENDY provide the years in which a spell begins and ends, while the variables BEGIN and END indicate for users' convenience the respective age of the respondent. The spell system for each individual in BIOMARSY always starts with the birth of the respondents. We thus created a first spell for each individual ever interviewed in the SOEP starting in the year of birth and continuing at least until the year in which a person turns age 16. The SPELLTYP of the first spell per definition is 'unmarried'. Even if a respondent reported an earlier marriage in the biography questionnaire we restricted its beginning to age 16, however, we marked the beginning of this spell as 'implausible'.

There are some missing values (-1 and -3) in BEGINY as well as in ENDY (resp. BEGIN and END) indicating that we do not know the exact year of a change in the marital status. This can have two reasons. First, it may simply indicate that the respondent did not report the year in which a marriage began or ended. In order to differentiate the reasons for missing information the user may utilize the variables REMARK and CENSOR. Second, within single case corrections some dates are set 'implausible (-3)' if overlapping of marriages appear unsolvable or contradictions with the reported year and i.e. the year of death exist.

REMARK indicates whether a spell was ‘edited’ or ‘added’ in the same way as in BIOMARSM (see above). Variable CENSOR indicates if a spell is left or right censored or censored on both tails. The coding of CENSOR (see Table 2 above) provides also information about the reasons for censoring. In addition to what was said before about gap spells in BIOMARSM, gaps or missing values in BEGIN or END may appear in BIOMARSY if a respondent reported a terminated first marriage and the beginning of a second marriage, but did not report the reason for and/or the year of the end of the first marriage. The last spell of each person is marked as right censored, irrespective that the person died, quit the SOEP or is still married currently.

7 BIOBIRTH: A Data Set on the Birth Biography of Female Respondents[25]

by Christian Schmitt

7.1 Population and purpose of the data set BIOBIRTH

The file BIOBIRTH provides information on fertility histories of adult respondents in the SOEP. Until 2014 (version 30, wave BD) the data was stored in two separate files: BIOBIRTH containing female fertility histories, and BIOBRTHM providing male fertility histories. It is important to note that the latter file only records the male fertility histories for respondents who entered the SOEP in 2001 or later (for more details see below). Since 2015 (version 31, wave BE) all fertility histories of new respondents as well as the old and continuously updated data of the fertility histories collected in BIOBIRTH (until 2014) and BIOBRTHM (until 2014) are stored in a single file - BIOBIRTH (this naming is identical to the previous fertility histories of women). The variable SEX (distinguishing male and female respondents) is added to the BIOBIRTH dataset since wave BE. Moreover, the file BIOBIRTH is also supplemented with the fertility histories of the "Familien in Deutschland" (FiD) panel survey, which was integrated into the SOEP data-base in 2015. Records from the FiD subsamples can be distinguished by inspecting the variable BIOVALID. For more details on integration and extension procedures see below.

Fertility histories in BIOBIRTH provide information on every woman (as well as every man with a panel entry since 2001) who has ever provided at least one successful SOEP interview. Note that the data is right censored for respondents who left the panel early in their fertile life-phase and it is left censored for persons who entered the SOEP at higher ages *without* ever filling in a biographical questionnaire. The variable BIOVALID provides information on whether individual level information is based merely on information derived from household composition and family relations, or on biographical questionnaire data. The variables EINTRITT and AUSTRITT in ppfad (panel entry and exit) provide information on censoring (left-censoring can be ignored if a biographical questionnaire exists for a given person)

For each of the mentioned adult respondents BIOBIRTH documents the fertility history. The annual update focuses on including new information on becoming a biological parent as based on data collected with the individual or the biographical questionnaire, respectively. Furthermore adults who have been interviewed for the first time but who have not yet provided information on their fertility histories are included. The latter case applies to either new adult household members, or teenagers who have reached the required minimum for a participation in the personal questionnaire (16 years). BIOBIRTH constitutes an accumulative data set, in which the entire birth biography of all SOEP respondents is presented.

[25] Information on female birth biographies was designed with reference to earlier works by Joachim R. Frick.

7.2 Structure of the data set

BIOBIRTH covers the following information:

(1) Person identifier (PERSNR), respondent's year of birth, status information on the origin of the included data, number of children derived from fertility histories, total number of children derived from fertility histories *and* subsequent consideration of changes in household structure up to the last date of interview.

(2) A sequence of 15 variables relating to 1 out of 15 children, including child's person identifier (KIDPNR[nn], provided the child could be identified within the SOEP's household structure), child's sex, child's year of birth, child's month of birth.

BIOBIRTH contains the following variables for all adult women (and men since 2001):

- HHNR Invariable number of the original household
- PERSNR Invariable personal identifier of the respondent
- SEX Respondent's sex
- GEBJAHR Respondent's year of birth
- BIOVALID Status of the birth biography: (Please mind: The codes in the variable BIOVALID have been extended in 2015 (wave BF)).

  10: no birth biographical entries

  20: youth biography questionnaire, no fertility histories on children

  30: birth biography questionnaire, no children in fertility history

  31: birth biography questionnaire, one or more children in fertility history

  40: FiD: no birth biography, no data on children

  41: FiD: no birth biography, information on existing children in FiD parent/couple questionnaire

  50: FiD: birth biography questionnaire, no children in fertility history

  51: FiD: birth biography questionnaire, one or more children in fertility history.

- BIOYEAR Survey year when birth biography questionnaire was completed (1985ff.).
  Code "-2" is assigned if the respondent never completed a birth biography questionnaire.

- BIOAGE — Age of the woman at the time of the birth biography survey. Code "-2" is assigned if the respondent never completed a birth biography questionnaire.
- SUMKIDS — Total number of children born (more precisely: total number of children identifiable within SOEP by merging all available data up to the time of the last observation (SUMKIDS=BIOKIDS+subsequent births during panel participation).
- BIOKIDS — Total number of children identified in the birth biography. Code "-2" is assigned if the respondent never completed a birth biography questionnaire.
- KIDGEB[nn] — Year of birth of the child [nn] (for the first child up to the fifteenth child).
- KIDSEX[nn] — Sex of the child [nn] (for the first child up to the 15th child).
- KIDPNR[nn] — Personal number of the child [nn] (for the first child up to the 15th child), given it is identifiable in the SOEP.
- KIDMON[nn] — Month of birth of a child [nn] (for the first child up to the 15th child).

  With respect to the variables KIDGEB[nn], KIDSEX[nn], KIDPNR[nn], and KIDMON[nn] identical missing codes apply: The code "-2" is assigned if there's no [nn]th child identified for a mother/father. The code "-1" applies if an [nn]th child can be identified but the information on the birth year and/or sex and/or personal identifier is unavailable, or if it could not be identified.

For every respondent a maximum of up to 15 children are considered. The sequence of children within BIOBIRTH is recorded with regards to the birth order in terms of age of the children. The order ranks from the oldest child specified under KIDPNR01 to the youngest child. If the age is missing it is listed in the first record (KIDPNR01), and in subsequent records following KIDPNR01 if more than one child's personal identifier remains missing.

7.3 Information basis of the birth biography

The main basis of the individual fertility history considered in BIOBIRTH is the information collected with the biography questionnaire[26], in which the number, birth year and sex of the biological children for every adult respondent are collected. For adults with information on

[26] The information collected over the course of the biography survey for every adult contains the number of children, the year of birth, the sex, residence status of the child, and, if applicable the year of death of the biological child. The biography data is stored in wave specific files ($LELA), which are not provided with the SOEP distribution.

children stemming from the biography questionnaire the BIOVALID code "31" is assigned. Women who completed this questionnaire, but did not report any biological children receive the code "30". In correspondence to this, the samples of the FiD panel integrated into the SOEP since 2015 (wave BE) contribute the codes 50 and 51 for respondents with biography questionnaires with, or without children, respectively. Codes 10 and 40 correspond accordingly for SOEP and FiD data. 41 is a qualitatively new code for parents in the FiD who did not fill in the birth biography but who have provided reliable information on their status as a parent in other sources of FiD (this relates to specific questions in the FiD parent and couple questionnaire).

A minority of respondents did not provide any information on fertility histories by filling in the biographical questionnaire for several reasons[27]. For these cases the variable BIOVALID is assigned the code 10 (original SOEP samples), and 40 (new FiD subsamples), respectively. This group is subject to a risk of underestimating the total number of births, particularly since births prior to the entry into the SOEP cannot be identified unless current household structure provides substantive evidence, usually reflected by parent-child co-residence. Respondents without a valid fertility history (codes 10 and 40) can be distinguished in three major groups:

- Respondents who were 16 years of age at the time of the first interview. In most cases these respondents participate in the biography survey at a later date. Thus, the parent-child relationship recorded earlier in BIOBIRTH (as based on household structure) can be verified and supplemented with data on their fertility histories at a later date.
- Respondents who were at about 30 years of age or younger at the time of first interview. In this sub-population, children are not yet adults and still reside in the parental home in most cases. Since information from the biographical questionnaire is missing, a final distinction in social and biological children is difficult particular for records considered in earlier SOEP waves when the intra-household relationships ($STELL) were less refined, compared to the recent setup. Hence these older records have a higher likelihood of mis-specifying a social as a biological parent-child relationship.
- Respondents who were well over 30 years of age at the time of the first interview. In these cases some of the children are likely to already have left the parental home, and therefore are no longer part of the survey population. For that reason, the number of biological children might be underestimated in this group of respondents to a larger extent as compared to younger women.

[27] Beside the reason 'refusal', the collection date of the life history biographies differ among SOEP sub-samples.

7.4 A new source of biographical information – the youth questionnaire

From wave T onwards the data within BIOBIRTH includes information of a further biographical instrument: the youth biography. The youth-questionnaire has been in circulation since the year 2000 (wave Q) for all young adults, one year after they have reached the required age for completing the individual-questionnaire. Apart from exceptions described in table 1, this means the age of 17. What is important for the BIOBIRTH dataset is that these individuals who fill in the youth-questionnaire complete this questionnaire *instead* of the biographical questionnaire. The age groups which instead fill in the youth-questionnaire of the biographical module differ slightly among the SOEP-subsamples (table 1):

Table 1: Target population of the Youth Questionnaire by year, sample and age

sample	2000	2001	2002	2003 and later
A-E	17 years	17 years	17 years	17 years
F		17-19 years	17 years	17 years
G, H, I, J, K, M				17 years

Source: SOEP v32, doi: 10.5684/soep.v32

The youth-biography does not contain any birth-biographical modules. Assuming that only very few women give birth before the age of 17 and these few can be identified in the household context (as long as they remain within the SOEP), this does not pose any problem for compiling the birth-biography of the respondents. Nevertheless, a few changes to the BIOBIRTH dataset have to be outlined:

- In the variable BIOVALID a new code ("20": "youth biography questionnaire completed") is added. As the youth questionnaire doesn't contain any information about own children the addendum "no children in biography" is always added to the code "20".
- While calculating the age at the time of the biographical questionnaire (BIOAGE), the age upon completion of the youth questionnaire is applied.
- The variable BIOKIDS always remains at zero as no biographical information on parenthood can be derived from the youth-biography (in this cases no missing code is applied in BIOKIDS).

7.5 The fertility histories of male respondents in BIOBIRTH

Since 2001 the fertility histories of male respondents are collected with the SOEP survey and have been integrated in a special data set BIOBRTHM since 2003 (wave T). Since 2015 (wave BE) this data is merged with the female fertility history and collected in the file BIOBIRTH, while the variable SEX distinguishes between the fertility histories of men and

women. Contents and updating procedures for male fertility histories are identical to those of their female pendants with two important exceptions:

- First: only information about men with at least one completed questionnaire *in 2001 or later* is considered in the BIOBIRTH file.
- Second: information from the birth-biography will only be added for *new* panel members who joined since 2000, as only these persons fill in a new biography interview (usually one wave after the first participation in the SOEP which in our case means in 2001 or later). Most of the members who have completed a questionnaire before 2000 have also already completed the biographical modules that are only collected once for every person.

The module collecting information on male fertility histories was introduced in 2001. Therefore, most men in subsample "F" (which started in 2000) have completed the birth-biography, since biographical questionnaires are usually completed one wave after the starting wave with only a marginal rate of non-response for this life-course related questionnaire. Note that for men within BIOBIRTH who did not complete a biography questionnaire (BIOVALID code 10), the information about fatherhood is underestimated as only the context of the household is available to determine the respondents biological children. Underestimation of this type is generally more severe for men than for women, since children have a distinctively higher likelihood to co-reside with their mothers after separation, in which case a father will appear to be childless, thus leading to a misspecification as being childless.

7.6 Integration of "Familien in Deutschland" – FiD

In 2015 (wave BE) data on respondents from the "Familien in Deutschland" (FiD) panel survey was integrated into the SOEP data-base. While the design of this survey was oriented on a close congruence to the SOEP questionnaires and data structure, FiD introduces some additional information. The focus of the FiD survey was put on couples, families and lone parents. As a consequence, there is even more information on parent-child relationships compared to the original SOEP samples. Hence the derivation of parent-child relationships is quite reliable for the FiD subsamples (see variable PSAMPLE in the file PPFAD for a distinction). Almost all adult respondents completed a specific biographical questionnaire including data on their fertility histories. These respondents were assigned BIOVALID codes 50 (without children), and 51 (with children). Additionally, particular questions in a couple, as well as a parent-child-questionnaire provide data on biological children. Hence, the original BIOVALID code 10, which corresponds to the code 40 for the FiD population, shows only few case numbers, while a new code 41 was introduced for parents, who lack detailed information on their fertility histories, but which provide data on children in FiD's couple or parent questionnaire. For the update of the BIOBIRTH population and data in subsequent

waves, the FiD population will be treated identical to any other of the SOEP subsamples (see 1.10 in particular).

7.7 Identification process of the children in the SOEP data base

The starting point for the process of identifying children is the relationship of a household member to the head of the household (HH) (variable $STELL in the file $PBRUTTO). Until wave 29, the variable $STELL had the following codes:

Code	Label
0	head of the household (HH)
1	spouse of HH
2	"life companion" of HH
3	daughter / son (including adopted/step-children) of HH
4	foster child of HH
5	daughter in law / son in law of HH
6	father / mother of HH
7	father in law / mother in law of HH
8	brother / sister / brother in law / sister in law of HH
9	grandchild of HH
10	other relation to HH
11	not related to HH
12	child of "life companion" of HH (included since 1999)

However, there are only certain combinations among household members in which a biological parent-child relationship between a female adult and another person can be assumed.

Table 1: Potential parent-child relationships as a combination of the variable $STELL

$STELL of the		Potential parent-child relationship
woman	another person	In this case the person is the…
0 1	3 3	Child of reference person (reference person = head of the household) Child of the wife of reference person
1 1	11 12	Child of the wife of reference person, but not child of reference person
2 2 2	3 11 12	Child of "life companion" of reference person and of reference person Child of "life companion" of reference person but not of reference person
3 4 5	9 9 9	Child of daughter of reference person Child of foster child of reference person Child of daughter in law of reference person (3 generation household)
6 6	0 8	Child is reference person, lives with his mother/father in the same household Child is the sister / brother of reference person, the siblings live with their mother/father in the same household
7 7	1 8	Child is spouse of reference person and lives together with spouse and mother/father in the same household Child is daughter / son of the mother/father in law of reference person, but not the

		spouse of the reference person rather the sister in law / brother in law of reference person
8	10	Child is niece / nephew of reference person, parent is sister / sister in law of reference person
9	10	Child is another relation to reference person, great grandchild of reference person
10	10	Mother/father and child have another relation to reference person
11	11	Child and mother/father are in no way related to reference person

Source: SOEP v32, doi: 10.5684/soep.v32

With SOEP wave 29 a more complex representation of the variable $STELL in the file $PBRUTTO has been implemented, describing the following types of relationships to the head of household:

Code	Label
0	Head Of Household
11	Spouse Of HH Head
12	Same-Sex Spouse
13	Life Partner
21	Son, Daughter
22	Stepchild (Child of the Partner)
23	Adoptive Child
24	Foster Child
25	Grandchild
26	Great-Grandchild
27	Son, Daughter-In-Law
31	Father, Mother
35	Parent-In-Law
36	Grandparents
41	Brother, Sister
42	Half-Brother, Half-sister
43	Stepbrother, Stepsister
51	Brother, Sister –in Law (Spouse/ Life Partner of Siblings)
52	Brother, Sister -in Law (Siblings of Spouse/ Life Partner)
61	Aunt, Uncle
62	Niece/ Nephew
63	Cousin/Cousine
64	Other Relative
71	Others
99	Relationship to Head of HH Unknown

Source: SOEP v32, doi: 10.5684/soep.v32

Based on these new representations of relationship patterns considered since wave BC, the algorithm, scanning for potential parent-child dyads has been adjusted and implemented in generating the birth-biographical information that relies solely on the household composition. It should be noted that the majority of parent-child relations provided with the BIOBIRTH biography file are still derived from personal information given in the biographical core-questionnaires (see below).

7.8 Identification of the children of parents with completed fertility histories

If a parent mentions either the existence of biological children, the year of birth of a child, the sex of a child or co-residence with this initiates a process of identification. In the first step, the algorithm scans the current household structure and aims to determine the parent's relationship to the reference person and scans potential combinations of parent child relationships within the current household structure. The principle is the same as outlined above (1.6). The goal is to supplement the biographical information on sex and year of birth of a child with a personal identifier, thus enabling more complex analyses of parent-child interaction for the researcher. If a fit is determined between the year of birth and the sex of a potential child in birth biography and household, the individual is considered as the child of the respondent and the child's personal identifier is assigned. Since the majority of the households with children present small nuclear families including one potential mother/father, this kind of identification process is broadly sufficient. In other, rather complex households a careful hand editing is conducted, in order to identify the 'right' child to the 'right' mother/father. The same is done, if the sex or the year of birth of a child mentioned in the biography questionnaire is unspecified, i.e. missing.

In the case of a successful identification the variable KIDPNR[nn] is been filled with the person identifier of this child. Children, for whom the parent in the biography questionnaire has reported that they were deceased or had moved out, are assigned the personal number (KIDPNR[nn]) "-1", for missing information, in BIOBIRTH.

7.9 Identification of the children for women who have no biography data/ not completed the biography questionnaire

To get as close as possible to the definition of a biological child, for this group of parents only, specific relationships among household members are considered. Since the key information from the biography questionnaire is not available, a careful analysis of the composition and the history of the household in which the children live is conducted in order to assign proper parent-child linkages.

7.10 Updating BIOBIRTH

As mentioned in section 6.1 the annual update of the data set BIOBIRTH is examined with respect to two dimensions. First, updating the birth biography of the BIOBIRTH population and second, extending BIOBIRTH by new persons. The latter are either new adult respondents joining the household or teenagers who have reached the required age for giving a first interview (16 years). Since the extension of BIOBIRTH follows the generation rules as described above, the following only summarizes the updating of the BIOBIRTH population.

New-born children in the SOEP study are documented in the variable $PZUG in the data set $PBRUTTO:

Code	Label
11	Born since the last survey
17	Born before the last survey, but only now first mentioned
31	Born two years ago

Source: SOEP v32, doi: 10.5684/soep.v32

For this group of new born children the parent-child ties are investigated by the same algorithm which has been described in section 1.6. The general logic of updating the BIOBIRTH birth biography follows this pattern:

1) New information from household composition (basis $STELL) updates older information contained in BIOBIRTH (i.e. new-born children, or children moving into the household or otherwise observed in the SOEP for the first time).

2) New or first time information from the biographical questionnaire is displayed as core part of the parent child-ties in BIOBIRTH, or replaces older household-based information contained in BIOBIRTH, in case household structure-based data is inconsistent with the data from the biographical questionnaire.

3) New information, derived from household composition (basis $STELL), may not replace earlier information in BIOBIRTH, in case the previous information on children stems from the biographical questionnaire, and the two sources cover the same period and are inconsistent. In other words: information from the biographical questionnaire always overrides household information..

Overview: Central variables in the file BIOBIRTH (Version 2015 / Up to Wave BF)

Biovalid

		Frequency	Percent	Cumulative Percent
Valid	10 No Birth Biography - No Kids From Bio	11170	15,3	15,3
	20 Youth Biography - No Kids From Bio	5464	7,48	22,78
	30 Birth Biography - No Kids From Bio	16234	22,23	45,01
	31 Birth Biography - Kids From Bio	29142	39,91	84,91
	40 FiD: No Birth Biography	1063	1,46	86,37
	41 FiD: Parent/Partner Questionnaire – Data on Kids	863	1,18	87,55
	50 FiD: Birth Biography - No Kids From Bio	642	0,88	88,43
	51 FiD: Birth Biography - Kids From Bio	8450	11,57	100
	Total	73028	100	

Source: SOEP v32, doi: 10.5684/soep.v32

BIOYEAR Year of Biography Survey

Year of Biography Survey	Frequency	Percent	Cumulative Percent
-2	13.096	17,93	17,93
1985	6.618	9,06	27
1986	83	0,11	27,11
1987	104	0,14	27,25
1988	220	0,3	27,55
1989	211	0,29	27,84
1990	202	0,28	28,12
1991	161	0,22	28,34
1992	2.635	3,61	31,95
1993	230	0,31	32,26
1994	592	0,81	33,07
1995	528	0,72	33,8
1996	255	0,35	34,14
1997	231	0,32	34,46
1998	203	0,28	34,74
1999	1.033	1,41	36,15
2000	361	0,49	36,65
2001	9.429	12,91	49,56
2002	899	1,23	50,79
2003	2.691	3,68	54,47
2004	822	1,13	55,6
2005	667	0,91	56,51
2006	530	0,73	57,24
2007	2.576	3,53	60,77
2008	594	0,81	61,58
2009	438	0,6	62,18
2010	9.597	13,14	75,32
2011	6.908	9,46	84,78
2012	3.227	4,42	89,2
2013	651	0,89	90,09
2014	4.606	6,31	96,4
2015	2.630	3,6	100
Total	73028	100,0	

Source: SOEP v32, doi: 10.5684/soep.v32

SUMKIDS Total Number of Births

Total Number of Children Born	Frequency	Percent	Cumulative Percent
0	26310	36,03	36,03
1	13921	19,06	55,09
2	19720	27	82,09
3	8782	12,03	94,12
4	2790	3,82	97,94
5	888	1,22	99,16
6	328	0,45	99,6
7	138	0,19	99,79
8	79	0,11	99,9
9	28	0,04	99,94
10	23	0,03	99,97
11	11	0,02	99,99
12	7	0,01	100
16	2	0	100
17	1	0	100
Total	73028	100.00	

Source: SOEP v32, doi: 10.5684/soep.v32

BIOKIDS Number of Births from Biography

Births from Birth Biography Questionnaire	Frequency	Percent	Cumulative Percent
-2	13096	17,93	17,93
0	22340	30,59	48,52
1	11293	15,46	63,99
2	15589	21,35	85,33
3	7169	9,82	95,15
4	2283	3,13	98,28
5	714	0,98	99,26
6	282	0,39	99,64
7	129	0,18	99,82
8	71	0,1	99,92
9	32	0,04	99,96
10	16	0,02	99,98
11	7	0,01	99,99
12	4	0,01	100
16	2	0	100
17	1	0	100
Total	73028	100.00	

Source: SOEP v32, doi: 10.5684/soep.v32

8 BIOTWIN: TWINS in the SOEP

by Christian Schmitt

8.1 Population and contents of the data set BIOTWIN

The file BIOTWIN contains all twins that were ever identified within the SOEP. To be classified as a twin, a person is required to:

- have exactly the same age as his or her sibling (year & month of birth),
- have a relationship to the head of the household that indicates that he or her and a second persons are siblings, and
- have the same mother (as far as a pointer to the mother is available).

Furthermore, it is not only twins that are recorded in the BIOTWIN data set, but also triplets or quadruple siblings. The following variables are stored within the BIOTWIN data set:

- HHNR Invariable number of the original household.
- PERSNR Invariable personal identifier of the first sibling.
- PNRTWIN Invariable personal identifier of the second sibling, the twin.
- PNRTRIP Invariable personal identifier of the third sibling.
- PNRQUAD Invariable personal identifier of the fourth sibling.
- PNRMOTH Pointer to the personal identifier of the mother of the twin-group.
- BIOMONOZ Monozygotic group? Information if the group is monozygotic.
- INFSOURC Source of information from which the status of being a twin is derived

The central variable PERSNR is assigned to the sibling with the lowest personal identifier in the twin group. The PNRTWIN and – in rare cases if available – PNRTRIP or PNRQUAD contain the personal identifier of second, and third or fourth sibling in the group. This means that every case in the data set consists of a *group* of twins (or triplets or quadruplets). The code "-2" is assigned to PNRTRIP and/or PNRQUAD if a third or fourth twin sibling doesn't exist. PERSNR and PNRTWIN however should always contain valid codes.

The variable PNRMOTH provides the link to the mother of the group and is derived from the data sets $KIND (reference to this $KIND was discontinued in with wave Z / 2009) and/or BIOBIRTH.

8.2 The twin survey of 2006

In 2006, a questionnaire was distributed among all households with potential twin groups, identified up till then. The aim was to validate that none of these twins had been identified by mistake. The variables INFOTWIN and BIOMONOZ contain new information which was derived from this survey.

The result of the survey could widely validate the selection of the twin population, contained in the BIOTWIN data set of the SOEP. More than 80% of households with potential twins as of 2006 could be contacted and were interviewed in the twin survey. Among these only 3 groups of twins turned out to be identified erroneously (those false positives were removed from the BIOTWIN data set). Thus the algorithms of identifying twins within the SOEP could prove to be widely reliable. Additional information that was collected with the twin survey contributed to identifying a number of mothers of twins, for whom the mother-child-link was missing previously. Furthermore the twin survey provided additional information on monozygotic respectively dizygotic twins. The variable BIOMONOZ was extended, in order to reflect this additional information (see below for more details).

8.3 Construction of variables in the data set BIOTWIN

The variable BIOMONOZ[28] indicates if the group is monozygotic. If the information could be validated in the twin-survey in 2006 the code is set to 1 for monozygotic twins and 2 for dizygotic twins. If the information on being mono- or dizygotic twins could *not* be validated in the twin survey, which was carried out in 2006, the code is set to 0 if the sex of all the siblings is identical, and this group thus *might* be monozygotic. Please pay attention to the fact that the labels and values of the variable BIOMONOZ from wave W onwards are not consistent with values and labels from previous waves.

