1076 2021

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

SOEP-Core – 2020: Sampling, Non-response, and Weighting in Living in Germany – Nationwide Corona Monitoring (RKI-SOEP)

Hans Walter Steinhauer, Sabine Zinn, Rainer Siegers

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, Sabine Zinn, Rainer Siegers. 2021. SOEP-Core – 2020: Sampling, Nonresponse, and Weighting in Living in Germany – Nationwide Corona Monitoring (RKI-SOEP). SOEP Survey Papers 1076: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

SOEP-Core – 2020: Sampling, Nonresponse, and Weighting in Living in Germany–Corona Monitoring (RKI-SOEP)

Hans Walter Steinhauer, Sabine Zinn, Rainer Siegers

Last updated: May 12, 2021

In autumn 2020, the Robert Koch Institute (RKI) joined forces with the Socio-Economic Panel (SOEP) to launch the study "Living in Germany–Corona Monitoring" on the prevalence of current and past SARS-CoV-2 infections in a sample of the adult population in Germany. The survey began shortly thereafter, in October 2020. Participation was voluntary and entailed completion of a questionnaire on COVID-19 infections and symptoms, testing, and health behavior, as well as self-administered tests for current COVID-19 infections (PCR test) and antibodies (DBS test). The results allow for estimation of seroprevalence in the population and identification of socio-economic differences in infection rates and health behavior. Because the study covers all adults in each participating household, it increases the potential for analysis. Selectivity is likely to occur on the household and individual level and may bias results. Here, the use of information from an ongoing panel allowed us to analyze and adjust for possible selectivities due to noncontact, attrition, and refusal at both levels. At the household level, we find characteristics related to the spread of COVID-19 as well as health-related variables to be the main drivers of noncontact, attrition, and refusal. At the individual level, we find age and household composition to be the main drivers of attrition and refusal.

1 Introduction

In December 2019, the virus SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) was found to be causing the disease known as COVID-19. By October 1, 2020, a total of 291,722 individuals had tested positive in Germany (Robert Koch-Institut 2020). The Robert Koch Institute (RKI) tracks and officially reports the numbers of infected people in Germany (see www.rki.de/covid-19). Their numbers are based on tests administered to people who exhibit certain symptoms typical of a COVID-19 infection.[1] At the start of the pandemic, one or more such symptoms were sufficient for individuals to receive a corona test. But as numbers rose, a larger set of symptoms had to be diagnosed to justify testing. This was deemed necessary to meet the increasing demand for testing with the limited capacities available. However, concerns were raised that following the tightening of restrictions on testing, the reported numbers would underestimate the number of people actually infected with COVID-19 (see, e.g., Rendtel et al. (2020)). This meant in particular that individuals who had milder disease progressions, with fewer or weaker symptoms, would no longer be tested. It was determined that in order to obtain a more accurate picture of infection rates, seroepidemiological studies were needed. These studies test for antibodies in the blood that indicate whether the subject was currently or had previously been infected with COVID-19. At that time, most seroepidemiological surveys were conducted either in "hot spots" with small numbers of subjects, or based on non-random samples (see, e.g., Santos-Hövener et al. (2020)). Neither of these not allow for generalization of findings to an entire population. To allow for an unbiased and efficient estimation of the seroprevalence in the general population, a nationwide seroepidemiological survey based on a random sample is necessary (Kritsotakis 2020).

In the RKI-SOEP Nationwide Corona Monitoring study (henceforth referred to as the RKI-SOEP study), we used the SOEP panel study as the basis for our survey. This provided several key advantages: First, it allowed us to invite more than 30,000 adults in over 19,000 households to participate within a very short period of time, enabling us to start the survey just two months after initial discussions. Second, the SOEP consists of random samples covering all of Germany. Third, the panel data available from earlier SOEP survey years provide a rich source of information that can be used to analyze nonresponse processes (noncontact, attrition, and refusal) in the RKI-SOEP study. Besides the panel data from earlier survey years, the SOEP also provides information at different regional levels (such as address, street, municipality, and administrative districts). The RKI provided information on the occurrence of infections at the level of administrative districts. We used all of this information to model nonresponse processes – noncontact, attrition, and refusal – on the household and individual level. The distinction between the levels is necessary for two reasons. First, the SOEP is a household survey, and this has to be accounted for in the modeling process. Second, nonresponse processes are likely to be driven by different characteristics on the two levels. Third, even in participating households, some adult household members may decline to participate. To correct for possible over- or under-coverage, we used margins from the German Microcensus on the household and individual level.

[1] These include, but are not limited to, cough, high temperature or fever, shortness of breath, loss of sense of smell or taste, runny nose or sneezing, sore throat, headache, limb pain, and general feeling of weakness.

In this paper, we present a weighting strategy for a national seroepidemiological study based on an existing panel survey. Section 2 describes the RKI-SOEP study, its aims, sample composition, and survey process. Section 3 outlines the weighting strategy for the study and elaborates on the distinct models of noncontact, attrition, and refusal. Section 4 presents the results, discusses limitations, and highlights potential avenues for future research.

2 Sample Composition of the RKI-SOEP Study

For the RKI-SOEP study, we used the SOEP-Core samples (Goebel et al. 2019), including the migrant samples M1 covering migration from 1995 to 2011 (Kroh et al. 2015) and M2 covering migration from 2009 to 2013 (Kühne and Kroh 2017) as well as the SOEP Innovation Sample (Richter, Schupp, and others 2015).[2] Using these samples as the basis for the study gave us access to 31,675 adults (aged 18 and older) living in 19,574 households in Germany (Hoebel et al. 2021).[3] Figure 1 shows that the SOEP covers all of the 401 districts in Germany. The two maps in the figure show the quartiles for the number of households (left map) and for the number of adults (right map) by districts in Germany. The minimum number of households and adults per district is one. The maximum number of households (adults) per district is 894 (1,281).

Figure 2 shows the composition of the SOEP sample by sex and age. Because of the exclusion of individuals under the age of 18, the figure is truncated at the bottom of the pyramid. The population pyramid for the SOEP is similar to that for the adult population of Germany, except for persons in their early to late 20s. The SOEP has 52.6 percent female respondents, and 47.4 percent male respondents. The average age of the adult population in the SOEP is 50.4 years; for women it is 50.3 years and for men 50.8 years.

The households were divided into four successive tranches (see Table 1) in order not to strain testing capacities. When forming the tranches, we considered the particularities of each federal state as well as infection rates at the district level. When assigning households to the first three tranches, we also took state-specific standards for corona testing into account.[4] In order to provide a sufficient number of observations, we pooled the following federal states: Berlin and Brandenburg, Bremen and Lower Saxony, and Saarland and Rhineland-Palatinate. Moreover, within each stratum of federal states, we also considered another stratification according to the cumulative number of infected people within the district to allow for detailed monitoring of antibody prevalence. To do so, we summed up the number of infected persons per district and correlated this with the district's population density per 100,000 inhabitants, resulting in the cumulative incidence. The numbers of infected persons were provided by the RKI.[5] Information on population density was provided by the Federal Statistical Office.[6] The households within each federal state

[2] In the following, we use "SOEP" without further distinctions to refer to these samples. If necessary, distinctions will be made.

[3] Further exclusion criteria, besides being younger than 18 years of age, include: not providing written consent, not being able to self-administer the tests, and lacking sufficient German skills to understand the materials provided.

