Research Institute
Social Cohesion

RDC

SOEP

SCP Documentation

German Social Cohesion Panel

SCP 2021-22 W1-2 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., Teichler, N., Traunmüller, R., & Zinn, S. (2026). SCP 2021-22 W1-2 - Codebook PHRF: Weights for Persons (English). *German Social Cohesion Panel 2021-2022 - Wave 1-2*. Bremen and Berlin: RDC-RISC/SOCIUM, SOEP/DIW Berlin. doi:10.60532/scp.2021-22.w1-2.v1

- **Authors:** Olaf Groh-Samberg, Julian B. Axenfeld, Jean-Yves Gerlitz, Carina Cornesse, Martin Kroh, Holger Lengfeld, Stefan Liebig, Lara Minkus, 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)
  Mohrenstr. 58
  10117 Berlin
  Germany
- **DOI:** 10.60532/scp.2021-22.w1-2.v1
- **Website:** www.fgz-risc-data.de

  www.diw.de

SCP Documentation

German Social Cohesion Panel

SCP 2021-22 W1-2 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		3
2110000302		1
2110000901		3
2110000902		3
2110001001		1
2110001201		1
2110001601		3
2110002001		3
2110002101		3
2110003201		1
2110003701		1
2110003901		3
2110004401		1
2110004402		1
2110004403		1
...	(17271 rows omitted)	34779
2113796701		2
2113796702		3
2113796703		3
2113797101		1
2113797102		2
2113797201		1
2113797301		1
2113797601		1
2113797801		2
2113797901		3
2113798501		2
2113798701		1
2113799101		3
2113799102		3
2113800001		1

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

hid – Current Household ID

21100003		4
21100009		6
21100010		1
21100012		1
21100016		3
21100020		3
21100021		3
21100032		1
21100037		1
21100039		3
21100044		3
21100045		1
21100049		3
21100050		1
21100058		1
...	(13029 rows omitted)	34778
21137972		1
21137973		1
21137976		1
21137978		2
21137979		3
21137985		2
21137987		1
21137991		6
21138000		1
22103378		1
22103896		1
22115150		1
22117540		1
22119085		1
22125622		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	4
21100009	6
21100010	1
21100012	1
21100016	3
21100020	3
21100021	3
21100032	1

21100037		1
21100039		3
21100044		3
21100045		1
21100049		3
21100050		1
21100058		1
...	(13023 rows omitted)	34759
21137960		1
21137961		1
21137963		6
21137964		6
21137967		8
21137971		3
21137972		1
21137973		1
21137976		1
21137978		2
21137979		3
21137985		2
21137987		1
21137991		6
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)	17027
2	[2] Wave 1, part 2 (2021/22)	9168
3	[3] Wave 2 (2022/23)	8642

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

4 Statistical Weighting Factors

design_ap – Inverse Sampling Probability AP

0	9391
905.88493688161	5907
905.884936881611	2050
1612.62063492063	1185
2297.77831821929	16304

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

0	274
70.1461868286133	2
113.673835754395	1
129.841339111328	23
151.398025512695	32
177.21418762207	4
181.577438354492	59
202.015487670898	24
209.343887329102	1
226.846572875977	262
230.22819519043	41
230.803298950195	3
255.753479003906	8
287.660064697266	4
302.295227050781	1271
322.924377441406	5
328.682861328125	16
383.379913330078	71
403.530364990234	68
453.192596435547	6914
459.955841064453	278
537.873657226562	142
574.819702148438	1392
766.259521484375	3252
806.560424804688	903
905.884948730469	2026
1149.13916015625	13571
1612.62072753906	400
2297.7783203125	3790

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

107.988438205446		1
110.746671602059		1
118.587694514195		1
121.05651621583		1
128.595096225792		1
129.651130466615		1
134.054318275203		1
135.242228712308		1
141.907949719763		1
144.300750545466		1
146.512111132012		1
151.558361014506		1
154.280512181138		1
157.852979851786		1
158.556929373648		1
…	(34788 rows omitted)	34807
69293.3876233723		1
69427.5129092607		1
69505.9614890907		1
69799.315694785		1
70233.3240777905		1
70274.680222308		1
71393.7605667988		1
72387.8714938801		1
72473.9290001437		1
72478.270015789		1
74533.9073422054		1
78992.6000245832		1
80925.347808723		1
86542.8035181267		1
87190.6729773467		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 also 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, educational attainment) of the German population.

