SOEP The Socio-Economic Panel

745

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP – The German Socio-Economic Panel at DIW Berlin 2019

SOEP-Core v34 – Biographical Information in the Meta File PPATH (Month of Birth, Immigration Variables, Living in East or West Germany in 1989)

Diana Schacht, Christian Schmitt, Antonia Scherz, Lisa Ulrich, and SOEP Group

DIW SOEP

Running since 1984, the German Socio-Economic Panel (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, Antonia Scherz, Lisa Ulrich, and SOEP Group. 2019. SOEP-Core v34 – Biographical Information in the Meta File PPATH (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989). SOEP Survey Papers 745: 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 v34 – Biographical Information in the Meta File PPATH (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989)

Diana Schacht, Christian Schmitt, Antonia Scherz, Lisa Ulrich, and SOEP Group

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

By Diana Schacht, Christian Schmitt, Antonia Scherz, and Lisa Ulrich

The files PPFAD, PPATH and PPATHL include—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 month of birth (GEBMONAT), and, second, generated demographic variables on the country of origin (GERMBORN, CORIGIN), year of last 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 / PPATH data set

1.1 Introduction

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)
  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,

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

$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.

1.2 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.

Table 1: GEBMONAT Month of Birth distribution

		Frequency	Percent	Cumulative Percent
Valid	-5 Not Present in Version of Questionnaire	17,243	12.43	12.43
	-3 Answer improbable	1	0	12.43
	-1 No Answer	25,179	18.14	30.57
	1 January	10,413	7.50	38.08
	2 February	7,802	5.62	43.70
	3 March	8,620	6.21	49.91
	4 April	7,816	5.63	55.54
	5 May	8,148	5.87	61.41
	6 June	7,652	5.51	66.93
	7 July	8,226	5.93	72.86
	8 August	7,901	5.69	78.55
	9 September	7,988	5.76	84.31
	10 October	7,646	5.51	89.82
	11 November	6,955	5.01	94.83
	12 December	7,177	5.17	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

Table 2: GEBMOVAL Month of Birth, Data Source distribution

		Frequency	Percent	Cumulative Percent
Valid	-5 Not Present in Version of Questionnaire	17,243	12.43	12.43
	-3 Not Valid	1	0.00	12.43
	-1 No Answer	25,179	18.14	30.57
	1 Generated, newborn's entry into household	1,633	1.18	31.75
	3 $KIND, Info from mother	6,519	4.70	36.45
	4 Info from SP	21,854	15.75	52.19
	5 Info from $LELA	45,560	32.83	85.03
	6 Info from bioage[n]	12,792	9.22	94.25
	7 Info from $PAGE17	7,986	5.75	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

2 Immigration information

2.1 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, PPATH and PPATHL). In detail, GERMBORN and CORIGIN give information on the country of birth, with the exception of persons who immigrated to Germany before

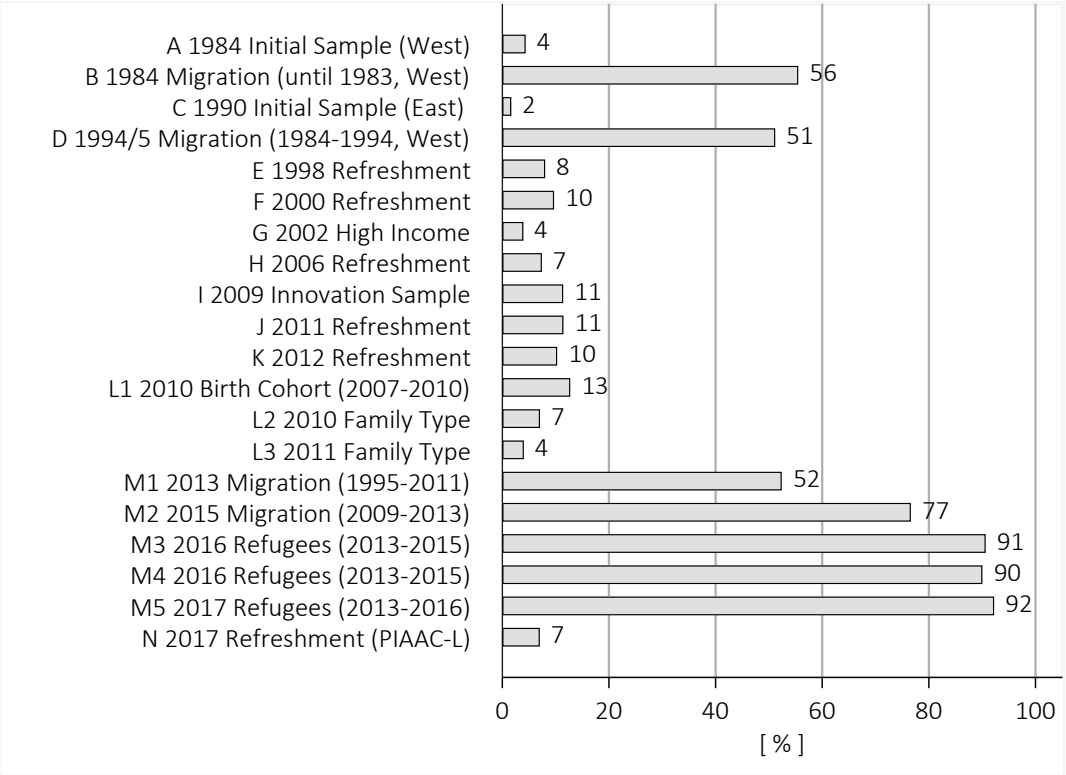
1950 who are considered to have been born in Germany (the Federal Republic of Germany was founded in 1949). IMMIYEAR specifies the last year of immigration to the Federal Republic of Germany for all persons considered not born in Germany, and MIGBACK is useful to identify immigrant descendants by combining 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, G, H, I, J, K, L1, L2, L3 and N) 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, M4 and M5).[2] *Graph 1* illustrates the share of persons who immigrated to Germany since 1950.

