Steinhauer, Hans Walter; Trübswetter, Parvati; Zinn, Sabine

Research Report

SOEP-Core - 2020: Sampling, nonresponse, and weighting in the IAB-SOEP migration studies M7 and M8

SOEP Survey Papers, No. 1105

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

Suggested Citation: Steinhauer, Hans Walter; Trübswetter, Parvati; Zinn, Sabine (2022) : SOEP-Core - 2020: Sampling, nonresponse, and weighting in the IAB-SOEP migration studies M7 and M8, SOEP Survey Papers, No. 1105, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
http://hdl.handle.net/10419/261441

1105 2022

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

SOEP-Core – 2020: Sampling, Nonresponse, and Weighting in the IAB-SOEP Migration Studies M7 and M8

Hans-Walter Steinhauer, Parvati Trübswetter, Sabine Zinn

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:

Hans Walter Steinhauer, Parvati Trübswetter, Sabine Zinn. 2022. SOEP-Core – 2020: Sampling, Nonresponse and Weighting in the IAB-SOEP Migration Studies M7 and M8. SOEP Survey Papers 1105: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

SOEP-Core – 2020: Sampling, Nonresponse, and Weighting in the IAB-SOEP Migration Studies M7 and M8

Hans Walter Steinhauer[1], Parvati Trübswetter[2], and Sabine Zinn[1,3]

[1]Deutsches Institut für Wirtschaftsforschung
[2]Institut für Arbeitsmarkt- und Berufsforschung
[3]Humboldt-Universität zu Berlin

July 18, 2022

Abstract
This paper provides details on sampling design, fieldwork, nonresponse and population adjustments for the 2020 samples M7 and M8 of the Socio-Economic Panel (SOEP). Sample M7 refreshes the SOEP core sample, especially samples M1 and M2, including households with household heads having a Bulgarian, Polish, or Romanian nationality. The sample M8 augments the SOEP core sample, sampling households of foreigners from third countries to evaluate the "Fachkräfteeinwanderungsgesetz." Obtaining nearly 1,000 household interviews and panel consent of households for each sample was complicated by the first wave of the Corona pandemic and the first lockdown. Nevertheless, nonresponse on the household level is driven by a variety of characteristics, such as nationality or regional contexts as well as information contained in the Integrated Employment Biographies.

1 Introduction

Starting with sample M1 in 2013 the Institute for Employment Research (IAB) in Nuremberg and the German Socio-Economic Panel (SOEP) at DIW Berlin began to jointly survey the migrant population in Germany (Liebig et al., 2021). In 2015, sample M2 became the follow-up sample, where both institutes continued their cooperation.
The sample M7 refreshes the SOEP core sample, especially samples M1 (Kroh, Kühne, Goebel, & Preu, 2015) and M2 (Kühne & Kroh, 2017), with households including persons with Bulgarian, Polish, or Romanian nationalities. As before, we use the Integrated Employment Biographies (IEB) as a nationwide sampling frame. The IEB is spell data based on IAB's employment history (BeH), IAB's benefit recipient history (LeH), the participants-in-measures data (MTG), and job search data originating from the applicants pool database (BewA). Thus, the IEB include observations of unemployment benefits, job search, and participation in active labor market programs, see Oberschachtsiek, Scioch, Seysen, and Heining (2009) for details. Beyond that, it covers socio-demographic information on gender, age, and nationality as well as geographic information, including, for example, postal codes, municipality, and regional classification. Moreover, the cooperation allows us to link information from the IEB to the SOEP data and use them, for example, in the nonresponse analyses. Moreover, the Federal Employment Agency (Bundesagentur für Arbeit) provides information on third-country nationals who applied for working in Germany as professionals ("Fachkräfte") based on the Residence Act (Zuwanderungsgesetz, ZuwG). This information is used to identify the population in the IEB data. The sample from this population is referred to as M8. It was sampled and surveyed in 2020 in order to provide a basis to evaluate the "Fachkräfteeinwanderungsgesetz" – Skilled Immigration Act – which became effective March 1st, 2020.

This paper documents the sampling design and the weighting strategy for the 2020 samples M7 and M8 of the SOEP. Therefore, section 2 provides details on the population. Sampling is described in section 3. Section 4 provides information on the fieldwork and its results. Weighting adjustments are presented in section 5. Finally, section 6 gives a brief summary.

