DIW Berlin / SOEP (Ed.)

Research Report

SOEP-Core v36 - PPATHL: Person-related meta-dataset

SOEP Survey Papers, No. 1048

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: DIW Berlin / SOEP (Ed.) (2021) : SOEP-Core v36 - PPATHL: Person-related meta-dataset, SOEP Survey Papers, No. 1048, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/248506

1048 2021

SOEP Survey Papers
Series D – Variable Descriptions and Coding

SOEP-Core v36 – PPATHL: Person-Related Meta-Dataset

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:
SOEP Group, 2021. SOEP-Core v36 – PPATHL: Person-Related Meta-Dataset. SOEP Survey Papers 1048: Series D – Variable Descriptions and Coding. Berlin: DIW Berlin/SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

The German Socio Economic Panel at DIW Berlin

SOEP-Core v36 – PPATHL: Person-Related Meta-Dataset

SOEP Group

2021

The file ppathl is part of a collection, which is released with doi:10.5684/soep.core.v36.

Contents

1 General Information

The path datasets should be the building block of any analysis. Path Files indicate the total population at the household and individual level (over time) and provide all IDs necessary to access further files at different levels (Krause/Glass/Reher 2019a,b). Path-Files are delivered in three data formats – in long-format [H|P-PATHL] (as the most comprehensive version including weighting variables), in wide-format [H|P-PFAD] (the traditional version), and in a short-version [H|P-PATH] as a reduced population file (indicating the total of population of households and individuals) Household Level [HID|SYEAR] {Navigation File: H-PATH-L (long-format)} Individual Level [PID|SYEAR] {Navigation File: P-PATH-L (long-format)} Household Level [HID] {Population File: H-PATH} Individual Level [PID] {Population File: P-PATH} Household Level [HID (HHNRAKT)] {Path File: HPFAD (wide-format)} Individual Level [PID (PERSNR)] {Path File: PPFAD (wide-format)}. The constituting SOEP population considers three levels – cases, households, and individuals. Due to the SOEP sampling and survey process, these levels follow an implicit hierarchy. All samples refer to primary source households – indicated by the household id at the time when the survey starts – the (fixed) Case ID [CID]. New Households may emerge from these original households during the longitudinal survey process by split-offs of family members – all (current) households are therefore indicated by a (variable) Household ID [HID]. IDs for individuals living in the households are derived from the households, where they were living when they were surveyed for the first time – the (fixed) Personal ID [PID]. It is recommendet to use the (almost) time-independent (demographic) information like sample membership, sex, year and country of origin are adjusted on a wave-by-wave basis in the framework of demographic testing.

2 Primary Key, Foreign Keys and Sample Information

pid – Never Changing Person ID

A person can be uniquely identified with variable PID. Together with SYEAR primary key in this file.

syear – Survey Year

1984		16252
1985		16737
1986		15868
1987		14974
1988		14596
1989		14000
1990		19666
1991		19713
1992		19552
1993		19240
1994		19469
1995		19947
1996		19527
1997		19064
1998		21175
...	(6 rows omitted)	188652
2005		30339

2006	32747
2007	30962
2008	29005
2009	31642
2010	45977
2011	50329
2012	50120
2013	55611
2014	51684
2015	50277
2016	57287
2017	64554
2018	62491
2019	62839

Together with PID primary key in this file.

hid – ID Household

The household the person belongs to in the corresponding year (SYEAR).
For more information, contact: Peter Krause (Tel. 030-89789-690)

psample – Sample Member

1	A 1984 Initial Sample (West)	277404
2	B 1984 Migration (until 1983, West)	95261
3	C 1990 Initial Sample (East)	130046
4	D 1994/5 Migration (1984-1994, West)	25893
5	E 1998 Refreshment	26921
6	F 2000 Refreshment	169823
7	G 2002 High Income	36373
8	H 2006 Refreshment	29661
9	I 2009 Innovation Sample	7130
10	J 2011 Refreshment	47322
11	K 2012 Refreshment	21608
12	L1 2010 Birth Cohort (2007-2010)	60811
13	L2 2010 Family Type (Low-Income, Single-Parent, Large Families)	64822
14	L3 2011 Family Type (Single-Parent, Large Families)	26088
15	M1 2013 Migration (1995-2011)	45389
16	M2 2015 Migration (2009-2013)	12718
17	M3 2016 Refugee (2013-2015)	18476
18	M4 2016 Refugee/family (2013-2015)	26515
19	M5 2017 Refugee (2013-2016)	14587
20	N 2017 Refreshment (PIAAC-L)	17353
21	O 2018 Social City	4083
22	P 2019 Top-Shareholder	5199
23	Q 2019 Lesbian-Gay-Bisexual (LGB) persons	813
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0

-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

The sample membership never changes.
For more information, contact: Peter Krause (Tel. 030-89789-690)

cid – Case-ID, Original Household Number

Case Id - Household Source Identifier. The fixed household source id points to the first SOEP ancestor household for this person. The person does not necessarily need to have ever lived in a household with this number. This is relevant for sampling information and calculating the weights.

persnr – Never Changing Person ID

Same as PID.