[4] See https://www.bundesregierung.de/breg-de/themen/coronavirus/corona-bundeslaender-1745198

[5] Daily updates are available from https://opendata.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0.csv. The figures used were from September 14, 2020.

[6] GENESIS Online (https://www-genesis.destatis.de/genesis/online) Table 12411-0015, accessed on De-

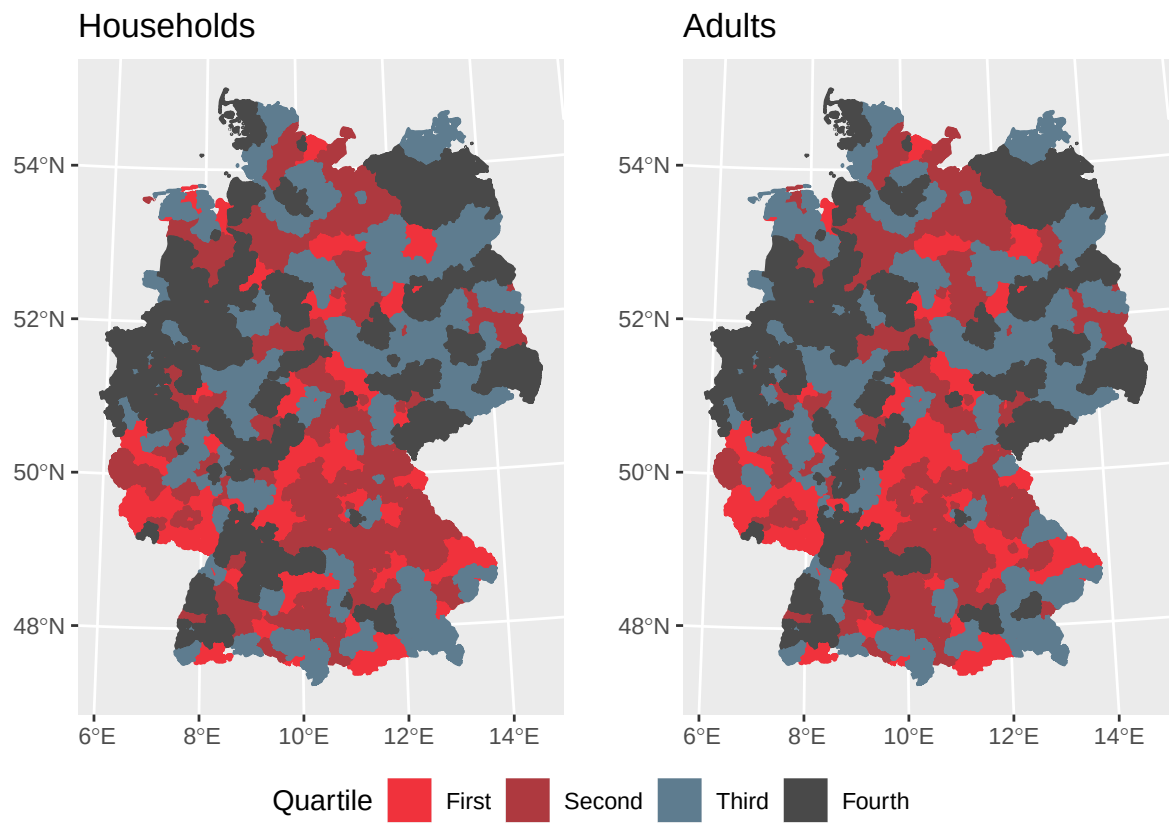

Figure 1: Quartiles for the numbers of households ($Q_{\{1;2;3\}} = \{22; 36; 56\}$) and individuals ($Q_{\{1;2;3\}} = \{35; 59; 93\}$) by district.

stratum were assigned to one of the three infection classes: districts with a low, middle, or high number of infected persons by cumulative incidence. The assignment was based on the corresponding tertiles of the cumulative incidence. Finally, the first tranches were formed in such a way that, for each federal state stratum and infection class, tranche 1 contained 50% of the households and tranches 2 and 3 each contained 25% of the households. The migration samples M1 and M2 formed the last tranche. This was due to the increased administrative effort resulting from our cooperation with the Institute for Employment Research in Nuremberg on these samples.

The households assigned to the four tranches received a letter informing them about the RKI-SOEP study, a letter from the German Ministry of Health urging them to participate, and further information about the study. A few days later, they received the invitation to take part in the study together with the privacy policy, declaration of consent, participation plan, a short questionnaire, and the test kits. The test kits were accompanied by detailed instructions for self-administration and packaging materials for safe return of the samples. Participants were also informed that they would receive another letter with the test results. Individuals who did not respond were sent a reminder two weeks later. As described above, participants received a questionnaire containing questions about their COVID-19 history as well as a polymerase chain reaction (PCR) and a dry blood spot (DBS) test. The PCR test identifies current infection with the SARS-CoV-2 virus, and

cember 31, 2019.

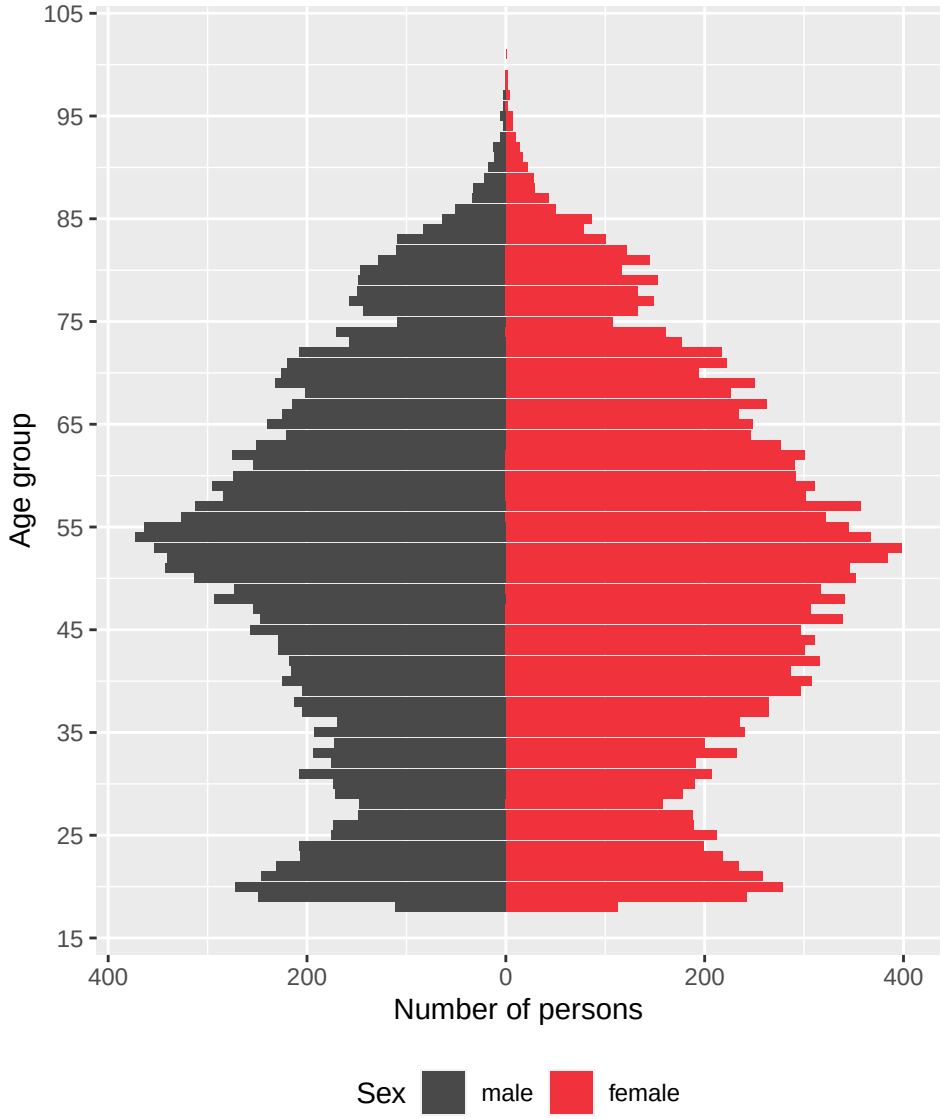

Figure 2: Number of individuals in the SOEP by sex and age.

the DBS test checks for Immunoglobulin G (IgG) antibodies in the blood, indicating a previous infection. The testing procedure is detailed in the study protocol (Hoebel et al. 2021). The survey covered a time period of 15 weeks starting on October 2, 2020, and ending on January 15, 2021 (see 1).

