Research Institute
Social Cohesion

RDC

SOEP

SCP Documentation

German Social Cohesion Panel

SCP 2021-23 W1-3 Codebook HHRF: Weights for Households (English)



Research Institute Social Cohesion

SOEP The Socio-Economic Panel

German Social Cohesion Panel

Established in 2021, the German Social Cohesion Panel (SCP) is a wide-ranging representative longitudinal study of private households in Germany, carried out in collaboration of the Research Institute Social Cohesion (RISC) and the German Socio-Economic Panel (SOEP).

The aim of the SCP Documentation is to thoroughly document the survey's data collection and data processing.

Recommended Citation

Groh-Samberg, O., Axenfeld, J. B., Gerlitz, J.-Y., Cornesse, C., Kroh, M., Lengfeld, H., Reinecke, J., Teichler, N., Traunmüller, R., & Zinn, S. (2026). SCP 2021-23 W1-3 - Codebook HHRF: Weights for Households (English). *German Social Cohesion Panel 2021-2023 - Wave 1-3*. Bremen and Berlin: RDC-RISC/SOCIUM, SOEP/DIW Berlin. doi:10.60532/scp.2021-23.w1-3.v1

- **Authors:** Olaf Groh-Samberg, Julian B. Axenfeld, Jean-Yves Gerlitz, Carina Cornesse, Martin Kroh, Holger Lengfeld, Jost Reinecke, Nils Teichler, Richard Traunmüller, Sabine Zinn

- **Contributors:** Cosima Adams, Anton Bochert, Martin Gerike, Josefine Kuhrmeier, Anna-Tabea Müller, Eric Nissen, Sebastian Rueda-Uribe, Rainer Siegers, Hans Walter Steinhauer, Knut Wenzig, Julia Witton (Project Members), infas (Data Collector)

- **Publisher:** RDC-RISC
  SOCIUM, University of Bremen
  P.O. Box 330 440
  28334 Bremen
  Germany

  SOEP
  DIW Berlin
  German Socio-Economic Panel (SOEP)
  Anton-Wilhelm-Amo-Str. 58
  10117 Berlin
  Germany

- **DOI:** 10.60532/scp.2021-23.w1-3.v1

- **Website:** www.fgz-risc-data.de

  www.diw.de

The text of this publication is published under the Creative Commons license CC BY-SA 4.0 Attribution-ShareAlike 4.0 International. The exact wording of the license CC BY-SA 4.0 can be found here:

https://creativecommons.org/licenses/by-sa/4.0/

SCP Documentation

German Social Cohesion Panel

SCP 2021-23 W1-3 Codebook HHRF: Weights for Households (English)

Contents

1 General Information 2

2 Identifiers 2
hid – Current Household ID 2
cid – Original Household ID 3

3 Survey Context 3
wave – Survey Wave 3

4 Statistical Weighting Factors 4
design – Inverse Sampling Probability 4
hhrf – Weighting Factor 5

5 Inverse Staying Probability 6
hbleib – Inverse Staying Probability 6

1 General Information

The HHRF dataset contains survey weights for households in the SCP. Each household (HID) that responded to the household questionnaire in a particular survey wave (WAVE) has one row in the dataset. Households whose anchor person did not take part in the survey wave or broke off the survey before the household questionnaire began are not included in the data set.
In some places in the documentation and in the data, year numbers are used, for example, for the names of variables and of the questionnaire instrument. These year numbers are always based on the field start of the data collection of the corresponding survey wave.

2 Identifiers

hid – Current Household ID

21100003		3
21100009		4
21100010		1
21100012		1
21100016		4
21100020		4
21100021		4
21100032		1
21100037		1
21100039		4
21100044		1
21100045		1
21100049		4
21100050		1
21100058		1
...	(13071 rows omitted)	31362
22250034		1
22250037		1
22250038		1
22250039		1
22250040		1
22250041		1
22250043		1
22250044		1
22250045		1
22250046		1
22250047		1
22250048		1
22250049		1
22250050		1
22250051		1

This identifier groups all individuals into their respective households at the time of the most recent wave. This means that a person's HID can change over time, for example if an adult

child moves out of their parents' home and starts their own household. The HID reflects the status prior to the fieldwork phase. Changes in household membership are only accounted for in the HID of the subsequent wave.

cid – Original Household ID

21100003		3
21100009		4
21100010		1
21100012		1
21100016		4
21100020		4
21100021		4
21100032		1
21100037		1
21100039		4
21100044		1
21100045		1
21100049		4
21100050		1
21100058		1
...	(13023 rows omitted)	31347
21137960		1
21137961		1
21137963		4
21137964		3
21137967		3
21137971		1
21137972		1
21137973		1
21137976		1
21137978		2
21137979		4
21137985		2
21137987		1
21137991		4
21138000		1

This identifier groups individuals into their original households at the start of the panel. That means that a person's CID is time-constant and will always relate them back to the household they initially belonged to, even if they moved out since.

3 Survey Context

wave – Survey Wave

1	[1] Wave 1, part 1 (2021/22)	13053
2	[2] Wave 1, part 2 (2021/22)	6669

3	[3] Wave 2 (2022/23)	6128
4	[4] Wave 3 (2023)	5562

This variable identifies the (partial) wave in which the data collection took place.