2 Target Population and Sampling Frame

The target population of sample M7 consists of persons living in private households having a Bulgarian, Polish, or Romanian nationality, who immigrated to Germany between January 1, 2016, and December 31, 2018. The target population of sample M8 consists of third-country nationals living in private households who were granted a permission to work in Germany as professionals between January 1, 2019, and January 30, 2020. To sample from these two populations, we make use of the IEB data, which is official data provided by the IAB. In total the IEB contain 917,835 persons belonging to the population of M7 and 112,945 to the population of M8. We further restrict the persons to those having a valid address in Germany. For M7 (M8) this reduces the number of persons by 146,743 (29) to a target population of 771,092 (112,916) persons. In case of the M7 sample, the reduction is mostly driven by persons working in Germany but having an address in regions near the border outside of Germany, for example, in Poland. The number of persons with valid address information within Germany contained in the IEB

data by group for the two samples is displayed in Table 1a for M7 and in Table 1b for M8.

Table 1: Number of target persons for in the IEB data by group.

(a) Subsample M7

Nationality	Number	Percent
Bulgarian	132,154	17.1
Polish	243,566	31.6
Romanian	395,372	51.3
Total	771,092	100.0

(b) Subsample M8

Application for	Number	Percent
Qualified employment	50,430	44.7
Unqualified employment	32,114	28.4
Education	14,268	12.6
Other	16,104	14.3
Total	112,916	100.0

The table clearly shows that the majority of the M7 target population has immigrated from Romania (slightly over 50%); the smallest share originates from Bulgaria (around 17%). These immigrants scatter unequally across Germany. Throughout the 8,171 German postal code areas, there are several regions along the border and within the country, especially in the eastern part of Germany, where there are fewer than 50 immigrants. These are most likely to be regions that people commute to rather than move there, if they have a job in these regions. The majority of immigrants is located in urban and agrarian-oriented areas in the western and southern regions of Germany as well as in Berlin and its neighboring regions in Brandenburg. In contrast to the M7 population, the majority of the M8 target population is located in urban areas. Here, their number is highly correlated to the number of people forming the M7 population.

3 Sampling Design

The sampling design can be summarized as a stratified multi-stage sampling design. Because the distribution of the immigrant population for M7 containing $N = 770{,}709$ individuals is unequally spread over Germany and within states, we form regional clusters of postal code areas as primary sampling units (PSU) stratified by federal states and a rural-urban-classification; strata $h = 1, \ldots, H$.[1] The PSU , indexed $j = 1, \ldots, M_h$, were constructed to cover a minimum of at least 600 and no more than 1,200 immigrants, that is, $600 \geq N_{jh} \leq 1{,}200$. In a first step, each postal code exceeding the minimum number of 600 immigrants became a PSU itself. This size is chosen in order to guarantee the minimum number of 60 immigrants from Bulgaria, who form the smallest group, compare Table 1. In a second step, a postal code was selected at random and the neighboring postal codes were attached until the minimum number was achieved. In the third step, all remaining postal codes as well as those PSU exceeding 1,200 immigrants were again split, then the second step was repeated. In the last step, all remaining postal codes were attached to the neighboring PSU that covered the least number of immigrants. This procedure clustered a total of 8,171 postal codes into $M = \sum_{h=1}^{H} = 773$ PSU. In the first stage, $m = 125$ PSUs had to be selected with systematic probability proportional to size

[1]The number of the target population is reduced by another 383 persons because they were born before 2002, thus most likely not yet of legal age.

sampling. To balance between urban and agrarian-oriented areas, the latter were reduced in weight for sampling by the factor $p_{jh} = 0.5$. The measure of size (x_{jh}) for PSU j in stratum h is $x_{jh} = p_{jh} \cdot N_{jh}$ and the stratum-specific measure of size x_h is $x_h = \sum_{j=1}^{M_h} x_{jh}$. The number of PSUs to select from each stratum was allocated proportional to the measure of size per stratum, that is $m_h = m \cdot \frac{x_h}{\sum_{h=1}^{H} x_h}$. Thus, the inclusion probability π_{jh} for PSU j in stratum h is given by

$$\pi_{jh} = m_h \cdot \frac{x_{jh}}{\sum_{j=1}^{M_h} x_{jh}} = m_h \cdot \frac{p_{jh} \cdot N_{jh}}{\sum_{j=1}^{M_h} p_{jh} \cdot N_{jh}} \tag{1}$$

