Survey Research Methods (2013)
Vol.7, No.2, pp. 91-101
ISSN 1864-3361
http://www.surveymethods.org

On the Impact of Response Patterns on Survey Estimates from Access Panels

Tobias Enderle, Ralf Münnich and Christian Bruch
University of Trier, Germany

Household and individual surveys increasingly gain importance in policy support and other areas. However, the raising number of surveys leads to reduced response rates. One way to overcome the problem of nonparticipation in surveys involving a non-response bias is to use access panels as a sampling frame. Though leading to expected higher response rates, the self-selection process at the recruitment stage urges the need for a bias correction. This can be done directly when extrapolating the estimates to the population of interest or when using response propensity scores. The latter implies a correct model specification on the recruitment stage.
Keywords: volunteer panel; self-selection; propensity weighting; calibration; variance estimation

1 Introduction

The willingness to participate in voluntary household and person surveys has declined in recent years (cf. De Heer 1999, de Leeuw and de Heer 2002 and Curtin et al. 2005). Besides the consequence of a resulting non-response bias, low response rates may additionally lead to an infeasible sample design (i.e., sampling zeroes). Time, costs and effort must be enlarged to achieve the aspired net sample size. However, since we have an increase in scientific and political demand for information from ad hoc surveys, such a proceeding lacks in its flexibility. Establishing access panels is one way to overcome this situation. Especially over the past few years, access panels have become popular in market, opinion and social research not least due to their multi purpose usage. That is, they serve as a sample frame for conducting multiple random samples in an efficient manner. A survey of great relevance in European Official Statistics is EU-SILC, a study about poverty, social exclusion and living conditions. In 2004 the German Official Statistics started establishing an access panel that serves as a sample frame for D-SILC, the German subsample of EU-SILC (cf. Nimmergut and Körner 2003). Other relevant access panels like the DFG (Deutsche Forschungsgesellschaft) funded access panels for the Priority Programme on Survey Methodology (PPSM) in Germany or the LISS panel in the Netherlands (cf. http://www.lissdata.nl/) are intended to carry out multiple surveys, each with its own individual sample design.

Multiple surveys from an access panel have in common that they rely on the same database involving a number of quite useful advantages that entail higher response rates. For instance, an access panel simplifies the field work of the research institute for a couple of reasons. Questionnaires for surveys can be kept quite short because the sociodemographic characteristics are already collected at the recruitment stage. They also lead to a reduced effort when recruiting sample units. Once a contacted person has answered a short recruitment interview, she or he generally does not have to be motivated to continue in the panel and respond at subsequently waves or further surveys (cf. Engel et al. 2004:148). So, for example, Amarov and Rendtel (2013) show that the willingness to continue in the German access panel among respondents who were already recruited is about 70 to 85 per cent. Though this is not least due to an intensive maintenance of the panel what turns out to be a remarkable requirement. Engel et al. (2004) give an overview of further effects such as a higher attachment to the field work institute which makes respondents more willing to participate.

However, there is still a lack of satisfactory basic research on access panels. Besides the response rate itself, also the response pattern (i.e. missing data pattern due to unit non-response) affects the participation. An important issue when examining the usability of access panels is the self-selection process due to the voluntary nature of participation. Since variables that influence the participation have an effect on the distribution of the variables of interest, procedures such as survey weighting for correcting the inherent selection bias have to be applied. One approach is propensity weighting. That involves direct estimation of response probabilities, so-called propensity scores (cf. Rosenbaum and Rubin 1983), by logistic regression modeling of the recruitment stage. Advantages of using response propensity scores on the quality of survey estimates depend on a correct model specification. That is why Schnell et al. (2005:314) contest the power of such an approach. Another common way to handle the self-selection is to extrapolate directly to the population of interest. In doing so, a more sophisticated modeling of response propensity scores can be ignored, especially in case of unobserved variables (e.g. para data that has not been measured) that account for the dropout.

Contact information: Tobias Enderle, Ralf Münnich and Christian Bruch, Economics and Social Statitsics Department, University of Trier, Germany, E-mail: {enderle,muennich,bruch}@uni-trier.de

The present paper will address the estimation of total values in different access panel participation scenarios using five possible estimation strategies including their accuracy assessment. Since an analytical consideration of these complex survey estimates can be conducted only in irrelevant cases, the most appropriate tool to handle the discussed issues are close-to-reality simulation studies. The scenarios of interest can be evaluated regarding the accuracy of possible survey estimates.

Section 2.1 briefly introduces design and recruiting methods of the access panel used for D-SILC. Consequently, point and variance estimators of total values for this survey are introduced in Section 2.2. Since the major aim of the study is to elaborate the usability of response propensity scores when included as survey weights, we will present five carefully selected estimation strategies in Section 2.3. Section 3 justifies the kind of Monte-Carlo simulation we apply in the present paper. We then implement and examine the impact of several participation patterns on the estimation strategies. Sample design and setup are presented in Section 4.1, comparisons and results in 4.2. A concluding discussion with directions for future research is given in Section 5.

2 Access panel based estimates

2.1 The access panel of German Official Statistics

One way to establish an access panel is to use a large and representative sample, a so-called master sample, and try to recruit members from this sample. This was done with the German Microcensus (MC) that serves as a recruitment pool for the "Dauerstichprobe" (DSP), the access panel of German Official Statistics. Due to the rotational design with four representative quarters, the households rotate out of the MC after 4 consecutive years of participation. At their last interview, the households that are willing to participate in further surveys (e.g. D-SILC) will be invited to join the DSP. Due to the voluntary participation, this access panel itself is based on a non-probability selection. Using the German MC as a master-sample for the DSP guarantees a representative sampling frame (i.e., besides negligible coverage and non-response bias). The MC is a probability-based one per cent sample of the German population with mandatory participation as illustrated in Figure 1. Hence, it is a high quality master sample that offers a large variety of variables for the explanation of response behavior and hence furnishes modeling response propensities for the participation in the DSP based on a variety of categorical variables.

Körner et al. (2006) expect the DSP to improve efficiency of conducting multiple surveys. A standardization of methods and procedures yields shorter lead times and a lower effort. The common field work and administration of addresses and attributes in the DSP allows synergy effects and harmonization between the different surveys. Additionally to D-SILC, the DSP also serves as a sample frame for the Statistics on Information and Communication Studies (ICT) and the Survey of Births (SB). With the DSP it is also desired to flexibly comply with short term demands for information by additional ad hoc surveys. Figure 1 illustrates the sampling stages and the usage of the DSP.

