1080 2021

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

SOEP-Core – 2019: Sampling, Nonresponse, and Weighting in the Sample P

Rainer Siegers, Hans Walter Steinhauer, Johannes König

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, Johannes König. 2021. SOEP-Core – 2019: Sampling, Nonresponse, and Weighting in the Sample P. SOEP Survey Papers 1080: 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 – 2019: Sampling, Nonresponse, and Weighting in Sample P

Rainer Siegers, Hans Walter Steinhauer, Johannes König

December 21, 2021

Abstract

This paper provides details on the sampling design, fieldwork results, and nonresponse, as well as population adjustments for the 2019 Sample P of the Socio-Economic Panel (SOEP). Sample P adds the population of entrepreneurs and shareholders with high-value company shareholdings to the SOEP. Assuming that very wealthy individuals almost always invest a significant share of their wealth in company shares, this addition will significantly increase the number of cases and improve analysis potential of the SOEP for analyses focusing on the right tail of the wealth distribution. In order to identify members of our target population we used the global company database ORBIS which was provided by the business information publisher Bureau van Dijk (BvD). From those shareholders with residence in Germany representing the top percentile of the adult population of Germany in terms of their estimated value of shareholdings, we used a multistage stratified sampling design, and were ultimately able to conduct 1,960 interviews with households of this population. In order to correct for disproportionalities resulting from sampling design and attrition, we provide weighting factors that allow the analysis of the new cases both alone and in combination with the other SOEP samples.

1 Introduction

High-net-worth individuals are usually underrepresented in surveys, including the SOEP (see Westermeier & Grabka, 2015). Yet the research interest in this population is increasing with the growing concentration of wealth at the right tail of the distribution. To address the need for data on this relatively small population, the SOEP has created a special sample of high-net-worth individuals in Germany. The sampling of Sample P is intended to close the gap that exists between the super-rich, as documented, for instance, in Manager Magazin, and the right tail of the wealth distribution as previously represented in the SOEP. For more information on the intentions behind drawing this sample, see Schröder, Bartels, Grabka, Kroh, and Siegers (2018).

In order to draw this special population, for which no register data are available in Germany, a new sampling strategy was necessary. By using the ORBIS company database, which is provided by the business information publisher Bureau van Dijk (BvD), we were able to identify and draw a sample of high-net-worth individuals. Sampling from the ORBIS database was carried out under the assumptions that

1. the vast majority of high-net-worth individuals holds at least a significant proportion of their assets in company shares, and
2. the data available on the ownership structures of the companies are sufficient in terms of quality and completeness, and
3. the value of the companies can be estimated from the available company data.

This data documentation provides an overview of the sampling and weighting of Sample P. First, Section 2 describes the target population and sampling frame as well as the preparatory work to identify eligible persons from the ORBIS database. Section 3 details the stratified two-stage cluster sampling procedure that was used and the design weights that can be derived from it. Third, a brief overview of the fieldwork is given in Section 4. The process of nonresponse analysis and creation of weighting factors for the standalone cross-sectional analysis of Sample P is described in Section 5. In Section 6 we give a short description of most important characteristics of the first wave weights of Sample P. Because the population of Sample P may partially overlap with the rest of the SOEP population, this required us to identify households in the other samples that are also included in the target population of Sample P. The integration procedure followed by a (joint) marginal adjustment is described in detail in Section 7.

2 Target Population and Sampling Frame

2.1 Target Population

The target population of Sample P consists of all private households in Germany in which there is at least one person who belongs to the top percentile of the adult population in terms of their estimated wealth from corporate holdings. This population was chosen as a proxy to achieve the initial goal of significantly increasing the number of high-net-worth individuals in the SOEP. Very high wealth is usually accompanied by ownership of business assets, which means that a large proportion of high-net-worth individuals are likely to be included in the target population used here.

When focusing on business ownership, as opposed to total wealth, we have a sampling frame at our disposal: the global company database ORBIS, which is provided by the business information publisher Bureau van Dijk (BvD).

Our objectives in selecting a target population for the Sample P were twofold:

1. We defined a target population to construct a sampling and weighting scheme allowing for an integration of Sample P in the SOEP panel.
2. We sampled from a population that contains a large fraction of wealthy individuals.

Numerous studies show that the very wealthy are heavily represented in the population of company shareholders. In fact, calculations based on the Panel of Household Finances (PHF) suggest that the share of households in Germany holding some of their wealth in business assets ranges from 20% in lower percentiles to above 80% in the top percentile of the wealth distribution (see Schröder et al., 2020). Relying on the Survey of Consumer Finances, Wolff (2017) shows that 94% of those in the top 10% of the net wealth distribution hold at least some business wealth. In addition, rich lists, such as those by Forbes or Manager Magazin, consist almost exclusively of people who own a company or hold considerable shares in at least one business.

Our sampling strategy relied on this empirical pattern. We also needed data from company registers, particularly about ownership and the value of these companies. Company registers allowed the base population to be defined: The population for Sample P consists of people residing in Germany, who, according to company registers, have significant shareholdings in companies worldwide. We focused on the 640,000 shareholders with the highest business wealth, who make up about one percent of the adult population in Germany.[1]

As noted above, our base population of shareholders was sampled from the company database ORBIS.[2] ORBIS contains information on the financial situation and ownership structures of more than 400 million companies[3] worldwide. According to ORBIS, there are 1.58 million shareholders residing in Germany.[4]

2.2 ORBIS as a sampling frame

The ORBIS database contains information on a large proportion of companies with shareholders residing in Germany and the companies' ownership structures. By providing owners and shareholders with unique IDs across the various entries, it offers the possibility of aggregating the information on an ownership basis as well. This provided us a sampling frame that allowed for sampling from this difficult-to-address population of entrepreneurs and shareholders and thus a large portion of the high-net-worth population.

[1] Only individuals who own at least 0.1% of company shares entered into the analysis. This is ORBIS's internal threshold for recording information on an owner. This also guarantees that small shareholders in listed companies are left out of in the target population.

[2] For more information on ORBIS see: https://www.bvdinfo.com/en-gb/our-products/data/international/ORBIS.