The variable INFOTWIN is introduced with wave W and provides information on the source from which the status of being a member of a twin group is derived from and whether this information could be validated in the twin-survey in 2006.

INFOTWIN can take the following characteristics:

1 Generated up to 2006 – basis: household co-residence, identical parent, year & month of birth – not validated by in the twin survey 2006

2 Possible Twin or Triplet – Information not revisable in twin survey 2006

3 Possible Twin or Triplet –Answer refused in twin survey 2006

[28] This variable existed before wave W but was restructured to reflect the additional information which became available with the 2006 twin questionnaire.

4 Twin or Triplet – Information validated by twin survey 2006

5 Twin or Triplet – New since 2006
(congruent year & month of birth)

6 Twin or Triplet – New since 2006
(congruent year of birth / missing month of birth)

The selection of twins within the SOEP, which compiles the data set BIOTWIN, is based on the either the month of birth, or an identical year of birth. Priority is given to congruent months of birth, as a woman might – in rare cases – give birth at two different times in a year. Hence the month of birth plays a central role in identifying potential twin-groups. According to that logic people with a) valid month of birth information or b) identical month of birth, or c) with an identical year of birth *and* missing data on the month of birth among both siblings are classified as twins.

In a second step, the relationship of these potential twins to the head of household is scanned ($STELL). If the relationship of both persons assures that they are siblings, then they are assumed to be twins.

In a third step the pointer to the mother is checked for both siblings with focus on the files $kind / BIOBIRTH. If this maternal link is identical for both siblings, it is transferred into the variable PNRMOTH.

An overview of central information in the file BIOTWIN (Version 2015 / Wave BF)

Table 1: Siblings in BIOTWIN[29]

Sibling type	n	Valid Mother Pointers
Twins	1428	1278
Triple	50	50
Quadruple	4	4

Source: SOEP v32, doi: 10.5684/soep.v32

[29] Please note: sibling groups contribute observations for each individual of the twin-pair/triplet, that is – a pair of twins would provide two entries to the data set, one for each twin.

Table 2: BIOMONOZ Gender Combination of Siblings

		Frequency	Percent	Cumulative Percent
Valid	(-1) No Answer	2	0,14	0,14
	0 Possibly Identical Twins	721	50,49	50,63
	1 Definitely Identical Twins	46	3,22	53,85
	2 Definitely Fraternal Twins	659	46,15	100
	Total	1428	100	

Source: SOEP v32, doi: 10.5684/soep.v32

Table 3: INFOTWIN Twin Status: Source of Information

		Frequency	Percent	Cumulative Percent
Valid	1 Generated - not in twin survey 2006	45	3,15	3,15
	2 Twinsurvey 2006 (answer not verified)	73	5,11	8,26
	3 Twinsurvey 2006 (answer refused)	2	0,14	8,4
	4 Twinsurvey 2006 (answer validated)	196	13,73	22,13
	5 Gen. since 2007 (basis: year of birth & month)	717	50,21	72,34
	6 Gen. since 2007 (basis: year of birth / month miss)	395	27,66	100
	Total	1428	100	

Source: SOEP v32, doi: 10.5684/soep.v32

9 BIOSIB: Information on siblings in the SOEP

by Josephine Kraft and Daniel D. Schnitzlein

9.1 General description of the data set

BIOSIB provides information on siblings living within the SOEP households. The data set contains the person numbers of all siblings in an observed family. It includes information on their sex, their year of birth, the number of siblings, the individual's position within the birth order, and on the relationship between the observed siblings.

9.2 Sources of information on siblings in the SOEP

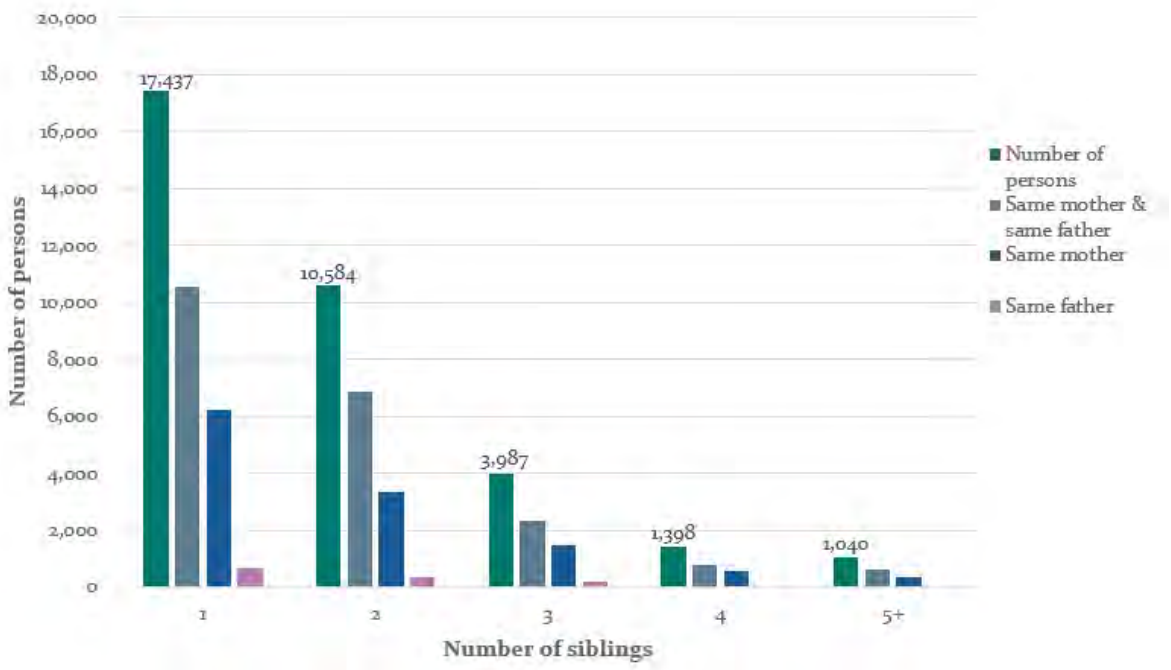
Information on siblings in the SOEP is available from three sources:

- First, the respondents are asked about their siblings in the biography questionnaire and the youth questionnaire. This information (for example *have or ever had siblings yes/no, number of sisters, number of brothers*) is stored in the file BIOPAREN (for detailed information please see the chapter on BIOPAREN).
- Second, in the years 1991, 1996, 2001, 2006, 2011 and 2013 all respondents were asked about their family relations. Among other questions on their family, the individuals were asked about their siblings (like above, *have or ever had siblings yes/no, number of sisters, number of brothers*) outside the SOEP household in 1991, 1996 and 2001. From 2006 on they were asked about all siblings within and outside the household. The information from these questions is stored in the $$p-files.
- Third, siblings within the SOEP households are observed directly. The aim of this file is to provide the never changing person ID of the siblings of each respondent as far as they can be identified in the SOEP. With this information it is possible to use the whole range of personal or household information for each sibling to carry out detailed sibling analyses. This file therefore adds to the information in BIOTWIN (for details see chapter on BIOTWIN), which provides the person IDs for all twins in the SOEP.

9.3 Overview on the number of siblings in BIOSIB

Figure 1 gives an overview on the information stored in BIOSIB. The data set contains 13,716 families with 34,484 individuals that have at least one sibling identified in the SOEP. 17,437 of them have one sibling, 10,584 have two siblings, 3,987 have three siblings, 1,398 have four siblings and 1,040 have five or more siblings identified in the data. As is apparent from Figure 1 in all cases most of the siblings are identified by having the same mother and the same father (for more details about the identification of siblings and the identification of biological siblings please see the description to SIBDEF1-SIBDEF11 on the subsequent pages).

Figure 1: Number of persons with information on their siblings, by number of siblings



Source: own calculation, based on SOEPv32, doi: 10.5684/soep.v32

9.4 Organization of the data in BIOSIB

Each row in the dataset represents one individual for which at least one sibling could be identified. Therefore a family with three siblings appears three times in BIOSIB, one time for each child. The person IDs of the siblings are ordered by birth order starting with the oldest sibling.

List of variables

HHNR	Original Household Number
PERSNR	Never Changing Person ID
SIBPNR1 – SIBPNR11	Person Number of 1st – 11th Sibling
SIBDEF1 – SIBDEF11	Sibling Relation to 1st – 11th Sibling
FAMCOUNT	Family Counter
POS_SIB	Position in the birth order
NUM_SIB	Number of observed siblings in the SOEP
SEX	Gender of Individual
GEBJAHR	Year of Birth of Individual
SEXSIB1 – SEXSIB11	Gender of 1st – 11th Sibling
GEBSIB1 – GEBSIB11	Year of Birth of 1st – 11th Sibling

The variables HHNR, PERSNR, SEX, GEBJAHR, SEXSIB1-SEXSIB11 and GEBSIB1-GEBSIB11 are generated from the information stored in PPFAD. The newly generated variables SIBPNR1-SIBPNR11, SIBDEF1-SIBDEF11, FAMCOUNT, POS_SIB, and NUM_SIB are described on the next pages.

Variable	SIBPNR1 – SIBPNR11
Label:	Person Number of 1st-11th Sibling
Values:	(-1) No answer (-2) Does not apply (-3) Answer improbable
Description:	The variables provide the never changing person IDs for the siblings of the individual identified by PERSNR. The sibling relationship is generated from the parent information in BIOBIRTH, BIOBRTHM and BIOPAREN (for detailed information on these files please see the relevant chapters above). Two persons are defined as siblings if they report both, the same mother and father, only the same mother, or only the same father. This information on the sibling relationship is stored in SIBDEF1-SIBDEF11. In the case of inconsistent information on parents in BIOBIRTH and BIOPAREN, BIOPAREN was assigned the lowest priority. Please note, that BIOPAREN uses a social definition of parenthood based on cohabitation. In contrast, BIOSIB contains both biological (BIOBIRTH/BIOBRTHM) and social siblings with a higher priority on biological relations.

Variable	SIBDEF1 – SIBDEF11
Label:	Sibling Relation to 1st-11th Sibling
Values:	(-1) No answer (-2) Does not apply (-3) Answer improbable (1) Same mother (B); same father (B) (2) Same mother (B); same father (nB) (3) Same mother (nB); same father (B) (4) Same mother (nB); same father (nB) (5) Same mother (B) (6) Same mother (nB) (7) Same father (B) (8) Same father (nB)
Description:	The variables provide the information on the sibling relationship between the individual identified by PERSNR and the respective sibling. Two siblings can have either, the same mother and the same father, only the same mother, or only the same father. The indicator further provides information if the identified parent is a biological (B) (indicator from BIOBIRTH) or non-biological parent (nB) (indicator from BIOPAREN). So for example, variable value (1) indicates that the two individuals share the same biological mother and the same biological father.

Variable	**FAMCOUNT**
Label:	Family Counter
Values:	(-1) No answer (-2) Does not apply (-3) Answer improbable
Description:	The variable contains a non-systematic counter of families occurring in BIOSIB. All siblings (biological and non-biological), who belong to one family, are assigned the same value of FAMCOUNT. The variable can be used, for example, in multilevel analyses to define the family level. Note: In the case a family splits up, children from the new partnerships of the parents are no siblings. Children of the early partnerships are siblings to all children in the new partnership.

Variable	POS_SIB
Label:	Position in birth order
Values:	(-1) No answer (-2) Does not apply (-3) Answer improbable
Description:	The variable contains the individual's position in the birth order of the observed siblings (biological and non-biological).

Variable	NUM_SIB
Label:	Number of siblings observed in SOEP
Values:	(-1) No answer (-2) Does not apply (-3) Answer improbable
Description:	The variable contains the total number of identified siblings, including the respondent, in the SOEP (biological and non-biological).

10 BIOAGE01, BIOAGE03, BIOAGE06, BIOAGE08, BIOAGE10, BIOAGE12: Generated variables from the "Mother & Child", "Parent", and "Pupils" questionnaires

by David Richter[30]

10.1 Introduction

The bioage data files are generated using information collected in the "Mother & Child" and "Parent" questionnaires. In 2014, for the first time, the children themselves answered the "Pupils" questionnaire. In addition, the data from the Families in Germany (FiD) study was integrated into the SOEP in 2014 as well. There are six different "Mother & Child" and "Parent" questionnaires and one "Pupils" questionnaire (see Table 1). Two additional questionnaires (biage02 and bioage10b) were only surveyed in FiD, but are integrated into the *bioagel* data set too.

Table 9: Overview of bioage data, corresponding age group, and respondents

Bioage	Classification by Age	Respondents	First Wave
Bioage01	0-1 years old	Mothers only	2003
Bioage02 (FiD only)	1-2years old	Mothers only	2010
Bioage03	2-3 years old	Mothers only	2005
Bioage06	5-6 years old	Mothers only	2008
Bioage08a	7-8 years old	Parent 1	2010
Bioage08b	7-8 years old	Parent 2	2010
Bioage10	9-10 years old	Parent 1 (SOEP & FiD)	2012 / 2010 (FiD)
Bioage10b (FiD only)	9-10 years old	Parent 2 (FiD only)	2010
Bioage12	11-12 years old	Children	2014

Source: SOEP v32, doi: 10.5684/soep.v32

The "Mother & Child", "Parent" and "Pupils" questionnaires aim to follow and observe future generations of the SOEP and collect all information in age-specific files, even though the data come from different survey years. As we try to make this process as comprehensive and gap-free as possible, we begin following and documenting the development of children in SOEP

[30] This documentation is based on earlier versions of documentation materials on bioage data sets and has benefited from previous work by Sebastian Frischholz, Anne Fromm, Stefanie Lenuweit, Katharina Mahne, Christian Schmitt, and Jürgen Schupp.

households from birth onwards. Since many questions overlap, all bioage data files are covered in this chapter, rather than dedicating a separate chapter of this document to each bioage data file. By doing so, we are able to provide you with an overview of all variables covered in the bioage data files.

Starting with the v31 data distribution, all available data is now provided in one data set: *bioagel*. The separate bioage data files are no longer part of the data distribution.

10.2 Respondents in the 'Bioagel' Data Set

The values of the BIOAGE variable in the *bioagel* data file reflects the age of the children when the respective questionnaire was taken by their parents (e.g., *bioage* = 6 covers children who turned six that survey year, producing a range in ages between 5 years and 1 month and 6 years and 11 months, depending on the birth month and interview month). For information on the questionnaires used with each age group, please see Table 1. Except for bioage08b and bioage12 (and – FiD only – bioage10b), it is usually the mother who completes the questionnaire. In the exceptional case that a mother cannot complete the questionnaire, the father does. The respondents' personal ID numbers (regardless if fathers, mothers or other care takers) are provided in the variable PERSNRESP in the *bioagel* data set. Note that while the parents are the actual respondents (except for bioage12, covering the "Pupils" questionnaire), the data are organized under the child's unchanging personal ID number (PERSNR).

Bioage01: "Mother and Child" Questionnaire, Children Aged 0-1 years

The questionnaire is given to all women who gave birth to a child in the current or previous survey year, and to all women whose non-biological child was born in the same period. The questionnaire contains information on pregnancy, childbirth, childcare, and child health. In the exceptional case that a mother is unable to complete the questionnaire, the father responds.

Bioage02: "Your child between the ages of one and two", Children Aged 1-2 years (FiD only)

The questionnaire is given to all mothers whose child turns two in the current survey year. In the exceptional case that a mother cannot complete the questionnaire, the father responds.

Bioage03: "Your child between the ages of two and three", Children Aged 2-3 years

The questionnaire is given to all mothers whose child turns three in the current survey year. In the exceptional case that a mother cannot complete the questionnaire, the father responds.

Bioage06: "Your child between the ages of five and six", Children Aged 5-6 years

The questionnaire is given to all mothers whose child turns six in the current survey year. In the exceptional case that a mother cannot complete the questionnaire, the father responds.

Bioage08a and Bioage08b: "Parent" Questionnaire, children aged 7-8 years

The questionnaire is given to *both* parents of children turning eight in the current survey year. Data of parent 1, which is usually the mother, can be found in *bioage08a*. Data of parent 2, which is usually the father, can be found in *bioage08b*.

Bioage10: "Your child between the ages of nine and ten", Children Aged 9-10 years (only FiD: Bioage10b)

The questionnaire is given to all mothers (FiD: and fathers) whose child turns ten in the current survey year. In the exceptional case that a mother cannot complete the questionnaire, the father responds. Data of parent 1, which is usually the mother, can be found in *bioage10a*. Data of parent 2 (FiD only), which is usually the father, can be found in *bioage10b*.

Bioage12: "Pupils between the ages of eleven and twelve", Children Aged 11-12 years

The questionnaire is given to all children themselves who turn twelve in the current survey year.

Number of Children and Twins in the *Bioage* Data Sets

The data set has grown over the years and now contains data on 12,661 children. At the moment (2015), there are 139 children for whom information has been provided through all of the "Mother & Child", "Parent", and "Pupils" questionnaires (*bioage01*, *bioage03*, *bioage06*, *bioage08a/b, bioage10, and bioage12*). Data are available for 234 children from six questionnaires, 694 children from five questionnaires, 1,405 children from four questionnaires, for 2,697 children from three questionnaires, and for 3,140 children from two questionnaires. For 4,352 children data are available from one questionnaire. An overview of the number of children in each data set is given in Table 3.

Table 10: Number of respondents to the "Mother & Child", "Parent", and "Pupils" questionnaires

Bioage File	Survey Year													Total
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	
$01	318	247	246	234	205	185	196	1503	353	378	349	444	312	4970
$02 (FiD)								787	647	568	187			2189
$03			257	222	237	246	186	1061	914	745	701	454	386	5152
$06						237	210	687	696	612	858	825	657	4782
$08a								646	703	715	647	629	779	4119
$08b								409	490	500	439	399	556	2793
$10								403	510	699	686	656	616	3570
$10b (FiD)								242	310	291	301			1144
§12												606	608	1214
Total	318	247	503	456	442	668	592	5738	4623	4508	4168	4013	3914	30190

Source: SOEP v32, doi: 10.5684/soep.v32

Table 4 gives an overview of the number of children per family in the specific age range. Note that there might be more children in the household than indicated here: these children are not in the age range of the "Mother & Child", "Parent", and "Pupils" questionnaires and therefore not covered by the *bioagel* data set.

Table 11: Number of children in the family in the specific age range

Bioage File	Number of Children in the Family					
	1 Child	2 Children	3 Children	4 Children	5 Children	6 Children
Bioage01	2909	815	113	16	2	3
Bioage02 (FiD only)	1671	232	18			
Bioage03	3332	858	92	16	3	2
Bioage06	2922	811	71	5	1	
Bioage08a	2607	624	84	3		
Bioage08b	1772	418	59	2		
Bioage10	2460	469	56	1		
Bioage10b (FiD only)	785	169	7			
Bioage12	1082	63	2			

Source: SOEP v32, doi: 10.5684/soep.v32

The data also includes twins. *Bioage01* contains 109 pairs, *bioage02* 57 pairs, *bioage03* 106 pairs, *bioage06* 64 pairs, *bioage08a* 73 pairs, *bioage08b* 45 pairs, *bioage10* 67 pairs, *bioage10b* 24 pairs, and *bioage12* 19 pairs of twins.

10.3 Topics and Variables

The *bioage* data contain information regarding

- Pregnancy and childbirth
- Child health
- Childcare situation
- Changes in living circumstances since birth of child
- Child's abilities
- Parenting experiences
- Expectations about the child's success in school
- Educational goals and aspirations of the parents
- Educational behavior of the parents
- Self-conception of the role of parents

The rules used to generate the variables from the questionnaires are consistent over the various *bioage* questionnaires. In the new integrated *bioage long* data set, the data is presented in "long" format, i.e. this dataset contains information from *bioage01*, *bioage03*, *bioage06*, *bioage08a* and *bioage08b, bioage10* as well as *bioage12*. In addition, all variables from the FiD-Study are included as well: if there was a corresponding SOEP Variable, the data was combined into the SOEP variable; if there was no corresponding SOEP Variable or if the data was coded differently, the information is provided in separate variables with the suffix *_fid*.

A few variables are generated by combining the information from two or more variables. The present documentation provides detailed information on variables generated using information from other files. An overview of the specific variables and the rules by which they were generated is given in section 10.4 "Generated Variables".

10.4 Generated Variables

This section provides additional information on variables that have been generated from combinations of variables from other datasets than the parent questionnaires.

AGE

Variable label	**"Child's age in months"**
Variable format	2-digit integer
Comment:	This variable provides the child's age in months as a combination of month of birth and interview month. As the exact day of birth remains unknown, information is only an approximation and may vary by one month. Note that the information concerning year and month of birth from the "Mother & Child" and "Parent" questionnaires proved to be partially inconsistent. Therefore, as of the beginning of 2012 the child's age in months is computed using birth information from *ppfad*. For further information, refer to the documentation on *ppfad*.

PREGY

Variable label	**"Mother: pregnant at interview in survey year X"**
Variable format	4-digit integer
Bioage File	01
Comment	This variable is based on information from the previous year's individual questionnaire provided by the mother on her pregnancy status at the time of the interview. If pregnancy was reported (or was unknown) and a child was born, the year in which the interview took place is contained in BCPREGY. Hence, this information is available only for those women in the sample for at least two years with a completed individual interview in the first year and a completed "Mother & Child" questionnaire in the second year. Please note that some mothers are not aware that they are pregnant in the early stages of pregnancy. The time of observation starts in survey year 2003 with the 2002 birth cohort.

PREGMO

Variable label	**"Mother: pregnancy month at interview"**
Variable format	2-digit integer
Bioage File	01
Comment	This variable is based on the exact month of birth (BCPREGMO), the duration of childbearing in weeks (BCSSW) and the interview month of the previous year's personal interview. Hence, this information is available only for those women in the sample for at least two years. As the exact day of birth is unknown, this variable remains a close approximation.

PREBEG

Variable label	**"Spell Begin Pregnancy (Month, 01.83=1)"**
Variable format	3-digit integer
Bioage File	01
Comment	The variable BCPREBEG contains information on the beginning of pregnancy (i.e., the month of conception). Information is given in the regular SOEP spell format: values start with 1 for January 1983 (e.g., the earliest spell in *bioage01*, survey year 2010, is 304, which equals April 2008). The variable is based on the exact month of birth (BCPREGMO) and the duration of pregnancy in weeks (BCSSW). Accordingly, information is available only for women who completed the "Mother & Child Questionnaire" and for whom the duration of the pregnancy is known. Note that the month of conception may vary by one month as the exact date of birth remains unknown.

PREEND

Variable label	**"Spell End Pregnancy, Birth (Month, 01.83=1)"**
Variable format	3-digit integer
Bioage File	01
Comment	The variable BCPREEND contains information on the end of pregnancy (i.e. the month of birth). Information is given in the regular SOEP spell format: values start with 1 for January 1983 (e.g., earliest spell in *bioage01*, survey year 2010, is 304, which equals April 2008). This variable is based on the exact month of birth (BCPREGMO) and the duration of pregnancy in weeks (BCSSW).

SEX

Variable label	**"Gender of child"**
Variable format	1-digit integer
Comment	The sex of the child is not asked for in the "Mother & Child" and "Parent" questionnaires. Information on this variable stems from the *ppfad*.

SEXRESP

Variable label	**"Sex of respondent (parent)"**
Variable format	1-digit integer
Bioage File	08a, 08b
Comment	This variable tells whether the mother or father answered the respective questionnaire. The information for this variable comes from the *ppfad* file.

11 BIOAGE17: The Youth Questionnaire[1]

by Marco Giesselmann, Mila Staneva and Tabea Naujoks[2]

A special group of first time respondents are young persons living in a panel household, who reach the surveying age of 17 years. From this specific group of panel entrants, we are able to obtain some more detailed information on youth and socialisation than from other new sample members. At the same time, certain life-course dimensions (as the partnership- or employment biography) have not yet developed in 17 year-olds. With regard to these specifics, the standard biography questionnaire is not appropriate to this group. Thus, we use an independent questionnaire for this special group of first time respondents: the Youth Questionnaire. This instrument is used since the year 2000 and can be understood as an alternative version of the Biography Questionnaire, collecting more comprehensive information on relationships with parents, leisure-time activities, and past achievements in school, as well as on personality characteristics. In addition, there are numerous prospective questions about educational plans and plans for further training, as well as questions about expectations for future career and family.

A number of statements regarding specific circumstances—including the expectations for the future mentioned above—are directly related to the time at which the questionnaire was completed. However, they provide a multifaceted background for long-term analyses since these young people will continue to be interviewed in subsequent years like other SOEP respondents. The Youth Questionnaire also contains retrospective questions, for example, at what age the teenager started his or her first job or first music lessons, what recommendations he or she received regarding choice of secondary school level, and which grades he or she repeated.

11.1 Genesis and Target Population of the Youth Questionnaire

The Youth Questionnaire is aimed at youths who have reached the surveying age of 17 years[3] and are therefore being interviewed for the first time. This questionnaire takes the place of the supplementary Biography Questionnaire, since the latter does not apply to the young people's family or career situations. As a rule, information on social origin can be obtained from the parents' Individual Questionnaire, in case the youth lives together with the respective parent. If the teenager does not live with either parent, the Youth Questionnaire collects information on the missing parent(s). Young people who immigrated to Germany are also given the

[1] In earlier SOEP-data releases BIOAGE17 was called BIOYOUTH.
[2] Replaces earlier versions by Henning Lohmann and Sven Witzke, Jürgen Schupp, and Michaela Frühling, Thorsten Schneider, and Bettina Isengard.
[3] More precisely, this refers to youths who live in an already existing panel household and are or will turn 17 years old in the year of the survey. They are therefore 16 or 17 years old at the time of the interview.

standard questions on immigration from the supplementary Biography Questionnaire. This guarantees that all important information collected in the Biography Questionnaire is also available on these young people.

