Siegers, Rainer; Steinhauer, Hans Walter; Zinn, Sabine

Research Report

Weighting the SOEP-CoV study 2020

SOEP Survey Papers, No. 989

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

Suggested Citation: Siegers, Rainer; Steinhauer, Hans Walter; Zinn, Sabine (2021) : Weighting the SOEP-CoV study 2020, SOEP Survey Papers, No. 989, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
http://hdl.handle.net/10419/235215

989 2021

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

Weighting the SOEP-CoV study 2020

Rainer Siegers, Hans Walter Steinhauer, 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:

Rainer Siegers, Hans Walter Steinhauer, Sabine Zinn. 2021. Weighting the SOEP-CoV study 2020. SOEP Survey Papers 989: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

Weighting the SOEP-CoV study 2020

Rainer Siegers, Hans Walter Steinhauer, Sabine Zinn

Last updated: August 6, 2020

Abstract
This paper details the sampling design, results from the field, analysis of selectivities, as well as raking procedures used in the SOEP-CoV study. The sample consists of a random selection of 12,000 households in the Socio-Economic Panel Study (SOEP). The sample was divided and surveyed in nine successive tranches. The largest tranches were surveyed first starting April 1, 2020. By the end of May, all of the smaller remaining tranches had been surveyed.

1 Overview

The Socio-Economic Panel (SOEP) is a longitudinal study conducted at the German Institute for Economic Research (DIW Berlin). The study began in 1984 with an annual survey of households and their members in Germany. As a result, SOEP data can be used to describe and analyze trajectories and changes that occur due to outside influences. In spring 2020, in addition to the regular face-to-face interview, SOEP households were surveyed about their experiences during the corona crisis using computer-assisted telephone interviewing (CATI). For more information on the design and content of the SOEP-CoV study, see Kühne, Kroh, Liebig, and Zinn (2020). Results from the study, some published in a Spotlights series, are available at www.soep-cov.de.

At the time of planning and later weighting the study, the most recently published SOEP scientific use file was version 35, which covers the survey years from 1984 up to and including 2018. At this time, the SOEP at DIW Berlin also had the data from 2019 but was still processing and publishing this data. Changes occur over time in the composition of the sample of households due to births, deaths, household members moving out, and new members moving in. In addition, households or individual household members may decline participation in a given survey year or may withdraw from the study altogether. Because of all these changes in household structure, the households selected for the SOEP-CoV study were those that had participated in at least one survey in 2018 or 2019 and had not explicitly declined participation before the start date for 2020 fieldwork. Of the remaining households, the following were also excluded:

- Households in the refugee samples M3, M4, and M5. These will be interviewed by telephone about their experiences in the corona crisis as part of a separate survey led by the Institute for Employment Research (IAB).
- Households in the samples that were interviewed for the first time in 2019 (i.e., subsamples P and Q) in order not to jeopardize their willingness to participate in the regular second wave.
- Households that are normally interviewed through what is known as "central processing". The SOEP survey institute (Kantar Public) uses this method for households that cannot or do not want to be contacted through the usual SOEP channels (by interviewers). The participants in "central processing" are usually contacted by telephone and complete the questionnaire independently or with assistance over the telephone. Thus, "centrally processed" households are those that already show a high propensity toward non-participation in the regular SOEP survey. These households should not be additionally burdened by special surveys.
- Households without a valid telephone number, since they cannot be interviewed by telephone as part of the SOEP-CoV study.

The sample of remaining households was updated in terms of composition and contact information by the SOEP survey institute in the period prior to March 2020 and returned to SOEP as a gross sample for the SOEP-CoV study. These households were then randomly distributed among a total of nine tranches. These were surveyed successively in order of size, with the largest surveyed first: The tranches were constructed in such a way that sample size decreased over time. This approach took into account the fact that people in Germany faced the greatest challenges and thus the most significant changes in their daily lives during the first weeks of the complete lockdown (and thus during fieldwork on the first four tranches).

The first four tranches were the largest, with a survey period of two weeks each. The remaining five tranches were smaller, and their interview periods were extended by one additional week. A few interviews could only be carried out with a few days' delay, which means that the interview periods for the tranches overlap to some extent. The interview periods and sample sizes by tranche are shown in Table 1. Fieldwork for the SOEP-

Table 1: Fieldwork and sample sizes by tranches.

	Fieldwork (2020)		Household status in sample		
Tranche	Start	End	Included	Contacted	Completed
1	April 1	April 18	2,756	2,068	1,689
2	April 14	May 2	3,296	2,450	1,932
3	April 27	May 16	1,767	1,310	978
4	May 11	May 30	1,183	871	632
5	May 25	June 6	608	443	309
6	June 2	June 13	629	450	303
7	June 8	June 20	578	409	288
8	June 15	June 27	598	433	298
9	June 22	July 4	584	405	265
1-9	April 1	July 4	11,999	8,839	6,694

CoV study started on April 1, 2020, and finished on July 4, 2020. The total number of households in tranches 1 to 9 was 11,999. Of these, 8,839 households could be contacted by telephone and 6,694 actually participated in the SOEP-CoV study.

A graphical presentation of the sample sizes by status (contactability as well as willingness to participate) and tranches can be found in Figure 1. The left part of the figure shows the distribution of contact and participation status by tranche in absolute numbers, the right part shows the respective percentages.

The figure on the right shows that the percentage of participating households fell slightly but steadily over time. Here, it is reasonable to assume that the initially high level of interest in the topic of corona among the population declined over time. The proportion of households that could not be contacted, however, remained virtually unchanged across the tranches.

2 Procedure for SOEP-CoV weighting

The weighting of the SOEP-CoV study was largely analogous to the weighting of the SOEP-Core study. This is described in detail by Kroh, Siegers, and Kühne (2015) and is documented for the current version 35 in Siegers, Belcheva, and Silbermann (2020).

The initial household weight was based on the most recent available household weight from SOEP-Core (hhrf), that is, usually the weight from wave *bi* (SOEP scientific use file, version v35). This was adjusted in the SOEP-CoV study for successive decision processes at the household level and afterwards with respect to various population distributions taken from the 2018 Microcensus.