The DSP consists only of persons or households with a high willingness to participate. The response rates in recent years are about 10%.[1] Amarov and Rendtel (2013) and Amarov and Rendtel (2011) highlight that the recruitment rates vary considerably amongst federal states due to different modes and field work effects. However, further groups in the population such as entrepreneurs or pensioners seem skeptical about surveys and less willing to participate. Since the DSP differs from the German MC and thus from the German population in a systematic way, the pattern of the self-selection process as well as the rate itself appear to be important when analyzing statistics drawn from the DSP and, hence, from access panel based surveys.

2.2 Estimators of interest within D-SILC

The focus in our study will be on the estimation of the population total τ_y of an outcome variable y where n is the sample size that will be drawn from the finite population of size N. A general estimator for population totals described by Horvitz and Thompson (1952) is the unbiased Horvitz-Thompson (HT) estimator where the sum of the weighted outcome variable over all units in the sample $\mathcal{S}$ is taken:

$$\widehat{\tau}_{y,HT} = \sum_{i\in\mathcal{S}} w_i \cdot y_i = \sum_{i\in\mathcal{S}} \underbrace{\frac{1}{\pi_{i,MC}\cdot\pi_{i,AP}\cdot\pi_{i,PA}\cdot\pi_{i,RS}\cdot\phi_i}}_{w_i} \cdot y_i. \quad (1)$$

In the original work of Horvitz-Thompson, only the design weights $w_i = 1/\pi_i$ were considered which refer to the inclusion probabilities in the sample of the corresponding units. According to the original sample design that is described by Münnich et al. (2005:3), the participation probabilities of household i at D-SILC must be separated into:

- $\pi_{i,MC}$ Microcensus sample probability (1%),
- $\pi_{i,AP}$ participation at the Access Panel (AP),
- $\pi_{i,PA}$ continuation at the AP (i.e. panel attrition),
- $\pi_{i,RS}$ D-SILC sample probability (i.e. stratified random sample) and
- ϕ_i participation at D-SILC (response).

In order to improve the precision of the estimates above, the design-based HT estimator can be extended using auxiliary information (cf. Deville and Särndal 1992) such as demographic variables. The generalized regression estimator (GREG) additionally calibrates the sample to the marginal totals of some auxiliary variables in a linear regression model. A well known expression for the GREG is the HT estimator plus an adjustment term, the so-called g-weights

[1] The reponse rate in the recent paper is defined as the number of households participating in the DSP divided by the total number of households in the MC. Contrarily to volunteer online panels, the sampling frame of the DSP is well-defined. The computation of response rates is not comparable to those of online panels (cf. Baker et al. 2010).

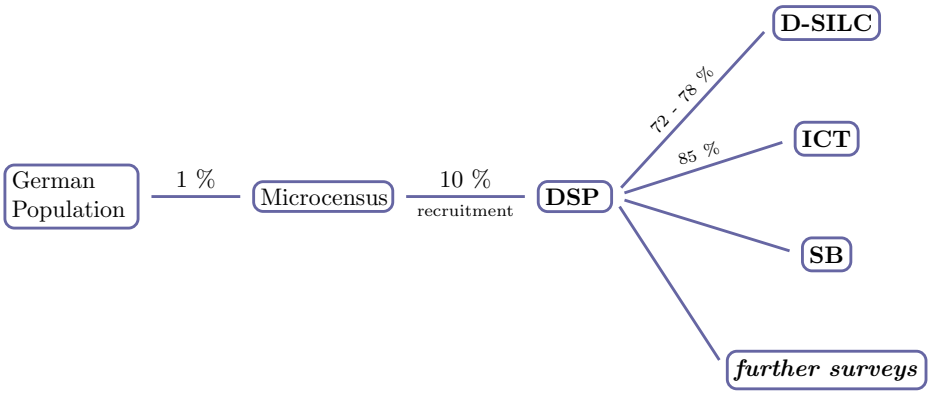

Figure 1. Usage of the access panel of German Official Statistics. Recruitment and response rates (in per cent). Source: Körner et al. (2006:452) and Amarov and Rendtel (2013)

$$\widehat{\tau}_{y,GREG} = \sum_{i \in \mathcal{S}} w_i \cdot y_i \cdot g_i.$$

The formal derivation of these g-weights is given in Särndal et al. (1992:232). Further, the GREG estimator can be seen in view of a more general class of calibration estimators with similar properties, which includes also the post-stratification estimator.

Besides the aim of reducing possibly occurring biases due to self-selection, one is interested in gaining information on the accuracy of the corresponding estimators. An essential basis for accuracy measurement for at least asymptotically unbiased estimators is variance estimation. Therefore, also the variance of the estimated population total, $V(\widehat{\tau}_y)$, has to be considered which in general also has to be estimated from the sample.

Two classes of variance estimation methods can be distinguished, direct and resampling based methods. In any case, the complexity of the sampling design has to be considered carefully in order to produce appropriate variances estimates. Since variance estimation is not the focus of the present study, only the core methods that will be applied later in the simulation study will be shortly presented. For further reading, we recommend Wolter (1985) and Shao and Tu (1995), as well as Münnich (2008) or Bruch et al. (2011).

Since direct methods suffer from the fact that any item of randomness has to be considered separately[2], we focus on presenting resampling techniques. For some of the estimation strategies, no direct variance estimator is available yet. Even so, once one strategy is evaluated as superior, one may wish to develop a specialized (approximate) direct estimator that considers all factors from Equation (1) which, finally, reduces the computational burden by far. Applying direct variance estimation methods that do not account for the complexity of the design or for multiple survey weights may result in very poor estimates (cf. Lee and Valliant 2009). The results of Lee and Valliant (2009) confirm the application of resampling techniques that are easier to implement in the present study.

The idea behind variance estimation using resampling is to draw a sufficient number of subsamples from a given sample and calculating estimates $\widehat{\tau}_y$ in each replication. These resampling estimates give evidence of the distribution of the estimate of interest. This is, the variance of the resampling estimates can be seen as the variance of the corresponding sample estimate.