Within each of the sampled PSU, a simple random sample of $n_s = 60$ immigrants was drawn from each nationality s. Thus, the inclusion probability for immigrant i of nationality s is

$$\pi_{is} = \frac{\min(n_s, N_{sjh})}{N_{sjh}} \tag{2}$$

such that the final inclusion probability π_{isjh} for immigrant i of nationality s sampled in PSU j in stratum h is

$$\pi_{isjh} = \pi_{is} \cdot \pi_{jh}. \tag{3}$$

This sampling procedure yields a maximum sample of $n = m \cdot 3 \cdot n_s = 125 \cdot 3 \cdot 60 = 22{,}500$

To realize the sample for M8 we make use of the high correlation in the number of individuals in the two populations. Because of this, we were able to use the same PSUs formed for sampling households for M7. A previous simulation study showed that using the same PSUs will provide a sample of sufficient size for M8, too. The final samples drawn from the two populations include 22,020 individuals for M7 and 21,552 for M8.

4 Fieldwork Results and Response Rates

After sampling, the addresses were handed over to KANTAR Public, the field work agency, and were validated. During the fieldwork, a total of 19,751 addresses were validated for M7 and 12,992 for M8. This left 2,269 addresses in M7 and 8,560 in M8 unused. The validation yielded a noticeable number of invalid or old addresses that were not eligible. Of these, the largest number was untraceable and a huge number also had moved abroad. We find this very likely for the following reasons. First, many individuals from Poland, Bulgaria, and Romania come to Germany as seasonal workers. Second, both populations are likely to have moved back to their home country because of the Covid-19 pandemic. For these reasons, only 8,173 addresses were visited by interviewers for the M7 subsample and 7,804 for M8. Table 3 displays the results for the fieldwork. In total, there were 783 complete or partial interviews in M7 and 1,096 in M8 resulting in a response rate on the household-level, calculated according to American Association for Public Opinion Research (2016), of $RR2_{M7} = 0.096$ for M7 and $RR2_{M8} = 0.141$. The response rate at the household-level is quite low, but as expected because of the underlying populations and the

Table 2: Number of target persons, postal codes and PSU by Federal State.

Federal	Target persons		Number of		
State	M7	M8	postal codes	PSU in population	PSU in sample
BB	14,933	1,594	215	18	2
BE	29,982	8,207	190	24	5
BW	120,514	23,434	1,194	131	20
BY	138,846	26,482	2,062	145	22
HB	7,345	,643	40	7	1
HE	70,303	11,544	544	74	11
HH	15,488	3,058	100	13	3
MV	9,717	1,027	189	10	2
NI	86,136	6,685	796	80	14
NW	158,092	17,429	865	154	26
RP	50,027	4,591	659	48	8
SH	23,477	2,241	445	23	4
SL	5,999	,545	69	6	1
SN	13,247	2,949	384	14	2
ST	12,687	1,159	201	12	2
TH	13,916	1,328	218	14	2
Total	770,709	112,916	8,171	773	125

Note: BW = Baden-Württemberg, BY = Bavaria, BE = Berlin, BB = Brandenburg, HB = Bremen, HH = Hamburg, HE = Hessen, MV = Mecklenburg-Vorpommern, NI = Lower Saxony, NW = North Rhine-Westphalia, RP = Rhineland-Palatinate, SL = Saarland, SN = Saxony, ST = Saxony-Anhalt, SH = Schleswig Holstein, TH = Thuringia.

Covid-19 pandemic. The refusal rate ($REF1$) is similar when compared to other samples / studies. For M7 the refusal rate is $REF1_{M7} = 0.225$ and for M8 it is $REF1_{M8} = 0.213$. For more detailed information on the fieldwork see Rathje and Glemser (2021).