[3] 270 million at the time of sampling

[4] Foundations, clubs, and non-commercial partnerships cannot appear in our query, since reporting standards are weaker for these entities and do not force them to publish basic financial information. Hence, they do not appear in the ORBIS data. Additionally, small business (*"Kleingewerbe"*) and liberal professions (lawyers, doctors, etc.) are not included in the company register and thus are not part of our target population.

Bureau van Dijk states that its database contains entries on approximately 400 million companies and entities worldwide and 162 million shareholders.[5] Among them they are also shareholders with smaller company shareholdings of at least 0.1% of all shares.

The ORBIS database is a commercial product that was not created primarily for research purposes. It bundles information, some of which is publicly available, but some of which has been collected and supplemented by the providers. In some cases, it is not entirely transparent which information on which entities was included in the database from which source at which point in time. Therefore, limitations regarding the completeness, timeliness and quality of the data must be accepted. As a result, a great deal of effort had to be put into preparing the data, e.g., by imputing missing values or conducting time-consuming research on outdated address data.

A detailed investigation of the quality of ORBIS data is provided by Bajgar, Berlingieri, Calligaris, Criscuolo, and Timmis (2020). They point out some weaknesses of ORBIS, but conclude that it provides relevant and relatively unbiased information if some requirements are met. In our case, these requirements seemed to be met, since we were particularly interested in larger and thus more valuable companies; since most of the holdings were in companies located in Germany and Europe, where the ORBIS data are quite complete; and since we imputed missing values before analysis.

Since the target population of Sample P includes only those shareholders who represent the top percentile of the adult population in Germany, extensive preparations were necessary. First, nested ownership structures of firms had to be identified in relation to each other. Second, a market price had to be estimated for each firm or shareholding on the basis of the different firm data available. And finally, the value of all of a person's holdings had to be aggregated.

The ORBIS database allowed us to identify and directly sample individuals from our intended target population. A previously conducted pretest (see Schröder et al., 2019) indicated that this new sampling approach was promising and would make it possible to obtain a representative sample of sufficient size with nationwide coverage of this population.

2.3 Estimation of the Business Wealth Proxy

The determination and stratification of the base population (by business wealth) required information on all of the shareholders' business wealth. However, the market capitalization of many of the companies these shareholders owned was not known because only a small fraction of all companies were listed on the stock market and this information was also not included in the companies' books.

Therefore, we had to proxy firm values (and thus the shareholders' business wealth). Our proxy was company turnover, which is available for most of the firms listed in ORBIS, and which correlates strongly with both company value and company profits (Pearson correlation coefficients of company turnover and either company value or company profits are above 0.8, see Schröder et al. (2019)). The advantage of this procedure is that the turnover was used as the central index in the so-called multiplier approach to asses company value (Krolle, Schmitt, & Schwetzler, 2005). If company turnover was missing

[5] https://www.bvdinfo.com/en-gb/our-products/data/international/ORBIS/ORBIS-infographic

in ORBIS, we imputed it by predictive mean matching with five nearest neighbors. We imputed within cells of the companies' industries (using information from NACE – the statistical classification of economic activities in the European Community) and the legal forms. The most recent records on company assets and the number of employees were used as independent variables.

The determination of the value of each shareholder's business wealth requires, in addition to the firm value, the individual shareholdings (in percent) across companies worldwide and also the ownership relationships between companies. Based on ownership percentages, we attributed company wealth of subsidiaries to parent companies until we reached the final parents. Then, we attributed the value of shareholdings to the shareholders based on the ownership percentages across all firms.

We did not need an exactly estimated value of business wealth according to ORBIS. It was sufficient to create an ordinal measure (based on the company turnovers) to sort shareholders in ascending order of their business wealth. We obtained business wealth directly from Sample P participants later as part of the wealth module in the survey. In Schröder et al. (2019) and Schröder et al. (2020) we showed that rankings based on observed company values and turnovers are very close. The most straightforward check is to verify whether the Top 100 of the Manager Magazin rich appear in the top percentiles of the distribution of monetized shareholdings. The average percentile of matched individuals from Manager Magazin was 99. Further, we compared household net wealth of the Sample P respondents across the three business wealth strata they belonged to. Here, the mean for the first stratum was roughly 1 million euros whereas for the third stratum it was about 3.3 million euros.

3 Sampling Design

There was no register from which we could directly sample the target population of interest.For this reason, the ORBIS company database, which contains company shareholders and their private addresses or, in some cases, at least business addresses, served as the sampling frame for Sample P. Our target population was determined as the percentile of the adult population with the highest estimated values of their company shares, in total about 640,000 individuals. From this target population, we planned to sample 31,000 individuals using a two-stage procedure. In the first stage, we drew regions, and in the second, individuals. We then submitted their addresses to the survey institute.

3.1 Sampling of the Primary Sampling Units (PSUs)

In order to make fieldwork more cost-effective and to be able to deploy interviewers more efficiently, it is common practice to draw random samples that are spatially clustered. To make this possible in the case of Sample P, we first created areas in which a similar number of people from the target population live. For this purpose, we determined the number of target persons on the basis of postal code areas. Subsequently, neighboring postal code areas were combined until a PSU was created that contained a sufficient number of target persons. For Sample P we aggregated postal codes so that they contained at least 400 and no more than 800 target persons. Doing so resulted in 1275 primary sampling units (PSUs). Due to limitations in the clustering algorithm, there remained 15 PSUs with fewer than 400 target persons that could not be merged with adjacent clusters without exceeding the maximum number of 800 target persons. Among them there were 5 PSUs with less than 360 target persons, which were therefore not considered in the sampling. Figure 1 shows the distribution of the number of targets per PSU.

