1192

2022

SOEP Survey Papers

Series G - General Issues and Teaching Materials

SOEPcompanion (v37), V.3

Selin Kara, Stefan Zimmermann, and SOEP Group

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:

Selin Kara, Stefan Zimmermann, and SOEP Group. 2022. SOEPcompanion (v37), V.3. SOEP Survey Papers 1192: Series G. Berlin: DIW/SOEP.

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

SOEPcompanion

Release 2022, v.3

Selin Kara, Stefan Zimmermann, and SOEP Group

Sep 26, 2022

CONTENTS

CHAPTER

ONE

PREFACE

SOEP-Core is the centerpiece of the Socio-Economic Panel, a wide-ranging representative longitudinal study of private households in Germany, based at the German Institute for Economic Research, DIW Berlin. SOEP-Core was started in 1984, and in 1990—shortly after German reunification—it was enlarged to include a representative sample from East Germany. This feature makes the SOEP unique among household panel surveys worldwide. Every year since 1984, individuals in households have been surveyed by the SOEP's fieldwork organization, infas Institut für angewandte Sozialwissenschaften GmbH. The data provide information on every member of every household taking part in the survey. Respondents include Germans living in both the former East and West Germany, foreign citizens residing in Germany, recent immigrants, and a new sample of refugees added in 2016. Some of the many topics include household composition, education, occupational biographies, employment, earnings, health, and satisfaction indicators.

The SOEPcompanion describes the current version of the SOEP-Core data (v37) and introduces users to the different SOEP-Core data structures. It also provides applications in Stata as well as instructions on how to use our various documentation services. We plan to revise the information in the SOEPcompanion annually to continue providing users a comprehensive, up-to-date introductory understanding of the SOEP.

We know that starting to use any new dataset is difficult, and this is especially true of panel data given their complexity. We hope that this introduction will help. We always welcome any feedback or tips on how to improve our documentation.

- Recommendation of our most recent version of a general short description of SOEP study: The German Socio-Economic Panel Study (SOEP)
- To the information system for efficient working with complex datasets: paneldata.org

CHAPTER

TWO

TOPICS OF SOEP-CORE

The topics of the SOEP questionnaires and the various modules they contain can be grouped into 11 areas. Some of the modules deal with aspects of life that tend to change from one year to the next, and are therefore repeated annually, while other modules are repeated at intervals of several years. How often a module is repeated is stated in the "module" column of our topic tables. Some SOEP modules are also adapted in different ways to the different questionnaires. The questions in the "Big Five" personality traits module, for instance, are formulated differently in the mother-child questionnaires than they are in the individual questionnaire.

Overview of Modules in Different SOEP Questionnaires

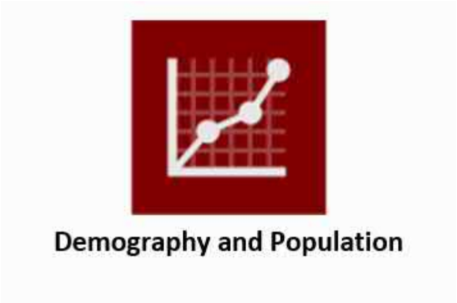
	Individual	*Youth*	*Mother-Child A*	*Mother-Child B*	*Mother-Child C*	*Parents D*	*Mother-Child E*
Affective well-being	x	x					
Big Five personality traits	x	x		x	x		x
Birth history		x	x	x	x	x	x
Childcare			x	x	x	x	x
Educational aspirations		x				x	x
Health of child			x	x	x		x
Height and weight of child			x	x	x		
Height and weight	x	x					
Language ability German / native language	x	x					
Leisure and activities (with child)				x	x		
Life satisfaction	x	x					
Language use				x			x
Locus of control	x	x					
Family background	x	x					
Parental interest in school performance		x					x
Allowance		x					x
Political orientation	x	x					
Risk aversion	x	x					
State of health	x	x					
Strengths and difficulties					x		x
Temperament			x	x			

The SOEP Scales Manual briefly describes the theoretical background and development of all of the scales used in the Socio-Economic Panel (SOEP) study. It also provides the relevant citations as well as the items belonging to the scales and the answer format, including the verbal anchors.

Note: The topic tables list the modules in a questionnaire, not the question items. The modules listed do not represent all of the variables in the SOEP, nor do they refer to all questionnaires. Specific information can also be found in our generated datasets.

2.1 Demography and Population

The demography and population topic provides information including the birth date and sex of each household member (including children) and of each interviewer; the place and history of births in the household; household size; and relationships among household members.



Questionnaire	Module	Years	Variables
Individual Questionnaire	Country of Origin, Acquisition Date German Citizenship	[2002-2020], [2002-2012,2014-2020], [2013]	plj0024_h, plj0024_v1, plj0024_v2
	Country of Origin, Birthcountry Parents	[2012-2018,2020]	plj0175
	Country of Origin, Citizenship	[1984-1993,1996-2020], [1984-1995]	plj0014_v1, plj0014_v2
	Country of Origin, German Citizenship	[1996-2020]	plj0014_v3
	Country of Origin, Second Citizenship	[2000-2020]	plj0022, plj0023_v1
	Country of Origin, Second/Other Citizenship	[2020]	plj0023_v2
	Gender Identification	[2019]	pla0047
Youth Questionnaire	Birth history, Birth date	[2000-2020], [2003-2020]	jl0233, jl0234
	Birth history, Born in Germany	[2000-2020], [2000-2005,2011-2020], [2006-2010]	jl0235_h, jl0235_v1, jl0235_v2
	Birth history, Country of Birth	[2000-2020]	jl0238
	Birth history, Year of Migration	[2000-2018]	jl0239
	Country of Origin, Acquisition Date German Citizenship	[2012-2018]	jl0419
	Country of Origin, German Citizenship	[2006-2018]	jl0241, jl0244
	Country of Origin, Nationality	[2006-2018]	jl0245
	Country of Origin, Residence status	[2000-2013], [2014-2018]	jl0246, jl0445
	Country of Origin, Second Citizenship	[2006-2018]	jl0242, jl0243

continues on next page

Table 1 – continued from previous page

Questionnaire	Module	Years	Variables
	Country of Origin, Status upon Migration	[2000-2018]	jl0240
Mother and Child Instruments	Birth history, Birth date	[2003-2020]	birthm, birthy

2.2 Work and Employment

The work and employment modules provide information on diverse employment-related topics including the respondent's first job, further training, changes in working conditions following parenthood, part-time work, and unemployment. Modules cover not only objective information such as working hours but also subjective perceptions of working conditions and feelings about work.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Care period (Pflegezeit)	[2011-2019], [2011-2014,2020], [2015-2019]	plb0020_h, plb0020_v1, plb0020_v2
	Change of job	[1985-2020], [1985-1993], [1994-2020]	plb0031_h, plb0031_v1, plb0031_v2
	Change of job, Early Retirement	[2011-2020]	plb0480
	Change of job, Interruption due to Children	[2010-2020]	plb0295
	Change of job, Month Change in Company	[1985-1993]	plb0033_v5
	Change of job, Month First Job	[1985-1993]	plb0033_v1
	Change of job, Month Position taken up	[1994-2020]	plb0033_v7
	Change of job, Month Self Employed	[1985-1993]	plb0033_v4
	Change of job, Month Taken Over	[1991-1993]	plb0033_v6
	Change of job, Month Work resumed	[1985-1993]	plb0033_v2
	Change of job, Month new Employer	[1985-1993]	plb0033_v3

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Change of job, Number of Changes	[2012-2020]	plb0478, plb0479
	Change of job, Start Current Position non-response	[2004-2020]	plb0034
	Change of job, Start new Position	[1994-2020]	plb0032
	Change of job, Type of Change	[1994-2020], [1994-2004], [2005-2020]	plb0284_h, plb0284_v1, plb0284_v2
	Changes in workplace tools / technologies	[2015-2018]	plb0595
	Changes in workplace tools / technologies, Influnces	[2015-2018]	plb0596, plb0597, plb0598, plb0599, plb0600
	Commuter Module, Distance	(irregular) [1985-2019]	plb0158
	Commuter Module, Frequency	[2015,2017,2019]	plb0591
	Commuter Module, No Cummute	[1985,1990], [1993,1995,1997-1999], [2000-2013,2015,2017,2019]	plb0159_v1, plb0159_v2, plb0159_v3
	Commuter Module, Second Residence	[2015,2017,2019]	plb0589, plb0590
	Commuter Module, Time	[2015,2017,2019]	plb0592
	Contract to Provide Specific Services (Werkvertrag)	[2013,2015]	plb0482
	Contractual working hours, Actual working hours	[1984-2020], [1984-1990], [1990-2019]	plb0186_h, plb0186_v1, plb0186_v2
	Contractual working hours, Desired working hours	[1985-1995,1997-2019], [1985-1995,1997-1999,2016], [2000-2015,2017-2019]	plb0241_h, plb0241_v1, plb0241_v2
	Contractual working hours, Not fixed	[1984-1986,1990-2020], [1987-1989]	plb0185_v1, plb0185_v2
	Contractual working hours, Working days	(irregular) [1990-2020]	plb0209, plb0210
	Current job, Civil Servant	[1984-2020]	plb0065
	Current job, Civil Service	[1984-2020]	plb0040
	Current job, Employment Agency	[2001-2020]	plb0041
	Current job, Fixed-Term Employment	[1985-2020], [1985-1999], [2000-2005], [2006-2020]	plb0037_h, plb0037_v1, plb0037_v2, plb0037_v3
	Current job, Industrial Worker	[1984-2020]	plb0058
	Current job, Month current Position	[1984-1988,1990-1993,1996-2020]	plb0035
	Current job, Number of Employees	[1984-1990], [1990], [1991-1998], [1999-2004], [2005-2015]	plb0049_v1, plb0049_v2, plb0049_v3, plb0049_v4, plb0049_v5
	Current job, Number of Employees Self Employed	[1984-2020], [2014-2020]	plb0057_h2, plb0057_v8

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Current job, Occupational Status	[2014-2020], [2019]	plb0568_v1, plb0568_v2
	Current job, Salaried employees	[1984-1990], [1991-2018]	plb0064_v1, plb0064_v2
	Current job, Self Employed	[1984-2020], [2014-2020]	plb0057_h1, plb0057_v1, plb0057_v2, plb0057_v3, plb0057_v4, plb0057_v5, plb0057_v6, plb0057_v7, plb0057_v9
	Current job, Trainee/Intern	[1984-1999], [1990]	plb0063_v1, plb0063_v2
	Current job, Working at current Employer	[1984-1988,1990-2020], [1984-1988,1990-1998], [1999-2020]	plb0036_h, plb0036_v1, plb0036_v2
	Employment / education calendar, Full-Time Employment	[1984], [1985-1989,1991-2020]	pab0001_v1, pab0001_v2, pab0001_v3
	Employment / education calendar, Further Education	[2000-2020], [2020]	pab0010_v1, pab0010_v2
	Employment / education calendar, Homemaker	[1984-2020]	pab0008
	Employment / education calendar, In School	[1984-2020]	pab0013
	Employment / education calendar, Maternity Leave	[1996-2020]	pab0006
	Employment / education calendar, Military/Civil Service	[1984-2020]	pab0007
	Employment / education calendar, Mini-Job	[2005-2020]	pab0011
	Employment / education calendar, Other	[1984-2020]	pab0012
	Employment / education calendar, Part-Time Employment	[1985-2020]	pab0002
	Employment / education calendar, Retired	[1984-2020]	pab0005
	Employment / education calendar, Short-Time Work	[2010-2014,2017-2018]	pab0003_v1
	Employment / education calendar, Unemployed	[1984-2020]	pab0004
	Employment / education calendar, Vocational Training	[2020]	pab0003_v2
	Employment status	[1984-2020], [1984], [1985-1990], [1990], [1991-1995], [1996-1998], [1999], [2000-2001], [2002-2015], [2016-2020]	plb0022_h, plb0022_v1, plb0022_v2, plb0022_v3, plb0022_v4, plb0022_v5, plb0022_v6, plb0022_v7, plb0022_v8, plb0022_v9
	Employment, October 2014	[2015]	plb0574
	Entitlement to paid breaks	[2015-2018]	plb0601, plb0602, plb0603

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Evening and weekend work, Evening	[1990], (irregular) [1995-2019], [2012], [2013]	plb0205_v1, plb0205_v2, plb0205_v3, plb0205_v4
	Evening and weekend work, Night	[1990], (irregular) [1995-2019], [2012], [2013]	plb0206_v1, plb0206_v2, plb0206_v3, plb0206_v4
	Evening and weekend work, Saturday	(irregular) [2005-2019]	plb0218
	Evening and weekend work, Sunday	(irregular) [2005-2019]	plb0219
	Financial compensation for overtime	[1984-1986,1988-2014,2018,2020], [1984-1986,1988-1995], [1996], [1997-2014,2018,2020]	plb0195_h, plb0195_v1, plb0195_v2, plb0195_v3
	Gross / net income, October 2014	[2015]	plb0584, plb0585
	Industry sector, occupational classification	[2013-2020], [1990-1993], [1999-2020]	p_isco08, p_nace, plb0072_v1, plb0072_v2, plb0072_v3, plb0073_h, plb0073_v1, plb0073_v2, plb0073_v3, plb0073_v4, plb0073_v5
	Job search	[1985-2020]	plb0358_h
	Job search, Active Search	[1989-2020]	plb0362
	Job search, Applied on Speculation	[1989-1998]	plb0358_v8
	Job search, Friends / Acquaintances	[1985-1998], [1985-1988]	plb0358_v3, plb0358_v5
	Job search, Job Centre	[1985-1998]	plb0358_v1
	Job search, Learn about current Position	[1999-2002], [2003-2013], [2014], [2015-2020]	plb0358_v10, plb0358_v11, plb0358_v12, plb0358_v13
	Job search, Newspaper	[1985-1998]	plb0358_v2
	Job search, Offer within Company	[1985-1988]	plb0358_v4
	Job search, Other	[1989-1998]	plb0358_v7
	Job search, Private Agent	[1995-1998]	plb0358_v9
	Job search, Self Employed	[1985-1998]	plb0358_v6
	Job search, motives	(irregular) [1994-2017]	plb0111
	Job search, preferences	(irregular) [1994-2017]	plb0426
	Labor intensity	[2015], [2016-2018], [2015-2018]	plb0593_v1, plb0593_v2, plb0594
	Leaving a job	[1985-2020], [1985-2000], [2001-2020]	plb0282_h, plb0282_v1, plb0282_v2
	Leaving a job, Abandonment of own business	[1985-1998]	plb0304_v8
	Leaving a job, Closure of operations	[1991-1998]	plb0304_v11
	Leaving a job, Compensation	[1991-2020], [1991-2001], [2002-2020]	plc0040, plc0041_h, plc0041_v1, plc0041_v2
	Leaving a job, Early Retirement	[1987-1998]	plb0304_v10

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Leaving a job, End Fixed-Term Contract	[1985-1998]	plb0304_v2
	Leaving a job, End Vocational Training	[1985-1998]	plb0304_v3
	Leaving a job, Exempted	[1991-1998]	plb0304_v12
	Leaving a job, Month	[1985-2020]	plb0298, plb0299
	Leaving a job, Months Worked	[1985-2020]	plb0302
	Leaving a job, Mutually agreed dissolution	[1985-1990]	plb0304_v5
	Leaving a job, Non-response	[2004-2020]	plb0300
	Leaving a job, Other	[1985-1998]	plb0304_v9
	Leaving a job, Own Resignation	[1985-1998]	plb0304_v4
	Leaving a job, Perspective after Leaving	[1999], [2000-2020]	plb0305_v1, plb0305_v2
	Leaving a job, Retirement	[1991-1998]	plb0304_v15
	Leaving a job, Termination by Employer	[1985-1998]	plb0304_v1
	Leaving a job, Transfer	[1985-1998]	plb0304_v7
	Leaving a job, Transfer at own request	[1985-1998]	plb0304_v6
	Leaving a job, Type of Leaving	[1985-2020], [1999-2000], [2001-2020]	plb0304_h, plb0304_v13, plb0304_v14
	Leaving a job, Years Worked	[1985-2020]	plb0301
	Maternity / parental leave	[1999-2000], [2001-2020]	plb0019_v1, plb0019_v2
	Occupational expectations, non-employed	(irregular) [1999-2020]	plb0427, plb0428, plb0429
	Overtime, October 2014	[2015]	plb0582, plb0583
	Paid breaks, October 2014	[2015]	plb0575
	Paid breaks, October 2015	[2015]	plb0576
	Paid breaks, October 2016	[2015]	plb0577
	Paid breaks, October 2017	[2015]	plb0578
	Performance evaluation by superior	[2004,2008,2011,2016]	plb0098, plb0099, plb0100, plb0101, plb0102
	Professional expectations	[1985,1987,1989-1994,1996,1998]	plb0432_v1, plb0433_v1, plb0434_v1, plb0435_v1, plb0436_v1, plb0437_v1, plb0438_v1, plb0439_v1, plb0440_v1, plb0441_v1, plb0442_v1
	Professional expectations, next two years	(irregular) [1999-2018]	plb0432_v2, plb0433_v2, plb0434_v2, plb0435_v2, plb0436_v2, plb0437_v2, plb0438_v2, plb0439_v2, plb0440_v2, plb0441_v2, plb0442_v2
	Registered unemployed	[1985-2020]	plb0021
	Self-employment, reasons	[2010,2015]	plb0333, plb0334, plb0335, plb0336, plb0337, plb0338

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Short-time compensation (Kurzarbeitergeld)	[1984-2001,2003-2005,2010-2011], [1984-2001,2003-2005], [2010-2011], [1984], [1985-2001,2003-2005,2010-2011]	plc0057_h, plc0057_v1, plc0057_v2, plc0058_v1, plc0058_v2
	Side jobs	[1998-2007], [1998]	plb0382_h, plb0382_v1
	Side jobs, Agriculture	[1999-2007]	plb0382_v2
	Side jobs, Construction	[1999-2007]	plb0382_v3
	Side jobs, Days	[1985-2016]	plb0396
	Side jobs, Gross Income	[1995-2016], [1995-2001], [2002-2016]	plc0062_h, plc0062_v1, plc0062_v2
	Side jobs, Helping Family Members out	[1986-2016]	plb0392
	Side jobs, Hours per Month	[1985-2014]	plb0397
	Side jobs, Hours per Week	[2015-2016]	plb0573
	Side jobs, Industrial Sector	[1999-2007]	plb0382_v4
	Side jobs, Iregual	[1985-2016]	plb0395
	Side jobs, Months	[2000-2013]	plb0398
	Side jobs, Occupational Classification ISCO08	[2013-2016]	p_isco08_sidejob, p_isco08_sidejob1, p_isco08_sidejob2, p_isco08_sidejob3
	Side jobs, Occupational Classification ISCO88	[1991-2016]	p_isco88_sidejob, p_isco88_sidejob1, p_isco88_sidejob2, p_isco88_sidejob3
	Side jobs, Other	[1985-2016]	plb0393
	Side jobs, Regular	[1985-2016]	plb0394
	Side jobs, Service Sector	[1999-2007]	plb0382_v5
	Standby duty	[2011,2014-2019]	plb0212, plb0213, plb0214, plb0215
	Start of working hours	(irregular) [2002-2019]	plb0180, plb0181, plb0182
	Starting a new job, Acceptable Position	[1984-2020]	plb0423
	Starting a new job, Active Job Search	[1994-1998], [1999-2020]	plb0424_v1, plb0424_v2
	Starting a new job, Desired Employment Type	[1984-2020]	plb0240
	Starting a new job, Expected Minimum Income	[1987-1989,1992-1994,1996-2001], [2002-2020]	plb0420_v1, plb0420_v2
	Starting a new job, Intention	[1984-1993], [1994-2020]	plb0417_v1, plb0417_v2
	Starting a new job, Non-response Salary	[1987-1989,1992-1994,1996-2020]	plb0421
	Starting a new job, Number of Hours	[2007-2020]	plb0422
	Starting a new job, Suitable Job	[1987-2020], [1987-2002], [2003-2020]	plb0419_h, plb0419_v1, plb0419_v2
	Starting a new job, Timing	[1984-2020]	plb0418

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Supervisory position	(irregular) [2007-2019], [2007,2009,2011,2013,2015,2017]	plb0067, plb0068, plb0069
	Use of professional skills in job	[1985-2007,2009]	plb0357
	Vacation entitlement, Carried over Vaccation	[2005,2010]	plb0275, plb0276
	Vacation entitlement, Contracted Days	[2000,2005,2010]	plb0269
	Vacation entitlement, Days on Vaccation	[1985-1990,2000,2005,2010]	plb0265
	Vacation entitlement, Expired Vaccation	[2005,2010]	plb0273, plb0274
	Vacation entitlement, Not specified	[2005,2010]	plb0270, plb0272
	Work council (Betriebsrat)	[2001,2006,2011,2016,2019]	plb0050
	Work from home	[1997,1999,2002,2009-2014,2020], (irregular) [1997-2020], [2012], [2013]	plb0095, plb0096_v1, plb0096_v2, plb0096_v3
	Work from home, Possibility	[1997,1999,2009-2014]	plb0097
	Work from home, Possibility in Contract	[2020]	plb0697
	Work in black economy	[2015-2016], [2015]	plb0571, plb0572
	Work time regulations	[2003,2005,2007,2009-2019]	plb0211
	Work, last 7 days	[1999-2020]	plb0018
	Working hours, October 2014	[2015]	plb0579, plb0579_h, plb0580, plb0581, plb0581_h
	Working overtime	[1997-2020]	plb0193
	Working overtime, Compensation Period	[2002-2020], [2020]	plb0194_v1, plb0194_v2
	Working overtime, Compensation period	[2002-2020], [2002-2017], [2018-2020]	plb0220_h, plb0220_v1, plb0220_v2
	Working overtime, Financial Compensation	[2015-2020]	plb0605
	Working overtime, Hours Last Month	[1986,1988-2020]	plb0197
	Working overtime, Last Month	[1986,1988-2020], [1986,1988-1996], [1997-2001], [2002-2020]	plb0196_h, plb0196_v1, plb0196_v2, plb0196_v3
	Working overtime, Paid Hours Last Month	[2001-2020]	plb0198
	Working overtime, Time taken off	[2013-2020]	plb0483, plb0484
	Workload (effort-reward imbalance), Career Prospects	[2006,2011-2012,2016]	plb0134, plb0135
	Workload (effort-reward imbalance), Interruptions	[2006,2011-2012,2016]	plb0120, plb0121

continues on next page

Table 2 – continued from previous page

Questionnaire	Module	Years	Variables
	Workload (effort-reward imbalance), Job at risk	[2006,2011-2012,2016]	plb0128, plb0129
	Workload (effort-reward imbalance), Poor Career Prospects	[2006,2011-2012,2016]	plb0124, plb0125
	Workload (effort-reward imbalance), Poor Working Conditions	[2006,2011-2012,2016]	plb0126, plb0127
	Workload (effort-reward imbalance), Problems Sleeping	[2006,2011-2012,2016]	plb0117
	Workload (effort-reward imbalance), Recognition by Superiors	[2006,2011-2012,2016]	plb0130, plb0131
	Workload (effort-reward imbalance), Recognition for Performance	[2006,2011-2012,2016]	plb0132, plb0133
	Workload (effort-reward imbalance), Sacrifices for Career	[2006,2011-2012,2016]	plb0115
	Workload (effort-reward imbalance), Salary	[2006,2011-2012,2016]	plb0136, plb0137
	Workload (effort-reward imbalance), Thinking about Work	[2006,2011-2012,2016]	plb0113, plb0114, plb0116
	Workload (effort-reward imbalance), Time Pressure	[2006,2011-2012,2016]	plb0112, plb0118, plb0119
	Workload (effort-reward imbalance), Work Volume	[2006,2011-2012,2016]	plb0122, plb0123
Youth Questionnaire	Jobs and money, Employment Form	[2000-2020]	jl0014
	Jobs and money, First Job	[2000-2020]	jl0017, jl0018
	Jobs and money, Job Search	[2006-2020]	jl0386
	Jobs and money, Own Earnings	[2000-2020]	jl0013
	Jobs and money, Paid Work	[2006-2020]	jl0385
	Jobs and money, Reason for Working	[2001-2020]	jl0019
	Jobs and money, Savings	[2000-2020], [2000-2001], [2002-2020], [2000-2020]	jl0023, jl0024_h, jl0024_v1, jl0024_v2, jl0025
	Jobs and money, Unemployment	[2006-2020]	jl0387

2.3 Income, Taxes, and Social Security

The income, taxes, and social security modules collect wide-ranging financial information from earnings and spending to public benefits, pensions, inheritances, taxes, and debts. They also cover assets such as real estate and other property.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Additional questions for employed people, 13th month payment prev. year	[1984-2020], [1984-2001], [2002-2020]	plc0042, plc0043_h, plc0043_v1, plc0043_v2
	Additional questions for employed people, 14th month payment prev. year	[1984-2020], [1984-2001], [2002-2020]	plc0044, plc0045_h, plc0045_v1, plc0045_v2
	Additional questions for employed people, Christmas Bonus prev. year	[1984-2020], [1984-2001], [2002-2020]	plc0046, plc0047_h, plc0047_v1, plc0047_v2
	Additional questions for employed people, No Bonus prev. year	[1984-2020]	plc0054
	Additional questions for employed people, Other Bonus prev. year	[1984-2020], [1984-2001], [2002-2020]	plc0052, plc0053_h, plc0053_v1, plc0053_v2
	Additional questions for employed people, Profit-sharing Bonus prev. year	[1985-2020], [1985-2001], [2002-2020]	plc0050, plc0051_h, plc0051_v1, plc0051_v2
	Additional questions for employed people, Vacation Bonus prev. year	[1984-2020], [1984-2001], [2002-2020]	plc0048, plc0049_h, plc0049_v1, plc0049_v2
	Additional questions for retirees / pensioners, Accident Insurance Retirement Pension	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0243_h, plc0243_v1, plc0243_v2
	Additional questions for retirees / pensioners, Accident Insurance Widow's Pension	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0286_h, plc0286_v1, plc0286_v2
	Additional questions for retirees / pensioners, Company Retirement Pension	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0240_h, plc0240_v1, plc0240_v2

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Additional questions for retirees / pensioners, Company Widow's Pension	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0283_h, plc0283_v1, plc0283_v2
	Additional questions for retirees / pensioners, Invalid Pension non-response	[2003-2020]	plc0251
	Additional questions for retirees / pensioners, Orphan Benefit non-response	[2003-2020]	plc0290
	Additional questions for retirees / pensioners, Other Retirement Pensions	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0249_h, plc0249_v1, plc0249_v2
	Additional questions for retirees / pensioners, Other Widow's Pensions	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0288_h, plc0288_v1, plc0288_v2
	Additional questions for retirees / pensioners, Private Retirement Pension	[2003-2020], [2018-2020]	plc0242_v1, plc0242_v2
	Additional questions for retirees / pensioners, Private Widow's Pension	[2003-2020]	plc0285
	Additional questions for retirees / pensioners, Retirement Pension Civil Servants	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0236_h, plc0236_v1, plc0236_v2
	Additional questions for retirees / pensioners, Riester Pension	[2015-2020]	plc0516, plc0517
	Additional questions for retirees / pensioners, Shareholder Company	[2019]	plc0572
	Additional questions for retirees / pensioners, Supplementary Pension Civil Servants	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0238_h, plc0238_v1, plc0238_v2
	Additional questions for retirees / pensioners, Supplementary Widow's Pension Civil Servants	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0281_h, plc0281_v1, plc0281_v2
	Additional questions for retirees / pensioners, War Victims Pension	[1986-2001,2003-2016], [1986-2001], [2003-2016]	plc0245_h, plc0245_v1, plc0245_v2
	Additional questions for retirees / pensioners, War Victims Widow's pension	[1986-2001,2003-2016], [1986-2001], [2003-2016]	plc0247_h, plc0247_v1, plc0247_v2

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Additional questions for retirees / pensioners, Widow's pension	[1986-2001,2003-2020], [2003-2020]	plc0268_h, plc0268_v1, plc0268_v2, plc0268_v3
	Additional questions for retirees / pensioners, Widow's pension Civil Servants	[1986-2001,2003-2020], [1986-2001], [2003-2020]	plc0279_h, plc0279_v1, plc0279_v2
	Asset balance	[2002]	plc0340
	Asset balance, Building Loan Contract (Bausparvertrag)	[2007,2012], [2017,2019], [2007,2012], [2017,2019]	plc0317_v1, plc0317_v2, plc0318_v1, plc0318_v2
	Asset balance, Building Society Savings	[2007,2012,2017,2019]	plc0315, plc0316, plc0319
	Asset balance, Cash Surrender	[2002]	plc0327, plc0335, plc0336, plc0337, plc0338
	Asset balance, Enterprise	[2002]	plc0341, plc0364, plc0365, plc0366, plc0367, plc0368, plc0369
	Asset balance, Financial Assets	[2002], [2002,2007,2012], [2002,2007,2012,2017,2019]	plc0314, plc0326, plc0328, plc0329, plc0330, plc0331, plc0332, plc0333, plc0334
	Asset balance, Financial Burden	[2002], [2002,2007,2012]	plc0408, plc0409, plc0411, plc0412, plc0413, plc0414, plc0415, plc0416, plc0417, plc0418, plc0419, plc0420
	Asset balance, Life Insurance	[2002,2007,2012,2017,2019]	plc0363
	Asset balance, Non self used Property	[2002,2007,2012,2017,2019], [2002,2007,2012], [2002,2007,2012,2017,2019]	plc0356, plc0357, plc0358, plc0359, plc0360, plc0361, plc0362
	Asset balance, Other Property	[2002,2007,2012,2017,2019]	plc0354
	Asset balance, Property non-response	[2007,2012,2017,2019]	plc0355
	Asset balance, Remaining Debt	[2002]	plc0410, plc0421, plc0422, plc0423, plc0424, plc0425
	Asset balance, Residential property	[2002]	plc0339, plc0342, plc0343, plc0344, plc0345, plc0346, plc0347, plc0348, plc0349