4 Statistical Weighting Factors

design – Inverse Sampling Probability

0	5
70.1461868286133	3
113.673835754395	1
129.841339111328	10
151.398025512695	15
177.21418762207	4
181.577438354492	40
202.015487670898	4
209.343887329102	1
226.846572875977	185
230.22819519043	13
230.803298950195	4
255.753479003906	9
287.660064697266	5
302.295227050781	1000
322.924377441406	6
328.682861328125	13
383.379913330078	49
403.530364990234	55
453.192596435547	6204
459.955841064453	187
537.873657226562	110
574.819702148438	999
766.259521484375	2425
806.560424804688	787
905.884948730469	2402
1149.13916015625	11883
1612.62072753906	481
2297.7783203125	4512

This variable contains the inverse sampling probabilities (design weights) for the SCP sample. They account for the unequal inclusion probabilities resulting from the sampling design.
The SCP has a two-stage probability sampling design. At the first stage, municipalities are sampled (primary sampling units; PSUs) stratified by region and degree of urbanity. At the second stage, individuals are sampled (secondary sampling units; SSUs) from the PSU's population registers. Generally, sampling was conducted proportional to size, except for deliberate oversampling of Eastern Germany. All selected individuals who participated in the

survey were asked to report their household members aged 18 years or older. These household members were subsequently also invited to the surveys. This results in a higher inclusion probability for larger households, which is also accounted for by the design weights. Due to rounding of decimal places, values may be summarized in the codebook.

hhrf – Weighting Factor

89.9704132080078		1
102.496826171875		1
103.33854675293		1
114.508514404297		1
117.915008544922		1
121.074142456055		1
121.479843139648		1
125.444351196289		1
132.868362426758		1
137.444686889648		1
138.32536315918		1
149.049224853516		1
150.570892333984		1
156.400650024414		1
156.875350952148		1
...	(31373 rows omitted)	31382
45724.4709507794		1
45834.347811575		1
45859.7743569513		1
46253.0062740231		1
46675.6061271224		1
46891.2510077426		1
47071.5742033779		1
47231.6730588036		1
48348.0996327961		1
49844.7931102933		1
49918.8162558279		1
50033.0723977376		1
50162.1334203015		1
50745.6955823896		1
53255.1817071805		1

This variable represents the household nonresponse weights for the SCP sample, which serve to mitigate bias due to unit nonresponse. This weighting factor is a combination of the inverse sampling probability, a nonresponse adjustment factor, and an extrapolation towards the survey target population.
The inverse sampling probability (see DESIGN variable) corrects for the unequal selection probabilities in the panel gross sample (e.g. the deliberate oversampling of people in Eastern Germany).
The nonresponse adjustment factor corrects for unit nonresponse. For its computation, survey participation probabilities are estimated from a logistic regression model, incorporating sampling frame data (age groups, gender, German citizenship status, federal states/Länder)

and micro-geographic data to predict response propensities. Missing data in these predictors were handled with multiple imputation. Predictors were selected using a mix of backward and forward selection, with cross-validation mean squared error as the selection criterion.
The extrapolation procedure is based on iterative proportional fitting (aka raking) using Microcensus information on the demographic composition (federal state, municipality size, and household size) of the German population.
The weights for waves from wave 1 part 2 onwards were generated by multiplying the initial household nonresponse weight at recruitment with the anchor person's (AP) inverse participation probability to the according subsequent survey wave. In wave 1 part 2, this was estimated through logistic regression. Predictor variables here also cover survey data from previous waves as well as interaction terms for all variables with respondent type (AP vs. household members). (This model of the staying probability is the same as the one used in phrf for estimating the staying probability if individual respondents.) In subsequent waves (from wave 2 onwards), the AP's participation probability was generated by multiplication based on two models: (1) the probability of a household to still be invited to this survey wave and (b) the AP's probability to respond if invited. As in previous waves, multiple imputation was used to deal with missing data and backward and forward selection was applied to select relevant predictor variables. Subsequently, the weights were raked again using Microcensus information.
Due to rounding of decimal places, values may be summarized in the codebook.

5 Inverse Staying Probability

hbleib – Inverse Staying Probability

0		14042
1.03273797035217		1
1.03302705287933		1
1.03336811065674		1
1.03418147563934		1
1.0357654094696		1
1.03591632843018		1
1.03751683235168		1
1.03825700283051		1
1.0383415222168		1
1.03835797309875		1
1.03866636753082		1
1.03944218158722		1
1.03976380825043		1
1.03996884822845		1
...	(14942 rows omitted)	17341
5.29517936706543		1
5.31301116943359		1
5.42971420288086		1
5.43652248382568		1
5.45314073562622		1
5.51391649246216		1
5.60983037948608		1
5.65867233276367		1

5.72668218612671	1
5.77419185638428	1
5.92092227935791	1
6.20641374588013	1
6.32693386077881	1
6.40488052368164	1
7.49871492385864	1

This variable contains the household inverse staying probability in waves after recruitment, which is equivalent to the anchor person's (AP) inverse staying probability, as modeled through logistic regression. Predictor variables cover survey data from previous waves, including interaction terms with respondent type (AP vs. household member). Missing data in these predictors were handled with multiple imputation. Predictors were selected using a mix of backward and forward selection, with cross-validation mean squared error as the selection criterion.