3 Survey History

eintritt – Year First Contacted, Netto=10-99

1984		253147
1985		6309
1986		7650
1987		7142
1988		7095
1989		7474
1990		106986
1991		9346
1992		8430
1993		8092
1994		17561
1995		17340
1996		7474
1997		6687
1998		30330
...	(6 rows omitted)	223421
2005		5747
2006		32561
2007		5020
2008		4142
2009		10787
2010		119623
2011		73374
2012		25513
2013		47719
2014		5347

2015	15729
2016	43328
2017	36172
2018	7493
2019	7257

The year a person joined the SOEP.
For more information, contact: Peter Krause (Tel. 030-89789-690)

erstbefr – Year First Surveyed, Netto=10-99

1984		189358
1985		7818
1986		8446
1987		7722
1988		6737
1989		7358
1990		79592
1991		8815
1992		8377
1993		8205
1994		14713
1995		14993
1996		8361
1997		8173
1998		25947
...	(7 rows omitted)	206396
2006		29476
2007		9192
2008		7033
2009		11041
2010		59850
2011		51502
2012		24030
2013		34625
2014		9824
2015		15241
2016		23819
2017		27541
2018		10288
2019		10015
-2		229808

The year of a person's first interview.
For more information, contact: Peter Krause (Tel. 030-89789-690)

austritt – Year Of Last Contact, Netto=10-99

1985	2607
1986	3906

1987		3068
1988		4819
1989		5029
1990		3524
1991		4369
1992		5143
1993		6742
1994		7955
1995		8242
1996		8674
1997		8088
1998		10553
1999		11483
...	(5 rows omitted)	73155
2005		16258
2006		24635
2007		24423
2008		25232
2009		38173
2010		38504
2011		42879
2012		31640
2013		36796
2014		34442
2015		41694
2016		45348
2017		42255
2018		51760
2019		502900

The last year of a person's SOEP appearance.
For more information, contact: Peter Krause (Tel. 030-89789-690)

letztbef – Year Of Last Survey, Netto=10-99

1984	3434
1985	3169
1986	3044
1987	4404
1988	4128
1989	4070
1990	4541
1991	5358
1992	5908
1993	7161
1994	7286
1995	7113
1996	8176
1997	10411
1998	11009

...	(7 rows omitted)	108505
2006		21623
2007		23756
2008		25669
2009		27690
2010		30257
2011		38915
2012		28369
2013		32293
2014		32318
2015		34565
2016		36383
2017		42690
2018		54003
2019		308240
-2		229808

The year of a person's most recent interview.
For more information, contact: Peter Krause (Tel. 030-89789-690)

netto – Current survey status

10	Interviewee With Succesful Interview (_P)	567068
12	Individual Questionnaire And Person Biography	94848
13	Individual Questionnaire And Youth Biography	318
14	Individual Questionnaire And Other Questionnaires	32
15	Individual Questionnaire And Experiments, Test	43317
16	Individual Questionnaire, First Time Surveyed, Age 17	5946
17	Youth Biography First Time Surveyed, Age 17	6804
18	Individual Questionnaire And Child under age 17	8
19	Individual Questionnaire Without Household Interview	782
20	Children in Succesfully Interviewed Households (_Kind)	199218
21	Children With Mother-Child Questionnaire_I, Age 0-1	7035
22	Children With Mother-Child Questionnaire_II, Age 2-3	7301
23	Children With Mother-Child Questionnaire_III, Age 5-6	7094
24	Children age 7-8, with parental questionnaire	6471
25	Children age 9-10, with parental questionnaire	6620
...	(29 rows omitted)	202061
91	Moved abroad	2722
92	Moved abroad (abroad)	177
93	Moved abroad (exit)	65
94	Person Gap with advices	424
97	advice to dead person (exit)	981
98	advice to dead person (_VP)	341
99	Has Died	4639
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	24
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0

-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

This variable indicates available information and files for the entire SOEP individuals. Nettocodes 10-19 (and 29) define the respondents population of PGEN, the codes 20-28 indicate children, 30-39 unit-non-responses in partially realized households, and the codes 90-99 describe permanent (or temporary) dropouts. Further differentiations point to the survey instruments (questionnaires). The Codes 10-39 describe the population in realized (and partially realized households).
For more information, contact: Peter Krause (Tel. 030-89789-690)

nett1 – Current survey status (old 1 digit)

0	Person Gap PBR_EXIT	13241
1	Successful Interview _P, _JUGEND	718335
2	Below Survey Age _KIND	242302
3	Did Not Participate _PBRUTTO	176618
4	Missing This Wave _PLUECKE	10552
5	Interviewee Without Household Interview	788
-1	No Answer	0
-2	Does not apply	2397
-3	Answer improbable	63
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

Short version of NETTO-variable
For more information, contact: Peter Krause (Tel. 030-89789-690)

casemat – Case-Match, combined panel households

0	CASE With HH Details	124
20605		2
20613		5
27367		14
27430		7
250996		3
272906		6
277908		1
283924		9
291102		7
292338		13
292621		7
344095		11
700495		3
701564		13
701670		3

701718		1
863637		1
3014079		2
3094650		7
3499749		4
3635481		9
3920330		4
-1	No Answer	0
-2	Does not apply	1164040
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

It is possible that Individuals from different original households (CID) move together in one common household. Then people with identical values for (HID) in one wave may have different values for CID. Only for those persons moving together CASEMAT contains the HID of the other household members. This information is not relevant when linking person and household data based on the current household number HID.
For more information, contact: Peter Krause (Tel. 030-89789-690)

piyear – Jahr des Interviews (Interview Year)

1984		16252
1985		16361
1986		15548
1987		14633
1988		14254
1989		13689
1990		19427
1991		19414
1992		19147
1993		18833
1994		19101
1995		19486
1996		19108
1997		18785
1998		20689
...	(8 rows omitted)	248440
2007		30460
2008		28453
2009		29655
2010		45151
2011		50254
2012		49401
2013		54819
2014		50734
2015		49465