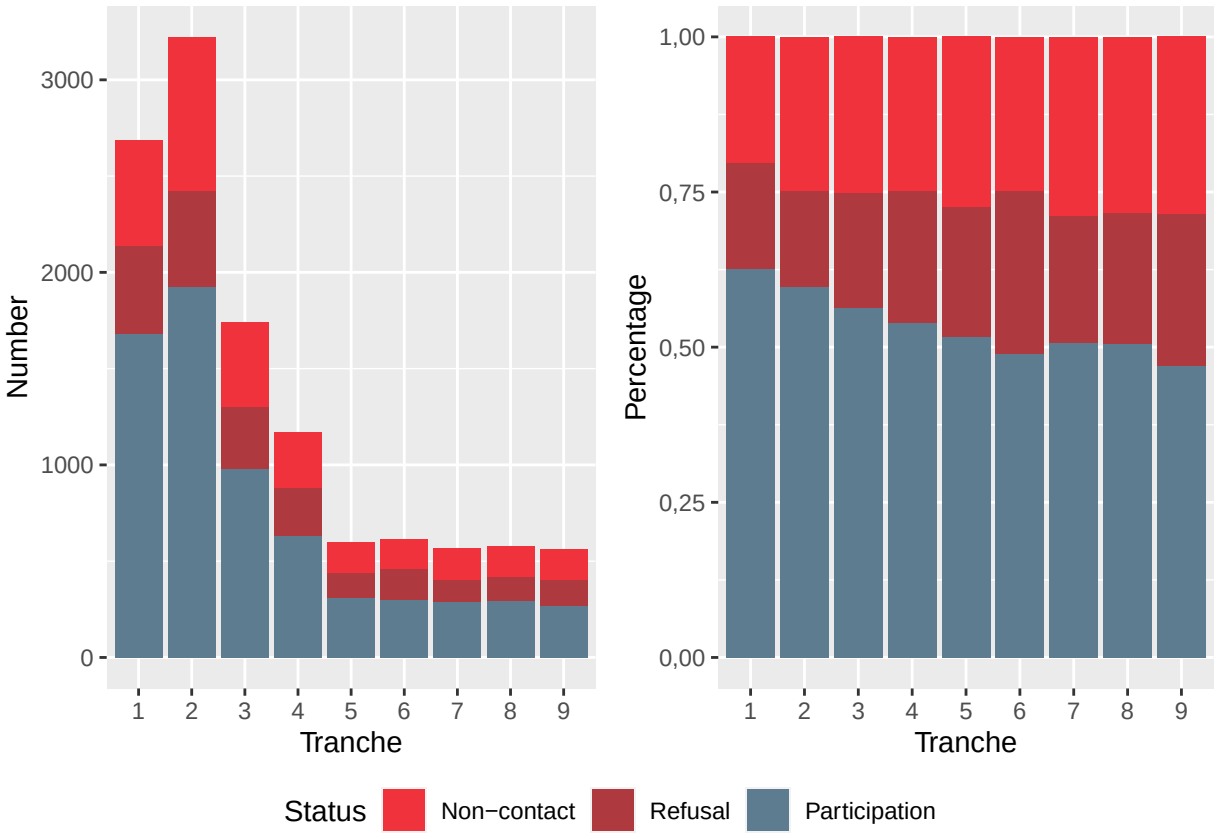

Figure 1: Included samples by tranche and status.

Based on these household weights, we generated weights for all persons in the participating households using another raking adjustment. For the person in the household who participated in the CATI survey, a final weighting step was performed to correct for any selection effects that occurred.

The following Figure 2 shows the weighting procedure. Specifically, in the first step, the initial weights were corrected for changes between the composition of the SOEP in 2018 and 2020. In this context, the 2018 SOEP household weights were adjusted for entries to (new members moving into existing households, newborns) and departures from (deaths in the household, panel attrition) the sample. In the subsequent step, adjustments were made for households that were excluded from participation in the SOEP-CoV study from the outset (see Section 1).

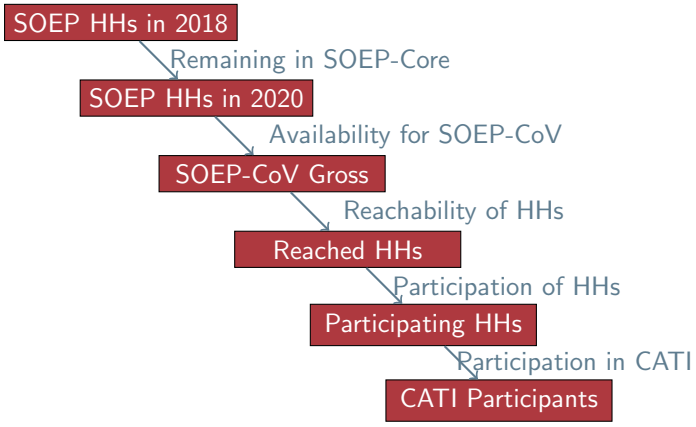

Figure 2: Steps of the weighting procedure for SOEP-CoV (HH: household).

In order to make the data available in a timely manner, the sample of the SOEP-CoV study was weighted after certain tranches had been completed. This took into account the use of tranches and extrapolated the households to the population for each of these subsets. In particular, the use of subsamples M1 and M2 (migration samples), which only took place from the second tranche onward, was taken into account in this step.

In order to achieve a sample of a variety of different household members, all households were called at different times of the day from 7 am to 9 pm. In general, it was also assumed that, due to exit restrictions and the increased proportion of people who worked from home as a result of the crisis, respondents would be easier to contact by telephone than before the crisis. The corresponding distribution of calls by day of the week, time of day, and connection is shown in Figure 3. Nevertheless, there remain between 25 and 31 percent of households that could not be reached in the respective survey period (see Figure 1 above). In the third step of weighting, we therefore corrected for the contactability of households within the respective survey periods.

Finally, in the fourth step, we corrected for the willingness of households to participate in the SOEP-CoV survey. Between 69 and 75 percent of the households in each tranche of SOEP-CoV could be reached. Averaging across tranches 1 to 9, 73 percent were reached. Of the households reached, between 65 and 82 percent of households in each tranche were interviewed successfully. Averaging across tranches 1 to 9, 72 percent were interviewed. Thus, the response rate according to AAPOR (The American Association for Public Opinion Research, 2016) is RR1 = 0.558. Within each tranche, this rate varies between 0.454