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Asset balance, Residential property Usage	[2002,2007,2012,2017,2019]	plc0350, plc0351, plc0352
	Asset balance, Tangible Assets	[2002,2007,2012,2017,2019]	plc0370, plc0371, plc0372, plc0373, plc0374
	Asset balance, Undeveloped Land	[2002,2007,2012]	plc0353
	Asset development	[2019]	plc0570i01, plc0570i02, plc0570i03, plc0570i04, plc0570i05, plc0570i06, plc0570i07, plc0570i08
	Benefits and bonuses from employer	(irregular) [2006-2020]	plc0026, plc0027, plc0028, plc0029, plc0030, plc0031, plc0032, plc0033, plc0034, plc0035, plc0036, plc0037, plc0038, plc0039
	Financial advantages from use of company car	[2016-2018,2020]	plc0532
	Financial support received, Children	[2009-2013]	plj0156, plj0157, plj0158, plj0159
	Financial support received, No Payments	[2009-2013]	plj0172
	Financial support received, Other Relatives	[2009-2013]	plj0164, plj0165, plj0166, plj0167
	Financial support received, Parents	[2009-2013]	plj0152, plj0153, plj0154, plj0155
	Financial support received, Spouse	[2009-2013]	plj0160, plj0161, plj0162, plj0163
	Financial support received, Unrelated Persons	[2009-2013]	plj0168, plj0169, plj0170, plj0171
	Financial support to relatives or others, Children	[1984-1991,1993,1995-2020]	plj0135, plj0136_h, plj0136_v1, plj0136_v2, plj0137_h, plj0137_v1, plj0137_v2, plj0137_v3
	Financial support to relatives or others, No Payments	[1984-1991,1993,1995-2020]	plj0151
	Financial support to relatives or others, Other Relatives	[1984-1991,1993,1995-2020]	plj0143, plj0144_h, plj0144_v1, plj0144_v2, plj0145_h, plj0145_v1, plj0145_v2, plj0145_v3
	Financial support to relatives or others, Parents	[1984-1991,1993,1995-2020]	plj0131, plj0132_h, plj0132_v1, plj0132_v2, plj0133_h, plj0133_v1, plj0133_v2, plj0133_v3

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Financial support to relatives or others, Spouse	[1984-1991,1993,1995-2020]	plj0139, plj0140_h, plj0140_v1, plj0140_v2, plj0142_h, plj0142_v1, plj0142_v2, plj0142_v3
	Financial support to relatives or others, Unrelated Persons	[1984-1991,1993,1995-2020]	plj0147, plj0148_h, plj0148_v1, plj0148_v2, plj0149_h, plj0149_v1, plj0149_v2, plj0149_v3
	Gross / net income, collective wage agreements	[1984-2020]	plc0013_h, plc0013_v1, plc0013_v2, plc0014_h, plc0014_v1, plc0014_v2, plc0502_v1, plc0502_v2, plc0507, plc0508, plc0509
	Income, Alimony	[2010-2015]	plc0181, plc0182, plc0183_v1, plc0183_v2, plc0184, plc0188_v1, plc0188_v2, plc0190_v1, plc0190_v2, plc0494, plc0496
	Income, Alimony Months	[2010-2017], [2012,2018-2020], [2010-2013,2015]	plc0189_v1, plc0189_v2, plc0495
	Income, Child Support	[2010-2015]	plc0177, plc0178, plc0488, plc0490
	Income, Child Support Months	[2010-2013,2015]	plc0489
	Income, Gross Selfemployed prev year	[1990-2020], [1990-2001], [2002-2020], [1995-2020], [2019-2020], [2000-2020], [2000-2001], [2002-2020]	plb0474_h, plb0474_v1, plb0474_v2, plc0073_v1, plc0073_v2, plc0075_h, plc0075_v1, plc0075_v2
	Income, Gross prev year	[1990-2020], [2017-2020]	plb0471_h, plb0471_v1, plb0471_v2, plc0015_h, plc0015_v1, plc0015_v2
	Income, Maternity benefit	[1995-2020], [2019-2020], [1995-2020], [2019-2020], [1995-2020], [1995-2001], [2002-2020], [1990-2020], [1990-2001], [2002-2020]	plc0126_v1, plc0126_v2, plc0152_v1, plc0152_v2, plc0153_h, plc0153_v1, plc0153_v2, plc0155_h, plc0155_v1, plc0155_v2
	Income, Maternity benefit Months	[1995-2020]	plc0154
	Income, Months of Second Job Income	[1995-2020]	plc0065
	Income, Months of Self-Employed Income	[1995-2020]	plc0074
	Income, Months of Wages	[1995-2020], [2000-2020], [2000-2001], [2002-2020]	plc0016, plc0017_h, plc0017_v1, plc0017_v2
	Income, No Other Income	[1995-2020], [2001-2020]	plc0116, plc0117

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Income, Retirement pension	[1995-2020], [2017-2020], [1995-2020], [1995-2001], [2002-2020], [1995-2020], [2017-2020]	plc0232_v1, plc0232_v2, plc0233_h, plc0233_v1, plc0233_v2, plc0234_v1, plc0234_v2
	Income, Retirement pension Months	[1995-2020]	plc0235
	Income, Second Job	[2015,2019-2020]	plc0515
	Income, Second Job prev year	[1990-2020], [1990-2001], [2002-2020]	plb0477_h, plb0477_v1, plb0477_v2
	Income, Self-Employment	[2015,2019-2020]	plc0514
	Income, Sideline Job prev year	[1995-2020], [2019-2020]	plc0064_v1, plc0064_v2
	Income, Student loans	[1995-2020], [1995-2001], [2002-2020], [1995-2020], [2017-2020], [1990-2020], [1990-2001], [2002-2020]	plc0168_h, plc0168_v1, plc0168_v2, plc0169_v1, plc0169_v2, plc0171_h, plc0171_v1, plc0171_v2
	Income, Student loans Months	[1995-2020]	plc0170
	Income, Support from outside the household	[1990-2020], [1990-2001], [2002-2020], [1995-2020], [2019-2020], [1995-2020], [1995-2001], [2002-2020], [1995-2020], [2019-2020]	plc0198_h, plc0198_v1, plc0198_v2, plc0202_v1, plc0202_v2, plc0203_h, plc0203_v1, plc0203_v2, plc0204_v1, plc0204_v2
	Income, Support from outside the household Months	[1995-2020], [2018]	plc0205_v1, plc0205_v2
	Income, Unemployment benefit	[1995-2020], [2017-2020], [1995-2020], [1995-2001], [2002-2020], [1995-2001], [2002-2020], [2018-2020], [1995-2020], [2017-2020], [1995-2020], [1990-2020], [1990-2001], [2002-2020]	plc0130_v1, plc0130_v2, plc0131_h, plc0131_v1, plc0131_v2, plc0132_v1, plc0132_v2, plc0132_v3, plc0135_v1, plc0135_v2, plc0136, plc0137_h, plc0137_v1, plc0137_v2
	Income, Unemployment benefit II	[1995-2020], [2018-2020], [1995-2020]	plc0138_v1, plc0138_v2, plc0139
	Income, Wages	[2015,2019-2020]	plc0513
	Income, Widow's pension	[1995-2020], [2019-2020], [1995-2020], [1995-2001], [2002-2020], [1995-2020], [2019-2020]	plc0273_v1, plc0273_v2, plc0274_h, plc0274_v1, plc0274_v2, plc0275_v1, plc0275_v2
	Income, Widow's pension Months	[1995-2020]	plc0276
	Inheritances	[2001,2019], [2017]	plc0375_v1, plc0375_v2
	Inheritances, First Inheritance, amount	[2001], [2017], [2001,2019], [2017]	plc0383_v1, plc0383_v2, plc0384_v1, plc0384_v2
	Inheritances, First Inheritance, last 15 years	[2017]	plc0376_v2, plc0377_v2, plc0378_v2, plc0379_v2, plc0380_v2, plc0381_v2, plc0382_v2

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Inheritances, First Inheritance, once/ever	[2001,2019]	plc0376_v1, plc0377_v1, plc0378_v1, plc0379_v1, plc0380_v1, plc0381_v1, plc0382_v1, plc0383_h
	Inheritances, From Whom	[2001]	plc0385, plc0395, plc0405
	Inheritances, Future	[2001]	plc0406, plc0407
	Inheritances, Second Inheritance, last 15 years	[2017]	plc0386_v2, plc0387_v2, plc0388_v2, plc0389_v2, plc0390_v2, plc0391_v2, plc0392_v2, plc0393_v2, plc0394_v2
	Inheritances, Second Inheritance, once/ever	[2001,2019], [2001], [2001,2019]	plc0386_v1, plc0387_v1, plc0388_v1, plc0389_v1, plc0390_v1, plc0391_v1, plc0392_v1, plc0393_v1, plc0394_v1
	Inheritances, Third Inheritance, amount	[2001], [2017], [2019], [2001,2019], [2017]	plc0403_v1, plc0403_v2, plc0403_v3, plc0404_v1, plc0404_v2
	Inheritances, Third Inheritance, last 15 years	[2017]	plc0396_v2, plc0397_v2, plc0398_v2, plc0399_v2, plc0400_v2, plc0401_v2, plc0402_v2
	Inheritances, Third Inheritance, once/ever	[2001,2019]	plc0396_v1, plc0397_v1, plc0398_v1, plc0399_v1, plc0400_v1, plc0401_v1, plc0402_v1, plc0403_h
	Pension entitlements, company	[2013,2018], [2013], [2018], [2013,2018]	plc0432, plc0433, plc0434_v1, plc0434_v2, plc0435, plc0441, plc0442, plc0443, plc0444_v1, plc0444_v2, plc0445
	Pension payments	[2013,2018], [2013], [2018]	plc0437, plc0438, plc0439_v1, plc0439_v2
	Riester / Ruerup pension plans	(irregular) [2004-2020]	plc0313_h, plc0313_v1, plc0313_v2, plc0430, plc0431
	Social security, Don't know	[2002,2007,2012,2017]	plc0009
	Social security, Financial Security	[2002,2007,2012,2017]	plc0111, plc0112, plc0113, plc0114
	Social security, Minimum Household Income	[1992,2002,2007,2012,2017], [1992], [2002,2007,2012,2017]	plc0001_h, plc0001_v1, plc0001_v2
	Wage tax classification	[1991,1993,2004,2016], [2004,2016]	plc0091_h, plc0091_v1, plc0091_v2, plc0091_v3, plc0091_v4, plc0091_v5, plc0091_v6, plc0091_v7, plc0091_v8, plc0091_v9

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
Household Questionnaire	Alimony	[2010]	hlc0091, hlc0092, hld0004, hld0005
	Credit burden	[2005-2016], [2005-2011], [2011-2016]	hlc0115_h, hlc0115_v1, hlc0115_v2
	Expenditures on Food, Month	(irregular) [1998-2020], [1998,2000-2001], (irregular) [2003-2020]	hlf0436_h, hlf0436_v1, hlf0436_v2
	Expenditures on Food, Week	(irregular) [1998-2020], [1998,2000-2001], (irregular) [2003-2020]	hlf0435_h, hlf0435_v1, hlf0435_v2
	Good/Low Income, Good Household Income	[1992,1997,2007,2018], [1992,1997], [2007,2018]	hlc0022_h, hlc0022_v1, hlc0022_v2
	Good/Low Income, Insufficient Household Income	[1992,1997,2007,2018], [1992,1997], [2007,2018]	hlc0020_h, hlc0020_v1, hlc0020_v2
	Good/Low Income, Just Sufficient Household Income	[1992,1997,2007,2018], [1992,1997], [2007,2018]	hlc0021_h, hlc0021_v1, hlc0021_v2
	Good/Low Income, Poor Household Income	[1992,1997,2007,2018], [1992,1997], [2007,2018]	hlc0019_h, hlc0019_v1, hlc0019_v2
	Good/Low Income, Very Good Household Income	[1992,1997,2007,2018], [1992,1997], [2007,2018]	hlc0023_h, hlc0023_v1, hlc0023_v2
	Good/Low Income, Very Poor Household Income	[1992,1997,2007,2018], [1992,1997], [2007,2018]	hlc0018_h, hlc0018_v1, hlc0018_v2
	Household income / expenses, Basic financial security in old age prev. Year	[2005-2020]	hlc0061_h, hlc0061_v1, hlc0061_v2, hlc0062, hlc0063, hlc0071
	Household income / expenses, Child Allowance prev. year	[1984-2000], [2001-2020], [1985-2020], [1985-2001], [2002-2020]	hlc0040, hlc0041, hlc0042_h, hlc0042_v1, hlc0042_v2
	Household income / expenses, Child Allowance today	[1995-2020], [2000-2009], [2010-2020], [1995-2020], [1995-2001], [2002-2020]	hlc0044_h, hlc0044_v1, hlc0044_v2, hlc0045_h, hlc0045_v1, hlc0045_v2
	Household income / expenses, Child Benifit	[1984-2020], [1985-1990], [1991-1995], [1996-2020]	hlc0039_h, hlc0039_v1, hlc0039_v2, hlc0039_v3
	Household income / expenses, Child Care Subsidy	[2009-2020]	hlc0046_h, hlc0046_v1, hlc0046_v2, hlc0046_v3, hlc0046_v4, hlc0047_h, hlc0047_v1, hlc0047_v2, hlc0124, hlc0125
	Household income / expenses, Child Care Subsidy prev. Year	[2009-2020]	hlc0049_h, hlc0049_v1, hlc0049_v2, hlc0050_h, hlc0050_v1, hlc0050_v2, hlc0051_h, hlc0051_v1, hlc0051_v2, hlc0121, hlc0122, hlc0123

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Household income / expenses, Compulsory Long Term Care Insurance	[2001-2020], [2001], [2002-2020], [1996-2020], [1996], [1997-1999], [1997-2000], [1997-1998], [2000-2009], [2010-2020], [1996-2020], [1996-2001], [2002-2020]	hlc0079_h, hlc0079_v1, hlc0079_v2, hlc0085_h, hlc0085_v1, hlc0085_v2, hlc0085_v3, hlc0085_v4, hlc0085_v5, hlc0085_v6, hlc0090_h, hlc0090_v1, hlc0090_v2
	Household income / expenses, Compulsory Long Term Care Insurance prev. Year	[2001-2020]	hlc0078
	Household income / expenses, Family Members Support	[2001-2020]	hlc0077
	Household income / expenses, Help with living costs	[1984-2009], [1984,1991-2009], [1985-1990], [1995-2020], [1995-1998,2010-2020], [1999-2009], [1995-2020], [1995-2001], [2002-2020]	hlc0066_h, hlc0066_v1, hlc0066_v2, hlc0067_h, hlc0067_v1, hlc0067_v2, hlc0068_h, hlc0068_v1, hlc0068_v2
	Household income / expenses, Help with living costs prev. Year	[1984-2020], [1984,1991,2010-2020], [1985-1990], [1992-2009], [1984-1991,2001-2020], [1984-1991,2001], [2002-2020]	hlc0055_h, hlc0055_v1, hlc0055_v2, hlc0055_v3, hlc0059_h, hlc0059_v1, hlc0059_v2
	Household income / expenses, Housing assistance	[1984-2020], [1984,1991-2020], [1985-1990], [1995-2020], [1995-1998,2010-2020], [1999-2009], [1995-2020], [1995-2001], [2002-2020]	hlc0080_h, hlc0080_v1, hlc0080_v2, hlc0083_h, hlc0083_v1, hlc0083_v2, hlc0084_h, hlc0084_v1, hlc0084_v2
	Household income / expenses, Housing assistance prev. Year	[1984-2020], [1984-2001], [2002-2020]	hlc0081, hlc0082_h, hlc0082_v1, hlc0082_v2
	Household income / expenses, Income Bracket	[1999-2000], [2001-2002], [2003-2020]	hlc0006_v1, hlc0006_v2, hlc0006_v3
	Household income / expenses, Monthly Household Income	[1984-2020], [1984-2001], [2002-2020]	hlc0005_h, hlc0005_v1, hlc0005_v2
	Household income / expenses, Reduction of earning capacity	[2005-2020], [2005-2009], [2010-2020]	hlc0070_h, hlc0070_v1, hlc0070_v2
	Household income / expenses, Special Circumstances Assistance	[1984-2009], [1984-1991,2001], [2002-2009]	hlc0056_h, hlc0056_v1, hlc0056_v2, hlc0056_v3, hlc0058, hlc0060_h, hlc0060_v1, hlc0060_v2
	Household income / expenses, Subsistence Support prev. year	[1984-1991,2001-2020]	hlc0057
	Household income / expenses, Unemplyoment Subsidy II	[2005-2020], [2005-2009], [2010-2020], [2005-2020]	hlc0064_h, hlc0064_v1, hlc0064_v2, hlc0065

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Household income / expenses, Unemplyoment Subsidy II prev. year	[2006-2020]	hlc0052, hlc0053, hlc0054
	Household income / expenses; Number of Children	[1995-2020]	hlc0043
	Income and expenses from renting / leasing	[1984-1990,1992-2020], [1984-1990,1992-2001], [2002-2020]	hlc0007, hlc0008_h, hlc0008_v1, hlc0008_v2
	Income and expenses from renting / leasing, Maintenance costs	[1984-1990,1992-2020], [1984-1990,1992-2001], [2002-2020], [2016-2020]	hlc0111_h, hlc0111_v1, hlc0111_v2, hlc0176
	Income and expenses from renting / leasing, Redemption payments	[1985-1990,1992-2020], [1985-1990,1992-2001], [2002-2020], [2016-2020]	hlc0112_h, hlc0112_v1, hlc0112_v2, hlc0177
	Income and expenses from renting / leasing, Tax Deduction	[2005-2020]	hlc0009, hlc0010
	Inheritance, present, lottery prize	[2016-2020]	hlc0178, hlc0179, hlc0180, hlc0181, hlc0182, hlc0183
	Investments, Building Society	[1984-2020]	hlc0105
	Investments, Combined Savings	[1990]	hlc0097
	Investments, Fixed Interest Securities	[1984-2020]	hlc0107
	Investments, Interest and Dividend Income	[1984-2001], [2002-2020], [1985-2020]	hlc0013_v1, hlc0013_v2, hlc0014
	Investments, Life Insurance	[1984-2020]	hlc0106
	Investments, No Securities	[1984-2020]	hlc0093
	Investments, Non-response	[2003-2020], [2016-2020]	hlc0096, hlc0184
	Investments, Operating Assets	[1984-2020]	hlc0104
	Investments, Other Securities	[2001-2020]	hlc0108
	Investments, Savings Account	[1984-2020]	hlc0098
	Investments, Tax-deductible Loan	[2005-2020]	hlc0094, hlc0095
	Ratio between income and expenditures	[2010-2013], [2016-2018], [2010]	hlc0024_v1, hlc0024_v2, hlc0030
	Ratio between income and expenditures, Cap shortfall	[2010-2013], [2016]	hlc0029_v1, hlc0029_v2
	Ratio between income and expenditures, Expenditure Surplus	[2010-2013], [2016-2018], [2010-2013], [2016]	hlc0027_v1, hlc0027_v2, hlc0028_v1, hlc0028_v2

continues on next page

Table 3 – continued from previous page

Questionnaire	Module	Years	Variables
	Ratio between income and expenditures, Income Surplus	[2010-2013], [2016-2018], [2010-2013], [2016]	hlc0025_v1, hlc0025_v2, hlc0026_v1, hlc0026_v2
	Repayments of loans	[1997-2020], [1997-2011], [2011-2020], [1997-2020], [1997-2001], [2002-2011]	hlc0113_h, hlc0113_v1, hlc0113_v2, hlc0114_h, hlc0114_v1, hlc0114_v2
	Savings	[1992-2020]	hlc0119_h, hlc0119_v1, hlc0119_v2, hlc0119_v3, hlc0119_v4, hlc0120_h, hlc0120_v1, hlc0120_v2, hlc0120_v3, hlc0120_v4

2.4 Family and Social Networks

As a household study, the SOEP offers rich information on family and social relationships and how these connections change in different stages of life. The modules dealing with family and social networks cover the entire life cycle beginning with pregnancy and childbirth and continuing through parenthood, family formation, friendships, marriage, divorce, and death, and also provide a wealth of additional information on important life events.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Childcare	[2003-2020]	suppartn
	Circle of friends, sociodemographics	[2011,2016]	pld0104, pld0105, pld0106
	Circle of friends, sociodemographics, age	[2006,2011-2012,2016]	pld0095, pld0096, pld0097
	Circle of friends, sociodemographics, education	[2006,2011,2016]	pld0101, pld0102, pld0103
	Circle of friends, sociodemographics, labor force status	[2006,2011-2012,2016]	pld0098, pld0099, pld0100
	Circle of friends, sociodemographics, relations	(unregelmaessig) [1988-2016], [2012]	pld0089_h, pld0089_v1, pld0089_v2, pld0090_h, pld0090_v1, pld0090_v2, pld0091_h, pld0091_v1, pld0091_v2

continues on next page

Table 4 – continued from previous page

Questionnaire	Module	Years	Variables
	Circle of friends, sociodemographics, sex	(unregelmaessig) [1988-2016]	pld0092, pld0093, pld0094
	Circle of friends, trustworthy persons	[2006,2011,2016,2019], [2006,2011,2016,2019], (unregelmaessig) [1991-2019], [1991,1996], [2001], [2006,2011,2016,2019], (unregelmaessig) [1991-2019], [1991,1996], [2001], [2006,2011,2016,2019]	pld0062_v1, pld0063_v1, pld0064_v1, pld0065_v1, pld0066_v1, pld0067, pld0068_v1, pld0069_v1, pld0070_v1, pld0071_v1, pld0072_v1, pld0073, plf0049_h, plf0049_v1, plf0049_v2, plf0049_v3, plf0050_h, plf0050_v1, plf0050_v2, plf0050_v3
	Circle of friends, trustworthy persons (M3-M5)	[2017-2019]	pld0062_v2, pld0063_v2, pld0064_v2, pld0065_v2, pld0066_v2, pld0068_v2, pld0069_v2, pld0070_v2, pld0071_v2, pld0072_v2
	Circle of friends, trustworthy persons: conflicts	[2006,2011,2013,2016,2019], [2006,2011,2013,2016,2019]	pld0077, pld0078, pld0079, pld0080, pld0081, pld0082
	Circle of friends, trustworthy persons: help	[2006,2011,2016,2019]	pld0074, pld0075, pld0076
	Circle of friends, trustworthy persons: unpleasant truths	[2006,2011,2016,2019], [2017-2018], [2006,2011,2016,2019], [2017-2018], [2006,2011,2016,2019], [2017-2018], [2011,2016,2019], [2017-2018], [2011,2016,2019], [2017-2018], [2006,2011,2016,2019]	pld0083_v1, pld0083_v2, pld0084_v1, pld0084_v2, pld0085_v1, pld0085_v2, pld0086_v1, pld0086_v2, pld0087_v1, pld0087_v2, pld0088
	Family changes	[1991,1996,2001], [1985-2020], [2003-2020]	pld0012, pld0013, pld0014, pld0038, pld0039, pld0040, pld0159, pld0160
	Family changes, childbirth	[1999-2020]	pld0152, pld0153, pld0154
	Family changes, death	[1999-2020]	pld0146, pld0147, pld0148, pld0161, pld0162, pld0163, pld0164, pld0165, pld0166, pld0167, pld0168, pld0169, pld0170, pld0171
	Family changes, divorce	[1999-2020]	pld0140, pld0141, pld0142
	Family changes, marriage	[1999-2020]	pld0134, pld0135, pld0136
	Family changes, moving in	[1999-2020]	pld0137, pld0138, pld0139
	Family changes, moving out (child)	[1999-2020]	pld0149, pld0150, pld0151
	Family changes, other	[1985-1995,1999-2008,2010-2020]	pld0155, pld0156, pld0158

continues on next page

Table 4 – continued from previous page

Questionnaire	Module	Years	Variables
	Family changes, separation	[1999-2020]	pld0143, pld0144, pld0145
	Family network, aunt	[2006,2011,2016]	pld0115, pld0116
	Family network, children	[1991,1996,2001,2006,2011,2016]	pld0025, pld0026, pld0027, pld0028, pld0301i01, pld0301i02, pld0302
	Family network, distance	[1991], [1991], [1996,2001], [2006,2011,2016]	plj0117_v1, plj0117_v2, plj0117_v3, plj0118_v1, plj0118_v2, plj0118_v3, plj0119_v1, plj0119_v2, plj0119_v3, plj0120, plj0121, plj0122_v1, plj0122_v2, plj0122_v3, plj0123_v1, plj0123_v2, plj0123_v3, plj0124_h, plj0124_v1, plj0124_v2, plj0124_v3, plj0125_v1, plj0125_v2, plj0125_v3, plj0126, plj0127_v1, plj0127_v2, plj0127_v3, plj0128, plj0129, plj0130_v1, plj0130_v2, plj0130_v3
	Family network, grandchildren	[1991,1996,2001,2006,2011,2016]	pld0033, pld0034
	Family network, grandparents	[2006,2011,2016]	pld0110, pld0111, pld0112, pld0113, pld0114
	Family network, other relatives	[1991,1996,2001,2006,2011,2016]	pld0035, pld0036
	Family network, parents	[1991,1996,2001,2006,2011,2016]	pld0023, pld0024
	Family network, siblings	[1991,1996,2001,2006,2011,2016], [1991,1996,2001,2003,2006,2011], [1991,1996,2001,2006,2011,2016], (unregelmaessig) [1991-2016]	pld0029, pld0030, pld0031, pld0032
	Family network, spouse	[1996,2001,2006,2011,2016], [1996,2001], [2006,2011,2016], [1991,1996,2001,2006,2011,2016], [2006,2011,2016]	pld0020, pld0021_h, pld0021_v1, pld0021_v2, pld0022, pld0107
	Family network, stepparents	[2006,2011,2016]	pld0108, pld0109
	Family network, uncle	[2006,2011,2016]	pld0117, pld0118
	Friends	(unregelmaessig) [2003-2020]	pld0047
	Leisure and activities (with child)	[2003-2020]	tvhrs, tvyn
	Marital / partnership status	[2019]	pld0131_v2, pld0131_v3, pld0132_v1, pld0132_v2, pld0133, pld0299, pld0300, plk0001_v2, plk0001_v3
	Pregnancy and childbirth	[2003-2020]	pregplan
	Sexual orientation	[2016]	pld0298_v1, pld0298_v2, pld0298_v3

continues on next page

Table 4 – continued from previous page

Questionnaire	Module	Years	Variables
Youth Questionnaire	Allowance (Pocket money)	, [2000-2020], [2002-2020]	, jl0022_h, jl0022_v2
	Allowance (Pocket money, Deutschmark)	[2000-2001]	jl0021_v1, jl0022_v1
	Childhood and parental home	[2000-2018]	jl0273, jl0279
	Childhood and parental home (parent's education, ISCO-08)	[2013-2018]	j_isco08_jobfather, j_isco08_jobmother
	Childhood and parental home (parent's education, ISCO-88)	[2000-2017]	j_isco88_jobfather, j_isco88_jobmother
	Childhood and parental home (parent's education, KldB 2010)	[2013-2018]	j_kldb2010_jobfather, j_kldb2010_jobmother
	Childhood and parental home (parent's education, KldB 92)	[2000-2017]	j_kldb92_jobfather, j_kldb92_jobmother
	Childhood and parental home, father	[2000-2018], [2014-2017], [2014-2018], [2015-2018]	jl0307, jl0309, jl0311, jl0313_v1, jl0313_v2, jl0315, jl0327_h, jl0327_v1, jl0327_v2, jl0506, jl0508, jl0510, jl0512, jl0514, jl0516, jl0518, jl0520, jl0522
	Childhood and parental home, mother	[2000-2018], [2014-2017], [2014-2018], [2015-2018]	jl0304, jl0308, jl0310, jl0312, jl0314_v1, jl0314_v2, jl0316, jl0328_h, jl0328_v1, jl0328_v2, jl0507, jl0509, jl0511, jl0513, jl0515, jl0517, jl0519, jl0521, jl0523
	Childhood and parental home, siblings	[2004-2012]	jl0274, jl0275, jl0276, jl0277, jl0278, jl0446, jl0447, jl0454, jl0455, jl0456, jl0457, jl0458, jl0459, jl0460, jl0461, jl0462, jl0463, jl0464, jl0465, jl0466, jl0467, jl0468, jl0469, jl0470, jl0471, jl0472, jl0473, jl0474, jl0475, jl0476, jl0477, jl0478, jl0479, jl0480, jl0481, jl0482, jl0483, jl0484, jl0485, jl0486, jl0487, jl0488, jl0489, jl0490, jl0491, jl0492, jl0493, jl0494, jl0495, jl1406, jl1407, jl1408, jl1409, jl1410, jl1411
	Parental interest in child's performance in school	[2000-2018]	jl0167, jl0168, jl0169, jl0170, jl0171, jl0172, jl0173, jl0174

continues on next page

Table 4 – continued from previous page

Questionnaire	Module	Years	Variables
	Relationship to family members	[2001-2020]	jl0026, jl0027, jl0028, jl0029, jl0030, jl0031, jl0032, jl0033, jl0034, jl0040, jl0041, jl0043, jl0044, jl0045, jl0046, jl0047, jl0048, jl0049, jl0050, jl0051, jl0052, jl0053, jl0054, jl0055, jl1043
	Relationship to family members, conflicts	[2001-2018], [2019-2020], [2001-2018], [2019-2020], [2001-2018], [2019-2020], [2001-2018], [2019], [2001-2018], [2019]	jl0035_v1, jl0035_v2, jl0036_v1, jl0036_v2, jl0037_v1, jl0037_v2, jl0038_v1, jl0038_v2, jl0039_v1, jl0039_v2
Mother and Child Instruments	Allowance (Pocket money)	[2003-2020]	allow, allowpm, allowpw
	Attitude toward parental role	[2003-2020]	bepar1, bepar10, bepar2, bepar3, bepar4, bepar5, bepar6, bepar8, bepar9
	Attitude towards maternal role	[2003-2020]	change1, change2, change3, change4, change5, change6, change7, change8, health
	Breastfeeding	[2003-2020]	breastf, breastfc, breastfm
	Childcare	[2003-2020]	care10h, care11h, care12h, care19, care1h, care3h, care4h, care5h, care6h, care7h, care8h, care9h, maincare
	Eating behavior (child)	[2003-2020]	eatsat1, eatsat2, eatsat3, eatson1, eatson2, eatson3, eatweek1, eatweek3
	Frequency of leisure and other activities (child)	[2003-2020]	freqact1, freqact10, freqact11, freqact12, freqact13, freqact14, freqact15, freqact16, freqact17, freqact18, freqact19, freqact2, freqact20, freqact3, freqact4, freqact5, freqact6, freqact7, freqact8, freqact9
	Friends (child)	[2003-2020]	frndadlt, frndchld
	Language use	[2003-2020]	language
	Leisure and activities (with child)	[2003-2020]	activ1, activ2, activ3, activ4, activ5, activ6, activ7, activ8, activ9
	Parental interest in child's performance in school	[2003-2020]	conscho1, conscho2, conscho3, conscho4, conscho5, conscho6, conscho7
	Parenting goals	[2003-2020]	edgoal1, edgoal10, edgoal11, edgoal12, edgoal13, edgoal14, edgoal15, edgoal16, edgoal17, edgoal18, edgoal2, edgoal3, edgoal4, edgoal5, edgoal6, edgoal7, edgoal8, edgoal9

continues on next page

Table 4 – continued from previous page

Questionnaire	Module	Years	Variables
	Parenting style	[2003-2020]	edbeh1, edbeh10, edbeh11, edbeh12, edbeh13, edbeh14, edbeh15, edbeh16, edbeh17, edbeh18, edbeh2, edbeh3, edbeh4, edbeh5, edbeh6, edbeh7, edbeh8, edbeh9
	Pregnancy and childbirth	[2003-2020]	birthpw, delivpl, nchild
	Relationship to other parent or child	[2003-2020]	biochild