The resampling methods used for this study only differ in the manner in which they draw the subsamples and in their complexity, speed and efficiency. In case of the delete-1 jackknife, subsamples are built by omitting one element from the original sample in each resample. We used an adopted version, the so called delete-a-group jackknife (DaGJK), that deletes groups of elements instead of single elements. To create a subsample by applying the Monte Carlo Bootstrap (Boot), n elements are drawn with replacement from the original sample. This method is designed for with replacement sampling and needs corrections in without replacement sampling. The method of Balanced Repeated Replication (BRR) assumes two elements respectively two groups of elements in each stratum. For each replication one of these elements respectively groups is drawn by using hadamard matrices to receive the so-called balanced sample which reduces the computational burden in contrast to considering all possible subsamples by far. As proposed in Rao and Shao (1996:344) a repetition of the random grouping of the elements in the stratum can improve the accuracy of variance estimation and is labeled by BRR2. For all methods, either simple programs or packages are available. All three resampling methods were considered in the simulation study below.

2.3 Estimation Strategies

In this paper we focus on the investigation of the statistical behavior of the estimators described above within several participation scenarios (i.e. response patterns as well as rates) using different strategies. To allow a well-arranged

[2]Davison and Sardy (2004:10), for instance, have to apply a complex linearization approach for the calibrated and imputed HT estimator in case of a stratified sampling.

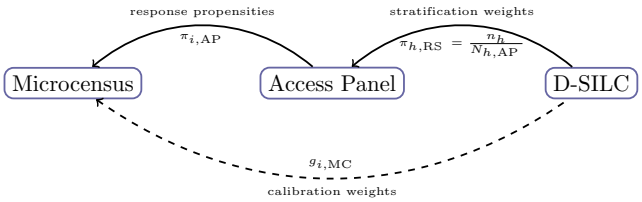

Figure 2. Strategies I and I.cal.MC* use response propensity scores (* additional calibration to the MC = inclusive the dashed line)

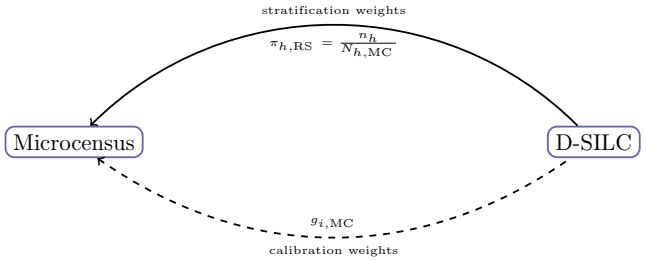

Figure 3. Strategies II and II.cal.MC* extrapolate directly to the Microcensus (* additional calibration to the MC = inclusive the dashed line)

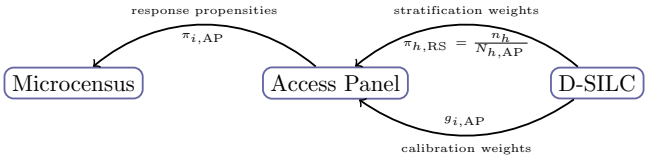

Figure 4. Strategy I.cal.AP uses response propensity scores but calibrates to the AP

comparison we keep the study straight forward and make some restrictions. Instead of replicating the entire sampling stages, we directly make inferences on the MC. Hence, there is no need to take the MC sample probability $\pi_{i,MC}$ into consideration. Furthermore, panel attrition within the AP as well as unit non-response within D-SILC will not be implemented into the simulation environment. That is,

$$\pi_{i,PA} = \phi_i = 1.$$

To obtain the above described total and variance estimates we can roughly classify our procedures into two different types of estimation strategies.

Strategy class I – Using propensity scores Within this class of estimators we estimate response propensity scores for $\pi_{i,AP}$. A common procedure to derive these weights is to estimate the response probability for each household by logistic regression modeling of the recruitment stage given a set of auxiliary variables X. By doing so, the self-selection process can directly be taken into account. Important is the knowledge of the relation between the variables to the self-selection process that causes the non-participation. Additionally, we include the design of D-SILC while adding stratification weights $\pi_{h,RS} = n_h/N_{h,AP}$. Using both inclusion probabilities the estimates can be derived by weighting the sample with their inverse values as described in Equation (1). We call this strategy I that is graphically presented in Figure (2) (omitting the dashed line).

As discussed before, when estimating response propensity scores it is necessary to include the relevant variables. Loosveldt and Sonck (2008) conclude that using basic variables does not make the survey sample more comparable to the population of interest. Alternatively, we compare this strategy to a less demanding naïve method which omits gaining propensities from complex models and, hence, a more sophisticated variable selection. We refer to it as strategy II.

Strategy class II – Direct extrapolation A common approach is to gross up directly the sample to the target population by using design weights (i.e. $N_{h,MC}$). Hence, the stratification weights are $\pi_{h,RS} = n_h/N_{h,MC}$. A disadvantage of this strategy II (see Figure 3) can be seen in case of an extremely non-representative self-selection that will be implemented in the study as worst case scenario. Ignoring the response propensities could lead to highly biased and, hence, inefficient estimates.

Especially in strategy II, one may wish to calibrate the estimates against the population of interest. This may help to compensate for the above described biases and improves the efficiency of the point estimates while applying the GREG estimator by using g-weights. We apply this to both strategies described above. The dashed lines in the upper two graphs of Figure 2 and 3 implicate these additional weights: strategies I.cal.MC and II.cal.MC. Important for the calibration approaches is the correct choice of appropriate auxiliary information. Strategy I.cal.MC closely resembles the official procedure of German Official Statistics (cf. Körner et al. 2006:463).

In applications where micro-level data access may be restricted due to legal constraints, one may also calibrate against the access panel rather than the population of interest, here MC. This strategy is referred to as I.cal.AP which is depicted in Figure 4.

3 Evaluation of Estimation Strategies

Theoretical findings in general help to set-up estimators properly. A comparative evaluation of these estimators, however, may be performed in very specific cases which may lack considerably in practical interest. Hence, we aim at comparing the estimators and strategies in a close-to-reality framework applying a large-scale Monte-Carlo simulation study that allows us to give advice on possible drawbacks in applications.

The simulation study in the present paper tries to generate estimation distributions in moderate sample sizes which allow to understand how to apply a method. These samples sizes were chosen in accordance to the settings in the DSP.

