1497 2025

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

IAB-SOEP-Migration – 2022: Sampling, Nonresponse, and Weighting in Sample M8b

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**

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:

Hans Walter Steinhauer, Parvati Trübswetter, Tanja Fendel, Boris Ivanov, Adriana Cardozo Silva, Felix Süttmann, Rainer Siegers, Sabine Zinn. 2025. IAB-SOEP-Migration – 2022: Sampling, Nonresponse, and Weighting in Sample M8b. SOEP Survey Papers 1497. Series C. Berlin: DIW/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)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

IAB-SOEP-Migration – 2022: Sampling, Nonresponse, and Weighting in Sample M8b

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

[1]German Institute for Economic Research (DIW Berlin)
[2]Institute for Employment Research (IAB)
[3]Humboldt-Universität zu Berlin

April 14, 2025

Abstract
This paper provides details on the sampling design, the fieldwork, as well as nonresponse and population adjustments for the 2022 sample M8b. The survey is conducted in cooperation between the Institute for Employment Research (IAB) and the Socio-Economic Panel (SOEP). Sample M8b contributes to the IAB-SOEP-Migration Samples. It refreshes the IAB-SOEP-Migration sample M8a by adding 2,333 households of foreigners from countries outside the European Union (EU) 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.

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 of the sample. Adding new participants into an ongoing panel compensates for attrition and captures changes in the population over time. This process is essential for maintaining the "representativeness" of the panel and 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. By including participants from different subgroups of the German population, the panel ensures diversity. Specifically, the IAB-SOEP-Migration samples M1, M2, M7, and M8a 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 sample M8a covers a specific subgroup of the population living in Germany, namely skilled labor immigration from non-EU countries to Germany; for details see Steinhauer, Trübswetter, and Zinn (2022). This sample provides a basis to evaluate the Skilled Labor Immigration Act (Fachkräfteeinwanderungsgesetz), which first 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 M8b. Moreover, sample M8b augments the previous migration samples M1, M2, M7, and M8a.

This paper details the approach applied to refresh the SOEP-Core covering the population of non-EU nationals 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 the sample M8b consists of skilled labor immigrants from non-EU countries for whom the Federal Agency for Employment (BA) granted permission to work in Germany between March 1, 2020, and October 30, 2021. The data of the ZuWG ("Fachverfahren Zuwanderungsgesetz") contains information on the legal work title for which the BA granted permission according to §39 German AufenthG. For some relevant titles there is no obligation to receive a BA-permission such as visas for job search or EU blue cards for jobs with an earning higher than certain thresholds. These titles are only included if permission was needed due to other reasons. Our sampling design excluded some titles not belonging to the sample's target population. Whereas many of those titles such as journalists, tour guides, or artists with short durations of stay, represent very low proportions in the data, the largest excluded group are qualified refugees with a tolerated

status. Due to data security obligations, information on work titles cannot be transferred to the survey data and the questionnaire includes separate questions on the work title at time of migration and at time of the survey. The Federal Agency for Employment listed 48,455 target persons in 3,624 municipalities with at least one person from the target population; see Table 1.

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

	Population		Sample	
Federal State	Number	Percent	Number	Percent
Schleswig-Holstein	1,130	2.3	364	1.8
Hamburg	1,563	3.2	625	3.0
Lower Saxony	3,407	7.0	1,279	6.2
Bremen	428	0.9	194	0.9
Northrhine-Westphalia	7,994	16.5	3,895	18.9
Hesse	4,145	8.6	1,770	8.6
Rhineland-Palatinate	1,841	3.8	789	3.8
Baden-Württemberg	7,890	16.3	3,845	18.6
Bavaria	9,875	20.4	3,819	18.5
Saarland	267	0.6	114	0.6
Berlin	4,985	10.3	1,994	9.7
Brandenburg	776	1.6	180	0.9
Mecklenburg Western Pomerania	700	1.4	313	1.5
Saxony	1,720	3.6	623	3.0
Saxony-Anhalt	755	1.6	372	1.8
Thuringia	979	2.0	479	2.3
Total	48,455	100.0	20,655	100.0

The actual number of skilled labor immigrants from non-EU countries to Germany between March 1, 2020, and October 30, 2021, was low, in part due to the worldwide COVID-19 confinement measures and travel restrictions. For this reason, we did not construct primary sampling units as we did for the M8a (Steinhauer et al., 2022), but selected the entire municipality. From experiences gained in the previous study, M8a, we decided to exclude municipalities with fewer than 30 persons of interest. From the remaining municipalities we sampled 40% of addresses in large municipalities with at least 500,000 inhabitants and 90% of addresses in municipalities with less than 500,000 inhabitants.