2.5 Health and Care

The modules on health and care cover doctor visits, sports and fitness, alcohol consumption, health insurance, health status, and grip strength, both on respondents themselves and on other individuals in the household, such as children and deceased household members.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Additional private insurance	(irregular) [1999-2020]	ple0128_h, ple0128_v1, ple0128_v2, ple0129, ple0130, ple0131, ple0132, ple0133, ple0134
	Alcohol consumption	[2006,2008,2010]	ple0090, ple0091, ple0092, ple0093, ple0177, ple0178
	Disabilities in everyday life (SF-12)	(irregular) [2002-2020]	ple0004, ple0005
	Electronic cigarette: liquid	[2020]	ple0195
	Health insurance	[1999-2020], [1999], [2000-2009], [1999-2020]	ple0097, ple0099_h, ple0099_v1, ple0099_v2, ple0099_v3, ple0099_v4, ple0099_v5, ple0104_h, ple0104_v1, ple0104_v2, ple0104_v3, ple0104_v4, ple0104_v5, ple0104_v6, ple0104_v7, ple0160
	Health insurance debts	[2017]	plc0567

continues on next page

Table 5 – continued from previous page

Questionnaire	Module	Years	Variables
	Health insurance, private	[1984-1986], (irregular) [1999-2020]	ple0098_v1, ple0098_v2, ple0098_v3, ple0098_v4, ple0098_v5
	Health restrictions	[2011-2013,2015-2020], [2012-2013,2015-2020]	ple0009, ple0162
	Height and weight	(irregular) [2002-2020]	ple0006, ple0007
	Hospital stays	[1984-1989,1991-1992,1994-2020]	ple0053, ple0055, ple0056
	Illnes	[2011,2013,2015,2017,2019], [2019]	ple0011, ple0012, ple0013, ple0014, ple0015, ple0016, ple0017, ple0018, ple0019, ple0020, ple0021, ple0022, ple0023, ple0024, ple0189
	Individual health services	[2016,2018,2020]	ple0186
	Nutritional awareness	[2016,2018,2020]	ple0179, ple0180, ple0181, ple0182
	Private supplementary care insurance	[2016,2018,2020]	ple0183, ple0184, ple0185
	Qualification for additional benefits	[1999-2011]	ple0121
	Reduced ability to work	[1984-1989,1991-1992,1994-2020]	ple0040, ple0041
	Sickness notifications to employer	[1985-1989,1991-1992,1994-2020]	plb0024_h, plb0024_v1, plb0024_v2, plb0024_v3, ple0044_h, ple0044_v1, ple0044_v2, ple0046, ple0048, ple0049, ple0050, ple0051, ple0052, ple0174, ple0175
	Smoking	[1998], (irregular) [2006-2020], [2016,2018,2020]	ple0080_v1, ple0080_v2, ple0080_v3, ple0081_h, ple0081_v1, ple0081_v2, ple0082, ple0083, ple0084, ple0085, ple0086_v1, ple0086_v2, ple0086_v3, ple0086_v4, ple0089, ple0176
	State of health	[1992,1994-2020]	ple0008
	Stress and exhaustion (SF-12)	(irregular) [2002-2020], (irregular) [1984-2020]	ple0026, ple0027, ple0028, ple0029, ple0030, ple0031, ple0032, ple0033, ple0034, ple0035, ple0036

continues on next page

Table 5 – continued from previous page

Questionnaire	Module	Years	Variables
	Visits to the doctor	[1988-1989,1991-1992,1995-2020], [1984-1989,1991-1992,1994-2020]	ple0072, ple0073
Youth Questionnaire	Height and weight	[2006-2020]	jl0219, jl0220
	State of health	[2006-2020]	jl0218
Household Questionnaire	Satisfaction with availability of care	[1997,2002,2008]	hlf0318
Mother and Child Instruments	Health of child	[2003-2020]	chhealth, lstmedex, medaid3mb
	Health of child, disorders	[2003-2020]	disord, disord1, disord2, disord3, disord4, disord5, disord6, disord7, disord8, disord9
	Health of child, hospital stays	[2003-2020]	hospital12m, hospital3mb
	Health of child, illnesses	[2003-2020]	ill0, ill10, ill11, ill12, ill13, ill14, ill15, ill2, ill31, ill32, ill4, ill5, ill6, ill7, ill8, ill9, illno
	Height and weight of child	[2003-2020]	height, weight, weightb
	Physical and mental health of mother	[2003-2020]	feeling1, feeling2, feeling3, feeling4

2.6 Home, Amenities, and Contributions of Private HH

The housing, amenities, and household expenses modules provide wide-ranging information on everyday life including the type of dwelling and whether it is a rental property or owner-occupied; expenditures on personal hygiene, transportation, and vacations; and the division of household labor.

Questionnaire	Module	Years	Variables
Household Questionnaire	Childcare costs	[2010-2013,2015,2017-2019], [2010-2012], [2013,2015,2017-2019]	ks_cost_h, ks_cost_v1, ks_cost_v2

continues on next page

Table 6 – continued from previous page

Questionnaire	Module	Years	Variables
	Childcare provider	[1987,1995,1997,2002,2005,2007], [2002,2005,2007]	kd_insta_h, kd_insta_v1, kd_insta_v2, kd_insta_v3, kd_insta_v4, kd_insta_v5, kd_insta_v6, kd_insta_v7
	Childcare situation	[1987,1997,1999-2002,2004-2020]	kc_care_h, kc_care_v1, kc_care_v2, kc_care_v3, kc_care_v4, kc_care_v5, kc_care_v6, kc_care_v7
	Dependence on childcare hours	[2002]	kd_rely
	Leisure activities, children	(unregelmaessig) [2006-2020]	ka06_art, ka06_mus, ka06_non, ka06_oth, ka06_spo, ka16_art, ka16_ctr, ka16_mus, ka16_non, ka16_org, ka16_sar, ka16_smu, ka16_sot, ka16_spo, ka16_ssp, ka16_sth, ka16_yth
	Leisure costs, children	(unregelmaessig) [2002-2019], [2002,2005,2007], [2010-2013,2015,2017,2019], [2017-2018], [2010-2013,2015,2017-2019], [2010-2012], [2013,2015,2017,2019], [2017-2018]	kk_amtp_h, kk_amtp_v1, kk_amtp_v2, kk_amtp_v3, kk_cost_h, kk_cost_v1, kk_cost_v2, kk_cost_v3
	Lunch, childcare	(unregelmaessig) [1997-2019], [1997,2002,2005,2007], [2010-2013,2015,2017,2019], [2017-2018]	kd_lunch_h, kd_lunch_v1, kd_lunch_v2, kd_lunch_v3
	Lunch, school	[2010-2013,2015,2017-2019]	ks_lunch
	School attendance by child	[1984-2020], [1984-1994], [1990], [1991], [1995-2020], [2017], [2016-2019]	ks_gen_h, ks_gen_v1, ks_gen_v2, ks_gen_v3, ks_gen_v4, ks_gen_v5, ks_spe
	School provider and costs	(unregelmaessig) [1987-2019], [1987,1996], [2010-2012], [2013,2015,2017-2019]	ks_amtp_h, ks_amtp_v1, ks_amtp_v2, ks_amtp_v3
Household Questionnaire	Change in residential situation	[1991-2020], [1991-1998], [1999-2020], [2015-2020]	hlf0106, hlf0107_h, hlf0107_v1, hlf0107_v2, hlf0523

continues on next page

Table 6 – continued from previous page

Questionnaire	Module	Years	Variables
	Changes in home fixtures and furnishings since last year	[2004,2006,2008,2010-2013]	hlc0116, hlc0117, hlf0159, hlf0164, hlf0165_h, hlf0165_v1, hlf0165_v2, hlf0166, hlf0167, hlf0223, hlf0224, hlf0225, hlf0226, hlf0227, hlf0228, hlf0229, hlf0230, hlf0231, hlf0232, hlf0233, hlf0234, hlf0235, hlf0236, hlf0237, hlf0238, hlf0244, hlf0245, hlf0246, hlf0247, hlf0248, hlf0249, hlf0250, hlf0251, hlf0252
	Changes in home fixtures and furnishings since last year: Internet	[2000,2002,2004,2006], (unregelmaessig) [2000-2013], [2000], [2002,2004,2006,2008,2010-2013]	hlf0169_v1, hlf0169_v2, hlf0169_v3, hlf0169_v4, hlf0169_v5, hlf0169_v6, hlf0169_v7, hlf0170_h, hlf0170_v1, hlf0170_v2
	Changes in home fixtures and furnishings since last year: car	(unregelmaessig) [2000-2013], [2000], [2002,2004,2006,2008,2010-2013], [2010-2013], [2010-2011]	hlf0209_h, hlf0209_v1, hlf0209_v2, hlf0210, hlf0211
	Changes in home fixtures and furnishings since last year: cell phone	(unregelmaessig) [2000-2020], (unregelmaessig) [2000-2013], [2010-2013]	hlf0241_h, hlf0241_v1, hlf0241_v2, hlf0241_v3, hlf0241_v4, hlf0241_v5, hlf0241_v6, hlf0241_v7, hlf0241_v8, hlf0242, hlf0243
	Changes in home fixtures and furnishings since last year: kitchen appliances	(unregelmaessig) [1998-2013]	hlf0214, hlf0215, hlf0216, hlf0217, hlf0218, hlf0219, hlf0220, hlf0221, hlf0222
	Changes in home fixtures and furnishings since last year: motorcycle, moped	(unregelmaessig) [2000-2013], [2000], [2002,2004,2006,2008,2010-2013], [2010-2013]	hlf0212_h, hlf0212_v1, hlf0212_v2, hlf0213
	Changes in home fixtures and furnishings since last year: phone	(unregelmaessig) [1990-2020], (unregelmaessig) [1990-2013], [2013,2015], [2014], [2016-2020], (unregelmaessig) [1998-2013]	hlf0239_h, hlf0239_v1, hlf0239_v2, hlf0239_v3, hlf0239_v4, hlf0240
	Cleaning or household help	[1991,1994,1999-2020], [2010-2020]	hlf0261, hlf0262
	Comparison of old and new home	[1985-2013,2015,2017,2019-2020]	hlf0126, hlf0127, hlf0128, hlf0129, hlf0130, hlf0131, hlf0132
	Consumption Module	[2010-2013]	hlf0172
	Consumption Module: Cars	(unregelmaessig) [1990-2013], [1991]	hlf0163_h, hlf0163_v1, hlf0163_v2
	Consumption Module: Clothes and Shoes	[2010-2013]	hlf0379, hlf0380, hlf0381, hlf0382

continues on next page

Table 6 – continued from previous page

Questionnaire	Module	Years	Variables
	Consumption Module: Cosmetics	[2010-2013]	hlf0383, hlf0384, hlf0385, hlf0386
	Consumption Module: Culture	[2010-2013]	hlf0399, hlf0400, hlf0401, hlf0402
	Consumption Module: Education	[2010-2013]	hlf0395, hlf0396, hlf0397, hlf0398
	Consumption Module: Food and Drinks	[2010-2013]	hlf0371, hlf0372, hlf0373, hlf0374, hlf0375, hlf0376, hlf0377, hlf0378
	Consumption Module: Furniture	[2010-2013]	hlf0427, hlf0428, hlf0429, hlf0430
	Consumption Module: Health	[2010-2013]	hlf0387, hlf0388, hlf0389, hlf0390
	Consumption Module: Hobby	[2010-2013]	hlf0403, hlf0404, hlf0405, hlf0406
	Consumption Module: Holiday	[2010-2013]	hlf0407, hlf0408, hlf0409, hlf0410
	Consumption Module: Insurance	[2010-2013]	hlf0411, hlf0412, hlf0413, hlf0414, hlf0415, hlf0416, hlf0417, hlf0418
	Consumption Module: Internet	[2010-2013]	hlf0168, hlf0171
	Consumption Module: Other	[2010-2013]	hlf0431, hlf0432, hlf0433, hlf0434
	Consumption Module: Repair	[2010-2013]	hlf0419, hlf0420, hlf0421, hlf0422
	Consumption Module: Telecommunication	[2010-2013]	hlf0391, hlf0392, hlf0393, hlf0394
	Consumption Module: Transportation	[2010-2013]	hlf0423, hlf0424, hlf0425, hlf0426
	Costs of comparable rental homes	[1984-2002,2005-2014]	hlf0094
	Costs of home ownership	[2010-2014,2016-2020]	hlf0084, hlf0090_h, hlf0090_v1, hlf0090_v2, hlf0601, hlf0602, hlf0603, hlf0604, hlf0605
	Dwelling / building type	[1986-2020], [1986-1990, 1991-2008, 2010-2018], [2009], [2016-2020]	hlf0155_h, hlf0155_v1, hlf0155_v2, hlf0596
	Financial burden of home ownership	[2016]	hlf0606
	Financial burden of home rental	[2016]	hlf0611
	Government-subsidized housing	[1984-2020], [1998-2015], [1986-2002,2008-2020]	hlf0011_h, hlf0011_v1, hlf0011_v2, hlf0011_v3, hlf0011_v4, hlf0073
	Hereditary lease interest	[1984-2013,2015-2020], [1986-2020], [1986-1987], [1988-1992], [1993-2020]	hlf0016, hlf0154_h, hlf0154_v1, hlf0154_v2, hlf0154_v3

continues on next page

Table 6 – continued from previous page

Questionnaire	Module	Years	Variables
	Home fixtures and furnishings	[1991], [2015-2020]	hlf0023, hlf0024, hlf0025, hlf0026, hlf0027, hlf0028, hlf0029, hlf0030, hlf0031, hlf0032, hlf0033, hlf0034, hlf0035, hlf0036, hlf0037, hlf0529, hlf0530, hlf0531
	Home ownership / rental	[1984-2020], [2003-2020]	hlf0001_h, hlf0001_v1, hlf0001_v2, hlf0001_v3, hlf0006, hlf0015
	Home ownership / rental, ownership acquisition	[1984-2020], [1984-1990,1999-2001], [1991-1998], [1991-1997]	hlf0007_h, hlf0007_v1, hlf0007_v2, hlf0007_v3
	Home ownership / rental, ownership transfer	[1999-2020]	hlf0007_v4, hlf0009
	Homeowner	[1990-2020], [1990-2002,2005-2012], [2003-2004], [2013-2020]	hlf0013_h, hlf0013_v1, hlf0013_v2, hlf0013_v3
	Loans, mortgages, building loan agreements	[1985-2020], [1985-1990, 1991-1998], [1999-2020], [1984-2020], [1984-1990, 1991-2001], [2002-2020]	hlf0087_h, hlf0087_v1, hlf0087_v2, hlf0088_h, hlf0088_v1, hlf0088_v2
	Material deprivation	(unregelmaessig) [2001-2015]	hlf0175, hlf0177, hlf0178_h, hlf0178_v1, hlf0178_v2, hlf0178_v3, hlf0178_v4, hlf0178_v5, hlf0179, hlf0180, hlf0181, hlf0183, hlf0185, hlf0186, hlf0187, hlf0188, hlf0189, hlf0190, hlf0191, hlf0192, hlf0193, hlf0194, hlf0195, hlf0444, hlf0613, hlf0622
	Modernization costs	[2016-2020]	hlf0599
	Monthly rent, heating, other expenses	[2016-2020]	hlf0607, hlf0608, hlf0610
	Monthly rent, heating, other expenses (Deutschmark)	[2002-2014,2016-2020]	hlf0081_v2
	Monthly rent, heating, other expenses: electricity	[2010-2014,2016-2020]	hlf0078, hlf0079
	Monthly rent, heating, other expenses: heating	[1986-2014,2016-2020], [1986-1990,1997-2001], [2002-2014,2016-2020]	hlf0069_h, hlf0069_v1, hlf0069_v5
	Monthly rent, heating, other expenses: heating and hot water	[1990,1996], [1991-1995]	hlf0069_v2, hlf0069_v3, hlf0069_v4
	Monthly rent, heating, other expenses: other	[1991-2014,2016-2020], [1991-2001], [1996-2014,2016-2020]	hlf0081_h, hlf0081_v1, hlf0082
	Monthly rent, heating, other expenses: rent	[1984-2020], [1984-2001]	hlf0074_h, hlf0074_v1

continues on next page

Table 6 – continued from previous page

Questionnaire	Module	Years	Variables
	Monthly rent, heating, other expenses: rent (Deutschmark)	[2002-2020]	hlf0074_v2
	Name and birth of children	[1984-2020], [2017,2020]	hlk0044_v1, hlk0044_v2
	Number of books in household	[2001,2006,2011,2016]	hlf0197
	Persons in household in need of care	[1984-2020], [2015-2020], [2016-2020]	hlf0291, hlf0292, hlf0300, hlf0301, hlf0302, hlf0303, hlf0304, hlf0315_h, hlf0315_v1, hlf0315_v2, hlf0315_v3, hlf0317_h, hlf0317_v1, hlf0317_v2, hlf0317_v3, hlf0319, hlf0320, hlf0321, hlf0322, hlf0331, hlf0332, hlf0369, hlf0370_h, hlf0370_v1, hlf0370_v2, hlf0446, hlf0448, hlf0595, hlf0631
	Pets	[2006,2011,2016], [2006,2011]	hlf0196, hlf0254, hlf0255, hlf0256, hlf0257, hlf0258, hlf0259
	Photovoltaic and solar thermal system	[2015-2016,2020]	hlf0532, hlf0535, hlf0536, hlf0537, hlf0538, hlf0539
	Reasons for moving	[1985-2013,2015,2017-2020]	hlf0108_h, hlf0108_v1, hlf0108_v10, hlf0108_v11, hlf0108_v12, hlf0108_v13, hlf0108_v14, hlf0108_v15, hlf0108_v2, hlf0108_v3, hlf0108_v4, hlf0108_v5, hlf0108_v6, hlf0108_v7, hlf0108_v8, hlf0108_v9, hlf0109, hlf0124, hlf0125
	Reasons for moving, comparison of old and new home	[2015,2017-2020], [2015,2017], [2015,2017,2019-2020]	hlf0524, hlf0525, hlf0526
	Residential area	(unregelmaessig) [1986-2019], [1994,1999,2004,2009,2014,2019], [2004,2009,2014,2016-2020], [1994,1999,2004,2009,2014,2019], [1986-2020], [1986-1987], [1988-1992], [1993-2020], [1986,1994,1999,2004,2009], [2014, 2019]	hlf0148, hlf0149, hlf0150, hlf0151, hlf0152, hlf0153_h, hlf0153_v1, hlf0153_v2, hlf0153_v3, hlj0004_v1, hlj0004_v2
	Residential area, distances	(unregelmaessig) [1986-2019]	hlf0135, hlf0136, hlf0137, hlf0138, hlf0139, hlf0140, hlf0141, hlf0142, hlf0143, hlf0144, hlf0145, hlf0146, hlf0147

continues on next page

Table 6 – continued from previous page

Questionnaire	Module	Years	Variables
	Residential area, neighbors	(unregelmaessig) [1986-2019]	hld0001, hld0002, hld0003
	Second Residence	[2011,2016], [2011], [2011,2016]	hlf0156, hlf0157, hlf0158
	Size and condition of home, rooms	[1984-2020]	hlf0021_h, hlf0021_v1, hlf0021_v2, hlf0021_v3
	Size and condition of home, size	[1986-2020]	hlf0018, hlf0019_h, hlf0019_v1, hlf0019_v2, hlf0019_v3, hlf0071_h, hlf0071_v1, hlf0071_v2, hlf0071_v3
	Type of energy used in household	[2015,2020]	hlf0591
	Type of energy used in household: biomass	[2015,2020]	hlf0582, hlf0583, hlf0584, hlf0585, hlf0586
	Type of energy used in household: coal	[2015,2020]	hlf0570, hlf0571, hlf0572_v1, hlf0572_v2, hlf0573, hlf0574, hlf0575
	Type of energy used in household: district heat	[2015,2020]	hlf0540, hlf0541, hlf0542, hlf0543, hlf0544
	Type of energy used in household: electricity	[2015,2020]	hlf0557, hlf0558, hlf0559_v1, hlf0559_v2, hlf0560, hlf0561, hlf0562, hlf0563
	Type of energy used in household: environmental heat	[2015,2020]	hlf0589, hlf0590
	Type of energy used in household: gas	[2015,2020]	hlf0545, hlf0546, hlf0547_v1, hlf0547_v2, hlf0548, hlf0549, hlf0550
	Type of energy used in household: heating oil	[2015,2020]	hlf0564, hlf0565, hlf0566_v1, hlf0566_v2, hlf0567, hlf0568, hlf0569
	Type of energy used in household: liquid gas	[2015,2020]	hlf0551, hlf0552, hlf0553_v1, hlf0553_v2, hlf0554, hlf0555, hlf0556
	Type of energy used in household: solar	[2015,2020]	hlf0587, hlf0588
	Type of energy used in household: wood/pellets	[2015,2020]	hlf0576, hlf0577, hlf0578_v1, hlf0578_v2, hlf0579, hlf0580, hlf0581

2.7 Education and Qualification

Education is one of the cornerstones of society today. The education, training, and qualifications modules provide extensive information on educational attainment and outcomes, the level of completed education and training, reasons for not completing education or training, educational goals, and much more, along with data on children's skill development, for instance, whether they are able to speak in full sentences or use scissors.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Completed education and training, Current year	[1993-2020]	plg0074
	Completed education and training, Degree Nonresponse	[2014-2020]	plg0268
	Completed education and training, Degree Recignized in Germany	[2015-2020]	plg0284
	Completed education and training, Education Completed	[2000-2020], [2005-2020]	plg0076, plg0077
	Completed education and training, Field of study	[1985-2009]	p_field
	Completed education and training, General school-leaving certificate	[1985-2020], [1991]	plg0078_h, plg0078_v1, plg0078_v2
	Completed education and training, Nonresponse	[2004-2020]	plg0075
	Completed education and training, Since previous year	[1985-2020], [1993-2020]	plg0072, plg0073
	Completed education and training, Type of college degree	[1985-2009]	p_degree
	Completed education and training, University school-leaving certificate	[1985-2008], [2009-2013], [2014-2020]	plg0079_v1, plg0079_v3, plg0079_v4
	Further education, course details and motives for participation, Course 1 Adjusting to new Demands	[1989,1993,2000,2004,2008]	plg0144

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Further education, course details and motives for participation, Course 1 Applicability	[2004,2008]	plg0117
	Further education, course details and motives for participation, Course 1 Career Change	[1989,1993,2000,2004,2008]	plg0135
	Further education, course details and motives for participation, Course 1 Duration	[1989,1993]	plg0120_v1, plg0120_v2, plg0120_v3, plg0120_v4
	Further education, course details and motives for participation, Course 1 During Working Hours	[1989,1993,2000,2004,2008]	plg0150
	Further education, course details and motives for participation, Course 1 Financial Support	[1989,1993,2000,2004,2008], [2000], [2004,2008]	plg0177_h, plg0177_v1, plg0177_v2, plg0177_v3, plg0177_v4, plg0177_v5, plg0177_v6
	Further education, course details and motives for participation, Course 1 Hours of Instruction	[1989,1993,2000,2004,2008]	plg0129
	Further education, course details and motives for participation, Course 1 Introduction New Job	[1989,1993,2000,2004,2008]	plg0138
	Further education, course details and motives for participation, Course 1 No Own Costs	[2000,2004,2008]	plg0174
	Further education, course details and motives for participation, Course 1 Organizer	[1989,1993,2000,2004,2008]	plg0155_h, plg0155_v1, plg0155_v10, plg0155_v2, plg0155_v3, plg0155_v4, plg0155_v5, plg0155_v6, plg0155_v7, plg0155_v8, plg0155_v9
	Further education, course details and motives for participation, Course 1 Other	[1989,1993,2000,2004,2008]	plg0147
	Further education, course details and motives for participation, Course 1 Own Costs	[1989,1993,2000,2004,2008], [1989,1993,2000], [2004,2008]	plg0169_h, plg0169_v1, plg0169_v2
	Further education, course details and motives for participation, Course 1 Participation Certificate	[1989,1993,2000,2004,2008]	plg0184

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Further education, course details and motives for participation, Course 1 Qualification for Promotion	[1989,1993,2000,2004,2008]	plg0141
	Further education, course details and motives for participation, Course 1 Start	[1989,1993,2000,2004,2008], [1989,1993], [2000,2004,2008], [1989,1993,2000,2004,2008]	plg0108_h, plg0108_v1, plg0108_v2, plg0111
	Further education, course details and motives for participation, Course 1 Telecourse	[1989,1993,2000,2004,2008]	plg0132
	Further education, course details and motives for participation, Course 1 pay off	[2004,2008]	plg0114
	Further education, course details and motives for participation, Course 2 Adjusting to new Demands	[1989,1993,2000,2004,2008]	plg0145
	Further education, course details and motives for participation, Course 2 Applicability	[2004,2008]	plg0118
	Further education, course details and motives for participation, Course 2 Career Change	[1989,1993,2000,2004,2008]	plg0136
	Further education, course details and motives for participation, Course 2 Duration	[1989,1993]	plg0121_v1, plg0121_v2, plg0121_v3, plg0121_v4
	Further education, course details and motives for participation, Course 2 During Working Hours	[1989,1993,2000,2004,2008]	plg0151
	Further education, course details and motives for participation, Course 2 Financial Support	[2000,2004,2008], [2000], [2004,2008]	plg0182_h, plg0182_v1, plg0182_v2
	Further education, course details and motives for participation, Course 2 Hours of Instruction	[1989,1993,2000,2004,2008]	plg0130
	Further education, course details and motives for participation, Course 2 Introduction New Job	[1989,1993,2000,2004,2008]	plg0139

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Further education, course details and motives for participation, Course 2 No Own Costs	[2000,2004,2008]	plg0175
	Further education, course details and motives for participation, Course 2 Organizer	[2000,2004,2008]	plg0164
	Further education, course details and motives for participation, Course 2 Other	[1989,1993,2000,2004,2008]	plg0148
	Further education, course details and motives for participation, Course 2 Own Costs	[2000,2004,2008], [2000], [2004,2008]	plg0171_h, plg0171_v1, plg0171_v2
	Further education, course details and motives for participation, Course 2 Participation Certificate	[1989,1993,2000,2004,2008]	plg0185
	Further education, course details and motives for participation, Course 2 Qualification for Promotion	[1989,1993,2000,2004,2008]	plg0142
	Further education, course details and motives for participation, Course 2 Start	[1989,1993,2000,2004,2008], [1989,1993], [2000,2004,2008], [1989,1993,2000,2004,2008]	plg0109_h, plg0109_v1, plg0109_v2, plg0112
	Further education, course details and motives for participation, Course 2 Telecourse	[1989,1993,2000,2004,2008]	plg0133
	Further education, course details and motives for participation, Course 2 pay off	[2004,2008]	plg0115
	Further education, course details and motives for participation, Course 3 Adjusting to new Demands	[1989,1993,2000,2004,2008]	plg0146
	Further education, course details and motives for participation, Course 3 Applicability	[2004,2008]	plg0119
	Further education, course details and motives for participation, Course 3 Career Change	[1989,1993,2000,2004,2008]	plg0137

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Further education, course details and motives for participation, Course 3 Duration	[1989,1993]	plg0122_v1, plg0122_v2, plg0122_v3, plg0122_v4
	Further education, course details and motives for participation, Course 3 During Working Hours	[1989,1993,2000,2004,2008]	plg0152
	Further education, course details and motives for participation, Course 3 Financial Support	[2000,2004,2008], [2000], [2004,2008]	plg0183_h, plg0183_v1, plg0183_v2
	Further education, course details and motives for participation, Course 3 Hours of Instruction	[1989,1993,2000,2004,2008]	plg0131
	Further education, course details and motives for participation, Course 3 Introduction New Job	[1989,1993,2000,2004,2008]	plg0140
	Further education, course details and motives for participation, Course 3 No Own Costs	[2000,2004,2008]	plg0176
	Further education, course details and motives for participation, Course 3 Organizer	[2000,2004,2008]	plg0165
	Further education, course details and motives for participation, Course 3 Other	[1989,1993,2000,2004,2008]	plg0149
	Further education, course details and motives for participation, Course 3 Own Costs	[2000,2004,2008], [2000], [2004,2008]	plg0172_h, plg0172_v1, plg0172_v2
	Further education, course details and motives for participation, Course 3 Participation Certificate	[1989,1993,2000,2004,2008]	plg0186
	Further education, course details and motives for participation, Course 3 Qualification for Promotion	[1989,1993,2000,2004,2008]	plg0143
	Further education, course details and motives for participation, Course 3 Start	[1989,1993,2000,2004,2008], [1989,1993], [2000,2004,2008], [1989,1993,2000,2004,2008]	plg0110_h, plg0110_v1, plg0110_v2, plg0113

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Further education, course details and motives for participation, Course 3 Telecourse	[1989,1993,2000,2004,2008]	plg0134
	Further education, course details and motives for participation, Course 3 pay off	[2004,2008]	plg0116
	Further education, course details and motives for participation, Course Subject / Content	[1989,1993]	plg0153
	Further education, course details and motives for participation, Financial Support	[1989,1993]	plg0167, plg0168
	Further education, course details and motives for participation, Initiative for taking Course	[1989,1993]	plg0166
	Further education, course details and motives for participation, Pay Off	[1989]	plg0187, plg0188, plg0189
	Further training measures, Further professional training	[2014-2015], [2016-2020]	plg0269_v1, plg0269_v2
	Further training measures, Trainig measures prev year	[2014-2020]	plg0270, plg0271
	Further training, financing	[2015-2018,2020]	plg0285, plg0286, plg0287, plg0288, plg0289, plg0290, plg0291
	Further training, reasons for not taking part	[2014]	plg0277, plg0278, plg0279, plg0280, plg0281
	Further training, suggested / provided by employer	[2014]	plg0274
	Further training, suggested / provided by employere	[2014]	plg0273
	Lifelong learning	[2014]	plg0266
	Vocational training, Currently in education / training	[1984-2020], [2020]	plg0012_v1, plg0012_v2
	Vocational training, General school	[1984-2015], [2016-2020]	plg0013_v1, plg0013_v3
	Vocational training, Scholarship	[2007-2020]	plg0015_h, plg0015_v1, plg0015_v2, plg0015_v3, plg0015_v4

