1546 2025

SOEP Survey Papers
Series D – Variable Descriptions and Coding

SOEP-Core v40.1 – PPATHL: Person-Related Meta-Dataset

SOEP Group

DIW BERLIN

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

Editors:
Dr. Jan Goebel, DIW Berlin
Dr. Christian Hunkler, DIW Berlin
Prof. Dr. Philipp Lersch, DIW Berlin and Humboldt-Universität zu Berlin
Dr. Levent Neyse, DIW Berlin and Berlin Social Science Center (WZB)
Prof. Dr. Carsten Schröder, DIW Berlin and Freie Universität Berlin
Prof. Dr. Sabine Zinn, DIW Berlin and Humboldt-Universität zu Berlin

Please cite this paper as follows:
SOEP Group, 2025. SOEP-Core v40.1 – PPATHL: Person-Related Meta-Dataset . SOEP Survey Papers 1546: Series D – Variable Descriptions and Coding. Berlin: DIW Berlin/SOEP

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
© 2025 by SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Anton-Wilhelm-Amo-Straße 58
10117 Berlin
Germany

soeppapers@diw.de

DIW BERLIN The German Socio Economic Panel at DIW Berlin

SOEP-Core v40.1 – PPATHL: Person-Related Meta-Dataset

SOEP Group

2025

The file ppathl is part of a collection, which is released with doi:10.5684/soep.core.v40.1o.

Contents

1 General Information 4

2 Primary Key, Foreign Keys and Sample Information 4
pid – Never Changing Person ID . . . 4
syear – Survey Year . . . 4
hid – Current Wave HH Number . . . 5
psample – Sample Member . . . 5
cid – Original Houshold Number, Case ID . . . 6
rv_id – ID SUF pension insurance . . . 6

3 Survey History 6
eintritt – Year First Contacted, Netto=10-99 . . . 6
erstbefr – Year First Surveyed, Netto=10-99 . . . 7
austritt – Year Of Last Contact, Netto=10-99 . . . 8
letztbef – Year Of Last Survey, Netto=10-99 . . . 8
netto – Current survey status . . . 9
nett1 – Current survey status (old 1 digit) . . . 10
casemat – Case-Match, combined panel households . . . 11
piyear – Jahr des Interviews (Interview Year) . . . 11
reclin_iab_match – Available in SOEP-CMI-ADIAB . . . 12

4 Basic Demographic Information 13
sex – Gender . . . 13
gebjahr – Birth Year, 4-digit . . . 13
gebmonat – Month Of Birth . . . 14
todjahr – Year Died, 4 Digits . . . 14
todinfo – Year Died, Information Source . . . 15
birthregion – Birth place: German Federal Land . . . 16
birthregion_ew – Birth place: German Federal Land (East-West Version) . . . 17
germborn – Born in Germany . . . 18
germborninfo – Germborn: Quality of information . . . 21
corigin – Country Born In . . . 21
corigininfo – Corigin: Quality of information . . . 23
immiyear – Year Moved to Germany . . . 24
immiyearinfo – Immiyear: Quality of information . . . 28
migback – Migration background . . . 29
miginfo – Migback: Quality of information . . . 30
arefback – Refugee Experience . . . 31
arefinfo – arefback: Source of Information . . . 31
loc1989 – Where did you live in 1989? . . . 32
locinfo – Loc1989: Source / Quality of information . . . 34
sampreg – Current place of residence (FRG/GDR - with respect to borders from 1989) . . . 34
pop – Sample Membership . . . 34
sexor – Sexual Orientation . . . 35
sexorinfo – Sexual Orientation:Source of information . . . 36
parid – Partner Person Number . . . 37
partner – Status Of Partnership . . . 37

5 Weighting **38**
pbleib – Inverse Staying Probability 38
phrf – Weighting factor 38
phrf0 – Weighting factor for new samples (wave 1 of new sample) 38
phrf1 – Weighting factor without new samples (wave 1) 38
prgroup – Random Groups 38

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
...	(10 rows omitted)	311705
2009		31642

2010	45977
2011	50329
2012	50120
2013	55611
2014	51684
2015	50277
2016	57287
2017	64554
2018	62491
2019	62829
2020	63063
2021	57713
2022	74842
2023	86362

Together with PID primary key in this file.

