SOEP The German Socio-Economic Panel study

584

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP – The German Socio-Economic Panel Study at DIW Berlin 2018

SOEP-Core v33.1 – Biographical Information in the Meta File PPFAD (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989)

Diana Schacht, Christian Schmitt, Luisa Hammer, Lisa Ulrich, and SOEP Group

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**

Please cite this paper as follows:

Diana Schacht, Christian Schmitt, Luisa Hammer, Lisa Ulrich, and SOEP Group. 2018. SOEP-Core v33.1 – Biographical Information in the Meta File PPFAD (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989). SOEP Survey Papers 584: Series D. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)
DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany
soeppapers@diw.de

SOEP-Core v33.1 – Biographical Information in the Meta File PPFAD (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989)

Diana Schacht, Christian Schmitt, Luisa Hammer, Lisa Ulrich, and SOEP Group

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 Diana Schacht, Christian Schmitt, Luisa Hammer, and Lisa Ulrich

The file PPFAD includes—aside from other, primarily survey-relevant variables such as response status—important demographic information on every person who has ever participated in at least one wave of the SOEP study. These variables are of two types: first, longitudinally checked data on sex (variable SEX) and date of birth (four-digit year of birth in variable GEBJAHR and month of birth in variable GEBMONAT), and, second, generated demographic variables on the year of death (TODJAHR), country of origin (GERMBORN, CORIGIN), year of first immigration to Germany (IMMIYEAR), migration background (MIGBACK), as well as the region in which a person lived prior to German unification (LOC1989). In the following section, the construction of these generated variables will be explained briefly.

1 Month of Birth in the PPFAD data set

Introducing the variables

From Wave T on (2003), the data set PPFAD contains not only the year of birth but also the month of birth (GEBMONAT). This new variable is accompanied by the supplementary variable GEBMOVAL, which indicates the data source for the month of birth.

GEBMONAT and GEBMOVAL may have the following characteristics:

- GEBMONAT: Month of birth;
  1 (January) to 12 (December)
- GEBMOVAL: Month of birth — data source
  1 Generated
  2 Info stored in PPFAD
  3 Info derived from data set $KIND
  4 Info derived from data set SP (self-reported)
  5 Derived from data set $LELA (self-reported)
  6 Derived from BIOAGE01 (mother-child questionnaire)
  (NEW with Wave W / survey year 2006)

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

7 Derived from Youth Questionnaire (self-reported)
(NEW with Wave Z / survey year 2009)

The month of birth was surveyed in Wave S in the individual questionnaire (SP). Furthermore, the month of birth was surveyed in the biography questionnaire starting with Wave T ($LELA, file not included in the SOEP data release). Additionally, for all children, the month of birth is available in the file $KIND (starting with Wave T). Starting with Wave W, additional sources of information are considered in the generation of the month of birth. The data is based on biographical information on children obtained from the mother-child questionnaire (completed by mothers of newborns, $MUKI), along with a number of additional biographical questionnaires in which parents report on their children's development at different age intervals ($MUKI2, $MUKI3, $MUKI5, $elt, $SCHOOL, $SCHOOL2, and the consolidated child biography BIOAGEL). All these biographical questionnaires ask for the child's age in years and months. Furthermore, information from the Youth Questionnaires (self-response, age 17) is also considered. Priority is given to self-reporting in the Youth Questionnaire over parent proxy information.

All these sources of information are used to derive the month of birth, and valid information is used to replace missing data in one or more of the sources mentioned. This procedure has been successful in providing the relevant information for most of the current panel members. The information remains missing for individuals who lack any of the above information. This applies mainly to temporary dropouts or respondents who exited in a previous wave, before providing data in any of the questionnaires mentioned above. For some of these individuals, the month of birth could be reconstructed (this refers primarily to newborns, for whom the month of entry into the household is considered as a proxy if 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 an individual's own response on date of birth is given preference over information derived from other sources, and parent responses are considered more reliable for younger children.

2 Construction of variables

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

- Generated (basis: $P, $PBRUTTO $KIND)
- $KIND
- SP
- $LELA
- Child-related biography files, including $MUKI, $MUKI2, $MUKI3, $MUKI5, $elt, $SCHOOL, $SCHOOL2, BIOAGEL
- Youth Questionnaire ($PAGE17)

whereby each subsequent file overrides the previous one.

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 entry 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 and the month of birth was assigned to the child (relevant file $P).

The generated data has been tested and adjusted in several steps. The results show that—in the cases in which the generated data was also collected by SP, $LELA, $KIND, $PAGE17, and BIOAGEL—the generated data is almost always consistent with the collected data and is therefore reliable.