2016	56395
2017	58307
2018	64115
2019	64614
2020	1115
-2	18191

Interview Year (indicates personal interviews realized also outside of standard SYEAR)

4 Basic Demographic Information

sex – Gender

1	Male	573535
2	Female	590670
-1	No Answer	82
-2	Does not apply	0
-3	Answer improbable	9
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

Respondent's (last) sex, plausibility longitudinally validated.
For more information, contact: Peter Krause (Tel. 030-89789-690)

gebjahr – Birth Year, 4-digit

1882		2
1888		4
1892		19
1893		8
1894		13
1895		22
1896		65
1897		58
1898		54
1899		156
1900		176
1901		162
1902		304
1903		201
1904		363
...	(101 rows omitted)	1068858
2006		9976
2007		12933
2008		12235
2009		10589

2010	10751
2011	6263
2012	5580
2013	5092
2014	4256
2015	3712
2016	3136
2017	2147
2018	1025
2019	296
-1	5840

Respondent's year of birth, plausibility longitudinally validated.
For more information, contact: Peter Krause (Tel. 030-89789-690)

gebmonat – Month Of Birth

1	January	92690
2	February	79320
3	March	86007
4	April	78134
5	May	81228
6	June	75731
7	July	82358
8	August	79700
9	September	80118
10	October	77037
11	November	69549
12	December	73071
-1	No Answer	100861
-2	Does not apply	2623
-3	Answer improbable	13
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	105856
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

The month of birth

- was asked starting with wave W (2006) in the mother-child-questionnaire for newborns
- was asked starting with wave T (2003) in supplementary biography questionnaire, resulting in file BIOL
- was asked in wave S individual questionnaire, resulting in file SP
- is recorded for all children within the file TKIND (wave T, 2003)
- can be approximately derived for newborn children from the month of moving into the household, stored in file PBRUTTO
- can be reported by parents in the personal questionnaire (which might simultaneously establish a link to the child), stored in file PL

whereas the former information is prefered over the latter. This means the generated infor-

mation (from TKIND, PBRUTTO or PL) 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. Several adjustments and tests of the generated data have been done which showed that – in the cases in which the generated data was also collected by PL, BIOL or $KIND – the data generation is almost always congruent with the collected data and therefore has proven to be reliable. The used source of information is stored in GEBMOVAL.
While this provides the relevant information for most of the current panel members, the information remains missing for some persons including temporary dropouts or people who exited in a previous wave. For some of them the month of birth could be reconstructed. This reconstruction remains an approximation and might differ from the true month of birth in individual cases.
For more information, contact: Christian Schmitt (Tel. 030-89789-603)

gebmoval – Month Of Birth, Data Source

1	Generated from gebmonth (parents)	9188
2	Ppfad, carry forward	0
3	$kind, Info from mother	64317
4	Info From Sp	419373
5	Info From $lela	276254
6	Info From bioage$$ (mother)	76164
7	Info from $PAGE17	109647
-1	No Answer	100861
-2	Does not apply	2623
-3	Answer improbable	13
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	105856
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

Indicates the data source for the month of birth (GEBMONAT).
For more information, contact: Christian Schmitt (Tel. 030-89789-603)

todjahr – Year Died, 4 Digits

1984	14
1985	153
1986	298
1987	404
1988	527
1989	636
1990	592
1991	865
1992	923
1993	1156
1994	1085
1995	1663

1996		1585
1997		1486
1998		1682
...	(8 rows omitted)	21442
2007		3350
2008		4259
2009		2765
2010		2825
2011		2513
2012		2289
2013		3258
2014		2919
2015		2775
2016		3570
2017		3424
2018		3432
2019		2335
-1		357
-2		1089714

The variable TODJAHR contains the four-digit year of death for persons whose death could be firmly established or a missing value code:

- (-2): persons, for whom it is unknown whether they are deceased (that is, both persons still living up to that wave, and persons whose exact whereabouts is unknown and have dropped out of SOEP)
- (-1): persons, for whom the fact of death is known, but the year of death is unknown.

todinfo – Year Died, Information Source

1	From Annual Survey (pbr_exit)	58626
2	survey about died person ($v)	318
3	survey about parents ($lela)	15
4	Infratest drop-out study 1992	17
5	Infratest drop-out study 2001	4044
6	Infratest drop-out study 2007	33
7	Infratest drop-out study 2008/9	8173
8	Modul Family changes [P]	2999
-1	No Answer	357
-2	Does not apply	1089714
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

For all persons who have been identified as deceased over the course of SOEP, the variable TODINFO gives the source of this information.
53 persons were identified as deceased in the Infratest Field Organization Study (Follow-up

study of drop-outs between 1984 and 1992) carried out from April – June 1992.
In the framework of the Infratest Field Organization Study (follow-up study of drop-outs) of 2001, a total of over 700 persons were identified as deceased. Among them were several with multiple entries for year of death, that is, persons who were already identified as deceased in the standard wave-to-wave follow-up procedure (stored in the file PBR_EXIT) or in the Infratest Field Organization Study of 1992. A generally very high level of correspondence was found between the information given in the standard follow-up procedure and the point of death established ex-post in the Infratest Field Organization Studies. For ten persons, the year of dropping out of SOEP was used to impute the missing year of death. In the third of those follow-up studies which has been conducted in 2007, another 21 individuals were identified as deceased between 2001 and 2005. For 18 of those persons a valid year of death could be investigated, for the remaining three observations for which the exact year of death is unknown, TODJAHR has been set to the standard missing code "-1".
When the data from the Infratest Field Organization Study contradicted the data from PBR_EXIT, the data from the Field Organization Study was used.

