Research Institute
Social Cohesion

RDC

SOEP

SCP Documentation

German Social Cohesion Panel

SCP 2021-22 W1-2 Codebook PHRF: Weights for Persons (German)

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 (German). *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 (German)

Inhaltsverzeichnis

1 Allgemeine Informationen

Der PHRF-Datensatz enthält Gewichtungsfaktoren für die einzelnen Befragten im SCP. Jede Person (PID), die in einer bestimmten Erhebungswelle (WAVE) an der Befragung teilgenommen hat, hat eine Zeile im Datensatz.
An einigen Stellen in der Dokumentation und in den Daten werden Jahreszahlen z. B. für die Bezeichnung von Variablen und des Fragebogeninstruments verwendet. Diese Jahreszahlen orientieren sich stets am Feldstart der Datenerhebung der entsprechenden Erhebungswelle.

2 Identifikatoren

pid – Unveränderliche Personen-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 Zeilen unterdrückt)	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

Die PID ist die unveränderliche Kennziffer einer Person, die über die Zeit und über alle Datensätze identisch gehalten wird.

hid – Aktuelle Haushalts-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 Zeilen unterdrückt)	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

Diese Kennziffer gruppiert Individuen in ihre zugehörigen Haushalte zum Zeitpunkt der aktuellen Erhebungswelle. Das bedeutet, dass die HID einer Person sich über die Zeit verändern kann, zum Beispiel wenn ein erwachsenes Kind aus dem elterlichen Haushalt auszieht und einen eigenen Haushalt eröffnet.

cid – Ursprüngliche Haushalts-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 Zeilen unterdrückt)	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

Diese Kennziffer gruppiert Individuen in ihre Ursprungshaushalte zu Beginn des Panels. Das bedeutet, dass die CID einer Person zeitkonstant gehalten wird und sie immer mit dem Haushalt verbinden wird, zu dem sie initial gehört hat, selbst wenn sie seitdem den Haushalt gewechselt hat.

3 Befragungskontext

wave – Erhebungswelle

1	[1] Welle 1, Teil 1 (2021/22)	17027
2	[2] Welle 1, Teil 2 (2021/22)	9168
3	[3] Welle 2 (2022/23)	8642

Diese Variable identifiziert die (Teil-)Welle, in der die Datenerhebung stattgefunden hat.

4 Statistische Gewichtungsfaktoren

design_ap – Inverse Ziehungswahrscheinlichkeit AP

0	9391
905.88493688161	5907
905.884936881611	2050

1612.62063492063	1185
2297.77831821929	16304

Diese Variable enthält die inversen Ziehungswahrscheinlichkeiten (Design-Gewichte) für die initiale Stichprobe von Ankerpersonen. Das Design-Gewicht berücksichtigt die aus dem Stichprobenziehungsdesign resultierenden ungleichen Ziehungswahrscheinlichkeiten, die aus der überproportionalen Ziehung von Personen in Ostdeutschland resultieren. Diese Design-Gewichte sollen für Analysen verwendet werden, die nur die initiale Ankerpersonenstichprobe ohne die weiteren Haushaltsmitglieder nutzen.
Das SCP hat ein zweistufiges Stichprobenziehungsverfahren. Auf der ersten Stufe werden, stratifiziert nach Region und Urbanitätsgrad, Gemeinden gezogen (primary sampling units; PSUs). Auf der zweiten Stufe werden Personen (secondary sampling units; SSUs) aus den Registern dieser Gemeinden gezogen. Generell erfolgte die Stichprobenziehung proportional zur Gemeindegröße. Eine Ausnahme ist die beabsichtigte überproportionale Ziehung in Ostdeutschland.
Aufgrund der Rundung von Nachkommastellen kann es im Codebuch zu einer Zusammenfassung von Werten kommen.

design – Inverse Ziehungswahrscheinlichkeit

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

Diese Variable enthält die inversen Ziehungswahrscheinlichkeiten (Design-Gewichte) für die SCP-Stichprobe. Das Design-Gewicht berücksichtigt die aus dem Stichprobenziehungsdesign resultierenden ungleichen Ziehungswahrscheinlichkeiten.
Das SCP hat ein zweistufiges Stichprobenziehungsverfahren. Auf der ersten Stufe werden, stratifiziert nach Region und Urbanitätsgrad, Gemeinden gezogen (primary sampling units; PSUs). Auf der zweiten Stufe werden Personen (secondary sampling units; SSUs) aus den Registern dieser Gemeinden gezogen. Generell erfolgte die Stichprobenziehung proportional zur Gemeindegröße. Eine Ausnahme ist die beabsichtigte überproportionale Ziehung in Ostdeutschland. Alle in die Stichprobe gezogenen Personen, die an der Befragung teilnehmen, wurden gebeten ihre weiteren volljährigen Haushaltsmitglieder anzugeben. Diese weiteren Haushaltsmitglieder werden dann ebenfalls befragt. Das führt zu einer höheren Ziehungswahrscheinlichkeit für größere Haushalte, was ebenso durch die Design-Gewichte berücksichtigt wird.
Aufgrund der Rundung von Nachkommastellen kann es im Codebuch zu einer Zusammenfassung von Werten kommen.