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

Table 1: GEBMONAT Month of Birth

		Frequency	Percent	Cumulative Percent
Valid	-5 Not Present in Version of Questionnaire	17.243	12,41	12,41
	-3 Answer improbable	1	0,00	12,41
	-1 No Answer	25.251	18,18	30,59
	1 January	10.426	7,51	38,10
	2 February	7.806	5,62	43,72
	3 March	8.625	6,21	49,93
	4 April	7.819	5,63	55,56
	5 May	8.154	5,87	61,43
	6 June	7.656	5,51	66,94
	7 July	8.232	5,93	72,87
	8 August	7.905	5,69	78,56
	9 September	7.991	5,75	84,31
	10 October	7.651	5,51	89,82
	11 November	6.958	5,01	94,83
	12 December	7.179	5,17	100,00
Total		138.897	100	

Source: PPFAD, SOEP v33.

Table 2: GEBMOVAL Month of Birth, Data Source

		Frequency	Percent	Cumulative Percent
Valid	-5 Not Present in Version of Questionnaire	17.243	12,41	12,41
	-3 Not Valid	1	0,00	12,41
	-1 No Answer	25.251	18,18	30,59
	1 Generated, newborn's entry into household	1.634	1,18	31,77
	3 $KIND, Info from mother	6.519	4,69	36,46
	4 Info from SP	21.854	15,73	52,20
	5 Info from $LELA	45.610	32,84	85,04
	6 Info from bioage[n]	12.799	9,21	94,25
	7 Info from $PAGE17	7.986	5,75	100,00
Total		138.897	100	

Source: PPFAD, SOEP v33.

3 Year of birth

(not generated)

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 from:

(1) the dropout file PBR_EXIT2, that is, as a result of the yearly fieldwork

(2) the 1992 Infratest *Verbleibstudie* (conducted by Infratest to follow up on dropouts)

(3) the 2001 Infratest *Verbleibstudie* (conducted by Infratest to follow up on dropouts)

(4) the 2006 Infratest *Verbleibstudie* (conducted by Infratest to follow up on dropouts)

(5) the 2008/9 Infratest *Verbleibstudie* (conducted by Infratest to follow up on dropouts)

Missing codes

(-2) Persons who are not deceased or no longer in the sample

Deaths of SOEP respondents are reported in the course of the yearly household interview, during which respondents are asked about the status of currently living household members as well as any births and deaths that occurred in the household since the previous year. Furthermore, the four follow-up studies that have been carried out so far to locate SOEP dropouts (Infratest *Verbleibstudien*) have been successful in identifying dropouts due to mortality and dropouts that moved abroad. The mortality information is used in generating the variable TODJAHR.

In the first *Verbleibstudie* conducted from April to June 1992, a total of 53 individuals were identified as deceased. In incorporating this information into the variable TODJAHR, an exact year of death was determined for only 35 of these individuals. An exact date of death was missing for 16 respondents; that is, only the qualitative information on their death was available. As a substitute for the year of death in these cases, the year of the wave in which the person dropped out of SOEP was used. For the remaining two individuals, implausible entries were corrected.

In the second Infratest *Verbleibstudie* conducted in 2001, over 700 individuals were identified as deceased. Included in this number were multiple identifications, i.e., individuals who were already

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

determined to be deceased through the standard follow-up procedure or in course of the first *Verbleibstudie* in 1992, mentioned above. This shows a very high correspondence of results between the standard follow-up and the ex-post determination of the time of death. For 10 individuals, 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 was conflicting information between the first two follow-up studies and the information from PBR_EXIT (formerly YPBRUTTO), the information from the *Verbleibstudie* was used.

In the third *Verbleibstudie*, another 21 individuals were identified as deceased between 2001 and 2005. For 18 of these individuals, a valid year of death was found, and the remaining three observations were set to the standard missing code "-1".

Again, some of these deaths have also been registered in the most recent of the Infratest *Verbleibstudien*, which was carried out in 2008. In this study. a total of 982 individuals were identified as deceased, with some of these deaths dating 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) 2006'

5 'Infratest Verbleibstudie (follow-up Study) 2008/9'