A preliminary version of the Youth Questionnaire was tested in 2000 in samples A-E on individuals born in 1983. An expanded and revised questionnaire entered the field one year later, in 2001, for all samples (A-F). In samples A-E, young people born in 1984 were surveyed, and in sample F, those born in the years 1982 to 1984. With the expansion of the number of birth cohorts, entries for the birth year 1983 are also collected for sample F (data previously existed only for samples A-E), which also creates a clear increase in the number of entries. In the following years, also the youths from additional samples have been interviewed. **In 2014, Data from the SOEP-related FID-study from the years 2010 to 2014 has been integrated in the SOEP (Sample L). Therefore, the number of cases in BIOAGE 17 has increased with SOEP Version 31 retrospectively for the years 2010 to 2013, compared with previous versions. You might want to use the psample-variable as filter, if you want to exclude these cases and reproduce your old sample.** For an overview of the target population in each survey year, see Table 1. In total, we have gathered interview data from 6,6641 analysable observations up to the present.

Table 1: Target Population for the Youth Questionnaire by year, sample and age

Suvey year	Sample									frequency
	A-E	F	G	H	I	J	K	L	M	
2000	17 yrs									232
2001	17 yrs	17-19 yrs								618
2002	17 yrs	17 yrs								352
2003	17 yrs	17 yrs	17 yrs							365
2004	17 yrs	17 yrs	17 yrs							373
2005	17 yrs	17 yrs	17 yrs							368
2006	17 yrs	17 yrs	17 yrs							307
2007	17 yrs	17 yrs	17 yrs	17 yrs						346
2008	17 yrs	17 yrs	17 yrs	17 yrs						261
2009	17 yrs	17 yrs	17 yrs	17 yrs						243
2010	17 yrs	17 yrs	17 yrs	17 yrs	17 yrs			17 yrs		404
2011	17 yrs	17 yrs	17 yrs	17 yrs		17 yrs		17 yrs		531
2012	17 yrs	17 yrs	17 yrs	17 yrs		17 yrs	17 yrs	17 yrs		537
2013	17 yrs	17 yrs	17 yrs	17 yrs		17 yrs	17 yrs	17 yrs		567
2014	17 yrs	17 yrs	17 yrs	17 yrs		17 yrs	17 yrs	17 yrs	17 yrs	577
2015	17 yrs	17 yrs	17 yrs	17 yrs		17 yrs	17 yrs	17 yrs	17 yrs	560

Status: up to wave BF (2015)

Source: SOEP v32, doi: 10.5684/soep.v32

In 2006, a new questionnaire on cognitive potential was introduced. Like the Youth Questionnaire, this instrument is aimed at youths who have reached the surveying age of 17 years. In order to keep the Interview at an acceptable length, the standard Individual Questionnaire is now left out for respondents of the Youth Questionnaire. The 2006 data on cognitive potentials was provided for secondary analysis in 2009 (dataset COGDJ).

11.2 Contents and Structure of the Data Set BIOAGE17

From a technical perspective, four different types of questions are asked in the Youth Questionnaire:

A) Questions used to complete certain biographical files (BIOIMMIG, BIOPAREN). These questions are identical to questions in the standard Biography Interview. This applies to the topic blocks 'Origin' (questions 60 to 71) and 'Childhood and Parents' House' (questions 72-87). The corresponding variables are *not* included in BIOAGE17, but combined with biographical information from non-youth new entrants in the files BIOIMMIG and BIOPAREN.

B) Questions that are similar to items in the standard Biography Interview, but go further into detail. This applies to the topic blocks 'Relationships' (questions 12-14), 'Free time and Sport' (questions 15-25) and 'Education and Career plans' (questions 26-55). These variables are stored in BIOAGE17. Corresponding Variables obtained from other new sample members (with a standard Biography Interview) are included in the dataset BIOSOC. Depending on the complexity and scope of the analysis, the user might want to combine corresponding data from BIOAGE17 and BIOSOC in order to access all panel members.

C) Questions that specifically relate to young persons and therefore have no equivalent in the standard Biography Interview. This applies to the topic blocks 'Residence' (questions 1-3), 'Jobs and Money' (questions 4-11), 'Future' (question 59) and 'Attitudes and Opinions' (questions 86-87). These Variables are stored in BIOAGE17 and have no equivalent for other panel entrants in BIOSOC.

D) Since 2006, selected time-variant questions from the unanswered regular individual questionnaire (which is not handed out to first time panel respondents from existing panel households, see 1.3) are added to the Youth Questionnaire. This refers to the questions 56 to 58, 91, 92 and the topic block 'personality' (questions 93 to 101). This data is *not* included in BIOAGE17[4], but stored in an additional dataset $PAGE17.

[4] The first ten items in question 92 are still stored in BIOAGE17, for details see 13.3.

The design of the dataset BIOAGE17 is patterned after the 2001 Youth Questionnaire, which is the standard version for subsequent years. As in the biographical data survey, every youth answers the Youth Questionnaire only once. The data is therefore presented in column form, just as it would be in a cross-sectional record. The variable SYEAR makes it possible to quickly identify the year of the survey. The entries to the questions that were only asked in 2000 and not in 2001 are not included in BIOAGE17. The complete dataset from 2000 is provided free of charge upon request. However, all entries from 2000 that are also included in 2001 are contained in BIOAGE17[5]! With the integration of questions from the Individual Questionnaire in 2006, some changes have occurred in the Youth Questionnaire, especially in the numbering of the questions.

Table 2 (at the end of this chapter) lists all of the variables for the dataset BIOAGE17. The first column contains the name of each variable, the second a brief specification of its content, and the third the number of the question as it appears in the Youth Questionnaire distributed in 2000, wave Q. The fourth column lists the corresponding questions in the Youth Questionnaires 2001 to 2005. In the last column, the question number from 2006 to 2015 youth questionnaires is noted. The variables containing the identification of the person surveyed and the interview situation have no corresponding number because they do not originate from the regular section of the Youth Questionnaire.

11.3 Special Features of Some Questions and Variables

The question regarding the support received by these young people from their parents (question 14) is based on the Supportive Parenting Scale of Simons et al. (1992)[6], which was transformed for Germany by Schwarz and Walper (1997)[7]. The instrument used to compile career orientation (question 54) was taken from Kracke (1996)[8].

Before 2006, problems arose with the question concerning school attendance (question 25 from 2001 to 2005) because of discrepancies between the information from the Youth Questionnaire and the information on the variable “type of general school attended” from the Individual Questionnaire. Since 2006, 17-year-olds no longer receive the Individual Questionnaire, so the question about school type has been integrated into the question on

[5] In the event that a question was asked in 2001 but not in 2000, the variable will have the value -3 for the persons who were surveyed in 2000.
[6] Simons, R.L., F.O. Lorenz, R.D. Conger and C.-I. Wu (1992): Support from spouse as mediator and moderator of the disruptive influence of economic strain on parenting. in: Child Development 63: 1282-1301.
[7] Schwarz, B. and S. Walper (1997): Erziehung aus Sicht von Eltern und Kindern. Erste Erfahrungen mit den Instrumenten der 1. Erhebung. Berichte aus der Arbeitsgruppe “Familienentwicklung nach der Trennung” #19/97. Ludwig-Maximilians-Universität München.
[8] Kracke, B. (1996): Fragebogen zur Berufsorientierung bei Realschülern. University of Mannheim, unpublished manuscript.

school attendance. For the previous years, the variable was generated using information from the Individual Questionnaire and questions 25 and 45[9] from the Youth Questionnaire.

If the question on school attendance in the Youth Questionnaire is answered with 'yes' when at the same time information from the regular Individual Questionnaire indicates that the person does not attend the general school system, or vice versa, a recoding is undertaken. In this case the variable BYSCHBES is changed to the value -3 (-3: Entry deleted after intensive examination). Another problem arises if a person states in the Youth Questionnaire that she attends school but does not specify school type in the Individual Questionnaire. In this case the variable BYSCHBES is given the value -1 (-1: no answer).

In question 51, young people are asked whether they know what career they would like to start. If they give a positive answer ('yes, with some certainty', 'yes, with a lot of certainty'), then they are asked to specify the occupation in plain text. This plain-text entry is coded according to the classification of occupations of the Federal Statistical Office, Germany, (Statistisches Bundesamt), version 1992, and according to the ISCO 1988. In addition, the values for Ganzeboom's International Socio-Economic Index of Occupational Status (ISEI), for Treiman's Standard International Occupational Prestige Scale (SIOPS) for Erikson's and Goldthorpe's Class Category (EGP)[10] as well as Wegener's Magnitude Prestige Scale (MPS)[11] are also given.

Since 2005 some respondents have a value of –3 in variables BYMUSART, BYMUSMW and BYSPRTMW. This means that they gave more than one answer to the question although only one answer was possible. Because of this, it was not possible to assign a single valid answer.

By extending the questions about personality, we meanwhile ask the questions regarding attitudes about life and the future on a seven-point scale instead of the four-point scale we started with in the earlier version of this battery. From 2006, the variables BYESVERL to BYESENGA are stored with the values 1 (no acceptance) to 7 (total acceptance) and with the values 11 (total acceptance) to 14 (no acceptance) for respondents of previous years. Thus, the normative decision on how to integrate these two scales is up to the user.

[9] For 2000 questions 24 and 45.
[10] For ISCO 88, SIOPS, ISEI and EGP see Ganzeboom, H.B.G. and D.J. Treiman (1996): Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. in: Social Science Research 25, 201-239.
[11] Frietsch, R. and H. Wirth (2001): Die Übertragung der Magnitude-Prestigeskala von Wegener auf die Klassifizierung der Berufe. in: ZUMA-Nachrichten, 48, 139-163.

Table 2: Description of the data set BIOAGE17

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[1][2]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
	Entries for surveyed person			
HHNR	Original household identifier (invariant)			
HHNRAKT	Actual household identifier			
PERSNR	Personal identifier			
BEFRPER	Respondent identifier			
SYEAR	Survey year			
SEX	Sex			
PSAMPLE	Subsample			
BYGEBJAH	Year of birth			
BYMNR	identifier of mother (taken form BIOPAREN social, not necessarily biological relationship)			
BYVNR	identifier of father (taken from BIOPAREN social, not necessarily biological relationship)			
	Residence			
BYWOELT	Residing in parents' household (HH)	01	01	01
BYWOZIM	Own room	02	02	02
BYWOWEI	Additional apartment outside of parents' HH	04	03	03
	Jobs and Money			
BYVDEIG	Own income	09	04	04
BYVDART	Type of income	10	05	05
BYJBFRUE	Worked before (on holiday or while in school)	13	06	06
BYJBALT	Age by first job (on holiday or while in school)	14	07	07
BYJBGRUN	Reason for working	-	08	08
BYTGELD	Allowance	15	09	09
BYTGELDW	Amount of allowance per week	16	10	10
BYTGELDM	Amount of allowance per month	16	10	10
BYSPAR	Saving money	17	11	11
BYSPARM	Amount saved every month	17	11	11
BYSPARUN	Sporadic saving	17	11	11
	Relationships			
	Importance of various persons:			
BYWIVA	Father	-	12	12
BYWIMU	Mother	-	12	12
BYWIBS	Brother, Sister	-	12	12
BYWIVW	Other related persons	-	12	12
BYWIFFR	Serious boy/girlfriend	-	12	12
BYWIBFR	Best friend	-	12	12
BYWILEHR	Teacher	-	12	12
BYWICLQ	Clique	-	12	12
BYWISON	Other person	-	12	12

[1] If no corresponding question/variable exists, it is assigned a minus sign; numbers/names in parentheses mean that there is no identical question/variable but a corresponding one.

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[12]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
	Frequency of fights with:			
BYSTRVA	Father	-	13	13
BYSTRMU	Mother	-	13	13
BYSTRBS	Brother, Sister	-	13	13
BYSTRFFR	Serious boy/girlfriend	-	13	13
BYSTRBFR	Best friend	-	13	13
BYBZ01MU	Talk with mother about personal experiences	-	14	14
BYBZ01VA	Talk with father about personal experiences	-	14	14
BYBZ02MU	Mother addresses problems	-	14	14
BYBZ02VA	Father addresses problems	-	14	14
BYBZ03MU	Mother asks opinion before a decision is made	-	14	14
BYBZ03VA	Father asks opinion before a decision is made	-	14	14
BYBZ04MU	Mother shows approval	-	14	14
BYBZ04VA	Father shows approval	-	14	14
BYBZ05MU	Solve problems together with mother	-	14	14
BYBZ05VA	Solve problems together with father	-	14	14
BYBZ06MU	Mother shows trust	-	14	14
BYBZ06VA	Father shows trust	-	14	14
BYBZ07MU	Mother asks opinion on family issues	-	14	14
BYBZ07VA	Father asks opinion on family issues	-	14	14
BYBZ08MU	Mother justifies decision	-	14	14
BYBZ08VA	Father justifies decision	-	14	14
BYBZ09MU	Mother shows love	-	14	14
BYBZ09VA	Father shows love	-	14	14
	Free time and Sport			
	Frequency of free time activities:			
BYFZFERN	TV, Video	-	15	15
BYFZPC	Computer games	-	15	15
BYFZMUSH	Listen to music	-	15	15
BYFZMUSS	Play music	-	15	15
BYFZSPRT	Do sports	-	15	15
BYFZTANZ	Dance, Theatre	-	15	15
BYFZTECH	Technical work, Programming	-	15	15
BYFZLESE	Read	-	15	15
BYFZEHRE	Volunteer activities	-	15	15
BYFZABH	Do nothing, hang around, day dream	-	15	15
BYFZMFFR	Spend time with boy/girlfriend	-	15	15
BYFZMBFR	Spend time with best friend	-	15	15
BYFZMCLQ	Spend time with clique	-	15	15
BYFZINT	Internet/chatting (2006-2013)	-	-	15
BYFZSINT	Frequency surfingin the internet (since2014)			15
BYFZSONW	Social online networks (since 2013)			15
BYFZJUGZ	visiting youth center	-	-	15
BYFZRELI	go to church/religious activities	-	-	15
BYMUSSP	Actively make music	-	16	16
BYMUSART	Style of music made	-	17	17
BYMUSMW	Play music with whom	-	17a	18

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[12]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
BYMUSALT	Age starting playing music	-	18	19
BYMUSUNT	Paid music lessons	-	19	20
BYSPRTTR	Participate in sports	20	20	21
BYSPRTAR	Favourite sport	21	21	22
BYSPRTAL	Age started favourite sport	23b	22	23
BYSPRTMW	Where and with whom favourite sport	23c	23	24
BYSPRTWE	Participation in competitions	23d	24	25
	Education and Career Plans			
BYSCHBES[3]	School attendance	24 / PF: 09	25 / PF[4]	26
BYKLASS	Current school year (since 2014)			26
BYSCHEND	Last year of school	25	26a	27
BYSCHABS	Type of school certificate	26	26b	28
BYSCHZUK	Strive for further school certificate	27	27	29
BYSCHZAR	Type of further school certificate	28	28	30
BYFMD1[5]	1. foreign language	32	29	31
BYFMD2	2. foreign language	32	29	31
BYSCHAUS	School attendance in foreign country	29	30	32
BYAUSAB	School abroad age begin (since 2014)			32
BYAUSAE	School abroad age end (since 2014)			32
BYSCHPRI	Attendance in a private school	-	31	33
	Activities in school:			
BYENKSPR	Class representative	34	32	34
BYENSSPR	School representative	34	32	34
BYENSZTG	School newspaper	34	32	34
BYENTHEA	Theatre, Dance group	34	32	34
BYENCHOR	Choir, Music	34	32	34
BYENSPRT	Sport group	34	32	34
BYENSONS	Other groups	34	32	34
BYENNEIN	No activities	34	32	34
BYZFINSG	Satisfaction with effort at school (overall)	31	33	35
BYZFDEUT	Satisfaction with effort in German	31	33	35
BYZFMATH	Satisfaction with effort in math	31	33	35
BYZFFMD1	Satisfaction with effort in 1. foreign language	31	33	35
BYEMPFEH	Recommendation after elementary school	-	34	36
BYNTDEUT	Last grade[6] in German	33	35	37
BYNTMATH	Last grade in math	33	35	37

[3] As mentioned in 13.3, for the years 2000 to 2005 the variable BYSCHBES is generated in consideration of information stemming from the personal questionnaire.
[4] The relevant question from the personal questionnaire differs from year to year: for 2001 question 11, for 2002 question 14, for 2003 question 33, for 2004 question 08, and for 2005 question 09.
[5] Additional category since 2006: value 7 “Spanish”.
[6] Students normally receive grades ranging from 1 to 6, whereby 1 is the best and 6 the worst. This system of assigning grades is used up to the 11th or 12th grade (level II of upper secondary or comprehensive school) depending on the federal state. After that, a new grading system is used. To make the data set more user-friendly, the information given for school grades and the information on points transformed into grades is stored in this variable. Note: No corrections have been made when a person has reported both grades and point scores and when the two types of information do not correctly correspond.

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[12]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
BYNTFMD1	Last grade in 1. foreign language	33	35	37
BYPTDEUT	Total points[7] in German	33	35	37
BYPTMATH	Total points in math	33	35	37
BYPTFMD1	Total points in 1. foreign language	33	35	37
BYGSDEUT	Level of German at comprehensive school[8]	33	35	37
BYGSMATH	Level of math at comprehensive school	33	35	37
BYLKDEUT	Complementary / main subject[9] in German	33	35	37
BYLKMATH	Complementary / main subject in math	33	35	37
BYLKFMD1	Complementary / main subject in 1. foreign language	33	35	37
BYKLWDJA	Class repeated	35	36	38
BYKLWD1	Class level 1. repeated	36	37	39
BYKLWD2	Class level 2. repeated	36	37	39
BYNACHHI	Paid tutor lessons	37	38	40
BYELKUEM	Parents care about efforts at school	39	39	41
BYELHAUS	Parents help with homework	40	40	42
BYELDIFF	Problems with parents because of effort at school	41	41	43
BYELABEN	Parents attend parents' evening	42	42	44
BYELSPRE	Parents go to parents' day	42	42	44
BYELLEHR	Parents go to see a teacher	42	42	44
BYELVERT	Active as parent representative	42	42	44
BYELNIDA	Parents do not participate in any of these activities	42	42	44
BYKLAUSL	Number of foreign classmates	(43)	43	45
BYBAABGE	Vocational education, Internship, training	44	44	46
BYBABGJ	Vocational introductory year ("Berufsgrundschul- / Berufsvorbereitungsjahr")	45	45	47
BYBABEGL	Vocational integration training ("Berufl. Eingliederungslehrgaenge")	45	45	47
BYBALEH	Vocational education, apprenticeship ("Berufsausbildung, Lehre")	45	45	47
BYBABFS	Full-time vocational school/ School for public health ("Berufsfachschule / Schule des Gesundheitswesen")	45	45	47
BYBAPRAK	Internship ("Praktikum, Voluntariat")	45	45	47
BYB???	Duales Studium (since 2014)			47
BYBAUNI	University of Applied Science/University (since			47

[7] From the 11th or 12th grade on, pupils are awarded points in upper secondary or comprehensive school ranging from 0 to 15, whereby 15 points are the best, 0 points the worst. The link between points and grades is as follows: 0 points: 6; grade of 1 to 3 points: grade of 5; 4 to 6 points: grade of 4; 7 to 9 points: grade of 3; 10 to 12 points: grade of 2; 13 to 15 points: grade of 1.

[8] The subjects German, math and the first foreign language are split up into different levels during the secondary school level I in comprehensive schools. Level A is the highest. The number of levels differs between the federal states.

[9] From the 11th or 12th grade on, pupils can choose their main subjects. At this stage, German, math and foreign languages can be downgraded from major to minor subjects.

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[12]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
	2013)			
BYZAJA	Vocational / university degree is aspired	46	46	48
	Type of aspired vocational / university degree:			
BYZALEH	Apprenticeship (“Lehre”)	47	47	49
BYZABFS	Full-time vocational school/ School for public health (“Berufsfachsule / Schule des Gesundheitswesen”)	47	47	49
BYZAFSC	Technical school, school for master of a trade (“Fachschule, Meister-, Technikerschule”)	47	47	49
BYZABEA	Training for civil servants (officer) (“Beamtenausbildung”)	47	47	49
BYZABAK	Approved vocational academy (“anerkannte Berufsakademie”)	47	47	49
BYZAFH	Advanced technical college (“Fachhochschule”)	47	47	49
BYZAUNI	University	47	47	49
BYSLBALT	Desired age for financial independence	48	48	50
BYSLBHEU	Already financially independent	48	48	50
BYBWUNJA	Occupation is aspired	49	49	51
	Occupation categories, encoded:			
BYKLAS	Classification of career according to the Federal Statistical Office, Germany, (Statistisches Bundesamt), version 1992	50	50	52
BYISCO88	International Standard Classification of Occupation 1988 (ISCO88)	50	50	52
BYEGP	Erikson and Goldthorpe’s Class Category (EGP)	50	50	52
BYISEI	International Socio-Economic Index of Occupational Status after Ganzeboom (ISEI)	50	50	52
BYSIOPS	Treiman’s Standard International Occupational Prestige Scale (SIOPS)	50	50	52
BYMPS	Magnitude Prestige Scale after Wegener (MPS)	50	50	52
BYZBINF	Information level of planned career	-	51	53
BYZBELT	Influence of the parents on career choice	-	52	54
BYZBLAS	No specific career in mind	-	52	54
BYZBBES	Intensive thoughts about various careers	-	52	54
BYZBRAU	Still looking for a career	-	52	54
	Important aspects for the career choice:			
BYWBSICH	Secure job	51	53	55
BYWBEINK	High income	51	53	55
BYWBAUF	Promotion opportunities	51	53	55
BYWBANE	Established profession	51	53	55
BYWBFREI	Enough free time	51	53	55
BYWBINT	Interesting activities	51	53	55
BYWBSELB	Working independently	51	53	55
BYWBKONT	Contact with persons	51	53	55
BYWBGSL	Relevant to society	51	53	55
BYWBGSND	Healthy conditions at work	51	53	55
BYWBFAM	Flexibility for family	51	53	55

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[12]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
BYWBHELF	Help others	51	53	55
	Future			
	Probability of future career related and private events:			
BYWAAUSP	To be accepted for a desired apprenticeship / place at university	52	54	59
BYWAERFA	To complete training/ university successfully	52	54	59
BYWAARBP	Job in desired career	52	54	59
BYWABERF	Job-related success	52	54	59
BYWAARBL	Longer unemployment	52	54	59
BYWAZURU	From family related reasons held back in career	52	54	59
BYWASELB	Self-employed	52	54	59
BYWAAUSL	Work in foreign country	52	54	59
BYWAHEIR	To marry	52	54	59
BYWAPART	Live together with partner (not married)	52	54	59
BYWAKID1	Have one child	52	54	59
BYWAKIDM	Have two or more children	52	54	59
	Attitudes and Opinions			
BYGLPART	Happiness: live with/without partner	82	79	86[10]
BYGLKIND	Happiness: with/without children	-	80	87[11]
	Success in FRG from			
BYEFFLEI	Studiousness	86	81	88[12]
BYEFAUSN	Exploitation of others	86	81	88
BYEFINT	Intelligence	86	81	88
BYEFFAM	Family's origin	86	81	88
BYEFFACH	Technical know-how	86	81	88
BYEFGELD	Money	86	81	88
BYEFSABS	School education	86	81	88
BYEFHART	Being inconsiderate and hard	86	81	88
BYEFBEZ	Networking	86	81	88
BYEFPOLI	Political activities	86	81	88
BYEFMANN	Sex/ 'being a man'	86	81	88
BYEFINI	Being dynamic and taking initiative	86	81	88
BYESVERL	What happens in life, depends on me	-	(82)	90[13]
BYESERRE	Did not reach, what I deserve	-	(82)	90
BYESGLUE	What you achieve, is a matter of luck	-	(82)	90
BYESAND	Others decide about my life	-	(82)	90
BYESHART	You have to work hard for success	-	(82)	90
BYESZWEI	By difficulties, doubt about own abilities	-	(82)	90
BYESSOZU	Chances are determined by social circumstances	-	(82)	90
BYESFAEH	Abilities are more important than efforts	-	(82)	90
BYESKNTR	Little control over events in my life	-	(82)	90

[10] Question 88 in 2012-2015
[11] Question 89 in 2012-2015
[12] Question 90 in 2012-2015
[13] Question 92 in 2012-2015

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire 2000[12]	Number of Question in Youth Questionnaire 2001-2005	Number of Question in Youth Questionnaire 2006 - 2015
BYESENGA	Change of social circumstances through social/political activities	-	(82)	90
	Specification of Interview Situation			
BYINTA	Type of interview			
BYDAUER1	Duration of personal interview			
BYDAUER2	Duration of interview filled out independently			
BYANW	Presence of other persons			
BYTAGIN	Day of the interview			
BYMONIN	Month of the interview			
INTID	Identifier of the interviewer			

Source: SOEP v32, doi: 10.5684/soep.v32

12 BIOSOC: Retrospective Data on Youth and Socialization

by Marco Giesselmann and Mila Staneva[1]

The standard supplementary Biography Questionnaire was expanded in 2000, and again in 2001 to include some specific questions on youth and early adulthood. Some of these questions are derived from the independent Youth Questionnaire (for detailed information on this questionnaire, see chapter 13). The expanded questionnaire asks respondents of all ages to describe aspects of their life at the age of 15, including their relationship with parents, grades in school, the federal state where they last attained educational qualifications, detailed information on vocational qualifications, as well as intentions to complete further education or vocational training (the latter questions were relevant mainly to younger respondents). Questions concerning military and alternative services are also included in this data set.

As these questions are a part of the standard Biography Questionnaire, they are only asked once. Some of these questions can, however, be followed up by the regular data collected in the Individual Questionnaire. For example, if someone was too young to have completed his military service when the Biography Questionnaire was conducted, the user can look at the data set ARTKALEN in later years, where labour force participation is recorded on a monthly basis. Here one can find out if somebody was doing military service at the time or not.

survey year	frequency
2000	246
2001	8,819
2002	552
2003	2,328
2004	450
2005	299
2006	223
2007	2,232
2008	336
2009	196
2010	2,038
2011	14,797
2012	2,855
2013	5,370
2014	552
2015	2,129
Total	43,422

sample	frequency
A: Germans (west)	930
B: Foreigners (west)	263
C: Germans (east)	634
D: Immigrants 1984- 1993	132
E: Supplement 1998	257
F: Innovation 2000	9.844
G: High-Income 2002	2.262
H: Supplement 2006	2.198
I: Incentivation 2009	1,870
J: Supplement 2011	5,409
K: Supplement 2012	2.579
L: Family Samples 11/12	9.956
M: Migrants 13/14	7,088
Status: up to wave BF (2015)	

[1] Replaces earlier versions by Henning Lohmann and Sven Witzke Jürgen Schupp and Michael Frühling Bettina Isengard and Thorsten Schneider.