hid – Current Wave HH Number

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)	287129
2	B 1984 Migration (until 1983, West)	96503
3	C 1990 Initial Sample (East)	136272
4	D 1994/5 Migration (1984-1994, West)	26956
5	E 1998 Refreshment	27316
6	F 2000 Refreshment	181987
7	G 2002 High Income	40392
8	H 2006 Refreshment	33310
9	I 2009 Innovation Sample	7130
10	J 2011 Refreshment	59625
11	K 2012 Refreshment	28346
12	L1 2010 Birth Cohort (2007-2010)	75017
13	L2 2010 Family Type (Low-Income, Single-Parent, Large Families)	78478
14	L3 2011 Family Type (Single-Parent, Large Families)	32563
15	M1 IAB-SOEP 2013 Migration (1995-2011)	58436
...	(9 rows omitted)	202674
25	M7 IAB-SOEP 2020 Migration	7545
26	M8 IAB-SOEP 2020 Migration Professionals	8072
27	R 2022 Refreshment	33648
28	M8b IAB-SOEP 2022 Refreshment	8827
29	M8c IAB-SOEP 2023 Refreshment non-EU countries	2063
30	M9 IAB-BAMF-SOEP 2023 Refugee	13977
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0

-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

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

cid – Original Houshold Number, Case ID

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.

rv_id – ID SUF pension insurance

-1	No answer	0
-2	Does not apply	1223278
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

3 Survey History

eintritt – Year First Contacted, Netto=10-99

1984	257725
1985	6395
1986	7812
1987	7264
1988	7241
1989	7618
1990	110344
1991	9582
1992	8678
1993	8303
1994	18122
1995	17838
1996	7775
1997	6928

1998		30902
...	(10 rows omitted)	289863
2009		11398
2010		142873
2011		89697
2012		32375
2013		59784
2014		7355
2015		21681
2016		66835
2017		67869
2018		17715
2019		31224
2020		30849
2021		3357
2022		43739
2023		17125

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		192908
1985		7977
1986		8650
1987		7858
1988		6880
1989		7535
1990		82232
1991		9070
1992		8566
1993		8430
1994		15255
1995		15462
1996		8646
1997		8388
1998		26387
...	(11 rows omitted)	273278
2010		70378
2011		60384
2012		28684
2013		41262
2014		11077
2015		18788
2016		33560
2017		47222
2018		18133
2019		22851

2020	20936
2021	9556
2022	27335
2023	13990
-2	334588

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		3064
1988		4819
1989		5027
1990		3524
1991		4369
1992		5116
1993		6738
1994		7955
1995		8242
1996		8674
1997		8074
1998		10553
1999		11467
...	(9 rows omitted)	163493
2009		38150
2010		38477
2011		42877
2012		31627
2013		36662
2014		34311
2015		41611
2016		45205
2017		42039
2018		51388
2019		61160
2020		57470
2021		52545
2022		53501
2023		561615

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		3048
1987		4404
1988		4129
1989		4035
1990		4545
1991		5354
1992		5912
1993		7161
1994		7286
1995		7131
1996		8183
1997		10426
1998		10983
...	(11 rows omitted)	205719
2010		29683
2011		38232
2012		27913
2013		31677
2014		31500
2015		33671
2016		35797
2017		41253
2018		43898
2019		40851
2020		112672
2021		52571
2022		83647
2023		213394
-2		334588

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)	583021
12	Individual Questionnaire And Person Biography	145112
15	Individual Questionnaire And Experiments, Test	84088
16	Individual Questionnaire, First Time Surveyed, Age 17	5992
19	Individual Questionnaire Without Household Interview	2147
20	Children in Succesfully Interviewed Households (_Kind)	3109
21	Children in Kidlong	285854
22	Children (only) in CHILDL	7474
30	Persons In Successfully Interviewed HH Without Individual Interview	80807
31	Successful Gap Interview (_LUECKE)	4423
32	Successfully Completed Biography Questionnaires	62
34	Successful Tests and Experiments	122

35	Part. Success, without HH interview	2733
40	Person in non completed gross HH	159767
41	Gap Interview (_LUECKE) in non completed HH	6634
...	(12 rows omitted)	67465
92	Moved abroad (abroad)	174
93	Moved abroad	65
94	Person Gap with advices	423
97	advice to dead person	981
98	advice to dead person (_VP)	1355
99	Has Died	4434
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	24
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