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

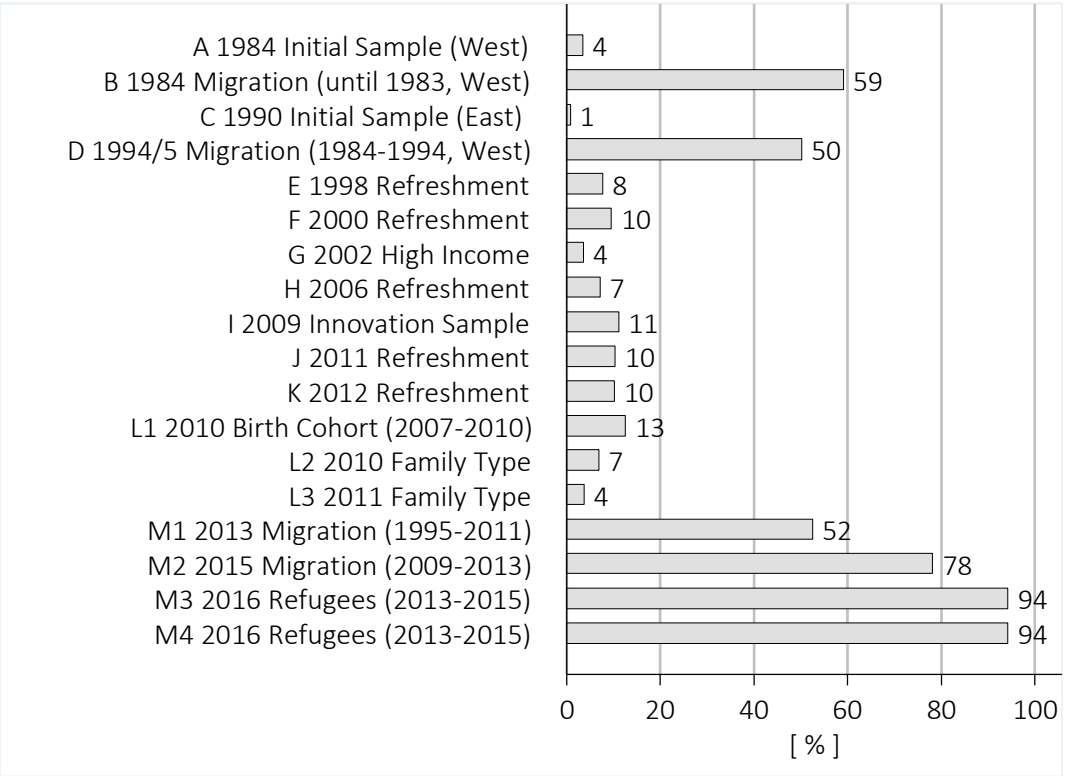
5 Immigration information

5.1 Introduction

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) 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 to have been born in Germany (the Federal Republic of Germany was founded in 1949). IMMIYEAR specifies the last year of immigration to the Federal Republic of Germany for all persons considered born in Germany, and MIGBACK is useful to identify immigrant descendants by combining information on respondents and their parents. In addition, GERMBORNINFO, CORIGININFO, IMMIYEARINFO, and MIGINFO indicate the quality of information given in GERMBORN, CORIGIN, IMMIYEAR and MIGBACK, respectively.

All SOEP samples include immigrants to Germany and their descendants. The shares vary, however, across samples depending on the target population covered. 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) contain a smaller number of immigrants and their descendants than the samples of foreigners and migrants (Sample B, D, M1, M2, M3 and M4).[3] *Graph 1* illustrates the share of persons who immigrated to Germany since 1950.

Graph 1: Distribution of foreign-born survey participants across SOEP samples (A to M4)



Source: All survey participants (n=126,804); row percentages, SOEP v33.

Information for GERMBORN, CORIGIN, IMMIYEAR and MIGBACK and the respective INFO variables (GERMBORNINFO, CORIGININFO, IMMIYEARINFO and MIGINFO) is collected primarily from the wave-specific individual questionnaires ($P, $PAUSL, $MIG or $REFUGEES) 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 3* lists information used for generating GERMBORN, CORIGIN, IMMIYEAR, and MIGBACK and the respective SOEP source files. In the following sections, the variables GERMBORN, CORIGIN, IMMIYEAR, and MIGBACK and their generation process are described in detail. Special attention is given to the filtering function of GERMBORN for CORIGIN, IMMIYEAR, and MIGBACK.

[3] 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: Information used for GERMBORN, CORIGIN, IMMIYEAR and MIGBACK

Information used	Data set(s)
Main indicators	
Born in Germany (yes/no)	BIOLELA / $LELA / $P / $MIG / $JUGEND / ELECTRONIC HOUESHOLD PROTOCOL M1
Country of birth	BIOLELA / $LELA / $P / $PAUSL / $MIG / $REFUGEES / $JUGEND / $PBRUTTO
Year of immigration to Germany	BIOLELA / $LELA / $P / $PAUSL / $MIG / $REFUGEES / $MIGSPELL / $JUGEND / ELECTRONIC HOUESHOLD PROTOCOL M1
East German, Ethnic German or migrated before 1949	
Immigration group (East German, emigrant of German descent from Eastern Europe, German who lived abroad, EU citizen, asylum seeker, other)	BIOIMMIG
Area of origin (GDR, FRG, former German territory, Europe, other)	BIOLELA / $P / LPBRUTTO
Displaced person between 1945 and 1950 (yes/no)	$LELA
Citizenship and legal status	
Citizenship	BIOLELA / $LELA / $KIND / $PBRUTTO / $PGEN / INFRATEST INFORMATION
German citizenship (yes/no)	$LELA / $P / $JUGEND
Current citizenship	$LELA / $P / $PAUSL / $MIG / $REFUGEES / $JUGEND
Previous citizenship	$LELA / $P / $MIG / $JUGEND
Dual citizenship	$LELA / $P / $MIG / $JUGEND
Citizenship: former GDR	GPOST
EU citizenship (yes/no)	$MIG
Naturalization	$LELA / $P / $MIG / $JUGEND
Residency permit in Germany	$LELA / $JUGEND
Migration history	
Place of residence before 1989	PPFAD
When first move from country of birth	$MIG / $REFUGGES
Moved to Germany or to other country (destination country)	$MIG / $REFUGGES
Moved back to country of origin or elsewhere at least once (yes/no)	$MIG / $REFUGEES
Moved back to Germany again/moved when?	$MIG / $REFUGEES
Family information	
Respondent: Year of birth	PPFAD
Mother/father pointer	BIOBIRTH
Mother/father: German citizenship (yes/no)	$LELA / $JUGEND / BIOPAREN
Mother/father: German citizenship (ethnic German, naturalized, since birth, no)	$MIG / $JUGEND
Mother/father: born in Germany (yes/no)	$LELA / $MIG / $JUGEND
Mother/father: country of birth	$LELA / $MIG / $JUGEND / BIOPAREN
Mother/father: year of immigration	$MIG
Mother/father: current citizenship	$MIG / $JUGEND
Parents: both born in Germany	$P / $MIG
Sample	
Relationship to head of household	$PBRUTTO