The data set BIOSOC contains information on 43,422 persons, of whom 8,819 stem from the year 2001. The reason for this is that the Biography Questionnaire was directed to sample F, as this was its second survey year. Consequently, the majority of the persons in this data set belong to sample F (27%) or the samples which were included in more recent years (sample G, H, I, J, K and M).

In 2014, Data from the SOEP-related FID-study from the years 2010 to 2014 has been integrated in the SOEP (Sample L). Therefore, the number of cases in BIOAGE 17 has increased with SOEP Version 31 retrospectively for the years 2011 to 2013, compared with previous versions. You might want to use the psample-variable as filter, if you want to exclude these cases and reproduce your old sample.

12.1 Structure of the Data Set BIOSOC

Respondents are given the Biography Questionnaire only once in a lifetime. Some of the information stored in the new data set BIOSOC is invariant (such as the relationship to parents at the age of 15) or is not surveyed to such an extent in the regular questionnaire (such as last school grades). Consequently there is only one record for each person and updates are not intended for this data set. The variable SYEARmakes it possible to quickly identify the year of the survey. Using the variable BSGEBJAH, which contains the year of birth, the user can determine the respondent's age. If the respondent is of a certain age, one can assume that some of the variables are constant. This applies to variables such as last school grades or military service.

In Table 3 (at the end of the chapter) all variables of the data set BIOSOC are listed. The first column contains the name of the variable, the second a brief specification of its content. The third column contains the number of the question as it appears in the Biography Questionnaire of wave Q (2000). Here, a minus sign means that the variable is not available in a given year and a question number in parenthesis indicates limited comparability. The fourth column contains the number of question in waves R to BF (2001-2015). As one can see, all listed questions were asked in the year 2001. In the last column the corresponding variable in the BIOAGE17 dataset is given, if available.

12.2 Special Features of Some Questions and Variables

The interviewees were asked if they did sports in their youth. If they answered in the affirmative, they were asked to include the sport they participated in most. This information was re-coded to a numeric variable and categorised. Some categories could easily be coded, such as soccer, whereas for others this was not possible. For users interested in specific

research questions on sports in youth, the original plain text answers can be provided upon request.

Individuals were asked: "When was the last year you attended school?" (Question 46). If they were still attending school, they had the opportunity to report that they were students. Unfortunately in the years 2001 to 2003, individuals who reported being students or who did not provide any answer to this question skipped over numerous questions due to the questionnaire design. Consequently, for these individuals there is no information on the number of foreign classmates[2], on the school degree aspired to, or planned vocational qualifications. Their record also lacks information on past vocational qualifications. However, this should prove less problematic since most of the students were enrolled in regular school programs throughout the entire time.

In 2011, the Biography Questionnaire was integrated in the Individual Questionnaire for the samples J and K. For these samples a new variable was introduced – BSDAUER3 – which shows the interview duration with the integrated questionnaire version. All other samples are coded with "-6" on this variable.

In 2012, sample M containing only respondents with migration background was added in SOEP ***(SOEP Survey Paper 216)***. This sample was surveyed with a modified version of the Biography Questionnaire which focuses on the immigration history and therefore misses some of the standard questions from the original Biography Questionnaire ***(SOEP Survey Paper 218)***. The last column in Table 3 gives an overview of the variables in BIOSOC available for the migration sample. A minus signs means that the variable is not available for sample M (in the Data, we used the missing code "-5" in such cases). A question number in parenthesis indicates that the question in the Migration Questionnaire is comparable, but not identical to the question in the original Biography Questionnaire.

[2] Most persons with no valid information on foreign classmates are not students but individuals who finished school abroad. This is because the question only targets those in German schools.

Table 3: Description of the data set BIOSOC

Variable Name	Content of the Variable	Number of Question in Biography Questionnaire…						Comparable Variable in BIOAGE17	Comparable Question in Migration Questionnaire 2013
		2000	2001-2010	2011-2012	2013	2014	2015		
	Entries for Surveyed Person								
HHNR	Original household identifier (invariant)							HHNR	
HHNRAKT	Actual household identifier							HHNRAKT	
PERSNR	Personal identifier							PERSNR	
SYEAR	Survey Year								
PSAMPLE	Subsurvey								
SEX	Sex								
BEFRPER	Respondent identifier							BEFRPER	
ERHEBJ	Survey year							ERHEBJ	
BSGEBJAH	Year of birth							BYGEBJAH	
	School								
BSELKUEM	Parents took care about efforts at school	-	29	30	32	36	38	BYELKUEM	-
BSNTDEUT	Last grade in German	-	30	31	33	37	39	BYNTDEUT	-
BSNTMATH	Last grade in maths	-	30	31	33	37	39	BYNTMATH	-
BSNTFMD1	Last grade in 1. foreign language	-	30	31	33	37	39	BYNTFMD1	-
BSPTDEUT	Total points[1] in German (last class)	-	30[2]					BYPTDEUT	-
BSPTMATH	Total points in maths (last class)	-	30[2]					BYPTMATH	-

Variable Name	Content of the Variable	Number of Question in Biography Questionnaire…						Comparable Variable in BIOAGE17	Comparable Question in Migration Questionnaire 2013
		2000	2001-2010	2011-2012	2013	2014	2015		
BSPTFMD1	Total points in 1. foreign language (last class)	-	30[2]					BYPTFMD1	-
BSGSDEUT	Level of German at comprehensive school[1] (last class)	-	30[2]					BYGSDEUT	-
BSGSMATH	Level of maths at comprehensive school (last class)	-	30[2]					BYGSMATH	-
BSGSFMD1	Level of 1. foreign language at comprehensive school (last class)	-	30[2]					BYGSFMD1	-
BSLKDEUT	Complementary / main subject[4] in German (last class)	-	30[2]					BYLKDEUT	-
BSLKMATH	Complementary / main subject in maths (last class)	-	30[2]					BYLKMATH	-
BSLKFMD1	Complementary / main subject in 1. Foreign language (last class)	-	30[2]					BYLKFMD1	-
	Relationships to Parents, Sport and Activities during Youth								
	Frequency of fights when respondent was 15. years old with:								
BSSTRVA	Father	-	31	32	34	38	40	BYSTRVA	-
BSSTRMU	Mother	-	31	32	34	38	40	BYSTRMU	-
BSSPRTTR	Participated in sports during youth	-	32/33	34	36	40	42	BYSPRTTR	57
BSSPRTAR	Favourite sport during youth	-	33/34	35	37	41	43	BYSPRTAR	-
BSSPRTWE	Participated in competitions during youth	-	34/35	36	38	42	44	BYSPRTWE	-
BSMUSSP	Played music or sang during youth	-	35/32	33	35	39	41	BYMUSSP	56

Variable Name	Content of the Variable	Number of Question in Biography Questionnaire…						Comparable Variable in BIOAGE17	Comparable Question in Migration Questionnaire 2013
		2000	2001-2010	2011-2012	2013	2014	2015		
	School Attendance								
BSSCHBES	Still at school	(34)	37	38	40	44	46	BYSCHBES	69
BSSCHEND	Year left school	-	37	38	40	44	46	BYSCHEND	69
BSSCHWO	Country of last school attendance	(34)	38	39	41	45	47	-	70
BSSCHLA	Federal State of last school attendance	-	41	42	44	48	50	-	73
BSKLAUSL	Number of foreign classmates	(37)	43	44	46	50	52	BYKLAUSL	75
BSSCHZUK	Strive for further school certificate	35	44	45	47	51	53	BYSCHZUK	76
BSSCHZAR	Type of further school certificate	36	45	46	48	52	54	BYSCHZAR	77
	Attained and Planed Vocational Qualification								
BSBADABG	Vocational / university degree acquired in Germany	38	46	47	49	53	55	(BYBAABGE)	80
	Type of vocational / university degree attained in Germany:								
BSBADLEH	Apprenticeship (“Lehre”)	39	47	48	50	53	55	(BYBABFS)	81
BSBADBFS	Full-time vocational school / School for public health (“Berufsfachsule / Schule des Gesundheitswesens”)	39	47	48	50	54	56	(BYBABFS)	81
BSBADFSC	Technical school, school for master of a trade (“Fachschule, Meister-, Technikerschule”)	39	47	48	50	54	56	-	81

Variable Name	Content of the Variable	Number of Question in Biography Questionnaire…						Comparable Variable in BIOAGE17	Comparable Question in Migration Questionnaire 2013
		2000	2001-2010	2011-2012	2013	2014	2015		
BSBADBEA	Training for civil servants (officer) ("Beamtenausbildung")	39	47	48	50	54	56	-	81
BSBADFHA	Advanced technical college ("Fachhochschule") or approved vocational academy ("anerkannte Berufsakademie")	39	47	48	50	54	56	-	81
BSBADUNI	University degree	39	47	48	50	54	56	-	81
BSBADPRO	Doctorate ("Promotion")	-		48[5]	50	54	56	-	81
BSBADSON	Other vocational qualification	39	47	48	50	54	56	(BYBABGJ, BYBABEGL)	81
BSBADEND	Year of attaining vocational / university degree in Germany	-	48	48	50	54	56	-	81
BSBAAABG	Vocational / university degree acquired abroad	40	49	49	51	55	57	-	82
	Type of vocational / university degree attained abroad:								
BSBAAFAN	Short-term training in a company	41	50		52	56		-	(83)
BSBAAFBA	Apprenticeship in a company	41	50		52	56		-	(83)
BSBAASCH	Vocational or professional school	41	50		52	56	58	-	(83)
BSBAAUNI	University degree	41	50		52	56	58	-	(83)
BSBAASON	Other vocational qualification	41	50		52	56	58	-	(83)
BSBAAEND	Year of attaining vocational / university degree abroad	-	51		53	57	59	-	(83)
BSBAAZEU	Certificate for abroad attained qualification	42	52		54	58	60	-	(84)
BSBAAABA	Applied for recognition					59	61		
BSBAAARA	Result recognition					59	61		
BSBAAAGA	Reason no recognition					59	61		
BSBAAAMB	Month of assessment					59	61		

Variable Name	Content of the Variable	Number of Question in Biography Questionnaire…						Comparable Variable in BIOAGE17	Comparable Question in Migration Questionnaire 2013
		2000	2001-2010	2011-2012	2013	2014	2015		
BSBAAAJB	Year of assessment					59	61		
BSBAAZEA	Recognition of aboard attained certificate	42	52		54	59	61	-	(85 to 89)
BSZAJA	Vocational / university degree is aspired	43	53		55	60	62	BYZAJA	95
	Type of aspired vocational / university degree:								
BSZALEH	Apprenticeship (“Lehre”)	44	54		56	60	62	BYZALEH	96
BSZABFS	Full-time vocational school/ School for public health (“Berufsfachschule / Schule des Gesundheitswesens”)	44	54		56	60	62	BYZABFS	96
BSZAFSC	Technical school, school for master of a trade (“Fachschule, Meister-, Technikerschule”)	44	54		56	60	62	BYZAFSC	96
BSZABEA	Training for civil servants (officer) (“Beamtenausbildung”)	44	54		56	60	62	BYZABEA	96
BSZABAK	Approved vocational academy (“anerkannte Berufsakademie”)	44	54		56	60	62	BYZABAK	96
BSZAFH	Advanced technical college (“Fachhochschule”)	44	54		56	60	62	BYZAFH	96
BSZAUNI	University degree	44	54		56	60	62	BYZAUNI	96
	Military and Voluntary Service								
BSDIGEL	Military or alternative service done (only men)	58	71/74	74	79	78	80	-	-

Variable Name	Content of the Variable	Number of Question in Biography Questionnaire…						Comparable Variable in BIOAGE17	Comparable Question in Migration Questionnaire 2013
		2000	2001-2010	2011-2012	2013	2014	2015		
BSDIART	Type of service (only men)	58	71/74	74	79	78	80	-	-
BSDIGRU	Reason for not serving (only men)	58	71/74	74	79	78	80	-	-
BSFSJ	Voluntary social service (“Freiwilliges Soziales Jahr”)	-	72/73	73	78	77	79	-	-
BSBFD	Federal voluntary service ("Bundesfreiwilligendienst")	-		3	78	77	79	-	-
BSFWD	Voluntary military service ("Freiwilliger Wehrdienst")	-		73	78	77	79	-	-
	Specification of Interview Situation								
BSINTA	Type of interview							BYINTA	
BSDAUER1	Duration of personal interview							BYDAUER1	
BSDAUER2	Duration of interview filled out independently							BYDAUER2	
BSDAUER3	Interview duration with the integrated questionnaire version							-	
BSTAGIN	Day of the interview							BYTAGIN	
BSMONIN	Month of the interview							BYMONIN	
INTID	Identifier of the interviewer							INTID	

[1] To make the data set more user-friendly, the information given on points are transformed into grades and stored in the corresponding variable. The link between points and grades is as follows: 0 points: grade of 6 1 to 3 points: grade of 5 4 to 6 points: grade of 4 7 to 9 points: grade of 3 10 to 12 points: grade of 2 13 to 15 points: grade of 1.
[2] Only in survey year 2001 (wave R)
[3] The subjects German, math and the first foreign language are split up into different levels during the secondary school level I of comprehensive schools. Level A is the highest one.
[4] From the 11th or 12th grade on students can choose their main subjects. At this stage, they can reduce German, maths and foreign languages from major to minor subjects.
[5] This item was integrated in the Biography Questionnaire in 2011.

Source: SOEP v32, doi: 10.5684/soep.v32

13 BIOPAREN: Biography Information for the Parents of SOEP-Respondents

by Anne Fromm, Sebastian Frischholz, Josephine Kraft and Daniel D. Schnitzlein[1]

13.1 Short summary

The aim of the data file BIOPAREN is to make the biography entries on the parents and on the social origin of the respondent available.

13.2 How biography information has been collected in the SOEP

In the third wave (1986) intergenerational aspects of the persons surveyed were included for the first time by means of a special group of questions in the Individual Questionnaire. These dealt with statements made about the education or professional training of the parents, the parents' residency, and their year of birth and death. For Sample B, only the education, residency, year of birth and death of the parents were asked. In 1988 the complete collection of biography questions (history of labor force participation, marriage and family biography, career start, and social origin) was included in the Individual Questionnaire for individuals surveyed for the first time. At the same time, a follow-up survey was given to those participants who had not yet received all or part of this collection of questions. This survey was continued in this form each of the following years until 1991, when the separate Biography Questionnaire was introduced. Since 1994, the biography was collected using the Personal History Questionnaire ('Lebenslauf-Fragebogen'), a slightly modified version of the Biography Questionnaire.

The Biography was included in Sample C in the third survey wave, that is, in 1992. The biographies of the persons in Sample D1 and D2 were collected during the first survey in 1994 and 1995. In 1999 the biography was collected for Sample E. In 2001 the follow-up survey was completed for Sample F and was followed by Sample G (High-Income) in 2003. The retrospective data of the sample H was collected in 2007. In 2009 the new subsample I

[1] *This documentation is based on earlier versions of the BIOPAREN documentation and has benefited from previous work by Charlotte Büchne, Stefanie Lenuweit, Katharina Mahne, Matthias Pollmann-Schult, Jürgen Schupp and Verena Tobsch.*

with valid interview data on about 2.500 adults was introduced in the SOEP. These new respondents did not fill in the biography questionnaire in order to reduce response burden in the first wave. Their data have been integrated in their second wave (2010). Starting with sample J which is introduced in 2011 the Biography Questionnaire has been reintegrated in the Personnel Questionnaire and is answered again in the first year. This is also true for the migration sample M1 for which the Biography Questionnaire is only a subset of a larger number of questions also covering own and parental migration history.

In addition to the Biography Questionnaire, there has been an independent questionnaire (Youth Questionnaire) in SOEP since 2000 for the group of survey participants who are 16-17 years old and are being interviewed for the first time (see also chapter BIOYOUTH). [2]

In SOEPv31 the former independent survey "Familien in Deutschland" (FiD) was integrated in the SOEP distribution (subsample L1-L3). As the latest FiD version (FiDv4.0) also contained a BIOPAREN file which was generated in a similar way to the SOEP version of BIOPAREN, it was possible to fully integrate the information of FiD in SOEP BIOPAREN.[3]

The biography information in FiD was collected as follows: since the first wave of FID in 2010 the respondents received a separate Biography Questionnaire in addition to the Individual Questionnaire. In 2010 it included only questions about the history of the surveyed persons themselves. In 2011 (the second wave of FID) another part of the Biography Questionnaire was handed out, where intergenerational aspects of the persons surveyed were included by means of a special group of questions. This deals with statements made about the education or professional training of the parents, the parents' residency, and their year of birth and death. In 2012 and 2013 the complete collection of biography questions was included in only one Biography Questionnaire for individuals surveyed for the first time. However, those who had only answered either of the two separate parts handed out in wave one and two were given the other, still unanswered part.

Therefore, there are persons who filled in the two Biography Questionnaire parts separately as well as persons who were given the complete questionnaire. In addition, there are respondents who only answered one of the two parts because they dropped out of the study at some point. This is the reason why there are persons who answer two Biography Questionnaires (but different parts of it) in two different years.

[2] A more precise representation of the development of the instruments used to collect the Personal History, including the social origin, can be found in the introduction to this documentation.

[3] A detailed documentation of FiD BIOPAREN is available in the FiD data documentation.

In addition to the Biography Questionnaire, similar to the SOEP, there was an independent questionnaire (Youth Questionnaire) for the group of survey participants who are 17 years old and are being interviewed for the first time.

Table 1 gives an overview on the development of the number of respondents that enter BIOPAREN in each year by subsample.

Table 1: Number of observations in BIOPAREN

Year of data collection	N	Samples															
		A	B	C	D	E	F	G	H	I	J	K	L1	L2	L3	M1	M2
1984	1,682	1,079	603	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1985	11,087	8,369	2,718	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1986	501	355	146	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1987	464	310	154	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1988	380	245	135	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1989	384	246	138	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1990	4,811	221	139	4,451	-	-	-	-	-	-	-	-	-	-	-	-	-
1991	506	219	118	169	-	-	-	-	-	-	-	-	-	-	-	-	-
1992	497	202	142	153	-	-	-	-	-	-	-	-	-	-	-	-	-
1993	471	201	134	136	-	-	-	-	-	-	-	-	-	-	-	-	-
1994	930	213	97	148	472	-	-	-	-	-	-	-	-	-	-	-	-
1995	1,067	217	76	135	639	-	-	-	-	-	-	-	-	-	-	-	-
1996	483	214	88	144	37	-	-	-	-	-	-	-	-	-	-	-	-
1997	487	194	98	144	51	-	-	-	-	-	-	-	-	-	-	-	-
1998	419	192	65	127	35	-	-	-	-	-	-	-	-	-	-	-	-
1999	2,047	206	62	130	35	1,614	-	-	-	-	-	-	-	-	-	-	-
2000	455	168	60	146	35	46	-	-	-	-	-	-	-	-	-	-	-
2001	9,433	156	46	116	33	46	9,036	-	-	-	-	-	-	-	-	-	-
2002	836	150	38	117	42	50	439	-	-	-	-	-	-	-	-	-	-
2003	2,677	155	53	125	33	56	258	1,997	-	-	-	-	-	-	-	-	-
2004	807	145	50	134	26	26	281	145	-	-	-	-	-	-	-	-	-
2005	663	136	50	94	24	36	248	75	-	-	-	-	-	-	-	-	-
2006	529	115	30	92	16	36	170	70	-	-	-	-	-	-	-	-	-
2007	2,577	148	43	89	29	41	202	47	1,978	-	-	-	-	-	-	-	-
2008	594	122	35	48	11	24	162	44	148	-	-	-	-	-	-	-	-
2009	439	103	28	47	19	22	134	32	54	-	-	-	-	-	-	-	-
2010	3,983	97	21	41	8	17	108	35	34	1,891	-	-	684	1,047	-	-	-
2011	12,133	99	30	34	8	30	109	30	30	-	5,161	-	3,041	3,281	280	-	-
2012	4,904	74	27	33	11	3	96	23	33	-	175	2,469	198	441	1,321	-	-
2013	5,930	70	16	43	15	-	102	30	36	-	152	103	44	261	94	4,964	-
2014	1,127	69	22	28	8	3	89	24	35	-	111	53	49	178	101	357	-
2015	2,627	66	18	25	5	2	74	18	17	-	81	54	60	223	118	155	1,711
Total	75,930	14,556	5,480	6,949	1,592	2,052	11,508	2,570	2,365	1,891	5,680	2,679	4,076	5,431	1,914	5,476	1,711

Source: own calculations based on SOEP v32 (PBIOSPE), doi: 10.5684/soep.v32

.

13.3 How is BIOPAREN generated?

The information available in BIOPAREN is obtained in two different ways. On the one hand, BIOPAREN includes the children's proxy entries on the parents from the Biography Questionnaire and the Youth Questionnaire. On the other hand, it contains the direct entries from the parents in the case the respondent lives in the same household as his parents. Every respondent is asked for information on the regional mobility of the children, as well as on the religious affiliation of the parents. However, information on the year of birth, as well as the education and occupational training of the parents, additional to the professional position and occupation of the father are not collected due to the filter command in the questionnaire when the parents (or the father) live in the same household as the child at the time of the survey. In this case, the direct entries of the parents are used.

The identification of the parents occurs first of all through the variable $$STELL (relationship to head of household). The possible values of the variable $$STELL (waves A-BB) are listed in Table 2. The combinations of these characteristics of the $$STELL-variable and their assigned interpretation for the generation of parent identifiers are describe in Table 3. Starting in wave BC the range of the $$STELL variable was extended to capture more complex household combinations. Table 4 shows the possible values of the new $$STELL and Table 5 lists which combinations of this new variable were used to generate parent identifiers.

The second source of information is the population of the file $$KIND, which includes all children under the age of 16. The file contains the personal number of the mother, as well as the personal number of the partner of the mother. Through both variables the latest (social) mother, as well as the latest partner of the mother are identified, ideally, at the time when the child is 16 years old and thus one year before the first survey of the child. In the case the parents could not be identified by the $$STELL variable, this information is used.

In a further step the biological mother is identified through the mother-child relationship in the file BIOBIRTH. In the event that still no personal number for the mother exists, the number from BIOBIRTH is used. Since 2001, an extra BIOBIRTH data-set exists for fathers (formerly BIOBRTHM, since SOEPv31 BIOBIRTH and BIOBRTHM were merged), which provides a way of identifying fathers of SOEP-respondents. For those cases where no social mother/father could be identified via $$STELL and/or $$kind, the biological mother or father identifier is used.

Please note that BIOPAREN focuses on the **social** parents. **Biological** parent identifier can be found in BIOBIRTH (see also chapter on BIOBIRTH).

Table 2: Characteristics of the variable $STELL "relationship of the person to the head of the household" (waves A-BB)

Code	Description
0	HH
1	Marital partner of the HH
2	partner of the HH
3	Daughter/son (also adopted/stepchild) of the HH
4	Foster child of the HH
5	Daughter/son–in-law of the HH
6	Father/mother of the HH
7	Father/mother-in-law of the HH
8	Brother/sister, brother/sister-in-law of the HH
9	Grandchild of the HH
10	Other relationship to the HH
11	Not related to the HH
12	Daughter/son of the partner of the HH
13	Marital partner of the HH (same sex)

Table 3: Possible Parent-Child Relationships based on $$STELL (waves A-BB)

Relationship of the child to the HH	Relationship of the parent to the HH	Person is ...
3	0	Child of HH
3	1 or 2	Child of marital/ partner of HH
4	0	Foster child of HH
4	1 or 2	Foster child of marital/ partner of HH
12	2 or 3	Child of partner of HH
9	3 or 4	Child of child/foster child of HH
0	6	Child is HH, lives with parents in same household
1 or 2	7	Marital partner/partner of HH (child of in laws of HH)
9	5	Grandchild of HH (child of son/daughter-in-law of HH)

Table 4: Characteristics of the variable $STELL "relationship of the person to the head of the household" (starting with wave BC)

Code	Description
0	HH
11	Marital partner of the HH
12	Marital partner (same sex) of the HH
13	Partner of the HH
21	Daughter/Son of the HH
22	Stepchild (or child of partner) of the HH
23	Adopted child of the HH
24	Foster child of the HH
25	Grandchild of the HH
26	Great-grandchild of the HH
27	Daughter/son–in-law (or partner of child) of the HH
31	Father/mother of the HH
35	Father/mother-in-law of the HH
36	Grandfather/Grandmother of the HH
41	Brother/sister of the HH
42	Halfbrother/Halfsister of the HH
43	Stepbrother/stepsister of the HH
52	Brother/sister-in-law of the HH
61	Uncle, aunt of the HH
62	Niece/nephew of the HH
63	Cousin of the HH
64	other relative of the HH
71	Not related to the HH

Table 5: Possible Parent-Child Relationships based on $$STELL (starting with wave BC)

Relationship of the child to the HH	Relationship of the parent to the HH
21, 22, 23, 24	0, 11, 12, 13
25	21, 22, 23
25	27
26	25
27	52
0, 41, 42	31
52	35
62	41, 42, 43
63	61

13.4 What's new in version v32?

- Minor corrections on missing values.