Instead of using real (survey) data where the real population value is unknown, a simulation with synthetically generated data offers a wide range of applicable sample designs and accuracy measurement methods with known population values and distributions. The simulation framework of the study is a synthetic data set that comes close to the general population and is limited to the key variables listed in the access panel database. Variable selection is not a crucial point of the present simulation study. The main emphasis is put on importance of using response propensities. We compare the introduced estimation strategies and evaluate their performances in different access panel scenarios using response propensity models that cover most realistic as well as more severe and extreme self-selection processes as possibly occuring scenarios.

Table 1: Response effects of the households whether they participate or not. The table shows simplified effects used for RPM A (labeled as variable set A)

Variable*	Negative Effect	Positive Effect
FED	Hamburg	Thuringia
	Bremen	Saxony
	Lower Saxony	Brandenburg
		Baden-Württemberg
INC	refusal	high income
HHT	other HH	single parent
AGE	70+	<19
FAM	widowed	divorced
NAT	non-German	German

* An explanation of abbreviations follows in Table 4.

4 Simulation Study

4.1 Simulation Setup

The master-sample used in the simulations for this paper is a synthetic household data set which is close to the real German MC in 2006 (following Münnich et al. 2012). The variables in the data set are on household level (e.g., federal state, household size and type, net income, recipients of retirement pensions per household). The set of variables is limited to the key variables listed in the access panel database. As in the DSP we generated further variables for the head of the household such as age, sex, family status, nationality, etc. The population size N_{MC} is approximately 305,000 and the variable of interest in the study is the number of recipients of retirement pensions (RRP).

At the first sampling stage, the recruitment, we rebuild the self-selection process that causes the dropout of units which are less likely to be in the DSP. A dichotomous vector indicates whether a household responds (R=1) or not (R=0). We do the computation in two steps:

1. First we have to find a response propensity model. Choose a set of variables X and determine which values influence the behavior to respond in which way by means of different coefficients β. The response probabilities for each unit of the master-sample can then be drawn from

$$\hat{\pi}_{\mathrm{AP}} = \frac{e^{\beta' X}}{1 + e^{\beta' X}}.$$

2. Out of these probabilities we then draw the dichotomous response indicator by rejection sampling

$$R = \begin{cases} 1 & \text{if} \quad \hat{\pi}_{\mathrm{AP}} \geq u, \\ 0 & \text{otherwise.} \end{cases}$$

where u is sampled from the uniform distribution over the unit interval.

To investigate the behavior of the estimates, we carry out several scenarios that differ in response rates and response patterns. Additionally to the given response rate of 10%, a worst case with 5% and benchmark case with 30 % are assumed. The different patterns can be done by variation in the first step from above:

RPM A Based on the empirical findings of Amarov and Rendtel (2013) we specified a model that comes close to the logit model estimated in the framework of the DSP research project of Münnich et al. (2006:41). Table 1 roughly illustrates the effects of the implemented logit model.

RPM B In order to elaborate possible misspecifications of a model, we augmented RPM A by a dichotomous variable *willingness to participate at surveys in general* (WTP). The additional variable is correlated to the variable of interest.

RPM C The variable of interest itself additionally to RPM A will be responsible for the behavior to respond. The higher the number of recipients of retirement pensions the less likely the household will be part of the DSP. The dropout can be seen as not missing at random (NMAR).

MCAR Finally, we applied a benchmark scenario where the dropout is generated completely at random (MCAR). Each household has the same probability to participate.

To be realistic the second sampling stage, the sampling of a survey, is nearly conducted as scheduled in the original design of D-SILC that is described in Körner et al. (2006). Whereas the original design assumes 4 stratification variables and up to 1,600 possible strata, the stratified random sampling design given within this study considers solely federal states that are merged to 4 regions and type of the household (HHT) that is merged to 3 categories. Reducing the number to 12 avoids a sophisticated collapsing of strata. Although more strata yield better point estimates at the expense of less accurate variance estimates, the decision for optimal stratification at D-SILC is not the purpose of the present

Table 2: Sample Allocation

Region	Sample Frame	Sample Size
Middle Germany	108,952	1,395
West Germany	69,362	886
East Germany	49,468	671
South Germany	76,869	1,148
$\sum$	304,651	4,100

Table 3: Model specifications for the four response patterns. Each scenario was done with 5, 10 and 30% response rate

	Variables used for	
Scenario	dropout	estimating π_{AP}
MCAR	MCAR	Set A
RPM A	Set A	Set A
RPM B.1	Set A + WTP	Set A
RPM B.2	Set A + WTP	Set A + WTP
RPM C.1	Set A + RRP2	Set A
RPM C.2	Set A + RRP2	Set A + RRP2
RPM C.3	Set A + RRP2	Set A + RRP2

Explanation: Variable set A is chosen according to Table 2, WTP is willingness to participate and RRP2 is a covariate. In RPM C.3 we augmented the calibration approach by a further covariate that is correlated to RRP2.

study.[3] Thus, further effects caused by different stratifications can be excluded. The allocation is done for every region separately according to the MC population. The sample sizes can be drawn from Table 2. Next, the sample sizes in every region are allocated proportionally to the access panel population of the second stratification variable HHT. A brief description of possible changes in the sampling design in order to allow a better stratification is given in Horneffer and Kuchler (2008).

For all scenarios, $R = 10{,}000$ samples each of size $n = 4{,}100$ are drawn. This is the aspired net sample size of D-SILC in 2005 (Horneffer and Kuchler, 2008:652). Age, social status and employment status of the head of the household are used as auxiliary variables for the GREG estimator. Furthermore, we introduced an observable variable, RRP2, that is highly correlated to the variable of interest.

The recruitment stage was modeled as described above in order to obtain different response patterns. Table 4 shows the self-selection process by means of margins. Since, for example, elder people are less willing to participate in surveys, the access panels we generated are not representative for that group (i.e. 70 and older).

4.2 Results of the total estimates

Point estimates. The results of the point estimates are presented in Figure 5. The several boxplots show the distribution of the 10,000 point estimates of each strategy in the different scenarios. The vertical reference line denotes the actual number of recipients of retirement pensions (RRP) in the general population (i.e., about 162,000) and the mean of the 10,000 point estimates is illustrated by diamonds. As long as the dropout is missing completely at random (MCAR) the point estimators of the different strategies are unbiased which can be seen in the first column in Figure 5. This happens to be the case when diamonds match exactly the corresponding line. Moreover, the distributions of the estimates of the different strategies are almost symmetric, since roughly the same number of point estimates are to the left and to the right of the benchmark which can be seen at the boxes, the whiskers and the outliers of the boxplot.