Member of household (in HH at least two years, moved from abroad, etc.)	$PBRUTTO
Subsample Identifier (German HH head, Turkish HH head, etc.)	$HBRUTTO / $H
Moved to Germany (Yes/No) (as reported by the anchor person)	Electronic household protocol M1 2013

Source: SOEP v33. The sub-headings differentiate between main indicators and auxiliary indicators such as information on ethnic Germans. The column "information used" gives an indication of information used to generate the respective (auxiliary) variables. For example, "Citizenship" and "Current citizenship" are differentiated since they are based on different wordings in the questionnaires. For more information on the wording of questions, generated variables, and topics within the SOEP see https://data.soep.de/.

5.2 GERMBORN "Born in Germany" and GERMBORNINFO

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 (the Federal Republic of Germany was founded in 1949; see also IMMIYEAR). In v33, all relevant information (see *Table 3*) available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD) was combined into a working dataset and compared to code GERMBORN. When information in this working dataset consistently indicated that a person was born either in Germany or abroad, GERMBORNINFO was coded with a *(1)* for *"consistent information"*. When inconsistent or no direct information on a SOEP person was available, GERMBORNINFO was coded with a *(2)* indicating *"inconsistent information"* or with a *(3)* indicating *"no information"*. GERMBORNINFO is thus an indicator of the quality of information given in GERMBORN. *Table 4* presents information about the GERMBORNINFO distribution in PPFAD. The vast majority of persons who have ever been part of a SOEP household (i.e., the population from PPFAD) gave *consistent information* on their country of birth (63%). For these 63% of the PPFAD population, GERMBORN could easily be coded according to the respondents' answers.

Table 4:GERMBORNINFO "GERMBORN: Quality of Information" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	Consistent information	80,043	63	63
2	Inconsistent information	2,079	2	65
3	No information	44,682	35	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

However, for another third of the dataset, no direct information on the person's country of birth was available (35% or around 45,000 PPFAD cases, see *Table 4*). GERMBORNINFO value *"(3) no information*" refers to persons who lived in a SOEP household but had not completed an individual, life history, or youth questionnaire up to the present date (67% children and 18% other household members). Another 15% of the "*(3) no information*" cases on GERMBORNINFO had given an interview but did not answer the question on their country of birth (item non-response).

Only in a few cases (2%, see *Table 4*) was "*inconsistent information"* provided. Over the course of the SOEP survey, some individuals may have stated on one occasion that they were born in Germany and on another that they were born abroad; such information was considered as *inconsistent* (value *(2)* on GERMBORNINFO). On average, persons with *"(2) inconsistent information*" answered seven questions on their country of birth in the SOEP study (from 2 to 18 answered questions).

For both, persons for whom "*(2) inconsistent* information" or "*(3) no information*" (GERMBORNINFO) was available, additional indicators were used to code the GERMBORN values. In this process, information on a respondent's citizenship and their parents' migration biography was used. We coded the values on GERMBORN in the following order (with descending priority):

- First, mothers' immigration history and their place of residence at the time of the respondents' birth were taken into account to determine the respondents' probable country of birth. For instance, when a respondent was born after or in the year of their mother's immigration to Germany, the respondent is considered to have been born in Germany. A more detailed differentiation by month of birth and mother's immigration month is not possible due to missing information. When a mother's immigration year was missing, the father's immigration history was used to code a respondent's country of birth. This procedure led to a coding of around 12% of the PPFAD cases (12% of the "no information" cases and 0.2% of the "inconsistent information" cases).
- In the next step, GERMBORN was coded for the remaining "*(2) inconsistent information"* cases. Respondents' information on their country of birth, their citizenship, and parental information was taken into account to identify a respondents' country of birth. For instance, a respondent who reported being born in Germany more often than being born abroad (*country of birth*), who had German citizenship (*citizenship*), and whose parents were both born in Germany (*parental information*) was considered to have been born in Germany.
- In a last step, GERMBORN was coded for the remaining "*(3) no information*" cases. Respondents' citizenship and parental information was used to approximate their most likely country of birth. By definition, information on their country of birth was missing. For instance, respondents with German citizenship whose parents were both born in Germany were coded as being born in Germany.