birthregion – Birth place: German Federal Land

1	Schleswig-Holstein	9508
2	Hamburg	6410
3	Lower Saxony	33508
4	Bremen	2990
5	North Rhine-Westphalia	70135
6	Hesse	21867
7	Rhineland-Palatinate	16973
8	Baden-Wuerttemberg	38928
9	Bavaria	50940
10	Saarland	3697
11	Berlin	14237
12	Brandenburg	16824
13	Mecklenburg-West Pomerania	11056
14	Saxony	34791
15	Saxony-Anhalt	20139
16	Thuringia	18363
-1	No Answer	96088
-2	Does not apply	697842
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

BITHREGION contains information about the German Federal State ("Bundesland") a person was born. In 2012 the SOEP asked all current respondents about the place of birth: "Where were you born? If there are other towns or cities with the same name, or if the town is very small, please state the nearest city. Please write the name of the town in the left blank and any additional information in the right blank. For example, write 'Düsseldorf', 'Frankfurt an der Oder', or 'Frankfurt am Main' in the left blank, and in the case of 'Roßdorf bei Schmalkalden', write 'Roßdorf' in the left blank and 'bei Schmalkalden' in the right blank.

Since then this question has been part of the biography questionnaire and a variable BIRTHREGION is povided in dataset PPATH, which has to be updated each year for new respondents. The answers is given in clear text and coded by Kantar at the level of municipalities for german cities or villages (including the geocodes for the city center). For places outside Germany, Kantar provides only the geocodes, if possible. However, the reponses could not all be assigned to a unique municipality, therefore multiple municipality codes are provided by Kantar (up to 19 in 2012). For the variable *birthregion* in *ppfad* only those answers are used, where a unique assignment of a German Federal State ("Bundesland"), based on the possible municipality codes, was possible. For persons born in a SOEP household (the household was responding in this year) the code of the respective Federal State of this year is used.
For more information, contact: Jan Goebel (Tel. +49 30-89789-377)

germborn – Born in Germany

1	born in Germany or immigr.<1950	954421
2	not born in Germany	209875
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

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 or 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 (grand-)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 residential population in Germany (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) or the sample of households in urban areas (Sample O).
Information for GERMBORN, CORIGIN, IMMIYEAR and MIGBACK and the respective INFO variables is collected primarily from the individual questionnaires (dataset PL) or the variations of the "biography / life history" questionnaires (integrated biographical data files and life-course information in dataset (BIOL) and from the additional youth questionnaire for 16-17-year-olds, in use since 2000 (dataset JUGENDL). In addition, information from the

electronic household protocol for M1 (2013) and the retrospective survey (2012) of early childhood in the context of war (dataset BCBFK) was used (both datasets are not included in the standard data distribution).
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 1: Information used for GERMBORN, CORIGIN, IMMIYEAR and MIGBACK) available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD, PPATH or PPATHL) was combined. The vast majority of persons who have ever been part of a SOEP household gave consistent information on their country of birth and GERMBORN was coded accordingly to the respondents' answers. For part of the population, no direct information on the person's country of birth was available. 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):

1. 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.

2. 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. 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.

3. 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. 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.

Table 1: Information used for GERMBORN, CORIGIN, IMMIYEAR and MIGBACK

Information used	Dataset (long format)
Main indicators	

Information used	Dataset (long format)
Born in Germany (yes/no)	BIOL / PL / JUGENDL / Electronic household protocol M1 / BCBFK
Country of birth	BIOL / PL / JUGENDL / PBRUTTO / BCBFK
Year of immigration to Germany	BIOL / PL / MIGSPELL / REFUGSPELL / JUGENDL / Electronic household protocol M1 / BCBFK
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
Area of origin (GDR, FRG, former German territory, Europe, other)	BIOL / PL / PBRUTTO
Displaced person between 1945 and 1950 (yes/no)	BIOL / BCBFK
Citizenship and legal status	
Citizenship	BIOL / KIDLONG / PBRUTTO / INFRATEST INFORMATION
German citizenship (yes/no)	BIOL / PL / JUGENDL
Current citizenship	BIOL / PL / JUGENDL
Previous citizenship	BIOL / PL / JUGENDL
Dual citizenship	BIOL / PL / JUGENDL / PBRUTTO
Citizenship: former GDR	PL
Residency permit in Germany	BIOL / JUGENDL
Place of residence before 1989	PPATHL
When first move from country of birth	BIOL
Moved to Germany or to other country (destination country)	BIOL
Moved back to country of origin or elsewhere at least once (yes/no)	BIOL
Moved back to Germany again/moved when?	BIOL
Month of immigration to Germany	BIOL
Travel time to Germany	BIOL
Family information	
Respondent: Date of birth	PPATHL
Mother/father pointer	BIOBIRTH / BIOPAREN / PBRUTTO
Mother/father: German citizenship (yes/no)	BIOL / JUGENDL
Mother/father: German citizenship (ethnic German, naturalized, since birth, no)	BIOL / JUGENDL
Mother/father: born in Germany (yes/no)	BIOL / JUGENDL
Mother/father: country of birth	BIOL / JUGENDL
Mother/father: year of immigration	BIOL
Mother/father: current citizenship	BIOL / JUGENDL
Maternal/paternal grandmother/grandfather pointer	BIOBIRTH / BIOPAREN / PBRUTTO
Sample	
Relationship to head of household	PBRUTTO
Member of household (in HH at least two years, moved from abroad, etc.)	PBRUTTO

