1536 2025

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

IAB-SOEP-Migration – 2023: Sampling, Nonresponse, and Weighting in Sample M8c

Hans Walter Steinhauer, Parvati Trübswetter, Tanja Fendel, Boris Ivanov, Adriana Cardozo Silva, Felix Süttmann, Rainer Siegers, 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, Tanja Fendel, Boris Ivanov, Adriana Cardozo Silva, Felix Süttmann, Rainer Siegers, Sabine Zinn. 2025. IAB-SOEP-Migration – 2023: Sampling, Nonresponse, and Weighting of the Sample M8c. SOEP Survey Papers 1536. Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

IAB-SOEP-Migration – 2023: Sampling, Nonresponse, and Weighting of the Sample M8c

Hans Walter Steinhauer[1], Parvati Trübswetter[2], Tanja Fendel[2], Boris Ivanov[2], Adriana Cardozo Silva[1,3], Felix Süttmann[1], Rainer Siegers[1], and Sabine Zinn[1,4]

[1]German Institute for Economic Research (DIW Berlin)
[2]Institute for Employment Research (IAB)
[3]University of Applied Sciences and Arts Hannover
[4]Humboldt-Universität zu Berlin

July 28, 2025

Abstract
This paper provides details on the sampling design, the fieldwork, as well as nonresponse and population adjustments for the 2023 sample M8c. The survey is conducted in cooperation between the Institute for Employment Research (IAB) and the Socio-Economic Panel (SOEP). Sample M8c contributes to the IAB-SOEP-Migration Samples. It augments the IAB-SOEP-Migration samples M8a and M8b by adding 970 households of foreigners from foreign countries to allow for evaluating the Skilled Labor Immigration Act (Fachkräfteeinwanderungsgesetz). The act came into effect on March 1, 2020. Its goal is to facilitate the immigration of skilled workers from non-EU countries to Germany.

Acknowledgments
This report is related to and relies on Steinhauer, Trübswetter, and Zinn (2022) and Steinhauer et al. (2025).

1 Introduction

Panel studies are widely recognized as valuable tools in social and behavioral research, offering insights into dynamic patterns and long-term trends within a given population. These studies rely heavily on the participation of individuals who are willing to contribute their time and experiences over an extended period of time. However, maintaining panel participation and ensuring that the sample remains "representative" of its population is challenging. A crucial aspect of panel studies is the refreshment and augmentation of the sample. Adding new participants into an ongoing panel compensates for attrition while also capturing changes in the population over time. This process is essential for maintaining the "representativeness" of the panel and for preserving the integrity of the findings.

The Socio-Economic Panel (SOEP) is one of the longest-running panel studies in the field of economic and social sciences. The SOEP provides data on various socio-economic aspects of individuals and households over an extended period since 1984. By including participants from different subgroups of the German population, the panel ensures diversity. Specifically, the IAB-SOEP-Migration samples M1, M2, M7, M8a, and M8b focus on migration to Germany motivated by job opportunities. More details on the IAB-SOEP-Migration samples are provided by Brücker et al. (2014). The SOEP samples M8a and M8b cover a specific subgroup of the population living in Germany, namely skilled labor immigration from non-EU countries to Germany; for details see Steinhauer et al. (2022) and Steinhauer et al. (2025). These samples provide a basis to evaluate the Skilled Labor Immigration Act (Fachkräfteeinwanderungsgesetz), which initially came into effect on March 1, 2020, with the goal of facilitating the immigration of skilled workers from non-EU countries to Germany. To further support this evaluation and compensate for the loss of households in the sample due to panel attrition, we introduce SOEP sample M8c. Moreover, sample M8c augments the previous migration samples M1, M2, M7, M8a, and M8b.

This paper details the approach applied to augment the SOEP-Core covering the population migrating to enter the German labor market. In Section 2, we provide information on the target population, the sampling frame, and details on the sampling design. The fieldwork process and its results are described in Section 3. Section 4 details the different steps of weighting. Characteristics of the final weights are displayed in Section 5, while Section 6 summarizes.

2 Sampling Design

The target population for sample M8c comprises households with persons from both EU and non-EU countries that show an employment spell for the first time in the IEB between November 1, 2021, and June 30, 2022. The Federal Agency for Employment listed 134,910 target persons in the Integrated Employment Biographies (IEB), see Table 1. From this population we draw a sample of 20,000 individuals with the design also applied in M8a, see Steinhauer et al. (2022) for details.

