1104 2022

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

Sampling, Nonresponse, and Weighting of the 2020 Refreshment Sample (M6) of the IAB-BAMF-SOEP Refugee Panel

Hans-Walter Steinhauer, Rainer Siegers, Manuel Siegert, Jannes Jacobsen, 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, Rainer Siegers, Manuel Siegert, Jannes Jacobsen, Sabine Zinn. 2022. Sampling, Nonresponse, and Weighting of the 2020 Refreshment Sample (M6) of the IAB-BAMF-SOEP Refugee Panel. SOEP Survey Papers 1104: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

Sampling, Nonresponse, and Weighting of the 2020 Refreshment Sample (M6) of the IAB-BAMF-SOEP Refugee Panel

Hans Walter Steinhauer[1], Rainer Siegers[1], Manuel Siegert[2], Jannes Jacobsen[3], and Sabine Zinn[1,4]

[1]Deutsches Institut für Wirtschaftsforschung
[2]Bundesamt für Migration und Flüchtlinge
[3]Deutsche Zentrum für Integrations- und Migrationsforschung
[4]Humboldt-Universität zu Berlin

July 18, 2022

Abstract
This paper provides details on the sampling design, fieldwork, nonresponse, and population adjustments for the 2020 sample M6 of the Socio-Economic Panel (SOEP). Sample M6 refreshes the SOEP core samples M3, M4, and M5, sampling households of refugees who arrived in Germany between January 2013 and the end of December 2016. The sample also augments the SOEP core by sampling households of refugees who arrived in Germany between January 2017 and June 2019. Obtaining nearly 1,000 household interviews and panel consent of households for each sample was complicated by the first wave of the Covid-19 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 interviewer characteristics.

1 Introduction

In June 2016, the Socio-Economic Panel (SOEP) started a cooperation with the Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung, IAB) and the Research Centre of the Federal Office for Migration and Refugees (Forschungszentrum des Bundesamtes für Migration und Flüchtlinge, BAMF-FZ) to survey refugees arriving in Germany from January 2013 onward. The study, referred to as the IAB-BAMF-SOEP Survey of Refugees, is designed as a panel study of refugee households and is incorporated into the SOEP as subsamples M3, M4, and M5. The samples M3 and M4 focus on individuals who arrived in Germany between January 2013 and January 2016 and who applied for asylum or were living in Germany as part of Federal Government or Federal States specific programs, independent of the asylum procedure's outcome or their legal status (Kroh, Kühne, Jacobsen, Siegert, & Siegers, 2017). Sample M5 comprises two distinct target populations not covered by M3 and M4. The first population covers adult refugees who arrived in Germany between January 2013 and January 2016 but were registered in the AZR after April 2016, referred to as late registrations. The second population covers adults who arrived in Germany after January 2016 but before December 31, 2016, and applied for asylum by January 2017, referred to as new arrivals (Jacobsen et al., 2019). During the course of the panel, subsamples M3 and M4 started with 3,273 households in 2016. In the latest release of the SOEP for 2019, only 1,764 households remain. Similarly, the number of households in subsample M5 decreased from 1,519 in 2017 to 929 in 2019; see Siegers, Steinhauer, and Dührsen (2021) for more detailed numbers. Therefore, sample M6 refreshes and augments samples M3, M4, and M5 by adding 1,141 households covering two different groups:

- Persons who entered Germany between January 2013 and December 2016 inclusive, filed an asylum application, and whose last change of asylum status took place between the beginning of 2013 and the end of 2016.
- Persons who entered Germany between January 2013 and the end of June 2019, filed an asylum application, and whose last change of asylum status took place between the beginning of 2017 and the end of June 2019.

This paper documents the sampling design and the weighting strategy for the 2020 sample M6 of the SOEP. Therefore, section 2 details the population and provides some further information on the population. Sampling is described in section 3. Section 4 provides detailed 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 M6 consists of two main groups:

1. Persons who entered Germany between January 2013 and December 2016 inclusive, filed an asylum application, and whose last change of asylum status took place between the beginning of 2013 and the end of 2016.
2. Persons who entered Germany between January 2013 and end of June 2019, filed an asylum application, and whose last change of asylum status took place in between the beginning of 2017 and the end of June 2019.

As a result, not only were the previous samples M3, M4, and M5 refreshed, but also expanded to include persons with entry between January 2017 and June 2019.

To sample from this population, we used the Central Register of Foreigners (Ausländerzentralregister, AZR), which is official data provided by the Federal Office for Migration and Refugees (Bundesamt für Migration und Flüchtlinge, BAMF). The Central Register of Foreigners documents all foreigners who are not German nationals or nationals of member states of the European Union and who are staying in Germany for a longer period of time –- at least for three month –- thus also covering the desired refugee populations. More details on the register is provided by Babka von Gostomski and Pupeter (2008). This register was filtered for persons meeting the population definitions presented above. In total the AZR contained 923,056 persons fulfilling these conditions (for details see Table 1).