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Vocational training, University	[1984-1995], [1999-2008], [2009-2012], [2013-2020]	plg0014_v1, plg0014_v2, plg0014_v3, plg0014_v4, plg0014_v5, plg0014_v6, plg0014_v7
Youth Questionnaire	Education and career plans	[2013-2019], [2000-2017], [2013-2019], [2000-2017]	j_isco08_jobwish, j_isco88_jobwish, j_kldb2010_jobwish, j_kldb92_jobwish
	Education and career plans, Apprenticeship	[2000-2020], [2001-2020]	jl0177, jl0182, jl0183, jl0203
	Education and career plans, Career Training	[2014-2020]	jl0438, jl0439
	Education and career plans, Engineering school	[2013-2020]	jl0440, jl0441
	Education and career plans, Exploring Skills	[2001-2020]	jl0205
	Education and career plans, Financial Independence	[2000-2020]	jl0197, jl0198
	Education and career plans, Informed about Future Occupation	[2001-2020]	jl0201
	Education and career plans, No Particular Plans	[2001-2020]	jl0204
	Education and career plans, Occupational Foundation	[2000-2020]	jl0179
	Education and career plans, Occupational Integration	[2000-2020]	jl0178, jl0180, jl0181
	Education and career plans, Parents Suggestions	[2001-2020]	jl0202
	Education and career plans, Preferred Occupation	[2000-2020]	jl0199
	Education and career plans, Vocational School	[2000-2020]	jl0184, jl0185
	Education and career plans, Volunteering	[2000-2020]	jl0186, jl0187
	Educational aspirations, Apprenticeship	[2014-2020]	jl0504
	Educational aspirations, Aspired school-leaving qualification	[2000-2020], [2000], [2001-2020], [2000-2020]	jl0130_h, jl0130_v1, jl0130_v2, jl0131
	Educational aspirations, Career Training	[2000-2020]	jl0193
	Educational aspirations, Civil Servant Training	[2000-2020]	jl0192

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	Educational aspirations, Completed Apprenticeship	[2000-2020]	jl0189
	Educational aspirations, Engineering school	[2000-2020]	jl0194
	Educational aspirations, Future Apprenticeship	[2000-2020], [2003-2020]	jl0188, jl0196
	Educational aspirations, Trade and Technical School	[2000-2020]	jl0191
	Educational aspirations, University	[2000-2020]	jl0195
	Educational aspirations, Vocational School	[2000-2020]	jl0190
	School, attendance & homework, Class Representative	[2000-2020]	jl0139
	School, attendance & homework, Course type	[2000-2020]	jl0162, jl0163
	School, attendance & homework, Extracurricular activities	[2000-2020]	jl0141, jl0142, jl0143, jl0144, jl0145, jl0146
	School, attendance & homework, First Foreign Language	[2000-2020], [2000], [2001-2005], [2006-2020]	jl0132_h, jl0132_v1, jl0132_v2, jl0132_v3
	School, attendance & homework, Grade / Year	[2014-2020]	jl0434
	School, attendance & homework, Grades / Points	[2000-2020]	jl0152, jl0153, jl0154, jl0155, jl0156, jl0157
	School, attendance & homework, Number classmates	[2000-2020], [2000], [2001-2018]	jl0176_h, jl0176_v1, jl0176_v2
	School, attendance & homework, Private School	[2001-2018]	jl0138
	School, attendance & homework, Satisfaction with grades	[2000-2020]	jl0147, jl0148, jl0149, jl0150
	School, attendance & homework, School attendance abroad	[2000-2020]	jl0137_h, jl0137_v1, jl0137_v2, jl0435, jl0436
	School, attendance & homework, School recommendation	[2001-2018]	jl0151
	School, attendance & homework, School-leaving certificate	[2000-2020], [2000-2011], [2012-2020]	jl0127_h, jl0127_v1, jl0127_v2
	School, attendance & homework, Second Foreign Language	[2000-2020], [2000], [2001-2005], [2006-2020]	jl0133_h, jl0133_v1, jl0133_v2, jl0133_v3

continues on next page

Table 7 – continued from previous page

Questionnaire	Module	Years	Variables
	School, attendance & homework, Still in School	[2000-2020], [2000-2002], [2003-2005], [2006-2020]	jl0125_h, jl0125_v1, jl0125_v2, jl0125_v3
	School, attendance & homework, Student Body President	[2000-2020]	jl0140
	School, attendance & homework, Year of leaving school	[2000-2020]	jl0126
	School, attendance & homework, Year repeated	[2000-2020]	jl0164, jl0165, jl0166
Mother and Child Instruments	Educational aspirations, Ideal school completion	[2003-2020]	idegrad1, idegrad2, idegrad3
	Educational aspirations, intermediate secondary	[2003-2020]	probgra2
	Educational aspirations, lower secondary	[2003-2020]	probgra1
	Educational aspirations, upper secondary	[2003-2020]	probgra3
	School and homework	[2003-2020]	scolcon1, scolcon2, scolcon3, scolcon4, scolcon5, scolcon6, scolcon7
	School and homework, Comprehensive school	[2003-2020]	curscol7
	School and homework, Grammar secondary class	[2003-2020]	curscol6
	School and homework, Intermediae secondary schol	[2003-2020]	curscol5
	School and homework, Last report mark	[2003-2020]	lamark, matmark, nomark
	School and homework, Other schoool	[2003-2020]	curscol8
	School and homework, Place	[2003-2020]	hwplace
	School and homework, Primary school	[2003-2020]	curscol1
	School and homework, Second general school	[2003-2020]	curscol4
	School and homework, Special pedagogic concept	[2003-2020]	curscol2
	School and homework, Special school	[2003-2020]	curscol3
	School and homework, Support	[2003-2020]	hwsupprt
	School enrollment	[2003-2020]	sclenrolm, sclenroln, sclenroly

2.8 Attitudes, Values, and Personality

The attitudes, values, and personality modules provide extensive information on respondents' personality traits, political orientations, concerns, satisfaction with different aspects of life, willingness to take risks, and much more.

Attitudes, Values and
Personality

Questionnaire	Module	Years	Variables
Individual Questionnaire	10,000-euro question	[2010,2017]	plh0134, plh0135, plh0136
	Affective well-being	[2007-2020]	plh0184, plh0185, plh0186, plh0187
	Anomie	(irregular) [1990-2018]	plh0188, plh0189, plh0190, plh0191
	Attitudes towards genders	[2019]	plh0395i01, plh0395i02, plh0395i03, plh0395i04, plh0395i05, plh0395i06
	Attitudes towards refugees	[2016,2018,2020]	plj0433, plj0434, plj0435, plj0436, plj0437, plj0438, plj0439, plj0440, plj0441, plj0442, plj0443
	Big Five personality traits	[2005,2009,2012-2013,2017,2019], [2009,2012-2013,2017,2019]	plh0212, plh0213, plh0214, plh0215, plh0216, plh0217, plh0218, plh0219, plh0220, plh0221, plh0222, plh0223, plh0224, plh0225, plh0226, plh0255
	Bundestag election	[2014,2018]	plh0333
	Depressive traits	[2016,2019]	plh0339, plh0340, plh0341, plh0342
	Discrimination	[2019]	plh0387i01, plh0387i02, plh0387i03, plh0387i04, plh0387i05, plh0387i06, plh0387i07, plh0387i08, plh0387i09, plh0387i10, plh0387i11
	Donation of blood	[2010]	plh0131_v1, plh0131_v2, plh0132, plh0133
	Donations	[2010,2015,2018,2020]	plh0129, plh0130
	Donations of goods	[2010,2020]	plj0108, plj0109, plj0110, plj0111, plj0112, plj0113, plj0114, plj0115
	Flourishing	[2015-2020]	plh0334

continues on next page

Table 8 – continued from previous page

Questionnaire	Module	Years	Variables
	Goals in life (Kluckhohn)	(irregular) [1990-2016], [2013,2017-2019], [2016]	plh0104, plh0105, plh0106, plh0107, plh0108, plh0109, plh0110, plh0111, plh0112, plh0343_v1, plh0343_v2
	Impulsivity, patience	[2008,2013,2018]	plh0253, plh0254
	Income justice, general	[2005]	plh0116, plh0117, plh0118, plh0119, plh0120, plh0121, plh0122, plh0123, plh0124, plh0125, plh0126, plh0127
	Life satisfaction	[1984-2020]	plh0182
	Locus of control	[1994-1996]	plh0369, plh0370, plh0371, plh0372, plh0373, plh0374, plh0375, plh0376, plh0377_v1, plh0378_v1, plh0379_v1, plh0380_v1, plh0381_v1, plh0382_v1, plh0383_v1, plh0384_v1, plh0385_v1, plh0386_v1
	Locus of control, rephrased	[2005,2010,2015-2016,2020]	plh0377_v2, plh0378_v2, plh0379_v2, plh0380_v2, plh0381_v2, plh0382_v2, plh0383_v2, plh0384_v2, plh0385_v2, plh0386_v2
	Loneliness	[2013,2016-2019], [2013,2016-2020]	plj0587, plj0588, plj0589
	Lottery question	[2004,2009,2014]	plh0203
	Money and account balance	[2016,2018]	plh0344, plh0345, plh0346
	Optimism/pessimism	[1999,2005,2009,2014,2019]	plh0244
	Organisational and community membership	(irregular) [1985-2019], (irregular) [1990-2019], (irregular) [1985-2019], [1985,1989,1993], (irregular) [1990-2019], (irregular) [2001-2019], [1998,2001,2003,2007,2011,2019], [2003,2007,2011]	plh0263_h, plh0263_v2, plh0264_h, plh0264_v1, plh0264_v2, plh0265, plh0266, plh0267
	Organizational and community membership	[1985,1989,1993]	plh0263_v1
	Policy objectives (Inglehart Index)	[1984-1986,1996,2006,2016]	plh0054, plh0056, plh0058, plh0061
	Political Tendency, Left-Right	[2005,2009,2014,2019]	plh0004
	Political influence	[2019]	plh0397i01, plh0397i02, plh0397i03, plh0397i04, plh0397i05
	Political orientation	[1985-2020]	plh0007

continues on next page

Table 8 – continued from previous page

Questionnaire	Module	Years	Variables
	Political orientation (Party Affiliation)	[1984-2020], [1984-1989], [1990], [1991], [1992], [1993], [1994-2020], [1987-1988]	plh0012_h, plh0012_v1, plh0012_v2, plh0012_v3, plh0012_v4, plh0012_v5, plh0012_v6, plh0013_v1
	Political orientation (Party Preference)	[1984-2020]	plh0011_h, plh0011_v1, plh0011_v2, plh0013_h, plh0013_v2
	Reciprocity	[2005,2010,2015-2020]	plh0206i01, plh0206i02, plh0206i03, plh0206i04, plh0206i05, plh0206i06
	Religious affiliation	(irregular) [1990-2020], [1990-1991,1997], [1990], [1991,1997], [1990-1991,1997], [2003], [2007,2011], [2015], [2013,2016-2020]	plh0258_h, plh0258_v1, plh0258_v10, plh0258_v11, plh0258_v12, plh0258_v13, plh0258_v2, plh0258_v3, plh0258_v4, plh0258_v5, plh0258_v6, plh0258_v7, plh0258_v8, plh0258_v9
	Risk aversion in different domains	[2004,2009,2014]	plh0197, plh0198, plh0199, plh0200, plh0201, plh0202
	Risk aversion in general	[2004,2006,2008-2020], [2013], [2004,2006,2008-2020]	plh0204_h, plh0204_v1, plh0204_v2
	Satisfaction with various aspects	(irregular) [1989-2019], [1984-2020], [2008-2020], [1984-2020], [1984-1990,1993-2020], [1984-2020], [2004-2020], [1984-2020], [1984-1989,1991-1994,1996-2020], [1990,1997-2020], [2006-2020], [2006,2011-2013,2016]	plh0164, plh0171, plh0172, plh0173, plh0174, plh0175, plh0176, plh0177, plh0178, plh0179, plh0180, plh0181
	Self-esteem	[2010,2015-2020]	plh0206i11
	Social justice	[2019]	plh0396i01, plh0396i02, plh0396i03, plh0396i04
	Social responsibility	[1997,2002,2017]	plh0016, plh0017, plh0018, plh0019, plh0020, plh0021, plh0022, plh0023, plh0024, plh0025, plh0026
	Tendency to forgive	[2010,2015-2016,2020]	plh0206i07, plh0206i08, plh0206i09, plh0206i10
	Trust, trustworthiness and fairness	[2003,2008,2013,2018]	pld0043, pld0044, pld0045, plh0192, plh0193, plh0194, plh0195, plh0196

continues on next page

Table 8 – continued from previous page

Questionnaire	Module	Years	Variables
	Wage justice	[2009,2011,2013,2015,2017,2019], [2015], [2017,2019], [2015], [2017-2019]	plh0138, plh0139, plh0140, plh0141, plh0337_v1, plh0337_v2, plh0338_v1, plh0338_v2
	Well-being aspects	[1994,1998-1999]	plh0091_v2, plh0092_v2, plh0093_v2, plh0094_v2, plh0095_v2, plh0096_v2, plh0097_v2, plh0098_v2, plh0099_v2, plh0100_v2, plh0101, plh0102, plh0103
	Well-being aspects, East Germany	[1990-1991]	plh0091_v1, plh0092_v1, plh0093_v1, plh0094_v1, plh0095_v1, plh0096_v1, plh0097_v1, plh0098_v1, plh0099_v1, plh0100_v1
	Worries	[1984-2020]	plh0032, plh0033, plh0034, plh0035, plh0038, plh0040, plh0042, plh0043, plh0046, plh0047, plh0335, plh0336
Youth Questionnaire	Affective well-being	[2007-2020]	jl0381, jl0382, jl0383, jl0384
	Attitudes and opinions	[2000-2020]	jl0329, jl0330, jl0360, jl0364
	Big Five personality traits	[2006-2020]	jl0365, jl0366, jl0367, jl0368, jl0369, jl0370, jl0371, jl0372, jl0373, jl0374, jl0375, jl0376, jl0377, jl0378, jl0379, jl0380
	Future	[2000-2020]	jl0222, jl0223, jl0224, jl0225, jl0226, jl0227, jl0228, jl0229, jl0230, jl0231, jl0232
	Life satisfaction	[2006-2020]	jl0392
	Locus of control	[2006-2020]	jl0350_v1, jl0351_v1, jl0352_v1, jl0353_v1, jl0354_v1, jl0355_v1, jl0356_v1, jl0357_v1, jl0358_v1, jl0359_v1
	Locus of control, rephrased	[2001-2005]	jl0350_v2, jl0351_v2, jl0352_v2, jl0353_v2, jl0354_v2, jl0355_v2, jl0356_v2, jl0357_v2, jl0358_v2, jl0359_v2
	Political orientation	[2006-2020]	jl0388, jl0389, jl0390, jl0391
	Risk aversion in general	[2006-2020]	jl0349
	Social justice	[2019-2020]	jl1909, jl1910, jl1911, jl1912
	Sources of social inequality	[2000-2020]	jl0337, jl0338, jl0339, jl0340, jl0341, jl0342, jl0343, jl0344, jl0345, jl0346, jl0347, jl0348
	Trust	[2006-2020]	jl0361, jl0362, jl0363
Mother and Child Instruments	Big Five personality traits	[2003-2020]	char10, char1a, char1b, char2, char3, char4, char5, char6, char7, char8, char9

continues on next page

Table 8 – continued from previous page

Questionnaire	Module	Years	Variables
	Strengths and difficulties questionnaire	[2003-2020]	behav1, behav10, behav11, behav12, behav13, behav14, behav15, behav16, behav17, behav18, behav2, behav3, behav4, behav5, behav6, behav7, behav8, behav9
	Temperament	[2003-2020]	temp1, temp2, temp3, temp4, temp5, temp6, temp7
	Vineland adaptive behavior scales (Movements)	[2003-2020]	mvmn1, mvmn3, mvmn4, mvmn5, mvmn6
	Vineland adaptive behavior scales (Playing)	[2003-2020]	sclr2, sclr3, sclr4, sclr5, sclr6
	Vineland adaptive behavior scales (Skills)	[2003-2020]	skll1, skll2, skll3, skll4, skll5
	Vineland adaptive behavior scales (Speaking and Listening)	[2003-2020]	spch3, spch5, spch6, spch7, spch8

2.9 Time Use and Environmental Behavior

The modules on time use and environmental behavior give information on time commitments, free time, and time planning as well as environmental awareness, for instance, the use of public transport and different energy sources, as well as what respondents think about renewable energies.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Computer usage: Private	[1997,2000-2001], [1997,2001], [2000], [1997,2001], [1997], [2001]	pli0066_h, pli0066_v1, pli0066_v2, pli0067_h, pli0067_v1, pli0067_v2
	Computer usage: Private (Internet)	[2001]	pli0068, pli0069
	Computer usage: Work	[1997,1999,2001], [2000], [1997,2001], [1997], [2001]	pli0070_v1, pli0070_v2, pli0071_h, pli0071_v1, pli0071_v2
	Computer usage: Work (Internet)	[2001]	pli0072, pli0073

continues on next page

Table 9 – continued from previous page

Questionnaire	Module	Years	Variables
	Leisure activities (long)	[2019]	plh0390, plh0391, plh0392, plh0393, plh0394, pli0083, pli0084, pli0085_v1, pli0085_v2, pli0087, pli0088, pli0089, pli0090_v1, pli0090_v2, pli0090_v3, pli0091_h, pli0091_v1, pli0091_v2, pli0165, pli0178, pli0182
	Leisure activities (long): Art and Music	(unregelmaessig) [1990-2019], (unregelmaessig) [2001-2017], (unregelmaessig) [1990-2019]	pli0093_h, pli0093_v1, pli0093_v2
	Leisure activities (long): Politics	(unregelmaessig) [1984-2019], [1984], (unregelmaessig) [1985-2017], (unregelmaessig) [1990-2019]	pli0097_h, pli0097_v1, pli0097_v2, pli0097_v3
	Leisure activities (long): Religion	(unregelmaessig) [1990-2019], (unregelmaessig) [1990-2017], (unregelmaessig) [1990-2019]	pli0098_h, pli0098_v1, pli0098_v2
	Leisure activities (long): Socializing	(unregelmaessig) [1990-2019], [1984], (unregelmaessig) [1985-2017], [1984], (unregelmaessig) [1985-2017]	pli0079, pli0080, pli0081, pli0082, pli0094_v1, pli0094_v2, pli0095_v1, pli0095_v2
	Leisure activities (long): Sports	(unregelmaessig) [1984-2019], [1984], (unregelmaessig) [1985-2017], (unregelmaessig) [1990-2019]	pli0092_h, pli0092_v1, pli0092_v2, pli0092_v3
	Leisure activities (long): Voluntary work	(unregelmaessig) [1984-2019], [1984], (unregelmaessig) [1985-2017], (unregelmaessig) [1990-2019]	pli0096_h, pli0096_v1, pli0096_v2, pli0096_v3
	Time use for different activities (Saturdays)	(unregelmaessig) [1990-2019], (unregelmaessig) [2001-2019], (unregelmaessig) [2003-2019], [2008-2013,2015,2017,2019-2020]	pli0003_h, pli0003_v1, pli0003_v2, pli0003_v3, pli0003_v4, pli0005, pli0036, pli0054, pli0055, pli0056, pli0060
	Time use for different activities (Saturdays): Childcare	(unregelmaessig) [1990-2019], (unregelmaessig) [1993-2019]	pli0019_h, pli0019_v1, pli0019_v2, pli0019_v3, pli0019_v4
	Time use for different activities (Saturdays): Chores	(unregelmaessig) [1990-2019], (unregelmaessig) [1990-2019]	pli0012_h, pli0012_v1, pli0012_v2, pli0012_v3

continues on next page

Table 9 – continued from previous page

Questionnaire	Module	Years	Variables
	Time use for different activities (Saturdays): Education	(unregelmaessig) [1990-2019], (unregelmaessig) [1990-2019]	pli0024_h, pli0024_v1, pli0024_v2, pli0024_v3
	Time use for different activities (Saturdays): Repair work	(unregelmaessig) [1990-2019], (unregelmaessig) [1993-2019]	pli0031_h, pli0031_v1, pli0031_v2, pli0031_v3, pli0031_v4
	Time use for different activities (Sundays)	[1984], (unregelmaessig) [1992-2019], (unregelmaessig) [1984-2019], (unregelmaessig) [1990-2019], (unregelmaessig) [2001-2019], (unregelmaessig) [2003-2019]	pli0002_v1, pli0002_v2, pli0007_v1, pli0007_v2, pli0007_v3, pli0007_v4, pli0007_v5, pli0010, pli0011, pli0057, pli0058
	Time use for different activities (Sundays): Childcare	(unregelmaessig) [1985-2019], (unregelmaessig) [1992-2019]	pli0022_h, pli0022_v1, pli0022_v2, pli0022_v3, pli0022_v4
	Time use for different activities (Sundays): Chores	(unregelmaessig) [1984-2019], (unregelmaessig) [1992-2019]	pli0016_h, pli0016_v1, pli0016_v2, pli0016_v3, pli0016_v4
	Time use for different activities (Sundays): Education	(unregelmaessig) [1984-2019], (unregelmaessig) [1992-2019]	pli0028_h, pli0028_v1, pli0028_v2, pli0028_v3, pli0028_v4
	Time use for different activities (Sundays): Repair work	[1984-1990], (unregelmaessig) [1992-2019]	pli0034_v1, pli0034_v2, pli0034_v3, pli0034_v4
	Time use for different activities (weekdays)	[1984], [1992-2020], [1990-2020], [2001-2020], [1984-2020], [2003-2020], [2008-2013,2015,2017,2019-2020]	pli0001_v1, pli0001_v2, pli0038_v1, pli0038_v2, pli0038_v3, pli0038_v4, pli0040, pli0046, pli0051, pli0052, pli0059
	Time use for different activities (weekdays): Childcare	[1985-2020], [1985-1991], [1990], [1992-2020]	pli0044_h, pli0044_v1, pli0044_v2, pli0044_v3
	Time use for different activities (weekdays): Chores	[1984-2020], [1984-1991], [1990], [1992-2020]	pli0043_h, pli0043_v1, pli0043_v2, pli0043_v3
	Time use for different activities (weekdays): Education	[1984-2020]	pli0047_v1, pli0047_v2, pli0047_v3
	Time use for different activities (weekdays): Repair work	[1984-2020], [1984-1991], [1990], [1992-2020]	pli0049_h, pli0049_v1, pli0049_v2, pli0049_v3
	Trip to work	[1985,1993,1995], [1998,2003], (unregelmaessig) [1985-2019]	plb0144, plb0145, plb0156_v1, plb0157_v1, plb0157_v2, plb0158
	Trip to work: Car passenger	[1985], [1998,2003]	plb0175_v1, plb0175_v2
	Trip to work: Cost	[1985,1993,1995]	plb0142, plb0143

continues on next page

Table 9 – continued from previous page

Questionnaire	Module	Years	Variables
	Trip to work: Starting year	[1997,1999], [1997], [1999]	plb0146_h, plb0146_v1, plb0146_v2
	Trip to work: Time	[1985,1990-1993,1995,1998,2003]	plb0147, plb0148, plb0149, plb0150, plb0151, plb0152, plb0153, plb0154, plb0155
	Trip to work: Way to work	[1998, 2003], [1985, 1990], [1993,1995,1997-1999], [2000-2013,2015,2017,2019]	plb0156_v2, plb0159_v1, plb0159_v2, plb0159_v3
	Use of transportation for errands	[1998], [2003, 2018], [1998], [2003,2018], [1998], [2003,2018], [1998], [2003, 2018], [1998], [2003, 2018], [1998,2003,2018]	pli0133_v1, pli0133_v2, pli0134_v1, pli0134_v2, pli0135_v1, pli0135_v2, pli0136_v1, pli0136_v2, pli0137_v1, pli0137_v2, pli0138
	Use of transportation for excursions	[1998], [2003,2018], [1998], [2003, 2018], [1998], [2003, 2018], [1998], [2003,2018], [1998], [2003,2018], [1998,2003,2018]	pli0139_v1, pli0139_v2, pli0140_v1, pli0140_v2, pli0141_v1, pli0141_v2, pli0142_v1, pli0142_v2, pli0143_v1, pli0143_v2, pli0144
	Use of transportation in general	[1998,2003,2018]	pli0106
	Use of transportation in general: Licenses	[1985,1991,1998,2003], [1985,1991,1995,1998,2003], [1985], [1991,1995,1998,2003]	pli0101, pli0102, pli0103, pli0104, pli0105_h, pli0105_v1, pli0105_v2
	Use of transportation in general: Public transportation	[1998,2003,2018]	pli0107, pli0108_h, pli0108_v1, pli0108_v2, pli0109_h, pli0109_v1, pli0109_v2, pli0110, pli0111_h, pli0111_v1, pli0111_v2, pli0112_h, pli0112_v1, pli0112_v2, pli0113, pli0114_h, pli0114_v1, pli0114_v2, pli0115_h, pli0115_v1, pli0115_v2, pli0116, pli0117_h, pli0117_v1, pli0117_v2, pli0118_h, pli0118_v1, pli0118_v2, pli0119, pli0120_h, pli0120_v1, pli0120_v2, pli0121_h, pli0121_v1, pli0121_v2, pli0122, pli0123, pli0124, pli0125, pli0126
	Use of transportation in leisure time	[1998], [2003, 2018], [1998], [2003, 2018], [1998], [2003,2018], [1998], [2003, 2018], [1998], [2003, 2018], [1998], [2003,2018]	pli0145_v1, pli0145_v2, pli0146_v1, pli0146_v2, pli0147_v1, pli0147_v2, pli0148_v1, pli0148_v2, pli0149_v1, pli0149_v2, pli0150_v1, pli0150_v2

continues on next page

Table 9 – continued from previous page

Questionnaire	Module	Years	Variables
	Use of transportation to take children places	[1998], [2003, 2018], [1998], [2003, 2018], [1998], [2003, 2018], [1998], [2003, 2018], [1998], [2003, 2018], [1998]	pli0151_v1, pli0151_v2, pli0152_v1, pli0152_v2, pli0153_v1, pli0153_v2, pli0154_v1, pli0154_v2, pli0155_v1, pli0155_v2, pli0156
	Use of transportation to work	[1998], [2003,2018,2020], [1998], [2003,2018,2020], [1998], [2003,2018,2020], [1998], [2003,2018,2020], [1998], [2003,2018,2020], [1998,2003,2018,2020]	pli0127_v1, pli0127_v2, pli0128_v1, pli0128_v2, pli0129_v1, pli0129_v2, pli0130_v1, pli0130_v2, pli0131_v1, pli0131_v2, pli0132
Youth Questionnaire	Leisure and hobbies	[2001-2020], [2001-2018], [2006-2020]	jl0058, jl0064, jl0065, jl0066, jl0067, jl0072, jl0073
	Leisure and hobbies: Internet	[2006-2013], [2014-2020], [2013], [2014-2020]	jl0060_v1, jl0060_v2, jl0060_v3, jl0060_v4
	Leisure and hobbies: Music	[2001-2020]	jl0061, jl0062, jl0074, jl0075, jl0076, jl0087, jl0104
	Leisure and hobbies: Socializing	[2001-2020], [2006-2020]	jl0068, jl0069, jl0070, jl0071
	Leisure and hobbies: Sports	[2001-2020]	jl0063, jl0105_h, jl0105_v1, jl0105_v2, jl0109, jl0112_v1, jl0112_v2, jl0116, jl0117
	Leisure and hobbies: Videogames	[2001-2020], [2001-2015], [2016-2020]	jl0059_h, jl0059_v1, jl0059_v2
Household Questionnaire	Private Vehicles	[1998,2003,2015,2020]	hli0005, hli0077, hli0078, hli0079, hli0080, hli0081, hli0082, hli0083, hli0084, hli0085
	Traffic and energy: E-Bike	[2015,2020]	hli0135, hli0136
	Traffic and energy: Electricity provider change	[2015,2020]	hli0139, hli0140, hli0141, hli0142
	Traffic and energy: Green electricity	[2015,2020], [2015]	hli0137, hli0138
	Traffic and energy: Private Vehicles (Biodiesel)	[2015,2020]	hli0114, hli0115, hli0116, hli0117, hli0118, hli0119, hli0120
	Traffic and energy: Private Vehicles (Diesel)	[2015,2020]	hli0100, hli0101, hli0102, hli0103, hli0104, hli0105, hli0106
	Traffic and energy: Private Vehicles (E10)	[2015,2020]	hli0093, hli0094, hli0095, hli0096, hli0097, hli0098, hli0099

continues on next page

Table 9 – continued from previous page

Questionnaire	Module	Years	Variables
	Traffic and energy: Private Vehicles (Electricity)	[2015,2020]	hli0121, hli0122, hli0123, hli0124, hli0125, hli0126, hli0127
	Traffic and energy: Private Vehicles (Gas)	[2015,2020]	hli0107, hli0108, hli0109, hli0110, hli0111, hli0112, hli0113
	Traffic and energy: Private Vehicles (Hydrogen)	[2015,2020]	hli0128, hli0129, hli0130, hli0131, hli0132, hli0133, hli0134
	Traffic and energy: Private Vehicles (Normal Petrol/Super)	[2015,2020]	hli0086, hli0087, hli0088, hli0089, hli0090, hli0091, hli0092

2.10 Integration, Migration, Transnationalization

Migration and forced migration are changing German society. The SOEP offers diverse migration samples and numerous specific migration questions that allow researchers to study migration-related questions in detail. The modules on integration, migration, and transnationalization provide data on migration histories, discrimination, inter-ethnic contact, education, cultural integration, transnational relations, identification with Germany, and the intention to stay in Germany.