phrf – Hochrechnungsfaktor

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 Zeilen unterdrückt)	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

Diese Variable enthält die individuellen Nonresponse-Gewichte für das SCP, die zur Reduzierung von Verzerrungen durch Unit-Nonresponse dienen. Dieser Gewichtungsfaktor ist eine Kombination aus inverser Stichprobenziehungswahrscheinlichkeit, einem Nonresponse-Adjustierungsfaktor und einer Extrapolation zur Zielpopulation der Befragung.
Die inverse Stichprobenziehungswahrscheinlichkeit (siehe DESIGN-Variable) korrigiert für die ungleichen Ziehungswahrscheinlichkeiten in die Panel-Stichprobe (z.B. durch das beabsichtigte Über-Ziehen von Ostdeutschen).
Der Nonresponse-Adjustierungsfaktor korrigiert zudem für Unit-Nonresponse. Für seine Berechnung wurden Teilnahmewahrscheinlichkeiten durch eine Verkettung mehrerer Regressionsmodelle geschätzt:

1. Ein logistisches Regressionsmodell der Teilnahmewahrscheinlichkeit der Ankerperson (AP), in der Sampling-Frame-Daten und zusätzliche mikrogeographische Daten als Prädiktoren berücksichtigt wurden. (Dieses Modell ist dasselbe wie das, das zur Generierung von hhrf genutzt wurde.) Fehlende Werte in den Prädiktoren wurden mittels multipler Imputation vervollständigt. Prädiktoren wurden durch eine Mischung aus Rückwärtselimination und Vorwärtsauswahl unter Verwendung des Kreuzvalidierungs-Mean-Squared-Errors als Selektionskriterium ausgewählt.

2. Ein fraktionelles Regressionsmodell des Anteils der Haushaltsmitglieder (HM), die durch die AP zur Teilnahme an der Studie genannt wurden. Dadurch wird eine Untererfassung der HM durch die zugehörige AP berücksichtigt. Hierbei werden zusätzlich zu Sampling-Frame- und mikrogeographischen Daten auch Befragungsdaten der AP als Prädiktoren genutzt. Wie in Modell (1) wurden fehlende Werte durch multiple Imputation ergänzt und relevante Prädiktorvariablen wurden via Rückwärtselimination und Vorwärtsauswahl ausgewählt.

3. Ein logistisches Regressionsmodell der Teilnahmewahrscheinlichkeit des HM. Hierbei wurden zusätzlich zu Sampling-Frame-Daten, mikrogeographischen Daten sowie Befragungsdaten der Ankerperson auch das Alter des HM, das Geschlecht des HM sowie die Beziehung des HM zur AP als Prädiktoren genutzt. Wie in Modell (1) wurden fehlende Werte durch multiple Imputation ergänzt und relevante Prädiktorvariablen wurden via Rückwärtselimination und Vorwärtsauswahl ausgewählt.

Für AP lässt sich die Teilnahmewahrscheinlichkeit direkt aus Modell (1) ableiten, während sich die Teilnahmewahrscheinlichkeit für HM aus der Multiplikation der vorhergesagten Wahrscheinlichkeiten der Modelle (1) bis (3) zusammensetzt.
Die Extrapolation basiert auf iterativem proportionalem Fitting (auch als “Raking” bezeichnet) mittels Mikrozensus-Daten zur demographischen Zusammensetzung (Alter, Geschlecht, deutsche Staatsbürgerschaft, Ost- vs. Westdeutschland, höchster Bildungsabschluss) der deutschen Bevölkerung.
Die Gewichte für Erhebungswellen ab Welle 1 Teil 2 wurden durch Multiplikation des initialen Nonresponse-Gewichts der Rekrutierung mit der inversen Teilnahmewahrscheinlichkeit an der entsprechenden Erhebungswelle generiert. In Welle 1 Teil 2 wurde hierfür ein logistisches Regressionsmodell geschätzt, das die Welle-2-Teilnahmewahrscheinlichkeit für alle Personen modelliert, die an Welle 1 Teil 1 teilgenommen hatten. Die Prädiktorvariablen umfassen hier zusätzlich Umfragedaten der Vorwellen, einschließlich Interaktionsterme für alle Variablen mit dem Befragtentyp (AP vs. HM). (Dieses Modell zur Schätzung der Bleibewahrscheinlichkeit ist dasselbe wie das Modell, das in hhrf zur Schätzung der Bleibewahrscheinlichkeit der Haushalte verwendet wurde.) Auch hier wurde Multiple Imputation

angewandt, um fehlende Daten zu vervollständigen, und relevante Prädiktorvariablen wurden via Rückwärtselimination und Vorwärtsauswahl ausgewählt. Anschließend wurden die Gewichte unter Verwendung von Mikrozensus-Informationen erneut geraket.
In Folgewellen (ab Welle 2) wird die Teilnahmewahrscheinlichkeit durch eine Reihe von Modellen bestimmt (auch hier jeweils mit Behandlung fehlender Werte durch Multiple Imputation und Prädiktorenauswahl durch Rückwärtselemination und Vorwärtsauswahl):

1. Ein logistisches Regressionsmodell der Verweilwahrscheinlichkeit eines Haushalts im Panel unter allen Haushalten, die an Welle 1 Teil 1 teilgenommen hatten
2. Ein fraktionelles Regressionsmodelldes Anteil im Panel verbleibender weiterer Haushaltsmitglieder in allen Haushalten, die im Panel verblieben sind
3. Ein logistisches Regressionsmodell der Teilnahmewahrscheinlichkeit der AP und HM im Panel
4. Ein logistisches Regressionsmodell der Teilnahmewahrscheinlichkeit der Neuen Haushaltsmitglieder (NHM).