As expected, the calibrated estimation strategies I.cal.MC and II.cal.MC yield the most efficient estimates. Since the boxplots of both of them differ not remarkably, computing response propensity scores can be avoided. When the dropout is MCAR, the response rates play a minor part.

However, these findings do not hold for scenario RPM A (i.e., column two in Figure 5) with more realistic response patterns. Hence, it will be necessary to consider the propensity scores. This holds especially for scenarios with a response rate of 10% or 5%. In scenarios with 30% response all strategies lead to unbiased estimates, except strategy II which shows a negative bias for all response rates. Strategy II.cal.MC overestimates the true value in case of 10% or 5%, especially in the last case significantly. The results of I.cal.MC are acceptable and have only a small positive bias in the case with the lowest response rate. But here, all strategies produce biased estimates which are significantly larger than the bias of I.cal.MC.

RPM B.1 gives a good example of the impact when omitting a variable which is responsible for the dropout in the estimation of the propensity scores. Here, the strategies which used the propensity scores lead to unacceptable results whereas the calibrated strategy II.cal.MC without considering the propensity scores has a reduced bias for all response rates. While considering the respective variable when computing the propensity scores in scenario RPM B.2, these strategies show better results especially with strategy I.cal.MC. Strategy I.cal.AP and strategy I also lead to unbiased estimates in case of high response rates (30% or 10%), while considerable underestimating occurs with low response (5%).

As mentioned above, RPM C consists of scenarios where the variable of interest itself is responsible for the dropout. Obviously, ignoring response propensity scores estimates may become extremely biased which can be seen in RPM C.1. Again, once we are not able to model the self-selection process correctly, strategy II.cal.MC would yield the best results which can be seen in Figure 5. However, using all appropriate variables when carrying out the logistic regression the estimates are improved enormously (scenario RPM C.2). All strategies using propensity weighting (i.e., I, I.cal.MC and I.cal.AP) appear unbiased in scenarios with a high response rate (30% or 10%), whereas strategies II and

[3] A study on the effects on survey estimates when applying different stratifications was carried out in a simulation by Münnich et al. (2005).

Table 4: Descriptive statistics (margins in per cent) on the self-selection process for important variables (in case of 10% response rates). Variables for the head of the household are denoted by an asterix

			DSP (Panel)			
		MC	MCAR	RPM A	RPM B	RPM C
Age (AGE*)	16 to 19	0.4	0.5	0.5	0.5	0.6
from ... to ...	20 to 29	10.2	10.3	9.7	10.6	13.2
	30 to 39	16.4	16.4	17.5	19.6	23.3
	40 to 49	20.5	20.8	20.7	23.5	27.7
	50 to 59	16.5	16.5	18.1	18.3	20.6
	60 to 69	15.9	15.7	16.2	13.6	8.2
	above 70	20.0	19.8	17.3	13.9	6.4
Sex (SEX*)	male	65.20	64.90	67.60	68.30	69.70
	female	34.80	35.10	32.40	31.70	30.30
Marital status (FAM*)	single	25.20	25.30	23.90	25.70	29.70
	married	47.30	46.90	51.30	51.40	50.60
	widowed	10.90	11.20	11.60	12.00	12.30
	divorced	13.80	13.70	10.40	8.10	4.00
	separated	2.90	2.90	2.80	2.80	3.30
Nationality (NAT*)	German	93.40	93.20	94.80	95.60	93.30
	German & further citizenship	0.90	1.00	0.60	0.60	0.90
	Non-German	5.70	5.80	4.50	3.80	5.80
Households with at least	yes	39.8	39.2	37.1	29.7	13.4
one retired person (RRP2)	no	60.2	60.8	62.9	70.3	86.8
Federal State (FED)	Baden-Württemberg	12.30	12.40	15.30	15.40	14.80
	Bavaria	16.20	16.20	14.40	14.60	15.20
	Berlin	5.00	5.10	5.10	4.90	5.40
	Brandenburg	3.30	3.30	4.70	4.70	4.20
	Bremen	0.90	1.00	0.60	0.50	0.60
	Hamburg	2.30	2.30	1.80	2.00	2.10
	Hesse	7.20	7.10	6.60	6.50	6.90
	Mecklenburg-Vorpommern	2.10	2.10	2.00	2.20	2.10
	Lower Saxony	9.40	9.40	6.80	6.70	7.00
	North Rhine-Westphalia	19.10	19.00	17.70	17.60	18.50
	Rhineland-Palatinate	4.90	5.10	6.40	6.50	6.10
	Saarland	1.20	1.20	1.20	1.50	1.20
	Saxony	6.20	6.10	8.90	8.60	7.50
	Saxony-Anhalt	3.20	3.20	2.30	2.30	2.30
	Schleswig-Holstein	3.50	3.60	3.40	3.30	3.40
	Thuringia	3.00	3.00	2.70	2.70	2.70
Households with a	under 900	13.80	14.20	11.40	10.60	11.80
monthly net income (INC), €	900 to 1,300	17.40	17.40	14.10	13.50	12.80
	1,300 to 2,600	42.20	42.20	43.10	43.10	40.90
	2,600 to 3,600	14.90	14.60	18.00	19.00	19.40
	above 3,600	11.70	11.60	13.40	13.80	15.20
Type of the household (HHT)	singles	38.00	38.30	33.40	32.20	32.30
	couples	29.20	28.90	30.80	28.60	24.10
	singles with kids	4.00	4.00	4.50	5.10	5.70
	couples with kids	17.60	17.60	19.70	22.40	26.20
	others	11.20	11.00	11.70	11.70	11.70