Questionnaire	Module	Years	Variables
Individual Questionnaire	Applying for German citizenship	(unregelmaessig) [1998-2018]	plj0021
	Circle of friends, percentage of migrants	[2013,2018]	plm0143
	Contacts abroad, thoughts about moving abroad	[2009,2014,2019]	plj0089, plj0090, plj0091, plj0092, plj0104, plj0105
	Country of origin	[2020]	plj0725

continues on next page

Table 10 – continued from previous page

Questionnaire	Module	Years	Variables
	Disadvantage / discrimination based on ethnic origins (detailed)	[2019]	plh0387i01, plh0387i02, plh0387i04, plh0387i05, plh0387i06, plh0387i07, plh0387i08, plh0387i09, plh0387i10, plh0387i11, plj0048_v1, plj0048_v2, plj0327, plj0328, plj0329, plj0330, plj0331, plj0332, plj0333, plj0334, plj0335, plj0336, plj0337, plj0338, plj0339
	Foreign language skills	[2013]	plm0135
	Integration indicators	(unregelmaessig) [1984-2018], (unregelmaessig) [1997-2019], [2020]	plj0078, plj0080_v1, plj0080_v2
	Intention to stay	[1996-2011,2015-2020], [2013], [1996-2011,2013,2015-2020], [2020], [1984-2011,2013,2015-2020], [1996-2011,2013,2015-2020]	plj0085_v1, plj0085_v2, plj0086_v1, plj0086_v2, plj0087, plj0088
	Language ability German	[2007-2011,2013-2020], [2010-2011,2013-2020]	plj0071, plj0072, plj0073
	Language ability native language	[2007-2011,2013-2019], [2010-2011,2013-2019]	plj0074, plj0075, plj0076
	Language use, media	[2014,2016], [2017-2020]	plj0226_v1, plj0226_v2
	Language use, newspapers	(unregelmaessig) [1988-2012]	plj0070
	Native language	[2007-2011,2013,2015-2019]	plj0009
	Native language (family)	[2013,2015-2020], [2013], [2015-2020]	plm0136_h, plm0136_v1, plm0136_v2
	Native language (friends)	[2013,2015-2020], [2013], [2015-2020]	plm0137_h, plm0137_v1, plm0137_v2
	Native language (workplace)	[2013,2015-2020], [2013], [2015-2020]	plm0138_h, plm0138_v1, plm0138_v2
	Regional attachment	[2009,2014,2019]	plj0043, plj0044, plj0045
	Sense of home	(unregelmaessig) [1988-2012], [2014]	plj0083, plj0340
	Translation help	[2013-2018]	p_buh1, p_buh10, p_buh2, p_buh3, p_buh4, p_buh5, p_buh6, p_buh7, p_buh8, p_buh9
	Visited country of origin in last 2 years	[2014,2016,2018,2020]	plj0322, plj0323
	Visiting / being visited by Germans and foreigners at home	(unregelmaessig) [2007-2019]	plj0060, plj0061, plj0062, plj0063
Youth Questionnaire	Language ability German	[2006-2018], [2010-2018]	jl0248, jl0442, jl0443, jl0444, jl1249

continues on next page

Table 10 – continued from previous page

Questionnaire	Module	Years	Variables
	Language ability native language	[2006-2013], [2010-2013]	jl0251, jl1251

2.11 Survey Methodology

Survey methodology modules offer diverse variables on imputation, weighting, SOEP-Core fieldwork, identifiers, interview methods, survey modes, and information about the respondent's exit from the survey.

Questionnaire	Module	Variables
Interviewer Questionnaire		
	Identificators	hhnr , intid , syear , wave
	Interview information	typint , lenghtinth , lenghtintp , lenghtintj
	Demography	gender , birth , marital , educ , modbula , modggk , ista1 , ibstam1 , ibstav1 , imusp , irel
	Interviewer history	startint , endint , experience , firstintm , firstintd , lastintm , lastintd
	Employment	iberuf , ioed , istell
	Interviewer activity	meancontacthh , responserate , amountinth , amountintp , amountintj , papi , capi , cawi , mail
	Patience	iged
	Health	iges
	Risk aversion	irisk
	Life satisfaction	izule
	Incentives	ibbarhon , ibbeval
	Optimism	ibopt
	Motivation & Fulfillment	igru01 - igru07 , ierf01 - ierf14
	Assessment of Participation	itebe01 - itebe13
	Interviewer Training	ibseval01 - ibseval04 , ibschul , ibschul02
	Big Five personality traits	iego01 - iego22
	Attitudes and social interaction	ibez01 - ibez05 , iverh01 - iverh06
	Political orientation	ipol1 - ipol4
	Worries	isor01 - isor14 , isor21 - isor22
	Working hours	ibwsist01 - ibwsist05 , ibwssol01 - ibwssol03
	Interviewer and other studies	ibsozer01 - ibsozer08 , ibsozerno , ibsozerso , ibef01 - ibef03

continues on next page

Table 11 – continued from previous page

Questionnaire	Module	Variables
	Foreign language skills	ispr01 - ispr10 , ibspre01 - ibspre10
	Flags (conflicts)	genderconfl , birthconfl , maritalconfl , educconfl , startintconfl , ista1confl

Important documents regarding this Topic are available here

Last change: May 12, 2022

CHAPTER

THREE

SURVEY DESIGN

3.1 SOEP Questionnaires

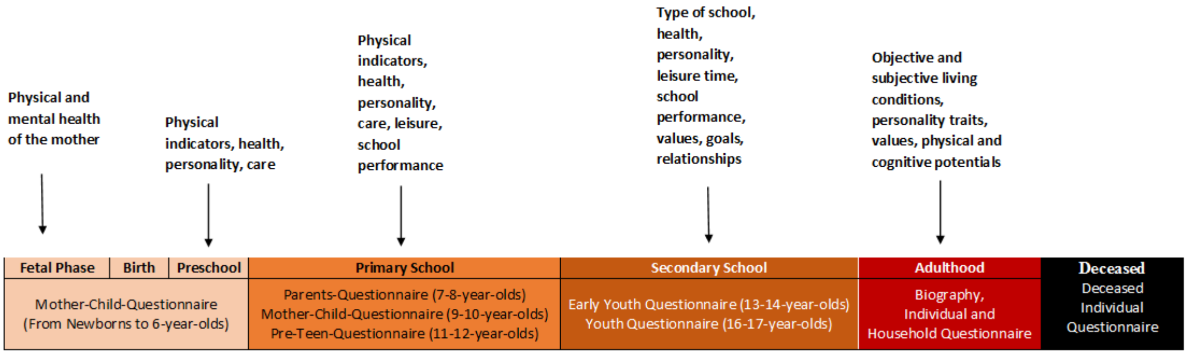
The interview methodology of the SOEP is based on a set of pre-tested questionnaires for households and individuals. Interviewers try to obtain face-to-face interviews with all members of a given survey household aged 16 and over. Thus, there are no proxy interviews for adult household members. Additionally, one person (the "head of household") is asked to answer a household-related questionnaire covering information on housing, housing costs, and different sources of income (e.g., social transfers such as social assistance or housing allowances). This questionnaire also includes questions on children up to the age of 16 in the household, mainly concerning daycare, kindergarten, and school attendance.

The questions in the SOEP are largely identical for all participants of the survey to ensure comparability across the participants within a given year, but of course there are differences across years. There are a few exceptions to this rule, which are due to different requirements in the target population. Up to 1996, the questionnaires for the sample of foreigners (B) and the immigrant sample (D) covered additional measures of integration or information on re-migration behavior. Between 1990 and 1992, i.e., during the first years of the German reunification process, the questionnaire for the East German sample (C) also contained some additional specific variables. From 1996 to 2012, all questionnaires were uniform and completely integrated for all of the main SOEP samples. For the IAB-SOEP Migration Sample, which was launched in 2013, specific questions were added to the SOEP questionnaires. The same is true of the IAB-BAMF-SOEP Survey of Refugees, which was launched in 2016.

Another special questionnaire is used for first-time respondents since some questions do not have to be repeated every year. Each respondent is asked to fill out a biographical questionnaire covering information on the life course up to the first SOEP interview (e.g., marital history, social background, and employment biography).

Additional information not provided directly by the respondent can be obtained from the "address logs", which are stored for every year in the $PBRUTTO and $HBRUTTO files. Every address log is filled in by the interviewer even in the case of non-response, thus providing very valuable information, e.g. for attrition analysis. For researchers interested in methodological issues, these data also contain information on the fieldwork process such as the number of contacts, reasons for drop-outs, and interview mode. For households that were contacted successfully, the address logs cover the size of the household, some regional information, survey status, etc. The individual data for all household members include the relationship to the household head, survey status of the individual, and some demographic information.

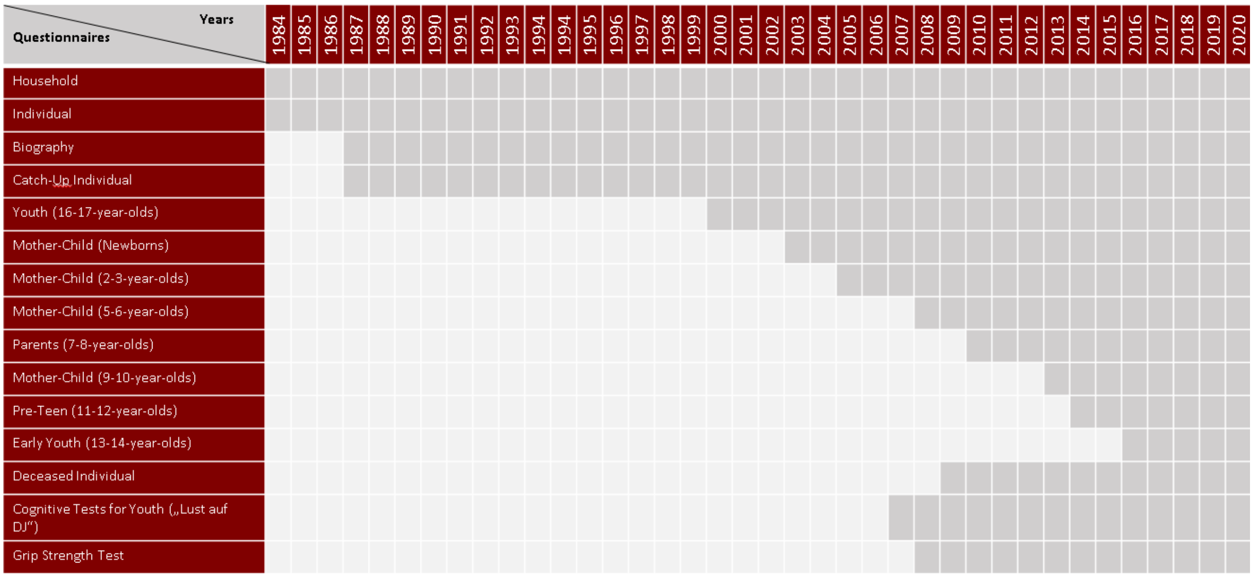
Life History


Physical and mental health of the mother
Physical indicators, health, personality, care
Physical indicators, health, personality, care, leisure, school performance
Type of school, health, personality, leisure time, school performance, values, goals, relationships
Objective and subjective living conditions, personality traits, values, physical and cognitive potentials

Fetal Phase	Birth	Preschool	Primary School	Secondary School	Adulthood	Deceased
Mother-Child-Questionnaire (From Newborns to 6-year-olds)			Parents-Questionnaire (7-8-year-olds) Mother-Child-Questionnaire (9-10-year-olds) Pre-Teen-Questionnaire (11-12-year-olds)	Early Youth Questionnaire (13-14-year-olds) Youth Questionnaire (16-17-year-olds)	Biography, Individual and Household Questionnaire	Deceased Individual Questionnaire

The SOEP questionnaires are designed so that people in a SOEP household can be analyzed from birth to adulthood and throughout the rest of their lives. In addition to the *Youth Questionnaire*, which was conducted for the first time in 2000/01, a series of questionnaires for specific cohorts of children living in SOEP households have been introduced since 2003. These have been completed annually since their year of introduction by mothers (in exceptional cases by fathers) with children of the appropriate age. In 2003, a questionnaire was developed for the mothers of newborn children, *Mother and Child Questionnaire (Newborns)*. The following instruments were developed in such a way that this starting cohort (born 2002/2003) can be followed up in their development and analyzed longitudinally. This was followed in 2005 by a questionnaire for mothers of 2-3-year-old children, *Mother and Child Questionnaire (2-3-year-olds)* and in 2008 by a questionnaire for 5-6-year-olds, *Mother and Child Questionnaire (5-6-year-olds)*. In 2010, the questionnaire for 7-8-year-old children, *Parents and Child Questionnaire (7-8-year-olds)*, completed by both mothers and fathers, was launched. In 2012, the questionnaire for 9-10-year-old children, *Mother and Child Questionnaire (9-10-year-olds)* was added as the last questionnaire to be answered by the mothers. This was followed by two youth instruments in which the children, aged 12, *Pre-Teen Questionnaire* and 14, *Early Youth Questionnaire*, answered questions about their own lives for the first time. These were introduced in 2014 and 2016, respectively. In 2018, the first cohort completed the entire battery of age-specific instruments and from then on, they will complete the annual questionnaires of the long-term SOEP study. Each person in a SOEP household receives the *Individual Questionnaire* as soon as they reach the age of 18, and the head of the household also receives the *Household Questionnaire*. If a respondent states in their interview that someone has died in the last year, regardless of whether the deceased person was part of a SOEP household, the *Deceased Individual Questionnaire* is given to the respondent providing the information.

3.1.1 Overview of the Questionnaires



Years Questionnaires	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Household	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Individual	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Biography				■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Catch-Up Individual				■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Youth (16-17-year-olds)																		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Mother-Child (Newborns)																					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Mother-Child (2-3-year-olds)																							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Mother-Child (5-6-year-olds)																										■	■	■	■	■	■	■	■	■	■	■	■	■
Parents (7-8-year-olds)																												■	■	■	■	■	■	■	■	■	■	■
Mother-Child (9-10-year-olds)																															■	■	■	■	■	■	■	■
Pre-Teen (11-12-year-olds)																																■	■	■	■	■	■	■
Early Youth (13-14-year-olds)																																		■	■	■	■	■
Deceased Individual																											■	■	■	■	■	■	■	■	■	■	■	■
Cognitive Tests for Youth („Lust auf DJ“)																									■	■	■	■	■	■	■	■	■	■	■	■	■	■
Grip Strength Test																										■	■	■	■	■	■	■	■	■	■	■	■	■

3.1.2 Household Questionnaire

The household questionnaire in its basic form has been an important part of the SOEP surveys since 1984 and has been improved and expanded continuously. The data collected and the questionnaire itself have become so complex that the original topics are no longer sufficient. Between 1984 and 2016, the number of questions more than doubled from 46 to 97. The multitude of questions offer users many options for analysis. Each year, the number of questions varies because new innovative question modules are added or because some questions are not asked every year. An overview of the modules included at different intervals can be found in the section *Topics of SOEP-Core*. The questions provide diverse information about the respondents' households that is stored in several hundred variables. Child-specific questions asked in the household questionnaire are found in the separate dataset $kind.

Availability: Since 1984

Dataset: $h (CS), *hl* (long)

Respondent: Head of household

The following question modules are part of the core program of the Household Questionnaire:

- Change of living situation
- Neighborhood
- Building type
- Size and condition of dwelling
- Amenities
- Type of dwelling
- Loans, mortgages, building-society loans
- Hereditary lease interest
- Modernization costs
- Ownership costs

- Photovoltaic and solar thermal system
- Owner debt
- Government-subsidized housing
- Home ownership
- Rental and expenses
- Tenant debt
- Cleaning or household assistance
- Persons in need of care
- Names and birth dates of children
- Child's school attendance
- Childcare situation
- Income and expenses from renting/leasing
- Loan repayment
- Debt
- Inheritances, gifts, winnings
- Investments
- Income/expenses household
- Savings
- Material deprevation
- Number of books
- Pets
- Cause of moving

where applicable:

+ migration-specific modules for the IAB-SOEP Migration Sample

- distinguishing repayment of loans, debt, income / expenses between Germany and foreign country

or where applicable:

+ refugee-specific modules for the IAB-BAMF-SOEP Sample of Refugees

- Information on shared accomodations
- Location preferences

3.1.3 Individual Questionnaire

The individual questionnaire has been a standard instrument since the beginning of the SOEP. In order to enable analysis over time, the individual questionnaire has a large number of question modules that are asked every year. There are also questions that do not have to be asked every year, as short-term changes are unlikely. In order to be able to react to current social changes, new topics are added to the individual questionnaire and repeated at intervals of more than one year.

Availability: Since 1984

Dataset: $p (CS), *pl* (long)

Respondent: Persons over 18 years in the household

The following question modules are part of the core program of the Individual Questionnaire:

- Satisfaction with various live aspects
- Satisfaction with current life situation
- Feelings
- Flourishing
- Risk aversion
- Political orientation
- Worrying
- Life satisfaction overall
- Ethnic/national origins
- Vocational training
- Completed level of education
- Higher education
- Family situation
- Family changes
- State of health
- Disability or severe disability
- Visits to the doctor
- Hospital stays
- Sick leave
- Health insurance
- Wages and collective wage agreements
- Additional questions for employees
- Additional questions for retirees/pensioners
- Government transfers
- Calendar
- Time use
- Second jobs

- Income
- Work, last 7 days
- Maternity/ parental leave
- Care period (Pflegezeit)
- Registered unemployed
- Quitting a job
- Employment status
- Start of job
- Change of job
- Job search
- Current profession
- Current job
- Working hours
- Overtime
- Optimism
- Religion
- Organization and Association membership
- Personality traits (Big Five)
- Anomie
- Life goals
- Locus of control
- Reciprocity
- Trust and Fairness
- Narcissism
- Lonelisness
- Impulsiveness and Patience
- Political Goals (Ingelhart-Index)
- Attitude towards refugees
- Just society
- Discriminatiom
- Bundestag election
- Social responsibility
- Influence on public decisions
- Friends
- LGBT-Status
- Child wish

- Gender stereotypes
- Attitudes towards gender
- On-Call occupation
- Commuting
- Home-Office
- Short-Time work payment
- Work council
- Payment equity
- Workload
- Occupational expectations
- Depressive traits
- Smoking and drinking
- Integration indicators
- Free time
- Leisure activities
- Donation

where applicable:

+ migration specific modules for the IAB-SOEP-Migrationsample

- First Job in Germany
- Job before immigration
- Language proficiency before and since immigration
- Partnership during immigration
- Living situation since immigration
- Religion and faith of parents
- Satisfaction in various areas of life before and after immigration

or where applicable:

+ refugee specific modules for the IAB-BAMF-SOEP-Sample of Refugees

- Legal status
- Religion and faith
- Language proficiency
- Integration courses and government measures
- Special questions for interviewers concerning language
- Recognition of qualifications

Re-Interviewed

- Cultural and political participation
- Application for recognition

- Trauma screener
- OK (Judgement of different actions)
- Citizenship (inkl. connection with country of origin/ Germany)
- Disadvantages
- Location preferences
- Willingness to participate in a tandem program
- Satisfaction in various areas before and after fleeing

New respondents

- Obtaining help and knowledge about advice services
- Assessment of current situation in country of origin
- Government, democracy and woman's position

3.1.4 Biography Questionnaire

Availability: Since 1987

Dataset: $lela (CS), *biol* (long)

Respondent: Supplementary, one-time data from the personal questionnaire of all persons aged 18 and over in the household.

Content:

- Nationality
- Country of Origin
- Childhood
- Parents
- Life course since the age of 15
- Education
- Occupation
- Partnership/marriage
- Information on children
- Siblings

where applicable:

+ migration specific modules for the IAB-SOEP-Migrationsample

- Travel to Germany
- Stays Abroad
- Citizenship
- Language proficiency
- Work before moving to Germany
- First job in Germany

- Relationship at the time of moving to Germany

or where applicable:

+ refugee-specific modules for the IAB-BAMF-SOEP Sample of Refugees

- Travel to Germany
- Questions concerning parents of respondent
- Lodging and living situation
- Language proficiency before moving to Germany

3.1.5 Mother and Child Instruments

Mother and Child Questionnaire (Newborns)

Mothers of newborn children answer questions dealing primarily with pregnancy, birth, breastfeeding, and the health of the newborn child. The questionnaire also asks to what extent the mother feels that her living situation changed after the birth of the child, how childcare is handled, and how mothers assess their baby's temperament (as a precursor to personality).

Availability: Since 2003

Dataset: $muki (CS), *bioagel* (long)

Respondent: Mother in household (child age 0-1)

Content:

- Course of pregnancy
- Childbirth
- Health screening
- Well-being
- Childcare
- Living situation

Mother and Child Questionnaire (2-3-year-olds)

Mothers of 2-3-year-old children answer questions about their child's health and how long they have been breastfeeding. The questionnaire asks again about the childcare situation and the child's temperament and includes a short scale on personality (the dimensions of agreeableness, extraversion, openness, and conscientiousness from the "Big Five"; McCrae and Costa 1987). In addition, it asks what language is spoken with the child and what activities they or the main caregiver engages in with their child (e.g., going to the playground, reading or telling stories, visiting other families with children). Mothers are asked to assess their children's adaptive behavior in the areas of communication, everyday skills, social relationships, and motor skills. This is based on a translated version of the Vineland Adpative Behavior Scale, which was reduced to 20 items for the SOEP to provide data on the child's stage of development in everyday life.

Availability: Since 2005

Dataset: $muki2 (CS), *bioagel* (long)

Respondent: Mother in household (child age 2-3)

Content:

- Personality of the child
- Well-being
- Childcare
- Language skills
- Development
- Abilities

Mother and Child Questionnaire (5-6-year-olds)

Mothers of 5-6-year-old children complete this questionnaire in the survey year when their child will turn six. It has many of the same topics as in previous years: health, childcare, a more comprehensive battery of items on the personality (from this age on, the “Big Five” dimension of neuroticism is also included) and activities that they or the main caregiver engages in with their child. In addition, the questionnaire includes a shortened version of the Strength and Difficulties Questionnaire (SDQ), a frequently used instrument to measure the mental health of children and adolescents, reduced here to 17 items of the German SDQ.

Availability: Since 2008

Dataset: $muki3 (CS), *bioagel* (long)

Respondent: Mother in household (child age 5-6)

Content:

- Personality of the child
- Activities with children
- Well-being
- Childcare

Parents and Child Questionnaire (7-8-year-olds)

This questionnaire on 7-8-year-old children is the only age-specific instrument that is completed by both parents, as long as they live together in the same household. In this age range, questions about school attendance (date of school enrolment) and parent’s aspirations for their children’s level of school completion become relevant for the first time. However, the focus is on parenting goals, parenting styles, and the roles of both parents. Parenting goals range between conformity and autonomy. Parenting styles are surveyed using 18 items, which can be divided into six scales: emotional warmth, inconsistent education, monitoring, negative communication, psychological control, strict control. The items were taken from the pairfam study, as were the 10 items on the role of parents, which can be divided into three scales: autonomy, hostile attributes, and willingness to make sacrifices.

Availability: Since 2012

Dataset: $elt (CS), *bioagel* (long)

Respondent: Parents in household (child age 7-8)

Content:

- Hopes and expectations for children’s educational attainment
- Parental goals
- Parental styles
- Parental role

- Childcare

Mother and Child Questionnaire (9-10-year-olds)

In addition to questions on health and child care, which are asked in almost all age groups, mothers of 9-10-year-old children are asked for more detailed information about the children's school situation. They are asked what level of schooling they would like their children to complete and what level they think is realistic, what their children's most recent grades were in their three main subjects, whether someone helps the child with homework, and whether the child likes going to school. Since friends and leisure activities are gaining in importance in this age group, some questions deal with these topics. Questions about allowance money are asked for the first time in this age group.

Availability: Since 2012

Dataset: $muki5 (CS), *bioagel* (long)

Respondent: Mother in household (child age 9-10)

Content:

- Hopes and expectations for children's educational attainment
- Education
- Parental involvement
- Leisure activities
- Family environment
- Social behavior of child
- Personality of Child
- Health of Child
- Supervision
- Allowance money

3.1.6 Youth Instruments

Pre-Teen Questionnaire

Young people complete a questionnaire for the first time themselves in the year they turn twelve. Here, as in the preceding questionnaires, the focus is on their school situation: what time their school day starts and ends on different days of the week, what type of school they attend, how many students are in their class, how many of their classmates are not from Germany, whether they feel discriminated against by their teacher, and what their grades were on their last report card in Math, German, and English. The questionnaire also asks how much time they spend on homework, where they do their homework, and who helps them with homework and studying. They are asked what level of schooling they would like to complete and what level they realistically expect to complete. Since friends play an important role at this age, pre-teens are asked how often they go to friends for support when they have problems. They are asked how many close friendships they have and how often their parents interfere in their choice of friends. They are asked about the educational aspirations of their three closest friends and three older siblings (if any). Several questions deal with their cultural capital and learning environment (e.g., books, musical instruments, and art in the household; whether they have a desk and a room of their own). They are asked about how they spend their free time, how much allowance money they get, and about their personality, willingness to take risks, and life satisfaction. Further questions deal with what languages are spoken with the child and who the child eats meals with.

Availability: Since 2014

Dataset: $school (CS), *biopupil* (long)

Respondent: 11-12-year-olds in the household

Content:

- Attitude
- Personality
- School (schedule, educational attainment, extra-curricular activities)
- Recreational activities
- Social and family surroundings
- Living situation

Early Youth Questionnaire

The questionnaire for early youth is designed similarly to the pre-teen questionnaire to provide important data on developmental psychology. There are fewer questions about homework and the learning environment and more questions on involvement in extra-curricular activities at school (e.g., student council, after-school clubs) since such activities build social capital. Early youth are asked about the importance of various family members and friends in their lives and about their own educational aspirations as well as those of their three best friends. They are asked how late they are allowed to stay out on school nights and weekends, and what types of activities they have taken part in without their parents (e.g., vacation, doctor visits, shopping, drinking alcohol, smoking cigarettes). They are asked how much allowance they get, and whether they have any savings. Another new topic in this age group is interest in politics and political orientations.

Availability: Since 2015

Dataset: $school2 (CS), *biopupil* (long)

Respondent: 13-14-year-olds in the household

Content:

- Self-perception
- School (schedule, educational attainment, extra-curricular activities)
- Recreational activities
- Friends
- Siblings
- Parents
- Allowance money
- Political party preferences
- Self-perceptions
- Willingness to take risks
- Life satisfaction
- Attitudes/opinions
- Future

Youth Questionnaire

In the SOEP, young people who turn 17 in the year of the survey are considered adult respondents. Like other first-time adult respondents, they receive a biography questionnaire and an individual questionnaire. Since part of the adult biography (e.g., employment history, relationships) does not yet apply to the young respondents, whereas other aspects such as relationships with parents, leisure activities, and school or vocational training play a greater role, a youth questionnaire was developed in 2000 to replace the biographical questionnaire in this age group. The content of this questionnaire corresponds in many respects to the adult biographical questionnaire so that the data can be used to supplement the information on parents (if parents do not live in the household; dataset: BIOPAREN). Health status, personality, willingness to take risks, locus of control, trust, time preferences, political preferences, knowledge of German, as well as information on the respondent's living situation, work situation, training, career plans, and educational aspirations are also covered in this questionnaire. For the period from 2000 to 2005, respondents in this age group completed the youth questionnaire and the individual questionnaire. Since 2006, they have only completed the youth questionnaire. The version used since then has been expanded to include a few additional indicators. A test was added to assess cognitive potential based on the I-S-T 2000R (Amthauer et al. 2001) using 20 subtasks each for the components of analogies, number series, and matrices (see Solga et al. 2005). The test measures fluid cognitive abilities, a strongly biologically determined dimension of cognitive abilities that is not influenced by education and is primarily based on reasoning, processing rate, and working memory capacity (Cattell 1971; Horn 1982). Although the format of the test differs from those usually used in surveys, young people's willingness to participate has been high (Schupp and Hermann 2009).

Availability: Since 2000

Dataset: $jugend (CS), *jugendl* (long)

Respondent: 16-17-year-olds in the household

Content:

- Living
- Relationships
- Leisure and sports
- School (educational attainment, foreign languages, extra-curricular activities)
- Allowance money
- Education
- Career plans
- Future
- Background
- Childhood and Upbringing
- Attitudes/opinions
- Self-Perception
- Life satisfaction
- Political party preferences

Cognitive Tests for Youth

In 2006, a separate questionnaire with cognitive tests for adolescents was used for the first time in the SOEP. It was named “Lust auf DJ” (or “interest in DJ”) as a play on disc jockey, but DJ stands for “Denksport und Jugend”, or mind sports and youth. The questionnaire was created for young people between the ages of 16 and 17.

Availability: Since 2007

Dataset: *cogdj* (CS)

Respondent: 16-17-year-olds in the household as a supplement to the youth questionnaire

Content:

- Assignment of word pairs
- Complete equations
- Assign figures

3.1.7 Additional Instruments

Catch-Up Individual Questionnaire

The Catch-Up or “Gap” (German:Lücke) questionnaire is given to respondents who failed to respond in the previous year of the study. They are asked to provide important data about the year they missed.

Availability: Since 1987

Dataset: pluecke (CS), *plueckel* (long)

Respondent: SOEP respondents who are temporarily unavailable.

Content:

All data refer to the previous survey year

- Status of the respondent
- Occupational change
- Receipt of social benefits within the last year
- Completion of education
- Type of educational attainment
- Change of family status

Deceased Individual Questionnaire

In 2009, for the first time in SOEP-Core, information was collected on former SOEP participants who had died since the last survey in 2008. The Deceased Individual questionnaire thus completes the life history information in the SOEP. The primary aim is to obtain as much information as possible about the causes and circumstances of death of former SOEP respondents. As the questionnaire also collects information on individuals who have never participated in the SOEP survey, this can be used together with the causes and circumstances of death in socio-scientific analysis.

Availability: Since 2009

Dataset: vp (CS), *vpl* (long)

Respondent: SOEP respondents who lost a loved one.

Content:

- Relationship to the deceased
- Was the deceased a survey respondent?
- Domestic environment of the deceased
- Cause and place of death
- Last will and testament
- Health of the deceased
- Life satisfaction of the deceased
- Influence of bereavement on respondent's own life

Grip Strength Test

Availability: Since 2008

Dataset: *gripstr* (long)

Respondent: Persons over 17 years in the household

Content:

This test measures hand grip strength, which is useful in assessing respondents' physical condition.

Interviewer Questionnaire

We derive basic demographical and employment information on interviewers from personnel data of the fieldwork organization. Since 2000, Kantar Public regularly updates these information. Additionally, at irregular intervals, the SOEP interviewers complete a short version of the standard individual questionnaire themselves, which is called the interviewer questionnaire.

