Survey Research Methods (2020)
Vol. 14, No. 2, pp. 103-108
doi:10.18148/srm/2020.v14i2.7751

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

Investigating selection bias of online surveys on coronavirus-related behavioral outcomes

Ines Schaurer
GESIS – Leibniz-Institute for the Social Sciences
Mannheim, Germany

Bernd Weiß
GESIS – Leibniz-Institute for the Social Sciences
Mannheim, Germany

The coronavirus SARS-CoV-2 outbreak has stimulated numerous online surveys that are mainly based on online convenience samples or commercial online access panels where participants select themselves. The results are, nevertheless, often generalized to the general population. In our paper we investigate the potential bias that is introduced by respondents' self-selection. The analysis is based on survey data of the "GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany", together with background information of the GESIS Panel. Our analyses show indication of a nonignorable amount of selection bias for measures of personality traits among online survey respondents. This provides some evidence that participating in an online survey and complying with measures that can minimize the risk of being infected with the SARS-CoV-2 virus are confounded. Hence, generalizing these results to the general population bears the risk of over- or underestimating the share of the population that complies with specific measures.

Keywords: COVID-19; online survey; nonprobability sample; self-selection bias; coronavirus; SARS-CoV-2; GESIS Panel

1 Introduction and background

With the coronavirus SARS-CoV-2 outbreak the demand for timely, non-clinical data that provides insight about various attitudinal and behavioral aspects of the crisis has increased enormously. Policy makers depend on this information, since any political intervention necessary to prevent the further spread of COVID-19 infection entails far-reaching restrictions for society. Moreover, the success of the restrictions depends on citizens' compliance.

Numerous surveys were implemented in order to gather the necessary information (for an overview see Matias & Leavit, 2020; Open Science Foundation, 2020; Rat für Sozial- und Wirtschaftsdaten, 2020). The vast majority of these surveys was conducted online to be able to collect timely data and also due to restriction and social distancing rules. They mostly rely on online convenience samples with self-selected participants or on quota samples from online access panels, which are subject to self-selection and potential bias due to noncoverage (Bethlehem, 2010a; Mercer, Kreuter, Keeter, & Stuart, 2017). Noncoverage occurs because about 12% of the German population still does not use the internet (Destatis, 2019), which is an issue for probability as well as nonprobability samples. Self-selection occurs due to the fact that respondents recruit themselves into the survey, rather than being selected (Baker et al., 2010; Kohler, Kreuter, & Stuart, 2019; Mercer et al., 2017). This leads to biased estimates if a variable is correlated with the outcome variable and the response propensity, respectively (Groves, 2006; Mercer et al., 2017). Post-hoc adjustment for selection bias is possible when specific assumptions are met (Groves, 2006), and is usually based on sociodemographic information; often without notable success with regard to bias correction (for a review see Cornesse et al., 2020a). With regard to estimating coronavirus-specific attitudes and behavioral measures, we are assuming additional confounding variables, for instance, personality traits or political attitudes that cannot be controlled or adjusted for because their distributions in the population are not known and they are usually not available in convenience web surveys or access panels. This fact increases the risk of biased estimates, in particular among point estimates of behavior or attitudes of the general population. In our paper we examine self-selection bias in web surveys for a selected set of potential confounding variables with regard to behavioral measures taken to decrease the risk of infection with SARS-CoV-2. From the universe of potential confounders we chose the BIG-5 measure of personality traits, a standard instrument in psychology that is also discussed as potential predictor of survey participation (Keusch, 2015). To describe potential selection bias, we will utilize data from a probability-based mixed-mode panel of

Contact information: Ines Schaurer, GESIS – Leibniz-Institute for the Social Sciences, B2,1, 68159 Mannheim (ines.schaurer@gesis.org)

the general population that also include respondents that do not use the internet.

2 Data and analysis strategy

Dataset

We use data from the GESIS Panel, a probability-based mixed-mode access panel that includes online and mail-mode respondents (Bosnjak et al., 2018). In particular, we will use the *GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany* (GESIS, 2020a), which was fielded between March 16th and 29th 2020, and can be linked to the GESIS Panel standard edition (GESIS, 2020b), which includes a wide range of background information.[1] Due to the necessity of timely data collection of the GESIS Panel special survey, only the subsample of GESIS Panel online respondents was invited. Overall, 3.765 panelists were invited, 3.176 completed the survey, resulting in a completion rate of 84.36%.