Figure 3 shows the number of adults in households and their allocation to the different statuses. We invited 31,675 adults in 19,574 households to participate in the RKI-SOEP survey, and a total of 15,122 adults in 9,783 households consented to do so. Participation requires a valid consent form: Only when this is returned can the tests and questionnaires be evaluated. Of those invited, 1,504 adults refused to participate. In addition to the usual reasons, there were 12 individuals who stated that they did not believe that COVID-19 is real (coronavirus deniers). Another 23 adults declined to participate because they had been tested already. 190 adults were not able to participate for reasons such as language problems or severe mental or health issues. 378 individuals had moved without providing a new address, and the survey agency was unable to find their new addresses.

Table 1: Field work periods, sample sizes and results by tranche.

	Field work periods		Sample size		
Tranche	Begin	End	Households	Persons	Participants
1	2020-10-02	2020-11-16	9,072	14,535	7,333
2	2020-10-26	2020-11-29	4,499	7,181	3,691
3	2020-11-10	2020-12-15	4,410	7,078	3,501
4	2020-12-14	2021-01-15	1,593	2,881	597
Total	-	-	19,574	31,675	15,122

The remaining 14,481 adults did not return any consent form, tests, or questionnaires.

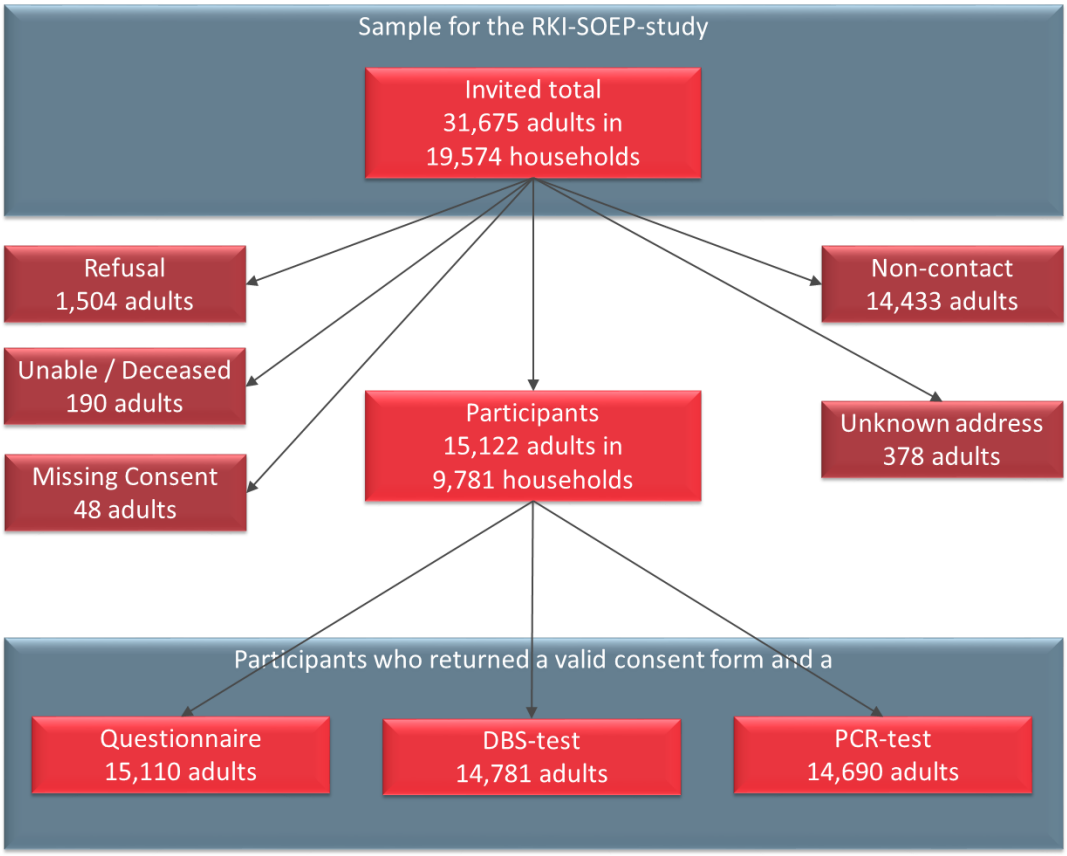

Figure 3: Flowchart of eligibility.

For the 15,122 adults participating, Figure 4 shows the cumulative number of tests (red) and questionnaires returned by date. The figure shows that the majority of tests and questionnaires were returned in October and November 2020. During that time, Germany was in the second wave of COVID-19. Also, in late December, no returned tests and questionnaires were registered because staff at the survey institute were on vacation. Starting in early January 2021, tests and questionnaires that had arrived during the holiday period were processed. We can see that tests and questionnaires were returned in small numbers

up to the beginning of March 2021.