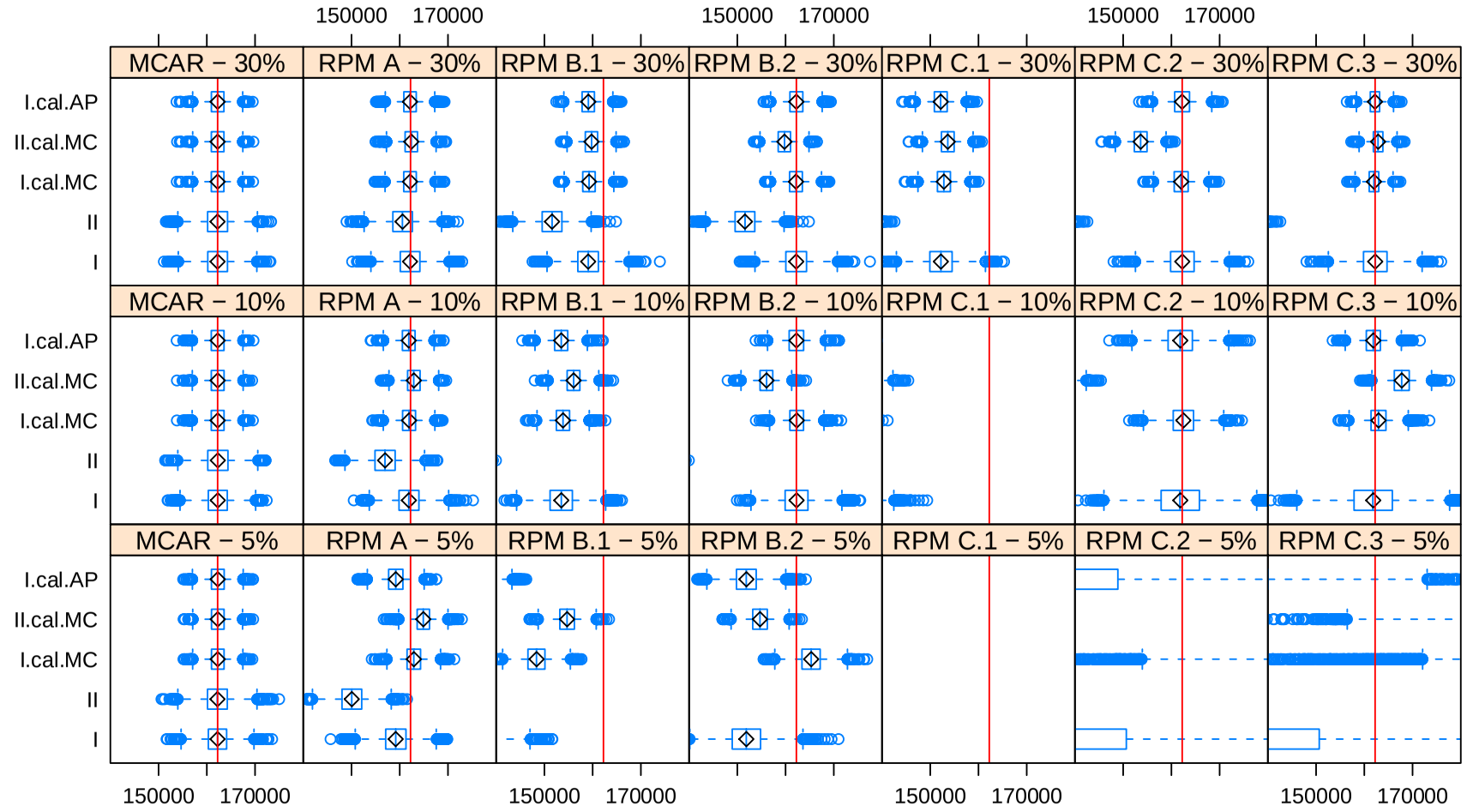

Figure 5. Totals of recipients of retirement pensions. The lower right cells cannot show the boxplots of the estimators due to scaling reasons of the whole graph. Table 3 gives an overview of the model specifications for the different response patterns.

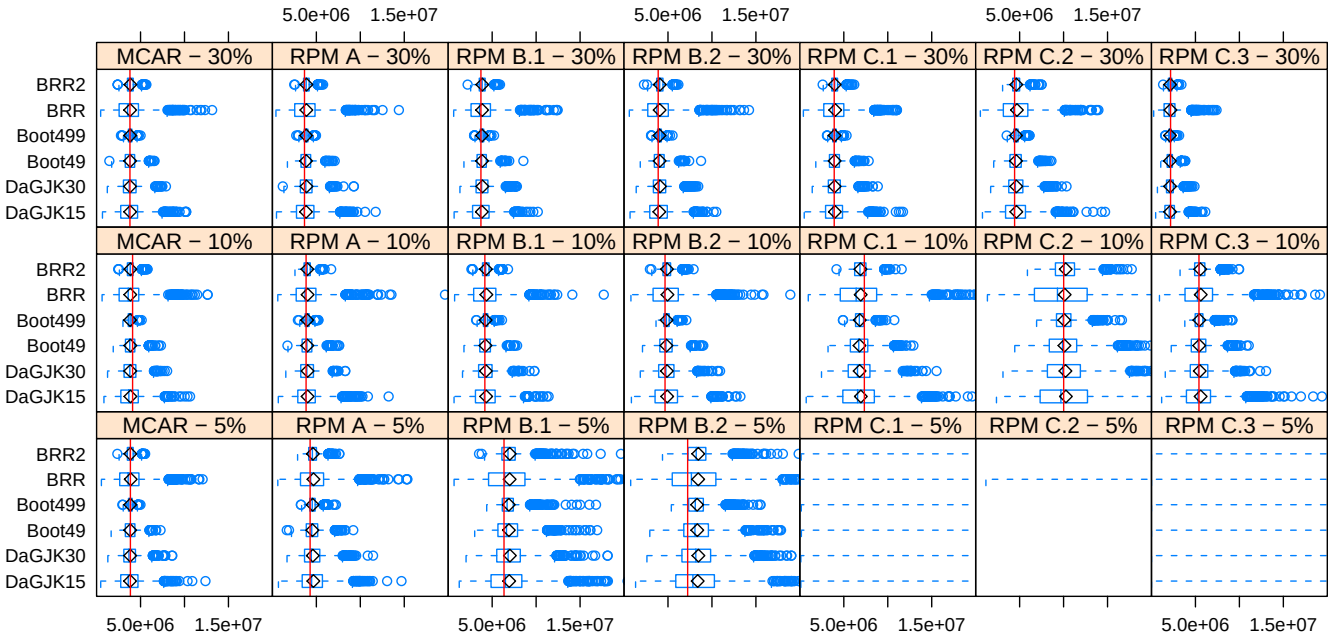

Figure 6. Variance estimates for strategy I.cal.MC. The red line denotes the benchmark (i.e., Monte-Carlo variance derived by the total estimates).