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). In addition, information from the electronic household protocol for M1 was used (answered by the household head and not included in the standard data distribution). *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.

[2] 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: Series C. Nebelin, J. and Petrenz, M. and Wenzig, K. (2019) Wegweiser zur Datenanalyse der IAB-SOEP Migrationsstichproben (M1, M2) und der IAB-BAMF-SOEP Stichproben von Geflüchteten (M3-M5). DIW Berlin: SOEP Survey Papers 600: Series G.

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



Source: All survey participants (n=138,767); row percentages, SOEP v34.

Table 3: Information used for GERMBORN, CORIGIN, IMMIYEAR and MIGBACK

Information used	Data set(s)	
	core	long
Main indicators		
Born in Germany (yes/no)	BIOLELA / $LELA / $P / $MIG / $JUGEND / Electronic household protocol M1	BIOL / PL / JUGENDL / Electronic household protocol M1
Country of birth	BIOLELA / $LELA / $P / $PAUSL / $MIG / $REFUGEES / $JUGEND / $PBRUTTO	BIOL / PL / JUGENDL / PBRUTTO
Year of immigration to Germany	BIOLELA / $LELA / $P / $PAUSL / $MIG / $REFUGEES / MIGSPELL / REFUGSPELL / $JUGEND / Electronic household protocol M1	BIOL / PL / MIGSPELL / REFUGSPELL / JUGENDL / Electronic household protocol M1
East German, Ethnic German or migrated before 1949		
Immigration group (Emigrant of German descent from Eastern Europe, German who lived abroad, EU citizen, asylum seeker, other)	BIOIMMIG	BIOIMMIG
Area of origin (GDR, FRG, former German territory, Europe, other)	BIOLELA / $P / LPBRUTTO	BIOL / PL / PBRUTTO
Displaced person between 1945 and 1950 (yes/no)	$LELA	BIOL
Citizenship and legal status		
Citizenship	BIOLELA / $LELA / $KIND / $PBRUTTO / INFRATEST INFORMATION	BIOL / KIDLONG / PBRUTTO / INFRATEST INFORMATION
German citizenship (yes/no)	BIOLELA / $LELA / $P / $JUGEND	BIOL / PL / JUGENDL
Current citizenship	$LELA / $P / $PAUSL / $REFUGEES / $JUGEND	BIOL / PL / JUGENDL
Previous citizenship	$LELA / $P / $MIG / $JUGEND	BIOL / PL / JUGENDL
Dual citizenship	$LELA / $P / $MIG / $JUGEND / $PBRUTTO	BIOL / PL / JUGENDL / PBRUTTO
Citizenship: former GDR	GPOST	PL
Naturalization	BIOLELA / $LELA / $P / $MIG / $JUGEND	BIOL / PL / JUGENDL
Residency permit in Germany	$LELA / $JUGEND	BIOL / JUGENDL
Migration history		
Place of residence before 1989	PPFAD / PPATH	PPATHL
When first move from country of birth	$LELA / $MIG / $REFUGEES	BIOL
Moved to Germany or to other country (destination country)	$LELA / $MIG / $REFUGEES	BIOL