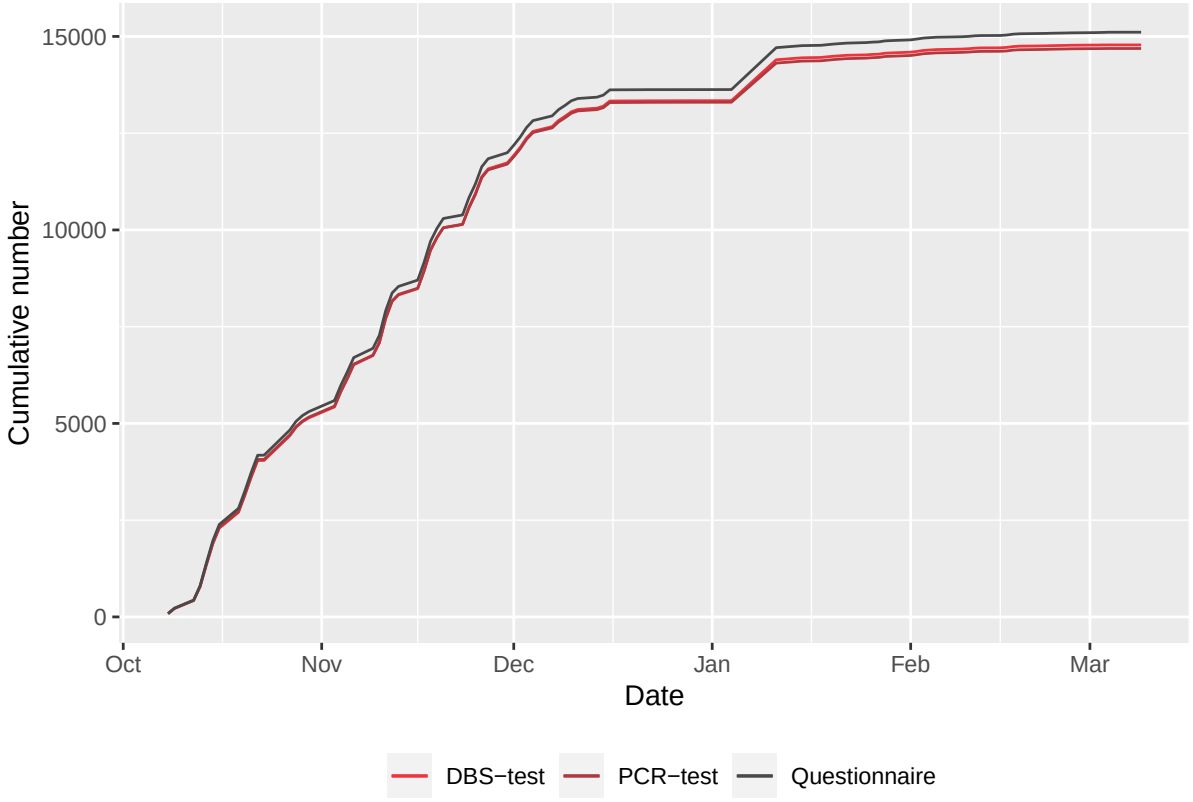

Figure 4: Cumulative number of valid tests and questionnaires received over time.

3 Weighting Strategy for the RKI-SOEP Study

The process of weighting a survey is usually done in three steps (Kalton and Kasprzyk 1986). In the first step, the design weights are derived from the sampling design. In the second step, referred to as sample weighting adjustment, the nonresponse adjustments are carried out to account for potential selectivity. In the last step, referred to as population weighting adjustment, estimates or distributions of the sample are adjusted to conform to those of the population to account for sampling error or under-coverage. The weighting strategy for the RKI-SOEP study resembles this three-step process and is largely similar to the weighting strategy for SOEP-Core detailed in Kroh, Siegers, and Kühne (2015). In order to maximize the number of observations, the studies SOEP-Core and SOEP-IS were integrated. The integration of the two samples was, first, at the household level and, second, at the individual level (see Section 3.3).

Because the SOEP is an ongoing panel survey and the RKI-SOEP study was outside the regular survey plan, we started with the last available observation of each panel household and its corresponding weight at that time. For most of the households, this was from the survey year 2019. This weight was then adjusted, in the sample weighting step, for successive decision processes at the household and individual levels. The decision processes included in the sample weighting adjustments at the household level covered: a)

a household was still in the SOEP panel after the last survey and was invited to participate in the RKI-SOEP study (see Section 3.2.1), b) the household refused to participate in the RKI-SOEP study (see Section 3.2.2), and c) the household participated in the RKI-SOEP study (see Section 3.2.3). Although b) and c) both apply to households that were included in the RKI-SOEP study, it is likely that refusal and agreement to participate are driven by different household characteristics. Thus, we dealt with the two processes separately. We adjusted the resulting weights in the population weighting step using a raking procedure so that sample distributions conformed to known population distributions from the 2019 Microcensus. Since there are fewer households in SOEP-IS than in SOEP-Core, these two samples are weighted separately because they have to "represent" the same number in the population. As a result, the average weight for a household in SOEP-IS is higher. To compensate for this imbalance, the weights for households in SOEP-IS had to be deflated and those for households in SOEP-Core had to be inflated. This was achieved by inflating / deflating them proportionally to the number of households in the federal states. These corrected household weights were then used in another raking step to compute the individual-level weights for all adults living in the participating households. For those adults who participated and returned a valid consent form, a final adjustment was done, correcting for potential selectivity in self-testing. Again, a population weighting adjustment was done, adjusting the participating adults to the population. The raking procedure as well as the distributions used in the population weighting adjustments are detailed in Section 3.3. Figure 5 gives a brief schematic overview of the different steps of the weighting process in the RKI-SOEP study.

3.1 Variables Considered in the Weighting Process

To analyze participation decisions, we chose complimentary log-log (cloglog) regression models to account for the skewed distribution. In modeling the different participation decisions, we regressed over 400 household and individual characteristics on the corresponding decision in a bivariate cloglog model. Most of the characteristics used stemmed from the previous wave of SOEP survey data on demographics, health behavior, education, family, finances, personality, migration, and political attitudes. Further, we used the reported number of people infected with COVID-19 on the district level on the day the DBS test was sent to the respondent.[7] Using the number of inhabitants from official statistics (GENESIS Online Table 12411-0015), we computed the cumulative incidence at the district level. We also used spatial information on the social structure of neighborhoods provided by Microm. This covered, for example, information on the average purchasing power per household in a certain area.

Not all of these variables entered into the corresponding models. The reason is obvious: Of these over 400 variables, only a few were likely to have a significant influence on the participation decisions. Besides that, it was possible that some of them might be highly correlated with each other. Using unnecessary explanatory variables in the model would 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 considered each of the variables in a bivariate model. If the variable turned out to have a significant ($p < 0.05$) influence on the participation decision modeled,

[7] The RKI provides these numbers on a daily basis. They can be downloaded from https://opendata.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0.csv.

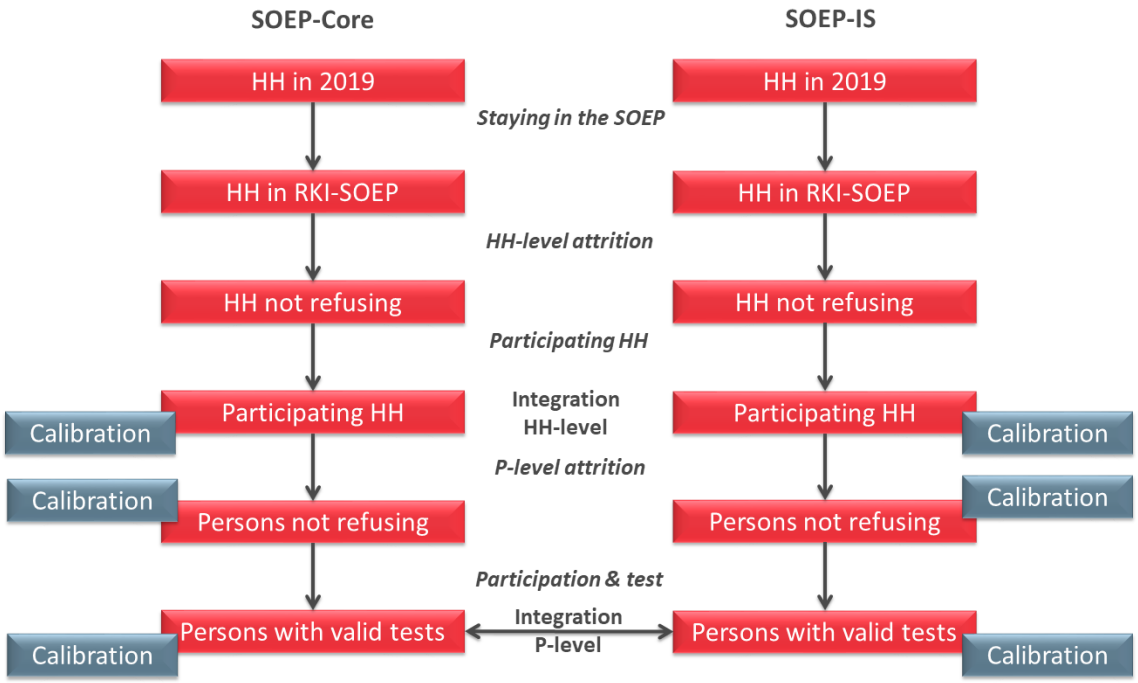

Figure 5: Brief overview on the steps in the weighting process.

it entered the set of significant variables. This set was then analyzed for correlation among each other. If variables showed an absolute correlation greater than 0.95, we chose the variable with the greater estimate from the bivariate model. The remaining set of variables entered the preliminary model. In order to reduce the number of explanatory variables to a minimum, we used a variable selection approach based on the Bayesian information criterion (BIC). This variable selection approach skips and adds variables in a stepwise algorithm, only skipping or keeping them if the model fit improves in terms of the BIC. This three-step procedure yields a final model for the estimation of participation probabilities used to adjust the weights. This procedure was applied in each step of weighting the RKI-SOEP study.