In the subsequent analyses, GESIS Panel offline respondents serve as a proxy group for respondents that are not willing or able to participate in an online survey (for details on the recruitment process see Bosnjak et al., 2018). We define all panelists that were invited to the last regular panel wave as active (n = 5.208). In our analyses, those that participated in the GESIS Panel special survey are referred to as participants (n = 3.176); the group of nonparticipants comprise of those that actually did not respond, and those that were not invited because they are usually participating via paper questionnaire (n = 2.032).

Analysis strategy

In this section, we briefly lay out our conceptual assumptions that guide our empirical analyses, which will be described thereafter. Figure 1 illustrates an assumed missingness mechanism that might introduce self-selection bias in, for example, the estimated prevalence of measures respondents have taken to decrease their infection risk (Y^*) (with ϵ representing the respective disturbance terms). Figure 1 is called an *m*-graph and graphically displays theoretical knowledge and assumptions about relationships among variables and missingness (R_Y) (Thoemmes & Mohan, 2015). Here, we concentrate on the identification of an unbiased estimate of an univariate parameter (e.g., the prevalence of Y in the population, observed as Y^*). With respect to online surveys and self-selection issues, the response indicator R_Y is assumed to be not missing at random (NMAR) (Little & Rubin, 2002). For instance, R_Y represents the participation or nonparticipation in an online survey. Z represents a set of variables that are known for participants and nonparticipants, usually sociodemographic information, which can be adjusted for. Furthermore, L represents a set of latent variables that are not known for nonparticipants (or for which no external benchmark information is available) and represent potential confounding variables. In order to (asymptotically) estimate a consistent parameter value in the presence of missing data, Y needs to be statistically independent of R_Y (and in the MAR case given Z and L), more formally, $Y \perp\!\!\!\perp R_Y | \{Z, L\}$ (Thoemmes & Mohan, 2015, p. 638).

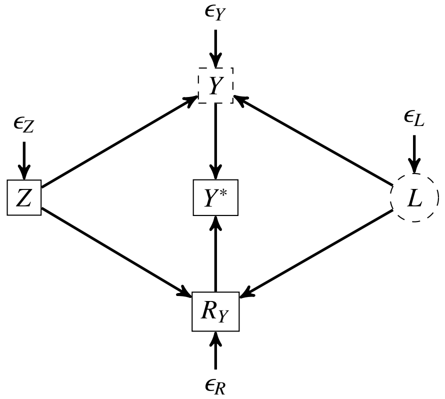

Figure 1. In the presence of missing data (R is a missingness indicator, e.g., nonparticipation), Y cannot be estimated by Y^* due to an open path from Y to R_Y via L (one or more latent variables) (Figure adapted from Thoemmes & Mohan, 2015, p. 635)

In our analyses, though, we assume that at least some relevant L-variables are known, i.e., they are no longer considered to be latent. Here, we chose personality information (BIG-5), which is usually not available in convenience online surveys or online access panels, and can hence be considered a latent variable.[2] So, the aim of our empirical analyses is to identify those confounding variables L that are correlated with the outcome of interest (Y) as well as the participation indicator (R_Y).

Due to the limited number of pages, we focus on only a few coronavirus-related outcome variables Y, namely a list of four measures respondents have taken to minimize their infection risk (wash hands more often, reduce social interactions, keep distance, avoid public places). Similarly, our set of L-variables that is assumed to explain both the outcome variable as well as the cause of missingness is limited to personality traits (BIG-5 inventory). We are utilizing the BFI-10 inventory (Rammstedt & John, 2007), a short scale that measures the BIG-5 with only ten instead of 44 questions.

Our empirical analyses are of exploratory nature and aim at identifying variables from the BFI-10 inventory (L) that are predictive for online survey participation (R_Y) *and* risk

[1] The linked data set will be available in the future in the GESIS data catalogue. As long as it is not published we provide it on request.

[2] In addition, it is an information that cannot be controlled or adjusted for since their distributions in the population are not known.

minimizing measures (Y). To estimate the bias due to self-selection, we would ideally be able to observe estimates of Y for respondents as well as nonrespondents. However, as mentioned above, information on risk minimizing measures (Y) is only available for the online sample of the GESIS Panel. Therefore, we apply the following two-fold strategy:

1. We identify relevant L variables by testing two independent models utilizing *two different* samples: (a) To investigate the path $L \rightarrow R_Y$, we will regress participation online (R_Y) on BFI-10 items (L), utilizing the offline and online sample of the GESIS Panel. (b) To investigate $L \rightarrow Y$, we will regress risk minimizing measures (Y) on BFI-10 items (L). In this case, we will be limited to the respondents of the GESIS Panel special survey. Details about the regression models as well as estimation results are omitted due to space constrains. In addition, it is important to note that we are assuming that the association between $L \rightarrow Y$ is not strongly biased by limiting our analyses to the online sample.[3]

2. Once we have identified relevant L-variables, we will calculate percentage differences with respect to R_Y, i.e., being online survey respondent or nonrespondent, and Y, i.e., four measures to minimize the risk of infection. For illustration purposes, the L-variables, i.e., the BFI-10 items, have been dichotomized at the respective mean values.