13.5 Complete list of variables in BIOPAREN

HHNR	Number of the original household
PERSNR	Personal number of the respondent (all persons)
VNR	Personal number of the father of the respondent
MNR	Personal number of the mother of the respondent
VGEBJ	Year of birth of the father
MGEBJ	Year of birth of the mother
VTODJ	Year of death of the father
MTODJ	Year of death of the mother
VAORT91	Residency of the father 1991 (Survey focus: family)
MAORT91	Residency of the mother 1991 (Survey focus: family)
VAORT96	Residency of the father 1996 (Survey focus: family)
MAORT96	Residency of the mother 1996 (Survey focus: family)
VAORT01	Residency of the father 2001 (Survey focus: family)
MAORT01	Residency of the mother 2001 (Survey focus: family)
VAORT06	Residency of the father 2006 (Survey focus: family)
MAORT06	Residency of the mother 2006 (Survey focus: family)
VAORT11	Residency of the father 2011 (Survey focus: family)
MAORT11	Residency of the mother 2011 (Survey focus: family)
VSBIL	Education of the father
MSBIL	Education of the mother
VBBIL	Vocational training of the father
MBBIL	Vocational training of the mother
VSINFO	Origin of the information on father's education
MSINFO	Origin of the information on mother's education
VBINFO	Origin of the information on father's vocational training
MBINFO	Origin of the information on mother's vocational training
VRELI	Religious affiliation of the father
MRELI	Religious affiliation of the mother
VNAT	Nationality of the father

MNAT	Nationality of the mother
VBSTELL	Professional position of the father (when the respondent was 15 years old)
VBSINFO	Origin of the information on the professional position of the father
MBSTELL	Professional position of the mother (when the respondent was 15 years old)
MBSINFO	Origin of the information on the professional position of the mother
VISCO88	Professional occupation of the father (when the respondent was 15 years old)
MISCO88	Professional occupation of the mother (when the respondent was 15 years old)
VISEI	Prestige score of father – concept of Ganzeboom
MISEI	Prestige score of mother – concept of Ganzeboom
VMPS	Prestige score of father – Magnitude scale – Wegener
MMPS	Prestige score of mother – Magnitude scale – Wegener
VSIOPS	Prestige score of father – Treiman standard score
MSIOPS	Prestige score of mother – Treiman standard score
VEGP	Prestige score of father – Erikson – Goldthorpe class category
MEPG	Prestige score of mother – Erikson – Goldthorpe class category
VBKLAS	Occupational coding scheme father according German statistical office
MBKLAS	Occupational coding scheme mother according German statistical office
ORTKINDH	Place of childhood
ORTKIND1	Still lives in place of childhood?
ORTKIND2	Year moved out of parents' household *(since 2000 no longer collected)*
ORTKIND3	Still lives in parents' household *(since 2000 no longer collected)*
LIVING1	No. of years living with both parents
LIVING2	No. of years living alone with mother
LIVING3	No. of years living with mother and new partner of mother
LIVING4	No. of years living alone with father
LIVING5	No. of years living alone with father and new partner of father
LIVING6	No. of years living with other relatives
LIVING7	No. of years living with foster parents
LIVING8	No. of years living in youth center

VSTREIT	Conflict with father
MSTREIT	Conflict with mother
VAORTAKT	Father‘s place of residence
MAORTAKT	Mother‘s place of residence
BIOYEAR	Year of the Biography Survey
BIO	Origin of the information
ALTER	Age of the respondents
VALTER	Age of the respondent’s father
MALTER	Age of the respondent’s mother
MORIGIN	Country of origin of the respondent’s mother
VORIGIN	Country of origin of the respondent’s father
GESCHW	Sibling yes/no
GESCHWUP	Year of update of GESCHW
NUMS	Number of sisters
NUMB	Number of brothers
TWIN	Twin sister/brother

Variable	VNR / MNR
Label:	Personal number of the father of the respondent / Personal number of the mother of the respondent
Values:	(-1) PERSNR father / mother unkown (-2) Does not apply (-3) Answer improbable

Description: The personal ID of the parents (VNR and MNR) is generated in three steps.

1. The parents of the respondent are identified by the relationship to the head of the household ($$STELL in $$BRUTTO). Ideally, the children's parents are identified at the time of the first survey of the child. Furthermore, the **social** parents and not necessarily the **biological** parents are identified.

2. The parents of the respondent are identified via the mother's ID as well as the mother's partner ID in $$KIND. By using these variables the "oldest" parents are identified. Ideally, these are the parents at the time the child is 16 years old (one year before the first survey).

3. The biological mother-ID of the respondent can be identified in BIOBIRTH and the father-ID in BIOBRTHM (from SOEPv31 onwards both are stored in BIOBIRTH).

As BIOPAREN aims at identifying the social parents that live in the household when the child is surveyed, the steps above are carried out in the hierarchy 1-3 with step 1 having the highest priority. If one is interested in only biological parents, please have a look at the information in BIOBIRTH and BIOBRTHM.

Variable **VGEBJ / MGEBJ**

Label: Year of birth of the father / Year of birth of the mother

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: In a first step the information of the year of birth comes from the Biography Questionnaire. Due to a filter command, the children's proxy entries are only available for these variables when the parents or one parent and the child do not live in the same household at the time of the survey.

After the parents' personal numbers have been identified the information can be compared with the entries in PPFAD. If there are differences of +/- two years the VNR / MNR will be set as missing.

For the missing entries the information of the parents' year of birth is taken from PPFAD.

Variable **VTODJ / MTODJ**

Label: Year of death of the father / Year of death of the mother

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables are generated as usual using the information from the Youth Questionnaire or the Biography Questionnaire and the parents' direct entries from PPFAD. As a next step the annual proxy information on a parent's death from the $$P-files are used. Furthermore, we use information of the month of death of a parent from the year before. That means we have information on the death of a father or a mother for the years 2002 onwards. With this data a wrong marking as "no death in 2002" / "death in 2003" can be corrected if there is data from 2003 indicating that one parent died e.g. in October 2002.

The variables VTODJ and MTODJ will be updated with new survey information. They are updated as long as the father or the mother is part of the SOEP sample. Since 2003 we additionally use the annual proxy information of respondents about reported life events of the last year.

Variable **VAORT91/ MAORT91 to VAORT11/ MAORT11**

Label: Residency of the father/ Residency of the mother in 1991, 1996, 2001, 2006 and 2011

Values:

(-1) No answer	(4) Lives Same Town
(-2) Does not apply	(5) Lives Other Town
(-3) Answer improbable	(6) Lives Elsewhere In Germany
(0) Has Died	(7) Lives Elsewhere
(1) Lives In Same HH	(8) Lives Else E Germany
(2) Lives In Same Housing	(9) Lives Else W Germany
(3) Lives Neighborhood	(10) Lives Foreign Country

Description: The information on the residency of the parents stems from the Youth and Biography Questionnaires as well as from the Person Questionnaire.

The information from the $$P-files have a higher priority. For more details see also the information on VAORTAKT / MAORTAKT.

Variable **VAORTAKT/ MAORTAKT**

Label: Father‘s place of residence /
Mother‘s place of residence

Values:

(-1) No answer	(4) Lives Same Town
(-2) Does not apply	(5) Lives Other Town
(-3) Answer improbable	(6) Lives Elsewhere In Germany
(0) Has Died	(7) Lives Elsewhere
(1) Lives In Same HH	(8) Lives Elsewhere E Germany
(2) Lives In Same Housing	(9) Lives Elsewhere W Germany
(3) Lives Neighborhood	(10) Lives Foreign Country

Description: The variables VAORTAKT and MAORTAKT contain the latest available information about the parents’ residence and on whether or not they are deceased, respectively.

For persons without identified parents who answered the biography questionnaire up to the year 2011, the most recent available information from the Person Questionnaire in 1991, 1996, 2001, 2006 or 2011 was assumed.

For those persons whose parents are identified in the SOEP, the information on the year of death in PPFAD was used for updating. If the year of death lies chronologically after the latest available information, VAORTAKT and MAORTAKT were put on “deceased”.[1]

[1] In gathering the information from different data sets, inconsistencies occurred. On the one hand, some parents had been reported as deceased in the early waves, while information about their residence at a later date was available. In this case, the information about the parents’ residence was not accepted.

Variable **VAORTUP / MAORTUP**

Label: Year of update of VAORTAKT/MAORTAKT

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variable contains the year, in which the information stored in VAORTAKT and MAORTAKT has been updated.

Variable **VSBIL / MSBIL**

Label: Education of the father / Education of the mother

Values:
(-1) No answer

(-2) Does not apply

(-3) Answer improbable

(0) Do Not Know

(1) Secondary School Degree

(2) Intermediate School Degree

(3) Technical School Degree

(4) Upper Secondary School Degree

(5) Other Degree

(6) No School Degree

(7) School Not Attended

Description: The parents' education is generated with information from the Youth Questionnaire, the Biography Questionnaire and direct entries from the $$PGEN-files. Due to the filter command, the children's proxy entries are only available for VSBIL / MSBIL when the parents or one parent and the child do not live in the same household at the time of the survey.

Along with other variables already contained in BIOPAREN, there will be an update with new survey information, insofar as no valid values exist in BIOPAREN.

Variable **VBBIL/ MBBIL**

Label: Vocational training of the father / Vocational training of the mother

Values:

(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(0) Do Not Know
(10) No Vocational Degree
(20) Vocational Degree
(21) Trained in Foreign Company
(22) Trained long Time in Foreign Company
(23) Foreign Vocational School
(24) Trade, Farming Apprentice
(25) Business Apprentice
(26) Health Care School
(27) Special Technical School
(28) Civil Service Training
(30) Tech Engineer School
(31) Foreign Collage
(32) College, University
(40) Other Training
(50) Currently in Vocational Training
(51) Currently in Schooling

Description: The parents' vocational training is generated the same way as the education variables (see VSBIL / MSBIL).

Variable **VSINFO/ MSINFO**

Label: Origin of the information on father's education /
Origin of the information on mother's education

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

(0) Do Not Know

(1) Biography-Proxy

(2) $P-Individual Info

Description: The variable contains the origin of the information on parental education.

Variable **VBINFO/ MBINFO**

Label: Origin of the information on father's vocational training /
Origin of the information on mother's vocational training

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

(0) Do Not Know

(1) Biography-Proxy

(2) $P-Individual Info

Description: The variable contains the origin of the information on parental vocational training.

Variable **VRELI/ MRELI**

Label: Religious affiliation of the father / Religious affiliation of the mother

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

(0) Do Not Know – Proxy

(1) Catholic

(2) Protestant

(3) Other Christian Denomination

(4) Islamic Denomination

(5) Other Denomination

(6) No Denomination

Description: The questions about the religious affiliation of the parents are only asked to children who are not living in the household of their parents. In order to provide as much information as possible we gather data from the Person Questionnaires. The reconstruction of the information on the parents' religion is restricted to the survey years 1997, 2003 and 2007 as they are the only years where information of a respondent's religious affiliation is available. In all survey years the question was formulated differently but the information is made as consistent as possible.

Variable **VNAT/ MNAT**

Label: Nationality of the father / Nationality of the mother

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

(1) German

(2) Other

Description: The information on the parents' nationality is generated similar to VRELI / MRELI. The question is only asked to children who are not living in the same household as their parents. $$PGEN information is used to compute a variable with data from 2006 onwards. In a further step the parents' personal numbers are used to match information on parents' nationality with data from the $$PGEN-files in the case if there are missing entries.

Variable **VBSTELL/ MBSTELL**

Label: Professional position of the father (when the respondent was 15 years old) / Professional position of the mother (when the respondent was 15 years old)

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The children's proxy entries on professional position and occupation of the father (VBSTELL) as well as VISCO88 and all prestige scores are available when the father and the child do not live in the same household at the time of the survey and if the father lived in Germany when the child was 16 years old. Since 2000, the same applies to the entries of the mother.

Besides the proxy entries parents' direct information from the $$P-files are used.

Variable **VBSINFO/ MBSINFO**

Label: Origin of the information on the professional position of the father / Origin of the information on the professional position of the mother

Values:

(-1) No answer

(-2) Does not apply

(-3) Answer improbable

(0) Do Not Know-Proxy

(1) Biography-Proxy

(2) $P-Individual Info

Description: The variables VBSINFO / MBSINFO are indicator variables. They tell whether the information is from the Biography or Youth or Person Questionnaires. This information is generated at the same steps as it is done with the VBSTELL / MBSTELL variables.

Variable **VISCO88/ MISCO88**

Label: Professional occupation of the father (when the respondent was 15 years old) / Professional occupation of the mother (when the respondent was 15 years old)

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables contain the ISCO88 code for the father and mother.

Variable **VISEI/ MISEI**

Label: Prestige score of father – concept of Ganzeboom /
Prestige score of mother – concept of Ganzeboom

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables contain the ISEI code for the father and mother.

Variable **VMPS/ MMPS**

Label: Prestige score of father – Magnitude scale – Wegener /
Prestige score of mother – Magnitude scale – Wegener

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables contain the prestige scores (magnitude scale - Wegener) for the father and mother.

Variable **VSIOPS/ MSIOPS**

Label: Prestige score of father – Treiman standard score /
Prestige score of mother – Treiman standard score

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables contain the prestige scores (Treiman standard score) for the father and mother.

Variable **VEGP/ MEGP**

Label: Prestige score of father – Erikson – Goldthorpe class category/
Prestige score of mother – Erikson – Goldthorpe class category

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables contain the prestige scores (EGP) for the father and mother.

Variable **VBKLAS/ MBKLAS**

Label: Occupational coding scheme father according German statistical office/
Occupational coding scheme mother according German statistical office

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables contain the occupational code for the father and mother according to the coding scheme of the German statistical office.

Variable **ORTKINDH / ORTKIND1 / ORTKIND2 / ORTKIND3**

Label:	ORTKINDH	Place of childhood /
	ORTKIND1	Still lives in place of childhood? /
	ORTKIND2	Year moved out of parents' household (since 2000 no longer collected) /
	ORTKIND3	Still lives in parents' household (since 2000 no longer collected)

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

ORTKINDH:	**ORTKIND1:**
(1) Large City	(1) Yes, Still
(2) Medium City	(2) Yes, Again
(3) Small City	(3) No
(4) Countryside	

Description: The variables provide information on the place of childhood.

Variable **LIVING1 / LIVING2 / LIVING3 / LIVING4 / LIVING5 /**

LIVING6 / LIVING7 / LIVING8

Label:	LIVING1	No. of years living with both parents
	LIVING2	No. of years living alone with mother
	LIVING3	No. of years living with mother and new partner of mother
	LIVING4	No. of years living alone with father
	LIVING5	No. of years living alone with father and new partner of father
	LIVING6	No. of years living with other relatives
	LIVING7	No. of years living with foster parents
	LIVING8	No. of years living in youth center

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variables show the total number of years for different categories of where the child lived during his childhood.

Variable **VSTREIT/ MSTREIT**

Label: Conflict with father /
Conflict with mother

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

(1) Very Often

(2) Often

(3) Sometimes

(4) Seldom

(5) Never

(6) Person Not Present

Description: The variables provide information on the frequency of conflicts with the parents.

Variable **BIOYEAR**

Label: Year of the Biography Survey

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: The variable BIOYEAR tells in which year the information was surveyed.

Variable **BIO**

Label: Form of Biography Questionnaire

Values:
(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(0) Participation before 2001
(1) Youth
(2) Biolela blue
(11) FiD Youth
(12) FiD Lela
(13) FiD Bio Part I
(14) FiD Bio Part II
(15) FiD Bio Part I a. II

Description: Since 2008 the variable BIO is generated to indicate the origin of the information in BIOPAREN. This information is valid for all cases from 2001 onwards.

Variable ALTER / VALTER / MALTER

Label: Age of the respondents / Age of the respondent´s father / Age of the respondent`s mother

Values:

(-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: Since 2008 the variables ALTER, VALTER and MALTER were added to BIOPAREN.

The variable ALTER gives the age of the respondent at the moment of the interview. VALTER gives the age of the respondents' father when the respondent answered the Biography Questionnaire or the Youth Questionnaire. The same was applied for the mothers with the variable MALTER. In order to generate the variables the information for the parents who are identified in the SOEP was gained with data from PPFAD. The proxy entries from BIOPAREN were used when there weren't any information of the respondents parents available. If the year of death lies chronologically before the latest available information, MALTER and VALTER were put on "deceased".

Variable V**ORIGIN/ MORIGIN**

Label: Country of origin of the respondent's father /
Country of origin of the respondent's mother

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: These variables give information about the country of origin of the respondents mother (MORIGIN) and father (VORIGIN). This information is collected in the Youth Questionnaires since 2007. Another source of information can be found in PPFAD by the direct-entries of the parents in the variable CORIGIN. These two kinds of information, proxy- and direct-entries, are used to generate MORIGIN and VORIGIN. In a first step we use the proxy-information for all the parents whose children made an entry in the Youth Questionnaire. For all the parents where there are no proxy-information available, we then use the direct-entries of the parents from the PPFAD-variable CORIGIN.

Variable **GESCHW**

Label: Siblings yes/no

Values:
(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(1) Yes
(2) No

Description: GESCHW contains the information if a respondent has siblings or not. The question is asked since 2003 in the Biography and Youth Questionnaire. In 2003 the question was asked in the Person Questionnaire.

For more Information about siblings see also BIOSIB, BIOTWIN, and the data in the $p-Files from the family focus questions in the Person Questionnaire (1991, 1996, 2001, 2006, 2011 and 2013).

Variable **GESCHWUP**

Label: Time of update - siblings

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: GESCHWUP contains the year, in which the sibling information was surveyed.

Variable **NUMS**

Label: Number of sisters

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: Nums contains the number of sisters.

Variable **NUMB**

Label: Number of brothers

Values: (-1) No answer

(-2) Does not apply

(-3) Answer improbable

Description: Numb contains the number of brothers.

Variable **TWIN**

Label: Twin sister/brother

Values:
(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(1) Yes, monozygotic
(2) Yes, dizygotic
(3) No

Description: Twin contains information whether the respondent has a twin sibling.

14 BIOIMMIG: Generated and Status Variables from SOEP for Foreigners and Migrants

by Jan Goebel and Katharina Strauch

14.1 Content

The variables contained in BIOIMMIG deal with questions related to foreigners in (and migrants to) Germany. Specifically, questions concerning desire to return to the home country, the presence of relatives in the home country, reasons for coming to Germany, and conditions upon initial arrival in Germany. A complete list of variables is shown in the table with German and English labels. Due to changes in the questionnaires over the years, some variables are not available for all years. The information on the yearly availability is shown in the different variable descriptions.

14.2 Status Variables and Carrying Forth of Information

The data available in this file are longitudinal, that is to say, the same variable name refers to different time periods, differentiated by the variable SYEAR. The data is stacked for each person, such that the unit of observation is a person-year. Thus for every person, there are as many observations as interviews given by this person. Much of the information was asked only once, and „carried“ forth in the following years. Frequencies can be found in SOEPINFO.

The sample in the dataset is defined by taking all available information and deleting all those persons who:

are born in Germany *and*
have German nationality *and*
have no valid BIOIMMIG information in any wave that they were observed.

As the data consists of person-year observations, if a person is excluded from the sample, then for all years. However if a person once belonged to the sample, then he is always included (say, even after receiving German citizenship).

List of Variables

Variable	German	English
PERSNR	Personennummer	Person Number
HHNR	Ursprüngliche HH-Nummer	Original HH Number
HHNRAKT	Aktuelle HH-Nummer für SYEAR	Current HH Number for ERHEBJ
SYEAR	Jahr/Erhebungsjahr	Current Year / Year Answered
BIIMGRP	BI: Status bei Einwanderung in Dt.	BI: Immigration Group
BIRESPER	BI: Status Aufenthaltserlaubnis	BI: Residence Status
BICAMP	BI: Aufnahmelager: J/N	BI: Refugee Residence Y/N
BICAMPW	BI: Aufnahmelager: Wochen	BI: Refugee Residence: Weeks
BICAMPM	BI: Aufnahmelager: Monate	BI: Refugee Residence: Months
BIWFAM	BI: Eingereist als Familienangehöriger	BI: Already had Family in Country
BIFAMC	BI: Vor Einreise Kontakte mit Pers.	BI: Contacts with Family in Germany
BIFAMCL	BI: Zuzug in Wohnort der Bekannten	BI: Moved to Same City/Town as Family
BIREASON	BI: Hauptgtund Zuzug (ab 2014)	BI: Main Reason Migrate (Since 2014)
BIRBETR	BI: Gründe Zuzug D: Besser	BI: Reason Migrate: Better
BIRMONEY	BI: Gründe Zuzug D: Geld	BI: Reason Migrate: Money
BIRFREE	BI: Gründe Zuzug D: Freiheit	BI: Reason Migrate: Freedom
BIRFAM	BI: Gründe Zuzug D: Familie	BI: Reason Migrate: Family
BIRPOOR	BI: Gründe Zuzug D: Armut	BI: Reason Migrate: Poor
BIRWAR	BI: Gründe Zuzug D: Krieg	BI: Reason Migrate: War
BIRJUST	BI: Gründe Zuzug D: Einfach So	BI: Reason Migrate: Just So
BIROTHR	BI: Gründe Zuzug D: Sonstiges	BI: Reason Migrate: Other
BIEXPR	BI: Vorstellungen von Dt.	BI: Expectations in Germany
BIEXPRLV	BI: Eigene Wohnung finden	BI: Expectations: Find Apt
BIEXPRAC	BI: Von Arbeitskollegen akzeptiert	BI: Expectations: Accepted by Coworker
BIEXPRAN	BI: Von Nachbarn akzeptiert	BI: Expectations: Accepted by Neighbor
BIRELH	BI: In Heimatland Familienmitglieder	BI: Family Abroad
BIRELHP	BI: In Heimat: Eltern	BI: Family Abroad: Parents
BIRELHGP	BI: In Heimat: Großeltern	BI: Family Abroad: Grandparents
BIRELHC	BI: In Heimat: Kinder	BI: Family Abroad: Children
BIRELHBS	BI: In Heimat: Bruder, Schwester	BI: Family Abroad: Brother/Sister
BIRELHDR	BI: In Heimat: Entferntere Verwandte	BI: Family Abroad: Distant Relatives
BIRELHSP	BI: In Heimat: Ehepartner, Verlobte(r)	BI: Family Abroad: Spouse
BIRELHFR	BI: In Heimat: Bekannte/Freunde	BI: Family Abroad: Friends
BIRELHMI	BI: Personen gern nach Dt. holen?	BI: Persons abroad bring to Germany
		...
BIRELHS2	BI: Ehepartner in Deutschland	BI: Spouse in Germany

Variable	German	English
BIRELHC2	BI: Kinder unter 18 J. nicht in D	BI: Underage Children not in Germany
BIGOBACK	BI: Rückkehr Heimat (ab 1994)	BI: Go back home ?
BISTAY	BI: Wunsch in D zu bleiben	BI: Desire to Stay in Germany
BISTAYY	BI: Dauer des geplanten Aufenthalts	BI: Years Desired to Stay in Germany
BISCGER	BI: In Dt. Schule besucht?	BI: Attended School in Germany
BISCGRAD	BI: In welche Klasse in dt. Schule	BI: Which Grade School
BISCGERC	BI: Besuch spezieller Vorbereitung	BI: Attended Special Foreigner Prep Class
BISCGC	BI: Auch dt. Schüler in Schulklasse	BI: Also German Pupils in Class
BISCGCF	BI: Wie viel Mitschüler Ausländer	BI: How many Pupils foreign
biscgcfn	BI: Eine oder mehrere Nationalität	BI: Mix of Nationalities in Class

14.3 Updating of Time-Dependent Information

The variables found in BIOIMMIG are created first using information from the SOEP biography files, the so-called BIOLELA, $LELA (starting with wave M) files. Additionally, starting in 2000 (wave Q), $JUGEND is collected of 16 and 17 year-olds, containing similar information to $LELA. In any given year, a person can have only information from $JUGEND or $LELA, but not both. If valid information is found in the $LELA or $JUGEND files for the given response year, then it is taken. Yearly valid update information is taken from the foreigner specific files APAUSL through $PAUSL and the foreigner specific questions in MP, NP, OP and onwards. Starting with wave M, the foreigner specific variables are found in the regular $P files, as the questionnaire is identical for natives and foreigners. Sometimes there is competing information in the biography and regular yearly person questionnaires. The most recent valid information is taken to be correct. First the $LELA or $JUGEND info is used and then updated with valid/non-missing information from the person questionnaire.

The data release v31 includes the complete data from “Familien in Deutschland” (Families in Germany, FiD), which is being retrospectively integrated into the SOEP. Hence new samples are retrospectively integrated into the bioimmig data set for the years 2010 to 2012. Due to that the value “-5” occurs in some variables for the FiD-Sample. This is the case when the variable was not part of the FiD questionnaire.

In 2013 the IAB-SOEP Migration Sample started. In that special sample immigrants in Germany are surveyed. Due to the fact that the survey for the immigrants is not identical to the regular questionnaire, information can only be used partly for the bioimmig file.

14.4 Using this File

The BIOIMMIG file can be used in cross-section or in panel. The usual matching variables are included: PERSNR (Person Number), HHNR (Original HH Number), HHNRAKT (Current HH Number for survey year given in SYEAR), SYEAR(Year). The data is sorted by HHNR, HHNRAKT, PERSNR, SYEARsuch that there are typically many person-year observations for every person. In that sense, the data are ready to be used/matched to a longitudinal dataset. However, simply by selecting on the appropriate year inSYEAR, the file can be used cross-sectionally as well.

The data structure looks like the following (using fictitious data in this example):

PERSNR	HHNR	HHNRAKT	ERHBEBJ	BIIMGRP	BIRESPER
101	19	19	1995	-2	-2
101	19	19	1996	2	1
101	19	19	1997	2	1
101	19	19	1998	2	1
102	19	19	1995	3	2
0102	19	19	1996	3	2
102	19	19	1997	3	2

14.5 Using BIOIMMIG as a Cross-Section

An example of how to use BIOIMMIG in a cross-section would be as follows:

(A) Open BIOIMMIG, keeping only those observations in BIOIMMIG for a particular year.

in Stata:
```
use bioimmig if syear==1984
```

(B) Rename all the desired variables with wave-specific information.

in Stata:
```
rename bicamp camp1984
rename bicampw campw1984
```

(C) Save the ID's and the renamed variables in a temporary file

in Stata:
```
sort hhnr persnr
save /tmp/bioim1984, replace
```

(D) Merge the temporary file to your main dataset

in Stata:

```
merge hhnr persnr using /tmp/bioim1984, nokeep
drop _merge
```

(E) Repeat starting at step (A) for all years of interest, i.e. `syear==1985`

14.6 Documentation of the Variables

Below, each variable is listed and its variable and value labels are displayed in both English and German. A list of the main source variables used in the generation is provided for reference purposes. Further, there is also information as to what question the variables correspond to in the Wave 13 -M-1996 Biography Questionnaire.