3.2 Sample Weighting Adjustments: Modeling Decision Processes

This section will present the models estimated using the procedure presented above.[8] Results are reported using coefficient plots. You will find the variables of the final model on the y-axis. Parallel to the x-axis, the estimated coefficient (red dot) is displayed together with its 95% confidence interval (red error bars). The dashed vertical line marks 0. Estimated coefficients are displayed in descending order, starting with the smallest estimate in the top left corner and ending with the greatest in the bottom right corner. Coefficients on the left-hand side of the dashed gray line indicate a negative influence on the corresponding decision modeled. Accordingly, coefficients on the right-hand side of the dashed gray line indicate a positive influence on the corresponding decision modeled.[9]

3.2.1 Households remaining in the SOEP after the 2019 wave

The first step of the sample weighting adjustment corrects for households that dropped out of the SOEP in the period between the end of the last wave in 2019 and the beginning of the RKI-SOEP study. In SOEP-IS, households in which a break-up with a partner had occurred in the last survey year were less likely to continue participating (see Figure 6). I n SOEP-Core as well as SOEP-IS, households whose head was 75 years or older were more likely to drop out of the panel. For SOEP-IS, a change of survey mode as well as a change of the interviewer led to a reduced probability to remain in the panel. The same was true of households in which the household head reported low satisfaction with health. If the household was located in a neighborhood with high housing turnover, it was more likely to remain in SOEP-IS. When a related household (e.g. the parent's household) dropped out either temporarily or permanently, this also reduced the other household's (e.g. the child's household) probability to remain in SOEP-Core. Households in which not all household members were surveyed in the last wave, that is, where partial unit nonresponse occurred, were less likely to remain in the panel. The same was true of households with a high

[8] For estimation we use the `glm` function provided by R version 4.0.2 (R Core Team 2021). In preparing the data, analyzing the data, and processing the results, we also used the packages broom (Robinson and Hayes 2020), fastDummies (Kaplan 2020), haven (Wickham and Miller 2020), here (Müller 2020), janitor (Firke 2020), kableExtra (Zhu 2019),
labelled (Larmarange 2020), sf (Pebesma 2018), snowfall (Knaus 2015), survey (Lumley 2020), and tidyverse (Wickham et al. 2019). This paper was created using rmarkdown (Xie, Allaire, and Grolemund 2018).

[9] In general, no estimated coefficient with a confidence interval including zero has a significant influence on the dependent variable. When using the procedure described above, only significant variables remain in the final models for weighting the RKI-SOEP study.

level of item nonresponse. Lower probabilities to remain in the panel were also found for households without Internet access and households with someone who had completed the biography questionnaire or was born abroad. Also, if the household head's satisfaction with life was low, the household was more likely to drop out. Higher staying probabilities were found for households located in big cities and those in which at least one person in the household was single.

Figure 6: Coefficient plot for the model used to correct for household-level attrition between 2019 and the RKI-SOEP study (y = 1 if household remains in the SOEP).

3.2.2 Households remaining in the RKI-SOEP study

The second correction was for households that dropped out of the panel after being asked to participate in the RKI-SOEP study. The coefficients of the model estimating the adjustment factors are shown in Figure 7. Here, several characteristics that impacted attrition between the last wave of the SOEP and the RKI-SOEP study also impacted household-level attrition within the RKI-SOEP study: namely, having a household head aged 75 years or older and not having Internet access in the household. The probability of staying in SOEP-IS was lower for households with at least one unemployed person and for single households; the latter also holds for SOEP-Core. In the case of the RKI-SOEP study, households located in neighborhoods with detached houses were more likely to remain. Households in SOEP-Core that lack access to green space had a lower staying probability in the RKI-SOEP study. Also, households with at least one non-German household member and households in which the household head had no close friends were

more likely to discontinue their participation in the SOEP. Not having made investments in the previous year as well as being located in a neighborhood with mostly single households reduced households' staying probability as well. Households with four or more insurance policies had a higher probability of remaining in the RKI-SOEP study. The same was true of households with four or more household members as well as households with at least one person who was not worried about immigration to Germany.

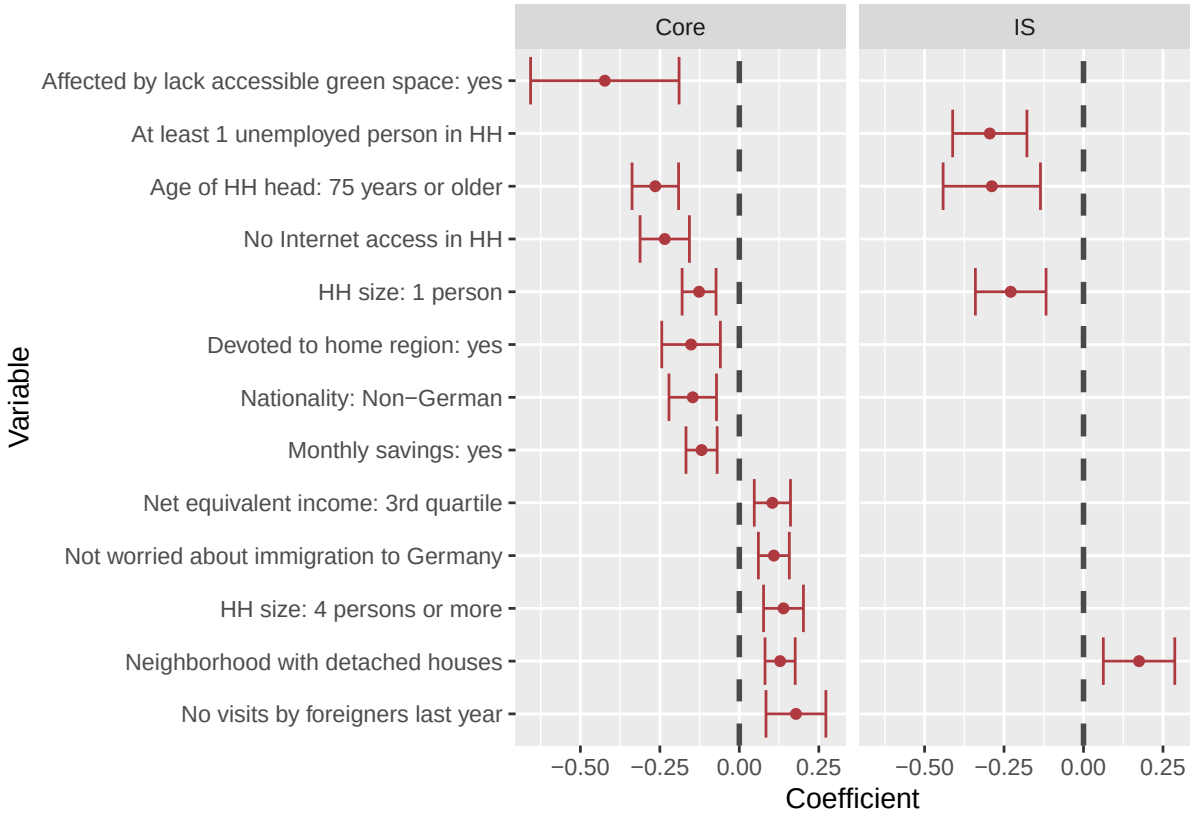

Figure 7: Coefficient plot for the model used to correct for household-level attrition in the RKI-SOEP study (y = 1 if household remains in the RKI-SOEP study).