Table 1: Number of target persons in the population by federal state.

	Population		Sample	
Federal State	Number	Proportion	Number	Proportion
Brandenburg	2,024	0.015	218	0.011
Berlin	15,488	0.115	4,630	0.232
Baden-Württemberg	21,094	0.156	2,379	0.119
Bavaria	27,618	0.205	3,403	0.170
Bremen	1,459	0.011	656	0.033
Hesse	11,768	0.087	1,940	0.097
Hamburg	3,809	0.028	687	0.034
Mecklenburg Western Pomerania	1,340	0.010	179	0.009
Lower Saxony	8,459	0.063	792	0.040
North Rhine-Westphalia	23,935	0.177	3,597	0.180
Rhineland-Palatinate	5,658	0.042	527	0.026
Schleswig-Holstein	2,625	0.019	320	0.016
Saarland	852	0.006	67	0.003
Saxony	4,110	0.030	191	0.010
Saxony-Anhalt	2,417	0.018	134	0.007
Thuringia	2,254	0.017	280	0.014
Total	134,910	1.000	20,000	1.000

3 Fieldwork results and Response Rates

The addresses sampled from official records at the Federal Agency for Employment were validated by infas (Institute of Applied Social Sciences) and deployed to the field in four independent tranches. The last tranche was not deployed to the field, because the desired number of households was nearly reached using the first three tranches. That is why the number of households reported here is reduced to 16,500. The interviews for the sample M8c were conducted between July 2023 and December 2023. Well in advance, the selected sample of households received letters via mail emphasizing that participation was voluntary. Table 2 details the results of the fieldwork on the household-level. In total, there were 970 complete or partial interviews, resulting in a response rate on the household-level, calculated according to American Association for Public Opinion Research (2023), of $RR2 = \frac{970}{9{,}804} = 0.099$. The number of households explicitly refusing to participate is rather low. The resulting refusal rate is $REF1 = \frac{690}{9{,}804} = 0.070$ and, thus, is similar to other samples and studies. Some addresses, although sampled and valid, were not deployed in the field because either interviewers did not contact them by the end of the field period or the numbers of desired interviews was already reached (see AAPOR code 3.11 in Table 2). Other addresses are out of sample, because the household has moved abroad or was screened out; see AAPOR codes 4.1 to 4.5 in Table 2.

Table 2: Fieldwork results on the household-level according to American Association for Public Opinion Research (2023).

Final Disposition	M8c			
	CAPI		CAWI	
Code	Number	Proportion	Number	Proportion
1. Interview				
(1.10) Complete	63	0.018	489	0.038
(1.20) Partial	56	0.016	362	0.028
2. Eligible, Non-Interview				
(2.10) Refusal	195	0.056	13	0.001
(2.12) Break-off	42	0.012	440	0.034
(2.20) Non-contact	440	0.126	1	0.000
(2.31) Dead	2	0.001	2	0.000
(2.32) Physically / mentally unable	7	0.002	0	0.000
(2.33) Language problems	76	0.022	1	0.000
(2.36) Miscellaneous	797	0.229	74	0.006
3. Unknown eligibility, non-interview				
(3.11) Not attempted or worked	347	0.021		
(3.20) Nothing ever returned			6,397	0.388
4. Not Eligible				
(4.10) Out of sample / screened out	244	0.070	1,254	0.096
(4.20) Moved abroad	35	0.010	4	0.000
(4.40) Untraceable	996	0.286	3,979	0.306
(4.50) Non-residential building	184	0.053	0	0.000
Total	3,484	1.000	13,016	1.000

Note: Proportions might not sum up to one because of errors due to rounding.

4 Cross-Sectional Weighting

According to Brick and Kalton (1996), the computation of weights typically involves three main steps. First, the design weights are calculated as the inverse of the inclusion probability (see Section 2). These are then adjusted to account for unit nonresponse, a process referred to as sample weighting adjustment by Kalton and Kasprzyk (1986). Finally, in the third step, the weights are calibrated to ensure that the estimates align with known population parameters, such as totals, ratios, or specific distributions. Kalton and Kasprzyk (1986) refer to this step as population weighting adjustment. For a comprehensive understanding of the general weighting strategy employed by the SOEP and the incorporation of new samples, please refer to Kroh, Siegers, and Kühne (2015).