In comparison to the GERMBORN coding from 2015 (v32), all PPFAD cases in 2016 (v33) have a value on GERMBORN. Previously (v32) missing cases (9,193 cases) were coded in this version (v33), because more longitudinal and detailed citizenship and parental information was used to generate GERMBORN (and GERMBORNINFO). For a few PPFAD cases, the new generation procedure led to a change of the GERMBORN value (1,084 cases). For the majority of these cases, *"(2) inconsistent information"* or *"(3) no information"* (see GERMBORNINFO) was available (65%) and a value change is therefore not surprising, particularly in light of the previous (v32 and earlier) GERMBORN generation procedure, which did not consider longitudinal GERMBORN information but only the latest answer given by the respondent or a third party. In addition, around 13,000 cases entered the SOEP in 2016 (v33), for instance, through one of the two refugee samples.[4] *Table 5* displays the GERMBORN distribution of persons in the latest PPFAD version (v33).

[4] For more information on sample sizes and panel attrition in the SOEP, see Kroh, M.; Kühne, S.; Jacobsen, J.; Siegert, M. and Siegers, R. (2017): Sampling, Nonresponse, and Integrated Weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4). SOEP Survey Papers 477: Series C -- Data Documentation. DIW Berlin.

Table 5:GERMBORN "Born in Germany" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	Born in Germany or immigrated < 1950	96,003	76	76
2	Not born in Germany	30,801	24	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

5.3 CORIGIN "Country of origin" and CORIGININFO

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. 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, $PAUSL, $MIG or $REFUGEES) 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). In addition, information from $PBRUTTO or the ELECTRONIC HOUSHOLD PROTOCOL for M1 was used (both answered by the household head).

In v33, all relevant information (see *Table 3*) available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD) was compiled into a working dataset and compared to code CORIGIN. CORIGININFO indicates whether "(*1) consistent*", "(*2) inconsistent*" or "(3*) no information*" was available on a respondent's country of birth after these comparisons. CORIGININFO is thus an indicator for the quality of information given in CORIGIN. The filtering of CORIGIN via GERMBORN was taken into account by implementing a separate category, *"(4) Filter GERMBORN*" on CORIGININFO for the persons who were considered being born in Germany on GERMBORN (for more information, see GERMBORN).

When information in this working dataset consistently indicated a specific country of origin, CORIGININFO was coded "(*1) consistent information*" and the respective country of origin was mentioned in CORIGIN. The SOEP team also considered information as "(*1) consistent*" in the following two additional cases (with descending priority):

- When state transformations (e.g., their founding or dissolution) may have led to respondents reporting different countries of birth over the course of the SOEP survey, information was considered consistent. For instance, respondents may have stated the Union of Soviet Socialist Republics (USSR) as their country of birth in 1987 but stated Russia in a later questionnaire. Other examples refer to the dissolution of the Socialist Federal Republic of Yugoslavia in 1992 and their temporary and contemporary successor states, such as *"(119) Croatia*", *"(120) Bosnia and Herzegovina"*, *"(121) Macedonia*", *"(122) Slovenia*", *"(165) Serbia*", *"(168) Montenegro*". In such cases, CORIGIN was coded with the most contemporary successor state mentioned by a respondent or third party. This may also include regions or ethnic groups that respondents mentioned, such as *"(140) Kosovo-Albanian*" or *"(149) Kurdistan*".
- When a respondent or third party mentioned a rather unspecific region of birth such as *"(12) Benelux"*, "(*222) Eastern European*" or *"(999) Ethnic minority*" and at another time mentioned a

more specific country of origin or citizenship within this region, information was considered consistent. The more specific country of origin was used in CORIGIN.

Table 6: CORIGININFO "CORIGIN: Quality of information" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	Consistent information	20,119	16	16
2	Inconsistent information	281	0	16
3	No information	10,401	8	24
4	Filter GERMBORN	96,003	76	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

The vast majority of respondents who have ever been a part of a SOEP household (i.e., the population from PPFAD) gave *consistent* information on their country of birth (16%, see *Table 6* or 65% for the foreign-born population, thus without Filter GERMBORN cases). For 8% of the dataset, *no direct information* on the person's country of birth was available (34% for the foreign-born population, thus without Filter GERMBORN cases). "(3*) No information*" either refers to persons who lived in a SOEP household but did not complete an individual, life history, or youth questionnaire up to now (64% children and 26% other household members) or to respondents who were interviewed but did not answer the questions on their country of birth (10% item-non-response). Over the course of the SOEP survey, only a few cases gave *"(2) inconsistent information*" with regards to their country of birth (around 300 PPFAD cases on CORIGININFO). On average, persons with *"(2) inconsistent information*" answered three questions on their country of birth in the SOEP (from 2 to 5 answered questions).