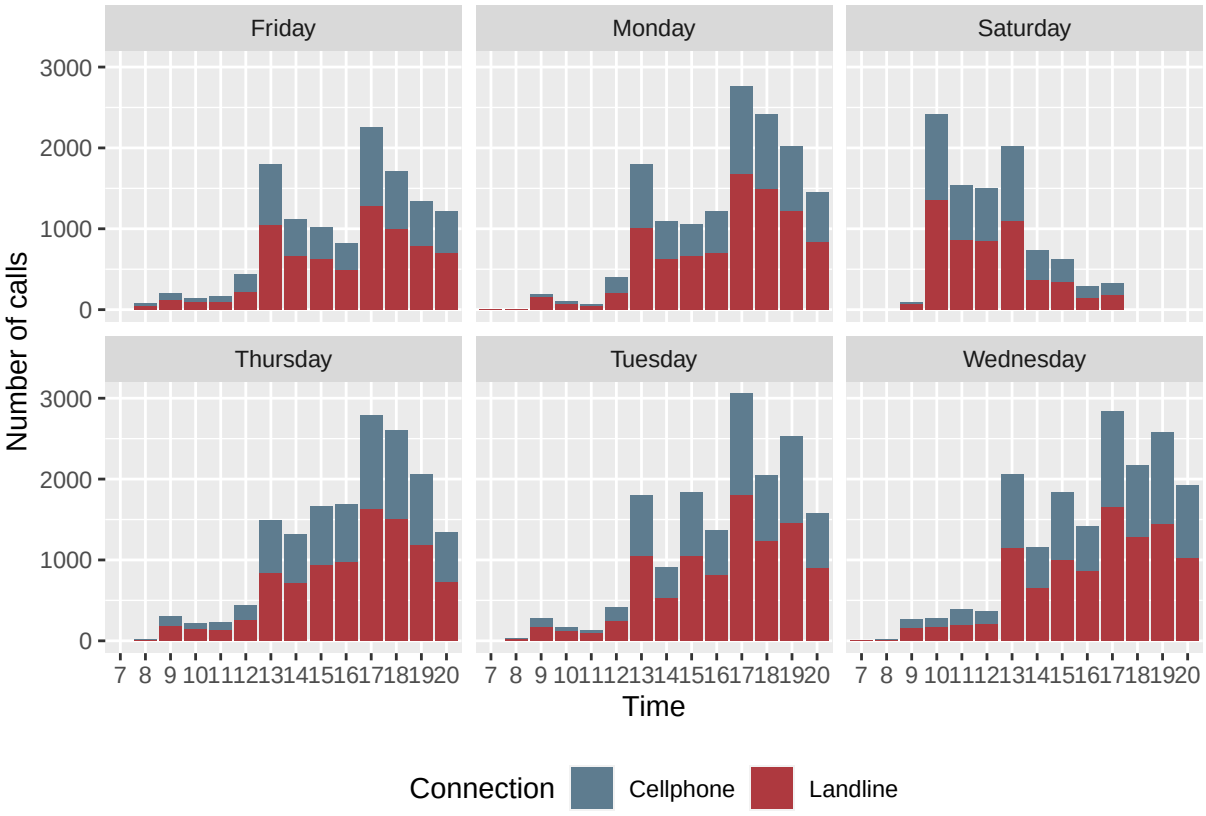

Figure 3: Number of calls by time of day, day of the week, and type of telephone connection.

and 0.613. This step was followed by a marginal adjustment to a variety of population distributions (see Section 5), which completes the household-level weighting.

Projection factors for individual household members were then created based on the household weights in another marginal adjustment step. The procedure and the marginal distributions used for this purpose are described in more detail in Section 5.

Based on this individual weight, we generated projection factors for the respondent in a participating household in a final step. In this step, we corrected for selective (self-)selection of the respondent for households with at least two adults.

3 Characteristics used in weighting

More than 400 characteristics at the household and personal level were used in the default models (cloglog regressions) of the SOEP-CoV weighting. The majority of the characteristics originate from previous waves of the *SOEP panel data*. Overall, variables from numerous SOEP survey areas were included, such as demographics, work, health, education, family, finances, personality, migration, and political attitudes. In addition, where reasonable and possible, personal characteristics aggregated at the household level were included in the default models. A list of characteristics used for weighting SOEP-Core version 35 can be found in Siegers et al. (2020, 63f, 70ff).

Information on contact history was also included in the weighting. The survey institute provided the SOEP with *contact logs of telephone histories* for a total of 86,069 calls. These include information on successful and unsuccessful contact attempts, information on the date and time of a contact attempt, whether a call was made using a landline or cellphone number, and the return code for the particular contact attempt. From this information, we created additional variables that indicate, for example, what type of telephone number (landline, cellphone, both) was used to contact a household and how often a household was contacted at certain times of the day.

Furthermore, the *current daily corona case numbers* (number of infected, deceased, recovered persons) at the county level on the day of the contact attempt or interview were used. The corresponding data are made publicly available by the Robert Koch Institute.[1] Using county-level population data provided by the Federal Statistical Office, the county-level corona incidence was calculated in addition to the above variables.[2] This incidence was also part of the weighting variables.

Similarly, small-scale information below the county level, predominantly on the *social structure of neighborhoods*, was incorporated into the default modeling. Corresponding data are provided by Microm.

Table A.1 in the online materials summarizes all variables that were tested in the different models for their influence on inclusion in the sample, accessibility, and participation.[3]

[1] The most recent data in each case can be downloaded from https://opendata.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0.csv.

[2] The data can be downloaded from GENESIS-ONLINE as Table 12411-0015. Data as of December 31, 2018.

[3] Online materials are available at www.soep-cov.de/Gewichtung

Not all variables are included in every attrition model. The reason for this is obvious: Of the over 400 available characteristics, as expected, many have no effect on the variable being explained (i.e., sample inclusion, contactability, or participation) and/or are highly correlated with each other. Including an unnecessarily large number of explanatory variables in a model creates a large scatter in the weighting factors to be generated (which are the inverse of the predicted inclusion, contact, and participation probabilities). This should be avoided in all cases for reasons of sampling efficiency.

Therefore, prior to any multivariate modeling, all variables were individually tested for their association with the variable being explained (i.e., inclusion in the sample, contactability, and participation). Only if this association was significant ($p<0.05$) was the corresponding variable included in the preliminary set of explanatory variables for the corresponding attrition model. For reasons of model efficiency, highly correlated characteristics were also excluded from the set of explanatory variables. For this purpose, the correlation of all explanatory variables among each other was determined. Of the characteristics with a correlation greater than 0.95, only the one that had the greatest (significant) influence on the variable to be explained (i.e., inclusion in the sample, contactability, or participation) was included in the corresponding model. This led to different sets of explanatory variables for the different attrition models.

In a final step, variable selection was performed using the Bayesian Information Criterion (BIC). Variables were iteratively removed from or added to the respective model if this change in the model led to a lower BIC and thus to better model quality. This three-step procedure for variable selection described here was applied to each of the attrition models estimated in the SOEP-CoV weighting.

4 Estimated weighting models

This section presents the models estimated for the above weighting steps.[4] The results are presented in the form of coefficient plots. Plotted on the y-axis are the characteristics that were included as explanatory variables in the respective weighting model. Parallel to the x-axis are the values of the estimated coefficients (red dots) together with their 95% confidence interval (red bars with vertical ends). The dashed vertical line marks the value 0. The estimated coefficients are sorted from the smallest (top left) to the largest (bottom right). Characteristics whose coefficient estimates lie to the left of the gray dashed line indicate a negative influence. Characteristics whose coefficient estimators lie to the right of the gray dashed line indicate a positive influence.[5]

[4] To estimate the models, the `glm` function of the free statistical software R is used in version: 4.1.0 (R Core Team, 2020). For the preparation of the results, the packages broom (Robinson & Hayes, 2020), gridExtra (Auguie, 2017), kableExtra (Zhu, 2019), and tidyverse (Wickham et al., 2019) are used. This paper was created using rmarkdown (Xie, Allaire, & Grolemund, 2018).

[5] In general, a coefficient estimator whose confidence interval includes zero has no significant influence on the variable being explained. However, with the variable selection method used here, all explanatory variables in the final attrition models of the SOEP-CoV weighting have a significant influence.

4.1 Attrition between 2018 SOEP and 2020 SOEP-CoV sample

Figure 4 shows the estimated coefficients and their confidence intervals for the model with cloglog link, which was used to correct for attrition between the 2018 SOEP wave *bi* and the 2020 gross sample of households. We find that non-participation in the SOEP in 2018 had a significant negative effect on the probability of remaining in the SOEP in 2020. Further, the use of translation aids in the migration samples in the last survey, as well as belonging to the migration samples M1 and M2, negatively affected the propensity to participate. Households with very young household members were significantly less likely to stay in the SOEP, as were households with older household heads.[6] Not having an Internet connection in the household also negatively affected the probability of remaining in the SOEP. If at least one person lived in the household who indicated that he or she was particularly attached to home, there was a negative effect on the probability of remaining. The same is true of characteristics that are related to missing values (specifically, partial unit non-response and a high proportion of item non-response at the household level). Finally, the fact that the last interview was conducted late in the field phase also had a negative effect on SOEP retention.