Problems:

If you encounter problems using this file, first-aid is available from the original STATA source code used to create this file, delivered with the regular SOEP data distribution.

BIIMGRP BI: Status bei Einwanderung in Dt.
BI: Immigration Group

BIO Question: Q5

Comment: The possible groups change in 2000, such that "[1] East German" and "[5] Non EU "are no longer identified starting 2000. However, as information can be carried forth from previous years, there may be valid [1] and [3] values starting 2000, but only if the information was collected before 2000.

German: Zu welcher der folgenden Zuwanderergruppen gehörten Sie, als Sie nach Deutschland kamen?

"[1] Ostdeutsche (bis 1995 erfragt) "
"[2] Aussiedler,Osteuropa"
"[3] Deutsche,Ausland"
"[4] EU-Mitgliedsstaat (bis 2009 EG) "
"[5] Asylbewerber,Fluechtling"
"[6] Sonstige Auslaender"
"[7] Sonstige"

English: Which immigrant group did you belong to, when you came to Germany?

"[1] East German (last time asked in 1995)"
"[2] Ethnic German"
"[3] German, Lived Abroad"
"[4] EU-Member (up to 2009 EG)"
"[5] Asylum Seeker"
"[6] Other Foreign Nationality"
"[7] Other"

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P070Z
1995 L	BIOLELA	P070Z
1996 M	MLELA	MB070Z
1997 N	NLELA	NB070Z
1998 O	OLELA	OB070Z
1999 P	PLELA	PB070Z
2000 Q	QLELA	QB070Z
2001 R	RLELA	RB070Z
2002 S	SLELA	SB05
2003 T	TLELA	TB05
2004 U	ULELA	UB05
2005 V	VLELA	VB05
2006 W	WLELA	WB05
2007 X	XLELA	XB05
2008 Y	YLELA	YB05
2009 Z	ZLELA	ZB05
2010 BA	BALELA	BAB05
2011 BB	BBLELA	BBB05
2012 BC	BCLELA	BCB05
2013 BD	BDLELA	BDB06
2014 BE	BELELA	BEB0601
2015 BF	BFLELA	BFB0601
2013 BD	BDP_MIG	BDPM_P_1002
		BDPM_P _1202
		BDPM_P _09
		BDPM_L _0201
		BDPM_L 01_1601
		BDPM_L01 _1701
		BDPM_L01_1801
		BDPM_L 01_2001
		BDPM_L *_2401
		BDMP_L*_2501
2014 BE	BEP_MIG	BEPM_L_08
		BEPM_L*_1201
		BEPM_L*_17
		...

Year	File	Variable
2015 BF	BFP_MIG	BFPM_L_08 BFPM_L*_1401 BFPM_L*_19 BFPM_L_*_22
	QJUGEND	QJ57
2001 R	RJUGEND	RJ59
2002 S	SJUGEND	SJ59
2003 T	TJUGEND	TJ59
2004 U	UJUGEND	UJ59
2005 V	VJUGEND	VJ59
2006 W	WJUGEND	WJ64
2007 X	XJUGEND	XJ64
2008 Y	YJUGEND	YJ64
2009 Z	ZJUGEND	ZJ64
2010 BA	BAJUGEND	BAJ64
2011 BB	BBJUGEND	BBJ64
2012 BC	BCJUGEND	BCJ64
2013 BD	BDJUGEND	BDJ64
2014 BE	BEJUGEND	BEJ65
2015 BF	BFJUGEND	BFJ65

BIRESPER BI: Status Aufenthaltserlaubnis
BI: Residence Status

BIO Question: Q6

Comment: The possible groups change in 2000 in QLELA and QJUGEND, such that "[3] German Citizen" is included in the original question. German citizens for the purpose of this question has been recoded to -2 (does not apply). German citizenship is however recorded in NATION$$ in $PGEN as usual.

German: Haben Sie heute eine unbefristete Aufenthaltserlaubnis bzw. Aufenthaltsberechtigung oder haben Sie eine befristete Aufenthaltserlaubnis?
"[1] Unbefristet"
"[2] Befristet"

English: Do you right now have a permanent or temporary residence permit?
"[1] Permanent"
"[2] Limited"

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P080Z
1995 L	BIOLELA	P080Z
1996 M	MLELA	MB080Z
1997 N	NLELA	NB080Z
1998 O	OLELA	OB080Z
1999 P	PLELA	PB080Z
2000 Q	QLELA	QB080Z
2001 R	RLELA	RB080Z
2002 S	SLELA	SB06
2003 T	TLELA	TB06
2004 U	ULELA	UB06
2005 V	VLELA	VB06
2006 W	WLELA	WB06
2007 X	XLELA	XB06
2008 Y	YLELA	YB06
2009 Z	ZLELA	ZB06
2010 BA	BALELA	BAB06
2011 BB	BBLELA	BBB06
2012 BC	BCLELA	BCB06
2013 BD	BDLELA	BDB07
2014 BE	BELELA	BEB07
2015 BF	BFLELA	BFB07
2013 BD	BDP_MIG	BDPM_L_15
2014 BE	BEP_MIG	BEPM_L_230
2015 BF	BFP_MIG	BFPM_L_50
2000 Q	QJUGEND	QJ58
2001 R	RJUGEND	RJ60
2002 S	SJUGEND	SJ60
2003 T	TJUGEND	TJ60
2004 U	UJUGEND	UJ60
2005 V	VJUGEND	VJ60
2006 W	WJUGEND	WJ69
2007 X	XJUGEND	XJ69
2008 Y	YJUGEND	YJ69
2009 Z	ZJUGEND	ZJ69
2010 BA	BAJUGEND	BAJ69
2011 BB	BBJUGEND	BBJ69
2012BC	BCJUGEND	BCJ71
2013 BD	BDJUGEND	BDJ71
2014 BE	BEJUGEND	BEJ73
2015 BF	BFJUGEND	BFJ73

BICAMP BI: Aufnahmelager: J/N
BI: Refugee Residence Y/N

BIO Question: Q7a

German: Haben Sie nach Ihrer Einreise zunächst in einem Aufnahmelager oder Übergangswohnheim gelebt?
"[1] Ja"
"[2] Nein"

English: After you arrived in Germany, did you live in temporary refugee/immigrant housing or residence?
"[1] Yes"
"[2] No"

See also: **BICAMP, BICAMPW, BICAMPM**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P090Z
1995 L	BIOLELA	P090Z
1996 M	MLELA	MB090Z
1997 N	NLELA	NB090Z
1998 O	OLELA	OB090Z
1999 P	PLELA	PB090Z
2000 Q	QLELA	QB090Z
2001 R	RLELA	RB090Z
2002 S	SLELA	SB0701
2003 T	TLELA	TB0701
2004 U	ULELA	UB0701
2005 V	VLELA	VB0701
2006 W	WLELA	WB0701
2007 X	XLELA	XB0701
2008 Y	YLELA	YB0701
2009 Z	ZLELA	ZB0701
2010 BA	BALELA	BAB0701
2011 BB	BBLELA	BBB0701
2012 BC	BCLELA	BCB0701
2013 BD	BDLELA	BDB0801
since 2014 BE	$LELA	n/a
2000 Q	QJUGEND	QJ5901
2001 Q	RJUGEND	RJ6101
2002 S	SJUGEND	SJ6101
2003 T	TJUGEND	TJ6101
2004 U	UJUGEND	UJ6101
2005 V	VJUGEND	VJ6101
since 2006 W	$JUGEND	n/a

BICAMPW BI: Aufnahmelager: Wochen
BI: Refugee Residence: Weeks

BIO Question: Q7b

German: Aufnahmelager: Wenn Ja, für wie lange (Wochen)?

English: Immigrant Residence: If so, then for how long (weeks)?

See also: **BICAMP, BICAMPW, BICAMPM**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P091Z
1995 L	BIOLELA	P091Z
1996 M	MLELA	MB091Z
1997 N	NLELA	NB091Z
1998 O	OLELA	OB091Z
1999 P	PLELA	PB091Z
2000 Q	QLELA	QB091Z
2001 R	RLELA	RB091Z
2002 S	SLELA	SB0702
2003 T	TLELA	TB0702
2004 U	ULELA	UB0702
2005 V	VLELA	VB0702
2006 W	WLELA	WB0702
2007 X	XLELA	XB0702
2008 Y	YLELA	YB0702
2009 Z	ZLELA	ZB0702
2010 BA	BALELA	BAB0702
2011 BB	BBLELA	BBB0702
2012 BC	BCLELA	BCB0702
2013 BD	BDLELA	BDB0802
Since 2014 BE	$LELA	n/a
2000 Q	QJUGEND	QJ5902
2001 R	RJUGEND	RJ6102
2002 S	SJUGEND	SJ6102
2003 T	TJUGEND	TJ6102
2004 U	UJUGEND	UJ6102
2005 V	VJUGEND	VJ6102
since 2006 W	$JUGEND	n/a

BICAMPM BI: Aufnahmelager: Monate
BI: Refugee Residence: Months

BIO Question: Q7c

German: Aufnahmelager: Wenn Ja, für wie lange (Monate)?

English: Immigrant Residence: If so, then for how long (months)?

See also: **BICAMP, BICAMPW, BICAMPM**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P092Z
1995 L	BIOLELA	P092Z
1996 M	MLELA	MB092Z
1997 N	NLELA	NB092Z
1998 O	OLELA	OB092Z
1999 P	PLELA	PB092Z
2000 Q	QLELA	QB092Z
2001 R	RLELA	RB092Z
2002 S	SLELA	SB0703
2003 T	TLELA	TB0703
2004 U	ULELA	UB0703
2005 V	VLELA	VB0703
2006 W	WLELA	WB0703
2007 X	XLELA	XB0703
2008 Y	YLELA	YB0703
2009 Z	ZLELA	ZB0703
2010 BA	BALELA	BAB0703
2011 BB	BBLELA	BBB0703
2012 BC	BCLELA	BCB0703
2013 BD	BDLELA	BDB0803
Since 2014 BE	$LELA	n/a
2000 Q	QJUGEND	QJ5903
2001 R	RJUGEND	RJ6103
2002 S	SJUGEND	SJ6103
2003 T	TJUGEND	TJ6103
2004 U	UJUGEND	UJ6103
2005 V	VJUGEND	VJ6103
since 2006 W	$JUGEND	n/a

BIWFAM BI: Eingereist als Familienangehoeriger

BI: Already had Family in Country

BIO Question: Q8

German: Als Sie einreisten, kamen Sie da als Familienangehöriger einer bereits in Deutschland lebenden Familie bzw. Person?
"[1] Ja"
"[2] Nein"

English: When you immigrated to Germany, was (at least one) a member of your family already living in Germany?
"[1] Yes"
"[2] No"

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P100Z
1995 L	BIOLELA	P100Z
1996 M	MLELA	MB100Z
1997 N	NLELA	NB100Z
1998 O	OLELA	OB100Z
1999 P	PLELA	PB100Z
2000 Q	QLELA	QB100Z
2001 R	RLELA	RB100Z
2002 S	SLELA	SB08
2003 T	TLELA	TB08
2004 U	ULELA	UB08
2005 V	VLELA	VB08
2006 W	WLELA	WB08
2007 X	XLELA	XB08
2008 Y	YLELA	YB08
2009 Z	ZLELA	ZB08
2010 BA	BALELA	BAB08
2011 BB	BBLELA	BBB08
2012 BC	BCLELA	BCB08
2013 BD	BDLELA	BDB09
Since 2014 BE	$LELA	n/a
2013 BD	BDP_MIG	BDPM_L01_1801 BDPM_L*_2501
2014 BE	BEP_MIG	BEPM_L_08
2015 BF	BFP_MIG	BFPM_L_09 BFPM_L*_1401
		...

Year	File	Variable
2000 Q	QJUGEND	QJ60
2001 R	RJUGEND	RJ62
2002 S	SJUGEND	SJ62
2003 T	TJUGEND	TJ62
2004 U	UJUGEND	UJ62
2005 V	VJUGEND	VJ62
since 2006 W	$JUGEND	n/a

BIFAMC BI: Vor Einreise Kontakte mit Pers.
BI: Contacts with Family in Germany

BIO Question: Q9

German: Hatten Sie vor der Einreise überhaupt Kontakte zu Verwandten oder Bekannte in Deutschland, an die Sie sich wenden konnten?
"[1] Ja"
"[2] Nein"

English: Before immigrating to Germany, did you have any contact with relatives or friends, who could possibly help you?
"[1] Yes"
"[2] No"

See also: **BIFAMC, BIFAMCL**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P110Z
1995 L	BIOLELA	P110Z
1996 M	MLELA	MB110Z
1997 N	NLELA	NB110Z
1998 O	OLELA	OB110Z
1999 P	PLELA	PB110Z
2000 Q	QLELA	QB110Z
2001 R	RLELA	RB110Z
2002 S	SLELA	SB09
2003 T	TLELA	TB09
2004 U	ULELA	UB09
2005 V	VLELA	VB09
2006 W	WLELA	WB09
2007 X	XLELA	XB09
2008 Y	YLELA	YB09
2009 Z	ZLELA	ZB09
2010 BA	BALELA	BAB09
2011 BB	BBLELA	BBB09
2012 BC	BCLELA	BCB09
2013 BD	BDLELA	BDB10
Since 2014 BE	$LELA	n/a
2000 Q	QJUGEND	QJ61
2001 R	RJUGEND	RJ63
2002 S	SJUGEND	SJ63
2003 T	TJUGEND	TJ63
2004 U	UJUGEND	UJ63
2005 V	VJUGEND	VJ63
since 2006 W	$JUGEND	n/a

BIFAMCL BI: Zuzug in Wohnort der Bekannten
BI: Moved to Same City/Town as Family

BIO Question: Q10

German: Sind Sie in den Ort in Deutschland gezogen, wo diese Verwandten bzw. Bekannten lebten?
"[1] Ja"
"[2] Nein"

English: Did you move to the same town/city in Germany where these relatives or friends lived?
"[1] Yes"
"[2] No"

See also: **BIFAMC, BIFAMCL**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P120Z
1995 L	BIOLELA	P120Z
1996 M	MLELA	MB120Z
1997 N	NLELA	NB120Z
1998 O	OLELA	OB120Z
1999 P	PLELA	PB120Z
2000 Q	QLELA	QB120Z
2001 R	RLELA	RB120Z
2002 S	SLELA	SB10
2003 T	TLELA	TB10
2004 U	ULELA	UB10
2005 V	VLELA	VB10
2006 W	WLELA	WB10
2007 X	XLELA	XB10
2008 Y	YLELA	YB10
2009 Z	ZLELA	ZB10
2010 BA	BALELA	BAB10
2011 BB	BBLELA	BBB10
2012 BC	BCLELA	BCB10
2013 BD	BDLELA	BDB11
Since 2014 BE	$LELA	n/a
2000 Q	QJUGEND	QJ62
2001 R	RJUGEND	RJ64
2002 S	SJUGEND	SJ64
2003 T	TJUGEND	TJ64
2004 U	UJUGEND	UJ64
2005 V	VJUGEND	VJ64
since 2006 W	$JUGEND	n/a

BIREASON {XE“BIREASON“\i} BI: Hauptgrund Zuzug D (seit 2014)
BI: Main Reason Migrate (since 2014)

BIO Question: Q100???

Comment: This variable comes from the IAB-SOEP Migrationsample and is available since 2014.

German: Welcher der folgenden Gründe war bei Ihnen der Hauptgrund nach Deutschland zu ziehen?
"[1] Partnerschaft"
"[2] andere familiaere Gruende"
"[3] wirts. Perspektive f. m. selbst"
"[4] wirts. Perspektive f. Kinder"
"[5] andere wirts. Gruende"
"[6] politische Gruende"
"[7] sonstige Gruende";

English Which oft he following was your main reason for moving to Germany?
"[1] Partnership"
"[2] Other family reasons"
"[3] Economic prospects for me"
"[4]Economic prospects form y children"
"[5] Other economic reasons "
"[6] Political reasons"
"[7] other";

Year	File	Variable
2014 2015	BEP_MIG BFP_MIG	BEPM_L_10001 BFPM_L_2501

BIRBETR BI: Gruende Zuzug D: Besser
BI: Reason Migrate: Better

BIO Question: Q11a

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle?
– Ich wollte ein besseres Leben haben: Besser wohnen, mehr kaufen können usw.
"[1] Besseres Leben"

English: There are many reasons to migrate to Germany. Did the following reason playa role?
– I wanted a better life. Live better, to be able to buy more etc.
"[1] Better Life"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P151Z
1995 L	BIOLELA	P151Z
1996 M	MLELA	MB151Z
1997 N	NLELA	NB151Z
1998 O	OLELA	OB151Z
1999 P	PLELA	PB151Z
2000 Q	QLELA	QB151Z
2001 R	RLELA	RB151Z
2002 S	SLELA	SB1401
2003 T	TLELA	TB1401
2004 U	ULELA	UB1401
2005 V	VLELA	VB1401
2006 W	WLELA	WB1401
2007 X	XLELA	XB1401
2008 Y	YLELA	YB1401
2009 Z	ZLELA	ZB1401
2010 BA	BALELA	BAB1401
2011 BB	BBLELA	BBB1401
2012 BC	BCLELA	BCB1401
2013 BD	BDLELA	BDB1501
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRMONEY BI: Gruende Zuzug D: Geld
BI: Reason Migrate: Money

BIO Question: Q11b

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle?
– Ich wollte arbeiten und Geld verdienen in Deutschland, um meine Familie zu unterstützen und Geld sparen.
"[1] Geld verdienen"

English: There are many reasons to migrate to Germany. Did the following reason play a role?
– I wanted to work and earn money to support my family and save money.
"[1] Earn money"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P152Z
1995 L	BIOLELA	P152Z
1996 M	MLELA	MB152Z
1997 N	NLELA	NB152Z
1998 O	OLELA	OB152Z
1999 P	PLELA	PB152Z
2000 Q	QLELA	QB152Z
2001 R	RLELA	RB152Z
2002 S	SLELA	SB1402
2003 T	TLELA	TB1402
2004 U	ULELA	UB1402
2005 V	VLELA	VB1402
2006 W	WLELA	WB1402
2007 X	XLELA	XB1402
2008 Y	YLELA	YB1402
2009 Z	ZLELA	ZB1402
2010 BA	BALELA	BAB1402
2011 BB	BBLELA	BBB1402
2012 BC	BCLELA	BCB1402
2013 BD	BDLELA	BDB1502
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRFREE BI: Gruende Zuzug D: Freiheit
BI: Reason Migrate: Freedom

BIO Question: Q11c

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle ?
– Ich wollte in der Freiheit leben.
"[1] In Freiheit leben"

English: There are many reasons to migrate to Germany. Did the following reason play a role?
– I wanted to live in freedom.
"[1] Live in freedom"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P153Z
1995 L	BIOLELA	P153Z
1996 M	MLELA	MB153Z
1997 N	NLELA	NB153Z
1998 O	OLELA	OB153Z
1999 P	PLELA	PB153Z
2000 Q	QLELA	QB153Z
2001 R	RLELA	RB153Z
2002 S	SLELA	SB1403
2003 T	TLELA	TB1403
2004 U	ULELA	UB1403
2005 V	VLELA	VB1403
2006 W	WLELA	WB1403
2007 X	XLELA	XB1403
2008 Y	YLELA	YB1403
2009 Z	ZLELA	ZB1403
2010 BA	BALELA	BAB1403
2011 BB	BBLELA	BBB1403
2012 BC	BCLELA	BCB1403
2013 BD	BDLELA	BDB1503
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRFAM BI: Gruende Zuzug D: Familie
BI: Reason Migrate: Family

BIO Question: Q11d

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle?
– Ich wollte mit meiner Familie zusammenleben (Ehepartner, Eltern, Kinder).
"[1] Mit Familie zusammen"

English: There are many reasons to migrate to Germany. Did the following reason play a role?
– I wanted to be together with my family (spouse, parents, children).
"[1] Live together with family"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P154Z
1995 L	BIOLELA	P154Z
1996 M	MLELA	MB154Z
1997 N	NLELA	NB154Z
1998 O	OLELA	OB154Z
1999 P	PLELA	PB154Z
2000 Q	QLELA	QB154Z
2001 R	RLELA	RB154Z
2002 S	SLELA	SB1404
2003 T	TLELA	TB1404
2004 U	ULELA	UB1404
2005 V	VLELA	VB1404
2006 W	WLELA	WB1404
2007 X	XLELA	XB1404
2008 Y	YLELA	YB1404
2009 Z	ZLELA	ZB1404
2010 BA	BALELA	BAB1404
2011 BB	BBLELA	BBB1404
2012 BC	BCLELA	BCB1404
2013 BD	BDLELA	BDB1504
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRPOOR BI: Gruende Zuzug D: Armut
BI: Reason Migrate: Poor

BIO Question: Q11e

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle?
– In meinem Heimatland herrschte Not und Armut.
"[1] Not/Armut in Heimat"

English: There are many reasons to migrate to Germany. Did the following reason play a role?
– In my native country there was poverty and hunger.
"[1] Poverty/Hunger at home"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P155Z
1995 L	BIOLELA	P155Z
1996 M	MLELA	MB155Z
1997 N	NLELA	NB155Z
1998 O	OLELA	OB155Z
1999 P	PLELA	PB155Z
2000 Q	QLELA	QB155Z
2001 R	RLELA	RB155Z
2002 S	SLELA	SB1405
2003 T	TLELA	TB1405
2004 U	ULELA	UB1405
2005 V	VLELA	VB1405
2006 W	WLELA	WB1405
2007 X	XLELA	XB1405
2008 Y	YLELA	YB1405
2009 Z	ZLELA	ZB1405
2010 BA	BALELA	BAB1405
2011 BB	BBLELA	BBB1405
2012 BC	BCLELA	BCB1405
2013 BD	BDLELA	BDB1505
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRWAR BI: Gruende Zuzug D: Krieg
BI: Reason Migrate: War

BIO Question*:* Q11f

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle?
– In meinem Heimatland konnte ich nicht in Sicherheit leben (Verfolgung, Krieg)
"[1] Krieg in Heimat"

English: There are many reasons to migrate to Germany. Did the following reason play a role?
– In my native country I could not live safely (Oppression, War).
"[1] War/Oppression at home"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P156Z
1995 L	BIOLELA	P156Z
1996 M	MLELA	MB156Z
1997 N	NLELA	NB156Z
1998 O	OLELA	OB156Z
1999 P	PLELA	PB156Z
2000 Q	QLELA	QB156Z
2001 R	RLELA	RB156Z
2002 S	SLELA	SB1406
2003 T	TLELA	TB1406
2004 U	ULELA	UB1406
2005 V	VLELA	VB1406
2006 W	WLELA	WB1406
2007 X	XLELA	XB1406
2008 Y	YLELA	YB1406
2009 Z	ZLELA	ZB1406
2010 BA	BALELA	BAB1406
2011 BB	BBLELA	BBB1406
2012 BC	BCLELA	BCB1406
2013 BD	BDLELA	BDB1506
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRJUST BI: Gruende Zuzug D: Einfach So
BI: Reason Migrate: Just So

BIO Question: Q11g

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle?
– Ich wollte einfach in Deutschland leben.
"[1] Einfach in D leben"

English: There are many reasons to migrate to Germany. Did the following reason play a role? – I just wanted to live in Germany.
"[1] Just wanted to live in Germany"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P157Z
1995 L	BIOLELA	P157Z
1996 M	MLELA	MB157Z
1997 N	NLELA	NB157Z
1998 O	OLELA	OB157Z
1999 P	PLELA	PB157Z
2000 Q	QLELA	QB157Z
2001 R	RLELA	RB157Z
2002 S	SLELA	SB1407
2003 T	TLELA	TB1407
2004 U	ULELA	UB1407
2005 V	VLELA	VB1407
2006 W	WLELA	WB1407
2007 X	XLELA	XB1407
2008 Y	YLELA	YB1407
2009 Z	ZLELA	ZB1407
2010 BA	BALELA	BAB1407
2011 BB	BBLELA	BBB1407
2012 BC	BCLELA	BCB1407
2013 BD	BDLELA	BDB1507
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIROTHR BI: Gruende Zuzug D: Sonstiges
BI: Reason Migrate: Other

BIO Question: Q11h

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Es gibt ja unterschiedliche Gründe, nach Deutschland zu ziehen. Welche der folgenden Gründe spielten bei Ihnen eine Rolle? – Sonstige Gründe.
"[1] Sonstige Gruende"

English: There are many reasons to migrate to Germany. Did the following reason play a role? – Other reasons.
"[1] Other reasons"

See also: **BIRBETR, BIRMONEY, BIRFREE, BIRFAM, BIRPOOR, BIRWAR, BIRJUST, BIROTHR**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P158Z
1995 L	BIOLELA	P158Z
1996 M	MLELA	MB158Z
1997 N	NLELA	NB158Z
1998 O	OLELA	OB158Z
1999 P	PLELA	PB158Z
2000 Q	QLELA	QB158Z
2001 R	RLELA	RB158Z
2002 S	SLELA	SB1408
2003 T	TLELA	TB1408
2004 U	ULELA	UB1408
2005 V	VLELA	VB1408
2006 W	WLELA	WB1408
2007 X	XLELA	XB1408
2008 Y	YLELA	YB1408
2009 Z	ZLELA	ZB1408
2010 BA	BALELA	BAB1408
2011 BB	BBLELA	BBB1408
2012 BC	BCLELA	BCB 1408
2013 BD	BDLELA	BDB1508
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIEXPR BI: Vorstellungen von D realisiert
BI: Expectations in Germany

BIO Question: Q12

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Haben sich Ihre Vorstellungen, mit denen Sie nach Deutschland gekommen sind, im grossen und ganzen erfuellt?
"[1] Ja"
"[2] Nur teilweise"
"[3] Nein, gar nicht"

English: Have your original expectations of Germany been fulfilled?
"[1] Yes"
"[2] Only partially"
"[3] No, not at all"