The rational behind this procedure is the assumption of an increased risk of selection bias in the estimate of the outcome in cases where L is predictive for R_Y *and* Y.

3 Results

In our analyses, we found that two out of ten BFI-10 items are associated with online participation (R_Y) and multiple risk minimizing measures (Y): (a) “I see myself as someone who tends to be lazy” and (b) “I see myself as someone who is generally trusting” (see Figure 2). Figure 2 is based on two 2×2 tables, i.e., “online participation (yes/no)” × “BFI-10 item (low/high)” and “risk minimizing measure (yes/no)” × “BFI-10 item (low/high)” (all statistical analyses were performed using R version 4.0.0, R Core Team, 2020). Reported are the percentage point differences between the two BFI-10 categories (e.g., tends to be lazy vs tends not be lazy) with respect to being a participant of the GESIS Panel special survey or not and the reported risk minimizing measures, respectively. Note that we report only percentage differences for those BFI-10 items that are statistically different from zero for online participation *and* risk minimizing measures. The differences in online participation for respondents that score high on the respective item are shown at the bottom of the two panels (see gray band). For instance, among participants of the GESIS Panel special survey the share of panelists that consider themselves as being “lazy” is by 10.59 percentage points higher, compared to the group of nonparticipants. The share of panelists that rate themselves as “being generally trusting”, is 3.83 percentage points higher among participants. Furthermore, the differences with respect to risk minimizing measures are denoted by points and the corresponding 95% confidence interval. We can see that “respondents that tend to be lazy” report less often to wash their hands or keep social distance. On the other hand, respondents that are “generally trusting” wash their hands more often or report more often that they reduced social interactions. All in all, we find considerable differences for certain coronavirus-related measures such as “washing hands” or “reducing social interaction” for two BFI-10 items. That is, respondents that consider themselves as “tending to be lazy” do not seem to comply well with risk minimizing measures. In contrast, respondents that are “generally trusting” seem more willing to adhere to these measures.

4 Discussion

In our paper we examined self-selection bias in web surveys for the BIG-5 measure of personality. These personality traits are assumed to serve as potential confounding variables with regard to participation in an online survey as well as behavioral measures taken to decrease the risk of infection with SARS-CoV-2. We based our analyses on data from a probability-based mixed-mode panel of the general population that also includes respondents that do not use the internet.

We were able to show that there is empirical indication of a nonignorable amount of selection bias for selected measures of personality among online survey respondents. Respondents that tend to be lazy have a higher participation probability *and* a lower probability to comply with risk-minimizing measures. This seemingly counter-intuitive result makes sense in view of our sample. We are analyzing members of a panel that are willing to participate in a survey in general. The online mode seems to be the most convenient way of participation. In sum, our results provide some evidence that participating in an online survey and complying to measures that can minimize the risk of being infected with the SARS-CoV-2 virus is confounded. In particular, when generalizing these results to the general population, it bears the risk of underestimating the share of the population that comply with specific measures.

The next step is to expand the set of coronavirus related behavioral outcome measures and potential confound-

[3] We also tested for the possibility that L lies on a path between Z and R_Y or Y, respectively, by running logistic regressions that include common Z variables (demographic information about age, highest school education, and sex.)

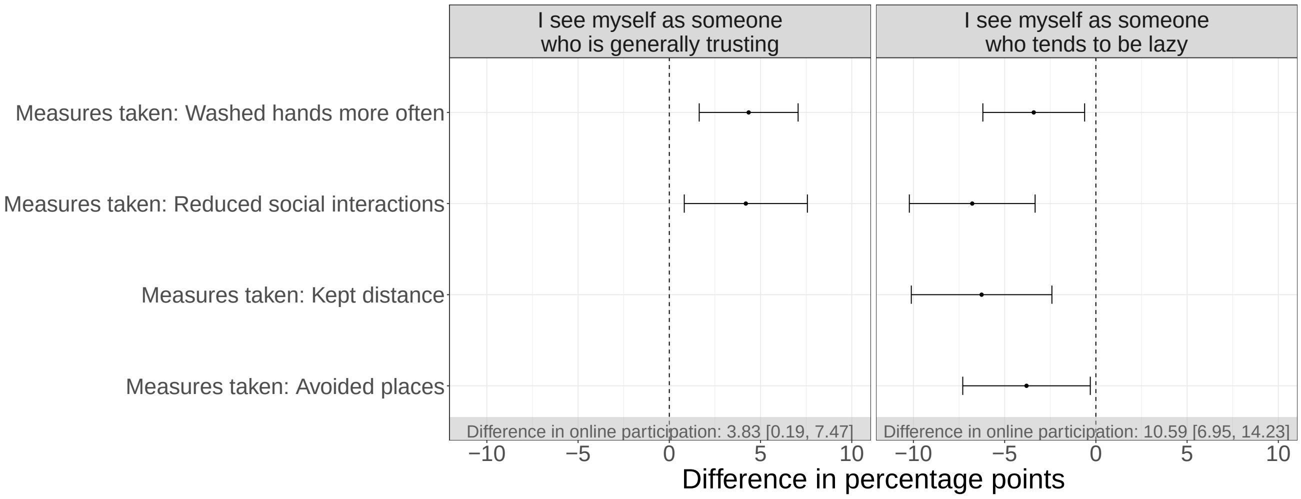

Figure 2. Differences (in percentage points) in online participation (R_Y, see bottom of panel) and risk minimizing measures (Y, see points with 95% confidence intervals) with respect to two BFI-10 items (Z).

ing variables. We could think of general risk behavior, health-related information on pre-existing conditions (e.g., in the vein of Schnell, Noack, & Torregroza, 2017) that increase the risk of a severe COVID-19 desease, or information on privacy issues or smartphone ownership, just to mention some.

Our analyses are based on several assumptions. First, we are assuming GESIS Panel offline panelists and nonrespondents to the special survey, to be comparable to the part of the population that does not participate in online surveys. As the nonrespondents in our analyses are members of a panel and willing to participate in surveys in general, the amount of selection bias is estimated conservatively and can be interpreted as the lower bound. Furthermore, the initial sample is based on a probability sample of the general population. In many of the surveys that are currently conducted we assume self-selection being much higher due to topic interest or accessibility of the survey. Second, we assume the existence of a selection bias, if L is predictive for Y and Ry, because we do not have any information for nonparticipants on Y. In a next step, we will be able to expand our analyses by including information about outcome measures for offline panelists. This will be possible because questions from the GESIS Panel special survey will be fielded again in the upcoming three GESIS Panel waves that include online and offline respondents.

We would like to end with a note of caution with regard to generalizing findings based on coronavirus-related surveys using convenient/self-selected online samples. Even though online surveys promise quick and affordable information, one should be aware of their restrictions and preconditions of the usefulness of data based on nonprobability samples (Cornesse et al., 2020a; Kohler et al., 2019; Mercer et al., 2017; Zack, Kennedy, & Long, 2019), especially with regard to descriptive inference about the general population.

Acknowledgement

The authors thank David Bretschi, Tobias Enderle, Michael Kühhirt, and the reviewers for their valuable comments on this paper.

References

Baker, R., Blumberg, S., Brick, J. M., Couper, M. P., Courtright, M., Dennis, M., … Zahs, D. (2010). *AAPOR report on online panels*. Retrieved May 6, 2020, from https://www.aapor.org/Education-Resources/Reports/Report-on-Online-Panels.aspx

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Bethlehem, J. (2010a). Selection bias in web surveys. *International Statistical Review*, *78*(2), 161–188. doi:10.1111/j.1751-5823.2010.00112.x

Bethlehem, J. (2010b). Selection bias in web surveys. *International Statistical Review*, *78*(2), 161–188. doi:10.1111/j.1751-5823.2010.00112.x

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS panel. *Social Science Computer Review*, *36*(1), 103–115.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., . . . Wenz, A. (2020a). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. doi:10.1093/jssam/smz041

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., . . . Wenz, A. (2020b). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. doi:10.1093/jssam/smz041

Destatis. (2019). Computer- und Internetnutzung im ersten Quartal des jeweiligen Jahres von Personen ab 10 Jahren. Retrieved May 9, 2020, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/IT-Nutzung/Tabellen/zeitvergleich-computernutzung-ikt.html

Ferri-Garcïa, R., & Rueda, M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS One*, *15*(4), e0231500. doi:10.1371/journal.pone.0231500

GESIS. (2020a). GESIS Panel special survey on the coronavirus SARS-CoV-2 outbreak in Germany. ZA5667 Datafile type: dataset. doi:10.4232/1.13485

GESIS. (2020b). GESIS Panel standard edition. ZA5665 Datafile Version 35.0.0 type: dataset. doi:10.4232/1.13436

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675. doi:10.1093/poq/nfl033

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken: John Wiley & Sons.