For those respondents who were not born in Germany and whose country of birth could not be determined (CORIGININFO value *(2)* and *(3)*), additional indicators were used to code their country of origin (CORIGIN). The generation process was conducted in the following order (with descending priority):

- Respondents' country of origin was considered to be the same as their country of citizenship if this was not German (for more information on the information used, see *Table 3* under the sub-heading citizenship). The citizenship variable was constructed on the basis of all information given on first, second, and previous citizenships as well as naturalizations, and includes the countries of citizenship a respondent reported most frequently and/or first. Since citizenship information is collected annually for all persons who lived in a SOEP household, it is based on much more detailed information than the *"(2) inconsistent information*" collected for the country of origin. Respondents whose information on country of origin is *"(2) inconsistent*" answered only three questions on their country of origin on average (from 2 to 5 answers).
- 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 (see also GERMBORN coding). If information on mothers' country of birth and the respondents' citizenship was missing, fathers' country of birth was used to code CORIGIN.
- For the few cases without citizenship and parental information (around 150 cases), respondents' most recently mentioned country of origin was used.
- For the few cases without citizenship, parental information and any information on their country of origin (CORIGININFO value *(3)*), respondents' legal status was used when it indicated that a person

moved to Germany from an "Eastern European" country, resulting in the coding of around 150 cases to "(*222) Eastern European*" on CORIGIN.

If the country of birth was still missing after this procedure, CORIGIN was coded *"(-1) don't know*". 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.

Table 7:CORIGIN "Country of Origin" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
-1	no answer / don't know	236	0	0
1	Germany	96,003	76	76
2	Turkey	3,117	2	78
…				
183	Niger	4	0	100
222	Unspecified Eastern European country	149	0	100
999	Ethnic minority	1	0	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

5.4 IMMIYEAR "Year of immigration" and IMMIYEARINFO

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 has been collected (up to 15 moves between countries, see MIGSPELL in this SOEP Survey Paper). 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. In addition, information from the ELECTRONIC HOUSEHOLD PROTOCOL for M1 was used, which was only answered by the household head.

In v33, all relevant information (see *Table 3*) available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD) was compiled into a working dataset and compared to code IMMIYEAR. IMMIYEARINFO indicates whether "(*1) consistent*", "(*2) inconsistent*" or "(3*) no informatio*n" was available on a respondent's country of birth after these comparisons. IMMIYEARINFO is thus an indicator for the quality of information given in IMMIYEAR. The filtering of IMMIYEAR via GERMBORN was taken into account by implementing a separate category *"(4) Filter GERMBORN*" on IMMIYEARINFO for individuals who were considered to have been born in Germany on GERMBORN (for more information, see GERMBORN).

Table 8: IMMIYEARINFO "IMMIYEAR: Quality of information" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	Consistent information	20,205	16	16
2	Inconsistent information	49	0	16
3	No information	10,547	8	24
4	Filter GERMBORN	96,003	76	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

When information in this working dataset consistently indicated a specific year of immigration, IMMIYEARINFO was coded "(*1) consistent information*" and the respective year of immigration was stated in IMMIYEAR. The vast majority of the persons who have ever been a part of a SOEP household (i.e., the population from PPFAD) gave *consistent* information on their year of immigration (16% for the PPFAD population or 66% for the foreign-born population, thus without Filter GERMBORN cases, see *Table 8*). For another 8%of the dataset (34% for the foreign-born population, without Filter GERMBORN cases) *no direct information* on the person's year of immigration was available (around 10,500 PPFAD cases, see *Table 8*). "(3*) No information*" either refers to persons who lived in a SOEP household but did not complete an individual, life history, or youth questionnaire up to now (62% children and 26% other household members) or to respondents who were interviewed but did not answer the questions on their year of immigration (12% item non-response). Over the course of the SOEP survey, only very few cases gave *"(2) inconsistent information*" with regard to their year of immigration (around 50 cases on IMMIYEARINFO). For these cases, their latest year of immigration was used in IMMIYEAR. The respondent's year of birth was used as their year of immigration if they mentioned a year of immigration that was before their year of birth (8 cases).

For those respondents who were not born in Germany and whose year of immigration could not be determined (IMMIYEARINFO value *(3)*), additional indicators were used to minimize the portion of missing values. These indicators were used in the following order (with descending priority):

- When a respondent entered the SOEP for the first time because they 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 a mother's year of immigration was missing, the father's year of immigration was used to code IMMIYEAR.

If the year of immigration was still missing after this procedure, IMMIYEAR was coded *"(-1) don't know*". IMMIYEAR includes more missing values than GERMBORN and CORIGIN due to cases in which it was not possible to determine a respondent's year of immigration.

Table 9: IMMIYEAR "Year of Immigration to Germany" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
-2	does not apply	96,003	76	76
-1	no answer / don't know	4,907	4	80
1950		12	0	80
…				
2016		522	0	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

However, users should be aware that the wording of questions on the year of immigration vary rather drastically over the course of the SOEP survey. *Table 10* gives an overview of the respective phrasings of the year of immigration questions. The column "*category*" gives an indication whether the question asked for the first or most recent year of immigration or whether the phrasing of the question did not specify which specific year of immigration may have been meant.