5 Cross-sectional Weighting

The computation of survey weights is usually performed in three steps (Brick & Kalton, 1996). In the first step, design weights are calculated as inverse of the inclusion probability, see Section 3. Second, these design weights are adjusted to correct for unit nonresponse. This step is referred to as sample weighting adjustment by Kalton and Kasprzyk (1986). Lastly, weights are calibrated so that estimates conform to known population parameters or to meet specific distributions. Kalton and Kasprzyk (1986) refer to this step as population weighting adjustment. For details on the general weighting strategy of the SOEP and the integration of new samples, see Kroh, Siegers, and Kühne (2015).

To account for possible selectivity due to nonresponse, we model the participation decision of the households using information on participating and nonparticipating households. Because there usually is only limited information available on nonparticipating households, we use area level information as well as interviewer observations on the residential environment. Information collected by the interviewer on the residential environment include:

Table 3: Fieldwork results on the household-level according to American Association for Public Opinion Research (2016).

Final Disposition Code	M7 Number	M7 Percent	M8 Number	M8 Percent
1. Interview				
(1.1) Complete	301	0.022	700	0.062
(1.2) Partial	482	0.035	396	0.035
Subtotal	*783*	*0.057*	*1,096*	*0.097*
2. Eligible, Non-Interview				
(2.11) Refusals	1,840	0.135	1,665	0.147
(2.20) Non-contact	3,960	0.290	3,712	0.328
(2.31) Dead	31	0.002	10	0.001
(2.32) Physically/mentally unable/incompetent	2	0.000	2	0.000
(2.33) Language	84	0.006	177	0.016
(2.36) Miscellaneous	403	0.029	419	0.037
Subtotal	*6,320*	*0.463*	*5,985*	*0.529*
3. Unknown eligibility, non-interview				
(3.11) Not attempted or worked	1,069	0.078	719	0.063
4. Not Eligible				
(4.0) Not Eligible	1,075	0.079	694	0.061
(4.2) Household moved abroad	2,194	0.161	1,150	0.102
(4.4) Household untraceable	2,223	0.163	1,680	0.148
Subtotal	*5,492*	*0.402*	*3,524*	*0.311*
Total	13,664	1.000	11,324	1.000

Note: Subtotals might not add up because of errors due to rounding.

problems with speaking German, condition of the housing area, condition of the house, access problems by barriers, access problems by intercom system, other access problems, safety of the housing area, composition of the housing area, and type of house (according to number of residential parties). Area level information is obtained from INKAR online (Indikatoren und Karten zur Raum- und Stadtentwicklung; www.inkar.de) on the district level. INKAR provides information on (un)employment, construction and housing, education, infrastructure, population characteristics, and other regional indicators. The time reference of INKAR data is 2017. Detailed documentation of the variables in the data is provided by (INKAR, 2019). Lower level information used in the nonresponse analysis is provided by Microm, typically on the street level (www.microm.de). Microm provides information about social structure of neighborhoods in Germany on the regional and local levels. Local level covers different aggregations; for instance, eight digit postal code areas (PLZ8) covering approximately 500 households, street level, or household cells aggregating a few households. Finally, we are able to link some information from the IEB data, such as the date of a person's first and last spell in the IEBs, the number of spells a person has in the IEBs, the persons date of birth (and the derived age), a person's nationality, the source a person's first, last, and most frequent spell originates from, a person's highest educational degree, and whether or not a person has an apprenticeship spell.

5.1 Sample Weighting Adjustment

When correcting the design weights in the second step, strong predictors for nonresponse are needed. For this purpose, we use the information detailed above. Not all of these variables enter the corresponding nonresponse model. The reason is obvious: of these variables, only a few turn out to significantly influence the participation decision. Further, some might also be highly correlated among each other. Using unnecessary explanatory variables in the model will only increase the variation in the computed adjustment factors, resulting from the inverse of the estimated probabilities. For reasons of efficiency, this should be avoided. Thus, we first consider each of the variables in a bivariate model. If the variable does turn out to have a significant ($p < 0.05$) influence on the participation decision modeled, it enters the set of significant variables. This set is then analyzed for correlation among each other. If variables show an absolute correlation greater than 0.95, we choose the variable with the greater estimate from the bivariate model. The remaining set of variables enters the preliminary model. In order to reduce the number of explanatory variables to a minimum, we use a variable selection approach based on the Bayesian information criterion (BIC). This variable selection approach skips and adds variables in a stepwise algorithm, only skipping or keeping them if the model fit improves in terms of the BIC. This three-step procedure yields a final model used in the estimation of participation probabilities used to adjust the weights. The models estimating the propensities for contact and participation used to derive weighting adjustments are presented in Table 4 and Table 5 for M7 as well as in Table 6 and Table 7 for M8.