In contrast, the presence of a party preference and a strong political interest on the part of at least one household member had a positive effect on the probability of remaining. It also had a positive effect if one person in the household was single or if at least one person in the household was an essential worker. Households with two adults but no children and households in which the additional mother-child survey questionnaire was completed in the last survey had a higher probability of remaining in the SOEP than households with more than two persons but no children and households in which the mother-child questionnaire was not completed. Members of subsample L3, which contained only the single-parent and multiple-child family types at the time of sample selection, also had a higher probability of remaining.

4.2 Cases included in tranches

For the SOEP-CoV survey, only households for which a current telephone number was available and which had not been part of the survey institute's "central processing" in the last survey were eligible (see Section 1). In the model described in the following (cp. Figure 5), the (potential) selective bias in the gross sample for SOEP-CoV compared to the 2018 SOEP sample is examined and quantified using information from the 2018 SOEP survey.

Figure 5 shows the estimated coefficients and their 95% confidence intervals for the associated default model with cloglog. Again, the characteristics whose coefficient estimators lie to the left of the gray dashed line are less present in the gross sample of SOEP-CoV than in the overall SOEP. Non-participation in the SOEP survey in 2018 as well as households with young household heads (below age 35) were less present in the gross sample. The same is true of households that at least one person has moved out of since 2018 and households in eastern Germany (households in Thuringia and Saxony-Anhalt, and households in subsample C forming the gross sample for households in eastern Germany from 1990). A high level of item non-response at the household level, as well as at the individ-

[6] Head of household is the person who is most familiar with household matters or the person who filled out the household questionnaire in the last interview.

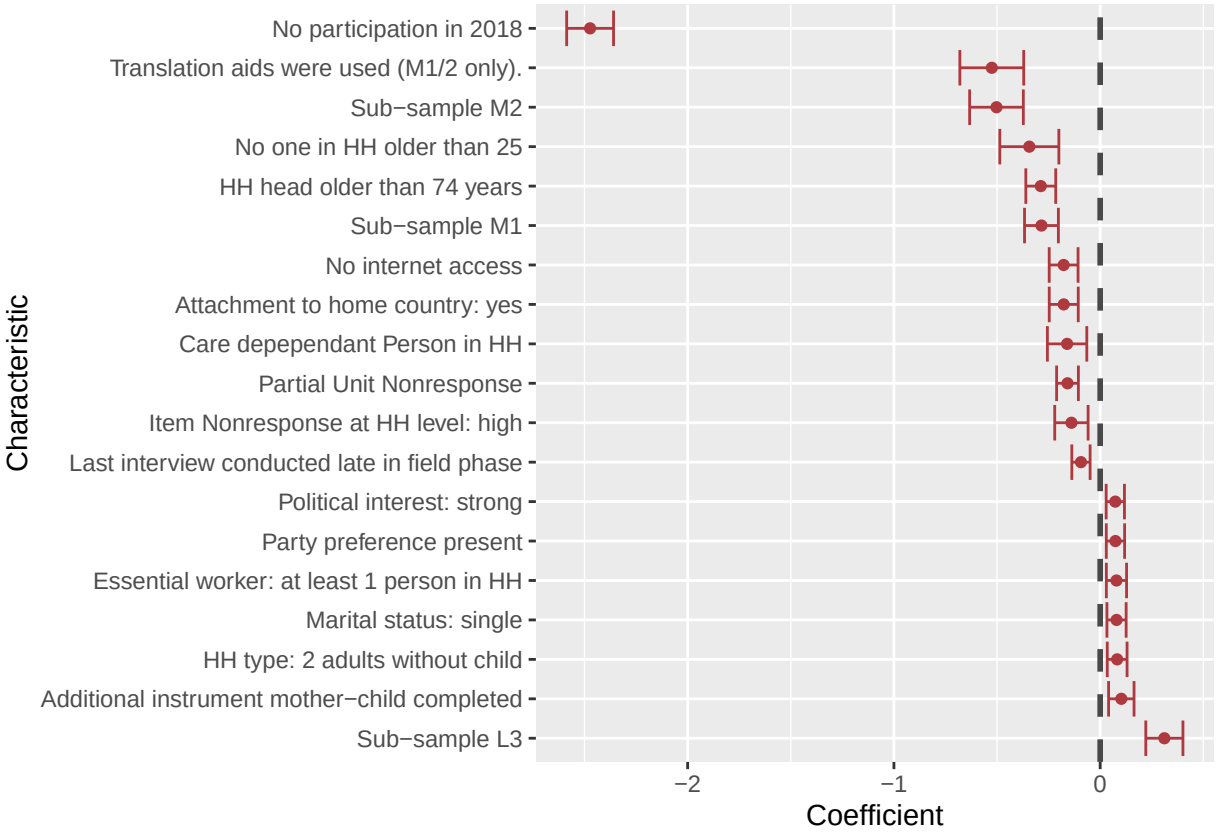

Figure 4: Coefficient plot of the model to correct for dropouts between the 2018 survey and the SOEP-CoV study. (HH: Household.)

ual level, led to a lower probability. Also, belonging to subsamples A (baseline sample West Germany; 1984) and O (households in Socially Integrative City areas; 2018) led to a lower probability of remaining in the sample. Finally, households with two adults and no children and "other" household compositions were also less likely. Finally, dissatisfaction with family life had a negative impact on remaining in the sample.

In contrast, households in which at least one person had more than 3 hours of free time on weekdays, in which the oldest household member was older than 65, in which at least one person was self-employed, in which the head of household was older than 74, and in which the head of household was not yet living in the household at the time of sampling were more likely to remain. Also disproportionately included in the gross sample were households for which the interview in the most recent survey was particularly long (fourth quartile of the survey duration distribution) or short (first quartile of the survey duration distribution). Also more likely to have remained in the sample were households in subsamples J (top-up from 2011), K (top-up from 2012), subsamples from 2010 and 2011 focusing on different family types L1 (birth cohorts from 2007 to 2010), L2 (low-income, single-parent, multi-child families), and L3 (single-parent, multi-child families). The same was true for the migration samples M1 from 2013 and M2 from 2015, as well as for subsample N (top-up from 2017).