Table 10: Variations of the SOEP questions regarding respondents' year of immigration (main indicators for IMMIYEAR and IMMIYEARINFO)

Category	Question	Data sets	Years
Respondent information			
First	What year did you move to the Federal Republic of Germany (including West Berlin) for the first time?	APAUSL-GPAUSL IPAUSL-JPAUSL	1984-1990 1992-1993
First	In which year did you move to Germany for the first time?	HPAUSL	1991
Unspecific	Since when have you lived in the area of the former FRG or West Berlin? If after 1949, since when?	GP-HP	1990-1991
Unspecific	Since when have you lived in the area of the former FRG or West Berlin or in the area of the former GDR and East Berlin? If after 1949, since when?	IP-JP	1992-1993
Unspecific	What year did you move to the Federal Republic of Germany (including West Berlin) for the first time?	BIOLELA	1984-1995
Unspecific	When did you move to the Federal Republic of Germany?	QJUGEND-BGJUGEND	2000-2016
Unspecific	When did you move to the Federal Republic of Germany?	MLELA-BCLELA	1996-2012
Unspecific	When did you move to Germany?	BDLELA	2013
Last	When did you move to Germany? If you have moved to Germany several times during your life, please refer to your most recent move to Germany.	BELELA-BGLELA	2014-2016
First	First of all we would like to know when you first moved away from your country of birth?	BDP_MIG-BGP_MIG	2013-2016
First	Which country did you move to?	BDP_MIG-BGP_MIG	2013-2016
First & Last	Did you move away from Germany again after that?	BDP_MIG-BGP_MIG	2013-2016
First & Last	When did you move to Germany?	BDP_MIG-BGP_MIG	2013-2016
First	First of all we would like to know when you first moved away from your country of birth?	BGP_REFUGEES	2016
First	Was Germany the first country you moved to, or was it another country?	BGP_REFUGEES	2016
First & Last	Did you move away from Germany again after that?	BGP_REFUGEES	2016
First & Last	Was Germany the first country you moved to after this, or was it another country?	BGP_REFUGEES	2016
First & Last	When did you move to Germany in this case?	BGP_REFUGEES	2016
Unspecific	When did you arrive in Germany?	BGP_REFUGEES	2016
Third party information			
Unspecific	Member of household: Moved into household from abroad.	BPBRUTTO-BGPBRUTTO	1985-2016
Unspecific	Did <first name> move to Germany? (reported by the anchor respondent)	ELECTRONIC HOUSEHOLD PROTOCOL M1 2013	2013
Unspecific	When did your father move to Germany?	BDP_MIG-BGP_MIG	2013-2016
Unspecific	When did your mother move to Germany?	BDP_MIG-BGP_MIG	2013-2016

Source: SOEP v33.

5.5 MIGBACK "Migration background" and MIGINFO

MIGBACK contains information on respondents' migration background for all persons who have ever been a part of a SOEP household (i.e., the population from PPFAD). In comparison to GERMBORN, the variable MIGBACK is useful to identify immigrants' descendants by combining information on respondents' country of birth (see GERMBORN) and parental information such as their country of birth and their citizenship. The information for this variable comes predominantly from PPFAD (GERMBORN), auxiliary citizenship variables (for more information, see *Table 3* under sub-heading "citizenship and legal status" and sub-heading "family information"), and the relevant biographical data sets (BIOPAREN, BIOIMMIG). The variables were also updated using information from the wave-specific individual questionnaires ($P, $PAUSL, $MIG or $REFUGEES), 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).

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"*.

In comparison to the MIGBACK coding from 2015 (v32), all PPFAD cases in 2016 (v33) have a value on MIGBACK. Previously (v32) missing cases (6,781 cases) or cases where the migration background could not be differentiated further (2,412 cases) were coded in this version (v33), because more information on respondent´s country of birth was provided (see GERMBORN). Please 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. *Table 11* displays the MIGBACK distribution of persons in the latest PPFAD version (v33).

Table 11: MIGBACK "Migration background" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	No migration background	81,134	64	64
2	Direct migration background	30,801	24	88
3	Indirect migration background	14,869	12	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

To provide the highest level of transparency possible, we include a variable for the sources used to create the migration background variable: MIGINFO. MIGINFO indicates the quality of information given in MIGBACK. MIGINFO provides information about the usage of parents' migration histories in the SOEP. Overall, MIGINFO can take on two different codes: *"(1) No parental information*" or *"(2) Parental information available"*. 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 (for more information, see *Table 3* under sub-heading "family information") and auxiliary citizenship

variables (for more information, see *Table 3* under sub-heading "citizenship and legal status" and sub-heading "family information").