3.2.3 Household-level participation in the RKI-SOEP study

Because of the large number of characteristics affecting participation decisions on the household level, we highlight only those related to health and the spread of COVID-19. The entire set of variables affecting household-level participation is shown in Figure 8. With respect to the spread of COVID-19, we find that households located in districts with higher incidences (cumulative incidence of 1,000–2,000 cases as well as 2,000 or more) were less likely to participate in the RKI-SOEP study. Also, households located in districts that were affected by a late second wave had a lower probability to participate in RKI-SOEP. Looking at variables related to health and health behavior, we find that households in which at least one person was a smoker had a lower probability to take part. Furthermore, households in which at least one person had chronic health issues or was diagnosed with migraines had a higher probability to participate. Finally, receiving individual health care benefits also increased the likelihood of participating in the RKI-SOEP study.

Figure 8: Coefficient plot for the model used to estimate household-level participation in the RKI-SOEP study (y = 1 if household participates in the RKI-SOEP study).

3.2.4 Individuals remaining in the RKI-SOEP study

After analyzing unit nonresponse in different participation decisions on the household level, we also looked at individual-level participation decisions within households that decided to participate in the RKI-SOEP study. Again, we drew a distinction between the explicit and implicit refusal to participate, as we did on the household level, because we expected the response mechanism behind this to be different. On the individual level, we used the information provided in the last survey wave available. This information mostly refers to the survey year 2019. For SOEP-IS, we found married people who were living with their spouse to be less likely to remain in the study. Adults who reported not being particularly impulsive had a lower propensity to remain in the SOEP. In SOEP-IS, respondents who reported being relatively nervous tended to decline participation more often. For SOEP-Core, we found that respondents with German citizenship were more likely to remain in the SOEP than non-Germans (see Figure 9). Also, younger persons aged 18 to 24 (categories 15-19 and 20-24) were more likely to decline participation in the RKI-SOEP study.

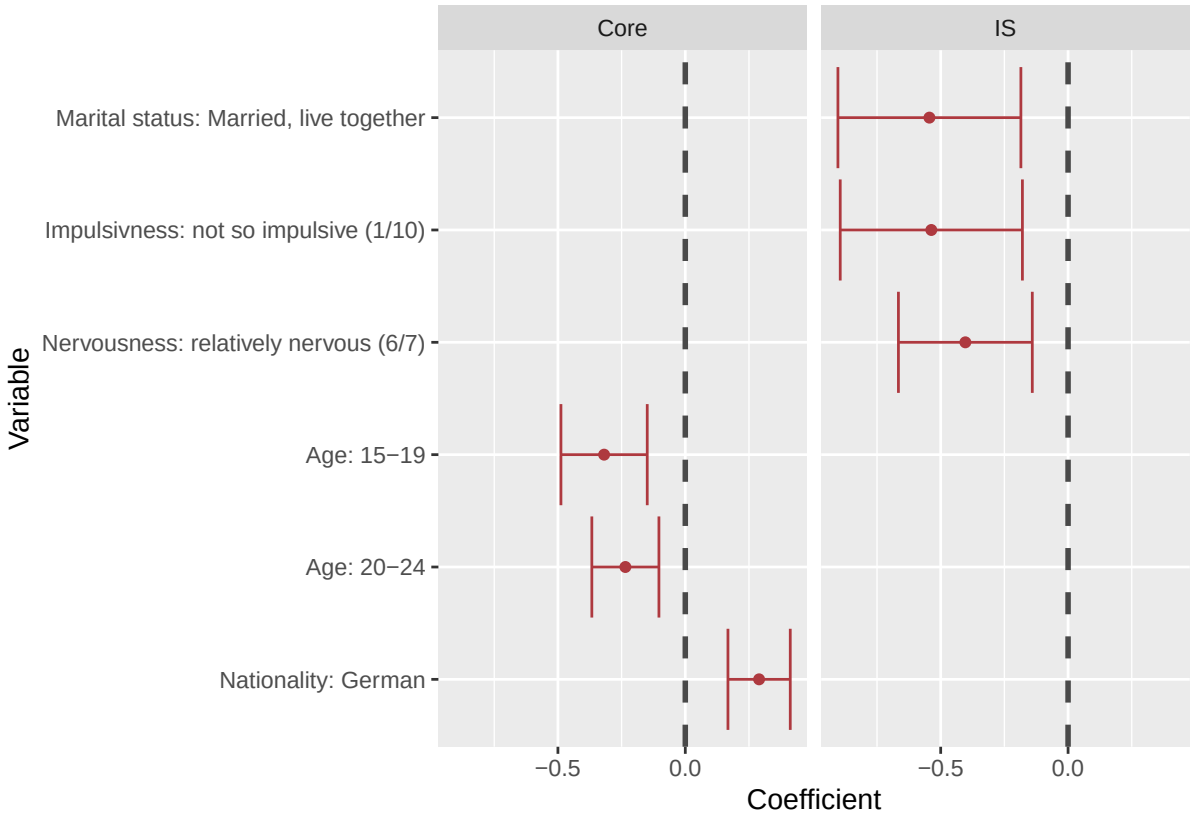

Figure 9: Coefficient plot for the model used to correct for individual-level attrition in the RKI-SOEP study (y = 1 if individual remained in the RKI-SOEP study).

3.2.5 Individual-Level Participation Decisions

Among those adults who did not explicitly refuse to participate, we further analyzed the final participation decision on the individual level. The coefficients of the corresponding model are displayed in Figure 10. In case of the SOEP-IS, we mostly found characteristics

that lowered the participation propensity. The only individual-level characteristics we found to negatively affect the participation propensity were the age group (category 20-24) and not being interested in politics at all. In SOEP-IS, women tended to have a higher probability of participation. The remaining characteristics describe the household context. Here, multi-generation households as well as households with children (regardless of their age) had a lower participation probability. For SOEP-Core, we found young adults aged 18 to 29 (categories 15-19, 20, 24, 25-29) to have a lower participation propensity. The same applies to adults living in a household with a partner and a child aged 16 or older. Also, adults who completed only general elementary education tended to be less likely to participate in the study. A self-rating for inquisitiveness of 4 (neither / nor) on a scale from 1 to 7, and a self-rating for the ability to forgive of 2 (not so easily) as well as being male affected the participation propensity negatively. Adults living in households without children tended to have a higher probability to participate. The same held for persons who were highly satisfied with their family life. Being employed in civil service increased the participation propensity. Moreover, being divorced as well as being German positively influenced the participation decision. Finally, individuals living alone (HH type: 1 adult, no child) tended to have a higher participation propensity.