Table 4 shows the coefficients for the model estimating the contact propensities for the sample M7. Persons having their first or last spell in the IEBs being related to a job-seeking activity or an employment are more likely to be successfully contacted. The older the last spell in the IEB data is, the less likely the person is to be successfully contacted. The timing of the last contact is also crucial. Here, persons were less successfully contacted during the $3rd$ quarter of 2020 (beginning of the field period) than throughout the 4^{th} quarter (mid field period). Persons in Hamburg and Schleswig-Holstein were less likely to be successfully contacted compared to persons from North Rhine-Westphalia. Further variables related to the building and the neighborhood (PLZ8-level, street-level) the person lived in also affect the successful contact.

Table 4: Model estimating contact propensities used to derive weighting adjustments for subsample M7.

	Contacted
(Intercept)	-0.569^{***}
	(0.068)
Last employment spell	0.416^{***}
Job-seeking history (XSozial)	(0.088)
Last employment spell	0.173^{***}
Job-seeking history	(0.041)
First employment spell	0.417^{***}
Job-seeking history	(0.073)
First employment spell	0.338^{***}
Employment history	(0.063)
Time of last spell	-0.778^{***}
2016 - quarter 1	(0.229)
Time of last spell	-0.658^{***}
2016 - quarter 2	(0.127)
Time of last spell	-0.807^{***}
2016 - quarter 3	(0.121)
Time of last spell	-0.800^{***}
2016 - quarter 4	(0.109)
Time of last spell	-0.722^{***}
2017 - quarter 1	(0.114)
Time of last spell	-0.540^{***}
2017 - quarter 2	(0.090)
Time of last spell	-0.568^{***}
2017 - quarter 3	(0.085)
Time of last spell	-0.488^{***}
2017 - quarter 4	(0.072)
Time of last spell	-0.410^{***}
2018 - quarter 1	(0.078)
Time of last spell	-0.487^{***}
2018 - quarter 2	(0.068)
Time of last spell	-0.423^{***}
2018 - quarter 3	(0.055)
Time of last contact attempt	-0.361^{***}
2020 - quarter 3	(0.035)
Time of last contact attempt	0.167^{***}
2020 - quarter 4	(0.030)
Federal state	-0.445^{***}
Hamburg	(0.080)
Federal state	0.289^{***}
North Rhine-Westphalia	(0.030)
Federal state	-0.298^{***}
Schleswig Holstein	(0.088)
Condition of building	0.096^{***}
No Peculiarities, Good Standard	(0.026)
Type of neighborhood	-0.268^{***}
Residential, commercial and industrial	(0.068)
Type of building	-0.260^{***}
Farm house	(0.066)
Type of building	0.144^{***}
Apartment in 5-8 unit building	(0.029)
Type of building	-0.275^{***}
Hostel for working persons	(0.069)
Type of building	-0.203^{**}
Other accommodation	(0.063)
Type of PLZ8-area	-0.145^{***}
Rural area	(0.037)
Fluctuation (PLZ8-level)	-0.139^{***}
slightly below average	(0.041)
Dominant migrant's milieu (PLZ8-level)	-0.208^{***}
Intellectual-cosmopolitan	(0.056)
Purchasing power parity (PLZ8-level)	0.231^{***}
far above average	(0.059)
Sinus-Geo-Milieu (street-level)	0.301^{***}
Traditionally ingrained	(0.078)
Dominant Microm group (street-level)	0.194^{***}
Pensioners in post-war buildings	(0.049)
Dominant Microm group (street-level)	0.140^{***}
Apartment towers and rental apartments	(0.031)
N	13,664

Notes: Dependent variable: household successfully contacted (1 = yes, 0 = no). Significance indicated by $*** \equiv p < 0.001$, $** \equiv p < 0.01$, and $* \equiv p < 0.05$. The model is estimated using the function `glm()` with a cloglog link function in R (R Core Team, 2020).