Please note that the MIGINFO coding from 2015 (v32) is further differentiated between the availability of direct and proxy information on respondents. We changed the MIGINFO coding due to the introduction of the GERMBORNINFO variable. The quality of information given in MIGBACK can thus only be assessed by combining the GERMBORNINFO and MIGINFO variables. MIGBACK information is considered to be highly reliable in cases coded *(2) "Parental information available"* on MIGINFO and *(1) "Consistent information"* on GERMBORNINFO (around 41% of the PPFAD cases). In contrast, the quality of information given on MIGBACK is considered relatively uncertain in cases where parental information (*(1) "No parental information"* on MIGINFO) and respondents' information was missing (*(3) "No information"* on GERMBORNINFO)).

In a few cases, *"(1) no parental information"* (see MIGINFO) was available but we were nonetheless able to identify respondents with an *"(2) indirect migration background"* (see MIGBACK). In these cases, respondents were born in Germany but further variables (for more information, see *Table 3* under sub-heading "citizenship and legal status" and sub-heading "East German, Ethnic German, or migrated before 1949") suggested that there was a migration background (e.g., ethnic Germans). MIGBACK may slightly underestimate the number of persons having an *"(3) indirect migration background",* since some of the respondents born in Germany with missing parental information and for whom no further indicators were available may be the descendants of immigrants.

Table 12: MIGINFO "MIGBACK: Quality of information" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	No parental information	43,425	34	34
2	Parental information available	83,379	66	100
Total		126,804	100	

Source: PPFAD, SOEP v33.

5.6 LOC1989 "Where did you live in 1989?" and LOCINFO

The variable LOC1989 in the meta-file PPFAD provides information about a person's residence *prior to* German reunification, distinguishing among "*(1) German Democratic Republic [GDR]*", *"(2) Federal Republic of Germany [FRG] (including West Berlin)",* and *"(3) abroad*". Respondents born after 1989 (GEBJAHR in PPFAD) were coded as *"(-2) does not apply*" on LOC1989. This information has been generated for all individuals who were ever a member of a SOEP household (the population of PPFAD).

LOC1989 combines information from two main sources: In 2003, the individual questionnaire included information on the place of residence before German reunification (TP). Since 2004, this question has been included in the biography questionnaires ($LELA). Along with these sources, the following indicators were used to code the variable LOC1989 (with descending priority):

- $HHNR in PPFAD: Place of residence in the former FRG before German reunification
- IMMIYEAR in PPFAD: Respondents who first immigrated to Germany after 1989 were coded as living "[3] abroad" in 1989
- IMMIYEAR, CORIGIN in PPFAD: Respondents who immigrated to Germany before 1990 were assumed to have been living in the "*(2)]* Federal Republic of Germany [FRG] (including West Berlin)" in 1989

- PSAMPLE in PPFAD: Respondent's sample affiliation in 1990, differentiating between members of the former West samples (A, B) and the former East sample (C)
- $SAMPREG in PPFAD & BRMOVEIN and SYEAR in BIORESID: Respondents who moved into their current dwelling in the former FRG or GDR before 1989
- GSAMPREG in PPFAD: Respondent living in the West or East sample region in 1990

The vast majority of information given in LOC1989 is based on information from these sources. For the remaining respondents, *indirect information* is derived from the following proxies to code their place of residence in 1989:

- $PZUG in $PBRUTTO: New entrants to the SOEP who previously lived in East Germany or abroad
- BSSCHEND and BSSCHWO in BIOSOC: Place and year of the last school attended
- LPGRUPPE in LPBRUTTO: Place of birth that was asked in 1995
- $P: Country of origin GDR
- KPNAT in KPBRUTTO: Citizens of (former) GDR
- $P: Place of residence in 1984
- BIOPAREN in PPFAD: Parental residence in 1989 for individuals younger than 18 in 1989

Table 13: *LOC1989 "Where did you live in 1989?" distribution*

Value	Value Label	Frequency	Percent	Cumulative Percent
-2	Does not apply	38,607	30	30
-1	No answer / don't know	18,018	14	45
1	German Democratic Republic [GDR]	14,642	12	56
2	Federal Republic of Germany [FRG] (including West Berlin)	47,191	37	93
3	Abroad	8,351	7	100
Total		126,809	100	

Source: PPFAD, SOEP v33.

The variable LOCINFO indicates the quality of information given in LOC1989, differentiating between direct and indirect information.

LOCINFO provides information about the use of proxy information in the process of generating LOC1989 due to missing values in respondents' and their parents' residence in 1989 in the SOEP. Overall, LOCINFO can take on three different codes: either "*(1) direct*" or "*(2) indirect information*" is available on respondents or they were "*(0) born after 1989"*. *Table 14* illustrates which variables were used to generate LOC1989 and their respective LOCINFO coding.

Table 14: LOCINFO "Loc1989: Source / quality of information" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
-1	No answer / don't know	18,013	14	14
0	Respondent born after 1989	38,607	30	45
1	Direct information	68,268	54	98
2	Indirect information	1,916	2	100
Total		126,804	100	

Source: PPFAD, SOEP v33