II.cal.MC are biased. For that reason propensity scores have to be considered, especially when the dropout is extremely asymmetric and caused by either the variable of interest or variables that can be used for dropout modeling. Alternatively, when including a variable that is correlated to the variable of interest in the calibration approach, the estimation with strategy II.cal.MC can also be improved, but still remains biased as seen in RPM C.3.

In almost all cases except MCAR, we could observe that very low participation rates cause biases in the estimations that have to be corrected properly.

Variance estimates. The results in the last section show that especially strategy I.cal.MC within the classes of using the propensity scores and strategy II.cal.MC within the classes without consideration of propensity scores are of special interest. For that reason only the variance estimators of these strategies are displayed in this section.

In general the results of variance estimation are quite similar for both strategies as shown in Figures 6 and 7. In case of scenario MCAR and RPM A almost all variance estimators are unbiased. For scenario B.1 and B.2 the variance estimation with resampling methods works well when the response rate is 10 or 30%. When the response rate is 5% the variance is overestimated in these scenarios for both strategies. In RPM C.1, C.2 and C.3 with a low response rate (5%) the variance of variances of all considered methods are extremely large and for that reason all these methods seem to be inappropriate here. For strategy I.cal.MC the variance estimation in RPM C.1 is biased because a variable which is responsible for the dropout is not considered. As a result the variance estimation with resampling methods leads to unbiased estimators in RPM C.2 and C.3 for response rates 10% or 30%. For 10%, the variance estimation for strategy II.cal.MC is unbiased in RPM C.3 but biased in RPM C.1 and C.2. This is because only in RPM C.3 a variable which is highly correlated to the dropout variable RRP2 is included in the calibration process.

Within the resampling methods the bootstrap with 499 replications has the smallest variance in all scenarios. As shown in the Figure 6 when comparing the bootstrap with 499 to 49 replications it is possible to reduce the variance of variances considerably when using enough replications. The ordinary BRR and the delete-a-group jackknife with 15 groups are the most inefficient variance estimators. The delete-a-group jackknife for such a complex survey needs more groups. Thus, when using 30 groups instead of 15 the variance of variances decreases significantly. The results of the BRR can also be improved significantly when repeating the random grouping (as done in BRR2) as suggested by Rao and Shao (1996:344).

The confidence interval coverage rates in Figure 8 further support the above findings. The coverage rates of the totals are derived by using the variance estimates of the three resampling methods delete-a-group jackknife, bootstrap with 499 replications and BRR with repeated grouping (BRR2). In MCAR scenarios the nominal rate is always met but when it comes to other scenarios, strategy I.cal.MC performs better, especially when the dropout is getting more asymmetric. Scenarios with low response rates tend to undercover the true value.

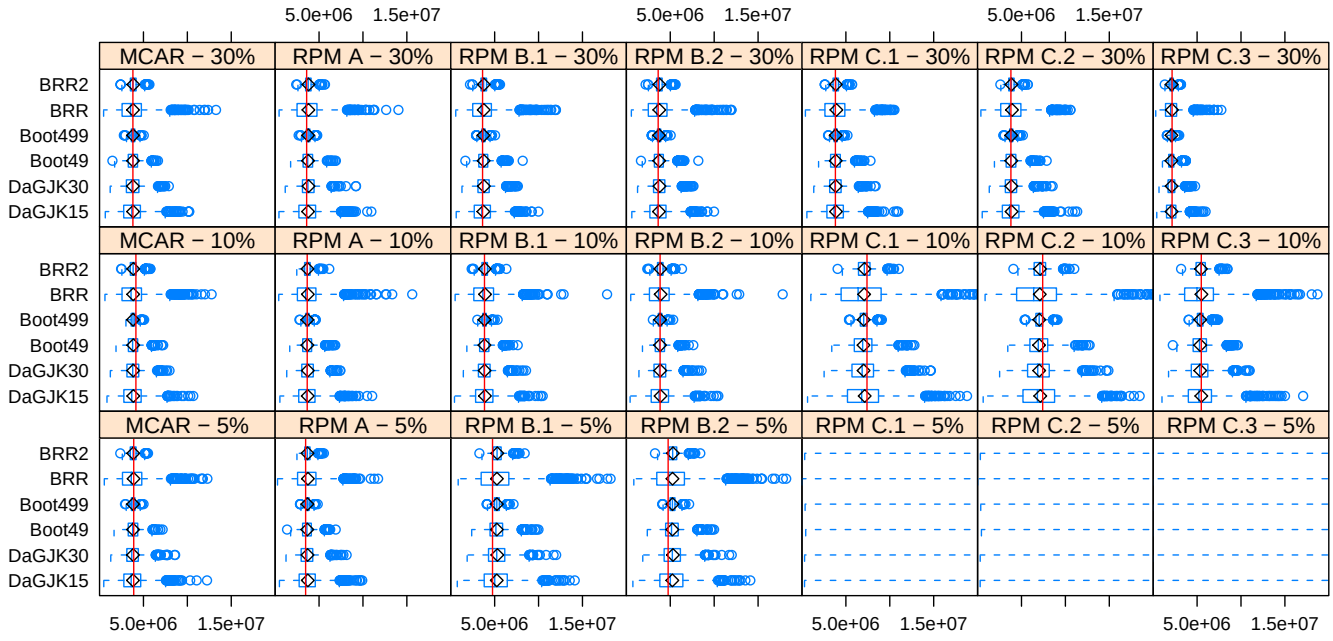

Figure 7. Variance estimates for strategy II.cal.MC

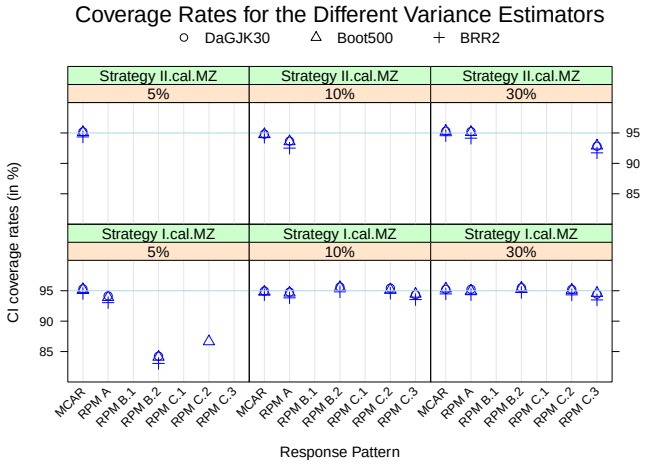

Figure 8. Coverage rates of the 95% confidence intervals for chosen resampling methods

5 Summary and outlook

The present study points out three important findings. First, it is strongly recommended to include propensity scores for the estimations. However, in practice, it is important to consider the variables that are responsible for the dropout when estimating propensity scores. In other cases, the estimates may become extremely biased and even a strategy without consideration of propensity scores can lead to better results. Third, very low response rates tend to yield extremely biased estimates. This certainly varies a little in accordance with the response pattern.