This variable indicates available information and files for the entire SOEP individuals. Netto-codes 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	16340
1	Successful Interview _P, _JUGEND	818213
2	Below Survey Age _KIND	296430
3	Did Not Participate _PBRUTTO	296162
4	Missing This Wave _PLUECKE	15903
5	Interviewee Without Household Interview	2147
-1	No answer	0
-2	Does not apply	1071
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	159
20605		2
20613		5
27367		14
27430		7
250996		3
272906		6
277908		1
283924		9
291102		7
292338		15
292621		7
344095		14
700495		3
701564		13
...	(4 rows omitted)	8
3014079		2
3094650		9
3416240		4
3499749		7
3635481		9
3920330		9
-1	No answer	0
-2	Does not apply	1445953
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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)

0	168
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
...	(13 rows omitted)	402848
2011		50254
2012		49401
2013		54819
2014		50734
2015		49465
2016		56395
2017		58307
2018		64115
2019		64604
2020		59421
2021		31000
2022		39103
2023		39946
2024		166
-2		131482

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

reclin_iab_match – Available in SOEP-CMI-ADIAB

0	No administrative data available	1112962
1	Administrative data available	333304
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

This variable indicates whether respondents could be successfully linked to administrative data of the Institute for Employment Research (IAB). From 2013 onward, SOEP respondents were asked for record linkage consent and, if consent was given, this information is written into the variable for the same person id over the past years. Respondents that were not asked for consent are assigned the missing value -5. For more information, refer to data product SOEP-CMI-ADIAB.

For more information, contact: Mattis Beckmannshagen (Tel. +49-30-89789-321, mbeckmannshagen@diw.de)

4 Basic Demographic Information

sex – Gender

1	Male	715231
2	Female	729180
-1	No answer	1732
-2	Does not apply	0
-3	Implausible value	123
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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		178
1901		164
1902		304
1903		201
1904		363
...	(105 rows omitted)	1347588
2010		15672
2011		10319
2012		9479
2013		9048
2014		8117
2015		7598
2016		7103
2017		6062

2018	4857
2019	3870
2020	2721
2021	1890
2022	1099
2023	357
-1	8875

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

gebmonat – Month Of Birth

1	January	134209
2	February	102902
3	March	112727
4	April	101708
5	May	106146
6	June	99819
7	July	108569
8	August	104340
9	September	105058
10	October	100465
11	November	91889
12	December	95165
-1	No answer	71983
-2	Does not apply	5454
-3	Implausible value	0
-4	Inadmissable multiple response	0
-5	Not included in this version of the questionnaire	105832
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of survey	0
-9	Missing due to a terminated interview	0

The month of birth of a respondent is created with information from net and gross datasets (i.a. information from questionnaires header or household protocol). For about 96% of the respondents, there was no discrepancy between this information. In 3.3% of the respondents, the mode of the month of birth is used when there is differing information. For the rest, the most plausible/newest information was used (for less than 1%). For about 20% of respondents there was no information about the month of birth at all, but for about 1000 persons it was possible to use indirect information, such as the move-in date of a newborn child.
For more information, contact: Andreas Franken (Tel. 030-89789-331)

todjahr – Year Died, 4 Digits

1984		14
1985		153
1986		298
1987		405
1988		527
1989		636
1990		592
1991		865
1992		923
1993		1156
1994		1085
1995		1663
1996		1585
1997		1486
1998		1682
...	(12 rows omitted)	34652
2011		2513
2012		2290
2013		3260
2014		2923
2015		2777
2016		3572
2017		3427
2018		3437
2019		2505
2020		2889
2021		1063
2022		3202
2023		2229
-1		357
-2		1362100

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	67898
2	survey about died person ($v)	434
3	survey about parents ($lela)	15
4	Infratest drop-out study 1992	17
5	Infratest drop-out study 2001	4045
6	Infratest drop-out study 2007	33
7	Infratest drop-out study 2008/9	8173

8	Modul Family changes [P]	3194
-1	No answer	357
-2	Does not apply	1362100
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	12759
2	Hamburg	8587
3	Lower Saxony	42757
4	Bremen	3873
5	North Rhine-Westphalia	90964
6	Hesse	28161
7	Rhineland-Palatinate	21217
8	Baden-Wuerttemberg	50131
9	Bavaria	65288
10	Saarland	4671
11	Berlin	18377
12	Brandenburg	21225
13	Mecklenburg-West Pomerania	13562
14	Saxony	42530
15	Saxony-Anhalt	25047
16	Thuringia	22608
-1	No answer	41023