See also: **BIEXPR, BIEXPRLV, BIEXPRAC, BIEXPRAN**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P160Z
1995 L	BIOLELA	P160Z
1996 M	MLELA	MB160Z
1997 N	NLELA	NB160Z
1998 O	OLELA	OB160Z
1999 P	PLELA	PB160Z
2000 Q	QLELA	QB160Z
2001 R	RLELA	RB160Z
2002 S	SLELA	SB15
2003 T	TLELA	TB15
2004 U	ULELA	UB1501
2005 V	VLELA	VB1501
2006 W	WLELA	WB1501
2007 X	XLELA	XB1501
2008 Y	YLELA	YB1501
2009 Z	ZLELA	ZB1501
2010 BA	BALELA	BAB1501
2011 BB	BBLELA	BBB1501
2012 BC	BCLELA	BCB1501
2013 BD	BDLELA	BDB16
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIEXPRLV BI: Eigene Wohnung finden
BI: Expectations: Find Apt

BIO Question: Q13a

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Auf welchen Gebieten war es leichter oder schwerer, als sie vorher gedacht hatten? – Eine eigene Wohnung zu finden.
"[1] Schwerer"
"[2] Wie erwartet"
"[3] Leichter"
"[4] TNZ"

English: In which areas was it harder or easier than you expected?
– to find your own apartment/housing.
"[1] Harder"
"[2] Just as expected"
"[3] Easier"
"[4] Not applicable"

See also: **BIEXPR, BIEXPRLV, BIEXPRAC, BIEXPRAN**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P171Z
1995 L	BIOLELA	P171Z
1996 M	MLELA	MB171Z
1997 N	NLELA	NB171Z
1998 O	OLELA	OB171Z
1999 P	PLELA	PB171Z
2000 Q	QLELA	QB171Z
2001 R	RLELA	RB171Z
2002 S	SLELA	SB1601
2003 T	TLELA	TB1601
2004 U	ULELA	UB1502
2005 V	VLELA	VB1502
2006 W	WLELA	WB1502
2007 X	XLELA	XB1502
2008 Y	YLELA	YB1502
2009 Z	ZLELA	ZB1502
2010 BA	BALELA	BAB1502
2011 BB	BBLELA	BBB1502
2012 BC	BCLELA	BCB1502
2013 BD	BDLELA	BDB1701
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIEXPRAC BI: Von Arbeitskollegen akzeptiert
BI: Expectations: Accepted by Coworker

BIO Question: Q13b

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Auf welchen Gebieten war es leichter oder schwerer, als sie vorher gedacht hatten? – Von den Arbeitskollegen akzeptiert zu werden.
"[1] Schwerer"
"[2] Wie erwartet"
"[3] Leichter"
"[4] TNZ"

English: In which areas was it harder or easier than you expected?
– to be accepted by your collegues at work.
"[1] Harder"
"[2] Just as expected"
"[3] Easier"
"[4] Not applicable"

See also: **BIEXPR, BIEXPRLV, BIEXPRAC, BIEXPRAN**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P172Z
1995 L	BIOLELA	P172Z
1996 M	MLELA	MB172Z
1997 N	NLELA	NB172Z
1998 O	OLELA	OB172Z
1999 P	PLELA	PB172Z
2000 Q	QLELA	QB172Z
2001 R	RLELA	RB172Z
2002 S	SLELA	SB1602
2003 T	TLELA	TB1602
2004 U	ULELA	UB1503
2005 V	VLELA	VB1503
2006 W	WLELA	WB1503
2007 X	XLELA	XB1503
2008 Y	YLELA	YB1503
2009 Z	ZLELA	ZB1503
2010 BA	BALELA	BAB1503
2011 BB	BBLELA	BBB1503
2012 BC	BCLELA	BCB1503
2013 BD	BDLELA	BDB1702
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIEXPRAN BI: Von Nachbarn akzeptiert
BI: Expectations: Accepted by Neighbor

BIO Question: Q13c

Comment: This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Auf welchen Gebieten war es leichter oder schwerer, als sie vorher gedacht hatten? – Von den Nachbarn akzeptiert zu werden.
"[1] Schwerer"
"[2] Wie erwartet"
"[3] Leichter"
"[4] TNZ"

English: In which areas was it harder or easier than you expected?
– To be accepted by your neighbors.
"[1] Harder"
"[2] Just as expected"
"[3] Easier"
"[4] Not applicable"

See also: **BIEXPR, BIEXPRLV, BIEXPRAC, BIEXPRAN**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P173Z
1995 L	BIOLELA	P173Z
1996 M	MLELA	MB173Z
1997 N	NLELA	NB173Z
1998 O	OLELA	OB173Z
1999 P	PLELA	PB173Z
2000 Q	QLELA	QB173Z
2001 R	RLELA	RB173Z
2002 S	SLELA	SB1603
2003 T	TLELA	TB1603
2004 U	ULELA	UB1504
2005 V	VLELA	VB1504
2006 W	WLELA	WB1504
2007 X	XLELA	XB1504
2008 Y	YLELA	YB1504
2009 Z	ZLELA	ZB1504
2010 BA	BALELA	BAB1504
2011 BB	BBLELA	BBB1504
2012 BC	BCLELA	BCB1504
2013 BD	BDLELA	BDB1703
Since 2014 BE	$LELA	n/a
2000 --	$JUGEND	n/a

BIRELH BI: Familienmitglieder im Heimatland oder außerhalb Deutschlands
BI: Family in the home country or abroad

BIO Question: Q14

Comment: From 2001 onwards the variable is only identified by the parents in $LELA and missing for $JUGEND. A distinction between abroad and home country is not consistently possible over time.

German: Haben Sie in dem Land, aus dem Sie kommen bzw. aus dem Ihre Familie kommt, noch Familienangehörige oder andere Ihnen nahstehende Menschen?
"[1] Ja"
"[2] Nein"

English: Do you have family members or close friends in the home country you (or your family) come from?
"[1] Yes"
"[2] No"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P18Z
1995 L	BIOLELA	P18Z
1996 M	MLELA	MB18Z
1997 N	NLELA	NB18Z
1998 O	OLELA	OB18Z
1999 P	PLELA	PB18Z
2000 Q	QLELA	QB18Z
2001 -	$LELA	n/a

BIRELHP BI: Im Ausland: Eltern

BI: Family Abroad: Parents

BIO Question: Q15a

Comment: This variable is used to identify any relatives starting 2001 for $LELA and missing for all $JUGEND starting 2001.

German: Personen in der Heimat: Was für Personen sind das? Eltern?
"[1] Eltern"

English: Persons in Native Country: Who are they? Parents?
"[1] Parents"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI, BIRELHS2, BIRELHC2, BIRELHSP, BIRELHC, BIRELH**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P191Z
1995 L	BIOLELA	P191Z
1996 M	MLELA	MB191Z
1997 N	NLELA	NB191Z
1998 O	OLELA	OB191Z
1999 P	PLELA	PB191Z
2000 Q	QLELA	QB191Z
2001 R	RLELA	RB0703V RB0705M
2002 S	SLELA	SB2101 SB2102
2003 T	TLELA	TB2101 TB2102
2004 U	ULELA	UB2101 UB2102
2005 V	VLELA	VB2101 VB2102
2006 W	WLELA	WB2101 WB2102
2007 X	XLELA	XB2101 XB2102
2008 Y	YLELA	YB2101 YB2102
2009 Z	ZLELA	ZB2101 ZB2102
2010 BA	BALELA	BAB2101 BAB2102
2011 BB	BBLELA	BBB22V01 BBB22M01
2012 BC	BCLELA	BCB2201 BCB2202
2013 BD	BDLELA	BDB2301 BDB2302
2014 BE	BELELA	BEB2701 BEB2702
2015 BF	BFLELA	BFB2701 BFB2702
2013 BD	BDP_MIG	BDPML_L_5901 BDPML_L_5902
2014 BE	BEP_MIG	BEPM_L_3501 BEPM_L_3502
2015 BF	BFP_MIG	BFPM_L_11101 BFPM_L_11102 ...

Year	File	Variable
2000 Q 2001 –	QJUGEND $JUGEND	QJ6701 n/a

BIRELHGP BI: In Heimat: Grosseltern
BI: Family Abroad: Grandparents

BIO Question: Q15b

Comment: This variable is not defined for new entrants starting 2001 (R).

German: Personen in der Heimat: Was für Personen sind das? Grosseltern?
"[1] Grosseltern"

English: Persons in Native Country: Who are they? Grandparents?
"[1] Grandparents"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P192Z
1995 L	BIOLELA	P192Z
1996 M	MLELA	MB192Z
1997 N	NLELA	NB192Z
1998 O	OLELA	OB192Z
1999 P	PLELA	PB192Z
2000 Q	QLELA	QB192Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ6702
2001 --	$JUGEND	n/a

BIRELHC BI: In Heimat: Kinder
BI: Family Abroad: Children

BIO Question: Q15c

Comment: This variable is not defined for new entrants starting 2001 (R).

German: Personen in der Heimat: Was für Personen sind das? Kinder?
"[1] Kinder"

English: Persons in Native Country: Who are they? Children?
"[1] Children"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI, BIRELHS2, BIRELHC2, BIRELHSP, BIRELHC, BIRELHP**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P193Z
1995 L	BIOLELA	P193Z
1996 M	MLELA	MB193Z
1997 N	NLELA	NB193Z
1998 O	OLELA	OB193Z
1999 P	PLELA	PB193Z
2000 Q	QLELA	QB193Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ6703
2001 --	$JUGEND	n/a

BIRELHBS BI: In Heimat: Bruder, Schwester
BI: Family Abroad: Brother/Sister

BIO Question: Q15d

Comment: This variable is not defined for new entrants starting 2001 (R).

German: Personen in der Heimat: Was für Personen sind das? Bruder/Schwester?
"[1] Bruder/Schwester"

English: Persons in Native Country: Who are they? Brother/Sister?
"[1] Brother/Sister"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P194Z
1995 L	BIOLELA	P194Z
1996 M	MLELA	MB194Z
1997 N	NLELA	NB194Z
1998 O	OLELA	OB194Z
1999 P	PLELA	PB194Z
2000 Q	QLELA	QB194Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ6704
2001 --	$JUGEND	n/a

BIRELHDR BI: In Heimat: Entferntere Verwandte
BI: Family Abroad: Distant Relatives

BIO Question: Q15e

Comment: This variable is not defined for new entrants starting 2001 (R).

German: Personen in der Heimat: Was für Personen sind das? Entferntere Verwandte?
"[1] Entferntere Verwandte"

English: Persons in Native Country: Who are they? Distant Relatives?
"[1] Distant Relatives"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P195Z
1995 L	BIOLELA	P195Z
1996 M	MLELA	MB195Z
1997 N	NLELA	NB195Z
1998 O	OLELA	OB195Z
1999 P	PLELA	PB195Z
2000 Q	QLELA	QB195Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ6705
2001 --	$JUGEND	n/a

BIRELHSP BI: In Heimat: Ehepartner, Verlobte(r)
BI: Family Abroad: Spouse

BIO Question: Q15f

Comment: This variable is not defined for new entrants starting 2001 (R), but for entrants from the IAB-SOEP Migrationsample starting in 2013. Here it is only asked whether the partner lives in native country.

German: Personen in der Heimat:
Was für Personen sind das? Ehepartner / Verlobte(r)?
"[1] Ehepartner/Verlobte(r) "

English: Persons in Native Country: Who are they? Spouse / Fiance(e)?
"[1] Spouse/Fiance(e) "

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI, BIRELHS2, BIRELHC2, BIRELHSP, BIRELHC, BIRELHP**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P196Z
1995 L	BIOLELA	P196Z
1996 M	MLELA	MB196Z
1997 N	NLELA	NB196Z
1998 O	OLELA	OB196Z
1999 P	PLELA	PB196Z
2000 Q	QLELA	QB196Z
2001 --	$LELA	n/a
2013 BD	BDP_MIG	BDPM_P_157
		BDPM_P_15301
		BDPM_P_154
2014 BE	BEP_MIG	BEPM_P_92
		BEPM_LA_8801
		BEPM_P_90
2015 BF	BFP_MIG	BFP_P_91
		BFP_1a_16801
		BFP_P_85
2000 Q	QJUGEND	QJ6706
2001 --	$JUGEND	n/a

BIRELHFR BI: In Heimat: Persoenliche Bekannte
BI: Family Abroad: Friends

BIO Question: Q15g

Comment: This variable is not defined for new entrants starting 2001 (R).

German: Personen in der Heimat: Was für Personen sind das?
Bekannte, Freunde ?
"[1] Persönliche Bekannte"

English: Persons in Native Country: Who are they? Friends?
"[1] Friends"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P197Z
1995 L	BIOLELA	P197Z
1996 M	MLELA	MB197Z
1997 N	NLELA	NB197Z
1998 O	OLELA	OB197Z
1999 P	PLELA	PB197Z
2000 Q	QLELA	QB197Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ6707
2001 --	$JUGEND	n/a

BIRELHMI BI: Personen gern nach Dt. holen?
BI: Persons abroad bring to Germany

BIO Question: Q16

Comment: This variable is not defined for new entrants starting 2001 (R).

German: Personen in der Heimat: Gibt es darunter Personen, die auch nach Deutschland kommen wollen bzw. die Sie gerne nachholen möchten?
"[1] Ja"
"[2] Nein"

English: Persons in Native Country: Among those mentioned above, do some want to come to Germany, or would you like them to come to Germany?
"[1] Yes"
"[2] No"

See also: **BIRELH, BIRELHP, BIRELHGP, BIRELHC, BIRELHBS, BIRELHDR, BIRELHSP, BIRELHFR, BIRELHMI**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P200Z
1995 L	BIOLELA	P200Z
1996 M	MLELA	MB200Z
1997 N	NLELA	NB200Z
1998 O	OLELA	OB200Z
1999 P	PLELA	PB200Z
2000 Q	QLELA	QB200Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ68
2001 --	$JUGEND	n/a

BIRELHS2 BI: Ehepartner in Deutschland
BI: Spouse in Germany

BIO Question: Q15f

Comment: This variable is not defined for new entrants starting 1996 (M) and missing for all $JUGEND.

German: Lebt Ihr Ehepartner in Deutschland?
"[1] D hier im HH"
"[2] D nicht im HH"
"[3] Nicht in D"

English: Does your spouse live in Germany?
"[1] Yes, here in the HH"
"[2] Yes, but NOT with me in HH"
"[3] Not in Germany"

See also: **BIRELHS2, BIRELHC2, BIRELHSP, BIRELHC, BIRELHP**

Year	File	Variable
1984 A	APAUSL	AP58A02
1985 B	BPAUSL	n/a
1986 C	CPAUSL	CP90A01
1987 D	DPAUSL	DP92A01
1988 E	EPAUSL	EP85A01
1989 F	FPAUSL	FP102A01
1990 G	GPAUSL	GP102A01
1991 H	HPAUSL	HP102A01
1992 I	IPAUSL	IP102A01
1993 J	JPAUSL	JP102A01
1994 K	KPAUSL	KP102A01
1995 L	LPAUSL	LP110A01
1996 --	$P	n/a
1984 --	BIOLELA	n/a
2000 --	$JUGEND	n/a

BIRELHC2 BI: Kinder unter 18 J. nicht in Deutschland
BI: Underage Children not in Germany

BIO Question: Q15c

German: Haben Sie Kinder unter 18 Jahren, die nicht in Deutschland leben?
"[1] Ja"
"[2] Nein"

English: Do you have children under 18, who do not live in Germany?
"[1] Yes"
"[2] No"

See also: **BIRELHS2, BIRELHC2, BIRELHSP, BIRELHC, BIRELHP**

Year	File	Variable
1984 A	APAUSL	AP66A01
1985 B	BPAUSL	BP95A01
1986 C	CPAUSL	CP86A01
1987 D	DPAUSL	DP88A01
1988 E	EPAUSL	n/a
1989 F	FPAUSL	FP98A01
1990 G	GPAUSL	n/a
1991 H	HPAUSL	HP98A01
1992 I	IPAUSL	n/a
1993 J	JPAUSL	JP98A01
1994 K	KPAUSL	n/a
1995 L	LPAUSL	LP106A01
1996 M	MP	MP7406
1997 N	NP	NP111A04
1998 O	OP	n/a
1999 P	PP	PP12904
2000 --	$PAUSL	n/a
1984- A-	BIOLELA	n/a
2000 --	$JUGEND	n/a

BIGOBACK BI: Rueckkehr Heimat (ab 1994)
BI: Go back home?

BIO Question: Q17

Comment: The question BIGOBACK (using BIOLELA) asks whether one intends to **return** home to the native country whereas BISTAY (using $PAUSL) asks wether one intends to **stay** in Germany. The wording and the answer possibilities are different in both questions. Further, there is no particular reason to believe that the two variables even are consistent. Starting 2001, this is not defined for new entrants.

German: Planen Sie selbst, in Ihr Herkunftsland wieder zurückzukehren?
"[1] Ja, ganz sicher"
"[2] Ja, wahrscheinlich"
"[3] Eher unwahrscheinlich"
"[4] Nein, sicher nicht"

English: Are you planning to go back to live in your native country?
"[1] Yes, certainly"
"[2] Yes, probably"
"[3] Probably not"
"[4] No, Certainly not"

See also: **BIGOBACK, BISTAY, BISTAYY**

Year	File	Variable
1984-93 A-J	BIOLELA	n/a
1994 K	BIOLELA	P230Z
1995 L	BIOLELA	P230Z
1996 M	MLELA	MB230Z
1997 N	NLELA	NB230Z
1998 O	OLELA	OB230Z
1999 P	PLELA	PB230Z
2000 Q	QLELA	QB230Z
2001 --	$LELA	n/a
2000 Q	QJUGEND	QJ69
2001 --	$JUGEND	n/a

BISTAY BI: Wunsch in D zu bleiben
BI: Desire to Stay in Germany

BIO Question: Q17

Comment: The question BIGOBACK (using BIOLELA) asks whether one intends to **return** home to the native country whereas BISTAY (using $PAUSL) asks wether one intends to **stay** in Germany. The wording and the answer possibilities are different in both questions. Further, there is no particular reason to believe that the two variables even are consistent. This variable is not defined for youths answering the $JUGEND biography questionnaire.

German: Wie lange wollen Sie in Deutschland bleiben?
"[1] Kehre innerhalb eines Jahres zurück"
"[2] Einige Jahre und zwar..."
"[3] Für immer in D bleiben"

English: How long would you like to stay in Germany**?**
"[1] Go back within 12 months"
"[2] Several years, specifically..."
"[3] Always stay in Germany"

See also: **BIGOBACK, BISTAY, BISTAYY**

Year	File	Variable
1984 A	APAUSL	AP67A01
1985 B	BPAUSL	BP96A01
1986 C	CPAUSL	CP87A01
1987 D	DPAUSL	DP89A01
1988 E	EPAUSL	EP77A01
1989 F	FPAUSL	FP99A01
1990 G	GPAUSL	GP96A01
1991 H	HPAUSL	HP99A01
1992 I	IPAUSL	IP99A01
1993 J	JPAUSL	JP99A01
1994 K	KPAUSL	KP96A01
1995 L	LPAUSL	LP107A01
1996 M	MP	MP101A01 MP100A
1997 N	NP	NP109A01 NP108A
1998 O	OP	OP11401 OP113
1999 P	PP	PP12601 PP125
2000 Q	QP	QP13401 QP133
2001 R	RP	RP12701 RP126
2002 S	SP	SP12601 SP125
2003 T	TP	TP13301 TP132
2004 U	UP	UP13501 UP134
2005 V	VP	VP14601 VP145
2006 W	WP	WP13601 WP135
2007 X	XP	XP14601 XP145
2008 Y	YP	YP14501 YP144
2009 Z	ZP	ZP14401 ZP143
2010 BA	BAP	BAP14601 BAP145
2011 BB	BBP	BBP14701 BBP146
2012 BC	BCP	n/a
2013 BD	BDP	BDP15001 BDP149
2014 BE	BEP	n/a
2015 BF	BFP	BFP16401 BFP163
2013 BD	BDP_MIG	BDPM_P_3701BDPM_P_36 n/a
2014 BE	BEP_MIG	BFPM_P_5201 BFPM_P_51
2015 BF	BFP_MIG	
2000 --	$JUGEND	n/a

BISTAYY	BI: Dauer des geplanten Aufenthalts BI: Years Desired to Stay in Germany
BIO Question:	Q17
Comment:	This variable is not defined for youths answering the $JUGEND biography questionnaire.
German:	Wie lange wollen Sie in Deutschland bleiben? Einige Jahre und zwar...
English:	How long would you like to stay in Germany? Several years, specifically...
See also:	**BIGOBACK, BISTAY, BISTAYY**

Year	File	Variable
1984 A	APAUSL	AP67A02
1985 B	BPAUSL	BP96A02
1986 C	CPAUSL	CP87A02
1987 D	DPAUSL	DP89A02
1988 E	EPAUSL	EP77A02
1989 F	FPAUSL	FP99A02
1990 G	GPAUSL	GP96A02
1991 H	HPAUSL	HP99A02
1992 I	IPAUSL	IP99A02
1993 J	JPAUSL	JP99A02
1994 K	KPAUSL	KP96A02
1995 L	LPAUSL	LP107A02
1996 M	MP	MP101A02
1997 N	NP	NP109A02
1998 O	OP	OP11402
1999 P	PP	PP12602
2000 Q	QP	QP13402
2001 R	RP	RP12702
2002 S	SP	SP12602
2003 T	TP	TP13302
2004 U	UP	UP13502
2005 V	VP	VP14602
2006 W	WP	WP13602
2007 X	XP	XP14602
2008 Y	YP	YP14502
2009 Z	ZP	ZP14402
2010 BA	BAP	BAP14602
2011 BB	BBP	BBP14702
2012 BC	BCP	n/a
2013 BD	BDP	BDP15002
2014 BE	BEP	n/a
2015 BF	BFP	BFP16402
2013 BD	BDP_MIG	BDPM_P_3702
2014 BE	BEP_MIG	n/a
2015 BF	BFP_MIG	BFPM_P_5202
2000 --	$JUGEND	n/a

BISCGER	BI: In Dt. Schule besucht? BI: Attended School in Germany
BIO Question:	Q18
Comment:	This question asks only if one has **ever** attended a (primary/secondary) school in Germany, but does **not** ask whether one received a certificate/diploma, such as the generated variable `$PSBIL` in the file **`$PGEN`**. Since 2014 and in the IAB-SOEP Migrationsample the question is about the last attended school.
German:	Haben Sie in Deutschland eine Schule besucht? "[1] Ja" "[2] Nein"
English:	Did you attend school in Germany? "[1] Yes" "[2] No"
See also:	**BISCGER, BISCGRAD, BISCGERC, BISCGC, BISCGCF, BISCGCFN**

Year	File	Variable
1984 A	APAUSL	AP06A01
1985 B	BPAUSL	BP100A01
1986 C	CPAUSL	CP100B01
1987 D	DPAUSL	DP97A01
1988 E	EPAUSL	EP90A01
1989 F	FPAUSL	FP107A
1990 G	GPAUSL	GP107A
1991 H	HPAUSL	HP107A
1992 I	IPAUSL	IP107A
1993 J	JPAUSL	JP107A
1984-93 A-J	BIOLELA	B46A
1994 K	BIOLELA	P280Z
1995 L	BIOLELA	P280Z
1996 M	MLELA	MB280Z
1997 N	NLELA	NB280Z
1998 O	OLELA	OB280Z
1999 P	PLELA	PB280Z
2000 Q	QLELA	QB280Z
2001 R	RLELA	RB280Z
2002 S	SLELA	SB11
2003 T	TLELA	TB11
2004 U	ULELA	UB11
2005 V	VLELA	VB11
2006 W	WLELA	WB11
2007 X	XLELA	XB11
2008 Y	YLELA	YB11
2009 Z	ZLELA	ZB11
2010-2012 B$	B$LELA	B$B11
2013 BD	BDLELA	BDB12
2014 BE	BELELA	BEB45
2015 BF	BFLELA	BFB47
2013 BD	BDP_MIG	BDPM_L_70
2014 BE	BEP_MIG	BEPM_L_61
2015 BF	BFP_MIG	BFPM_L_136
2000 Q	QJUGEND	QJ63
2001 R	RJUGEND	RJ65
2002 S	SJUGEND	SJ65
2003 T	TJUGEND	TJ65
2004 U	UJUGEND	UJ65
2005 V	VJUGEND	VJ65
2006 --	$JUGEND	n/a

BISCGRAD	BI: In welche Klasse in dt. Schule BI: Which Grade School
BIO Question:	Q19
Comment:	The question here is **not** on the highest schooling achieved, but rather what was the grade or class when one **first** came to Germany.
German:	In welche Klasse sind Sie in Deutschland in die Schule gekommen?
English:	Which class/grade did you attend when you came to Germany?
See also:	**BISCGER, BISCGRAD, BISCGERC, BISCGC, BISCGCF, BISCGCFN**

Year	File	Variable
1984 A	APAUSL	n/a
1985 B	BPAUSL	n/a
1986 C	CPAUSL	n/a
1987 D	DPAUSL	n/a
1988 E	EPAUSL	n/a
1989 F	FPAUSL	FP108A
1990 G	GPAUSL	GP108A
1991 H	HPAUSL	HP108A
1992 I	IPAUSL	IP108A
1993 J	JPAUSL	JP108A
1984-93 A-J	BIOLELA	B47A
1994 K	BIOLELA	P290Z
1995 L	BIOLELA	P290Z
1996 M	MLELA	MB290Z
1997 N	NLELA	NB290Z
1998 O	OLELA	OB290Z
1999 P	PLELA	PB290Z
2000 Q	QLELA	QB290Z
2001 R	RLELA	RB290Z
2002 S	SLELA	SB12
2003 T	TLELA	TB12
2004 U	ULELA	UB12
2005 V	VLELA	VB12
2006 W	WLELA	WB12
2007 X	XLELA	XB12
2008 Y	YLELA	YB12
2009 Z	ZLELA	ZB12
2010-2012 B$	B$LELA	B$B12
2013 BD	BDLELA	BDB13
Since 2014 BE	BELELA	n/a
2000 Q	QJUGEND	QJ64
2001 R	RJUGEND	RJ6601
2002 S	SJUGEND	SJ6601
2003 T	TJUGEND	TJ6601
2004 U	UJUGEND	UJ6601
2005 V	VJUGEND	VJ6601
2006 --	$JUGEND	n/a

BISCGERC BI: Besuch spezieller Vorbereitung
BI: Attended Special Foreigner Prep Class

BIO Question: Q20

German: Haben Sie vorher eine spezielle Vorbereitungsklasse für Ausländer in Deutschland besucht?
"[1] Ja"
"[2] Nein"

English: Did you attend a special preparation class for foreigners in Germany?
"[1] Yes"
"[2] No"

See also: **BISCGER, BISCGRAD, BISCGERC, BISCGC, BISCGCF, BISCGCFN**

Year	File	Variable
1984 A	APAUSL	n/a
1985 B	BPAUSL	n/a
1986 C	CPAUSL	n/a
1987 D	DPAUSL	n/a
1988 E	EPAUSL	n/a
1989 F	FPAUSL	FP109A
1990 G	GPAUSL	GP109A
1991 H	HPAUSL	HP109A
1992 I	IPAUSL	IP109A
1993 J	JPAUSL	JP109A
1984-93 A-J	BIOLELA	B48A
1994 K	BIOLELA	P300Z
1995 L	BIOLELA	P300Z
1996 M	MLELA	MB48A
1997 N	NLELA	NB48A
1998 O	OLELA	OB48A
1999 P	PLELA	PB48A
2000 Q	QLELA	QB48A
2001 R	RLELA	RB48A
2002 S	SLELA	SB13
2003 T	TLELA	TB13
2004 U	ULELA	UB13
2005 V	VLELA	VB13
2006 W	WLELA	WB13
2007 X	XLELA	XB13
2008 Y	YLELA	YB13
2009 Z	ZLELA	ZB13
2010-2012 B$	B$LELA	B$B13
2013 BD	BDLELA	BDB14
Since 2014 BE	BELELA	n/a
2000 Q	QJUGEND	QJ65
2001 R	RJUGEND	RJ6602
2002 S	SJUGEND	SJ6602
2003 T	TJUGEND	TJ6602
2004 U	UJUGEND	UJ6602
2005 V	VJUGEND	VJ6602
2006--	$JUGEND	n/a

BISCGC BI: Auch dt. Schueler in Schulklasse
BI: Also German Pupils in Class

BIO Question: Q21a

Comment: This variable is not defined for new entrants starting 2000 (Q) and $JUGEND.

