Research Institute
Social Cohesion

RDC

SOEP

SCP Documentation

German Social Cohesion Panel

SCP 2021/22 W1 Codebook PHRF: Weights for Persons (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., Liebig, S., Minkus, L., Reinecke, J., Richter, D., Teichler, N., Traunmüller, R., & Zinn, S. (2024). SCP 2021/22 W1 - Codebook PHRF: Weights for Persons (English). *German Social Cohesion Panel 2021/22 - Wave 1*. Bremen and Berlin: RDC-RISC/SOCIUM, SOEP/DIW Berlin. doi:10.60532/scp.2021_22.w1.v1

- **Authors:** Olaf Groh-Samberg, Julian B. Axenfeld, Jean-Yves Gerlitz, Carina Cornesse, Martin Kroh, Holger Lengfeld, Stefan Liebig, Lara Minkus, Jost Reinecke, David Richter, Nils Teichler, Richard Traunmüller, Sabine Zinn
- **Contributors:** Cosima Adams, Anton Bochert, Martin Gerike, Josefine Kuhrmeier, Anna-Tabea Müller, Eric Nissen, 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)
  Mohrenstr. 58
  10117 Berlin
  Germany

- **DOI:** 10.60532/scp.2021_22.w1.v1

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

  www.diw.de

SCP Documentation

German Social Cohesion Panel

SCP 2021/22 W1 Codebook PHRF: Weights for Persons (English)

Contents

1 General Information

The PHRF dataset contains survey weights for the individual respondents in the SCP. Each person (PID) who participated in the survey in a particular survey wave (WAVE) has one row in the dataset.
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

pid – Never Changing Person ID

2110000301		2
2110000302		1
2110000901		2
2110000902		2
2110001001		1
2110001201		1
2110001601		2
2110002001		2
2110002101		2
2110003201		1
2110003701		1
2110003901		2
2110004401		1
2110004402		1
2110004403		1
...	(16997 rows omitted)	26150
2113796701		1
2113796702		2
2113796703		2
2113797101		1
2113797102		2
2113797201		1
2113797301		1
2113797601		1
2113797801		2
2113797901		2
2113798501		2
2113798701		1
2113799101		2
2113799102		2
2113800001		1

The central individual identifier across time is PID, which is fixed over time (and of course datasets).

hid – Current Household ID

21100003		3
21100009		4
21100010		1
21100012		1
21100016		2
21100020		2
21100021		2
21100032		1
21100037		1
21100039		2
21100044		3
21100045		1
21100049		2
21100050		1
21100058		1
...	(13023 rows omitted)	26135
21137960		1
21137961		1
21137963		4
21137964		4
21137967		5
21137971		3
21137972		1
21137973		1
21137976		1
21137978		2
21137979		2
21137985		2
21137987		1
21137991		4
21138000		1

This identifier groups all individuals into their respective households at the time of the most recent wave (i.e. 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).

cid – Original Household ID

21100003	3
21100009	4
21100010	1
21100012	1
21100016	2
21100020	2
21100021	2
21100032	1

21100037		1
21100039		2
21100044		3
21100045		1
21100049		2
21100050		1
21100058		1
...	(13023 rows omitted)	26135
21137960		1
21137961		1
21137963		4
21137964		4
21137967		5
21137971		3
21137972		1
21137973		1
21137976		1
21137978		2
21137979		2
21137985		2
21137987		1
21137991		4
21138000		1

This identifier groups individuals into their original households at the start of the panel. That means, 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)	17027
2	[2] Wave 1, part 2 (2021/22)	9168

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

4 Statistical Weighting Factors

design_ap – Inverse Sampling Probability Anchor Persons

0	6623
905.88493688161	4547
905.884936881611	1562
1612.62063492063	917
2297.77831821929	12546

This variable contains the inverse sampling probabilities (design weights) for the initial sample of anchor persons. They account for the unequal inclusion probabilities from the sampling design due to the oversampling of persons in Eastern Germany. These design weights are intended to be used when analyzing only the initial sample of AP without their household members.
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.
Due to rounding of decimal places, values may be summarized in the codebook.