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

Group	Number	Percent
1	518,127	56.13
2	404,929	43.87
Total	923,056	100.00

Source: AZR, special analysis.

3 Sampling Design

The design of M6 can be summarized as a stratified two-stage sampling design. The population was stratified according to refreshment and augmentation. Within each of the two strata, regional clusters of immigration offices were built to form primary sampling units (PSUs). For a detailed description of the clustering see Jacobsen et al. (2019). For each of the two groups, 100 PSUs were sampled. In the refreshment stratum, anchor persons were sampled within the PSUs using simple random sampling. In the augmentation stratum, anchor persons were sampled disproportionally. For this purpose, a second stratification was implemented within these PSUs. The three strata here contain, first, refugees from western Africa (Nigeria and Guinea) and second, refugees from eastern Africa (Somalia and Eritrea). The third stratum covers refugees from all remaining countries, where the majority comes from Turkey and Syria. PSUs were sampled with replacement, thus some were sampled more than once. A total of 159 unique PSUs was sampled initially. Within these 159 PSUs, we sampled a total of 10,207 individuals, with 6,484 anchor persons for the augmentation and the remaining 3,723 persons for the refreshment.

4 Fieldwork Results and Response Rates

After sampling, the addresses were handed over to KANTAR Public, the field work agency, to be validated. During the fieldwork a subsample of 3,000 addresses was validated. This left 7,207 addresses unused. Interviews were conducted by Kantar Public between August 2020 and February 2021. The households sampled were provided with information sent to them via mail in advance. These letters emphasized that participation is voluntary and

information provided by respondents would not impact any legal proceedings. For more detailed information on the fieldwork see Rathje and Glemser (2021). Table 2 details results of the fieldwork on the household-level. In total, there were 1,141 complete or partial interviews, resulting in a response rate on the household-level, calculated according to American Association for Public Opinion Research (2016), of $RR2 = \frac{1,141}{3,000} = 0.380$. The refusal rate is $REF1 = \frac{638}{3,000} = 0.213$ and, thus, is similar to other samples / studies. The numbers for non-contact as well as for untraceable households are slightly lower than the earlier samples M3, M4, and M5.

Table 2: Fieldwork results on the household-level according to American Association for Public Opinion Research (2016).

Final Disposition Code	M6	
	Number	Percent
1. Interview		
(1.1) Complete	608	20.3
(1.2) Partial	533	17.8
Subtotal	*1,141*	*38.0*
2. Eligible, Non-Interview		
(2.11) Refusals	638	21.3
(2.20) Non-contact	692	23.1
(2.31) Dead	4	0.1
(2.32) Physically or mentally unable/incompetent	21	0.7
(2.33) Language problem	55	1.8
(2.36) Miscellaneous	16	0.5
(2.4) New address after field period	10	0.3
Subtotal	*1,436*	*47.9*
3. Unknown eligibility, non-interview		
(3.11) Not attempted or worked	34	1.1
(3.18) Unable to locate address	1	0.0
Subtotal	*35*	*1.2*
4. Not Eligible		
(4.2) Household moved abroad	18	0.6
(4.4) Household untraceable	370	12.3
Subtotal	*388*	*12.9*
Total	3,000	100.0

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

5 Cross-sectional Weighting

The computation of survey weights is usually performed in three steps (Brick & Kalton, 1996). In a 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 is usually little 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: 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. A 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, mostly 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 like, for example, eight-digit postal code areas (PLZ8) covering approximately 500 households, street level, or household cells aggregating a few households. Finally, we were able to link some information from the AZR data, such as sex, age, nationality, residence status (aufenthaltsrechtlicher Status), and asylum status (Asylstatus).

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. Beyond that, some might also be highly correlated with 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 step-wise algorithm only skipping or keeping them if the model fit improves in terms of the BIC. This three-step procedure yields the final model used in the estimation of participation probabilities that is used to adjust the weights. The models estimating the propensities for contact and participation used to derive weighting adjustments are

presented in Table 3 and Table 4.

When estimating the models for contactability, we find refugees arriving in the second quarter of 2017 as well as refugees with different residence status are harder to locate. Further, refugees with subsidiary protection are harder to locate. The information from the Microm data indicates that refugees in a more settled neighborhood are easier to locate. Additionally, field information points at the fact that refugees in a not well-off neighborhood seem to be harder to locate. The same holds for refugees living in Berlin. Finally, older interviewers as well as interviewers who recently joined Kantar are less likely to locate refugees at their addresses; see Table 3.