German: Gab es in der Schulklasse, die Sie zuletzt in Deutschland besucht haben, auch deutsche Schüler?
"[1] Ja"
"[2] Nein"

English: Were there also German children present in the class you last attended?
"[1] Yes"
"[2] No"

See also: **BISCGER, BISCGRAD, BISCGERC, BISCGC, BISCGCF, BISCGCFN**

Year	File	Variable
1984 A	APAUSL	n/a
1985 B	BPAUSL	n/a
1986 C	CPAUSL	n/a
1987 D	DPAUSL	n/a
1988 E	EPAUSL	n/a
1989 F	FPAUSL	FP110A01
1990 G	GPAUSL	GP110A01
1991 H	HPAUSL	HP110A01
1992 I	IPAUSL	IP110A01
1993 J	JPAUSL	JP110A01
1984-93 A-J	BIOLELA	B49A
1994 K	BIOLELA	B49A
1995 L	BIOLELA	B49A
1996 M	MLELA	MB49A
1997 N	NLELA	NB49A
1998 O	OLELA	OB49A
1999 P	PLELA	PB49A
2000 --	$LELA	n/a
2000 --	$JUGEND	n/a

BISCGCF BI: Wieviel Mitschueler Auslaender
BI: How many Pupils foreign

BIO Question: Q21b

German: Wie viele Ihrer Mitschueler waren Auslaender?
"[1] Die meisten"
"[2] Etwa 1/2"
"[3] Etwa 1/4"
"[4] Weniger als 1/4"
"[5] Ausser mir niemand"

English: How many of your fellow students were foreigners?
"[1] Most of them"
"[2] Around 1/2"
"[3] Around 1/4"
"[4] Less than 1/4"
"[5] I was only one"

See also: **BISCGER, BISCGRAD, BISCGERC, BISCGC, BISCGCF, BISCGCFN**

Year	File	Variable
1984-88 A-E	$PAUSL	n/a
1989 F	FPAUSL	FP110A02
1990 G	GPAUSL	GP110A02
1991 H	HPAUSL	HP110A02
1992 I	IPAUSL	IP110A02
1993 J	JPAUSL	JP110A02
1984-93 A-J	BIOLELA	B50A
1994 K	BIOLELA	B50A
1995 L	BIOLELA	B50A
1996 M	MLELA	MB50A
1997 N	NLELA	NB50A
1998 O	OLELA	OB50A
1999 P	PLELA	PB50A
2000 Q	QLELA	QB50A
2001 R	RLELA	RB50A
2002 S	SLELA	SB43
2003 T	TLELA	TB43
2004 U	ULELA	UB43
2005 V	VLELA	VB43
2006 W	WLELA	WB43
2007 X	XLELA	XB43
2008 Y	YLELA	YB43
2009 Z	ZLELA	ZB43
2010 BA	BALELA	BAB43
2011 BB	BBLELA	BBB43
2012 BC	BCLELA	BCB44
2013 BD	BDLELA	BDB46
2014 BE	BELELA	BEB50
2015 BF	BFLELA	BFB52
2013 BD	BDP_MIG	BDPM_L_75
2014 BE	BEP_MIG	BEPM_L_66
2015 BF	BFP_MIG	BFPM_L_143
2000 Q	QJUGEND	n/a
2001 R	RJUGEND	RJ43
2002 S	SJUGEND	SJ43
2003 T	TJUGEND	TJ43
2004 U	UJUGEND	UJ43
2005 V	VJUGEND	VJ43
2006 W	WJUGEND	WJ45
2007 X	XJUGEND	XJ45
2008 Y	YJUGEND	YJ45
2009 Z	ZJUGEND	ZJ45
2010-2015 B$	B$JUGEND	B$J45

BISCGCFN BI: Eine oder mehrere Nationalitaet
BI: Mix of Nationalities in Class

BIO Question: Q21c

Comment: This variable is not defined for new entrants starting 2000 (Q) and $JUGEND.

German: Gab es in dieser Klasse nur Schüler Ihrer Nationalität oder waren verschieden Nationalitäten gemischt?
"[1] Nur meine Nationalitaet"
"[2] Gemischt"

English: Were there only children of your nationality, or were the nationalites mixed?
"[1] Only my nationality"
"[2] Mixed"

See also: **BISCGER, BISCGRAD, BISCGERC, BISCGC, BISCGCF, BISCGCFN**

Year	File	Variable
1984 A	APAUSL	n/a
1985 B	BPAUSL	n/a
1986 C	CPAUSL	n/a
1987 D	DPAUSL	n/a
1988 E	EPAUSL	n/a
1989 F	FPAUSL	FP110A03
1990 G	GPAUSL	GP110A03
1991 H	HPAUSL	HP110A03
1992 I	IPAUSL	IP110A03
1993 J	JPAUSL	JP110A03
1984-93 A-J	BIOLELA	B51A
1994 K	BIOLELA	B51A
1995 L	BIOLELA	B51A
1996 M	MLELA	MB51A
1997 N	NLELA	NB51A
1998 O	OLELA	OB51A
1999 P	PLELA	PB51A
2000 --	$LELA	n/a
2000 --	$JUGEND	n/a

15 BIORESID: Variables on Occupancy and Second Residence

by Marco Giesselmann, Mila Staneva and Tabea Naujoks[1]

In 1994 questions with a focus on occupancy were introduced to the Biographical Questionnaire asking for the duration of residence in the current dwelling and any second residence. Questions on the second residence were also asked before 1994, but those were collected in the (blue version of the) Individual Questionnaire and therefore the corresponding variables are part of the $P files. The information surveyed in the Biographical Questionnaire is stored in the new file BIORESID.

The variables of BIORESID are based on following questions:

Question I

When did you move into this home?

☐☐☐☐

Year

Question II

Do you have another home in which you yourself reside or spend your vacation? [2]
(0) No (1) Yes => continue with question

Is this second home in western Germany (including West Berlin), in eastern Germany (including East Berlin) or abroad?
Germany[3]
Western Germany[3]
Eastern Germany[3]
Abroad[3]

Which home is your main residence?
This one
The other one
I use both about the same

[1] Replaces earlier versions by Henning Lohmann and Sven Witzke/ Jürgen Schupp and Michael Frühling / Thorsten Schneider.
[2] In the years 1994 and 1995 the question was "*Do you have another home, in Germany, in which you yourself reside in?*"
[3] The new category "abroad" was added in 1996.
[3] From 1994 to 2013 Western Germany and Eastern Germany were treated as two different categories in the Question. Since 2014, these two categories are combined in one Variable ("Germany").
[4]Requested only from 1994 to 2013

From which residence do you usually go to work?[4] From this one From the other one Not applicable

Since 2014, Western

15.1 Sources of Variables

The information for the years 1994 and 1995 stem from the file BIOLELA. Information for later years are taken from the wave-specific data sets $LELA.

In principle, SOEP respondents answer the Biography Questionnaire only once, so every person has only one record with wave-specific information in BIORESID. For fieldwork-related reasons, very few people have answered the Biography Questionnaire twice. For these, the first interview is taken as relevant for BIORESID. Further cases are dropped if their information stems from an interview completed before 1994.

15.2 Population of Interest

The BIORESID dataset as of wave 2015 contains information on 46,457 individuals, stemming from samples A-M. The data set is supplemented every year by new respondents filling in the supplementary Biography Questionnaire.

Table 1: Survey Year in BIORESID

Survey Year	n
1994	993
1995	1,075
1996	471
1997	480
1998	415
1999	2,039
2000	243
2001	8,816
2002	508
2003	2,319
2004	449
2005	299
2006	223
2007	2,232
2008	336
2009	196
2010	9,845
2011	6,860
2012	2,993
2013	398
2014	277
2015	5,050
Total	46,457

Status: up to wave BF (2015)

Source: SOEP v32, doi: 10.5684/soep.v32

Table 2: Samples in BIORESID

Sample	n
A Germans (West)	2,131
B Foreigners (West)	738
C Germans (East)	1,430
D Immigrants 1984-93	1,399
E Supplement 1998	1,871
F Innovation 2000	9,844
G High Income 2002	2,262
H Supplement 2006	2,198
I Incentivation 2009	1,870
J Supplement 2011	5,409
K Supplement 2012	2,579
L Family Types	9,956
M Migration Sample	4,770

Source: SOEP v32, doi: 10.5684/soep.v32

The information in BIORESID is treated as time-invariant. Although, in principle, it is possible to update the information on occupancy for some individuals on the basis of more recent information, we abstain from doing so for selectivity reasons.

15.3 Variable List of the Data Set BIORESID

Table 3: Description of the Data Set BIORESID

Variable Name	Content of the Variable
	Entries for Surveyed Person
HHNR	Original household number (invariant)
HHNRAKT	Current wave HH number (wave of biography interview)
PERSNR	Never changing person ID
SYEAR	Survey year
	Occupancy
BRMOVEIN	Year person moved in current dwelling
	Second Residence
BRSECHOM	Having a second residence
BRSECREG	Region of second residence
BRSECUSE	Use of second residence
BRSECWOR	Second residence at place of work
	Specification of Interview Situation
BRINTA	Type of interview
INTID	Identifier of the interviewer

15.4 Recent Changes in the Data Set

In 2012 the Biography Questionnaire was integrated in the Individual Questionnaire. This revised questionnaire version was used on the new sample K and did not contain the question about the use of the second residence. As a result respondents from sample K who have a second residence are coded with "-5" on the variable BRSECUSE.

In 2013 the new sample M was interviewed with a special version of the Biography Questionnaire which does not contain the questions on occupancy. Therefore, in 2013 and 2014 sample M is not part of the BIORESID dataset. In 2015 the sample M was interviewed on occupancy and is now a part of the BIORESID dataset. In 2014, Data from the SOEP-related FID-study from the years 2010 to 2014 has been integrated in the SOEP (Sample L). Therefore, the number of cases in BIORESID has increased with SOEP Version 31 retrospectively for the years 2010 to 2013, compared with previous versions.

16 BIOEDU: Data on educational participation and transitions

by Henning Lohmann and Sven Witzke

The Socio-Economic Panel Study (SOEP) contains a broad range of variables which cover early child education and care, educational participation, educational degrees and other related topics. However, the respective questions are included in different questionnaires (e.g., personal questionnaire, household questionnaire, youth questionnaire) and the variables are not always in a format which is suited for longitudinal analyses. For instance, transitions such as school enrolment or entry into tertiary education are not documented in a single variable but can only be reconstructed by comparing the status in a wave t with the status in a wave t+1 (e.g., a transition into tertiary education took place if a person was not in university in wave t but is in university in wave t+1). Generating such variables is time-consuming and prone to errors. It is the aim of the BIOEDU dataset to provide ready-made variables on educational transitions and related topics in order to support analyses in a longitudinal perspective.

The BIOEDU dataset is primarily based on prospectively collected information. Therefore, it contains most information for those persons who have been part of the survey population at the time when they have attended school or other educational institutions. In total the dataset contains information on 90,734 persons. This is the part of the SOEP sample for which we have observed an educational transition and/or an educational degree. For the larger part of this group we have observed an educational degree only (n=65,016). These are persons who have not been a part of the sample at the time when they participated in education or experienced educational transitions. The smaller part of the sample is more interesting for longitudinal analyse of educational participation. These are persons who lived in a survey household at the time of educational participation.[1] Depending on the age of the individual the dataset contains variables on:

- early child education and care (ECEC)
- entry into primary school
- transition to secondary school
- first exit from secondary school
- secondary school attendance after first exit from school
- first entry into and exit from vocational training

[1] Accordingly the first group is much older than the second group. At the time of the first observation in the sample the first group is on average 45 years old while the second group has an average age below 9 years.

- vocational training participation after first
- first entry into and exit from tertiary education
- tertiary education participation after first exit
- highest ever obtained educational degrees and last observed educational participation

The SOEP as a general household panel study is not specifically directed at the analysis of educational life courses. Nevertheless, right from the beginning of the panel in 1984 the survey instruments contained questions on the educational attainment of the respondents (aged 17 and older) and children younger than 17 years living in survey households. After more than 30 years of survey duration these data provide a precious source for the reconstruction of educational life courses. In the following we describe how we use these data to reconstruct educational transitions starting before school enrolment and up to post-secondary education.

The reconstruction of transitions is primarily based on yearly information on educational participation (i.e. entry and exit reconstructed from changes in participation). For later transitions there is some more information as explicit questions on the end of general school, vocational training and tertiary education are part of the questionnaires (changes during the year prior to the survey, only for persons aged 17+ years, exception: already obtained degrees before age 17 in youth questionnaire).

One remark on the variable naming conventions: The variable names always begin with "be" which stands for "biography education" (in analogy to other biography datasets). The third and fourth letter denote the type of transition or similar. For instance, t0 stands for variables on the first and t1 for the last year in child care. Variables on starting school contain a t2 and so on (up to t8= exit from tertiary education). Variables containing an x as the third letter contain information on the last observed year in education or on the highest educational degrees ever obtained (x4, x6, x8).

Using this dataset you should keep in mind that most of the information covered by the dataset is not directly asked in the SOEP questionnaires but has been derived from the combination of several variables. In the process of reconstruction assumptions have been made which we try to describe as detailed as possible in our exhaustive documentation of the dataset (see below). The more these assumptions are based on additional knowledge, e.g. provided by strict institutional regulations, the better for the reconstruction of the transitions.

The dataset covers transitions starting in early childhood up to tertiary education. For a part of the sample only one of these transitions or episodes is observed, for others the whole sequence from elementary education until the exit from tertiary education. The variable *beinfo*

provides an overview on the frequencies of these different patterns. In total the dataset contains information on more than 90,000 persons. This is the part of the SOEP sample for which we have observed an educational transition and/or an educational degree. For 597 cases we have full information (pattern 811111111).

We have provided a number of variables where we documented the process of data generation and the sources where the data stem from (betXinfo, variables with suffixes _s or _g). You could use these variables as indicators of the degree of uncertainty in the process of the reconstruction of educational transitions. The less the variables could be reconstructed just using the basic algorithm (e.g., bet2info<>"0000|0|0000"), the higher is the degree of uncertainty. The same applies to long durations between an observed exit and the observation of a matching educational degree (e.g., a high value in bet6cert_g). It is certainly advisable to check if certain deviations in the process of data generation "explain" substantial results. E.g., if children living in households where interviewed in August (this information is provided in betXinfo) have a much higher propensity of starting school late (bet2agemo), this might just be a data artefact because it is difficult to decide if the information the household provided referred to the school year which just started in August or to the school year which just ended at the time of the interview. In general, you should expect that there are no such systematic measurement errors in the reconstructed variables. But if you want to have a closer look on potential biases you could use the respective variables which document the data generation process. This documentation describes a beta version of the dataset (v32_0.1). If you have comments or encounter while using the dataset, please let us know.

This is just a brief introduction to the dataset. Way more detailed information (especially concerning the algorithms used to reconstruct information) is provided in the following publication which can be easily found on the DIW website. It is highly recommended for people interested in working with BIOEDU to have a look at it.

Lohmann, Henning / Witzke, Sven (2011): BIOEDU (beta version): Biographical data on educational participation and transitions in the German Socio-Economic Panel Study (SOEP), DIW Data Documentation 58, Berlin.

17 LIFESPELL: Information on the Pre- and Post-Survey History of SOEP-Respondents

by Martin Kroh and Hannes Kröger

Prospective panel surveys typically face the problem that no information is available on units of analysis after respondents have left the survey. The SOEP team therefore regularly conducts drop-out studies to identify the whereabouts of attritors. These studies draw on official register data and allow us to determine whether a person is still living in Germany, is deceased, or has moved abroad since the last SOEP interview. The information is combined in a spell file LIFESPELL. This dataset reports all available information on the pre- and the post-survey history of all persons who have ever been a member of a SOEP household. The LIFESPELL file lends itself particularly to mortality research, migration research, and non-response research. It extends the period under investigation from the last SOEP interview to the last drop-out study and thus reduces the problem of selective observational probabilities of units of analysis. For users less familiar with spell files, we also provide a STATA code for converting a spell file into a long format file (person x year) at the end of this chapter.

The file includes spells for every person ever living in a SOEP household. Each spell indicates one of the following states:

- Living abroad
- Living in Germany
- Living in Germany + part of a SOEP HH[1]
- Deceased
- No information about status

The information comprised by the LIFESPELL file includes the following:

- Year of birth

[1] This code does not distinguish between active and inactive respondents, such as children and temporary non-respondents. As long as interviewers can contact the household (even though an interview might not take place), we have information on the vital status of respondents. Therefore, all HH members are coded as being in the SOEP until the HH finally drops out of the SOEP and is no longer followed up by the interviewer.
* indicates that the fact is only necessary for subgroups of the population

- Year of immigration*
- Year of entry into SOEP
- Year of exit out of SOEP
- Year of emigration*
- Year of death

* indicates that the fact is only necessary for subgroups of the population

The year of birth, and year of immigration are self-reported (retrospective) information from personal interviews (e.g., p-files). The year of entry and exit are taken from gross information reported by the interviewer (e.g., pbrutto-files). The year of emigration and the year of death can either come from related persons in the household (e.g., deceased-files), the interviewer (e.g., pbrutto-files), or, finally, from drop-out studies (i.e., population registers).

The following register-based drop-out studies have been conducted in the past:

a) **1992 drop-out study** The study provides information about the dates of death or emigration, or one can infer that a person was still living in Germany until 1992.

b) **2001 drop-out study** The study provides information about the dates of death or emigration, or one can infer that a person was still living in Germany until 2001. Additionally, the study allows us to analyze regional mobility within Germany. It includes HH drop-outs from 1985-1998.

c) **2006 drop-out study** This study is not a register study, but a survey among non-respondents. Attritors were contacted by mail and responded by mail. It allows us to include the year of death for those whose letters returned by post with the mark "deceased" or whose letters were answered by relatives or friends. Those who answered themselves can also be assumed to still have been living in Germany in 2006. No information is available for cases where there was no response. No information about emigration is given. The study includes drop-outs from 2001-2004.

d) **2008 drop-out study** The latest of the studies provides information about the year of death and those still living in Germany in 2008. Emigration is captured insufficiently. It includes drop-outs from 1985-2006.

Year of study	1992	2001	2006	2008
Population at risk				
Attritors between…	1984-1990	1984-1998	2001-2004	1985-2006
Overall	1,089	8,082	4,982	5,838
Identified	1,044	7,106	962	5,591
Information available				
Death	+	+	(+)	+
Emigration	+	+	(+)	(+)
Mobility within Germany		+		

LIFESPELL contains the following variables. Comments are added to clarify certain assumptions made when coding the variables.

SPELLNR

Var Label: SPELLNR **“Number of spell”**

Var format: SPELLNR (I2)

Comment: This variable indicates the spell number.

SPELLTYP

Var Label: SPELLTYP **“Type of spell”**

Value Label: SPELLTYP (1)“Living in Germany”
(2)“SOEP”
(3)“Living abroad”
(4)“Deceased”
(5)“Gap”

Comment: The variable gives information about the current residence and status of the person. “Living in Germany” means that the person lives in Germany, but is not yet or no longer part of a SOEP HH. Persons who are part of a SOEP HH are given the “SOEP” code on this variable. Note that the code refers to the person being part of the gross sample, which also includes children and covers periods of temporary drop-outs as well as the final year when respondents refuse to participate. If a spell is coded “Abroad,” this means a person has not yet immigrated to Germany or has already emigrated. “Deceased” indicates the spell with the year of death and finally “Gap” means that for the time marked by the spell no information about the person’s status can be determined.

BEGIN

Var label: BEGIN **“Year in which spell begins”**

END

Var label: BEGIN **“Year in which spell ends”**

Comment: Two consecutive spells may overlap for one year in individual cases. This happens, for instance, when refusal to respond coincides with emigration in the same year.

CENSORING

Var label: CENSORING **“Censoring: How and why”**

Value Label: CENSORING

(10) “left censored”
(11) “left censored: gap”
(20) “right censored”
(21) “right censored: gap”
(22) “rc: no information after SOEP”
(23) “rc: no information after 1992 study”
(24) “rc: no information after study 2001”
(25) “rc: no information after study 2006”
(26) “rc: no information after study 2008”
(27) “rc: end of survey”
(40) “spell does not exist”
(50) “not censored”
(60) “right+left censored: gap”

Comment: The variable “censoring” gives information for each spell if all necessary information is available or if the spell is (left or right) censored, and why. If some information is missing, it is indicated by “Gap.” There are several instances when one or several facts are missing. For example, not every person reports a valid year of birth and year of immigration. In addition, sometimes we are able to report that a person is deceased but not the year of death. In these cases, a spell called ”Gap” is inserted between the last information which is certain (e.g., exit from SOEP) and the latest point in time by which the change in status must have happened. (For example, death must have occurred before the study was conducted, so the year of the study is the year of death; between the last SOEP participated in and the year of the study is the gap spell for which no information is certain except that the respondent’s death must have happened during this period.)

INFO

Var label: INFO “**Origin of information of spell**”

Value label:

(1) “SOEP survey”
(2) “SOEP fieldwork”
(3) “1992 study”
(4) “2001 study”
(5) “2006 study”
(6) “2008 study”
(7) “no info”

Comment: The variable indicates the information source for a spell. This can be a retrospective self-report (SOEP survey), fieldwork information provided by the interviewer (SOEP fieldwork), or a register-based drop-out study (1992, 2001, 2006, and 2008 studies).

STUDY$$$$

Var label: STUDY$$$$ “**Study $$$$**”
Value label:

(1) “not in study”
(3) “found in study”
(4) “not found in study”

Comment: The variable indicates if a person was part of the drop-out study in $$$$. A further distinction is made whether or not the person could be identified in register data.

FLAG

Var label: FLAG **"Spell information deviates from other SOEP information"**

Comment: In case of conflicting information between the drop-out studies and SOEP-based information, we give higher priority to the information from the drop-out studies. If two or more studies provide different information, that from the latest study is used. In just a few cases (110), the data set includes spells that contain information deviating from other SOEP data ("flag"). This happens, for instance, if register information says persons emigrated one year prior to refusal to take part in the SOEP. In this case, both spells end/begin in the same year with the register information being corrected by one year. A change in the original register information is coded 1 on the flag variable.

The following two tables illustrate the data structure for two individuals in the SOEP. Both individuals were born abroad, moved to Germany, became part of the SOEP sample, refused to participate after several waves. The one individual finally deceased while the other re-migrated. While in the first case, full information is available (no censoring), in the second case the exact year of immigration and the exact year of emigration are missing and therefore "gaps" had to be added (left and right censoring).

persnr	spellnr	spelltyp	begin	end	censoring	info
7016603	1	Living abroad	1903	1989	not censored	SOEP survey
7016603	2	Living in Germany	1990	1993	not censored	SOEP survey
7016603	3	SOEP	1994	1999	not censored	SOEP fieldwork
7016603	4	Living in Germany	2000	2003	not censored	2008 study
7016603	5	Deceased	2004	2004	not censored	2008 study

persnr	spellnr	spelltyp	begin	end	censoring	info
515802	1	Living abroad	1949	1949	right censored: gap	SOEP survey
515802	2	Gap	1950	1982	right+left censored: gap	No info
515802	3	Living in Germany	1983	1983	left censored: gap	SOEP survey
515802	4	SOEP	1984	1985	not censored	SOEP fieldwork
515802	5	Living in Germany	1986	2001	right censored: gap	Study 2001
515802	6	Gap	2002	2007	right+left censored: gap	No info
515802	7	Living abroad	2008	2008	right+left censored: gap	Study 2008

Users less familiar with analyzing spell data can easily convert the file into a long-format file (person x year). This is an example of STATA coding:

```
use lifespell.dta, clear
gen spellduration=(end-begin)+1
expand spellduration
bysort persnr spellnr: gen n=_n
gen year = begin+n-1
move year spellnr
keep persnr year spellnr spelltyp zensor info study1992 study2001 study2006
study2008 flag
compress

save lifelong.dta, replace
```