design – Inverse Sampling Probability

70.1461868286133	2
113.673835754395	1
129.841339111328	19
151.398025512695	23
177.21418762207	4
181.577438354492	47
202.015487670898	16
209.343887329102	1
226.846572875977	200
230.22819519043	28
230.803298950195	2
255.753479003906	8
287.660064697266	3
302.295227050781	964
322.924377441406	5
328.682861328125	11
383.379913330078	57
403.530364990234	53
453.192596435547	5246
459.955841064453	217
537.873657226562	108
574.819702148438	1074
766.259521484375	2460
806.560424804688	682
905.884948730469	1525
1149.13916015625	10200
1612.62072753906	311
2297.7783203125	2928

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.

phrf – Weighting Factor

138.076012832869		1
180.275392341203		1
191.814935987255		1
198.18234981009		1
202.519416418032		1
205.922105072994		1
225.538315268451		1
225.66939124434		1
230.998437933074		1
231.575753584113		1
242.620946761035		1
250.467474531022		1
252.662663340351		1
254.528631438953		1
255.406094077172		1
...	(26124 rows omitted)	26165
60540.2697879398		1
60553.0048254172		1
60651.5267019675		1
60926.7044585749		1
61361.9541063377		1
61490.091541196		1
61693.9423795728		1
62316.1498544427		1
63344.4531750494		1
63985.7215170963		1
64328.3586230575		1
64630.1889132602		1
65160.1943149763		1
66936.6426181433		1
67220.5262768564		1

This variable represents the individual 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 initial nonresponse adjustment factor corrects for unit nonresponse. For its computation, survey participation probabilities were estimated from chain of regression models:

1. A logistic regression model of the anchor person's (AP) participation probability, incorporating sampling frame data (age groups, gender, German citizenship status, federal states/Länder) and micro-geographic data to predict response propensities. (This is the same model as the one used for hhrf.) 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.

2. A fractional regression model of the share of household members (HM) named by the AP to participate in the study. This was done to account for underreporting of HM by the AP. Here, in addition to sampling frame and microgeographic data, predictor variables also covered survey data from the AP. As in model (1), missing data was multiply imputed, and backwards and forwards selection was applied to select relevant predictor variables.

3. A logistic regression model of the HM's participation probability. Here, in addition to sampling frame data, microgeographic data, and AP survey data, the predictor variables also covered the HM's age, gender, and relation to the AP as reported by the AP. As in model (1), missing data was multiply imputed, and backwards and forwards selection was applied to select relevant predictor variables.

For AP, the overall participation probability can be derived directly from model (1), while for HM, the overall participation probability is inferred from multiplying the predicted probabilities from model (1) through (3).
The extrapolation procedure is based on iterative proportional fitting (aka raking) using Microcensus information on the demographic composition (age, gender, German citizenship, East vs. West Germany) of the German population.
The weights for waves from wave 1 part 2 onwards were generated by multiplying the initial nonresponse weight at recruitment with the inverse participation probability to the according subsequent survey wave, as estimated through logistic regression. Predictor variables here also cover survey data, including interaction terms for all variables with respondent type (AP vs. HM). (This model of the staying probability is the same as the one used in hhrf for estimating the staying probability of households.) Again, 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

pbleib – Inverse Staying Probability

0	17027
1.11346220970154	1
1.12490344047546	1
1.12930381298065	1
1.13066411018372	1
1.13112485408783	1
1.13122057914734	1
1.13646674156189	1
1.13921761512756	1

1.1408588886261		1
1.14317750930786		1
1.14333009719849		1
1.14349722862244		1
1.14467191696167		1
1.14593815803528		1
...	(8960 rows omitted)	9139
6.88516664505005		1
6.88868188858032		1
6.91688442230225		1
6.92782402038574		1
7.00820970535278		1
7.04780006408691		1
7.04848480224609		1
7.21949863433838		1
7.34550809860229		1
7.46248769760132		1
7.49639701843262		1
7.84853172302246		1
8.08865928649902		1
9.17348957061768		1
10.321307182312		1

This variable contains the individual inverse staying probability in waves after recruitment as modeled through logistic regression. Predictor variables cover survey data from previous waves, including interaction terms with respondent type (anchor person 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.