Information used	Dataset (long format)
Subsample Identifier (German HH head, Turkish HH head, etc.)	HBRUTTO / HL
Moved to Germany (Yes/No) (as reported by the anchor person)	Electronic household protocol M1 2013

Source: v35
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

germborninfo – Germborn: Quality of information

1	consistent information	906933
2	inconsistent information	34053
3	no answer	223310
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

GERMBORNINFO indicates the quality of information given in GERMBORN. As in previous years, all relevant information available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD, PPATH or PPATHL) 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". 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 information (value (2) on GERMBORNINFO). The 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 or they had given an interview but did not answer the question on their country of birth (item non-response).
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

corigin – Country Born In

1	Germany	954421
2	Turkey	28358
3	Ex-Yugoslavia	4484
4	Greece	8433
5	Italy	12336
6	Spain	5194
7	Ex-GDR (only as country of origin)	0
10	Austria	2127
11	France	1236
12	Benelux	81

13	Denmark	226
14	Great Britain	720
15	Sweden	223
16	Norway	44
17	Finland	139
...	(166 rows omitted)	140570
190	Djibouti	5
193	Qatar	3
196	Kosovo	19
222	Eastern Europe	3218
333	Other Unspecified Foreign Country	0
444	EU-Member State (unspecif.)	0
999	ethnic minority	2
-1	No Answer	2457
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

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. 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 (dataset PL) or the variations of the "biography / life history" questionnaires (dataset BIOL) and from the additional questionnaire for 16-17-year-olds in use since 2000 (dataset JUGENDL). In addition, information from PBRUTTO (Person-Related Gross File), the electronic household protocol for M1 or the retrospective survey of early childhood in the context of war and the post-war period (dataset BCBFK) was used. To code CORIGIN, all relevant information (see Table 1) available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD, PPATH or 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. 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):

1. 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".
2. 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.

The vast majority of the foreign-born population (see GERMBORN) who have ever been a part of a SOEP household (i.e., the population from PPFAD, PPATH or PPATHL) gave consistent information on their country of birth (see CORIGININFO). For around a third of the foreign-born population, no direct or inconsistent information on the person's country of birth was available. 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):

1. The respondents' country of birth which occurred most frequently, in other words the mode, was used.
2. Respondents' country of citizenship was used as their country of birth if both were not German. 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).
3. 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. Grandparents' country of birth and grandparents' citizenship were additionally used if information on mothers' and fathers' country of birth and citizenship were missing.
4. 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 a few 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. To provide the highest level of transparency possible, we include a variable for the quality of information used to create the country of birth variable: CORIGININFO.
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

corigininfo – Corigin: Quality of information

1	consistent information	161866

2	inconsistent information	4229
3	no answer	43780
4	filter germborn	954421
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

CORIGININFO indicates the quality of information given in CORIGIN. As in previous years, all relevant information available on persons who have ever been a part of a SOEP household (i.e., the population from PPFAD, PPATH or PPATHL) 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, contact: Selin Kara (Tel. +49 30-89789-345)

immiyear – Year Moved to Germany

1950		321
1951		348
1952		139
1953		189
1954		225
1955		132
1956		348
1957		458
1958		604
1959		586
1960		984
1961		1173
1962		1439
1963		1332
1964		2087
...	(42 rows omitted)	114129
2007		1461
2008		1502
2009		1935
2010		2077
2011		2473
2012		2898
2013		5551
2014		9596
2015		28912

2016	7159
2017	994
2018	236
2019	66
-1	20521
-2	954421

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 or PPATHL) and who were not born in Germany (see GERMBORN). The information on this variable was collected from the wave-specific individual questionnaires (dataset PL) or the variations of the "biography / life history" questionnaires (dataset BIOL) and from the additional questionnaire for 16-17-year-olds in use since 2000 (dataset JUGENDL). Since sample M (starting in 2013), 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 or the retrospective survey of early childhood in the context of war and the post-war period (dataset BCBFK) was used (both datasets are not included in the standard data distribution).
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 or PPATHL) gave consistent information on their year of immigration. For another part of the dataset no direct information on the person's year of immigration was available. "(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 or to respondents who were interviewed but did not answer the questions on their year of immigration. Over the course of the SOEP survey, only very few cases gave "(2) inconsistent information" with regard to their year of immigration. 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.
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):

1. 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.

2. 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.

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.

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 (see Table 2). To provide the highest level of transparency possible, we include a variable for the quality of information used to create the year of immigration variable: IMMIYEARINFO.