Availability: 2006, 2012, 2016

Dataset: *interviewer* (long)

Respondent: SOEP interviewers

Content:

- Basic Demography
- Occupational History
- Personality
- Motivation
- Interviewer Training
- Worries
- Language Skills

Last change: May 12, 2022

3.2 Survey Concepts and Modes

Measuring stability and detecting changes means repeating (almost) identical measures over time. Furthermore, the SOEP questions capture stability and change by varying with regard to the time dimension, that is, asking about events in the past, the present, and the future. Conceptually, different measurements of time are used:

- Questions about a point in time (present), e.g., current employment status or current levels of satisfaction
- Retrospective questions about certain events in the past, e.g., how often have you changed jobs in the last ten years?
- Retrospective life event history since the age of 15 (in the past), e.g., employment or marital history
- Monthly calendar information on income and labor market participation (in the past), e.g., employment status January through December of last year
- Questions about a period of time (in the past), e.g., demographic changes since the last interview such as marriage or death of spouse
- Questions about the future, e.g., expected satisfaction with life five years from now, or job expectations

Survey Modes

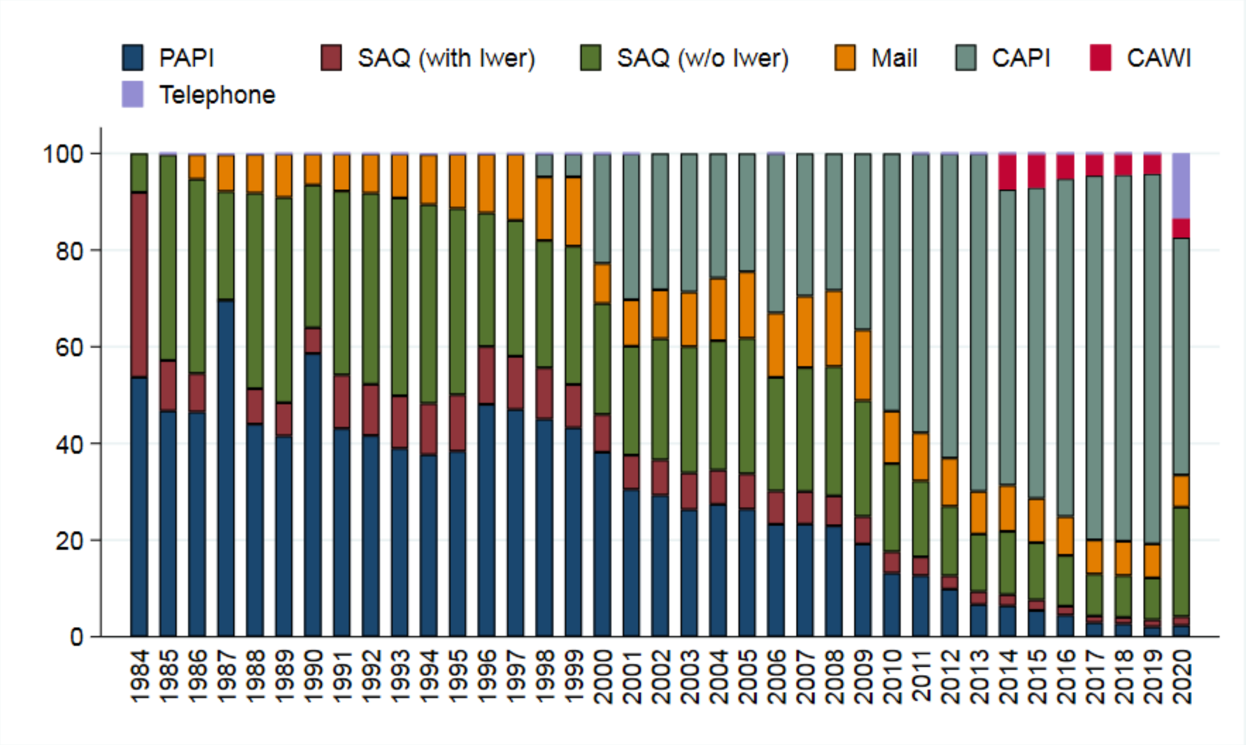
The SOEP uses several different modes to collect the data. Originally, the respondent's answers were always recorded by an interviewer who filled in the answers in a paper questionnaire, the "pen-and-paper interview" or PAPI. The personal contact between interviewer and respondent is important for the success of the survey; however, before losing a respondent due to a scheduling conflict between interviewer and respondent, the SOEP has allowed respondents to mail in the questionnaire since the second wave of subsamples A-I. This is not the same as the concept of a regular mail survey, because the interviewer still maintains personal contact with the household and schedules appointments with respondents if possible. Starting with subsample J, only "computer-assisted personal interviews" (CAPI) are allowed, and thus it is no longer possible to mail in the questionnaires.

When visiting a household, the interviewer interviews household members one at a time and can also give questionnaires to other household members to complete without the interviewer's assistance (self-administered questionnaires, SAQ). This is a time-efficient approach because it allows different household members to complete their questionnaires at the same time.

In 1998, computers were used for the first time in the SOEP for computer-assisted personal interviews (CAPI). Compared to PAPI, the CAPI mode is much more efficient in converting the data into an electronic format, which was an important asset especially with the extensions to the panel starting in the year 2000. The CAPI mode was first used parallel with PAPI, meaning that interviewers and respondents were free to chose how they wanted to do the interview. This was important for the "older" sample members (respondents as well as interviewers), who were used to the PAPI concept. Only in the most recent samples (starting in subsample J) is CAPI the sole interview mode. The figure depicts the development of modes up to 2011, showing that the CAPI mode has gained importance since its implementation.

Since the questionnaires have to be identical in both modes, CAPI is implemented in a relatively simple way in the SOEP and does not utilize all the technical possibilities of this interview mode. For example, the SOEP basically does not use any form of dependent interviewing (i.e., referring to respondent data from previous waves), because this cannot be easily implemented in the PAPI mode. Also, the filtering structure is very simple in the SOEP, because a respondent must be able to follow the interview path on paperon her/his own. Still, some technical features like the control of value ranges (e.g. month of birth, year of first marriage) or the randomization of scale items are implemented in the CAPI version of the questionnaire.

In the future, new modes will be introduced into the SOEP as they develop. The computer-assisted web interview (CAWI) is close to implementation, but will not be used as a replacement of the current CAPI and PAPI modes, but rather as an extension the respondents may use, similar to the mail-in or self-administered questionnaires. The core interview concept of the SOEP survey, the personal contact between respondent and interviewer, will not change.



`Download STATA Code to create figure.`

Last change: May 12, 2022

3.3 Panel Care

To cope with panel attrition and to keep longitudinal response rates high, the SOEP has implemented "panel care" efforts to maintain personal contact between respondents and the survey. Panel care can be divided into incentives given directly to the respondent and other measures undertaken to keep the respondent in the study.

Respondents have been given gifts as tokens of appreciation since the very beginning of the study. Most of these gifts are small in-kind incentives like flowers, for which the interviewers have their own budget. In addition, the interviewers are asked to hand out a brochure with recent results from the study. Up to 2007, respondents also received a lottery ticket as a thank-you upon completion of their interview. Proceeds from the lottery benefit social projects in Germany. Since 2008, the lottery ticket has been included with the contact letter that is sent out about two weeks prior to the interview. It is thus given unconditionally, as long as the person participated in the previous wave. After a successful interview, the respondent receives a thank-you letter from survey institute along with one postage stamp as a small additional gift.

In 2009, different incentive schemes were tested in the new subsample I to increase the first-wave response rates. The basic experiment included four randomized groups of households: (1) those with the default setup of the conditional lottery ticket; (2) those with a "low" cash incentive of 5 euros per household and 5 euros per adult respondent; (3) those with a "high" cash incentive of 5 euros per household and 10 euros per adult respondent; and (4) those with a choice between a "low" cash incentive and a lottery ticket. The results showed slightly higher response rates in the cash groups, although the extra money in group (3) did not pay off. The current incentive strategy for the different SOEP samples is shown here:

Samples	A-H	J,K, L1, N, O, Q	L2, L3	P	M1/M2
Incentives for adults	Lottery ticket (5,427 households) and cash (636 households)	5 euros (households) and 10 euros (adults)	5 euros (households), 5 euros (adults), 10 euros (bonus payment)	Lottery Ticket	5 euros (households) and 10 euros (adults)
Incentives for other respondents	Power bank (*Youth Questionnaire*), Bicycle repair kit (*Early Youth Questionnaire*), Small clock (*Pupils Questionnaire* CAPI/PAPI), Puzzle (*Pupils Questionnaire* MAIL)	Power bank (*Youth Questionnaire*), Bicycle repair kit (*Early Youth Questionnaire*), Small clock (*Pupils Questionnaire* CAPI/PAPI), Puzzle (*Pupils Questionnaire* MAIL)	5 euros (*Youth Questionnaire*), (*Youth Questionnaire*, *Early Youth Questionnaire*, *Pupils Questionnaire*) and 5 euros (*Mother-Child Instruments*)		Power bank (*Youth Questionnaire*), USB-Stick (*Early Youth Questionnaire*), Bicycle repair kit (*Pupils Questionnaire*)

The survey institute also does additional work to keep response rates high. Addresses are checked throughout the year to ensure that current addresses are on file. This is done, for instance, by sending out brochures about recent research based on the SOEP data and seasonal greeting cards.

Face-to-face interviews also ensure a personal relationship between interviewer and respondent, which increases the likelihood that respondents will stay in the survey. Keeping the same interviewer over time is therefore an important goal of the survey. Some SOEP respondents have in fact had the same interviewer since the beginning in 1984.

Last change: Sep 26, 2022

CHAPTER

FOUR

TARGET POPULATION AND SAMPLES

The target population covered in the SOEP is defined as the population of private households residing within the current boundaries of the Federal Republic of Germany (FRG). Because of changes in these boundaries (in 1990) and changes in the population due to migration, various adaptations have been made to the initial sampling structure to maintain the sample's representativity. In addition, certain groups have been oversampled to increase the statistical power.

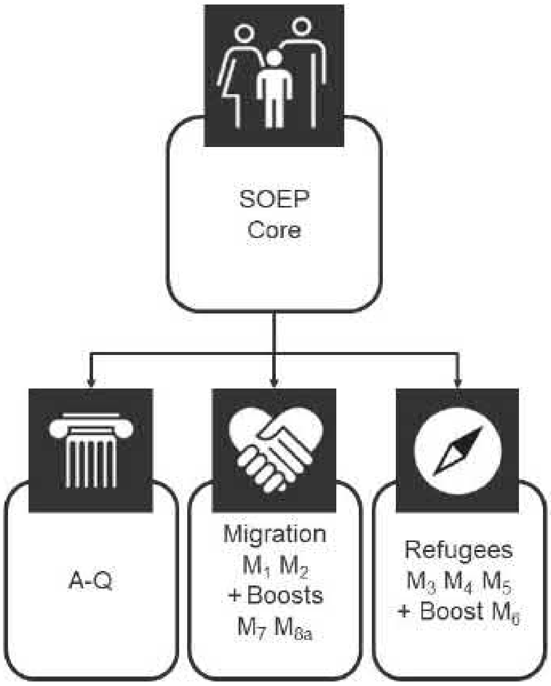
The different SOEP-Core subsamples constitute the centerpiece of the SOEP.

1. Within SOEP-Core, **samples A-Q** form the heart of the SOEP. They contain the oldest samples, beginning with the founding sample in A from 1984 and the highest number of participating households. Fieldwork traditionally starts at the beginning of February, and its questionnaires serve as a master for the other SOEP-Core subsamples.
2. The SOEP migration sample with it`s **samples M1 and M2** was established in 2013 and is designed to improve the representation of migrants living in Germany. Fieldwork started in April, using the questionnaires from samples A-Q, supplemented by translated questionnaires for five different languages.
3. In order to map recent migration and integration dynamics, SOEP refugee **samples M3 to M5** were installed beginning in the year 2016. In 2020, fieldwork began in August with a questionnaire that was tailored to issues of recent refugees while containing many questions from the SOEP samples A-Q as well.
4. **Sample M6** – a boost sample of refugees targeted the same population as the older refugee sample M5 - adult refugees who have applied for asylum in Germany since 1 January 2013 and are currently living in Germany – and the same sample design and sample frame were used.
5. The two boost samples, **samples M7 and M8a**, were added the SOEP migration sample system. Like the older migration samples M1 and M2, the Integrated Employment Biographies Sample (IEBS) of the Federal Employment Agency (BA) served as the sampling frame for both boost samples. Boost sample M7's goal was to capture migration dynamics and processes from 2016 to 2018 with a focus on EU migration. To ensure that statistically significant group comparisons can be made, sampling was restricted to the three most significant countries of origin in that time period: Romania, Bulgaria, and Poland. M8a, on the other hand, was designed to help evaluate the skilled worker immigration law (Fachkräfteeinwanderungsgesetz), which came into effect March 1, 2020, and targeted migrants from third countries that came to Germany between 2017 and 2018, sampling them as a control group for a treatment group that will be sampled at a later date.

In 1984, the survey started with a sample covering the entire population of then West Germany (FRG), where the five biggest groups of foreigners ("guest workers") were oversampled.

The SOEP was expanded to the territory of the German Democratic Republic in June 1990, only six months after the fall of the Berlin Wall. In 1994/95, a boost sample of migrants who came to Germany after 1984 was added to take the influx of ethnic Germans from former Soviet countries into account. In 2013 another sample of migrants which includes individuals who immigrated to Germany after 1995 or second-generation immigrants was added. Since then, multiple migration or refugees samples were added in cooperation with the IAB (Institut für Arbeitsmarkt- und Berufsforschung) or the BAMF (Bundesamt für Migration und Flüchtlinge)

Now and then samples that were representative of the entire population in Germany were added to counter effects of panel attrition and to increase the overall sample size.



The different samples in the SOEP are identified by letters: sample "A" refers to the German sample drawn in 1984, "C" to the East Germans from 1990, and so on. Even though these samples are kept separate, the respondents have received identical questionnaires for the most part, and distinctions by sample are usually not necessary in an analysis.

However, one of the ideas of the SOEP is that the users have full information available about survey methodological issues and survey design, which in this case means that you can identify the corresponding sample for each observation. In the following section, we present details on each of the samples, which unless stated otherwise are multi-stage random samples with regional clusters. The households are selected by random-walk routines.

For an extensive discussion on sampling (and weighting), see: Survey methods.

4.1 The SOEP Samples in Detail

Sample A "Residents of the Federal Republic of Germany" covers individuals in private households with a household head who does not belong to one of the main groups of "guest workers" (i.e., Turkish, Greek, Yugoslavian, Spanish, or Italian households). Because only a few foreigners are in Sample A, it is often called the "West German Sample" of the SOEP. In 1984 it covered 4,528 households with a sampling probability of about 0.0002.

Sample B "Foreigners in the Federal Republic of Germany" adds individuals in private households with a Turkish, Greek, Yugoslavian, Spanish, or Italian household head, who in 1984 constituted the main groups of foreigners in the FRG. Compared to Sample A, the population of Sample B is oversampled with a sampling probability of about 0.002. In the first wave, Sample B included 1,393 households.

Sample C "German Residents of the German Democratic Republic (GDR)" consists of individuals in private households in which the household head was a citizen of the German Democratic Republic (GDR). This meant that approximately 1.7% of the residential population of the GDR in June 1990 was excluded from the sample as foreigners (most of whom were living in "institutionalized" housing). In total, the sample started with 2,179 households with a sampling probability of about 0.0005.

Sample D "Immigrants" started in 1994/95 with two different samples. In 1994, the first sample, D1, had 236 households and in 1995, the second sample, D2, had 295 households, leading to a total of 531 households (D1 and D2) in

1995. This sample consisted of households in which at least one household member had moved from abroad to West Germany after 1984. The sampling probability is about 0.0002.

Sample E "Refresher" was added in 1998, selected from the entire population of private households in Germany. The households were chosen independently of the ongoing panel and its subsamples A through D. The aim was to increase the number of observations of the general population and to preserve its representativity. The selection scheme used for sample E essentially resembles the one used in subsample A. The number of households in the first wave of subsample E was $1,060$, with a sampling probability of about 0.00005. With the 2012 data release, parts of subsample E were extracted into the SOEP Innovation Sample. It is also the first sample in which Computer-Assisted Personal Interview (CAPI) was used. At that time, interviews in Samples A-D were being conducted entirely using Paper-and-Pencil-Interviews (PAPI). To study mode effects, households from sample E were randomly allocated to either CAPI or PAPI.

Sample F "Refresher" was selected independently of all other subsamples from the population of private households in 2000. The selection scheme was slightly altered compared to the previous addition in Sampl' E: while the "German" households (all adults aged 16 or older in the household have German nationality) were selected with a sampling probability of 0.00028, the 'non-German' households (at least one adult does not have German nationality) were oversampled with a probability of 0.0005. Overall, the number of added households in subsample F's first wave amounts to 6,043.

Sample G "High-Income" entered the SOEP in 2002 independently from all other subsamples. The original selection scheme required that the responding households had a monthly income of at least DM 7,500 (EUR 3,835), which - due to the lack of an adequate sampling frame - were identified using a screening procedure. This sample of a total of 1,224 households increased the potential for analysis in the high-income bracket, which was previously difficult to study because of the low case numbers. The derived sampling probability is about 0.0014. Starting with Wave 2 in 2003, the selection scheme for this subsample was changed such that only households with a net monthly income of at least EUR 4,500 were followed.

Sample H "Refresher" started in 2006 as a random sample, again independently of all previous subsamples, covering all residential households in Germany. The added 1,506 households were sampled with a probability of 0.0001.

Sample I "Incentive Sample" started in 2009, where in the first wave, a new incentive scheme was tested to increase participation rates (see also [sec:PanelCare]. The sampling was independent of all other SOEP samples, adding a total number of 1,531 households to the SOEP. The sampling probability was 0.00013. This sample remained in the main data release for its first two waves (2010 and 2011, or waves Z and BA). With the 2012 data release, subsample I was extracted into the SOEP Innovation Sample.

Sample J "Refresher Sample" started in 2011 as a random sample, independently of all previous subsamples, covering residential households in Germany. The added3,136 households were sampled with a probability of 0.0002.

Sample K "Refresher Sample" started in 2012 as a random sample, drawn independently of all previous subsamples, covering the residential households in Germany. The added 1,526 households were sampled with a probability of 0.0001.

Sample L1 "Cohort Sample" covers private households in Germany in which at least one household member was born between January 2007 and March 2010 and was therefore a child at that time. Again, migrants identified were oversampled using an onomastic procedure. Sample L1 (as well as L2 and L3) was part of the SOEP-related study "Families in Germany" (FiD), which was integrated into the SOEP in 2014. As part of an evaluation project by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) and the Federal Ministry of Finance (BMF), the study focused on public benefits in Germany for married people and families. Therefore, the survey instruments used in waves BA to BD differ in some respects from those used in the other samples.

Sample L2 "Family Types I" covers private households in Germany that meet at least one of the following criteria for household composition: single parents, low-income families, and large families with three or more children. Similar to Sample G, we face the problem that the eligible sub-population is relatively small and an adequate sampling frame is lacking. So again, a preceding telephone screening procedure identifies eligible households.

Sample L3 "Family Types II" covers private households in Germany that meet at least one of the following criteria for household composition: single parents or large families with three or more children. It is conducted analogously to Sample L2 to increase the number of cases in these sub-populations.

Sample M1 "Migration Sample" is a new migration sample added in 2013 with around 2,700 households drawn using register information from the German Federal Employment Agency. It includes individuals who immigrated to Germany after 1995 or second-generation immigrants.

Sample M2 "Migration Sample" was another migration sample added in 2015 with around 1,100 households drawn using register information from the German Federal Employment Agency. It includes individuals who immigrated to Germany between 2010 and 2013.

Sample M3 "Refugee Sample" was a new refugee sample added in 2016 for the IAB-BAMF-SOEP Refugee Survey in which roughly 1,769 refugee households were interviewed repeatedly. Respondents aged 18 and older who entered Germany between January 2013 and December 2016 and who had filed an asylum application by April 2016 (regardless of their current legal status) were interviewed along with the other members of their households.

Sample M4 "Refugee Family Sample": the 2016 "IAB-BAMF-SOEP Survey of Refugees" (Samples M3 and M4) is a joint project of the Institute for Employment Research (IAB), the Research Center of the Federal Office for Migration and Refugees (BAMF-FZ) and the Socio-Economic Panel (SOEP). The target population of the samples consists of 1,769 households with individuals who arrived in Germany between January 2013 and January 2016 and had applied for asylum by June 2016 or were hosted as part of specific programs of the federal states (irrespective of their asylum procedure and their current legal status). The first part of the sample (M3) was financed with funds allocated to the IAB from the research budget of the Federal Employment Agency (BA) . Sample M4 was funded by the Federal Ministry of Education and Research (BMBF) and has a focus on refugee families.

Sample M5 "Refugee Sample" M5 is the third boost sample of refugee households. The population of M5 covers adult refugees who applied for asylum in Germany between January 1, 2013, and December 31, 2016, and are currently living in Germany. The first wave of M5 was conducted in 2017. M5 added another 1,519 households of refugees who have migrated to Germany since 2013 to the SOEP framework.

Sample N "Refresher Sample (PIAAC-L)": Sample N integrated 2,314 households of former participants in the Program for the International Assessment of Adult Competencies (PIAAC and PIAAC-L) in 2017. This is the most recent addition to the SOEP-Core samples. Fieldwork in sample N was conducted between mid-March and mid-August and thus slightly later than the majority of samples A–L1.

Sample O "Social City Sample": Sample O includes 935 households located primarily in bigger cities. It was designed to enhance the potential of the data for analysis by incorporating more city-specific environments. The sample was selected in cooperation with BBSR using a new sampling design based on regional data in areas where the "Soziale Stadt" (social city) urban development project is being carried out. Based on the digital data available on the boundaries of the "Soziale Stadt" areas, it was possible to create a new variable going back to the year 2000 that shows whether or not a household's address is within an area covered by the project.

Sample P "Top Shareholder Sample": Sample P was conceptualized as a sample of highly affluent households in Germany. Against the backdrop of increasing income and wealth inequality in Germany, despite economic growth in recent decades, a lack of data on wealthy populations has become increasingly evident in the social sciences. Goals to be accomplished with sample P were to improve the empirical basis of the poverty and wealth report of the German government as well as laying the foundation for medium and long-term cross-sectional and longitudinal analyses. The gross sample of sample P consisted of 23,259 households.

Sample Q "LGB*": Sample Q is a boost sample of a hard-to-survey population: lesbians, gays, bisexuals, transgender people, and those who identify as non-binary. While the actual percentage of LGBTQ+ people in the general population is unknown, this population was too scarcely represented in the SOEP to meaningfully analyze this group. 835 households were recruited via an approximately 9-month long telephone screening process. Of these households 477 participated between April and November.

Sample M6 "Refugee Sample": M6 is the acronym for the fourth top-up sample for households that represents refugees. The population of M6 covers two groups: firstly, adult refugees who arrived in Germany between January 1, 2013 and December 2016 ("Refreshment") and secondly adult refugees who came to Germany between January 1, 2017 and June 2019 ("Enlargement") with a strongly disproportionate oversampling of refugees from East- and West-Africa.

Sample M7 "Migration Sample": Like the older migration samples M1 and M2, the Integrated Employment Biographies Sample (IEBS) of the Federal Employment Agency (BA) served as the sampling frame for both boost samples.

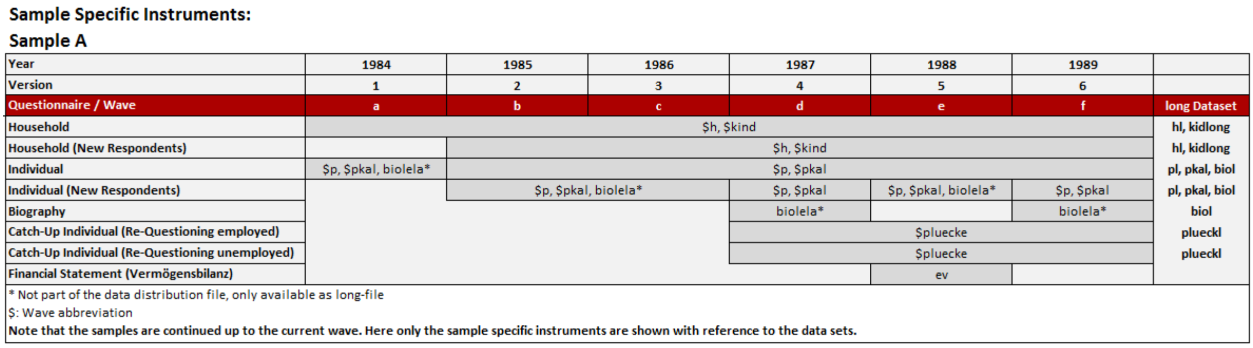
Boost sample M7's goal was to capture migration dynamics and processes from 2016 to 2018 with a focus on EU migration. To ensure that statistically significant group comparisons can be made, sampling was restricted to the three most significant countries of origin in that time period: Romania, Bulgaria, and Poland.

Sample M8a "Migration Sample": Like the older migration samples M1 and M2, the Integrated Employment Biographies Sample (IEBS) of the Federal Employment Agency (BA) served as the sampling frame for both boost samples. Boost sample M8a was designed to help evaluate the skilled worker immigration law (Fachkräfteeinwanderungsgesetz), which came into effect March 1, 2020, and targeted migrants from third countries that came to Germany between 2017 and 2018, sampling them as a control group for a treatment group that will be sampled at a later date.

More information about "Sample Sizes and Panel Attrition" can be found here

4.1.1 Sample-Specific Questionnaires

In SOEP it is common for special samples to receive extended, adapted, and/or integrated questionnaires in the first few years. This ensures that sample-specific questions that do not play a role in the main SOEP can also be included. In the following tables you can see which questionnaires the respective samples received, which years they ran, which raw data set they were included in, and which "long" data set they went into.

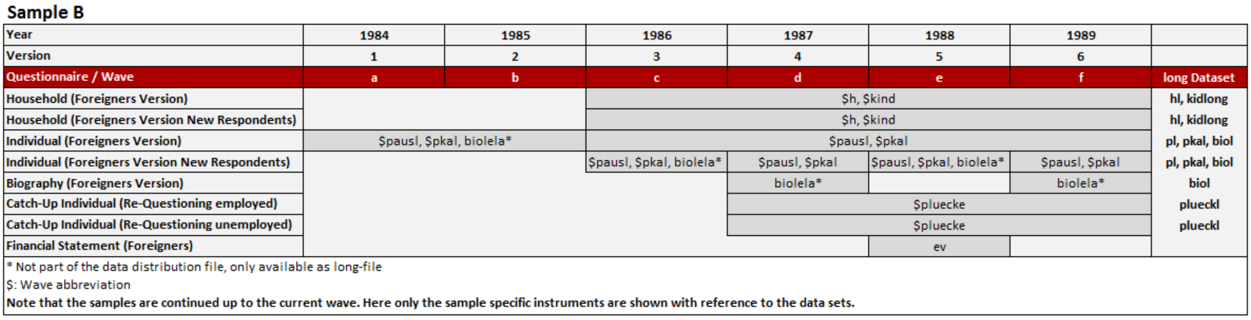


Sample Specific Instruments:

Sample A

Year	1984	1985	1986	1987	1988	1989	
Version	1	2	3	4	5	6	
Questionnaire / Wave	a	b	c	d	e	f	long Dataset
Household	$h, $kind						hl, kidlong
Household (New Respondents)		$h, $kind					hl, kidlong
Individual	$p, $pkal, biolela*	$p, $pkal					pl, pkal, biol
Individual (New Respondents)		$p, $pkal, biolela*		$p, $pkal	$p, $pkal, biolela*	$p, $pkal	pl, pkal, biol
Biography				biolela*		biolela*	biol
Catch-Up Individual (Re-Questioning employed)				$pluecke			plueckl
Catch-Up Individual (Re-Questioning unemployed)				$pluecke			plueckl
Financial Statement (Vermögensbilanz)					ev		

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

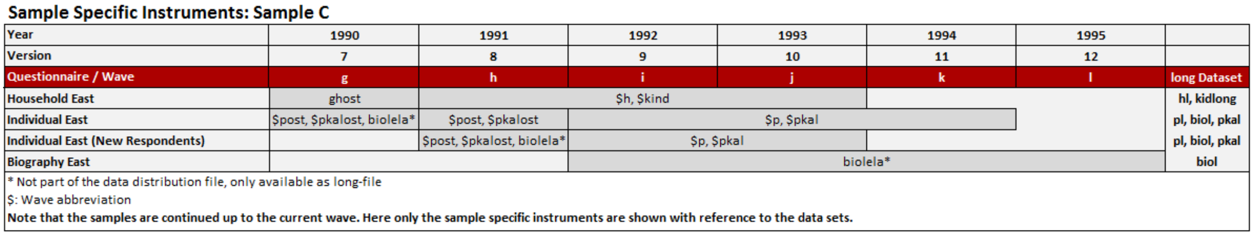


Sample B

Year	1984	1985	1986	1987	1988	1989	
Version	1	2	3	4	5	6	
Questionnaire / Wave	a	b	c	d	e	f	long Dataset
Household (Foreigners Version)			$h, $kind				hl, kidlong
Household (Foreigners Version New Respondents)			$h, $kind				hl, kidlong
Individual (Foreigners Version)	$pausl, $pkal, biolela*		$pausl, $pkal				pl, pkal, biol
Individual (Foreigners Version New Respondents)			$pausl, $pkal, biolela*	$pausl, $pkal	$pausl, $pkal, biolela*	$pausl, $pkal	pl, pkal, biol
Biography (Foreigners Version)				biolela*		biolela*	biol
Catch-Up Individual (Re-Questioning employed)				$pluecke			plueckl
Catch-Up Individual (Re-Questioning unemployed)				$pluecke			plueckl
Financial Statement (Foreigners)					ev		

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

From the start of Sample B (foreigners), respondents could complete the individual questionnaire in German or in the respective foreign language. Starting with wave 2 of the panel, there were "old" and "new" survey units (households, persons), and there were survey units with or without certain changes (e.g., households that had or had not moved; individuals who had or had not changed careers). The questionnaires took these changes into account for all subgroups. Survey procedures and tools were designed to ensure that each subgroup received the right questionnaire for them. This technique as well as the bilingual design of the foreigner questionnaires was retained for waves 3-6. In addition, retrospective information and missing information on temporary drop outs was collected. The "financial statement", which is now a survey module, was a separate questionnaire in the year 1988.