Moved back to country of origin or elsewhere at least once (yes/no)	$LELA / $MIG / $REFUGEES	BIOL
Moved back to Germany again/moved when?	$LELA / $MIG / $REFUGEES	BIOL
Month of immigration to Germany	BIOLELA / $LELA / $MIG / $REFUGEES	BIOL
Travel time to Germany	$LELA / $REFUGEES	BIOL
Family information		
Respondent: Date of birth	PPFAD / PPATH	PPATHL
Mother/father pointer	BIOBIRTH / BIOPAREN / $PBRUTTO	BIOBIRTH / BIOPAREN / PBRUTTO
Mother/father: German citizenship (yes/no)	$LELA / $JUGEND	BIOL / JUGENDL
Mother/father: German citizenship (ethnic German, naturalized, since birth, no)	$LELA / $JUGEND	BIOL / JUGENDL
Mother/father: born in Germany (yes/no)	$LELA / $JUGEND	BIOL / JUGENDL
Mother/father: country of birth	$LELA / $JUGEND	BIOL / JUGENDL
Mother/father: year of immigration	$LELA	BIOL
Mother/father: current citizenship	$LELA / $JUGEND	BIOL / JUGENDL
Maternal/paternal grandmother/grandfather pointer	BIOBIRTH / BIOPAREN / $PBRUTTO	BIOBIRTH / BIOPAREN / PBRUTTO
Sample		
Relationship to head of household	$PBRUTTO	PBRUTTO
Member of household (in HH at least two years, moved from abroad, etc.)	$PBRUTTO	PBRUTTO
Subsample Identifier (German HH head, Turkish HH head, etc.)	$HBRUTTO / $H	HBRUTTO / HL
Moved to Germany (Yes/No) (as reported by the anchor person)	Electronic household protocol M1 2013	Electronic household protocol M1 2013

Source: SOEP v34. 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/.

2.2 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). To code GERMBORN, 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, PPATH and PPATHL) was combined into a working dataset. 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 PPATH. The vast majority of persons who have ever been part of a SOEP household gave *consistent information* on their country of birth (66%). For these 66% of the PPATH 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	91,188	66	66
2	Inconsistent information	1,778	1	67
3	No information	45,801	33	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

However, for another third of the dataset, no direct information on the person's country of birth was available (33% or around 46,000 PPATH 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 (64% children and 21% 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 (1%, 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 were 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. For the coding of a few cases, more detailed information on respondents' month of birth and mother's immigration month was available and used. 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 13% of the PPATH cases (13% 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. While last year the latest information was considered more reliable, this year the mode was calculated for inconsistent information on respondents' and parental country of birth. In case of varying modes, higher values were given a preference when coding, to be more sensible to foreign countries 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 reported more often to be born in Germany than being born abroad (*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. In contrast to last year, the mode of parents' country of birth and citizenship was used for the coding of GERMBORN, too. For instance, respondents with German citizenship whose parents reported more often to be born in Germany than being born abroad were coded as being born in Germany.

For a few PPATH cases, the new generation procedure led to a change of the GERMBORN value (1,385 cases). For the majority of these cases, *"(2) inconsistent information"* or *"(3) no information"* (see GERMBORNINFO) was available (95%) and a value change is therefore not surprising. In addition, around 13,000 cases entered the SOEP in 2017 (v34), for instance, through the refugee sample M5.[3] *Table 5* displays the GERMBORN distribution of persons in the latest PPATH version.

Table 5:GERMBORN "Born in Germany" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	Born in Germany or immigrated < 1950	102,924	74	74
2	Not born in Germany	35,843	26	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

2.3 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, PPATH and PPATHL). 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 household protocol for M1 was used (both answered by the household head).

To code CORIGIN, 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, PPATH and PPATHL) was compiled into a working dataset. 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

[3] 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.

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	23,547	17	17
2	Inconsistent information	289	0	17
3	No information	12,007	9	26
4	Filter GERMBORN	102,924	74	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

The vast majority of respondents who have ever been a part of a SOEP household (i.e., the population from PPFAD, PPATH and PPATHL) gave *consistent* information on their country of birth (17%, see *Table 6* or 66% for the foreign-born population, thus without Filter GERMBORN cases). For 9% 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 27% 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 PPATH 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):

- The respondents' country of birth which occurred most frequently, in other words the mode, was used.

- Respondents' country of citizenship was used as their country of birth if both were 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. 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 on average three questions on their country of origin (from 2 to 5 answers).

- Mothers' country of birth and citizenship were 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, mothers' citizenship and the respondents' citizenship was missing, fathers' country of birth and fathers' citizenship were used to code CORIGIN. In comparison to the CORIGIN coding from the last wave, in the latest version grandparents' country of birth and grandparents' citizenship were additionally used if information on mothers' and fathers' country of birth and citizenship were missing.

- For the few cases without citizenship, (grand-)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 170 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	390	0	0
1	Germany	102,924	74	74
2	Turkey	3,032	2	77
…				
183	Niger	5	0	100
222	Unspecified Eastern European country	169	0	100
999	Ethnic minority	1	0	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

2.4 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, PPATH and PPATHL) 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 and REFUGSPELL in the SOEP Survey Paper Series). 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.