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. The interviews were conducted between May 2022 and February 2023. The households sampled received letters via mail in advance emphasizing that participation was voluntary. Table 2 details results of the fieldwork on the household-level. In total, there were 2,333 complete or partial interviews, resulting in a response rate on the household-level, calculated according to American Association for Public Opinion Research (2023), of $RR2 = \frac{2,333}{8,036} = 0.290$.

The refusal rate is $REF1 = \frac{2{,}832}{8{,}036} = 0.352$ and, thus, is similar to other surveys. 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 number 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 Code	M8b Number	M8b Percent
1. Interview		
(1.1) Complete	1,480	0.072
(1.2) Partial	853	0.041
2. Eligible, Non-Interview		
(2.1) Refusal	2,832	0.137
(2.2) Non-contact	1,924	0.093
(2.31) Dead	3	0.000
(2.32) Physically / mentally unable / incompetent	15	0.001
(2.33) Language	284	0.014
(2.36) Miscellaneous	645	0.031
3. Unknown eligibility, non-interview		
(3.11) Not attempted or worked	7,287	0.353
4. Not Eligible		
(4.1) Out of sample / screened out	417	0.020
(4.2) Moved abroad	46	0.002
(4.4) Untraceable	4,666	0.226
(4.5) Non-residential building	195	0.009
Total	20,647	1.000

Note: Totals might not add up 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). They 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).

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.

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 the 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. We find that having the first contact in the afternoon and evening is positively related, compared to a first contact earlier in the day, see column *Contact* in Table 3. First contact attempts on Mondays and Wednesdays are more often successful than on other days of the week. Moreover, contact was successful more frequently when the first contact was in August, September, or October, the first three months of the survey. Approaching the households in person for the first time negatively impacts contactability. Interviewers who were pensioners, employed, or held other occupations were more successful in contacting households compared to student interviewers. Additionally, female interviewers had lower contact success rates. In terms of residence attributes, a lower probability of living in exclusive neighborhoods increased the likelihood of successfully making contact with the individuals to be interviewed. A low share of children aged 10 to 15 living in a block of buildings around a given address lowers the probability of successful contact. The same is true for other types of settlement blocks, compared to pure residential, commercial, or mixed areas. In contrast, a very low density of constructions positively influences the success of contact. Finally, a high value (9^{th} decile) in the index of socio-economic deprivation positively impacts the probability of successful contact.

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

Variable Value	Contact Estimate (Std. Error)	Participation Estimate (Std. Error)
(Intercept)	−2.597***	−1.138***
	(0.097)	(0.145)
1. Attributes of first contact attempt		
Timing	1.317***	−0.286***
afternoon	(0.058)	(0.043)
Timing	1.692***	
evening	(0.059)	
Weekday	0.184***	
Monday	(0.039)	
Weekday	0.228***	
Wednesday	(0.038)	
Weekday		−0.199**
Thursday		(0.066)
Weekday		0.805***
Sunday		(0.067)
Month	0.865***	−0.508***

Table 3 continued.