-2	Does not apply	923242
-3	Implausible value	0
-4	Inadmissable multiple response	0
-5	Not included in this version of the questionnaire	10244
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of survey	0
-9	Missing due to a terminated interview	0

BIRTHREGION 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)

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

21	West-Germany	309870
22	East-Germany	137077
-1	No answer	122066
-2	Does not apply	834548
-3	Implausible value	0
-4	Inadmissable multiple response	0
-5	Not included in this version of the questionnaire	42705
-6	Version of questionnaire with modified filtering	0
-7	Only available in less restricted edition	0
-8	Question this year not part of survey	0
-9	Missing due to a terminated interview	0

BIRTHREGION_EW contains information about whether a person is born in eastern or western part of Germany. East Germany contains federal states of: - Mecklenburg-West Pomerania - Brandenburg - Saxony - Saxony-Anhalt
- Thuringia - Berlin (East & West Berlin)
West Germany contains the federal states of: - Schleswig-Holstein - Hamburg - Lower Saxony - Bremen
- North Rhine-Westpalatinate - Baden-Wuerttemberg - Bavaria - Saarland

For more information, contact: Marvin Petrenz (Tel.: +49 30 89789 463)

germborn – Born in Germany

1	born in Germany or immigr.<1950	1124719
2	not born in Germany	321547
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	
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	

Information used	Dataset (long format)
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
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	1084637
2	inconsistent information	62877
3	no answer	298752
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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

4	Afghanistan	14035
8	Albania	2491
10	Antarctica	0
12	Algeria	332
16	American Samoa	7
20	Andorra	0
24	Angola	161
28	Antigua and Barbuda	0
31	Azerbaijan	768
32	Argentina	348
36	Australia	198
40	Austria	2783
44	Bahamas	2
48	Bahrain	7
50	Bangladesh	408
...	(228 rows omitted)	1406937
862	Venezuela	251

876	Wallis and Futuna	0
882	Samoa	29
887	Yemen	200
894	Zambia	21
900	Kosovo	4634
-1	No answer	12654
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	223985
2	inconsistent information	16186
3	no answer	81376
4	filter germborn	1124719
-1	No answer	0

-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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		322
1951		356
1952		147
1953		195
1954		225
1955		134
1956		374
1957		492
1958		645
1959		623
1960		1012
1961		1188
1962		1523
1963		1377
1964		2154
...	(46 rows omitted)	135382
2011		3300
2012		4052
2013		7923
2014		14660
2015		49236
2016		16119
2017		6496
2018		5129
2019		4103
2020		1275
2021		3024

2022	4293
2023	182
-1	55606
-2	1124719

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

Category	Question	Dataset(long format)	Sample	Years
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
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	221689
2	inconsistent information	2380
3	no answer	97478

4	filter germborn	1124719
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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 SOEP 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	986702
2	direct migration background	321547
3	indirect migration background	138017
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	499299
2	At least 1 (grand-)parental information available	946967
-1	No answer	0
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	1267641
2	with evidence of direct refugee experience	125197
3	with evidence of indirect refugee experience	21090
-1	No answer	32338
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

Complementing the more general variables on migration background (MIGBACK and MIGINFO), the variables AREFBACK and AREFINFO provide biographical information on individual asylum and refugee background. These variables are generally time-independent and have been defined for the entire SOEP population since 1984. They are available in all path files (PPFAD, PPATH and PPATHL).
In addition to the data sources described in more detail in AREFINFO, direct and indirect refugee background is partly identified from path files using migration-related variables, specifically GERMBORN, CORIGIN and MIGBACK. Please note that respondents with a migration background may have either a direct or indirect refugee background, or neither. Additionally, respondents with a refugee background may not always be categorised as having a migration background, since these categories are based on distinct criteria. For example, respondents born in Germany without a migration background may have refugee status in relation to the former German Democratic Republic (GDR).
For more information, contact: Philippa Cumming (Tel. 030-89789-385)

arefinfo – arefback: Source of Information

0	without evidence of refugee experience	1267641
1	residence permit status (current)[current year]	9720
2	residence permit status (current)[past years]	44204
3	residence permit status (bioimmig)	13014