Table 2: Variations of the SOEP questions regarding respondents' year of immigration (main indicators for IMMIYEAR and IMMIYEARINFO)

Category	Question	Dataset(long format)	Sample	Years
Respondent information				
First	What year did you move to the Federal Republic of Germany (including West Berlin) for the first time?	PL	B	1984-1990 1992-1993
First	In which year did you move to Germany for the first time?	PL	B	1991
Unspecific	Since when have you lived in the area of today´s FRG or West Berlin? If after 1949, since when?	PL	A	1990
Unspecific	Since when have you lived in the area of the former FRG or West Berlin? If after 1949, since when?	PL	A	1991-1993
Unspecific	Since when have you lived in the area of the former GDR or East Berlin? If after 1949, since when?	PL	C	1992-1993
Unspecific	What year did you move to the Federal Republic of Germany (including West Berlin) for the first time?	BIOL	B, D	1994

Category	Question	Dataset(long format)	Sample	Years
Unspecific	When did you move to the Federal Republic of Germany?	BIOL	B, D	1995
Unspecific	Since when have you lived in the area of the former FRG or West Berlin? If after 1949, since when?	BIOL	A	1994-1995
Unspecific	When did you move to the Federal Republic of Germany?	BIOL	A-L3	1996-2012
Unspecific	When did you move to Germany?	BIOL	A-H, J-L2	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.	BIOL	A-H, J-L3, N-O	2014-2018
First	First of all we would like to know when you first moved away from your country of birth?	BIOL	M1-M2	2013-2018
First	Which country did you move to?	BIOL	M1-M2	2013-2018
First & Last	When did you move to Germany?	BIOL	M1-M2	2013-2018
First	First of all we would like to know when you first moved away from your country of birth?	BIOL	M3-M5	2016-2018
First	Was Germany the first country you moved to, or was it another country?	BIOL	M3-M5	2016-2018
First & Last	Did you move away from Germany again after that?	BIOL	M3-M5	2016-2018

Category	Question	Dataset(long format)	Sample	Years
First & Last	When did you move to Germany in this case?	BIOL	M3-M5	2016-2018
Unspecific	When did you arrive in Germany?	BIOL	M3-M5	2016-2018
Unspecific	When did you move to the Federal Republic of Germany?	JUGENDL	A-H, J-O	2000-2018
Unspecific	When did you move to Germany?	BCBFK	A-J	2012
Third party information				
Unspecific	Member of household: Moved into household from abroad.	PBRUTTO	all samples of SOEP	1985-2018
Unspecific	Did <first name> move to Germany? (reported by the anchor respondent)	ELECTR. HH. PROTOCOL	M1	2013
Unspecific	When did your father move to Germany?	BIOL	M1-M5	2013-2018

Source: v35
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

immiyearinfo – Immiyear: Quality of information

1	consistent information	159954
2	inconsistent information	988
3	no answer	48933
4	filter germborn	954421
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

IMMIYEARINFO indicates the quality of information given in IMMIYEAR. As in previous years, all relevant information available on persons who have ever been a part of a SO-

EP household (i.e., the population from PPFAD, PPATH or PPATHL) was compiled into a working dataset and compared to code IMMIYEAR. IMMIYEARINFO indicates whether "(1) consistent", "(2) inconsistent" or "(3) no information" 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). 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.
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

migback – Migration background

1	no migration background	838049
2	direct migration background	209875
3	indirect migration background	116372
4	migration background, not differentiable	0
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

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 or 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 1 under sub-heading "citizenship and legal status" and sub-heading "family information"), and the relevant biographical data sets (dataset BIOIMMIG). The variables were also updated using information from the wave-specific individual questionnaires (dataset PL), the variations of the "biography / life history" questionnaires (dataset BIOL), and the additional questionnaire for 16-17-year-olds in use since 2000 (dataset JUGENDL).
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". Grandparental information were additionally used if information on mothers' and fathers' migration background 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.
In a few cases, "(1) no (grand-)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 1 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.
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

miginfo – Migback: Quality of information

1	No (grand-)parental information	345436
2	At least 1 (grand-)parental information available	818860
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

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 (grand-)parental information" or "(2) At least 1 (grand-)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 1 under sub-heading "family information") and auxiliary citizenship variables (for more information, see Table 1 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) "At least 1 (grand-)parental information available" on MIGINFO and (1) "Consistent information" on GERMBORNINFO (around half of the PPATH cases). In contrast, the quality of information given on MIGBACK is considered relatively uncertain in cases where parental information ((1) "No (grand-)parental information" on MIGINFO) and respondents' information was missing ((3) "No information" on GERMBORNINFO)).
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

arefback – Refugee Experience

1	without evidence of refugee experience	1075203

2	with evidence of direct refugee experience	63350
3	with evidence of indirect refugee experience	12792
-1	No Answer	12951
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

The variable indicates asylum seekers and refugees – like for MIGBACK it is differentiated according to direct and indirect (later born children) background. (more detailed information on generation and usage can be found in Krause/Glass 2019).
For more information, contact: Peter Krause (Tel. 030-89789-690)

arefinfo – arefback: Source of Information

0	without evidence of refugee experience	1075203
1	residence permit status (current)[current year]	13379
2	residence permit status (current)[past years]	0
3	residence permit status (bioimmig)	15708
4	Refugees Samples [M.] target person	10127
5	Refugees Samples [M.] direct refugee experience	22681
6	Partner information	540
7	children[MUM], direct refugee experience	795
8	children[P-MUM], direct refugee experience	14
9	children[HV], direct refugee experience	10
10	children[geby<=immy+5] indirect refugee experience	4110
11	children[geby<=immy+10] indirect refugee experience	1791
12	children[geby<=immy+10] indirect refugee experience	3187
13	Refugees Samples [M.] indirect refugee experience	3704
14	HH Head info [household entrance year]	39
15	GER with direct refugee experience [biimgrp]	57
-1	No Answer	12951
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

The variable indicates further differentiations for asylum seekers and refugees [15 categories] – (more detailed information on generation and usage can be found in Krause/Glass 2019).
For more information, contact: Peter Krause (Tel. 030-89789-690)

loc1989 – Where did you live in 1989?