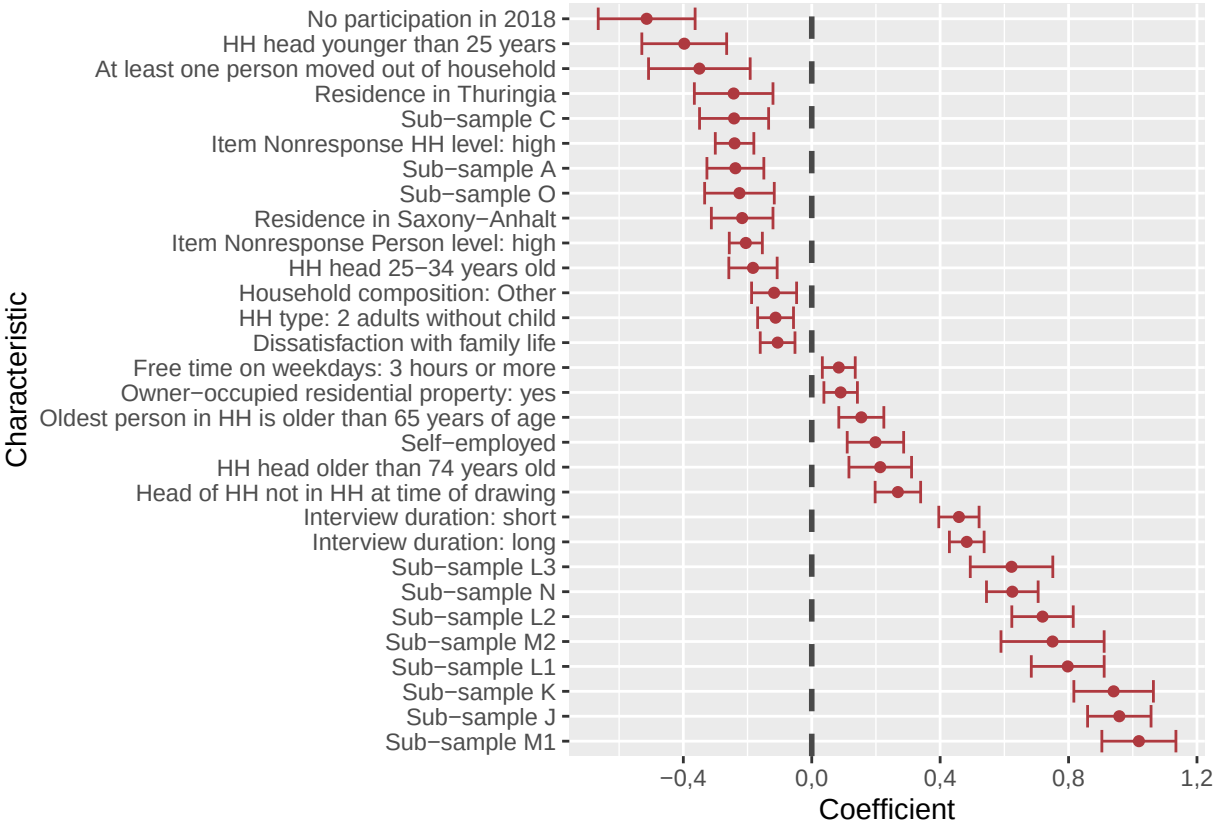

Figure 5: Coefficient plot of the model to correct for the design-related omission of households from "central processing" or without a known telephone number. (HH: Household.)

4.3 Telephone reachability of households

In contrast to the previous SOEP survey, which was conducted mainly by means of a personal computer-assisted (CAPI) or paper-and-pencil (PAPI) interview, this study was conducted as a telephone survey (CATI). There were a number of reasons why households could not be reached by telephone in this survey: for instance, because their home phone, work number, or fax number were wrong or because individuals within the household had died or moved abroad since the last survey. In addition, a small part of the sample had a blocking notice for telephone interviews with the ADM (Association for Interest Representation, Self-Regulation and Standards in German Market and Social Research, www.adm-ev.de) and were therefore not allowed to be contacted by telephone. Other households could not be reached for other reasons during the fieldwork for the respective tranche.

Figure 6 shows the estimated coefficients and their confidence intervals for the model with cloglog used to control for household reachability. Data on times and frequency of telephone contacts were used to describe the contactability of households; see also Figure 3. Some households were particularly difficult to contact and were therefore called frequently (11-25 calls) on landlines and cell phones and especially in the afternoon to evening. Some households that were called less often on landlines or at other times were difficult to reach. The same was true of households in migration samples M1 and M2. Households that had no assets in the previous year and in which at least one person smoked also had a lower probability of being reached.

Households in which at least one person has a preference for a particular political party had an increased probability of being reached by telephone. Households in which at least one retired person lived were also easier to reach. Likewise, households in owner-occupied housing and households with at least one person employed in the public sector were easier to reach. Finally, households that were contacted exclusively by landline also showed higher reachability.

4.4 Household participation in the SOEP-CoV study

The households that could be reached by telephone during the respective survey periods then decided whether or not to participate in the SOEP-CoV study. Figure 7 shows the estimated coefficients and their confidence intervals for the model with cloglog used to correct for refusals to participate in the SOEP-CoV study. Among the factors that negatively influenced household participation decisions, the predictors with the strongest influence were not having an Internet connection in the household, one or more respondents within the household not participating in the last SOEP survey, and the household head being older than 74 years of age. In addition, we observe a lower probability of participation for households in which there was at least one person of non-German nationality or households in which at least one person believes that refugees are bad for the economy. The same was true of households in which at least one person was born abroad and in which at least one person receives Unemployment Benefit II. Finally, the probability of participation was reduced if at least one person in the household did not have a high school diploma.

On the other hand, it had a positive effect if the household was in a neighborhood with a high proportion of households subscribing to national newspapers or if at least one

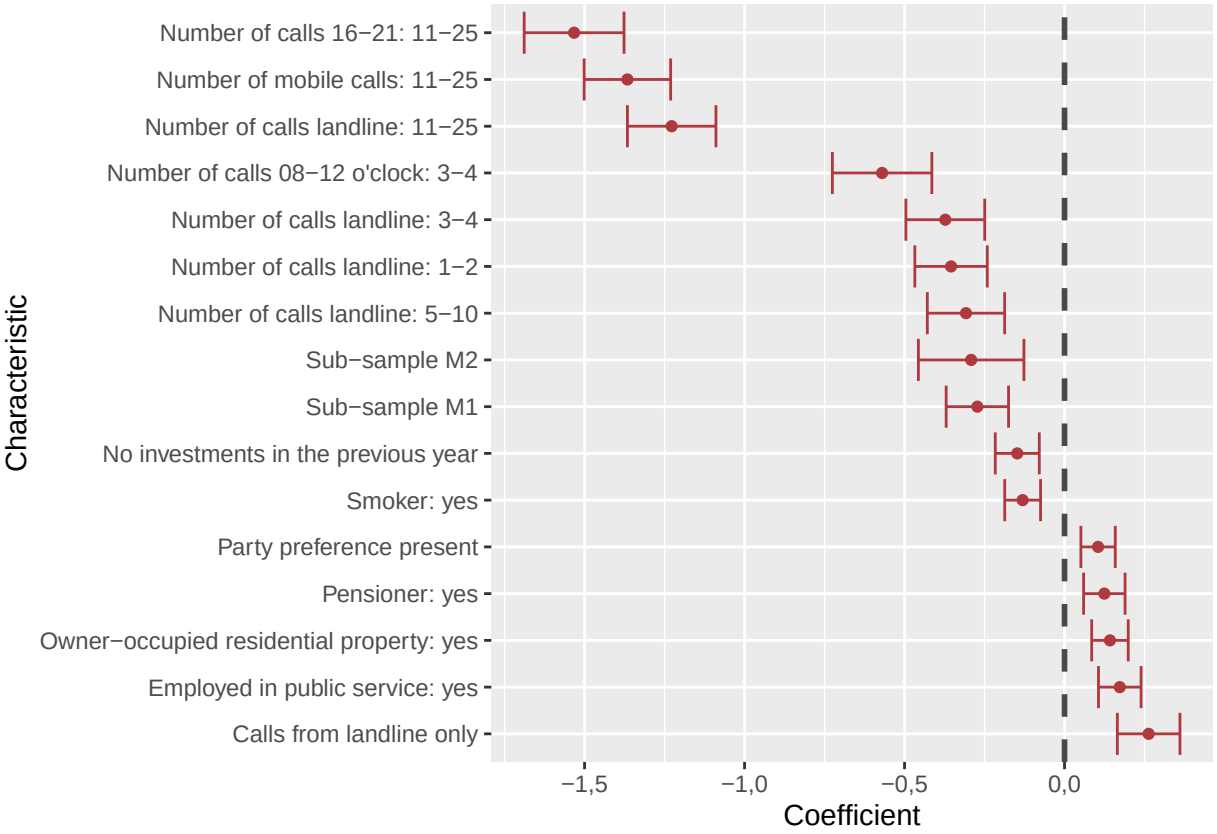

Figure 6: Coefficient plot of the model to correct for unreachable households in the SOEP-CoV study. (HH: Household.)