Table 5 displays the coefficients of the model estimating the participation propensity for sample M7. It shows that persons with a Polish nationality were less likely to participate in the survey. In contrast persons having a higher education entrance qualification tend to be more likely to participate. The closer the last spell in the IEB data is to the survey period the more likely persons were to participate. Persons located in Rhineland Palatinate were less likely to participate compared to persons located in North Rhine-Westphalia. If there were no language barriers the participation propensity was higher, too. Again, variables related to the neighborhood are related to participation, as are characteristics of the interviewer concerning their full-time occupation and their highest educational degree.

Table 5: Model estimating participation propensities used to derive weighting adjustments for subsample M7.

	Participated
(Intercept)	-5.166^{***}
	(0.239)
Nationality	-0.361^{***}
Polish	(0.079)
Highest educational degree	0.540^{***}
higher education entrance qualification	(0.123)
Time of last spell	0.684^{***}
2018 - quarter 4	(0.113)
Time of last spell	0.907^{***}
2019 - quarter 1	(0.168)
Time of last spell	0.696^{***}
2019 - quarter 2	(0.147)
Federal state	0.319^{***}
North Rhine-Westphalia	(0.086)
Federal state	-0.658^{***}
Rhineland Palatinate	(0.199)
Language barriers	0.862^{***}
none	(0.076)
Safety of the residential area	1.687^{***}
Safe	(0.176)
Safety of the residential area	1.846^{***}
Very safe	(0.179)
Type of neighborhood	0.335^{***}
Residential area, mostly old buildings	(0.076)
Type of PLZ8-area	-0.486^{**}
Small town fringe area	(0.171)
Type of PLZ8-area	0.435^{***}
rural area	(0.096)
Occupation of interviewer	0.571^{***}
full-time	(0.142)
Highest educational degree of interviewer	-0.455^{***}
University without a degree	(0.126)
Highest educational degree of interviewer	-0.345^{***}
Secondary school	(0.100)
N	7104

Notes: Dependent variable: Participation of the household (1 = yes, 0 = no). Significance indicated by $***\equiv p < 0.001$, $**\equiv p < 0.01$, and $*\equiv p < 0.05$. The model is estimated using the function `glm()` with a cloglog link function in R (R Core Team, 2020).

Table 6 shows the coefficients for the model estimating contact propensities used to derive weighting adjustments for subsample M8. Persons whose most frequent spell is in job seeking reduces the likelihood of successful contact. The timing of a person's first or last spell also influences the successful contact negatively, except for having a very recent last spell. Contacting a person was more successful in the fourth quarter than in the third, compared to the earlier contact attempts. Also people from Bangladesh and the Philippines were harder to contact compared to other nationalities. Compared to other states, foreigners were easier to contact in Bremen and harder to contact in Berlin. Looking at the regional information on PLZ8- and street-level provided by Microm, we see different economic indicators affecting the probability to be successfully contacted in different ways. Additionally, the type of building and neighborhood indicating persons

not living residential areas and buildings are harder to contact within the population of M8.

Table 6: Model estimating contact propensities used to derive weighting adjustments for subsample M8.