1	East Germany (DDR) incl. East Berlin	193548
2	West Germany (FRG) incl. West Berlin	562436
3	Abroad (Ausland)	77312
-1	No Answer	50934
-2	Does not apply	280066
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

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

1. HID in PPATHL: Place of residence in the former FRG before German reunification
2. IMMIYEAR in PPATH: Respondents who first immigrated to Germany after 1989 were coded as living "(3) abroad" in 1989
3. IMMIYEAR, CORIGIN in PPATH: Respondents from countries holding agreements on labor recruitment with the FRG 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
4. IMMIYEAR, CORIGIN in PPATH: Respondents from countries holding agreements on labor recruitment with the GDR who immigrated to Germany before 1990 were assumed to have been living in the "(1) German Democratic Republic [GDR]" in 1989
5. 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)
6. SAMPREG in PPATHL & BRMOVEIN and SYEAR in BIORESID: Respondents who moved into their current dwelling in the former FRG or GDR before 1989
7. SAMPREG in PPATHL: 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:

1. PZUG in PBRUTTO: New entrants to the SOEP who previously lived in East Germany or abroad
2. BSSCHEND and BSSCHWO in BIOSOC: Place and year of the last school attended
3. PGRUPPE in PBRUTTO: Place of birth that was asked in 1995
4. PL: Country of origin GDR
5. PNAT_V2 in PBRUTTO: Citizens of (former) GDR
6. PL: Place of residence in 1984

7. BIOPAREN and PPATH: Parental residence in 1989 for individuals younger than 18 in 1989

For more information, contact: Selin Kara (Tel. +49 30-89789-345)

locinfo – Loc1989: Source / Quality of information

0	Respondent born after 1989	275855
1	Direct information	814520
2	Indirect information	15239
-1	No Answer	51425
-2	Does not apply	7257
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

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".
For more information, contact: Selin Kara (Tel. +49 30-89789-345)

sampreg – Current sample region (Berlin, West-East)

1	West-Germany	930503
2	East-Germany	231147
-1	No Answer	0
-2	Does not apply	2646
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

Place of residence in East- or West-Germany with regard to borders of 1990 in corresponding year (SYEAR). (For Berlin East-West-assignements are approximated by zip-codes)

pop – Sample Membership

1	Private HH, German HH-Head	840288
2	Private HH, Foreign HH-Head	164874
3	Institutional. HH, Collective accommodation, German HH-Head	3360
4	Institutional. HH, Collective accommodation, Foreign HH-Head	9159

5	Not Compl. Private HH, German HH-Head	101039
6	Not Compl. Private HH, Foreign HH-Head	33997
7	Not Compl. Institutional. HH, Collective accommodation, German HH-Head	705
8	Not Compl. Institutional. HH, Collective accommodation, Foreign HH-Head	696
-1	No Answer	0
-2	Does not apply	10178
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

POP was derived from WUM2 in HBRUTTO as well as PNAT_V* and STELL_V* (nationality and relationship to head of household in PBRUTTO). Missing values were imputed based on the person's history. Thus, the only admissible missing value is –2, meaning not applicable. This variable is therefore particularly important, as it enters into the determination of cross-sectional weights. The variable corresponds with HPOP in HPATHL. See also the description of NETTO.

sexor – Sexual Orientation

0	probably heterosexual	747735
1	probably bi/homosexual	10543
2	insufficient information	406018
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

The variable SEXOR combines information on the sexual orientation of respondents from various sources in the SOEP. In 2016 (wave BG), (1) a direct question about sexual orientation was introduced (self-rep). Questions on marital status in the SOEP distinguish between same-sex civil unions and different-sex marriages. This distinction has been introduced in the household questionnaire since waves 2002 (wave S), in the person questionnaire since 2011 (wave BB), and in the biographical questionnaire since 2012 (wave BC). Starting with these years respectively, we use information of (2) the head of household on marital status of all household members (civil-hh), information on the marital status (3) reported by individuals in the person questionnaire (civil-p), as well as reported (4) in the partnership biography (civil-bio). Finally, the SOEP team provides pointers to the partner of each person in the SOEP households since 1984 (see pgpartnr in pgen documentation or parid in ppathl documentation). Combining information on the gender of both partners cohabitating in the SOEP household provides (5) the final source of information on the sexual orientation of adults in the SOEP (pointer).
Self-reports on sexual orientation surveyed in 2016 distinguish between the response options heterosexual, bisexual, and homosexual. It is however impossible to clearly identify bisexual

respondents from data on same-sex and different-sex partnerships even in longitudinal studies like the SOEP. This is because some bisexual respondents may be observed at periods of no-cohabitation, only same-sex, and only different-sex partnerships. Without any observed change in the partner's gender, we are unable to identify respondents as bisexual. Our approach to this problem is as follows: first, we do not seek to distinguish between homo- and bisexuals in the generated SEXOR variable. That is, we code individuals with (at least) one observation of a same-sex partnership as homo/bisexual. We code individuals with information from at least two years (arbitrary threshold) on only different-sex relationships as heterosexual. Since bisexuals in stable/multiple different-sex partnerships are misclassified as heterosexuals instead of homo/bisexuals, we add the label "probably" to our generated variable to indicate that this information is potentially erroneous. In the case of no information on partnerships or only one year of information on different-sex partnerships we consider this insufficient to make any inferences on sexual orientation in these individuals on the basis of their observed partnerships.
Finally, the sexor variable integrates both the self-reported as well as the partnership-obtained information on sexual orientation.