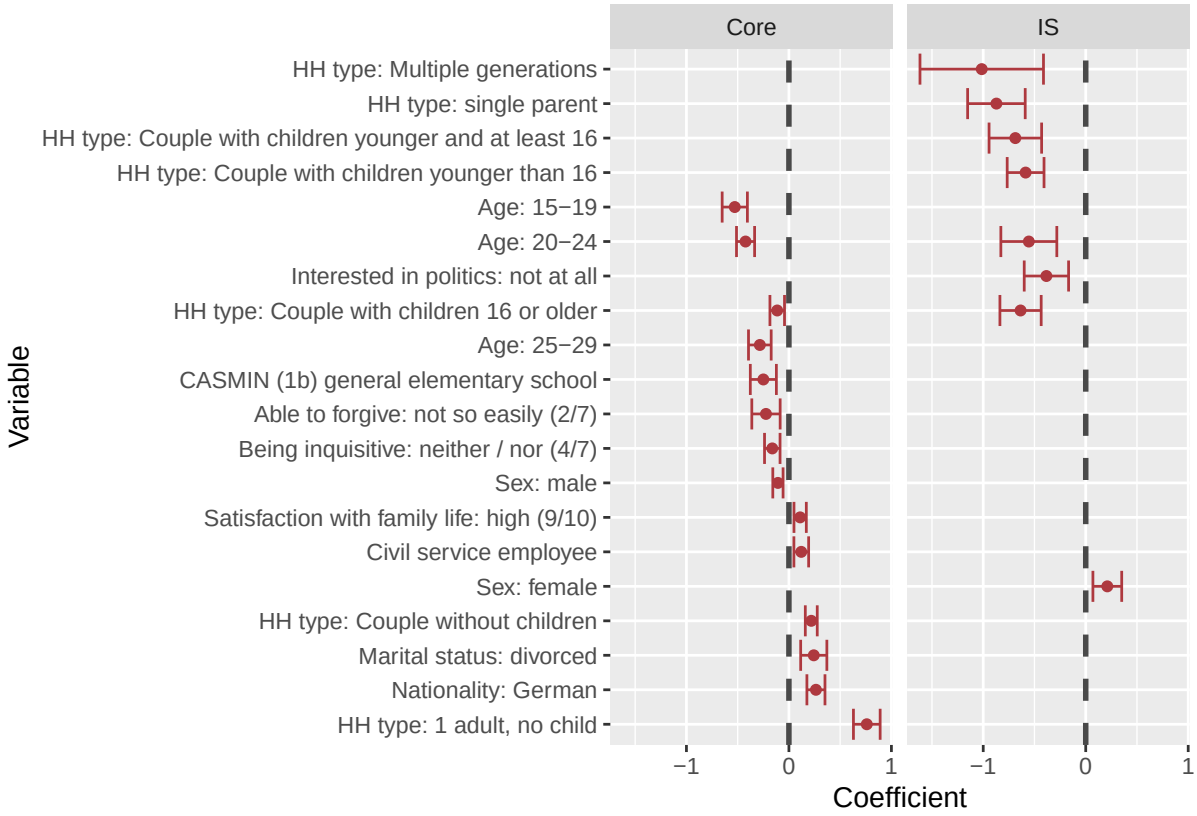

Figure 10: Coefficient plot for the model used to estimate individual-level participation in the RKI-SOEP study (y = 1 if individual participated in the RKI-SOEP study).

3.3 Integration and Population Weighting Adjustments

For the joint analysis of SOEP-Core and SOEP-IS, the weights for the samples needed to be integrated in a way that would allow for projection to the adult population of Germany. This integration step was carried out on the household and individual level (see Figure 5). After adjusting the weights on the household level, they were integrated according to the number of households in the federal state. This household-level weight was then used again to derive individual-level weights for further adjustments. After adjusting the weights on the individual level, they were again integrated according to the number of individuals for the joint distribution by age groups and sex for the adult population.

In order to avoid sampling error and under-coverage, the weights were adjusted to the population. In this last step of the weighting process, we used the raking procedure detailed by Deville, Särndal, and Sautory (1993) to adjust the weights to meet known joint and marginal distributions. Because no marginal distributions for the year 2020 were available from official statistical sources at the time of writing, we used those provided for the year 2019. As the RKI-SOEP study only covers the adult population, the corresponding margins were estimated from the SOEP data from 2019.

Margins used in the population weighting adjustments on the household level covered the number of households per federal state, household size, municipality size classes, and owner-occupied property. These adjustments followed the sample weighting adjustment correcting for participation on the household level. On the individual level, population weighting adjustments were performed on age groups, by sex, and by nationality (German vs. non-German). This adjustment was made for all participating adults living in participating households. In one final step, adjustments were made to clusters describing different courses of the spread of COVID-19 in the second wave, provided by the RKI.

4 Discussion and Summary

One prerequisite for estimating the number or percentage of people infected with SARS-CoV-2 is the use of random samples. Although random samples are frequently used in these kinds of studies, only a few of them provide details on the nonresponse processes involved. The information provided ranges from participation rates and eligibility figures to detailed modeling of nonresponse processes. The amount of unit nonresponse has an impact on the variance of an estimate, and selectivity will impact bias. More attention should therefore be paid to the analysis of, adjustment for, and documentation of selection processes wherever possible. Studies in which samples are derived from population registers administered at regional level often have very little information available on respondents and non-respondents. Nevertheless, Radon et al. (2020) and Warszawski et al. (2021) provide examples of how information on both groups can be enriched and used to provide deeper insights into the response process. We contribute to this line of investigation using previous waves of information from an ongoing panel study and augmenting this with information on regional contexts from official statistics as well as data on the spread of COVID-19 throughout Germany. The RKI-SOEP study is based on a random sample at the national level and thus offers a rich set of information on numerous topics other than COVID-19 from previous SOEP survey waves. This information enables research on how COVID-19 has affected different households and individuals in different ways and will also provide insight into the adverse effects of COVID-19 over the long run.

In this paper, we used the information already available on the household and individual level to take a close look at how unit nonresponse reshapes the sample used for estimating the prevalence of COVID-19 in the adult population of private households in Germany. In doing, so we accounted for different decision processes with respect to timing (before and during the RKI-SOEP study) and hierarchies in the sample (household and individual level).

References

Deville, Jean-Claude, Carl-Erik Särndal, and Olivier Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88 (423): 1013–20. https://doi.org/10.1080/01621459.1993.10476369.

Firke, Sam. 2020. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goebel, Jan, Markus M Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp. 2019. "The German Socio-Economic Panel (SOEP)." *Jahrbücher Für Nationalökonomie Und Statistik* 239 (2): 345–60. https://doi.org/10.1515/jbnst-2018-0022.

Hoebel, Jens, Markus A Busch, Markus M Grabka, Sabine Zinn, Jennifer Allen, Antje Gößwald, Jörg Wernitz, et al. 2021. "Seroepidemiologische Studie Zur Bundesweiten Verbreitung von Sars-Cov-2 in Deutschland: Studienprotokoll von Corona-Monitoring Bundesweit (Rki-Soep-Studie)." *Journal of Health Monitoring* 6 (S1): 2–17. https://doi.org/10.25646/7852.