Für AP bestimmt sich die kombinierte Teilnahmewahrscheinlichkeit aus einer Multiplikation der Schätzwerte aus Modell (1) und Modell (3), bei HM aus einer Multiplikation der Schätzwerte aus Modell (1) bis Modell (3) und für NHM aus einer Multiplikation der Schätzwerte aus Modell (1) bis (4). Die Gewichte ergeben sich aus der Multiplikation der Gegenwerte der kombinierten Teilnahmewahrscheinlichkeiten mit den Gewichten zum Rekrutierungszeitpunkt. Auch diese Gewichte wurden unter Verwendung von Mikrozensus-Informationen erneut geraket.
Aufgrund der Rundung von Nachkommastellen kann es im Codebuch zu einer Zusammenfassung von Werten kommen.

phrf_ap – Hochrechnungsfaktor 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 Zeilen unterdrückt)	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

Diese Variable enthält die individuellen Nonresponse-Gewichte für das SCP speziell für die initiale Zufallsstichprobe der Ankerpersonen (ohne weitere Haushaltsmitglieder) für Analysen, die sich nur auf Ankerpersonen beziehen. Dieser Gewichtungsfaktor ist eine Kombination aus inverser Stichprobenziehungswahrscheinlichkeit, einem Nonresponse-Adjustierungsfaktor und einer Extrapolation zur Zielpopulation der Befragung.
Die inverse Stichprobenziehungswahrscheinlichkeit der Ankerpersonen (siehe DESIGN_AP-Variable) korrigiert für die ungleichen Ziehungswahrscheinlichkeiten in die Panel-Stichprobe durch das beabsichtigte Überziehen von Ostdeutschen.
Der Nonresponse-Adjustierungsfaktor korrigiert zudem für Unit-Nonresponse. Hierfür wurden logistische Regressionsmodelle der Teilnahmewahrscheinlichkeit geschätzt. In Welle 1 Teil 1 wurden dabei Sampling-Frame-Daten und zusätzliche mikrogeographische Daten als Prädiktoren berücksichtigt. (Dieses Modell ist dasselbe wie das, das in Welle 1 Teil 1 zur Generierung von hhrf genutzt wurde.) Ab Welle 1 Teil 2 wurden zudem Befragungsdaten aus Vorwellen berücksichtigt. Fehlende Werte in den Prädiktoren wurden mittels multipler Imputation vervollständigt. Prädiktoren wurden durch eine Mischung aus Rückwärtselimination und Vorwärtsauswahl unter Verwendung des Kreuzvalidierungs-Mean-Squared-Errors als Selektionskriterium ausgewählt. Das Gewicht bestimmt sich in Welle 1 Teil 1 aus der Multiplikation der inversen geschätzten Teilnahmewahrscheinlichkeit mit dem Ankerpersonen-Designgewicht (DESIGN_AP). Anschließend wurde eine Randanpassung basierend auf iterativem proportionalem Fitting (auch als "Raking" bezeichnet) mittels Mikrozensus-Daten zur demographischen Zusammensetzung (Alter, Geschlecht, deutsche Staatsbürgerschaft, Ost- vs. Westdeutschland, höchster Bildungsabschluss) der deutschen Bevölkerung durchgeführt. In Folgewellen (ab Welle 1 Teil 2) bestimmt sich das Gewicht aus der Multiplikation der inversen geschätzten Teilnahmewahrschienlichkeit mit dem Nonresponse-Gewicht aus Welle 1 Teil 1. Auch hier wurde jeweils eine erneute Randanpassung vorgenommen.
Aufgrund der Rundung von Nachkommastellen kann es im Codebuch zu einer Zusammenfassung von Werten kommen.

5 Inverse Bleibewahrscheinlichkeit

pbleib – Inverse Bleibewahrscheinlichkeit

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 Zeilen unterdrückt)	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

Diese Variable enthält die individuelle inverse Bleibewahrscheinlichkeit in den Wellen nach der Rekrutierung entsprechend einer Modellierung durch logistische Regression. Die Prädiktorvariablen umfassen Erhebungsdaten aus früheren Wellen, einschließlich Interaktionsterme mit dem Befragungstyp (Ankerperson vs. Haushaltsmitglied). Fehlende Werte in den Prädiktoren wurden mittels multipler Imputation vervollständigt. Prädiktoren wurden durch eine Mischung aus Backward- und Forward-Selection unter Verwendung des Kreuzvalidierungs-Mean-Squared-Errors als Selektionskriterium ausgewählt.