	Contacted
(Intercept)	-0.284^{***}
	(0.050)
Most frequent employment spell	-0.299^{**}
Job-seeking history	(0.093)
Time of first spell	-1.075^{**}
2012 - quarter 1	(0.407)
Time of first spell	-0.397^{**}
2015 - quarter 4	(0.126)
Time of last spell	-0.741^{***}
2019 - quarter 2	(0.180)
Time of last spell	0.297^{***}
2019 - quarter 4	(0.045)
Time of last contact attempt	-0.126^{***}
2020 - quarter 3	(0.032)
Time of last contact attempt	0.224^{***}
2020 - quarter 4	(0.032)
Nationality	-0.559^{**}
Bangladesh	(0.186)
Nationality	-0.433^{***}
Philippines	(0.109)
Federal state	0.300^{***}
Bremen	(0.084)
Federal state	-0.264^{***}
Berlin	(0.046)
Federal state	0.219^{***}
North Rhine-Westphalia	(0.033)
Life phase by socio-economic status (PLZ8-level)	0.321^{***}
Financially week families	(0.097)
Life phase by socio-economic status (PLZ8-level)	0.281^{***}
Financially well-off families	(0.068)
Life phase by socio-economic status (PLZ8-level)	0.236^{***}
Financially well of single elderly	(0.068)
Type of PLZ8 area	-0.326^{**}
Holiday area	(0.110)
Dominant car brand (PLZ8-level)	0.368^{***}
Peugeot	(0.108)
Sinus-Geo-Milieu (PLZ8-level)	-0.186^{***}
Adaptive-pragmatical oriented	(0.040)
Life phase by socio-economic status (street-level)	-0.620^{**}
Financially well-off older multi-person household	(0.208)
Purchasing power parity (street-level)	-0.158^{***}
far above average	(0.043)
Type of neighborhood	-0.365^{***}
Mainly commercial and industrial area	(0.084)
Type of building	-0.334^{***}
Other accommodation	(0.062)
Type of building	-0.265^{***}
Hostel for working persons	(0.074)
N	11,324

Notes: Dependent variable: household successfully contacted (1 = yes, 0 = no). Significance indicated by $*** \equiv p < 0.001$, $** \equiv p < 0.01$, and $* \equiv p < 0.05$. The model is estimated using the function `glm()` with a cloglog link function in R (R Core Team, 2020).

Table 7 details the coefficients for the model estimating participation propensities used to derive weighting adjustments for subsample M8. Of the persons successfully contacted, those who's last spell stems from the employment history are less likely to participate. Different timings of the first and last spell also have a negative effect on the participation propensity. Having a higher education entrance qualification influences the participation decision in a positive way. Among the nationalities in the sample, persons from Gambia, Brazil, India, and Iraq are more likely and persons from China are less likely to participate. Additionally, people living in Thuringia have a higher participation propensity. On the regional level, the information provided by Microm show that households in areas with high numbers of young children are more likely to participate. On the PLZ8-level, the dominant migrant's milieu and, on the street-level, the dominant Microm group influence participation propensities negatively. Households where no language barriers were present show a higher willingness to participate. The same is true for safe and very safe residen-

tial areas. Moreover different types of neighborhoods increase a household's willingness to participate. Finally, full-time occupied interviewers were more successful in getting households to participate.

Table 7: Model estimating participation propensities used to derive weighting adjustments for subsample M8.

	Participated
(Intercept)	-7.704^{***}
	(0.534)
Last employment spell	-0.430^{***}
Employment history	(0.110)
Highest educational degree	0.272^{***}
higher education entrance qualification	(0.067)
Time of first spell	-1.228^{**}
2015 - quarter 3	(0.413)
Time of last spell	-1.100^{**}
2019 - quarter 3	(0.414)
Nationality	1.663^{***}
Gambia	(0.478)
Nationality	0.611^{***}
Brazil	(0.185)
Nationality	0.551^{***}
India	(0.106)
Nationality	1.068^{***}
Iraq	(0.297)
Nationality	-0.850^{**}
China	(0.302)
Federal state	0.884^{***}
Thuringia	(0.239)
Number of inhabitants	1.016^{***}
aged 3 up to 6 years	(0.163)
Dominant migrant's milieu (PLZ8-level)	-0.463^{***}
Multicultural performer's milieu	(0.103)
Dominant Mmicrom group (street-level)	-0.595^{**}
Elderly Persons in surrounding municipalities	(0.194)
Language barriers	1.343^{***}
none	(0.063)
Safety of the residential area	1.107^{***}
Very safe	(0.183)
Safety of the residential area	0.876^{***}
Safe	(0.180)
Type of neighborhood	2.137^{***}
Mainly commercial area	(0.307)
Type of neighborhood	1.702^{***}
Residential area, mostly old buildings	(0.259)
Type of neighborhood	1.694^{***}
Mixed commercial and residential	(0.264)
Type of neighborhood	1.518^{***}
Residential area, mostly new buildings	(0.266)
Type of building	0.522^{**}
Hostel for working persons	(0.162)
Occupation of interviewer	0.573^{***}
full-time	(0.149)
N	7081

Notes: Dependent variable: Participation of the household (1 = yes, 0 = no). Significance indicated by $*** \equiv p < 0.001$, $** \equiv p < 0.01$, and $* \equiv p < 0.05$. The model is estimated using the function `glm()` with a cloglog link function in R (R Core Team, 2020).