Please note that the IEBs lists individuals without providing information about their household context. However, the IAB-SOEP-Migration Survey is a household panel survey in which all adults are interviewed. Consequently, a household with two individuals, for example, has a higher probability of selection compared to a single-person household. To determine a household's sampling probability, we assign sampling probabilities to all members of existing households, even though these individuals were not initially sampled

as anchor persons. This involves accounting for various characteristics used in stratifying and clustering the sample. Once the sampling probabilities for each household member are identified, we calculate the joint probability the household to be part of the sample. For formal documentation of this procedure; see Kroh, Kühne, Jacobsen, Siegert, and Siegers (2017).

To adjust for unit nonresponse, we make use of the Integrated Employment Biographies (IEB), which is administrative data provided by the IAB. 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 includes socio-demographic information on gender, age, and nationality as well as geographic information, including, for example, regional classifications.

Please note that the IEB lists individuals without providing information about their household context. However, the SOEP is a household panel survey in which all adults are interviewed. Consequently, a household with two individuals, for example, has twice the probability of selection compared to a single-person household. To determine a household's sampling probability, we assign sampling probabilities to all members of the existing households, even though these individuals were not initially sampled as anchor persons. This process requires accounting for characteristics used in clustering the sample. Once the sampling probabilities for each household member are identified, we calculate the household sampling probabilities by summing these probabilities within each household. The household weights are then derived as the inverse of these household sampling probabilities. It is important to note that the number of households in the target population is unknown and cannot be determined from the sampling frame, as the register does not include household identifiers.

To address potential selectivity resulting from fieldwork and nonresponse, we employ two models that capture the success of contacting a household as well as the decision-making process of households regarding participation. The models incorporate information on

a) all households that have been deployed to the field and have been approached/ contacted; and

b) both participating and nonparticipating households.

The first model estimates the probability of successfully contacting a household. Given successful contact, the second model estimates the probability of the household participating in the panel. Given the limited availability of data on households in the initial sample, we use area-level information regarding the residential environment provided by infas360 (see https://datenkatalog.infas360.de/). Additionally, we use design information and data on fieldwork processes. The latter include information on the interviewer alongside attributes of the first contact attempt. This information only applies to the CAPI field, because the CAWI does not require interviewers.

4.1 Sample Weighting Adjustments

In the second step of correcting the design weights, it is essential to identify strong predictors of nonresponse. To accomplish this, we conduct an iterative process that involves

examining information listed above. We select variables that demonstrate significant influence on the participation decision through bivariate regression analysis. Subsequently, we remove variables from the set of significant variables if their absolute correlation value with each other is greater than or equal to 0.95. This step ensures that highly correlated variables are not duplicated in the analysis. The remaining variables then form the basis for a preparatory nonresponse model. To obtain the final model, we employ a variable selection procedure in both forward and backward directions, using the Bayesian Information Criterion (BIC) as the selection criterion. This approach allows us to arrive at a more parsimonious model, retaining only the most relevant variables. The resulting models estimating the probability to be successfully contacted as well as the response propensities used for deriving weighting adjustments, are presented in Table 3.

Various factors influence the success of contacting households for interviews. These factors include the timing of the first contact, the interviewer characteristics, specific attributes of the residence and neighborhood, as well as socio-economic characteristics. Besides that, we separate the two main survey modes: CAWI and CAPI for modelling contact and participation.

Table 3: Models estimating the probability to be successfully contacted as well as the response propensities used for deriving weighting adjustments.

	Contact		Participation	
	CAPI	CAWI	CAPI	CAWI
Variable	Estimate			
Value	(Std. Error)			
(Intercept)	-0.807^{***}	-2.377^{***}	-2.609^{***}	-0.047
	(0.065)	(0.062)	(0.129)	(0.048)
1. Attributes of first contact attempt				
Timing	0.355^{***}	0.174^{**}		
weekend	(0.071)	(0.058)		
2. Attributes of the anchor person				
Age group		0.465^{***}		
30 - 39		(0.056)		
Nationality				-0.674^{***}
Albania				(0.199)
Nationality	0.467^{***}			
Bosnia and Herzegovina	(0.113)			
Nationality		-0.614^{**}		
North Macedonia		(0.216)		
Nationality		-0.485^{**}		
Kosovo		(0.158)		
Nationality		0.690^{***}		
Brazil		(0.115)		
Nationality		0.563^{**}		
Mexico		(0.186)		
Sex	0.232^{***}	0.230^{***}		

Table 3 continued.