To code IMMIYEAR, 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, PPATH and PPATHL) was compiled into a working dataset. IMMIYEARINFO indicates whether "(*1) consistent*", "(*2) inconsistent*" or "(3*) no information*" was available on a respondent's year of immigration 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	23,446	17	17
2	Inconsistent information	62	0	17
3	No information	12,335	9	26
4	Filter GERMBORN	102,924	74	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

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, PPATH and PPATHL) gave *consistent* information on their year of immigration (17% for the PPATH population or 65% for the foreign-born population, thus without Filter GERMBORN cases, see *Table 8*). For another 9% 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 12,000 PPATH 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 (13% 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 60 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 a mother's and father's year of immigration were missing, the maternal and paternal grandparents' year of immigration were used respectively.

Table 9: 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-BHJUGEND	2000-2017
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-BHLELA	2014-2017
First	First of all we would like to know when you first moved away from your country of birth?	BDP_MIG-BGP_MIG, BHLELA	2013-2017
First	Which country did you move to?	BDP_MIG-BGP_MIG, BHLELA	2013-2017
First & Last	Did you move away from Germany again after that?	BDP_MIG-BGP_MIG, BHLELA	2013-2017
First & Last	When did you move to Germany?	BDP_MIG-BGP_MIG, BHLELA	2013-2017
First	First of all we would like to know when you first moved away from your country of birth?	BGP_REFUGEES, BHLELA	2016-2017
First	Was Germany the first country you moved to, or was it another country?	BGP_REFUGEES, BHLELA	2016-2017
First & Last	Did you move away from Germany again after that?	BGP_REFUGEES, BHLELA	2016-2017
First & Last	When did you move to Germany in this case?	BGP_REFUGEES, BHLELA	2016-2017
Unspecific	When did you arrive in Germany?	BGP_REFUGEES, BHLELA	2016-2017
Third party information			
Unspecific	Member of household: Moved into household from abroad.	BPBRUTTO-BHPBRUTTO	1985-2017
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?	BDLELA-BHLELA	2013-2017
Unspecific	When did your mother move to Germany?	BDLELA-BHLELA	2013-2017

Source: SOEP v34.

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 10: IMMIYEAR "Year of Immigration to Germany" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
-2	does not apply	102,924	74	74
-1	no answer / don't know	4,554	3	77
1950		18	0	77
…				
2017		297	0	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

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.

2.5 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, PPATH and PPATHL). In comparison to GERMBORN, the variable MIGBACK is useful to identify immigrants' descendants by combining information on respondents' country of birth (see GERMBORN) and (grand-)parental information such as their country of birth and their citizenship. The information for this variable comes predominantly from PPATH (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 (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 last wave, in the latest version grandparental information were considered when an indirect migration background and parental information were missing. . 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 a foreign 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 PPATH version.

Table 11: MIGBACK "Migration background" distribution

Value	Value Label	Frequency	Percent	Cumulative Percent
1	No migration background	85,217	61	61
2	Direct migration background	35,843	26	87
3	Indirect migration background	17,707	13	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

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 (grand-)parents' migration histories in the SOEP. Overall, MIGINFO can take on two different codes: *"(1) No parental information*" or *"(2) Parental information available"*. The (grand-)parental information refers to any information on the migration background of the respondents' mother, father or grandparents. 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 in 2016 (v33). 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 / PPATH 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 were 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 (grand-)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	41,120	30	30
2	Parental information available	97,647	70	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

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

The variable LOC1989 in the meta-file PPFAD / PPATH 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 PPATH) 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 PPATH).

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 PPATH: Place of residence in the former FRG before German reunification
- IMMIYEAR in PPATH: Respondents who first immigrated to Germany after 1989 were coded as living "[3] abroad" in 1989
- IMMIYEAR, CORIGIN in PPATH: 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 PPATH: Respondent's sample affiliation in 1990, differentiating between members of the former West samples (A, B) and the former East sample (C)
- $SAMPREG in PPATH & BRMOVEIN and SYEAR in BIORESID: Respondents who moved into their current dwelling in the former FRG or GDR before 1989
- GSAMPREG in PPATH: 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 and PPATH: 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	44,620	32	32
-1	No answer / don't know	17,358	13	45
1	German Democratic Republic [GDR]	15,467	11	56
2	Federal Republic of Germany [FRG] (including West Berlin)	49,393	36	91
3	Abroad	11,929	9	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34.

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	17,358	13	13
0	Respondent born after 1989	44,620	32	45
1	Direct information	74,873	54	99
2	Indirect information	1,916	1	100
Total		138,767	100	

Source: PPFAD / PPATH, SOEP v34