5.2 Population Weighting Adjustment

In the last step of the weighting process, we use raking to adjust the weights from the previous step to meet different joint and marginal distributions. The weights resulting from this step are the basis for cross-sectional and longitudinal weights derived for wave 2 and beyond. The population parameters and distributions used in the population weighting adjustments were provided by the Federal Statistical Office based on the German Micro-census. At the household-level the following distributions were used:

- Number of households by federal state
- Number of households by municipality size

- Number of households by household size
- Number of households by household type

At the individual level the following marginal and joint distributions were used:

- Number of Persons by immigration year
- Number of Persons by nationality
- Number of Persons by age group and gender

5.3 Characteristics of Weights

Table 8: Characteristics of weights after the steps of the weighting process (rounded to integer values).

		Quantiles							
Step	Min.	10%	25%	50%	75%	90%	Max.	Mean	SD
	Subsample M7								
DW	2	8	15	26	45	65	159	32	24
SWA	12	64	125	247	531	1,070	11,487	507	920
PWA	2	22	45	102	239	468	1,589	197	259
	Subsample M8								
DW	2	3	4	6	8	9	12	6	3
SWA	4	12	17	31	64	120	1,915	59	106
PWA	4	12	19	35	71	155	539	66	86

Abbreviations: SD = standard deviation, DW = design weighting, SWA = sample weighting adjustment, PWA = population weighting adjustment.

Due to stratification and disproportional allocation of households, there is some variance in the design weights. Multiplying design weights with the inverse of estimated participation probabilities increases variation in the second weighting step. The population weighting adjustments reduce the magnitudes as well as the variation of weights for the participating households.

6 Summary

M7 and M8 secure and expand the previous analysis potential of the SOEP's immigrant samples M1 and M2. The focus of the two samples is distinct. M7 refreshes the former immigration samples by households and individuals mostly with foreigners from within the European Union with an emphasis on eastern Europe. In contrast, the M8 sample augments the immigration samples by third-country nationals joining the German labor force.

References

American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). AAPOR.

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical methods in medical research*, *5*(3), 215–238. doi: 10.1177/096228029600500302

INKAR. (2019). *Indikatorenübersicht – Indkatoren Raum- und Zeitbezüge.* Retrieved from https://www.inkar.de/documents/Indikatoren%20Raum-%20und%20Zeitbezuege.pdf

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey methodology*, *12*(1), 1–16.

Kroh, M., Kühne, S., Goebel, J., & Preu, F. (2015). *The 2013 IAB-SOEP Migration Sample (M1): Sampling Design and Weighting Adjustment* (SOEP Survey Papers No. 271). Berlin: DIW/SOEP.

Kroh, M., Siegers, R., & Kühne, S. (2015). Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP). In *Nonresponse bias* (pp. 409–444). Springer.

Kühne, S., & Kroh, M. (2017). *The 2015 IAB-SOEP Migration Study M2: Sampling Design, Nonresponse, and Weighting Adjustment* (SOEP Survey Papers No. 473). Berlin: DIW/SOEP.

Liebig, S., Brücker, H., Goebel, J., Grabka, M. M., Schröder, C., Zinn, S., … Deutsches Institut Für Wirtschaftsforschung (DIW Berlin) (2021). *IAB-SOEP Migrationsstichprobe 2019.* SOEP Socio-Economic Panel Study. doi: 10.5684/SOEP.IAB-SOEP-MIG.2019

Oberschachtsiek, D., Scioch, P., Seysen, C., & Heining, J. (2009). *Stichprobe der Integrierten Erwerbsbiografien IEBS* (FDZ-Datenreport No. 03/2009). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung. Retrieved from http://doku.iab.de/fdz/reporte/2009/DR_03-09.pdf

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rathje, M., & Glemser, A. (2021). *SOEP-Core – 2020: Report of Survey Methodology and Fieldwork* (SOEP Survey Papers No. 1050). Berlin: DIW/SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.824248.de/diw_ssp1050.pdf