Accuracy measurement for the given estimation strategies is feasible, unless very complex sampling designs have to be considered. Applying resampling methods urges the needs for a high number of replications which certainly may become computer-intensive. In this study, the bootstrap has shown appropriate results once the response rate is not too low. However, when focusing on single estimation strategies, one may wish to develop (approximate) direct estimators in order to reduce the computational burden.

As long as the self-selection process or recruitment stage is specified correctly, the approach using response propensity scores outperforms other strategies. In cases, when the variables that are responsible for the self-selection are not known, the performance of the estimators using the corresponding weights may become considerably poor. In real applications, using misspecified response propensity models may yield highly biased estimates.

The focus of the present study was on the German DSP which is based on the MC. The MC is based on a mandatory 1% sample of the German population and furnishes a considerable response propensity modeling. Further, the participants already have four years experience in surveys. Nevertheless, we think that the findings shall also be valuable for other (commercial) access panels as long as a considerable number of variables for response propensity modeling is included. In all cases, the participation in access panels results from a self-selection process which has to be considered in the estimation later on.

Due to the rotational design of D-SILC and the DSP, further research will focus on longitudinal aspects of access panel based surveys (i.e. panel effects, estimation of change over time). In this case, adequate variance estimation techniques with respect to longitudinal aspects should be consid-

ered. Finally, besides compensating for unit non-response (i.e. the consideration of response propensities) one also has to deal with item non-response using adequate imputation techniques that take cross-sectional as well as longitudinal aspects into account.

Acknowledgements

The research of the present paper was done within the Priority Programme 1292 on Survey Methodology (cf. http://www.survey-methodology.de) of the German Research Foundation, DFG. Special thanks go to our project partners Professor Ulrich Rendtel and Boyko Amarov, Freie Universität Berlin, for their empirical research and findings on response propensities within the German access panel. Finally, we thank an associate editor and two anonymous reviewers for very valuable comments which helped to improve the readability of the paper.

References

Amarov, B., & Rendtel, U. (2011). Selectivity of Access Panel Recruitment, Survey Nonresponse and Estimation Strategies for the German Subsample of EU-SILC. *The Third Interantional Workshop on Internet Survey Methods, Statistics Korea*, 271-288.

Amarov, B., & Rendtel, U. (2013). The Recruitment of the Access Panel of German Official Statistics from a Large Survey: Empirical Results and Methodological Aspects. *Survey Research Methods*, *7*(2).

Baker, R., Blumberg, S., Brick, M., Couper, M., Courtright, M., Dennis, M., et al. (2010). *AAPOR task force report on online panels* (Tech. Rep.). American Association for Public Opinion Research.

Bruch, C., Münnich, R., & Zins, S. (2011). *Variance Estimation for Complex Surveys.* (Research Project Report No. WP3 - D3.1). FP7-SSH-2007-217322 AMELI. Available from http://ameli.surveystatistics.net.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in Telephone Survey Nonresponse over the past Quarter Century. *Public Opinion Quarterly*, *69*(1), 87-98.

Davison, A. C., & Sardy, S. (2004). *Resampling Methods for Variance Estimation.* (Research Project Report No. WP5 - D5.1). IST-2000-26057-DACSEIS. Available from http://www.dacseis.de/.

de Leeuw, E., & de Heer, W. (2002). *Trends in household survey nonresponse: A longitudinal and international comparison* (R. Groves, D. Dillman, J. Eltinge, & R. Little, Eds.). New York: Wiley.

De Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, *15*, 129-142.

Deville, J. C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Engel, U., Pötschke, M., Schnabel, C., & Simonson, J. (2004). *Nonresponse und Stichprobenqualität: Ausschöpfung in Umfragen der Markt- und Sozialforschung*. Verlagsgruppe Deutscher Fachverlag.

Horneffer, B., & Kuchler, B. (2008). Drei Jahre Panelerhebung EU-SILC: Erfahrungen und methodische Weiterentwicklungen. *Wirtschaft und Statistik*, *8*, 650-661.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663-685.

Körner, T., Nimmergut, A., Nökel, J., & Rohloff, A. (2006). Die Dauerstichprobe befragungsfreiwilliger Haushalte - Die neue Auswahlgrundlage für freiwillige Haushaltsbefragungen. *Wirtschaft und Statistik*, *5*, 451-467.

Lee, S., & Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Social Methods & Research*, *37*, 319-343.

Loosveldt, G., & Sonck, N. (2008). An Evaluation of the Weighting Procedures for an Online Access Panel Survey. *Survey Research Methods*, *2*(2), 93-105.

Münnich, R. (2008). Varianzschätzung in komplexen Erhebungen. *Austrian Journal of Statistics*, *37*(3 & 4), 319-334.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P., & Kolb, J.-P. (2012). *Stichprobenoptimierung und Schätzung im Zensus 2011* (Vol. 21). Statistisches Bundesamt, Statistik und Wissenschaft.

Münnich, R., Huergo, L., Magg, K., & Ohly, D. (2005). *Konzeption und Test von Varianzschätzung für Erhebungen auf Basis befragungsbereiter Haushalte* (Tech. Rep.).

Münnich, R., Knobelspieß, M., & Ohly, D. (2006). *Erhebungsdesign und Varianzschätzung im Access Panel am Beispiel von EU-SILC* (Tech. Rep.). University of Trier.

Nimmergut, A., & Körner, T. (2003). Zu den Möglichkeiten der Nutzung einer Dauerstichprobe befragungsbereiter Haushalte in der amtlichen Statistik. *Wirtschaft und Statistik*, *5*, 391-401.

Rao, J. N., & Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, *91*, 343-348.

Rosenbaum, P. R., & Rubin, D. R. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Schnell, R., Hill, P., & Esser, E. (2005). *Methoden der empirischen Sozialforschung* (Vol. 7). Oldenbourg.

Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer.