940 2021

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

SOEP-Core – 2019: Design, Nonresponse, and Weighting in the Sample Q (Queer)

Lisa de Vries, Mirjam Fischer, Martin Kroh, Simon Kühne, David Richter

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:

Lisa de Vries, Mirjam Fischer, Martin Kroh, Simon Kühne, David Richter. 2021. SOEP-Core – 2019: Design, Nonresponse, and Weighting in the Sample Q (Queer). SOEP Survey Papers 940: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

Design, Nonresponse, and Weighting in the 2019 Sample Q (Queer) of the Socio-Economic Panel

Lisa de Vries, Mirjam Fischer, Martin Kroh,
Simon Kühne, David Richter

April 20, 2021

Abstract
This report details the design of the boost Sample Q to the Socio-Economic Panel (SOEP) in 2019. Sample Q supplemented the SOEP Core sample by *queer* households, including gender and sexual minorities such as lesbian, gay, bisexual, and trans* respondents (LGBTQ*). Recruitment of these households was performed by a random telephone screening of adults living in Germany. Sample Q comprises of 477 newly recruited LGBTQ*queer households and this boost sample thus augments the 405 already existing queer households in the SOEP to a total of 882 households with 1,237 respondents identifying as sexual and gender minority.

1 Introduction

Diversity in sexual orientation and gender identity (i.e., the personal sense of one's own gender) represents an aspect of life that is becoming increasingly important in individuals, families, societies, and politics. The social situation of lesbians, gays, bisexuals, trans, inter and queer persons (LGBTQ*) thereby touches upon various life domains such as education, the labor market, health and well-being, family life, and many more. In 1994 the abolishment of Section 175 of the German Criminal Code (known formally as §175 StGB) which made homosexual acts between males a crime, has been a mere beginning in correcting decades long structural marginalization of sexual and gender minorities.[1] Since then, legal changes in Germany continue to occur incrementally. In 2017, the institution of marriage was opened to same-sex couples; in 2019, a third gender option officially acknowledges the diversity in genders.

The importance of gender identity and sexual orientation as a personal characteristic is underlined by the General Equal Treatment Act (AGG), in which they have the same relevance as, for example, ethnic origin, age, or religion. Despite the importance of the topic and the growing need to take it into account in social reporting and research, the empirical data infrastructure for investigating the living conditions of LGBTQ* is severely limited.

These recent legal changes in Germany have both increased the visibility of sexual and gender minorities and, simultaneously, highlighted the severely limited empirical data infrastructure for investigating the living conditions of sexual and gender minorities in Germany. Traditionally, the field has relied on convenience samples and recruited sexual and gender minorities from LGBTQ* venues, social clubs, interest and citizen groups and health facilities that tend to the needs of the queer community. However, there is an increasing need for an empirical data basis, which is lends itself for cross-sectional and longitudinal inferences beyond the limited generalizability of convenience samples (OECD, 2019; Valfort, 2017). This has been partially addressed by adding sexual orientation and gender identity items to existing population surveys like the UK Understanding Society Study (Booker, Riger, & Unger, 2017), the Australian Household, Income and Labour Dynamics Survey (Wooden, 2014), and the SOEP (Goebel et al., 2019).

Alternatively, an albeit limited indicator of sexual orientation in existing multidisciplinary surveys can be inferred indirectly from the reported gender-constellation of respondent's current and previous partnership(s). Researchers can draw on self-reported gender and the self-reported relationship between two respondents of a household in order to identify same-sex and opposite-sex couples (for a comparison of self-reported and partnership-inferred sexual orientation in household surveys see Kühne, Kroh, & Richter, 2019). As the group of sexual and gender minority people is numerically small (an estimated 2-6% of the general population; Gates, 2011), both strategies often do not yield enough observations to conduct meaningful analyses.

To address these shortcomings in the data infrastructure, the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) funded a boost sample of sexual and gender minorities in Germany (SOEP-LGB).[2] The boost sample of queer households was established in 2019 and integrated into the SOEP.

[1] http://dipbt.bundestag.de/extrakt/ba/WP19/2390/239069.html

[2] Grant numbers 01UW1803A and 01UW1803B: "Supplementing the SOEP Data Infrastructure with a Sample of Lesbians, Gays, and Bisexuals (SOEP-LGB)"

The collection of quality survey data of sexual and gender minority people is not without challenges as lesbians, gays, bisexuals, transgender persons, and those who identify as non-binary represent a so-called hard-to-survey population. Specifically, we are dealing with a numerically small group compared to the cis-heterosexual majority for whom no complete list is available to serve as a sampling frame. This is true for many population segments of interest to social science (e.g., religious and ethnic groups). However, in the case of sexual and gender minority people such a list would not be desirable in the first place, as it would infringe on protecting the privacy of a vulnerable group in society.

In addition, the actual percentage of persons identifying as member of a sexual and gender minority in the general population is unknown. This paper documents how these challenges were addressed in the sampling design and the weighting strategy for the 2019 Sample Q of the SOEP.

2 Sampling Design

The target population of Sample Q was the population of households in Germany with at least one resident aged 18 and up who identifies as lesbian, gay, bisexual, queer, pansexual, asexual, polysexual, demisexual, or trans* (LGBTQ*). According to estimates of Eurostat, virtually all private households in Germany can be contacted by landline and/or mobile phone.

For Sample Q, a telephone screening of the adult population in Germany was administered in the context of nationwide omnibus surveys of Kantar Public, one of the largest fieldwork organizations in Germany. The sample was drawn using the dual-frame method (e.g., Buskirk & Best, 2012), which makes it possible to infer sampling probabilities based on a mixture of tranches of mobile and landline samples. The approach also allows to reach the increasing number of respondents that only have a mobile phone and no landline access (estimated at about 14% in 2019).

In the dual-frame approach, the sampling is not performed based on one, but two selection frames, which together cover the population completely: one with telephone numbers exclusively from the (regionally stratified) landline network and the other with synthetically generated mobile phone numbers. To ensure regionally representative sampling, mobile phone numbers are overrepresented when the screening is started and reduced during the process as landline numbers from so-far underrepresented regions are added. The CATI screening interviews were conducted in the Kantar Public CATI omnibus survey and in a second CATI screening survey, exclusively conducted to boost the gross sample of SOEP Sample Q.

Persons identifying as LGBTQ* were identified by means of the following survey items used in the screening interview. First, people were asked if they would be willing to answer questions about sexuality and gender identity. Second, among those who were willing to provide information, respondents were asked to indicate whether they subscribe to a heterosexual, lesbian/gay, bisexual, or other sexual orientation. In the open answers we consider queer, pansexual, asexual, polysexual, or demisexual as categories of the target population. Respondents who did not subscribe to a heterosexual sexual orientation were screened-in as part of the target population.

Gender was surveyed according to an internationally used two-step method (Tate, Ledbetter, & Youssef, 2013). Thereby, people indicate whether they were assigned female or male

in their birth certificate. At the time of birth of people who are now 18 years and older, there was no option to have anything but female or male entered into birth certificates. In the following question, people were asked to indicate their current gender identity, which may be female, male, transgender, or none of these (again with an open answer option). People who identified as transgender or none of these genders were screened in, as well as people whose current gender identity did not align with the gender they were assigned at birth. In order to reflect this diversity, the Asterix after trans* is used when summarizing this group. The exact wording of the questions can be found in the appendix.

When people identified as members of the target population, they were invited to partake in the face-to-face SOEP survey. After elimination of probable false positive screenings, a gross sample of 835 households was obtained.

3 Fieldwork Results and Response Rates

Overall, roughly 75,000 screening interviews (including pretest screenings) were conducted between September 2018 and August 2019. Table 1 displays the results of the random telephone screening. In total, the number of randomly screened households lies at 74,998. Of those households, 53,497 indicating their willingness to provide information on their sexual and gender identity. Based on the subsequent answers on sexual orientation and gender identity, 2,824 respondents were identified to be part of the target population of sexual and gender minority people. Therefore, the proportion of persons identifying as member of a sexual and gender minority in our screening sample (5.3% of the households that provided answers to the questions on sexual orientation and gender identity) lies at the upper end of the estimated range of 2-6% of the general population.

Of those 2,824 persons, 1,093 agreed to participate in SOEP (38.7% of the target group) and 1,023 provided their contact information for face-to-face interviewing (36.2% of the target group). After exclusion of false positive screenings (6.7%), the final gross boost sample of households with complete contact information was 835 households (29.6% of the target group).

Table 1: Results of the random telephone screening.

	N	Percent of Total	Percent of Valid
Total screening interviews	**74,998**	**100.0**	
Refusal to answer questions on sexual orientation and gender identity	21,501	27.7	
Completed screening interviews	**53,497**	**71.3**	**100.0**
In LGBTQ* target group	2,824	3.8	5.3
Intent to participate in SOEP	1,093	1.5	2.0
Contact information provided	1,023	1.4	1.9
Final gross boost Sample Q	835	1.1	1.6
False positive screenings	188	0.3	0.4

Roughly seven months after the CATI screening was started, the field work of the SOEP started and face-to-face interviews were conducted from April to November 2019. Most households were interviewed in September (28.6%) and July (18.1%).

Sampled households were provided with information that was sent out by mail prior to the actual interview. A sample specific leaflet was designed to encourage participation and to outline the scientific purpose of the survey. It was sent along with the advance letter detailing the survey process and announcing that an interviewer would contact the household. In addition, households were informed that they would receive a cash incentive of 5 euros for a completed household interview and 10 euros for each completed individual interview. Interviewers were also provided with declarations on data protection and contact cards which could be left if no one was present at the interviewer's first contact attempt.

Table 2 displays the results of the fieldwork. Overall, 477 households were successfully interviewed, which corresponds to a response rate of 57.1%. The response rate adjusted for households in which the last household member had died, households that moved abroad, untraceable households and quality neutral sampling losses was 60.5%. A total of 170 households (20.4%) refused to participate in the survey. Another 14.9% could not be contacted during the fieldwork period.

Table 2: Fieldwork results on household level (with AAPOR disposition categories).

	N	In Percent
Gross sample	**835**	**100.0**
Interview	**447**	**57.1**
Complete	331	39.6
Partial	146	17.5
Non-interview	**358**	**42.9**
Non-contact	124	14.9
Temporarily, physically, or mentally unable	0	0.0
Language problem	4	0.5
Permanent refusal	170	20.4
Permanently physically or mentally unable	7	0.8
Deceased	2	0.2
Moved abroad	3	0.4
Household untraceable	25	3.0
Quality-neutral sampling losses	17	2.0

The questionnaires in Sample Q were mostly identical to the standard SOEP questionnaires of 2019 with few additional population-specific questions, such as outing and minority stress. The 477 households participating in the survey comprised of 741 adults, 14 youths, and 58 children. 564 adults (76.1%) and 7 youths were successfully interviewed. 177 adults residing in participating households could not be interviewed temporarily. Partial unit non-response (PUNR: the share of households with more than one household member with at least one missing individual questionnaire) was at 67.9% and relatively high. In part, this can be attributed to the anchor-person design, where the focus is on the anchor while other persons in the household, who may or may not belong to the target population, might not feel explicitly addressed.

4 Cross-Sectional Weighting

Cross-sectional weights of Sample Q were constructed in consecutive steps (Brick & Kalton, 1996). First, design weights consider the probability of inclusion to the CATI bus surveys. Also, design weights correct for the number of persons from the target population per household. Second, based on the gross sample, we estimated stepwise response propensities that resulted in weights adjusting for refusals. Finally, the combination of design and non-response weights lead to the cross-sectional weights of Sample Q in 2019. For details on the general weighting strategy of the SOEP and the integration of new samples see Kroh et al. (2015).

4.1 Design Weights

During the screening, 52,675 landline numbers and 22,324 mobile phone numbers were successfully contacted (30 percent mobile). In a first step, design weights of Sample Q describe the selection probability in the dual frame setup (e.g., Buskirk & Best, 2012). This approach requires information of sampled households on household size and the number of mobile and landline phones. Similarly, the approach requires population estimates on the number of households by size and contactability.

According to the estimates of Kantar Public, 50.7 Million landline and 72.4 Million mobile phones were used in 2017. The selection based on landline and mobile phones leads to a person-centered selection frame. Consequently, households with two, three, or more persons older than 14 years have a two-, three-, or manyfold probability to be selected into the sample. According to the Federal Statistical Office of Germany (Destatis), the total number of adult residents in private households in Germany in 2019 was 69.3 Million and total number of households was 41.5 Million.

Based on Häder (2016), the selection probably for individual respondents was calculated by the *number of landline phones of the person* (k_i^F) multiplied with the *size of landline sample* (m^F) divided by the *size of landline frame* (M^F) multiplied by *1* divided by the *number of persons in the household* (z_i) and the *number of mobile phones of the person* (k_i^C) multiplied with the *size of mobile phone sample* (m^C) divided by the *size of mobile phone frame* (M^C).

Thus, the inclusion probability π_i of person i arises as

$$\pi_i = k_i^F \frac{m^F}{M^F} \cdot \frac{1}{z_i} + k_i^C \frac{m^C}{M^C}, i = 1, \ldots, N$$

The dual frame approach results in sampling probabilities - and thus design weights - at the person level. In addition, the approach is based on estimates of the total population of residents in Germany. Sample Q, however, is a household sample of a smaller target group, namely queer households. Hence, we consider the sampling probability of persons outside the target population to be zero. Moreover, we multiply the selection probability of households by the number of household residents who are members of the target population. Therefore, households with same-sex couples receive two times the selection probability and households with one heterosexual and one homosexual resident, for instance, receive the selection probability of a single queer household.

4.2 Non-Response Weights

To correct for potential bias due to selective non-response, the participation of households is modelled in two steps. The first step involves modelling non-response to the questions on sexual orientation and gender identity in the CATI screening. A refusal to answer the questions on sexual orientation and gender identity prevents us from assessing whether a respondent might fall into the target group of LGBTQ* households.

For this modelling step, information on both the participating and the non-participating households is needed. Generally, such information is scarce. Since the questions on sexual orientation and gender identity were part of a larger omnibus survey placed toward the end of the screening interview, however, there are some information on the individual level - age, education, and employment status - and information on the household - number of adults and children in the household and household income - that can be used.

In addition, we adjust for which part of the sample the person originates from (telephone only, dual frame telephone, dual frame mobile or mobile only). Moreover, we use municipality size and regional level data at the county level to estimate non-response propensities. Detailed documentation on the data is provided by INKAR (2019).

The second step models participation in the SOEP face-to-face survey among those of the CATI screening who are part of the target population. In addition to the individual level, household, and regional data described before, we additionally know whether respondents identify as lesbian, gay, or bisexual and whether they have a trans* identity. This information can be added to the second model, which estimates non-participation in the SOEP survey after the survey institute Kantar Public contacted them based on the information, they provided in the screening interview.

Table 3 display the results of logistic modeling of response to the sexual orientation and gender identity questions asked in the screener survey. We see that respondents who were contacted by mobile phone were more likely to answer the questions about their sexual orientation and gender identity than those who were reached via landline numbers. Younger respondents, respondents in single or two-person households, employed respondents, respondents with a tertiary education, and respondents living in urban areas were also more likely to answer the questions on sexual orientation and gender identity.

Table 4 display the results of logistic modeling of successful participation in SOEP survey. Being younger, living in single or two-person households, and having a lesbian, gay, or bisexual identity considerably increases the likelihood to partake in the survey.

Variables also tested but not systematically and robustly related to the propensity to answer the questions about sexual orientation and gender identity or to the probability to take part in the SOEP survey were household income, and several regional level indicators.

Table 3: Logistic modeling of SOGI response in screener survey.

Variable	Estimate	Std. Error
Individual level information		
3 person household	-0.016^{**}	0.006
4 person household	-0.040^{***}	0.006
5+ person household	-0.061^{***}	0.008
Tertiary education	0.064^{***}	0.003
Age 30-59 years	-0.082^{***}	0.007
Age 60+ years	-0.135^{***}	0.007
Employed	0.035^{***}	0.005
No landline phone	-0.166^{***}	0.012
No mobile phone	-0.088^{***}	0.007
Screener mobile only	0.352^{***}	0.014
Screener pretest	0.041^{***}	0.007
Municipality size < 5000	-0.019^{***}	0.005
Municipality size 50.000 – 500.000	0.026^{***}	0.004
Municipality Size > 500.000	0.044^{***}	0.005
County level information (INKAR)		
High population increase 65+	0.017^{***}	0.004
High number of divorces	0.023^{***}	0.004
High number of day-care-centers	0.019^{***}	0.004
Observations	73,169	

Notes: Dependent variable SOGI response indicates whether sexual orientation and gender identity (SOGI) questions were answered in screener survey (1 = yes). Coefficients estimated with logistic regression models in Stata 15. Significance indicated by $*** \equiv p < 0.001$, $** \equiv p < 0.01$, and $* \equiv p < 0.05$.

4.3 Weighting Strategy

Multiplying the sampling probability (see 4.1) and the response probabilities (see 4.2) of the households results in the observational probabilities of the households of Sample Q, the inverse is the cross-sectional weight of Sample Q. Based on this self-weighting scheme, we conclude the number of households in Germany with at least one adult member of the queer community – i.e. the total of cross-sectional weights of Sample Q – to be 1.5 million (about 4 percent of all private households) in 2019. The number of adult persons residing in these households is estimated to be 2.2 million.

Since size and marginal distribution of the sexual and gender minority population in Germany is unknown due to the lack of an exhaustive lists and census data, we refrained to further post-stratify cross-sectional weights. Although the micro-census estimates the number of households with same-sex partnerships to be 142,000 in 2019, our much higher estimated population sizes in Sample Q and SOEP-Core appear to us more plausible in comparison to cross-national estimates (see, for instance, Kühne et al., 2019).

The distribution of the weights on the household-level resulting from each step are displayed in Figure 1. Mean value of the final weights on the household-level was 3,175.46 with a standard deviation of 2,192.28 and a range from 581.67 to 12,733.69.

Table 4: Logistic modeling of successful participation in SOEP survey.

Variable	Estimate	Std. Error
Individual level information		
4 person household	-0.100^{*}	0.045
5+ person household	-0.153^{*}	0.075
Age 60+ years	-0.058^{***}	0.023
Homo/Bisexual	0.130^{***}	0.036
Screener mobile only	-0.101^{***}	0.026
County level information (INKAR)		
Few job vacancies for skilled labor	0.010^{***}	0.020
Low levels of loans	0.059^{**}	0.022
Metropolitan area	0.060^{**}	0.021
Observations	1,962	

Notes: Dependent variable SOEP response indicates successful participation in the SOEP survey among those in the target population (1 = yes). Coefficients estimated with logistic regression models in Stata 15. Significance indicated by $*** \equiv p < 0.001$, $** \equiv p < 0.01$, and $* \equiv p < 0.05$.