Sample Specific Instruments: Sample C

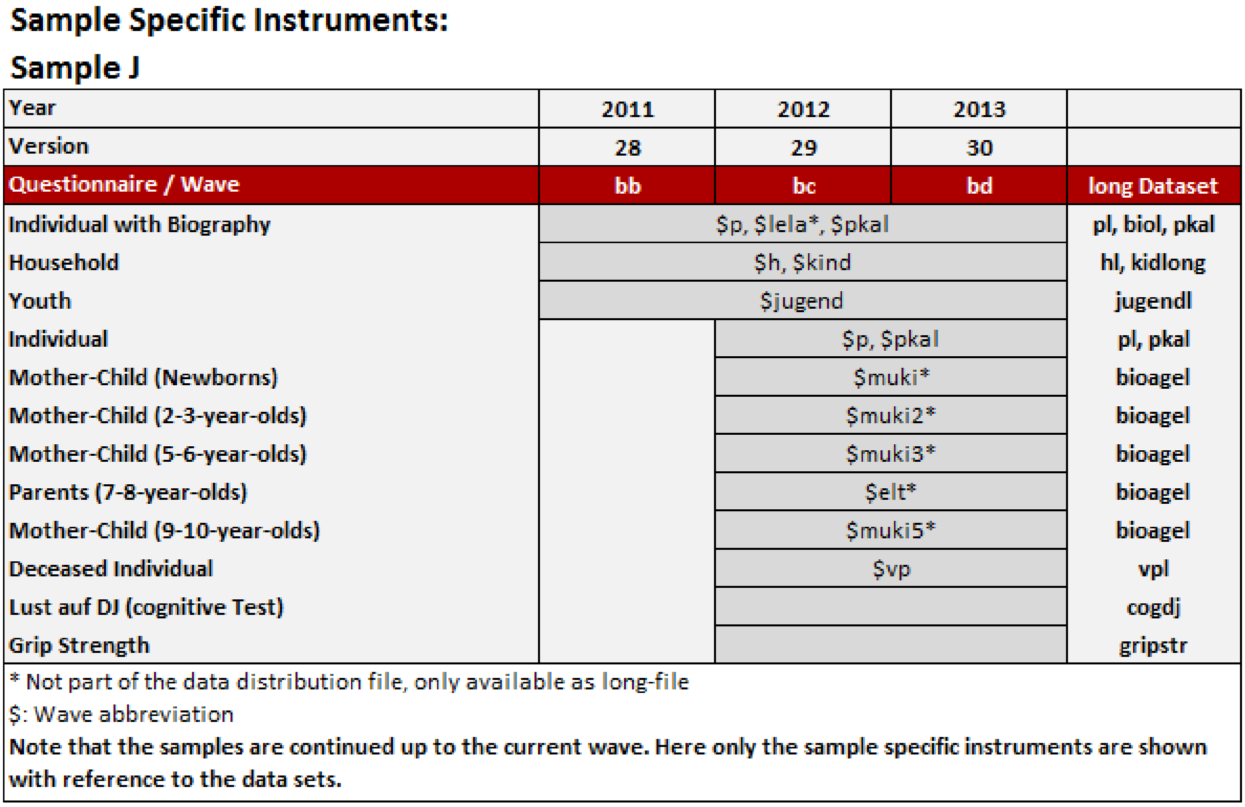
Year	1990	1991	1992	1993	1994	1995	
Version	7	8	9	10	11	12	
Questionnaire / Wave	g	h	i	j	k	l	long Dataset
Household East	ghost	$h, $kind					hl, kidlong
Individual East	$post, $pkalost, biolela*	$post, $pkalost	$p, $pkal				pl, biol, pkal
Individual East (New Respondents)		$post, $pkalost, biolela*	$p, $pkal				pl, biol, pkal
Biography East			biolela*				biol

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

SOEP researchers were determined to seize the historic opportunity of German reunification to obtain a first baseline measurement of incomes in the "old" GDR currency. The questionnaire was prepared by an East-West working group including DIW Berlin, WZB, Collaborative Research Centre 3, and the ISS at the Academy of Sciences in the GDR, with the participation of Infratest and its partner organization in the GDR. The result was a questionnaire that covered many of the same themes and questions and was structured similarly to the West SOEP questionnaire, but which focused more on the specific situation in the GDR (e.g., the housing situation).



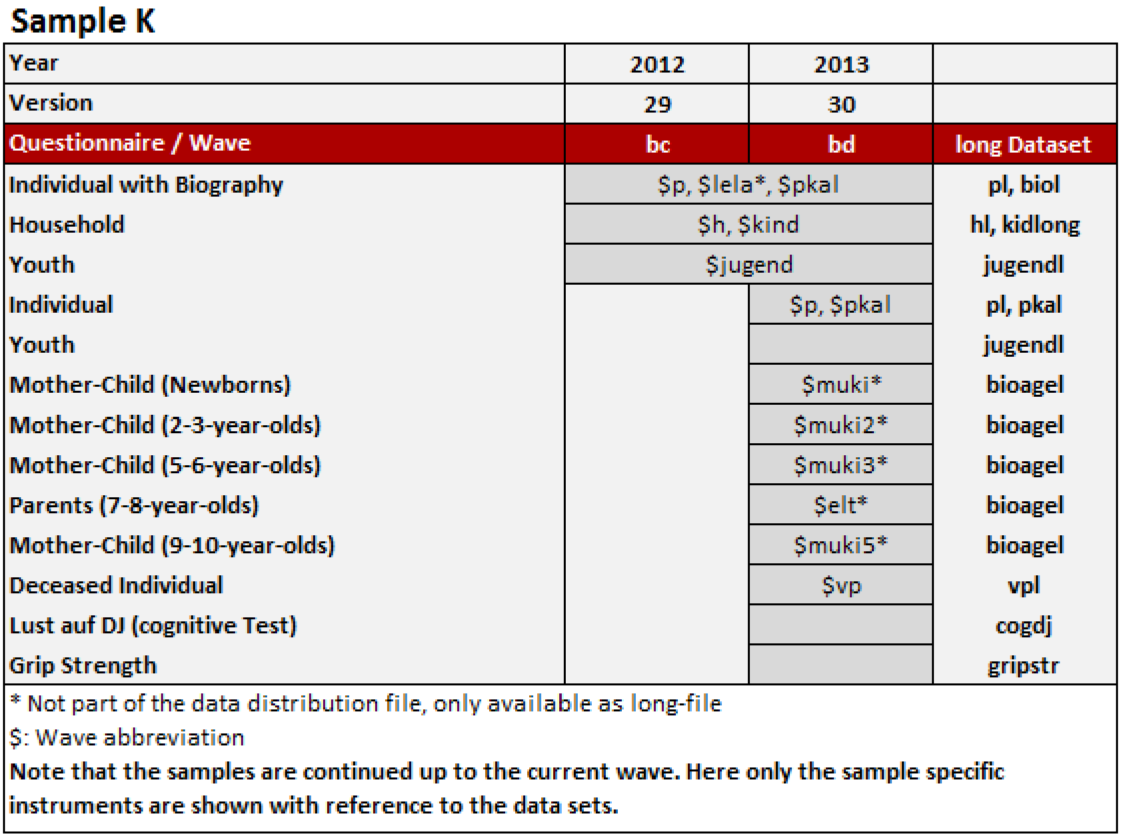
Sample Specific Instruments:
Sample J

Year	2011	2012	2013	
Version	28	29	30	
Questionnaire / Wave	bb	bc	bd	long Dataset
Individual with Biography	$p, $lela*, $pkal			pl, biol, pkal
Household	$h, $kind			hl, kidlong
Youth	$jugend			jugendl
Individual		$p, $pkal		pl, pkal
Mother-Child (Newborns)		$muki*		bioagel
Mother-Child (2-3-year-olds)		$muki2*		bioagel
Mother-Child (5-6-year-olds)		$muki3*		bioagel
Parents (7-8-year-olds)		$elt*		bioagel
Mother-Child (9-10-year-olds)		$muki5*		bioagel
Deceased Individual		$vp		vpl
Lust auf DJ (cognitive Test)				cogdj
Grip Strength				gripstr

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.



Sample K

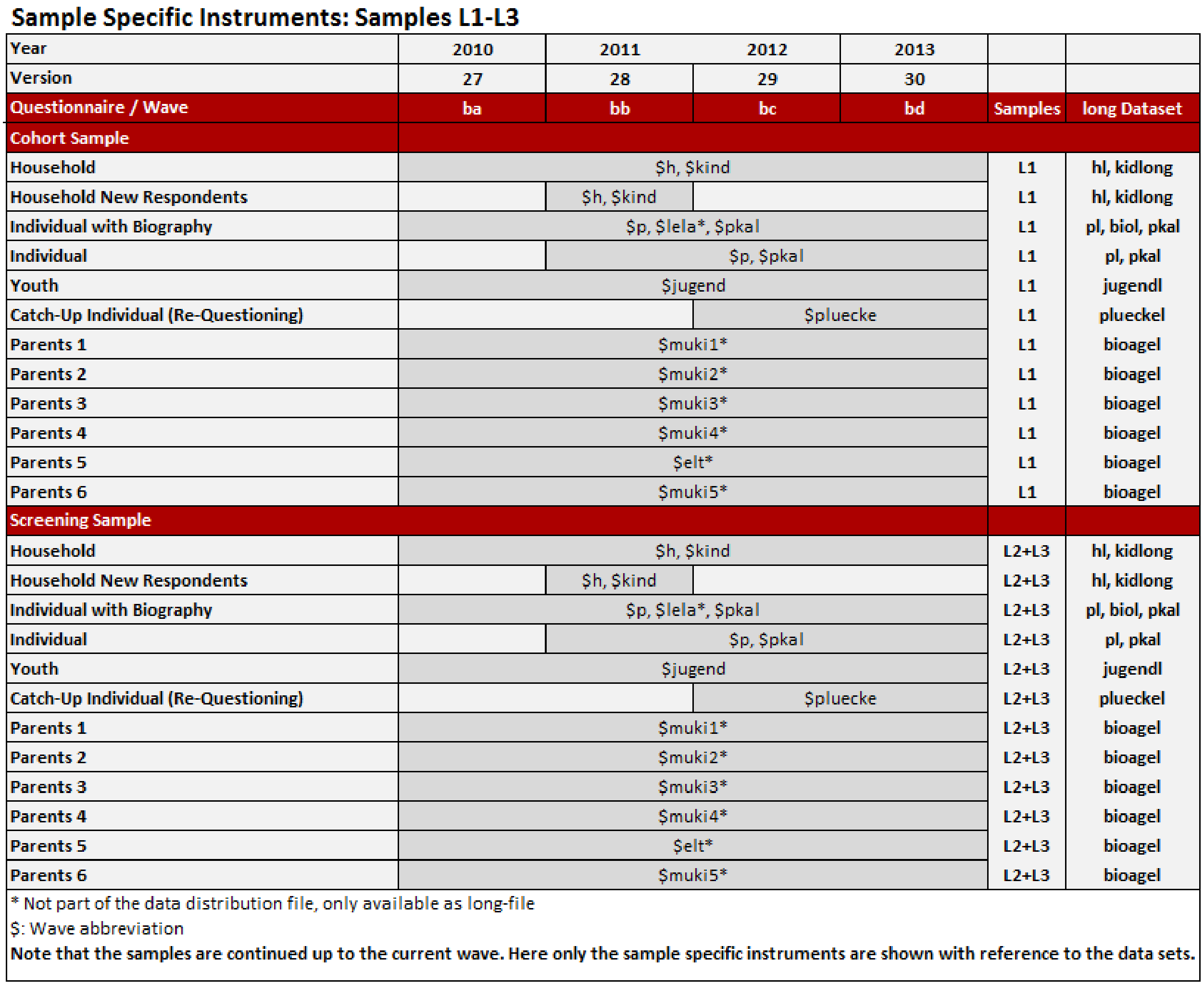
Year	2012	2013	
Version	29	30	
Questionnaire / Wave	bc	bd	long Dataset
Individual with Biography	$p, $lela*, $pkal		pl, biol
Household	$h, $kind		hl, kidlong
Youth	$jugend		jugendl
Individual		$p, $pkal	pl, pkal
Youth			jugendl
Mother-Child (Newborns)		$muki*	bioagel
Mother-Child (2-3-year-olds)		$muki2*	bioagel
Mother-Child (5-6-year-olds)		$muki3*	bioagel
Parents (7-8-year-olds)		$elt*	bioagel
Mother-Child (9-10-year-olds)		$muki5*	bioagel
Deceased Individual		$vp	vpl
Lust auf DJ (cognitive Test)			cogdj
Grip Strength			gripstr

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

A major shift in the design of SOEP questionnaires took place with Sample J. Due to the increased panel mortality from wave 1 to wave 2 that was observed for the refresher samples F (2000- 2001), H (2006-2007), and I (2009-2010), the biographical module, with an average interview length of 17 minutes, was integrated into wave 1. If this had not been done, no biographical data would have been collected for approximately 20% of all SOEP respondents who would probably not have participated in wave 2. In comparison to the longitudinal samples, data collection in the first wave was focused on the main three questionnaires: the household, the individual, and the youth questionnaire. As the fieldwork in these refresher samples was conducted exclusively by CAPI, it was feasible to include complex modules with event-triggered question loops.



Sample Specific Instruments: Samples L1-L3

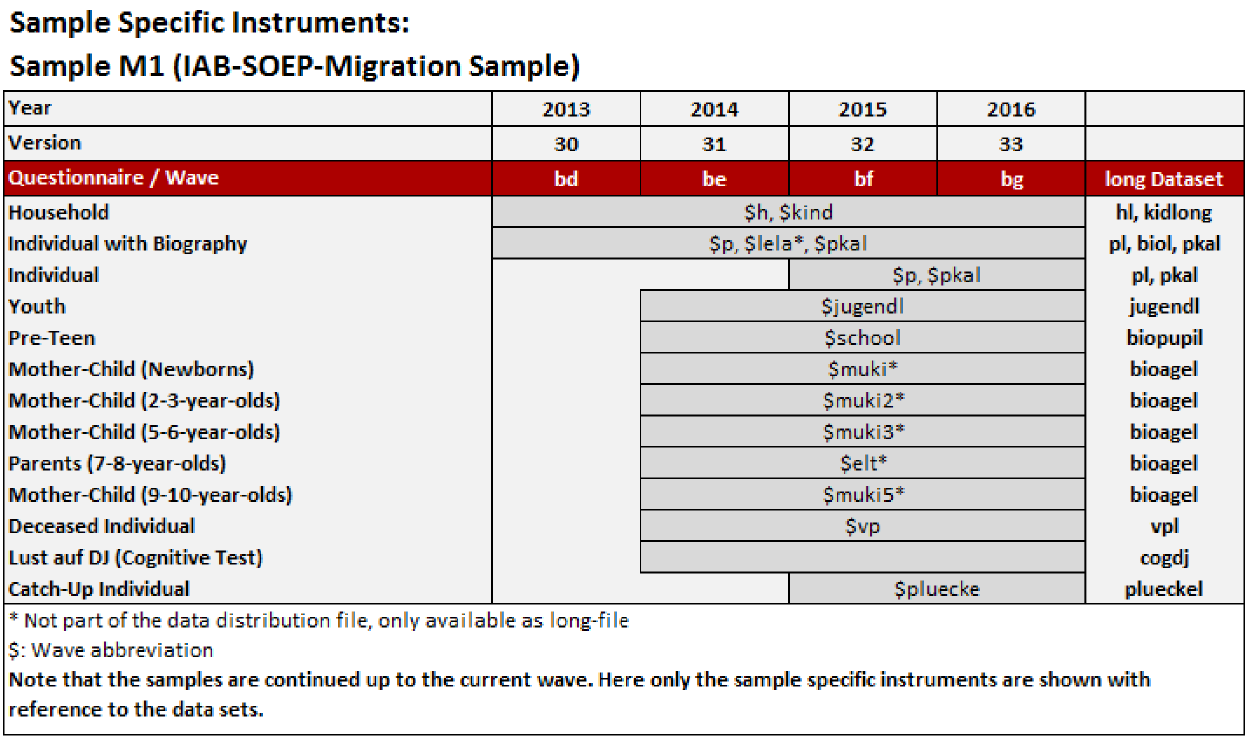
Year	2010	2011	2012	2013		
Version	27	28	29	30		
Questionnaire / Wave	ba	bb	bc	bd	Samples	long Dataset
Cohort Sample						
Household	$h, $kind				L1	hl, kidlong
Household New Respondents		$h, $kind			L1	hl, kidlong
Individual with Biography	$p, $lela*, $pkal				L1	pl, biol, pkal
Individual		$p, $pkal			L1	pl, pkal
Youth	$jugend				L1	jugendl
Catch-Up Individual (Re-Questioning)			$pluecke		L1	plueckel
Parents 1	$muki1*				L1	bioagel
Parents 2	$muki2*				L1	bioagel
Parents 3	$muki3*				L1	bioagel
Parents 4	$muki4*				L1	bioagel
Parents 5	$elt*				L1	bioagel
Parents 6	$muki5*				L1	bioagel
Screening Sample						
Household	$h, $kind				L2+L3	hl, kidlong
Household New Respondents		$h, $kind			L2+L3	hl, kidlong
Individual with Biography	$p, $lela*, $pkal				L2+L3	pl, biol, pkal
Individual		$p, $pkal			L2+L3	pl, pkal
Youth	$jugend				L2+L3	jugendl
Catch-Up Individual (Re-Questioning)			$pluecke		L2+L3	plueckel
Parents 1	$muki1*				L2+L3	bioagel
Parents 2	$muki2*				L2+L3	bioagel
Parents 3	$muki3*				L2+L3	bioagel
Parents 4	$muki4*				L2+L3	bioagel
Parents 5	$elt*				L2+L3	bioagel
Parents 6	$muki5*				L2+L3	bioagel

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

The main focus of Families in Germany (FiD) was on the families and children – the parental questionnaires (filled out by parents about their children) were about twice as long as the comparable questionnaires in SOEP-Core, and questionnaires for the 1-2-year-olds and the 9-10-year-olds were added (as of 2012, SOEP-Core had added a questionnaire for 9-10-year-olds that is partly comparable to the FiD version). In large part, FiD resembled the SOEP. Each adult was asked to answer an individual questionnaire, which, in the first two years, included retrospective questions on childhood, education, and early work experience. In addition, there were several questions designed to capture the challenges families face with regard to the return of mothers into the labor market – with respect to workplace, work schedule, overtime, daycare options, etc.



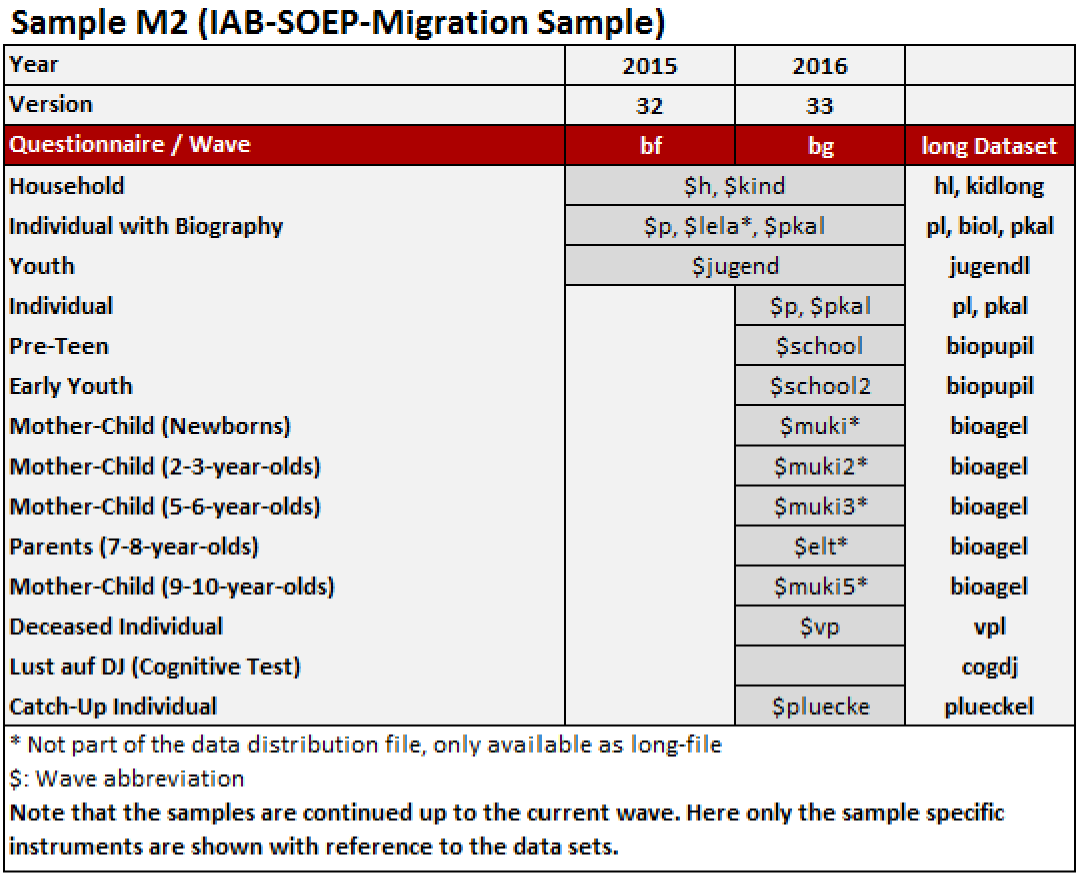
Sample Specific Instruments:
Sample M1 (IAB-SOEP-Migration Sample)

Year	2013	2014	2015	2016	
Version	30	31	32	33	
Questionnaire / Wave	bd	be	bf	bg	long Dataset
Household	$h, $kind				hl, kidlong
Individual with Biography	$p, $lela*, $pkal				pl, biol, pkal
Individual			$p, $pkal		pl, pkal
Youth		$jugendl			jugendl
Pre-Teen		$school			biopupil
Mother-Child (Newborns)		$muki*			bioagel
Mother-Child (2-3-year-olds)		$muki2*			bioagel
Mother-Child (5-6-year-olds)		$muki3*			bioagel
Parents (7-8-year-olds)		$elt*			bioagel
Mother-Child (9-10-year-olds)		$muki5*			bioagel
Deceased Individual		$vp			vpl
Lust auf DJ (Cognitive Test)					cogdj
Catch-Up Individual			$pluecke		plueckel

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.



Sample M2 (IAB-SOEP-Migration Sample)

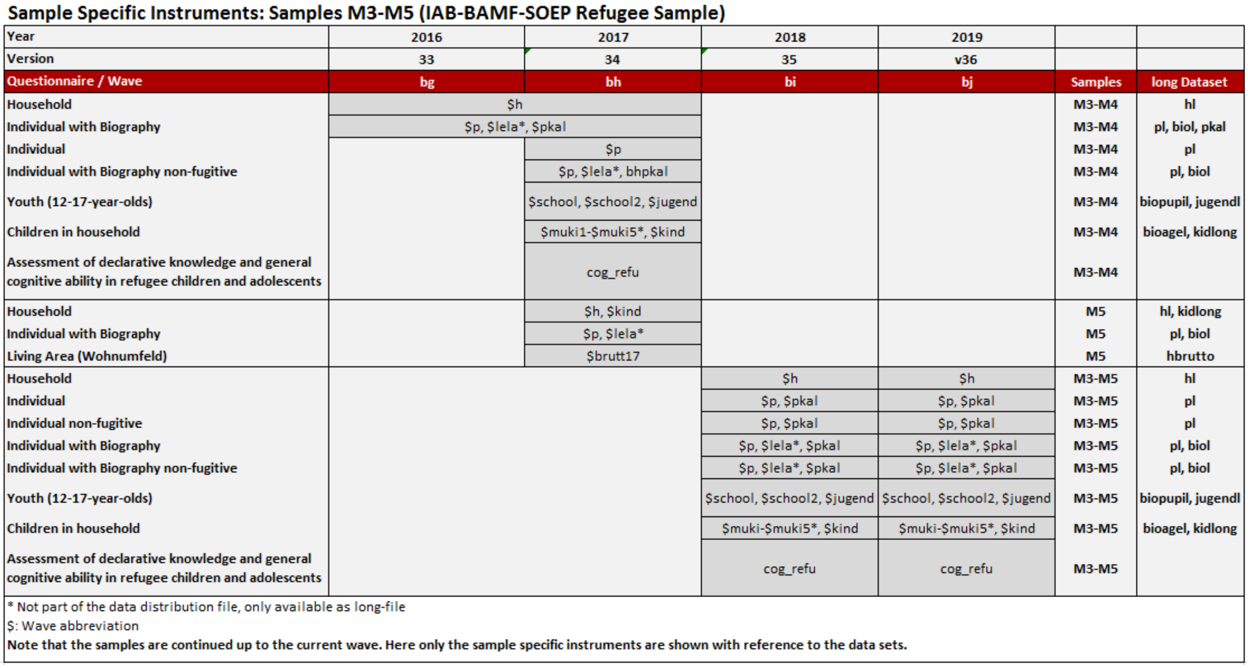
Year	2015	2016	
Version	32	33	
Questionnaire / Wave	bf	bg	long Dataset
Household	$h, $kind		hl, kidlong
Individual with Biography	$p, $lela*, $pkal		pl, biol, pkal
Youth	$jugend		jugendl
Individual		$p, $pkal	pl, pkal
Pre-Teen		$school	biopupil
Early Youth		$school2	biopupil
Mother-Child (Newborns)		$muki*	bioagel
Mother-Child (2-3-year-olds)		$muki2*	bioagel
Mother-Child (5-6-year-olds)		$muki3*	bioagel
Parents (7-8-year-olds)		$elt*	bioagel
Mother-Child (9-10-year-olds)		$muki5*	bioagel
Deceased Individual		$vp	vpl
Lust auf DJ (Cognitive Test)			cogdj
Catch-Up Individual		$pluecke	plueckel

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Following the design shift for refresher samples since Sample J in 2011, respondents have been surveyed on their life history using the “biography questionnaire”, which was integrated into the individual questionnaire from wave 1. This ensures that biographical information will be available for all target persons who provided an individual interview in participating households. Other supplementary questionnaires were not included in the survey instruments given to first-wave respondents to avoid “overburdening” respondents with an extremely lengthy first-wave interview. Questionnaires for the migration boost samples include questions that have been part of SOEP-Core for the last three decades. In addition, the survey covers each respondent’s complete migration history, education, training, and employment history in Germany and abroad, and numerous aspects of cultural and living environments relevant to the social integration of migrants. The household questionnaire is identical to the questionnaire used in the SOEP-Core sample.



Sample Specific Instruments: Samples M3-M5 (IAB-BAMF-SOEP Refugee Sample)

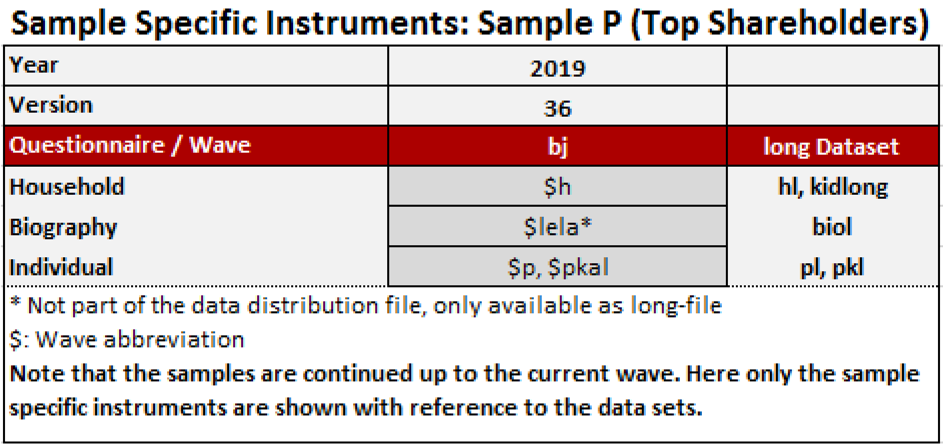
Year	2016	2017	2018	2019		
Version	33	34	35	v36		
Questionnaire / Wave	bg	bh	bi	bj	Samples	long Dataset
Household	$h				M3-M4	hl
Individual with Biography	$p, $lela*, $pkal				M3-M4	pl, biol, pkal
Individual		$p			M3-M4	pl
Individual with Biography non-fugitive		$p, $lela*, bhpkal			M3-M4	pl, biol
Youth (12-17-year-olds)		$school, $school2, $jugend			M3-M4	biopupil, jugendl
Children in household		$muki1-$muki5*, $kind			M3-M4	bioagel, kidlong
Assessment of declarative knowledge and general cognitive ability in refugee children and adolescents		cog_refu			M3-M4	
Household		$h, $kind			M5	hl, kidlong
Individual with Biography		$p, $lela*			M5	pl, biol
Living Area (Wohnumfeld)		$brutt17			M5	hbrutto
Household			$h	$h	M3-M5	hl
Individual			$p, $pkal	$p, $pkal	M3-M5	pl
Individual non-fugitive			$p, $pkal	$p, $pkal	M3-M5	pl
Individual with Biography			$p, $lela*, $pkal	$p, $lela*, $pkal	M3-M5	pl, biol
Individual with Biography non-fugitive			$p, $lela*, $pkal	$p, $lela*, $pkal	M3-M5	pl, biol
Youth (12-17-year-olds)			$school, $school2, $jugend	$school, $school2, $jugend	M3-M5	biopupil, jugendl
Children in household			$muki-$muki5*, $kind	$muki-$muki5*, $kind	M3-M5	bioagel, kidlong
Assessment of declarative knowledge and general cognitive ability in refugee children and adolescents			cog_refu	cog_refu	M3-M5	

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

As with every other previously established subsample of migrants in the SOEP (M1 and M2), there was a clear need for several deviations from standard SOEP-Core questionnaires to reflect the special characteristics of the target group. Several additional questions concerning migration and integration were incorporated into the individual questionnaire to better field the range of research questions and research goals of the project partners. These included topics such as ethnic background, experiences en route to Germany, language skills, integration courses in Germany, job experience, current occupation, educational background, health, attitudes, and values. The household questionnaire was much more SOEP-related than the individual questionnaire in order to establish longitudinal information on the households.