	Contact		Participation	
	CAPI	CAWI	CAPI	CAWI
Variable	Estimate			
Value	(Std. Error)			
female	(0.060)	(0.056)		
Highest school education		0.496***	0.611**	
high school diploma		(0.057)	(0.187)	
Number of years in IEB		−0.810**		
2 or more		(0.279)		
3. Attributes of the residence				
Distance to bus stop	−0.210**			
nearby	(0.070)			
Location within municipality		0.565***		
boarder of town		(0.148)		
4. Attributes of the residential area (block of buildings)				
People aged 10 to 14		−0.392***		0.291**
low share		(0.078)		(0.093)
People aged 15 to 17		−0.229**		
high share		(0.071)		
People aged 75 and older	−0.248**			
low share	(0.078)			
Foreigners	0.249***			
low share	(0.069)			
People with a PhD				0.242**
share higher than 2.5%				(0.077)
Dominant building type		−0.677***		
commercial buildings		(0.125)		
5. Attributes of the residential area (neighborhood)				
Socio-economic deprivation			0.735**	
7^{th} decile			(0.226)	
Socio-economic deprivation	0.355***			
8^{th} decile	(0.095)			
Socio-economic deprivation	0.687***			
9^{th} decile	(0.133)			
People commuting in		−0.287***		
low share		(0.073)		
People with a Yugoslavian mig. back.	−0.242**			
high share	(0.074)			
Price index for property ($€/m^2$)	0.258***			
low	(0.063)			
Land use - forest	0.273***			
high share	(0.067)			

Table 3 continued.

Variable Value	Contact CAPI	 CAWI	Participation CAPI	 CAWI
	Estimate (Std. Error)			
N	2,672	11,756	1,228	1,312

Notes: Dependent variable: Success in contacting the household (1 = yes, 0 = no), 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, 2025).

When it comes to contacting the households in the first place, doing so over weekends results in higher success rates in both survey modes, as compared to other days of the week. Looking at attributes of the anchor person, we find a higher likelihood to successfully contact the age group between 30 and 39 years old in the CAWI survey mode. Further, some nationalities are more likely to be successfully contacted. In CAPI mode, this is the case for people from Bosnia and Herzegovina, whereas in CAWI mode it is for people from Brazil and Mexico. In contrast, people from North Macedonia and Kosovo are less likely to be successfully contacted in CAWI mode. For both modes, female anchor persons have higher rates of being contacted successfully. For anchor persons in CAWI mode, a high school diploma affects the chances of a successful contact positively while having a history of two or more years in the IEB data lowers the chance of a successful contact. Households located close to bus stops have a lower propensity for being successfully contacted in CAPI mode, whereas being located at the boarding region of a town increases the propensity for a successful contact. The demographic composition of the neighborhood has different effects for each of the two survey modes. For the CAPI survey mode, households have lower number of successful contacts in residential areas with a low share of older people and higher contact rates in areas with a low share of foreigners. In contrast, households located in areas with low shares of younger people, have a lower likelihood of being successfully contacted in CAWI mode. Also in CAWI mode, for households located in areas where commercial buildings dominate, households are less likely to be successfully contacted. Finally, some attributes of the neighborhood, such as a higher level of Socio-economic deprivation, a low price index for buying property, and higher share of forestry land use have a positive effect on being contacted successfully in CAPI mode. A higher share of people with a Yugoslavian migration background has a negative effect on successful contacts in CAPI mode. In CAWI mode, households located in neighborhoods with a low share of people commuting in, have a lower rate of being successfully contacted.

For households that have been successfully contacted, there are only a few attributes that affect their participation propensity. Looking at the attributes of the anchor persons, households whose anchor has an Albanian nationality have a lower propensity to participate in the CAWI mode. Households whose anchor person has a high school diploma have a higher likelihood to participate in the CAPI survey mode. In CAWI mode attributes of the residential area have a positive effect on the propensity to participate. A low share of children aged 10 to 14 as well as a share of people with a PhD beyond 2.5% influence the decision to participate positively. In CAPI mode, households located in neighborhoods with a higher index of socio-economic deprivation, have a higher likelihood of participating in the survey.