The weights for survey waves from wave 1 part 2 onwards were generated by multiplying the initial non-response weight of the recruitment by the inverse probability of participation in the corresponding survey wave. In wave 1 part 2, a logistic regression model was estimated for this purpose, which models the wave 2 participation probability for all persons who had participated in wave 1 part 1. The predictor variables here also include survey data from the previous waves, including interaction terms for all variables with the respondent type (AP vs. HM). (This model for estimating lead probability is the same as the model used in hhrf for estimating household lead probability). Again, multiple imputation was used to fill in missing data, and relevant predictor variables were selected via backward elimination and forward selection. The weights were then rearranged using microcensus information.

In subsequent waves (from wave 2 onwards), the probability of participation is determined by a series of models (also here with treatment of missing values by multiple imputation and predictor selection by backward elimination and forward selection):

1. A logistic regression model of a household's probability of staying in the panel among all households that had participated in wave 1 part 1

2. A fractional regression model of the proportion of additional household members remaining in the panel among all households that remained in the panel

3. A logistic regression model of the probability of participation of the AP and HM in the panel

4. A logistic regression model of the participation probability of new household members (NHM).

For AP, the combined participation probability is determined from a multiplication of the estimated values from model (1) and model (3), for HM from a multiplication of the estimated values from model (1) to model (3) and for NHM from a multiplication of the estimated values from model (1) to (4). The weights result from multiplying the countervalues of the combined participation probabilities with the weights at the time of recruitment. These weights were also re-raked using microcensus information.
Due to rounding of decimal places, values may be summarized in the codebook.

phrf_ap – Weighting Factor AP

0		9391
919.233618271771		1
938.85815836844		1
965.780871018614		1
991.430262770771		1
992.732641241291		1
1010.33898763984		1
1027.36534489125		1
1044.28171923075		1
1057.81339488129		1
1073.971875086		1
1079.74292691667		1
1089.79833480965		1
1104.04668055725		1
1115.69815578474		1
...	(25412 rows omitted)	25417
94462.1565254744		1
98560.7743045842		1
98958.1678742411		1
100051.991644491		1
104080.903084607		1
106328.437153628		1
106960.174846649		1
107504.714049928		1
107675.727182931		1
109545.700029483		1
113834.686684411		1
115126.652417338		1
117608.203450486		1
123478.657222699		1
124303.053529464		1

This variable contains the individual nonresponse weights for the SCP specifically for the initial random sample of anchor persons (without other household members) for analyses that only refer to anchor persons. This weighting factor is a combination of inverse sampling probability, a nonresponse adjustment factor and an extrapolation to the target population of the survey.

The inverse sampling probability of the anchor persons (see DESIGN_AP variable) corrects for the unequal sampling probabilities in the panel sample due to the intended oversampling of East Germans.

The nonresponse adjustment factor also corrects for unit nonresponse. Logistic regression models of the probability of participation were estimated for this purpose. In wave 1 part 1, sampling frame data and additional microgeographical data were considered as predictors. (This model is the same as the one used in wave 1 part 1 to generate hhrf). From wave 1 part 2 onwards, survey data from previous waves were also taken into account. Missing values in the predictors were completed using multiple imputation. Predictors were selected by a mixture of backward elimination and forward selection using the cross-validation mean squared error as a selection criterion. In wave 1 part 1, the weight is determined by multiplying the inverse estimated response probability by the anchor person design weight (DESIGN_AP). Subsequently, iterative proportional fitting (also known as "raking") was carried out using microcensus data on the demographic composition (age, gender, German citizenship, East vs. West Germany, highest level of education) of the German population. In subsequent waves (from wave 1 part 2), the weight is determined by multiplying the inverse estimated participation probability by the non-response weight from wave 1 part 1. These weights were also re-raked using microcensus information in each wave.

Due to the rounding of decimal places, values may be summarized in the codebook.

5 Inverse Staying Probability

pbleib – Inverse Staying Probability

0		17869
1.02874624729156		1
1.0300555229187		1
1.03659904003143		1
1.03835904598236		1
1.03903603553772		1
1.03943228721619		1
1.03991448879242		1
1.04124534130096		1
1.0427041053772		1
1.04332339763641		1
1.04353272914886		2
1.04391014575958		1
1.04431486129761		1
1.04558312892914		1
...	(14510 rows omitted)	16938
5.31301116943359		1
5.35156965255737		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 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.