sexorinfo – Sexual Orientation:Source of information

0	insufficient information	406018
1	pointer	129901
2	civil-self	1101
3	pointer, civil-self	611
4	civil-hh	13387
5	pointer, civil-hh	166781
6	civil-self, civil-hh	137
7	pointer, civil-self, civil-hh	82990
8	BIO	562
9	pointer, bio	91
10	civil-self, bio	82
11	pointer, civil-self, bio	23
12	civil-hh, bio	477
13	pointer, civil-hh, bio	211
14	civil-self, civil-hh, bio	0
...	(10 rows omitted)	353949
25	pointer, bio, self-rep	328
26	civil-self, bio, self-rep	233
27	pointer, civil-self, bio, self-rep	51
28	civil-hh, bio, self-rep	8
29	pointer, civil-hh, bio, self-rep	41
30	civil-self, civil-hh, bio, self-rep	0
31	pointer, civ-self, civ-hh, bio, self-r.	7314
-1	No Answer	0
-2	Does not apply	0
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

This integer variable indicates which sources of information coincide with the value of SEXOR for the respective respondent. Its digits in binary representation are to be interpreted as binary flags, according to the following scheme: 1=Pointer, 2=Marital status, 4=Relation to head of household, 8=Biography, 16=Self-reported. If SEXORINFO has the value 5=1x1+0x2+1x4, this means that partnership pointers and relationship to head of household variables indicate the sexual orientation which is coded in SEXOR. Similarly, a value of 16 indicates that the inference was drawn from the direct question about sexual orientation. The variable is labeled accordingly.

parid – Partner Person Number

Partner indicators have the purpose of defining couples in SOEP households and thus to make possible analyses on the dyadic level. Persons without spouse and (cohabitating) partner receive a missing code "-2" (=does not apply). Also, the variable PARTNER is coded 0, 3, 4, 5 in these cases. In couples, partner is the value of the unchanging person ID number (=PID) of the partner. The assignment of the partner ID within households is based on four sources of information: A question in the person-file, that asks (unmarried) respondents to identify their partner in the household (bhppnr in 2017) (plk0001 in pl), the household matrix reported by the head of household at the beginning of the interview (bhstell in 2017) (stell_v1 stell_v2 stell_h in pbrutto), the partnership biography in the lifehistory calendar reported by new respondents (see also, biomars), and self-reports on marital status and life events, such as marriage, move in with partner, separation, etc. In unclear cases, due to temporal non-response for instance, we also consider longitudinal information from previous and prospective waves. Moreover, PARID is self-consistent between two individuals. For analyses of partner relationships, this information can be used to link all persons with their respective partners, and all information on both partners can also be stored in a common dataset.

partner – Status Of Partnership

0	No partner	495639
1	Spouse, registered partner	458720
2	Partner	68912
3	Probably spouse, registered partner	1077
4	Probably partner	1980
5	not clear	3701
-1	No Answer	0
-2	Does not apply	134267
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

The variable PARTNER generated in the context of the partner identifier (PARID) to describe whether a person in a SOEP household has a partner in that household, and if so, the type of relationship existing between the partners. Relationships with persons outside the SOEP household are not covered by this variable. Code 0 is assigned to all single persons living in households and those with partners outside the household. Codes 1 to 4 describe relationships. To assign Codes 1 and 2, the partnership has to be definable from the perspective of

both partners unanimously. If conflicting information exists between partners, the codes 3 or 4 are assigned. If it is unclear whether an individual has no partner or whether she forms a couple with one other household member, we assign the code 5. Registered partnerships (civil unions) for same-sex couples were introduced in Germany in 2001. Though, registered partnerships are legally not equal to marriage, they are listed in the same category.

5 Weighting

pbleib – Inverse Staying Probability

inverse probability weights

phrf – Weighting factor

standard individual weights

phrf0 – Weighting factor for new samples (wave 1 of new sample)

individual weights for first wave of new samples

phrf1 – Weighting factor without new samples (wave 1)

individual weights without first wave of new samples

6 ADD TO CODEBOOK.CSV

birthregion_ew – Birth place: German Federal Land (East-West Version)

21	West-Germany	251913
22	East-Germany	114437
-1	No Answer	0
-2	Does not apply	797946
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

NA
NA
For more information, contact: NA

prgroup – Random Groups

0	174
1	143821
2	144046
3	147126
4	140666

5	146492
6	149317
7	146820
8	145834

NA
NA
For more information, contact: NA

rv_id – ID SUF pension insurance

90000013		36
90000021		36
90000030		34
90000048		36
90000056		28
90000064		29
90000072		24
90000080		36
90000099		36
90000102		33
90000110		36
90000129		36
90000137		36
90000145		36
90000153		36
...	(8119 rows omitted)	94794
90081366		8
90081374		8
90081382		8
90081390		8
90081404		8
90081412		8
90081420		8
-1	No Answer	0
-2	Does not apply	1068938
-3	Answer improbable	0
-4	Inadmissible multiple response	0
-5	Not included in this version of the questionnaire	0
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of Survey program	0

NA
NA
For more information, contact: NA

7 SUPERFLOUS IN CODEBOOK.CSV

There are variables mentioned in codebook.csv, which cannot be found in the dataset. You should delete those lines from codebook.csv or add them to the dataset. The variables are: phrfe, pbleibe, phrfe0, phrfe1.