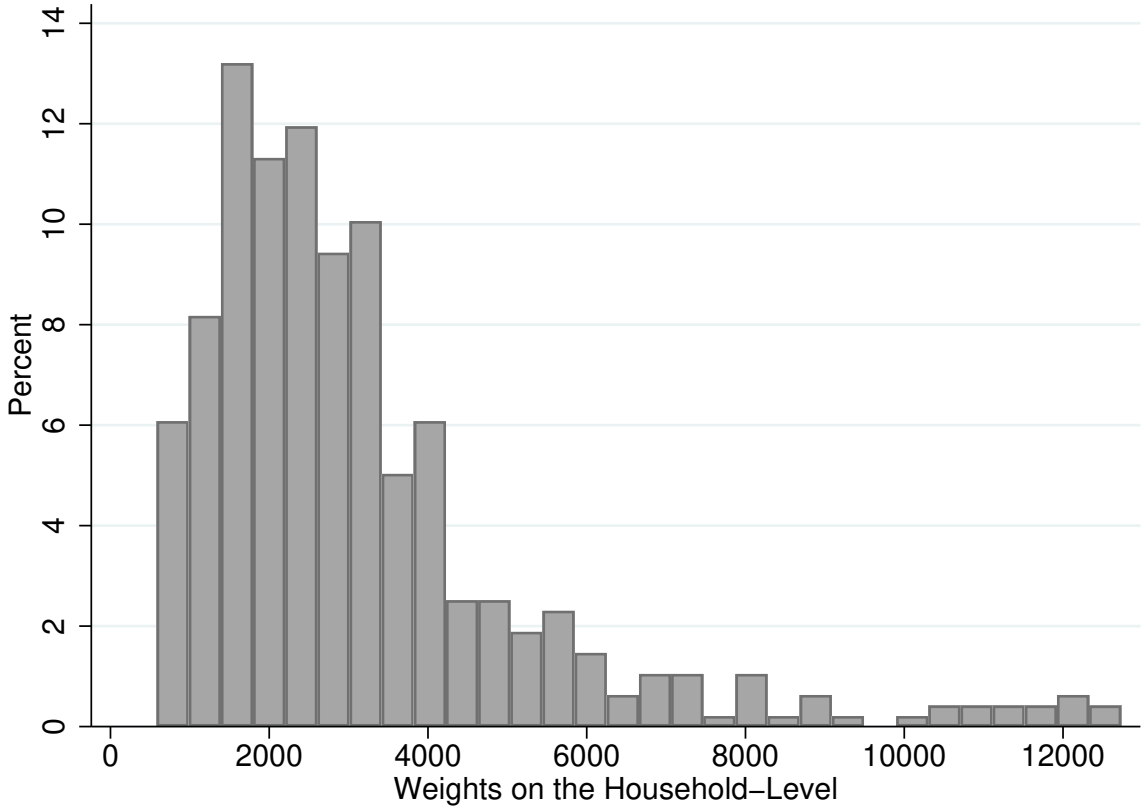

Figure 1: Distribution of weights for the first wave of Sample Q

5 Summary

Sample Q enriches the SOEP by 477 households of sexual and gender minority people. Combined with the existing 405 households in the panel, this is the first reliable empirical data source on this population in Germany. In addition, the SOEP is now one of few randomly selected samples with more the 1000 respondents identifying as part of a sexual and gender minority group, allowing direct comparison with the non-LGBTQ* population due to identical questionnaires.

The sample was obtained by means of a random telephone screening of German-speaking adults living in Germany. Albeit costly, this provides us with the much-needed empirical basis to study the social situation of LGBTQ* people in various life domains such as education, the labor market, health and well-being, family life, and many more (for first evidence see de Vries, 2020). The integration of this sample into the SOEP core sample will ensure the continuous fruits of this highly unique investment. It will be the first panel of it's kind.

References

Booker, C. L., Riger, G., & Unger, J. B. (2017). Sexual orientation health inequality: Evidence from Understanding Society, the UK Longitudinal Household Study. *Preventive Medicine*, *101*, 126–132. doi: 10.1016/j.ypmed.2017.06.010

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

Buskirk, T. D., & Best, J. (2012). Venn Diagrams, Probability 101 and Sampling Weights Computed for Dual Frame Telephone RDD Designs. *American Statistical Association - Proceedings of the Survey Research Methods Section*, 3696–3710.

de Vries, L. (2020). *Regenbogenfamilien in Deutschland. Ein Überblick der Lebenssituation von homo- und bisexuellen Eltern und deren Kindern. Expertise für den Neunten Familienbericht der Bundesregierung.* Retrieved from https://www.dji.de/9_familienbericht

Gates, G. J. (2011). *How many people are lesbian, gay, bisexual and transgender?* Los Angeles: The Williams Institute.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics*, *239*(2), 345–360. doi: 10.1515/jbnst-2018-0022

Häder, S. (2016). *Sampling in Practice. GESIS Survey Guidelines.* Mannheim: GESIS – Leibniz Institute for the Social Sciences. doi: 0.15465/gesis-sg_en_014

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

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. doi: 10.1007/978-3-658-10459-7_13

Kühne, S., Kroh, M., & Richter, D. (2019). Comparing self-reported and partnership-inferred sexual orientation in household surveys. *Journal of Official Statistics*, *35*, 777–805. doi: 10.2478/jos-2019-0033

OECD. (2019). *Society at a Glance 2019: OECD Social Indicators.* Paris: OECD Publishing.

Tate, C. C., Ledbetter, J. N., & Youssef, C. P. (2013). A two-question method for assessing gender categories in the social and medical sciences. *Journal of Sex Research*, *50*, 767–776. doi: 10.1080/00224499.2012.690110

Valfort, M.-A. (2017). LGBTI in OECD Countries: A review. In *OECD Social, Employment and Migration Working Papers, No. 198.* Paris: OECD Publishing. doi: 10.1787/d5d49711-en

Wooden, M. (2014). *The measurement of sexual identity in wave 12 of the HILDA Survey (and associations with mental health and earnings).* Melbourne: Melbourne Institute for Applied Economic and Social Research.

Appendix

In the context of relationships, the question of sexual orientation arises. Would you describe yourself as... ?

1. Heterosexual (that is attracted to the opposite sex)
2. Homosexual (gay or lesbian, that is attracted to the same sex)
3. Bisexual (that is, attracted to both sexes)
4. Other, namely… [only in SOEP-LGB, not in SOEP-core, yet]
5. No Answer

Which sex was assigned to you on your birth certificate?

1. Female
2. Male
3. No answer

How would you yourself describe your gender?

1. Female
2. Male
3. Transgender
4. None of these, …
5. No answer