person in the household had a strong interest in politics. Households in which at least one person reported having no concerns about foreigners or in which at least one person held a university degree were also more likely to participate. Households in Bavaria had a higher probability of participation, as did households with a female household head. The grouping into tranches was also accompanied by negative effects, however, which can be explained by the fact that in the later tranches, male participants were specifically asked to participate in a telephone interview. Finally, belonging to subsamples L2 (family types: low-income, single-parent, multi-child families) and L3 (family types: single-parent, multi-child families) had a positive effect on the participation decision.

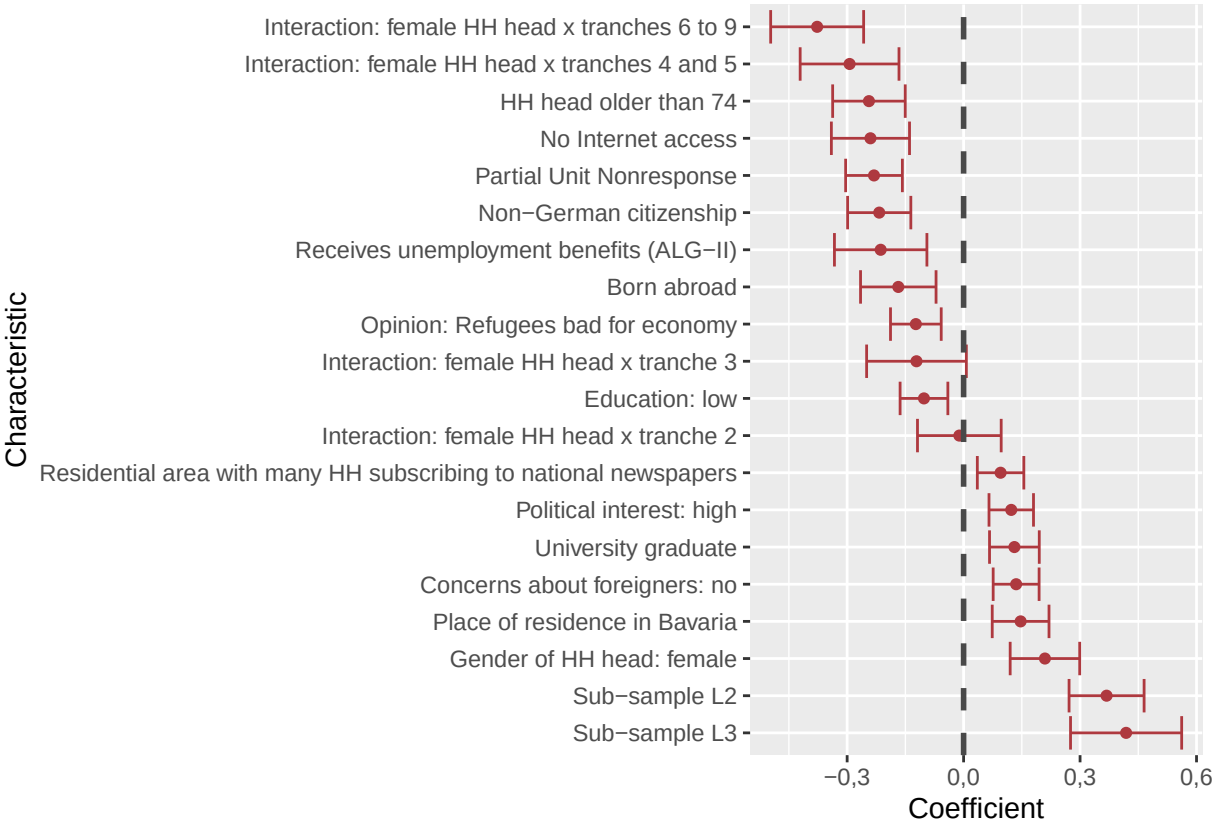

Figure 7: Coefficient plot of the model to correct for refusal of contacted households to participate. (HH: household.)

4.5 Contact for the telephone interview

In SOEP-CoV, only one person per household was interviewed. Although this individual provided some proxy information about other household members, they mainly gave information on themselves. The selection of the contact person was not systematic, but depended on who answered the phone at the time called and was willing to participate in the interview. In general, calls were made throughout the day, but more frequently in the late afternoon and evening, so that working people could also be interviewed (see also Figure 3). To reduce bias with respect to the gender of the respondent, both the head of household and a regular male household member were asked. Since participation in the CATI for the SOEP-CoV study required that the person to be interviewed was at least 18 years old at the time of the interview, only SOEP household members who met this

criterion were also included in the modeling. In addition, only persons from households in which at least two adults lived were included in the modeling, since in successfully contacted single-member or single-parent households, it is clear which person answers the questions.

Figure 8 shows the estimated coefficients and their confidence intervals for the model with cloglog link used to correct for individual-level bias. Looking at selection within participating multi-person households, it appears that persons aged 18 to 24 were less likely to participate in the CATI survey than older individuals. Similarly, individuals with a high school diploma and persons in the age groups "65 to 69" and "70 years and older" showed a lower probability of participation than persons without a high school diploma or individuals aged 25 to 68. The same was true of men and individuals in full-time employment.

On the other hand, individuals with a university degree and those in essential occupations were more likely to participate in the CATI survey. The same was true of individuals in two-person households compared to those in larger households. Middle-aged individuals were also more likely to participate in CATI, as were individuals who had already been tested for Covid-19 with a negative result. Finally, the head of household in the 2018 survey had a higher likelihood of participating in the CATI survey.