When estimating participation propensities, we find refugees from Syria (also being the biggest group) to be more likely to participate than refugees from other countries; see Table 4. Microm information suggests that refugees living in an urban neighborhood are less likely to participate. In contrast, neighborhoods with mostly disciplined people living there as well as neighborhoods with OPEL being the dominant brand of car seem to influence the participation propensity positively. Field information shows that the propensity to participate is higher in North Rhine-Westphalia and in neighborhoods that are residential areas with mostly old buildings, in good condition and where the interviewer feels very safe. Finally, refugees are less likely to participate when being interviewed by a person without a university or only an intermediate school degree.

Table 3: Model estimating contact propensities used to derive weighting adjustments.

	Estimate (Std. Err.)
(Intercept)	0.877***
	(0.043)
AZR information	
Time of arrival in Germany	−0.435**
2017 - Quarter 2	(0.140)
Residence status: suspension of deportation	−0.345**
because of other reasons	(0.119)
Residence status: suspension of deportation	−0.352***
because of missing documents	(0.100)
Residence status	−0.228***
temporary resident permit	(0.066)
Residence status	−0.206**
subsidiary protection acc. §4 Abs.1 AsylG	(0.066)
Microm data on PLZ8-area	
Phase of live by socio-economic status	−0.451**
financially solid young couples	(0.162)
Fluctuation in neighborhood	0.305***
slightly below average	(0.087)
Type of PLZ8-area	0.253***
Residential neighborhood in fringe area	(0.072)
Dominant migrant milieu	0.275***
Disrooted milieu	(0.074)
Dominant sub-typology	0.223**
Consumption oriented	(0.072)
Field information	
Interviewer's year of birth	−0.442**
1950 – 1960	(0.149)
Condition of the neighborhood	−0.236**
Mixed impression, partly derelict state	(0.075)
Federal state	−0.372***
Berlin	(0.078)
Interviewer for Kantar since	−0.207***
2020	(0.061)
N	2978

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, 2021).

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

Table 4: Model estimating participation propensities used to derive weighting adjustments.

	Estimate (Std. Err.)
(Intercept)	-1.711^{***}
	(0.169)
AZR information	
Nationality	0.377^{***}
Syria	(0.064)
Microm data on PLZ8-area	
Type of PLZ8-area	-0.427^{**}
City center area	(0.145)
Primary limbic type	0.225^{**}
Disciplined	(0.075)
Dominant car brand	0.245^{**}
OPEL	(0.081)
Field information	
Federal state	0.356^{***}
North Rhine-Westphalia	(0.064)
Condition of the neighborhood	0.272^{***}
In very good condition	(0.066)
Safety of the neighborhood	0.720^{***}
Very safe and pleasant	(0.167)
Description of neighborhood	0.298^{***}
Residential area with old buildings	(0.067)
Interviewer: educational level	-0.489^{***}
College / University without a degree	(0.116)
Interviewer: educational level	-0.445^{***}
Intermediate school degree	(0.106)
N	2597

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, 2021).

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 Office for Migration and Refugees based a special analysis of the Central Register of Foreigners. At the individual level, the following marginal and joint distributions have been used:

- Number of Persons by federal state
- Number of Persons by date of immigration
- Number of Persons by nationality
- Number of Persons by age group and gender

5.3 Characteristics of Weights

Table 5: 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	5	6	13	61	89	182	1033	77	87
SWA	7	15	51	125	230	475	2535	201	252
PWA	5	40	115	246	559	1171	12506	504	870

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 add to the variation and magnitude of weights.

6 Summary

M6 secures and expands the previous analysis potential of the IAB-BAMF-SOEP Survey of Refugees. Firstly, by refreshing the previous samples M3, M4, and M5, thus preserving the analysis potential on persons who came to Germany as protection seekers between January 2013 and the end of December 2016. In addition, M6 also makes it possible to look at persons who came to Germany from 2017 onward. Finally, M6 allows for a more differentiated view of persons originating from Africa who have sought protection in Germany.

References

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

Babka von Gostomski, C., & Pupeter, M. (2008). Zufallsbefragung von Ausländern auf Basis des Ausländerzentralregisters: Erfahrungen bei der Repräsentativbefragung "Ausgewählte Migrantengruppen in Deutschland 2006/2007" (RAM). *Methoden, Daten, Analysen*, *2*(2), 149–177.

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

Jacobsen, J., Kroh, M., Kühne, S., Scheible, J. A., Siegers, R., & Siegert, M. (2019). *Supplementary of the IAB-BAMF-SOEP Survey of Refugees in Germany (M5) 2017* (SOEP Survey Papers No. 605). Berlin: DIW/SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.615884.de/diw_ssp0605.pdf

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.

R Core Team. (2021). 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

Siegers, R., Steinhauer, H. W., & Dührsen, L. (2021). *SOEP-Core v36 – Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2019)* (SOEP Survey Papers No. 960). Berlin: DIW/SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.816375.de/diw_ssp0960.pdf