Variable Value	Contact Estimate (Std. Error)	Participation Estimate (Std. Error)
August	(0.072)	(0.092)
Month	0.616***	
September	(0.060)	
Month	1.004***	
October	(0.058)	
Mode	−0.372***	
CAPI	(0.031)	
2. Attributes of the interviewer		
Sex	−0.191***	
female	(0.032)	
Native language		0.675***
German		(0.139)
Occupation	0.366***	−0.700***
pensioners	(0.083)	(0.066)
Occupation	0.690***	
other	(0.091)	
Occupation	0.407***	
works	(0.078)	
CASMIN		−0.756***
1		(0.133)
3. Attributes of the residence		
Prob. for demanding concept of living	0.107***	
low	(0.028)	
Quality of area		0.195***
upper midfield		(0.058)
4. Attributes of the residential area (block of buildings)		
Share of inhabitants 10-15	−0.129***	
low	(0.033)	
Share of inhabitants under 18		−0.235***
high		(0.051)
Type of settlement block	−0.371***	−0.496***
other type of settlement	(0.070)	(0.149)
Dominant building type		−0.248**
mostly commercial buildings		(0.079)
Density of constructions	0.564***	
very low	(0.130)	
5. Attributes of the residential area (neighborhood)		
Unemployment rate		0.230***

Table 3 continued.

Variable Value	Contact Estimate (Std. Error)	Participation Estimate (Std. Error)
low		(0.052)
Index of Socio-economic Deprivation		0.304***
high (7^{th} decile)		(0.061)
Index of Socio-economic Deprivation	0.272***	−0.567***
high (9^{th} decile)	(0.066)	(0.149)
Share of persons with Soviet mig. back.		−0.214***
high		(0.054)
Share of persons with Muslim mig. back.		0.222***
high		(0.051)
Share of persons with Italian mig. back.		0.180***
high		(0.051)
N	12,894	6,035

Notes: Abbreviations are Prob. for probability, mig. back. for migration background. 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, 2023).

In terms of participating in the survey, we found that the timing of the first contact in the afternoon negatively affects the decision to participate; see column *Participation* in Table 3. When the first contact happened on a Thursday, households were less likely to participate. In contrast, on Sundays, households were more likely to participate. In the early phase of the field in August, only a few households were willing to participate. Households interviewed by interviewers whose native language is German, have a higher probability of participating. When the interviewer is a pensioner or has a low education (indicated by a low CASMIN), the household is less likely to participate. Households located in an area of upper midfield quality have a higher propensity to participate in the panel. When the block of buildings surrounding the household has a high share of children under the age of 18, the probability of participating is lower. The same is true when the household lives in a settlement block other than residential, commercial, or mixed settlement blocks. A household's propensity to participate is lower when the dominant building type surrounding the address is commercial. There is a higher likelihood for households to participate if they are located in a neighborhood with a low unemployment rate or high shares of persons with a Muslim or Italian migration background. In contrast, households located in a neighborhood with a high share of persons with a Soviet migration background are less likely to participate. Findings for the index of socio-economic deprivation are inconclusive. While a household located in a neighborhood with a high value (9^{th} decile) of the index has a lower probability of participating, households located in a neighborhood of the 7^{th} decile of the index have a higher likelihood of participating.

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 have been provided by the Federal Statistical Office, drawing upon data from the German Microcensus 2022. Margins used in the post-stratification process are:

Number of households with at least one person of the M8b population who immigrated to Germany in 2020 or 2021 by

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

Number of persons of the M8b-population who immigrated to Germany in 2020 or 2021 by sex, nationality, and age groups

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 add to the variation and magnitude of weights. Resulting weights are provided in the variable `hhrf0` included in the data set `hpathl`.

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	1.06	1.63	1.76	2.07	3.61	3.86	45.98	2.73	2.10
SWA	2.11	4.60	6.76	10.35	16.68	27.96	272.10	15.54	20.67
PWA	2.10	8.30	13.04	22.02	39.04	72.43	344.61	35.58	45.96

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 2022. Using the resulting standard SOEP weighting factors (`hhrf` included in `hpathl` and `phrf` included in `ppathl`), the sample M8b cases can then be analyzed jointly and comparatively in combination with all other SOEP cases.

6 Summary

The new Sample M8b is a refresher sample adding 2,333 households to the SOEP. It secures and expands the previous analysis potential of the SOEP's immigrant samples M1, M2, M7, and M8. The sample augments the immigration samples by non-EU nationals joining the German labor force. Thereby, it can be used to evaluate the Skilled Labor Immigration Act (Fachkräfteeinwanderungsgesetz) together with sample M8a. The households of this new sample have been seamlessly integrated into SOEP-Core.

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., 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. (2023). 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., & 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.