Kalton, Graham, and Daniel Kasprzyk. 1986. "The Treatment of Missing Survey Data." *Survey Methodology* 12 (1): 1–16.

Kaplan, Jacob. 2020. *FastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. https://CRAN.R-project.org/package=fastDummies.

Knaus, Jochen. 2015. *Snowfall: Easier Cluster Computing (Based on Snow)*. https://CRAN.R-project.org/package=snowfall.

Kritsotakis, Evangelos I. 2020. "On the Importance of Population-Based Serological Surveys of Sars-Cov-2 Without Overlooking Their Inherent Uncertainties." *Public Health in Practice* 1: 100013. https://doi.org/10.1016/j.puhip.2020.100013.

Kroh, Martin, Simon Kühne, Jan Goebel, and Friederike Preu. 2015. "The 2013 IAB-SOEP Migration Sample (M1): Sampling design and weighting adjustment." SOEP Survey Papers 271.

Kroh, Martin, Rainer Siegers, and Simon Kühne. 2015. "Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP)." In *Nonresponse Bias: Qualitätssicherung Sozialwissenschaftlicher Umfragen*, edited by Jürgen Schupp and Christof Wolf, 409–44. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-10459-7_13.

Kühne, Simon, and Martin Kroh. 2017. "The 2015 Iab-Soep Migration Study M2: Sampling Design, Nonresponse, and Weighting Adjustment." SOEP Survey Papers 473.

Larmarange, Joseph. 2020. *Labelled: Manipulating Labelled Data*. https://CRAN.R-project.org/package=labelled.

Lumley, Thomas. 2020. "Survey: Analysis of Complex Survey Samples."

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Radon, Katja, Elmar Saathoff, Michael Pritsch, Jessica Michelle Guggenbühl Noller, Inge Kroidl, Laura Olbrich, Verena Thiel, et al. 2020. "Protocol of a Population-Based Prospective Covid-19 Cohort Study Munich, Germany (Koco19)." *BMC Public Health* 20 (1036): 1–9. https://doi.org/10.1186/s12889-020-09164-9.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rendtel, Ulrich, Reinhard Meister, Jan Goebel, Antje Gößwald, Markus G Grabka, Jens Hoebel, Martin Schlaud, et al. 2020. "Ein interdisziplinäres Studienkonzept zur Dynamik von COVID-19 auf der Basis der prospektiv erhobenen Daten der Kohorten des Sozio-oekonomischen Panels (SOEP)." SOEPpapers on Multidisciplinary Panel Data Research 1094.

Richter, David, Jürgen Schupp, and others. 2015. "The Soep Innovation Sample (Soep Is)." *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift Für Wirtschafts-Und Sozialwissenschaften* 135 (3): 389–400.

Robert Koch-Institut. 2020. "Täglicher Lagebericht Des Rki Zur Coronavirus-Krankheit-2019 (Covid-19)." Situationsbericht 01.10.2020. Robert Koch-Institut. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Okt_2020/2020-10-01-de.pdf.

Robinson, David, and Alex Hayes. 2020. *Broom: Convert Statistical Analysis Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Santos-Hövener, Claudia, Markus A Busch, Carmen Koschollek, Martin Schlaud, Jens Hoebel, Robert Hoffmann, Hendrik Wilking, et al. 2020. "Seroepidemiologische Studie Zur Verbreitung von Sars-Cov-2 in Der Bevölkerung an Besonders Betroffenen Orten in Deutschland–Studienprotokoll von Corona-Monitoring Lokal." *Journal of Health Monitoring* 5 (S5). https://doi.org/10.25646/7052.4.

Warszawski, Josiane, Nathalie Bajos, Muriel Barlet, Xavier de Lamballerie, Delphine Rahib, Nathalie Lydié, S Durrleman, et al. 2021. "A National Mixed-Mode Seroprevalence Random Population-Based Cohort on Sars-Cov-2 Epidemic in France: The Socio-Epidemiological Epicov Study." *medRxiv*. https://doi.org/10.1101/2021.02.24.21252316.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Zhu, Hao. 2019. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Appendix

Providing a valid DBS test

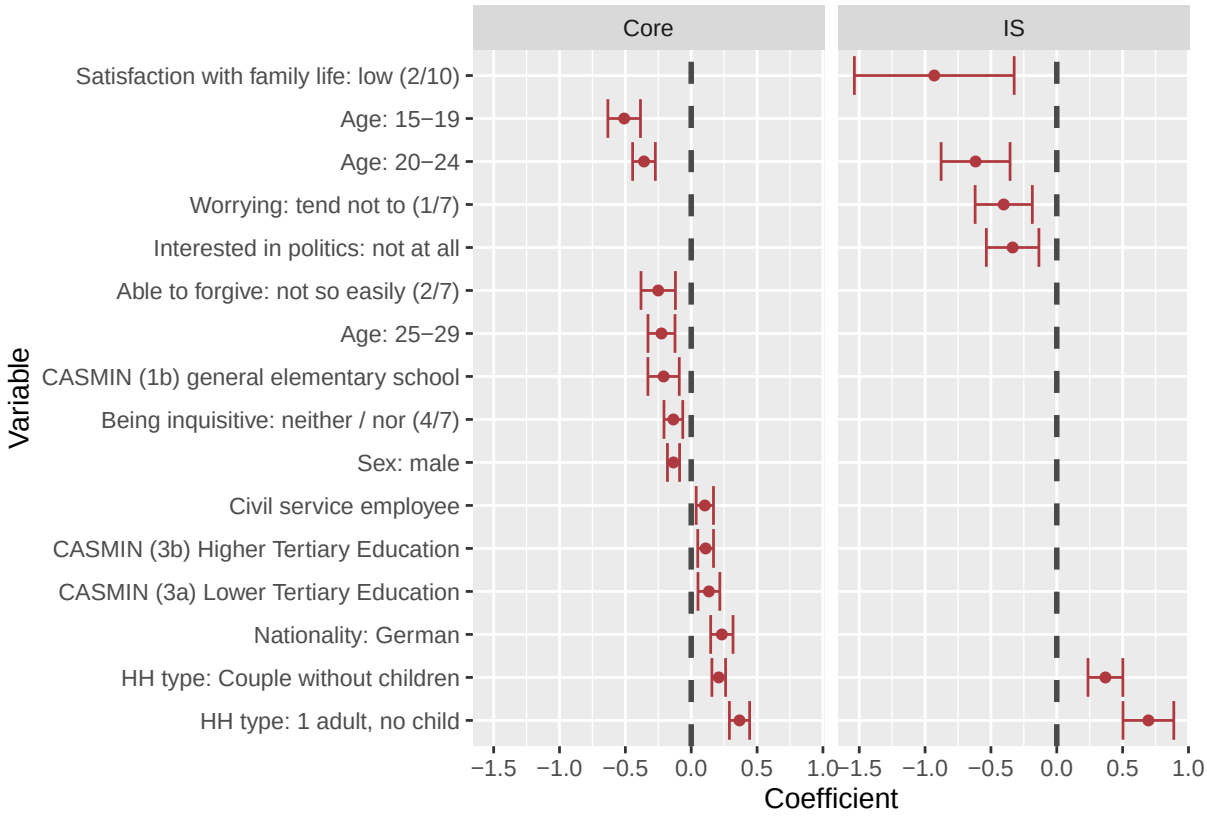

Figure 11: Coefficient plot for the model used to estimate individual-level provision of a valid DBS test in the RKI-SOEP study (y = 1 if individual returned a valid DBS test).