4	Refugees Samples [M.] target person	8319
5	Refugees Samples [M.] direct refugee experience	34130
6	Partner information	915
7	children[MUM], direct refugee experience	8345
8	children[P-MUM], direct refugee experience	4380
9	children[HV], direct refugee experience	1664
10	children[geby<=immy+5] indirect refugee experience	8371
11	children[geby<=immy+10] indirect refugee experience	2740
12	children[geby<=immy+10] indirect refugee experience	3987
13	Refugees Samples [M.] indirect refugee experience	5992
14	HH Head info [household entrance year]	397
15	GER with direct refugee experience [biimgrp]	109
-1	No answer	32338
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

The variable AREFINFO illustrates the more distinct categories that comprise the categories on refugee background. In addition to the migration variables listed above, direct refugee background is further determined using biographical information on legal status from the BIOIMMIG dataset, as well as yearly information on asylum application status and residence permits from the PL dataset. Additional indicators are taken from the PBRUTTO dataset (nationality, head of household and household anchor persons), the PPFAD dataset (sample type), the PGEN dataset (pgstatus_asyl and pgstatus_refu variables for the years 2016 and 2017) and partner information from the PPATHL dataset. This information defines categories 1 to 8 and 14 and 15, all of which are classified as having a direct refugee background. The sample type is also used alongside GERMBORN to identify respondents born in Germany with an indirect refugee background (category 13).
Children with a direct and indirect refugee background are identified using the dataset KIDLONG, utilising migration information relating to the mother, her partner and the head of the household. Information on the child's country of birth and the year of birth relative to the year of immigration of the mother, her partner and the head of the household is used to distinguish between those with direct (category 9) and indirect (10 – 12) refugee backgrounds.
For more information, contact: Philippa Cumming (Tel. 030-89789-385)

loc1989 – Where did you live in 1989?

1	East Germany (DDR) incl. East Berlin	221360
2	West Germany (FRG) incl. West Berlin	621487
3	Abroad (Ausland)	115424
-1	No answer	74961
-2	Does not apply	413034
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	413034
1	Direct information	937343
2	Indirect information	20928
-1	No answer	74961
-2	Does not apply	0
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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 place of residence (FRG/GDR - with respect to borders from 1989)

1	West-Germany	1161215
2	East-Germany	282324
-1	No answer	0
-2	Does not apply	2727
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	0

Place of residence in East- or West-Germany (with regard to borders of 1989) in corresponding year (SYEAR). (For Berlin East-West-assignements are approximated by zip-codes.)

pop – Sample Membership

1	Private HH, German HH-Head	957409
2	Private HH, Foreign HH-Head	224771
3	Institutional. HH, Collective accommodation, German HH-Head	4383
4	Institutional. HH, Collective accommodation, Foreign HH-Head	16199
5	Not Compl. Private HH, German HH-Head	152039

6	Not Compl. Private HH, Foreign HH-Head	71793
7	Not Compl. Institutional. HH, Collective accommodation, German HH-Head	1124
8	Not Compl. Institutional. HH, Collective accommodation, Foreign HH-Head	7030
-1	No answer	0
-2	Does not apply	11518
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	854723
1	probably bi/homosexual	14897
2	insufficient information	565946
-1	No answer	0
-2	Does not apply	10700
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	565946
1	pointer	106276
2	civil-self	1241
3	pointer, civil-self	567
4	civil-hh	65027
5	pointer, civil-hh	155597
6	civil-self, civil-hh	8520
7	pointer, civil-self, civil-hh	76488
8	BIO	582
9	pointer, bio	91
10	civil-self, bio	86
11	pointer, civil-self, bio	23
12	civil-hh, bio	529
13	pointer, civil-hh, bio	184
14	civil-self, civil-hh, bio	97
...	(11 rows omitted)	443941
26	civil-self, bio, self-rep	314
27	pointer, civil-self, bio, self-rep	58
28	civil-hh, bio, self-rep	20
29	pointer, civil-hh, bio, self-rep	134
30	civil-self, civil-hh, bio, self-rep	31
31	pointer, civ-self, civ-hh, bio, self-r.	9814
-1	No answer	0
-2	Does not apply	10700
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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	749842
1	Spouse, registered partner	590428
2	Partner	93089
3	Probably spouse, registered partner	1571
4	Probably partner	2221
5	not clear	5413
-1	No answer	0
-2	Does not apply	3702
-3	Implausible value	0
-4	Inadmissable 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	0
-9	Missing due to a terminated interview	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

prgroup – Random Groups

0	809
1	178657
2	178975
3	182701
4	175513
5	181380
6	184866
7	182435
8	180930