4.2 Population Weighting Adjustments

In the final step of the weighting process, we employ post-stratification and raking techniques to adjust the weights obtained in the previous step. This adjustment is necessary to align the weights with known population totals, as well as joint and marginal distributions. The specific method chosen for this adjustment depends on the available data for the population. A comprehensive overview of these methods is provided by Kalton and Flores-Cervantes (2003). The resulting weights from this step serve as the foundation for deriving cross-sectional and longitudinal weights for subsequent waves, starting from Wave 2 onwards.

The population parameters and distributions utilized in the population weighting adjustments are provided by the Federal Statistical Office, drawing upon data from the German Microcensus 2023. Margins used in the post-stratification process are:

Number of households with at least one person of the M8c-population who immigrated to Germany in 2022 by

- household typology (single vs. other);
- municipality size; and
- Regions (federal states categorized by north, east, south, and west)

Number of persons of the M8c-population who immigrated to Germany in 2022 by

- sex, and
- age group.

5 Characteristics of Weights

Due to the sampling design, 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; compare Table 4. The population weighting adjustments then again lower the variation and magnitude of weights. Resulting weights are provided in the variable `hhrf0` included in the data set `hpathl` as well as in the variable `bnhhrfm8c` in the data set `hhrf`.

Table 4: 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
DW	1.03	1.48	1.96	3.34	7.23	13.91	144.94	6.04	8.46
SWA	4.30	12.79	21.29	45.28	109.81	219.44	4,244.21	97.09	195.21
PWA	4.59	14.56	23.32	43.95	86.19	167.50	688.13	76.36	98.41

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

After the integration step, a further post-stratification step was carried out in which the weights (previously nonresponse-adjusted, if necessary post-stratified and integrated) of all SOEP samples were adjusted with respect to the standard marginal distributions used by SOEP, as taken from the Microcensus 2023. Using the resulting standard SOEP weighting factors (`hhrf` included in `hpathl` and `phrf` included in `ppathl`), the sample M8c cases can then be analyzed jointly and comparatively in combination with all other SOEP cases.

6 Summary

The new Sample M8c is an augmentation sample adding 970 households to the SOEP. Like to previous migration samples of the IAB-SOEP-migration survey it is drawn from the IEBs. Concerning field work, we find a few possible selectivities in the CAPI as well as in the CAWI field. However, they differ in terms of their characteristics. In the CAPI field, we mainly find that attributes of the residential area affect contactability. In contrast, in the CAWI field, we find that successful contacts are primarily driven by the attributes of the anchor person. Using the information available from regional attributes of the household as well as attributes of the anchor person from the IEB, we account for possible selectivities due to contact and participation.

Sample M8c augments the samples of the IAB-SOEP-Migration survey by EU and non-EU nationals joining the German labor force. The sample secures and expands the previous analysis potential of samples M1, M2, M7, M8a, and M8b. Thereby, it can be used to evaluate the Skilled Labor Immigration Act (Fachkräfteeinwanderungsgesetz) together with samples M8a and M8b. The households of this new sample have been seamlessly integrated.

References

American Association for Public Opinion Research. (2023). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (10th ed.). Retrieved from https://aapor.org/wp-content/uploads/2024/03/Standards-Definitions-10th-edition.pdf

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

Brücker, H., Kroh, M., Bartsch, S., Goebel, J., Kühne, S., Liebau, E., … Schupp, J. (2014). *The new IAB-SOEP migration sample. An introduction into the methodology and the contents.* (SOEP Survey Papers No. 216). Berlin: DIW-Berlin.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of official statistics*, *19*(2), 81–97.

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

Kroh, M., Kühne, S., Jacobsen, J., Siegert, M., & Siegers, R. (2017). *Sampling, nonresponse, and integrated weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4)—revised version* (SOEP Survey Papers No. 477). Berlin: DIW/SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.572346.de/diw_ssp0477.pdf

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.

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. (2025). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Steinhauer, H. W., Trübswetter, P., Fendel, T., Ivanov, B., Cardozo Silva, A., Süttmann, F., … Zinn, S. (2025). *IAB-SOEP-Migration – 2022: Sampling, Nonresponse, and Weighting in Sample M8b* (SOEP Survey Papers No. 1497). Berlin: DIW-Berlin.

Steinhauer, H. W., Trübswetter, P., & Zinn, S. (2022). *SOEP-Core - 2020: Sampling, Nonresponse, and Weighting in the IAB-SOEP Migration Studies M7 and M8* (SOEP Survey Papers No. 1105). Berlin: DIW-Berlin.