Homolak, J., Kodvanj, I., & Virag, D. (2020). Preliminary analysis of COVID-19 academic information patterns: A call for open science in the times of closed borders. Working paper. Retrieved from https://publons.com/publon/31471914/

Keusch, F. (2015). Why do people participate in web surveys? applying survey participation theory to internet survey data collection. *Management Review Quarterly*, *65*(3), 183–216. doi:10.1007/s11301-014-0111-y

Kohler, U. (2019). Possible uses of nonprobability sampling for the social sciences. *Survey Methods: Insights from the Field*, 1–12. doi:10.13094/SMIF-2019-00014

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, *6*(1), 149–172. doi:10.1146/annurev-statistics-030718-104951

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

Matias, N. J., & Leavit, A. (2020). COVID-19 social science research tracker. Retrieved from https://github.com/natematias/covid-19-social-science-research

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys. *Public Opinion Quarterly*, *81*(S1), 250–271. doi:10.1093/poq/nfw060

Open Science Foundation. (2020). Coronavirus outbreak research collection. Retrieved May 25, 2020, from https://osf.io/collections/coronavirus/discover

R Core Team. (2020). R: A language and environment for statistical computing. R version 4.0.0 patched (2020-04-25 r78297). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, *41*(1), 203–212. doi:10.1016/j.jrp.2006.02.001

Rat für Sozial- und Wirtschaftsdaten. (2020). Forschung zur Corona-Pandemie. Retrieved April 26, 2020, from https://www.ratswd.de/themen/corona

Schnell, R., Noack, M., & Torregroza, S. (2017). Differences in general health of internet users and non-users and implications for the use of web surveys. *Survey Research Methods*, *11*(2), 105–123. doi:10.18148/srm/2017.v11i2.6803

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 631–642. doi:10.1080/10705511.2014.937378

WHO. (2020). Global research in coronavirus disease (COVID-19). Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

Zack, E. S., Kennedy, J., & Long, J. S. (2019). Can nonprobability samples be used for social science research? A cautionary tale. *Survey Research Methods*, *13*(2), 215–227. doi:10.18148/srm/2019.v13i2.7262

Commentary

This paper addresses an important issue, namely that findings generated from online surveys that do not rely on a probability sampling (PSg) are falsely generalized to entire populations. Within the first weeks and months of the COVID-19 pandemic, the number of "rapid" online surveys that relied on nonprobability sampling (NPSg) skyrocketed in both the social and the natural sciences (WHO, 2020). This has led to the production of a considerable number of academic papers that have received considerable media attention (e.g. Homolak, Kodvanj, & Virag, 2020, for an analysis of COVID-19 related research in general).

In their comparison of the personality characteristics of online and mail-mode respondents in the GESIS Panel, Schaurer and Weiß found that both selection into the survey and compliance with COVID-19 safety measures were correlated with some latent personality traits (BIG 5), which leads to biased estimates (see Bethlehem, 1988, 2010b, for a more general discussion on the issue of self-selection and biased estimates). Presumably, the results by Schaurer and Weiß provide a lower-bound estimate of such biases, as biases in their study only arise due to the online-only mode and nonresponse (and not self-selection into the sample) (see Bethlehem, 1988, 2010b). The bottom line that researchers and policy makers should draw from this result (as well as many others on this topic) is that estimates, particularly descriptive estimates based on (online) nonprobability samples, can be quite misleading (e.g., Cornesse et al., 2020b).

We should, however, not throw the baby out with the bath water. First of all, in some instances—as it has been the case at the outset of the Corona crisis—it might still be better to have flawed data than no data at all. These data, however, need to be used and interpreted with caution and statistical knowledge. This requires education among decision makers who use these data to inform policies. Moreover, it requires appropriate tools among those who provide and analyze the data (e.g., Ferri-Garcïa & Rueda, 2020), and both access to these data and transparency about the data generation process for the research community (Cornesse et al., 2020b provide an overview on further recommendations). Second, NPSg can be appropriate for certain purposes (Groves et al., 2009). Kohler (2019), for example, outlined a couple of "research scenarios" in which valid information can be derived from NPSg. In particular, he argues that valid estimates can be derived from NPSg data under the assumption of homogeneous research units and for answering causal questions that do not aim at identifying population average treatment effects. These insights should be used to generate meaningful findings on the transmission of the SARS-CoV2 virus and the consequences of the COVID-19 pandemic, even in the absence of probability samples.

Lena Hipp
Berlin Social Science Center (WZB), and
University of Potsdam

Mareike Bünning, Stefan Munnes, and Armin Sauermann
Berlin Social Science Center (WZB)