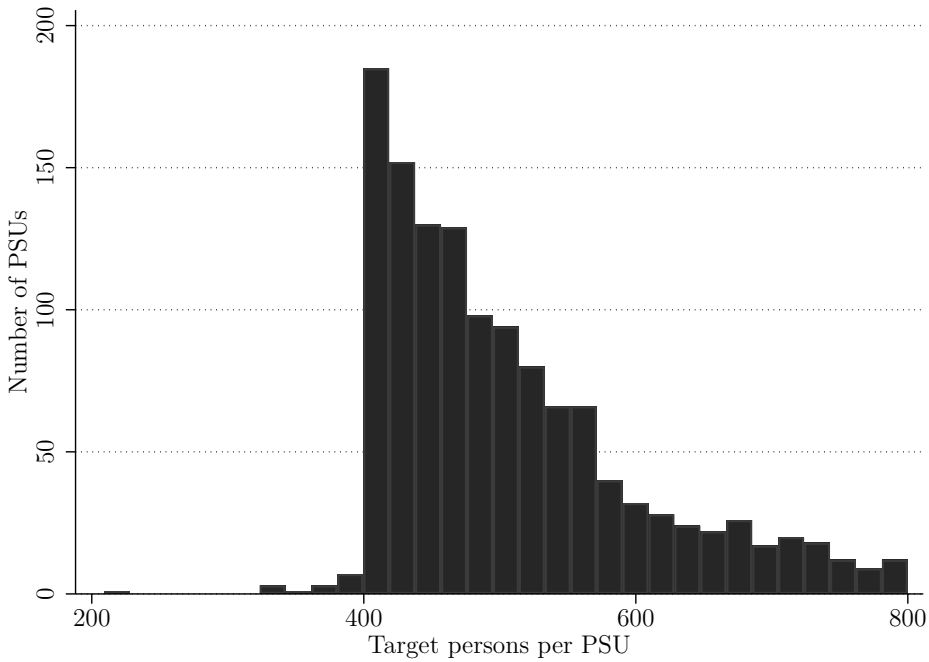

Figure 1: Distribution of the number of target persons per PSU

Figure 2 shows deciles of share of the target population with respect to the total population in each PSU. PSUs with a low proportion of target persons and/or a generally lower population density are larger in size in order to reach the same number of target persons per PSU.

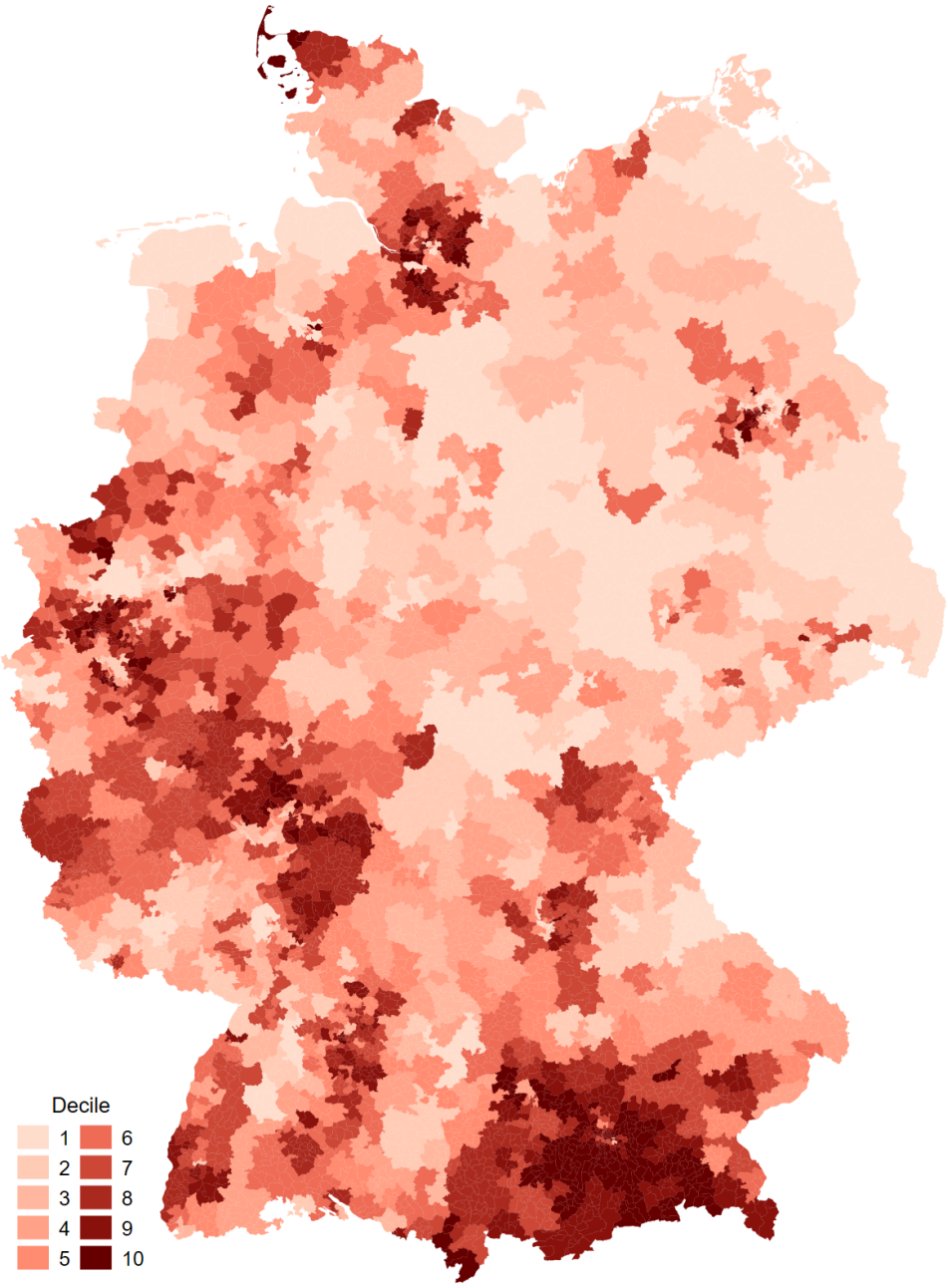

Figure 2: Share of target persons in each PSU (Deciles). Note: We refrain from providing the boundaries of primary sampling units (PSUs). Thus connecting PSUs with same color cannot be distinguished.

In the next step, we allocated the PSUs to strata defined by federal state and population density[6]. We did not categorize by population density in federal states where

[6]We divided The PSUs into two categories, high and low population density (referred to as urban and rural), so that half of the target population fell into each of the two categories. The threshold here was a population density of 188 inhabitants per km^2

one density category was dominant (Hamburg, Bremen, Berlin, Saarland, Brandenburg, Mecklenburg-Western Pomerania, Saxony-Anhalt, and Thuringia).). Then we drew 250 PSUs in stratified sampling. The number of PSUs drawn per stratum is shown in Table 1. PSUs were drawn proportionally to the number of target persons per PSU (proportional-to-size sampling) supplemented by an oversampling of the eastern German states by a factor of 2.

Table 1: Stratification of PSUs

	Stratum	PSUs
$s = 1$	Baden-Wuerttemberg, urban	16
$s = 2$	Baden-Wuerttemberg, rural	14
$s = 3$	Bavaria, urban	16
$s = 4$	Bavaria, rural	28
$s = 5$	Berlin	9
$s = 6$	Brandenburg	11
$s = 7$	Bremen	1
$s = 8$	Hamburg	6
$s = 9$	Hessen, urban	11
$s = 10$	Hessen, rural	8
$s = 11$	Mecklenburg-Western Pomerania	7
$s = 12$	Lower Saxony, urban	3
$s = 13$	Lower Saxony, rural	14
$s = 14$	North Rhine-Westphalia, urban	37
$s = 15$	North Rhine-Westphalia, rural	13
$s = 16$	Rhineland-Palatinate, urban	3
$s = 17$	Rhineland-Palatinate, rural	8
$s = 18$	Saarland	3
$s = 19$	Saxony, urban	6
$s = 20$	Saxony, rural	12
$s = 21$	Saxony-Anhalt	8
$s = 22$	Schleswig-Holstein, urban	2
$s = 23$	Schleswig-Holstein, rural	5
$s = 24$	Thuringia	9

3.2 Sampling of the Secondary Sampling Units (SSUs)

Within each PSU, we drew 124 individuals as secondary sampling units (SSUs). We drew a total of 31,000 target persons using stratified simple random sampling. We stratified the sample by age (younger than 55 years, at least 55 years), gender (female, male), and the level of estimated value of cumulative company shares (tercile membership). We used different sampling fractions between those strata.

On the one hand, a pretest had shown that the higher the estimated value of the company shares, the more difficult it was to survey target individuals. For this reason, we determined that $\frac{4}{7}$ of the gross sample within the target population should come from the highest tercile of estimated values, $\frac{2}{7}$ from the middle tercile, and $\frac{1}{7}$ from the lowest

tercile. The case was similar for the age distribution. The pretest had shown that a lower participation rate was to be expected for younger entrepreneurs and shareholders than for older ones, and that therefore, 60% of the gross sample should come from the younger group (younger than 55 years) of the target population and 40% from the older group (at least 55 years). On the other hand, the aim was to have enough cases to be able to analyze female entrepreneurs specifically. We therefore determined that 25% of the gross sample should be female.

The combination of the different characteristics resulted in different factors for adjusting the sampling probabilities for 12 different strata, as shown in Table 2. The correction factor should be understood in such a way that the inclusion probability of a target person from a given stratum is changed by this factor compared to the average inclusion probability of all target persons.

Thus, for example a young man from the bottom tercile of the estimated value of business holdings ($h = 3$) had an approximately 50% reduced probability of being selected into the gross sample. The maximum spread of these factors is between older male entrepreneurs from the lower tercile ($h = 4$) and younger female entrepreneurs from the highest tercile ($h = 9$). Compared to the former, the latter had an increased probability of being drawn by a factor of $\frac{2.689}{0.284}$=9.5.

Table 2: Stratification of SSUs with adjusting factors

	Stratum	adj. factor
$h = 1$	lower tercile of estimated values, female, young	.730
$h = 2$	lower tercile of estimated values, female, old	.414
$h = 3$	lower tercile of estimated values, male, young	.501
$h = 4$	lower tercile of estimated values, male, old	.284
$h = 5$	middle tercile of estimated values, female, young	1.515
$h = 6$	middle tercile of estimated values, female, old	.859
$h = 7$	middle tercile of estimated values, male, young	1.039
$h = 8$	middle tercile of estimated values, male, old	.589
$h = 9$	upper tercile of estimated values, female, young	2.698
$h = 10$	upper tercile of estimated values, female, old	1.795
$h = 11$	upper tercile of estimated values, male, young	2.172
$h = 12$	upper tercile of estimated values, male, old	1.231

We did not authorize the survey institute to interview all of the target persons supplied to them. Rather, in a first tranche, 62 of the 124 target persons from each PSU were selected for interview by simple random sampling. Depending on the success of the fieldwork (see Section 4), additional cases were made available as required for interviews in a second tranche.

Four classes were created for this purpose. In the first class it was not necessary to release further cases for fieldwork. Here, only the cases from the first tranche were contacted. In the second class, 22 additional target persons per PSU were contacted, in the third 31 and in the fourth class, with the most difficult fieldwork, 46 additional cases were contacted. In total, the addresses of 23,259 anchor persons were submitted to the survey institute as a gross sample.

Basically, the chosen design would have been self-weighting with a proportional-to-size sampling of the PSUs on one side and sampling a fixed number of SSUs within the selected PSUs on the other side. However, under the given circumstances, the sampling probabilities varied by region of origin (East / West), gender (female / male), age (younger than 55 years / at least 55 years), estimated value of company shares (terciles), and success of the fieldwork within a PSU (four classes). Combining these characteristics, 96 groups with different sampling probabilities were formed.

Since SOEP is designed as a household sample, design weights are reported at the household level. In the sampling frame used, no information was available on the household composition, so that Sample P was drawn at the individual level. However, the other household members were also interviewed. Consequently, a household with several persons from the target population had a higher probability of being drawn than a household with only one person from the population of interest. For the construction of design weights, it was therefore checked whether any further persons of the household belonged to the target population and to which group of sampling probabilities they would have been assigned. Some more details on the identification of such individuals are given in Section 7.1. The joint inclusion probability of a household is thus the sum of the individual probabilities of the possible target persons in a household reduced by the probability that more than one of these persons were selected at the same time. Design weights are the inverse of this sampling probability at the household level.

Finally, the design weights of the interviewed households have 124 different values. They have a mean of 325 with a standard deviation of 261, a minimum weight of 53 and a maximum weight of 1,742. The highest design weight is thus higher than the lowest by a factor of 33. Such a spread is unusually high and due to the many specifications in sampling.

4 Fieldwork Results and Response Rates

Of the 23,259 anchor persons in Sample P, 531 were eliminated in advance, either because they were "quality-neutral" or because they had moved abroad or died. 21,841 anchor persons could be found of which 1,960 could be interviewed, resulting in a response rate on the household-level of RR2 = 0.086, calculated according to The American Association for Public Opinion Research (2016). Table 3 displays the results of the fieldwork for the 23,259 anchor persons.

The problem of non-contactability is greater in Sample P than in other SOEP samples. This is primarily due to the limited quality of the adresses in the ORBIS database. While some of the adresses were no longer up-to-date, it turned out in the course of processing the population data that many other adresses were not private but company addresses. Kantar, as the survey institute for this sample, made great efforts to find missing addresses or to contact the respondents through the companies. Nevertheless, just under 4% of anchor persons who could not be contacted. This is a relatively high proportion.

The 9% of anchor persons who were willing to participate in the survey is an unusually low proportion and can be explained by the nature of this special population. This is a major reason why Sample P was created to add a significant number of high-wealth households to the SOEP. In order to compensate for selective dropout in statistical analyses, weighting factors (so-called nonresponse adjustment factors) can be used, as described in Section 5.

Table 3: Fieldwork results on the household-level.

	Number	Proportion
Interview		
Partial interview	1,290	5.55
Complete interview	670	2.88
No Participation		
Refusal	10,639	45.74
Not available during entire field phase	6,579	28.29
Households not processed	1,851	7.96
No time, language problems, other reasons	669	2.88
Illness/hospital/physically or mentally unable	143	0.61
Not found		
Household could not be located	887	3.81
Quality neutral drop-out		
Deceased	225	0.97
Moved abroad	99	0.43
Quality neutral drop-out	207	0.89
Total	23,259	100.00

5 Nonresponse Analysis and Cross-Sectional Weighting

According to Brick and Kalton (1996) the computation of weights is usually performed in three steps. In the first step, design weights are calculated as the inverse of the inclusion probability, see Section 3. Second, these design weights are adjusted to correct for unit nonresponse. Kalton and Kasprzyk (1986) refer to this step as sample weighting adjustment. Finally, in a third step, weights are calibrated so that estimates conform to known population parameters, for example, totals or ratios, or to meet specific distributions. This step is referred to by Kalton and Kasprzyk (1986) 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 modelled the participation decision of the households using information on participating and nonparticipating households. For the anchor persons, we have the information on age, gender and estimated value of the company shares from the ORBIS database, which was also used for the sampling. In addition, little information is available on non-participating households. We therefore used additional regional-level information, interviewer observations (mainly on the target persons' home environments), and information from the field.

Information collected by the interviewer includes: problems speaking German, condition of the neighborhood, condition of the building, access problems due to physical barriers (e.g. locked doors, fences), access problems due to intercom system, other access problems, safety of the neighborhood, composition of neighborhood, type of building (with regard to number of households).

Community-level information was obtained from INKAR online (Indikatoren und Karten zur Raum- und Stadtentwicklung; www.inkar.de). INKAR provides information on housing, (un)employment, construction and education, infrastructure, population character-

istics, and other regional indicators. The time reference for the data is 2015. Detailed documentation on the variables in the data is provided by INKAR (2019).

Lower-level regional information used in the nonresponse analysis was provided by Microm (www.microm.de). Microm provides estimated information about the social structure of neighborhoods in Germany on the regional and the local level. The local level refers to different aggregations such as eight-digit postal code areas covering approximately 500 households, street-level, or household cells aggregating a few households.

5.1 Sample Weighting Adjustment

As described above, no valid contact address could be found for 4% of the anchor persons. Of the remaining cases, 9% were successfully interviewed. Both dropout processes were modeled based on the available information for both respondents and nonrespondents using a bivariate regression[7]. It was assumed that dropouts due to untraceable addresses and due to refusal to participate are independent processes. On the basis of the estimated models, a participation probability was predicted for each successfully interviewed household, which is the product of both estimated probabilities. The reciprocal of this probability was used to adjust the design weight for unit nonresponse.

[7] with a cloglog link function

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

Variable	estimate (std. error)	Variable	estimate (std. error)
Inkar			
High share of 25 to under 30 year olds	0.127*** (0.038)	High probability of singles in the area	-0.064* (0.031)
Low share of unemployed aged 55 and older	-0.082** (0.031)	High probability of advertising refusers in the area	-0.081* (0.034)
Near an international airport	-0.086** (0.028)	House not in exclusive residential environment	-0.164*** (0.026)
High share of people receiving financial support at the age of 65 and older	-0.093** (0.035)		
		Field Information	
		Anchor person born 1975 or later	-0.181*** (0.025)
Microm		House in very good, upscale condition	0.116*** (0.027)
High probability of families with children in the area	0.102** (0.035)		
Low fluctuation house	0.097** (0.030)	Rather run-down house, makes neglected impression	-0.327*** (0.049)
High fluctuation area	-0.103** (0.032)	Rather run-down residential complex, makes neglected impression	-0.130* (0.053)
High probability of less loyal customers	0.095** (0.031)	Residential building with three or more apartments	-0.209*** (0.029)
High probability of individualists in the area	0.074** (0.028)	Unpleasant feeling in the residential street and the residential complex itself	-0.276*** (0.066)
Single-family house	0.071** (0.026)	(intercept)	1.315*** (0.031)
N			22728

Notes: Dependent variable: Address of the household traceable (1 = yes, 0 = no). Significance indicated by $***\equiv p<0.001$, $**\equiv p<0.01$, and $*\equiv p<0.05$. The model was estimated using the command cloglog in Stata. Sources of the variables used in the model are underlined.

We tried to find the strongest predictors of response for each of the models by iterating through all variables included in ORBIS information, interviewer observations, INKAR, and Microm, and selecting those that significantly influenced either the locatability or the decision to participate. In a second step, we omitted those variables from the set of significant variables with an absolute value of correlation among each other of greater than or equal 0.95. Finally, the remaining variables entered a preparatory nonresponse model. To obtain the final model, we ran variable selection in both directions using the BIC as a selection criterion. This yields a more parsimonious model. The models finally estimating the locatability and response propensities used to derive weighting adjustments are presented in Table 4 and Table 5.

The model for estimating locatability propensities contained variables from all three sources—Inkar, Microm, and field information. Factors that had an especially positive effect on the locatability of the target person were: if many younger people or families with children lived in their neighborhood, and if the building at the address was in particularly good condition. Locatability was reduced especially by factors indicating run-down buildings at the survey address or a run-down and unsafe neighborhood. In addition, age was found to play an important role in locatability as an individual characteristic of the target person. Younger anchor persons were more difficult to locate than older ones.

The model for estimating response propensities also contained variables from all three sources. With regard to regional information from Inkar and Microm, the picture was not entirely clear. It was striking, especially in comparison to other samples, that factors indicating a rather poor social structure in the environment of the target person seemed to lead to a more positive response behavior and vice versa. This is an unusual finding and may be due to the special population of Sample P, which consists mainly of financially privileged persons. It is possible that the greater contrast with the immediate residential environment played a role here and led to a greater willingness to participate in such a scientific study.

For the field information, we had large positive effects for two regions, Saarland and rural Hesse. However, these may also be underlying interviewer effects, especially for a small state like Saarland, where very few interviewers were deployed. A positive impression of the building and its surroundings had a positive effect on willingness to participate. Belonging to the second tranche or processing by the central office as well as appreciating the increased efforts of the survey institute in the final phase of fieldwork also had a positive effect.

Regarding individual information about target persons, some explanatory variables were significant, suggesting that younger people were less likely to participate, whereas women and target persons from the lowest tercile (of the top percentile) of company share values had greater response probabilities.

Table 5: Model estimating response propensities used to derive weighting adjustments.

Variable	estimate (std. error)	Variable	estimate (std. error)
Inkar			
High share of people who have a pharmacy near them	0.296*** (0.080)	High share of students per 100 inhabitants from 18 to under 25 years of age	-0.311*** (0.083)
High moving rate from the area	0.231*** (0.057)	High share of people receiving financial support at the age of 65 and older	-0.317*** (0.095)
Region with high negative internal migration balance	0.227*** (0.060)	Low share of apartments with 5 or more rooms	-0.344*** (0.087)
		Microm	
Little recreational area per inhabitant	0.204** (0.063)	High share of multi-person households with older people with low income in the area	0.485* (0.191)
Commuters who travel 150 km or more to work	0.165* (0.070)	High share of older couples with low income in the area	0.393* (0.177)
High share of women in local councils	0.152* (0.059)	High probability of young couples in the area	0.346** (0.114)
Low share of unemployed aged 55 and older	0.140* (0.070)	Low probability of couples in the area	0.187*** (0.056)
Little power generated by wind energy per inhabitant	-0.137* (0.064)	Low probability of single seniors in the area	0.147** (0.055)
Near the highway	-0.139* (0.057)	High probability of single seniors in the area	0.122* (0.056)
Small town or rural community	-0.150* (0.074)	Traditionally orientated people as the dominant group in the area	0.356** (0.114)
Low population development	-0.176** (0.066)	High-rise buildings and simple rental apartments as the dominant form of housing	0.256** (0.093)
High employment rate	-0.181** (0.061)	Low share of Audis among passenger vehicles in the area	0.166*** (0.047)
High voter turnout	-0.225*** (0.067)	Low employment rate in the area	0.152* (0.064)
Low share of children under 3 years of age in daycare facilities	-0.250*** (0.063)	Pure residential street or street with some stores	0.116* (0.050)

Continued on next page

Table 5 – *Continued from previous page*

Variable	estimate (std. error)	Variable	estimate (std. error)
House not in exclusive residential environment	0.112* (0.057)	Older female anchor person from higher tercile of business values	0.218* (0.092)
High share of higher-powered cars in the area	-0.183** (0.067)	Pleasant feeling on the street and in the residential complex itself	0.529*** (0.051)
Low share of people with health-conscious and sustainable lifestyles	-0.215*** (0.054)	House in a very good, upscale condition	0.183*** (0.053)
Harmony-oriented middle class as the dominant group in the area	-0.227* (0.106)	Household is part of the second tranche	0.130** (0.048)
Consumption-oriented as the dominant group in the area	-0.285* (0.137)	Processing by the central office of the survey institute	0.111* (0.050)
Field Information			
Region: Saarland	0.733*** (0.202)	Anchor person born between 1970 and 1974	-0.173* (0.069)
Region: Hessen, rural	0.261* (0.123)	Anchor person born 1975 or later	-0.305*** (0.063)
Younger female anchor person from lower tercile of business values	0.421*** (0.124)	The size of the community is between 20 000 and 50 000 inhabitants	-0.174** (0.064)
Older male anchor person from lower tercile of business values	0.285** (0.106)	Purely residential area with predominantly new buildings	-0.265*** (0.051)
Older female anchor person from middle tercile of business values	0.270* (0.120)	(intercept)	-2.730*** (0.091)
N			21841

Notes: Dependent variable: household participated in survey (1 = yes, 0 = no). Significance indicated by $***\equiv p<0.001$, $**\equiv p<0.01$, and $*\equiv p<0.05$. The model was estimated using the command cloglog in Stata.

5.2 Population Weighting Adjustment

In the last step of the weighting process, we used post-stratification and raking to adjust the weights from the previous step to meet known totals as well as joint and marginal distributions. The weights resulting from this step are the basis for cross-sectional and longitudinal weights derived for wave 2 and beyond.

The marginal distributions that we used for this purpose were directly taken from the ORBIS database and thus reflect the distributions of the sampling frame. In detail, we

used the distributions for age, gender, estimated value of business holdings, and region of residence (federal state and population density). Since the ORBIS database does not contain information on household contexts, raking must be done at the individual level. Thus, all individuals identified as belonging to the target population were subjected to post-stratification (see Section 7.3). In the following step, the household weight was defined as the mean value of all eligible persons in a household.

Usually in the SOEP, weights of individuals would be derived on the basis of the household weights in a subsequent step by first giving the household weight to each person in the household, which would then be adjusted in a post-stratification step at the individual level. In some cases, such as in the SOEP samples M1-5, no information on the household context was available for post-stratification, but marginal distributions existed for the entire population. In these cases, as in Sample P as well, raking was performed on the individual level. The resulting weights could be used for individual-level analyses. Household-level weights were derived from this as the mean of the individual-level weights in a household.

In the case of Sample P, margins are available only for entrepreneurs, i.e., for the anchor persons and other persons identified as eligible. However, weights are needed for all members of households in which at least one eligible person lives, that is, also for spouses and partners, children, and so on. The latter are not included in the marginal distributions used. In the absence of further information for this group of individuals, they were assigned the household weight as their individual weight directly, without adjusting it by raking.

6 Characteristics of Weights

The weights for the first wave of Sample P were derived in three steps (design weighting, sample weighting adjustment, and population weighting adjustment). The characteristics for the weights on the household-level resulting from each step are displayed in Table 6. As can be seen, the sample weighting adjustment increases the dispersion of the weights: for example, the coefficient of variation increased from 0.81 to 0.93. The population weighting adjustment, on the other hand, lowered the dispersion a little. The coefficient of variation dropped to 0.90.

Table 6: Characteristics of weights after the steps of the weighting process.

		Quantiles							
Step	Min.	10%	25%	50%	75%	90%	Max.	Mean	SD
DW	5	10	13	22	31	54	165	27	22
SWA	21	91	139	242	426	732	2,571	343	320
PWA	18	88	136	233	398	663	2,548	322	291

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

7 Integration into the SOEP

In order to be able to analyze the special population of Sample P in combination with the other SOEP samples and in comparison with other population groups, the weighting factors of both data sets, Sample P and the rest of the SOEP, had to be combined. In the case of completely disjoint populations, as was the case, for example, with the sample of households from East Germany that was added to the SOEP in 1990, the weighting factors for both individual studies can be used in joint analyses without further adjustment. This is not the case for Sample P because the existing SOEP samples contain or may contain persons who, as entrepreneurs and shareholders, are also eligible for Sample P. Thus, a correction of the weights was necessary for Sample P and the other SOEP samples. This prevents cases from the target population of Sample P from being over-represented in the analysis set in joint, weighted analyses.

In SOEP, this correction consists principally of three steps:

1. Identification of households from overlapping populations,
2. Construction of a suitable integration variable and adjustment of the weighting factors in its categories, and
3. Post-stratification by joint raking.

7.1 Identification of households in the overlapping population of Sample P and other SOEP Samples

Individuals from the other SOEP samples (and, in addition to the anchor persons, other household members from Sample P) who were also eligible for Sample P's target population were identified directly from the ORBIS database on the basis of their characteristics. For privacy reasons, the SOEP team had no direct access to these personal data. Nevertheless, to enable matching, the SOEP team developed an algorithm to scan the address databases of ORBIS and the SOEP respondents to identify identical individuals. This algorithm was then applied by the survey institute, as the holder of the addresses in the sample. The survey institute then provided information on whether a SOEP respondent from the older samples or a non-anchor person from Sample P was included in the ORBIS database.

The matching procedure utilized the following common variables in both data sets: last name, first name, gender, year of birth, and postal code. Assuming that the information about age, place of residence, etc. of all individuals in both data sources was perfect, the overlap would then be determined. However, several data quality issues prevented us from achieving this ideal. For the match to work well, we needed to balance our assumptions to minimize false positive as well as false negative assignments. Information about addresses could not be used for the matching procedure, as in the ORBIS database the address sometimes corresponded to the address of the company and addresses could not be guaranteed to be up-to-date.

To assess the quality of the match, we also linked the ORBIS data to the anchor persons from Sample P (1,960 individuals). For those persons, the matching procedure should have been able to find a perfect match by definition. These cases were therefore important to assess the quality of the procedure.

The matching procedure encompassed the following steps:

1. Both data sets were joined based on last names, that is, all matches based on last names were generated. This generated many multiple matches, which we restricted in subsequent steps. Roughly 98% of the anchor persons in Sample P were matched in this step.
2. Matches were kept if
   a) the gender matched,
   b) the birth year matched within a tolerance of +/-1 year,
   c) first names matched according to Jaro-Winkler score (string similarity measure) of 0.8 or better.
3. Remaining multiple matches were eliminated by keeping only those with the most similar postcode (computed based on a distance measure). After this step, about 94% of the anchor persons in Sample P still had a match.
4. Next, for a match to be successful and consistent with our sampling strategy, we required that potential matching candidates in the SOEP either had some business wealth or had been registered in a company register. This restricted the matches from the SOEP (excluding Sample P) to 8% of the original match population obtained after step 3 in our matching procedure. After this final step, roughly 250 observations from SOEP had been matched and constituted the overlapping population.

Data quality issues that prevented perfect matches were manifold. First among them was the issue of inaccuracy in the first and last names in ORBIS. ORBIS was based on automated processing of name entries into company registers. For example, prefixes and suffixes of names, such as "von" and "zu", could sometimes not be correctly coded into the first or last name variables; sometimes they were missing altogether. While the survey institute could generally process these erroneous entries by hand and locate the person, we could not do this during the matching procedure because the data set was simply too large. Further, names could have changed between the time the ORBIS entry was made and the time of the survey. This occurred mainly following marriage and divorce. First names could also have been incorrectly recorded: when individuals have several first names, ORBIS may just record one or only an abbreviation of them, while the field institute records all of them without abbreviations. This could drive the string similarity score down and lead us to drop a match erroneously. The data quality issues, naturally, also could affect the other variables recorded in ORBIS, like year of birth or gender.

7.2 Construction of a suitable integration variable and adjustment of the weighting factors in its categories

For cases from other SOEP samples that were identified as eligible for the Sample P target population, and for cases deriving from Sample P, the weights had to be reduced for a joint analysis, since additional cases are available for the population that remains unchanged. Usually the SOEP accomplishes this by creating an integration variable (see Kroh et al., 2015) that has categories tailored to the population being integrated. Within each category of this variable, we count what proportion of cases are from the new sample. The

pre-existing stand-alone weights for the new sample and the older SOEP samples are then adjusted in such a way that the weighted numbers of cases subsequently correspond to the same proportion per category. To fulfill the condition, the weights from the overlapping population of both data sets are scalarly reduced.

In the specific case of the integration of Sample P into the SOEP, it must be taken into account that another newly drawn sample, Sample Q[8] (see de Vries, Fischer, Kroh, Kühne, & Richter, 2021), had to be integrated at the same time. Unlike in the past, two samples whose target populations could possibly overlap were integrated into the SOEP in the same wave.

For this reason, the integration variable in the case at hand also includes categories related to Sample Q and therefore also less differentiated categories for Sample P. Table 7 shows the specifications of the integration variable and the correction factors per sample of origin.

Table 7: Correction factor by sample and specification of the integration variable

	Correction factor		
Eligibility for new Samples	Sample A-O	Sample P	Sample Q
1. Eligible for Q only, no partner in household	.471		.529
2. Eligible for Q only, partner in household	.608		.392
3. Eligible for P only	.093	.907	
4. Eligible for both P and Q	.119	.761	.119
5. Neither eligible for either P or Q	1.000		

For Sample P, only categories 3 and 4 are relevant. After integration, 90.7% of the cases (weighted as well as unweighted) in category 3 (households from the target population of P but not of Q) originated from Sample P and 9.3% from Samples A-O. Finally, in category 4 (households from the joint target population of Samples P and Q), 76.1% of cases are from Sample P, 11.9% are from Samples A-O, and 11.9% are from Sample Q.

7.3 Post-stratification by joint raking

Subsequent to the integration step, a further post-stratification step was carried out in which the weights (previously nonresponse-adjusted and if necessary post-stratified and integrated) of all SOEP samples were adjusted with respect to the standard marginal distributions used by SOEP which were taken from the Microcensus 2019 as described in Siegers, Steinhauer, and Dührsen (2021). Using the resulting standard SOEP weighting factors, the Sample P cases can be analyzed jointly and comparatively in combination with all other SOEP cases.

[8]The target population of Sample Q are households in which persons live who identify as lesbian, gay, or bisexual

8 Summary

Sample P adds 1.960 households containing individuals who own high levels of business assets to the SOEP. Assuming that the majority of high-net-worth individuals hold at least part of their wealth in shares, this addition makes an important contribution to considerably expanding the SOEP's analysis potential at the right tail of the wealth and income distribution. Experience has shown that the "rich" are a very difficult-to-reach population in surveys. It therefore seems legitimate to break new ground for sampling this group, even if this means accepting somee limitations regarding the SOEP's usual quality requirements for sampling frames and response rates.

Sampling this population from the commercial business database ORBIS required extraordinary effort in identifying eligible individuals, locating current residences, and contacting the households. Through weighting, we were able to address the unique challenges and to make the new cases usable for analyses in combination with the SOEP. As a result, we have succeeded in adding a valuable sample to the SOEP that covers a population that is extremely difficult to survey.

References

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., & Timmis, J. (2020, May). *Coverage and representativeness of Orbis data* (OECD Science, Technology and Industry Working Papers No. 2020/06). OECD Publishing. Retrieved from https://ideas.repec.org/p/oec/stiaaa/2020-06-en.html doi: 10.1787/c7bdaa03-en

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

de Vries, L., Fischer, M., Kroh, M., Kühne, S., & Richter, D. (2021). *Soep-core - 2019: Design, nonresponse, and weighting in the sample q (queer)* (SOEP Survey Papers No. 940). Berlin: DIW Berlin / SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.818176.de/diw_ssp0940.pdf

INKAR. (2019). *Indikatorenübersicht – Indkatoren Raum- und Zeitbezüge*. Retrieved from https://www.inkar.de/documents/Indikatoren%20Raum-%20und%20Zeitbezuege.pdf

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

Kroh, M., Siegers, R., & Kühne, S. (2015). Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP). In *Nonresponse bias* (pp. 409–444). Springer.

Krolle, S., Schmitt, G., & Schwetzler, B. (2005). *Multiplikatorverfahren in der unternehmensbewertung: Anwendungsbereiche, problemfälle, lösungsalternative*. Schäffer-Poeschel.

Schröder, C., Bartels, C., Göbler, K., Grabka, M. M., König, J., Siegers, R., & Zinn, S. (2020). *Improving the coverage of the top-wealth population in the socio-economic panel (soep)* (SOEPpapers on Multidisciplinary Panel Data Research No. 1114). Berlin. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.806917.de/diw_sp1114.pdf

Schröder, C., Bartels, C., Grabka, M., König, J., Kroh, M., & Siegers, R. (2019, 12). A novel sampling strategy for surveying high net-worth individuals—a pretest application using the socio-economic panel. *Review of Income and Wealth*, *66*. doi: 10.1111/roiw.12452

Schröder, C., Bartels, C., Grabka, M. M., Kroh, M., & Siegers, R. (2018). *A Novel Sampling Strategy for Surveying High-Worth Individuals - An Application Using the Socio-Economic Panel* (SOEPpapers on Multidisciplinary Panel Data Research No. 978). DIW Berlin, The German Socio-Economic Panel (SOEP). Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.596118.de/diw_sp0978.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 Berlin / SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.816375.de/diw_ssp0960.pdf

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

Westermeier, C., & Grabka, M. M. (2015). Große statistische Unsicherheit beim Anteil der Top-Vermögenden in Deutschland. *DIW Wochenbericht*, *82*(7), 123-133. Retrieved from https://ideas.repec.org/a/diw/diwwob/82-7-3.html

Wolff, E. N. (2017, November). *Household Wealth Trends in the United States, 1962 to 2016: Has Middle Class Wealth Recovered?* (NBER Working Papers No. 24085). National Bureau of Economic Research, Inc. Retrieved from https://ideas.repec.org/p/nbr/nberwo/24085.html