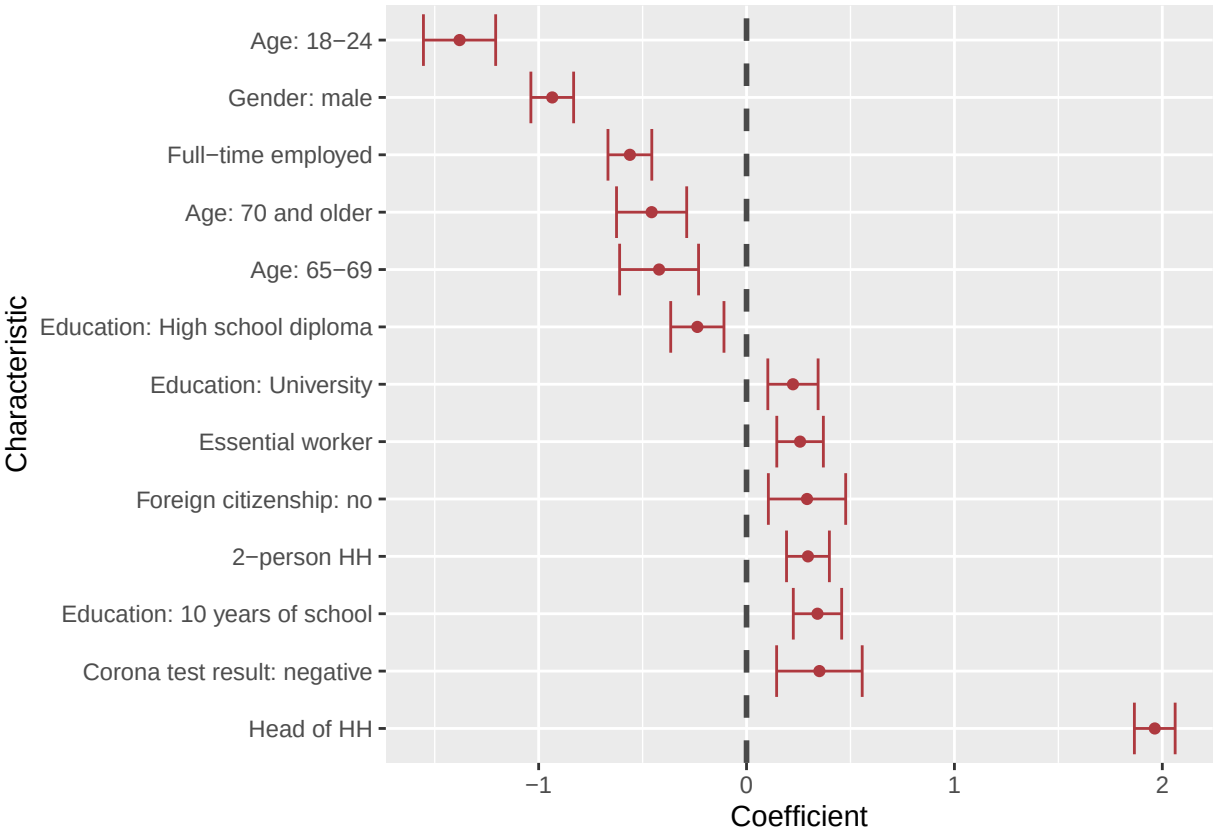

Figure 8: Coefficient plot of the model for correction to CATI participation in the SOEP-CoV study. (HH: Household.)

5 Trimming and raking

With the goal of improving the statistical efficiency of weighted analyses, weights were trimmed. Trimming the weights reduces the variance and thus counteracts a possible bias of weighted analyses due to individual observations with large weights. Here, the weights were not capped at a certain value, but redistribution according to the "weight distribution" method (see Potter, 1990).

This method is based on the parametric assumption that the weights w follow an inverse beta distribution with distribution function F_w. The two parameters of the distribution are estimated from the weights and a maximum value τ is calculated such that $1 - F_w(\tau) = 0.99$. Weights exceeding this value τ are trimmed at this maximum value and the excess mass is distributed among the remaining weights. Now a new maximum value $\hat{\tau}$ is calculated for the weights trimmed in this way, analogously to the procedure above. If there are weights larger than $\hat{\tau}$, they are trimmed at the new maximum value and the remaining mass is again redistributed to all weights smaller than $\hat{\tau}$. This procedure is iteratively repeated until none of the trimmed weights is larger than the new maximum value or, in other words, until $\tau = \hat{\tau}$. Trimming the weights was applied first at the household level and second at the individual level in the CATI weighting step.

To compensate for sampling errors and undercoverage, all weights were adjusted to known marginal distributions in a final step. For this purpose, the raking procedure described in Deville, Särndal, and Sautory (1993) was applied. Since marginal distributions cannot yet be provided by the Federal Statistical Office for 2020 (e.g., by the corresponding microcensus), the last available marginal distributions of the microcensus from 2018 were used for the marginal adjustments at the household level and for all persons of the household. A third weighting factor provided, extrapolates only contacted persons. Since these were all adults and we did not have marginal distributions from the Microcensus for this population, the corresponding marginal distributions for adults were estimated using the 2018 SOEP data.

At the household level, distributions for the number of households by state, household size, community size class, owner-occupied property, household type, and the last immigration year a new household member moved in from abroad were used for the margin adjustment. The corresponding margin adjustment step occurred after the weighting step, which adjusts for household-level bias in an interview that was completed successfully in a household, and trimming of the weights. The household-level margins, along with their characteristics and associated frequencies, are shown in Table A.2 in the online materials.[3]

At the individual level, distributions for the number of persons in the population by age, gender, citizenship (German vs. other) were used for the marginal adjustment of the weights. This margin adjustment was done on the individual weights for all household members in an interviewed household. The margins at the individual level in successfully interviewed households, together with their characteristics and the associated frequencies, are shown in Table A.3 in the online materials.[3] For margin fitting following the CATI weighting step, the margins from Table A.4 in the online materials were used.[3]

6 Summary of the weights

Table 2 shows the number of households and individuals who participated in the SOEP-CoV study for each tranche. Since only one person was interviewed per household, the number of persons participating in CATI is identical to the number of households. In addition, the table contains information on how many households and individuals living in them have a weight with the value 0. Since only one person per household participated in the CATI survey, the CATI weights for the remaining persons in the household also have the value 0. Weights with the value 0 occur because a snowball procedure was used in subsample D (1994/5 migration (1984-1994, West)). Due to this, no inclusion probabilities and thus no weights could be calculated for certain households. Here, household weights are marked with hhrf, weights for all household members with phrf, and the weights of persons who could be interviewed by CATI in the SOEP-CoV study with phrf_cati.

Table 2: Summary information on the weighting data.

	Number of		Weights equal to 0		
Tranche	Households	Persons	hhrf	phrf	phrf_cati
1	1,689	4,126	7	14	2,444
2	1,932	4,947	9	21	3,024
3	978	2,443	1	1	1,466
4	632	1,584	1	4	953
5	309	723	0	0	414
6	303	756	3	5	456
7	288	750	1	3	463
8	298	722	5	11	429
9	265	665	0	0	400
1-9	6,694	16,716	27	59	10,049

The following Table 3 shows the distribution of the different weights (phrf, phrf and phrf_cati) for the numbers of cases reported in Table 2. Weights with a value of 0 were excluded when calculating the corresponding statistics.

7 Deriving your own weighting factors

With the SOEP-CoV data, a large number of analyses are possible on a wide variety of analysis sets. It is not feasible for us to provide weights for each potential analysis set. Nevertheless, the weights provided for the entire SOEP-CoV sample should and must be used for statistical analyses aimed at making broader statements about the entire population, if only to check whether the weights are relevant for the calculation of population statistics (e.g., by simply comparing weighted and unweighted statistics). The SOEP-CoV weights were generated for the entire sample (of the nine SOEP-CoV tranches) of households or individuals who participated in the CATI survey. They therefore represent extrapolation factors for precisely this sample or for a random selection from this sample. This means that for each analysis set that does not meet this requirement, adjustment factors must be calculated so that extrapolations to the population of the SOEP-CoV sample are possible.

Table 3: Distribution of weights by tranches.

Tranche	Minimum	Median	Mean	Maximum	Std. dev.	Sum
			Weight: hhrf			
1	48	3,697	6,279	62,921	7,595	10,562,046
2	8	3,193	5,473	59,144	6,563	10,524,192
3	35	3,931	6,371	62,995	7,741	6,224,776
4	80	3,688	6,537	58,421	8,154	4,125,110
5	131	3,713	6,894	56,348	8,812	2,130,310
6	49	3,521	6,098	38,746	7,227	1,829,350
7	18	3,630	6,745	49,683	8,130	1,935,906
8	20	4,436	7,372	51,321	8,691	2,159,963
9	77	3,617	7,118	65,067	9,037	1,886,347
1-9	8	3,581	6,206	65,067	7,592	41,378,000
			Weight: phrf			
1	43	2,692	4,956	75,018	6,798	20,378,307
2	6	2,449	4,250	77,311	5,579	20,936,930
3	29	2,900	5,165	54,870	6,904	12,613,619
4	74	2,667	5,237	76,366	7,663	8,274,771
5	107	2,916	5,655	57,986	7,883	4,088,392
6	46	2,539	4,722	49,384	6,330	3,545,887
7	17	2,770	5,534	64,162	7,644	4,133,597
8	16	3,157	5,797	60,224	7,634	4,121,793
9	63	2,571	5,293	57,744	7,532	3,519,703
1-9	6	2,648	4,900	77,311	6,727	81,613,000
			Weight: phrf_cati			
1	60	5,674	10,254	92,106	12,478	17,246,453
2	3	5,463	9,574	98,090	11,601	18,410,223
3	49	6,263	10,790	92,106	12,925	10,542,234
4	159	6,131	10,754	80,088	12,865	6,785,625
5	206	6,118	11,460	98,090	13,668	3,540,989
6	48	5,954	10,487	65,970	12,309	3,146,210
7	26	6,186	11,244	92,106	13,966	3,227,107
8	38	7,091	11,622	72,292	13,117	3,405,221
9	127	6,909	12,082	72,292	13,812	3,201,754
1-9	3	5,862	10,425	98,090	12,552	69,505,815

- In order to check in a first step whether the SOEP-CoV weights can be used for a subsample of the SOEP-CoV sample and - if this is not readily possible - to derive appropriate adjustment factors, a selectivity analysis must be performed:
- Here, at minimum all variables to be included in the planned analysis must be included as explanatory variables in a logistic regression model (or a probit or cloglog regression).
- The dependent variable of this selection model is an indicator (coded to 0 and 1) that indicates whether, compared to the entire SOEP-CoV sample, a row of data is part of the analysis set (y = 1) or not (y = 0).
- The selection model thus includes as many data rows as there are observations in SOEP-CoV.
- Now, if none of the analysis variables shows a significant (i.e., $p<0.05$) and at the same time meaningful effect (i.e., $\beta > 0.01$) with respect to the assignment to the analysis set, the subsample under consideration is a random selection from the entire SOEP-CoV sample with respect to the analysis variables. The original SOEP-CoV weights can be used to extrapolate this subsample to the population. It should be noted that weighted data do not add up to the total population size, of course, but only to the subpopulation to which the analysis refers.
- However, if the selectivity analysis reveals distortions in the subsample with regard to the analytical variables (i.e., if there are significant and meaningful effects in the selectivity analysis), a correction of the SOEP-CoV weights is necessary before they can be used for extrapolation purposes. This correction of the SOEP-CoV weights is done by multiplying them by an adjustment factor, which in turn results from the selectivity analysis performed.
- In concrete terms, this means that all analytical variables that turned out to be both significant and meaningful are included in a new selectivity analysis. Analytical variables that were not significant and/or meaningful in the previously calculated selectivity analysis are disregarded (to avoid an unnecessary increase in variance in the adjustment factors to be generated). The dependent variable of the new selectivity analysis is identical to that of the previously calculated one, and the sample size also remains unchanged.
- Based on the estimated (new) selectivity analysis, probabilities must now be estimated (or predicted) for each row of data to belong to the analysis set. This can be done in Stata with the command `predict pr` and in R with the command `predict()` considering the argument `type = "response"`. Now the predicted probabilities for belonging to the original SOEP-CoV sample are fed to the analysis set. The inverse of these probabilities gives the adjustment factor to be multiplied by the SOEP-CoV weights to correct for bias relative to the weighted original SOEP-CoV study sample. In other words, multiplying the SOEP-CoV weights belonging to the analysis set by the inverse predicted probability yields the sought adjusted weight that can be used to calculate population statistics.
- *Note*: In any case, it is advisable to check how well the calculated selection model can discriminate between belonging and not belonging to the analysis set, e.g. by using appropriate boxplots: A boxplot indicates the distribution of (predicted) prob-

abilities for the analysis set and a boxplot shows the (predicted) probabilities for the part of the SOEP-CoV sample that is not part of the analysis set. In general, the first box plot should show a distribution close to 1, the second one a distribution close to 0, and the inter-quartile ranges of both box plots should have as little overlap as possible in their range of values. If this is not the case, the model used does not discriminate well and the addition of further explanatory variables that (better) describe the selection mechanism that generated the analysis set makes sense.

Acknowledgments:
We would like to thank Lennart Dührsen for the initial translation of the German version of this paper and Deborah Anne Bowen for her feedback on the English translation.

References

Auguie, B. (2017). gridextra: Miscellaneous functions for "grid" graphics [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=gridExtra (R package version 2.3)

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*(423), 1013–1020. doi: 10.1080/01621459.1993.10476369

Kroh, M., Siegers, R., & Kühne, S. (2015). Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP). In J. Schupp & C. Wolf (Eds.), *Nonresponse bias: Qualitätssicherung sozialwissenschaftlicher umfragen* (pp. 409–444). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from https://doi.org/10.1007/978-3-658-10459-7_13 doi: 10.1007/978-3-658-10459-7_13

Kühne, S., Kroh, M., Liebig, S., & Zinn, S. (2020, Jun.). The Need for Household Panel Surveys in Times of Crisis: The Case of SOEP-CoV. *Survey Research Methods*, *14*(2), 195-203. Retrieved from https://ojs.ub.uni-konstanz.de/srm/article/view/7748 doi: 10.18148/srm/2020.v14i2.7748

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (p. 225-230). Retrieved from http://www.asasrms.org/Proceedings/papers/1990_034.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/

Robinson, D., & Hayes, A. (2020). broom: Convert statistical analysis objects into tidy tibbles [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=broom (R package version 0.5.6)

Siegers, R., Belcheva, V., & Silbermann, T. (2020). *SOEP-Core v35 Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2018)* (SOEP Survey Papers No. 826). Berlin: DIW/SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.745900.de/diw_ssp0826.pdf

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Xie, Y., Allaire, J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Boca Raton, Florida: Chapman and Hall/CRC. Retrieved from https://bookdown.org/yihui/rmarkdown (ISBN 9781138359338)

Zhu, H. (2019). kableextra: Construct complex table with "kable" and pipe syntax [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=kableExtra (R package version 1.1.0)