Sample Specific Instruments: Sample P (Top Shareholders)

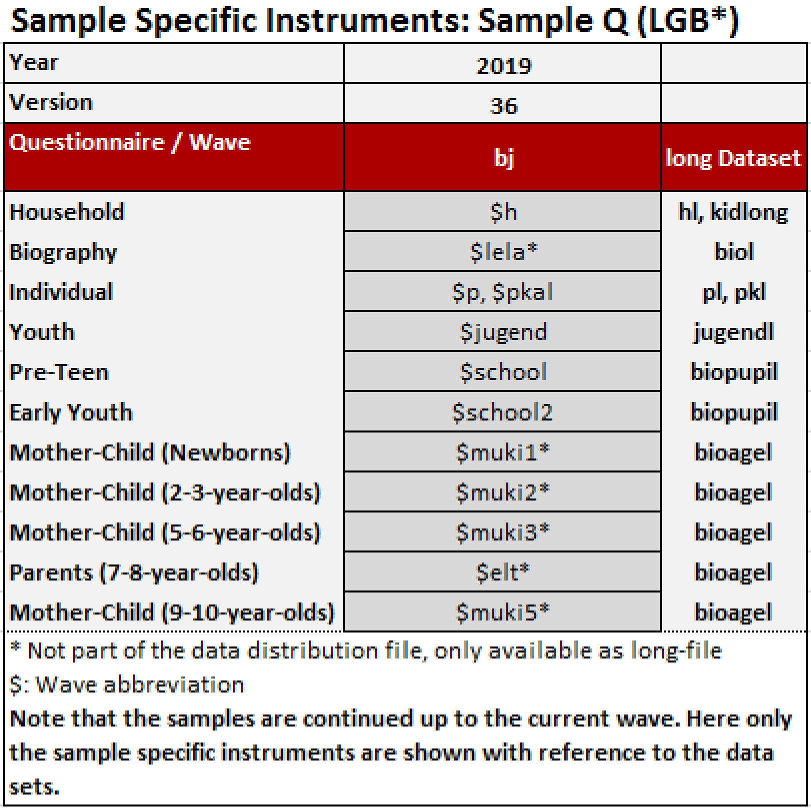
Year	2019	
Version	36	
Questionnaire / Wave	bj	long Dataset
Household	$h	hl, kidlong
Biography	$lela*	biol
Individual	$p, $pkal	pl, pkl

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Three different questionnaires were used to collect data in sample P. Apart from the regular household and individual questionnaires, a life-history questionnaire module was used to collect background information of all respondents. Computer-assisted personal interviewing (CAPI) was applied alongside paper questionnaires (PAPI or SELF) for all questionnaires. While the life history questionnaire was integrated into the individual questionnaire in the CAPI, it was administered as a separate questionnaire in the PAPI and SELF modes



Sample Specific Instruments: Sample Q (LGB*)

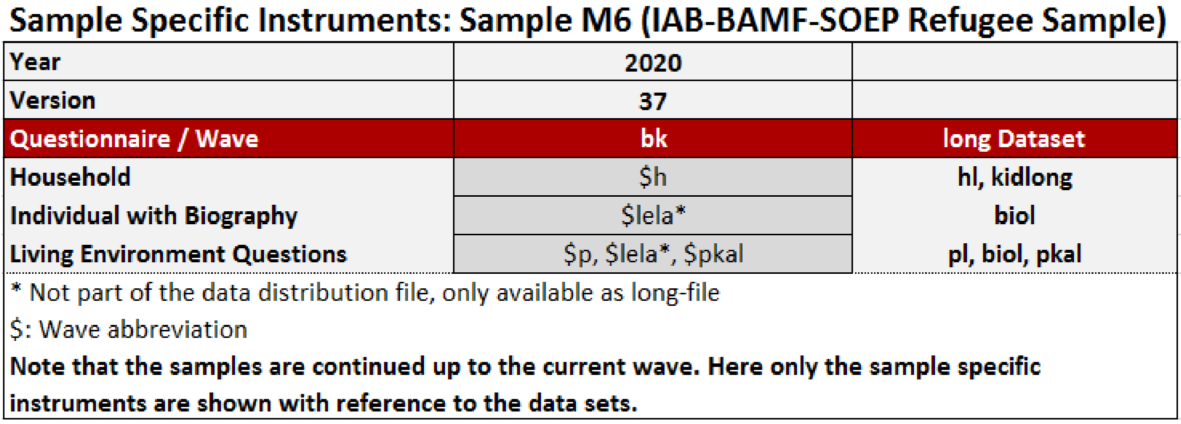
Year	2019	
Version	36	
Questionnaire / Wave	bj	long Dataset
Household	$h	hl, kidlong
Biography	$lela*	biol
Individual	$p, $pkal	pl, pkl
Youth	$jugend	jugendl
Pre-Teen	$school	biopupil
Early Youth	$school2	biopupil
Mother-Child (Newborns)	$muki1*	bioagel
Mother-Child (2-3-year-olds)	$muki2*	bioagel
Mother-Child (5-6-year-olds)	$muki3*	bioagel
Parents (7-8-year-olds)	$elt*	bioagel
Mother-Child (9-10-year-olds)	$muki5*	bioagel

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Eleven different questionnaires were used to collect data in sample Q. Apart from the regular household and individual questionnaires, a life-history questionnaire module was used to collect background information of all respondents. A special module regarding their sexual orientation was added in the individual questionnaire. Adolescents of the age 16 or 17, 13 or 14 and 11 or 12 were interviewed using specific youth questionnaires. Additionally, all mother and child /parent questionnaires were administered in this boost sample. Computer-assisted personal interviewing (CAPI) was applied exclusively for all questionnaires.



Sample Specific Instruments: Sample M6 (IAB-BAMF-SOEP Refugee Sample)

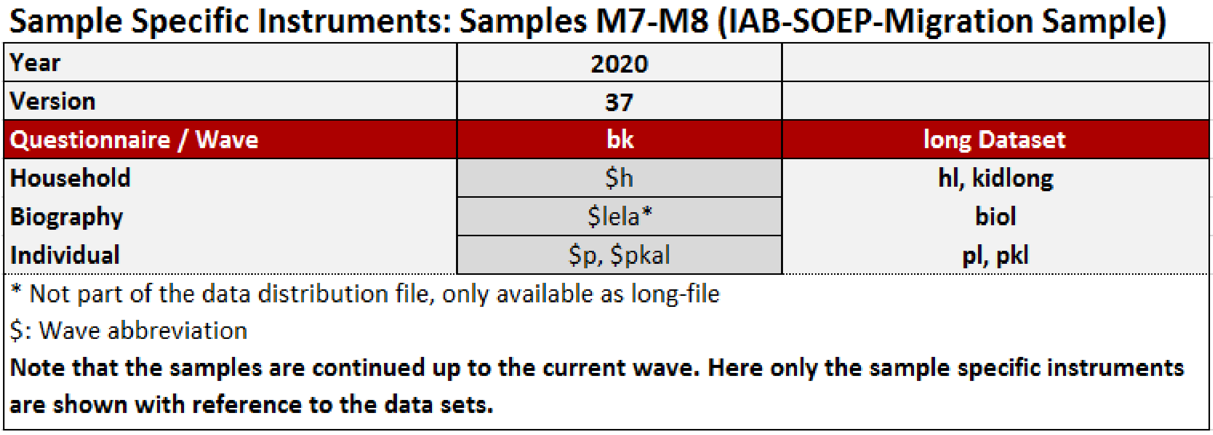
Year	2020	
Version	37	
Questionnaire / Wave	bk	long Dataset
Household	$h	hl, kidlong
Individual with Biography	$lela*	biol
Living Environment Questions	$p, $lela*, $pkal	pl, biol, pkal

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

In the first wave of M6 three questionnaires were fielded: the individual questionnaire for first time respondents (including additional biographical questions) for all adult household members, which was administered in separate versions for refugees and for Germans or migrants respectively, and the household questionnaire for the anchor respondent. Like for the other refugee samples M3-5, a special SOEP individual and life-history questionnaire was developed that includes issues specific to refugees. The version for Germans and migrants was identical to the individual and life-history questionnaire in samples A-Q and M1/2. As is the usual approach for boost samples, no youth or child questionnaires were fielded in sample M6. All questionnaires were solely available in CAPI mode and provided in seven different language versions, although a small percentage of interviews was conducted via telephone in the CAPI environment.



Sample Specific Instruments: Samples M7-M8 (IAB-SOEP-Migration Sample)

Year	2020	
Version	37	
Questionnaire / Wave	bk	long Dataset
Household	$h	hl, kidlong
Biography	$lela*	biol
Individual	$p, $pkal	pl, pkl

* Not part of the data distribution file, only available as long-file

$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

In the first waves of M7 and M8 three questionnaires were fielded: the individual questionnaire for first time respondents (including additional biographical questions) for all adult household members, which had the life-history module integrated in the CAPI-instrument and the household questionnaire for the anchor respondent. In addition to these instruments, anchor-respondents had to answer a short screening questionnaire in order to clarify their membership in the target populations of M7 and M8a respectively. Respondents had to have been born outside of Germany, their stay should not be temporary, and they were to have moved to Germany no earlier than 2016 (M7) or 2017(M8a) respectively. All questionnaires were solely available in CAPI mode. Translation aides were provided only in paper form in four additional languages. With regards to questionnaire content, the household and individual questionnaires were almost identical to the ones used in samples M1/2.

Last change: May 12, 2022

4.2 Eligibility and Follow-up

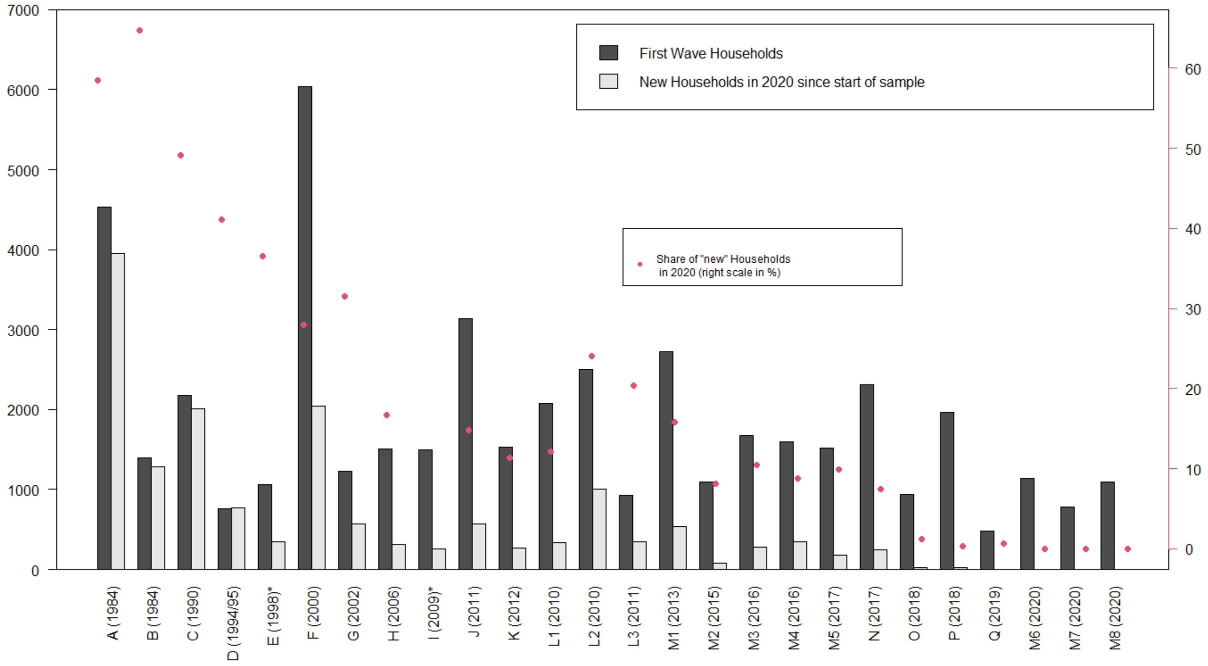
As mentioned, the SOEP's goal is to be representative of the residential population of Germany. All household members 16 and older are eligible for a personal interview, starting with the youth questionnaire for their age group, followed by "regular" individual questionnaires thereafter. As years go by, the children from the first wave reach age eligibility and become panel members. If they move out and start their own families, they and their new family members are also part of the survey. "New" individuals become part of the SOEP population by being born into SOEP households or as a result of residential mobility. If a person enters a SOEP household after the initial wave in which that household was surveyed, this person is asked to fill out the regular individual questionnaire if age-eligible or will be asked to participate once old enough. In the absence of panel attrition, this would make the SOEP a self-sustaining survey.

The concept of how to follow respondents and sample members over time is important for the representativeness of the study. The basic principle for follow-up in the SOEP is that all persons participating in a wave of any subsample will be surveyed in the following years as long as they stay within the boundaries of Germany. This rule also extends to respondents who entered a SOEP household after the first wave in which it was surveyed due to residential mobility or birth. If there is a "split-off", that is, if someone moves out of the household in which they were last interviewed, the members of the new household receive a new household identifier. The table conceptualizes how new sample members and households are surveyed in the SOEP. The figure shows that as a result of the follow-up concept, several thousand "new" households had become part of the SOEP population.

Individuals or households that could not be interviewed in a given year are termed "temporary drop-outs". They are followed until there are two consecutive waves of missing interviews for all household members or until the entire household refuses to participate further. In the case of a temporary drop-out, in which a respondent participates in the survey again after not participating in the previous wave, the respondent is asked to fill out an additional short questionnaire covering key information about employment and demographics in the year of their absence.

	Existing Households	New Households
Existing Individuals	Classic case: without change of address entire household moves	Respondent moves out of existing household and forms a new household
New Individuals	Born into household, move into household	Moved into or born into HH formed after a respondent moved out of existing HH

Changes to the Sample: Old and new household in the SOEP



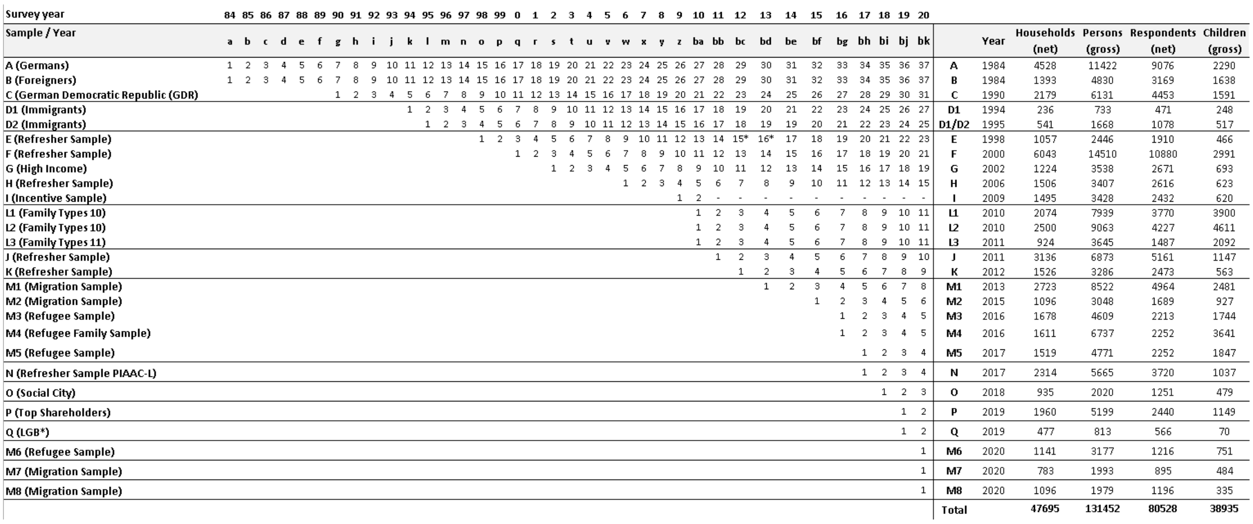
`Download R Code to create figure`

Last change: May 12, 2022

4.3 Development of Sample Sizes

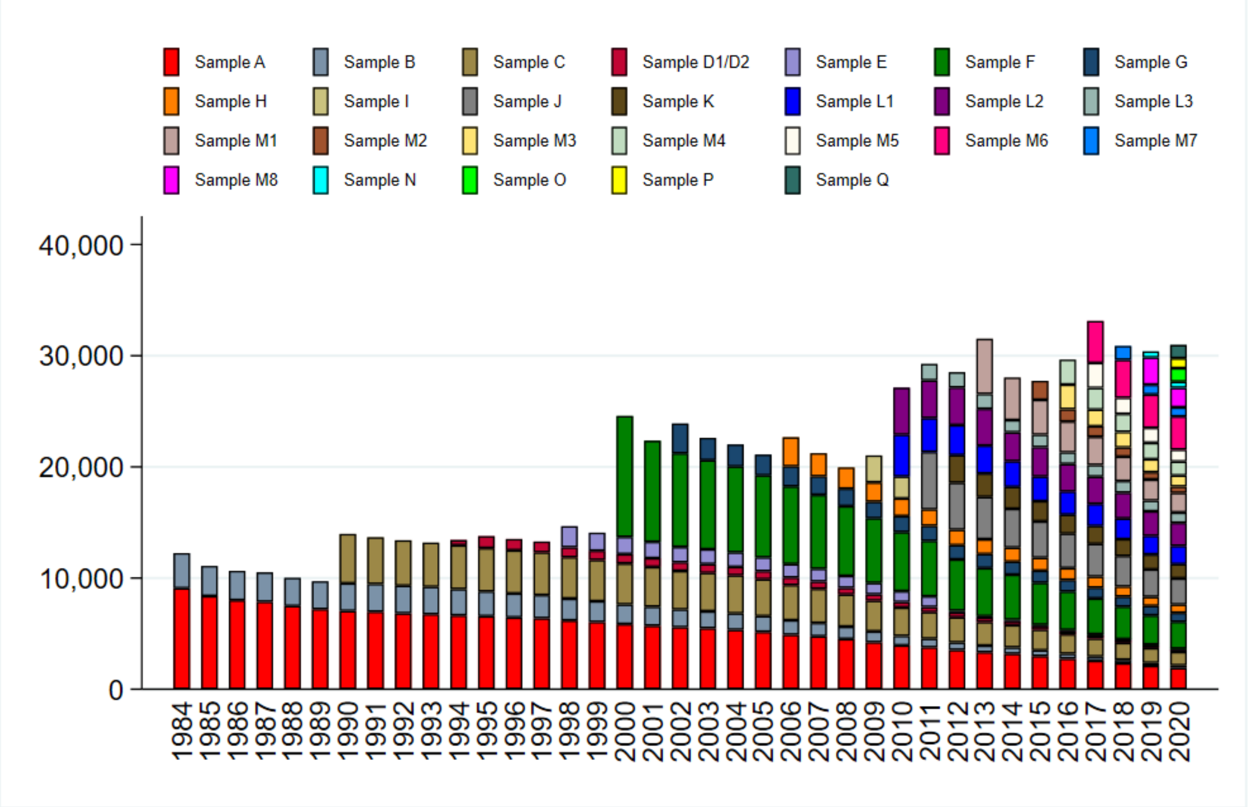
Individuals who decline to take part in the survey or are not available for an interview are kept in the so-called "gross" sample of the study as long as they continue to live in households with at least one participating respondent. If the entire household declines to participate in two consecutive waves, all individuals in the household are removed from the SOEP. The table shows the starting sample sizes of samples A through M4, the years when the samples were first collected, as well as the percentage of those persons who were eligible for an interview but declined participation ("partial unit non-response", PUNR) in the first wave. The figure illustrates the development of the number of successful person interviews since 1984. The reduction in the population size for all individual samples is mainly the result of individual-level drop-outs, refusals, moving abroad, etc. However, due to new persons moving into already existing households and children reaching the age of 16 and thereby increasing the sample size, this negative development is offset somewhat.

Starting Sample Size of the SOEP Samples



Survey year	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20						
Sample / Year	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	ba	bb	bc	bd	be	bf	bg	bh	bi	bj	bk		Year	Households (net)	Persons (gross)	Respondents (net)	Children (gross)
A (Germans)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	A	1984	4528	11422	9076	2290
B (Foreigners)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	B	1984	1393	4830	3169	1638
C (German Democratic Republic (GDR)							1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	C	1990	2179	6131	4453	1591
D1 (Immigrants)											1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	D1	1994	236	733	471	248
D2 (Immigrants)												1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	22	23	24	25	D1/D2	1995	541	1668	1078	517
E (Refresher Sample)															1	2	3	4	5	6	7	8	9	10	11	12	13	14	15*	16*	17	18	19	20	21	22	23	E	1998	1057	2446	1910	466
F (Refresher Sample)																	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	F	2000	6043	14510	10880	2991
G (High Income)																			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	G	2002	1224	3538	2671	693
H (Refresher Sample)																							1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	H	2006	1506	3407	2616	623
I (Incentive Sample)																										1	2	-	-	-	-	-	-	-	-	-	-	I	2009	1495	3428	2432	620
L1 (Family Types 10)																											1	2	3	4	5	6	7	8	9	10	11	L1	2010	2074	7939	3770	3900
L2 (Family Types 10)																											1	2	3	4	5	6	7	8	9	10	11	L2	2010	2500	9063	4227	4611
L3 (Family Types 11)																											1	2	3	4	5	6	7	8	9	10	11	L3	2011	924	3645	1487	2092
J (Refresher Sample)																												1	2	3	4	5	6	7	8	9	10	J	2011	3136	6873	5161	1147
K (Refresher Sample)																													1	2	3	4	5	6	7	8	9	K	2012	1526	3286	2473	563
M1 (Migration Sample)																														1	2	3	4	5	6	7	8	M1	2013	2723	8522	4964	2481
M2 (Migration Sample)																																1	2	3	4	5	6	M2	2015	1096	3048	1689	927
M3 (Refugee Sample)																																	1	2	3	4	5	M3	2016	1678	4609	2213	1744
M4 (Refugee Family Sample)																																	1	2	3	4	5	M4	2016	1611	6737	2252	3641
M5 (Refugee Sample)																																		1	2	3	4	M5	2017	1519	4771	2252	1847
N (Refresher Sample PIAAC-L)																																		1	2	3	4	N	2017	2314	5665	3720	1037
O (Social City)																																			1	2	3	O	2018	935	2020	1251	479
P (Top Shareholders)																																				1	2	P	2019	1960	5199	2440	1149
Q (LGB*)																																				1	2	Q	2019	477	813	566	70
M6 (Refugee Sample)																																					1	M6	2020	1141	3177	1216	751
M7 (Migration Sample)																																					1	M7	2020	783	1993	895	484
M8 (Migration Sample)																																					1	M8	2020	1096	1979	1196	335
																																						Total		47695	131452	80528	38935

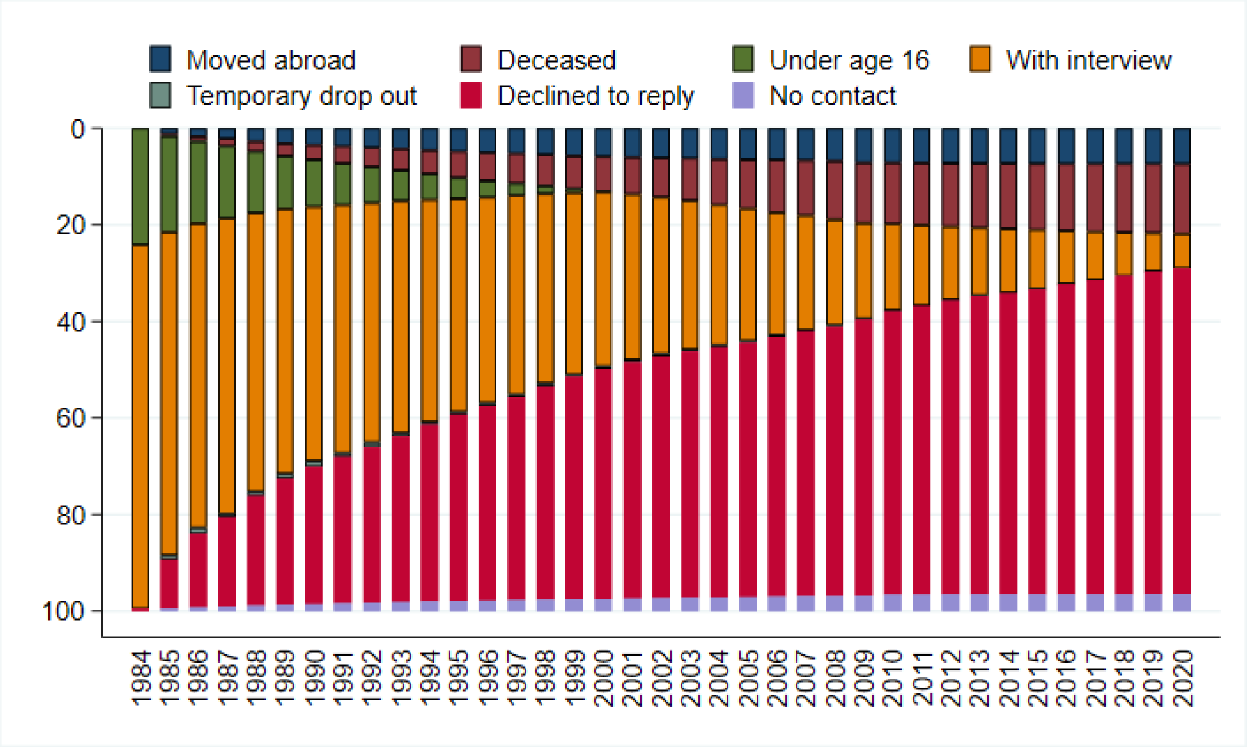
Cross-Sectional Development of Sample Size (Respondents)



Download Stata Code to create figure

This cross-sectional view is insufficient when examining the longitudinal development of the sample, which is influenced by different demographic and fieldwork-related factors. As already shown, demographic reasons for entering the panel are birth and residential mobility. Analogously, the demographic reasons for a panel exit are death and moving abroad. Fieldwork-related reasons are different, in that they relate to the interaction between the interviewer and the responding household. Respondents are either not reached for an interview (non-contact) or they decline to participate for the current year. The figure illustrates the longitudinal development of first-wave respondents in 1984, as well as their children, of samples A and B.

Longitudinal Development of the 1984 Population



`Download Stata Code to create figure`

Last change: May 12, 2022

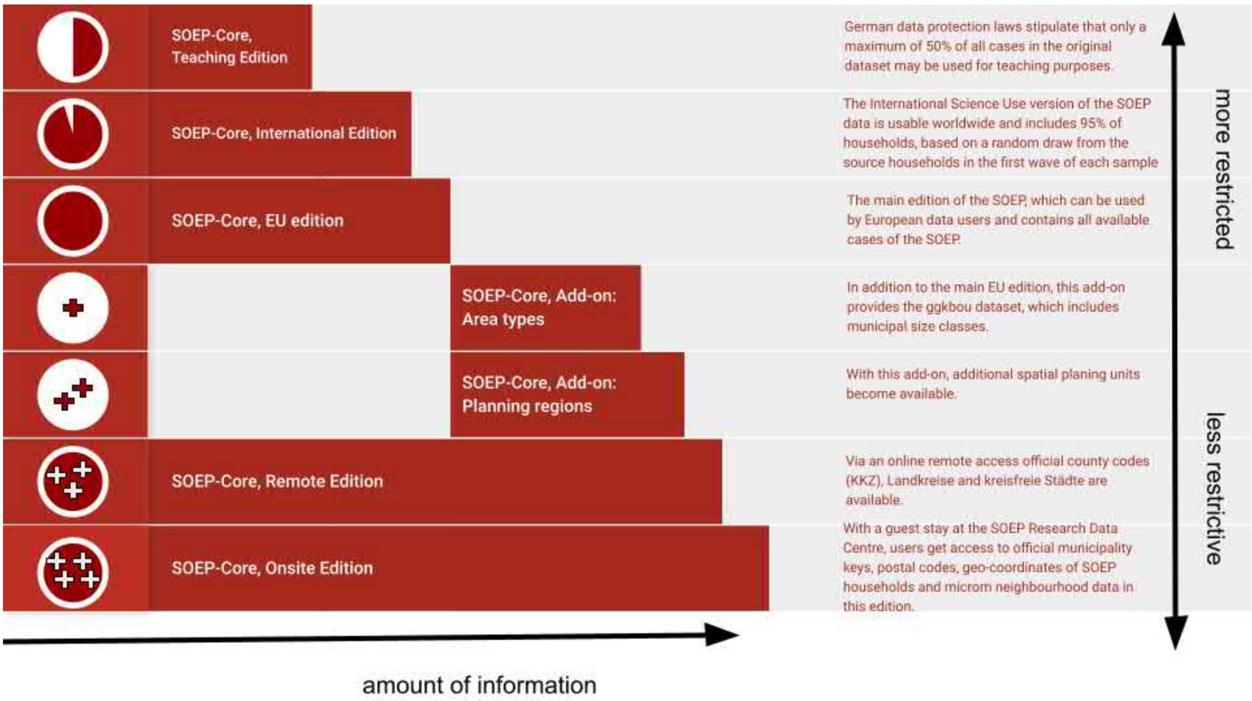
CHAPTER

FIVE

DATA STRUCTURE OF SOEP-CORE

5.1 Data Editions of SOEP-Core

Access to SOEP data is provided in compliance with the highest security standards to protect respondents' confidentiality and maintain their trust in the survey. The data are also provided solely for scientific research purposes, that is, they are only made available to members of the scientific community. This means that researchers are only given access to SOEP data after they have signed a data distribution contract with DIW Berlin. Different data packages, called "editions", reflect these requirements and can be differentiated by the amount of information contained in them, the level of data protection, and the mode of data access. The EU Edition is considered the standard edition. More restricted editions provide less information; less restricted editions provide more information but are only available under more restrictive conditions. The Teaching, International, and EU Editions *Teaching, International, and EU Edition* are made available as downloads under the standard data distribution contract, while the two add-ons Area Types and Planning Regions *Add-ons: Area Types and Planning Regions* require additional contracts. The Remote Edition *Remote Edition* can only be accessed through remote execution, and the Onsite Edition *Onsite Edition* can only be accessed on site at the SOEP Research Data Center at DIW Berlin.


SOEP-Core, Teaching Edition
SOEP-Core, International Edition
SOEP-Core, EU edition
SOEP-Core, Add-on: Area types
SOEP-Core, Add-on: Planning regions
SOEP-Core, Remote Edition
SOEP-Core, Onsite Edition
more restricted
less restrictive
amount of information

German data protection laws stipulate that only a maximum of 50% of all cases in the original dataset may be used for teaching purposes.

The International Science Use version of the SOEP data is usable worldwide and includes 95% of households, based on a random draw from the source households in the first wave of each sample

The main edition of the SOEP, which can be used by European data users and contains all available cases of the SOEP.

In addition to the main EU edition, this add-on provides the ggkbou dataset, which includes municipal size classes.

With this add-on, additional spatial planing units become available.

Via an online remote access official county codes (KKZ), Landkreise and kreisfreie Städte are available.

With a guest stay at the SOEP Research Data Centre, users get access to official municipality keys, postal codes, geo-coordinates of SOEP households and microm neighbourhood data in this edition.

In this figure, "more restrictive" means that existing variables from the EU Edition are left blank for reasons of data protection or not all cases are included. For example, variables that provide information at the federal state level are not available in the International or Teaching Editions (which only distinguish between East and West Germany). A higher level of data protection makes it possible to provide more information with fewer restrictions. This makes the editions less restrictive in terms of the information available. In most cases, as more sensitive information is added to an edition, access to the data edition changes and the requirements for its use also change.

5.1.1 Teaching, International, and EU Edition

Only the standard data distribution contract is required for the EU Edition and the International Edition. The EU Edition includes 100% of all observations, the German federal states, and the urban/rural variable. This edition is only available to users from research institutions in the EU and countries with an "adequacy decision" (Angemessenheitsbeschluss)—Switzerland, Japan, Canada, Israel, and a few others.

The International Edition is available to users from research institutions in all other countries than those listed above. This edition contains 95% of all households from the first wave of each SOEP subsample based on a random sampling of the original households in each subsample and only the East/West versions of variables normally containing the federal states. The original variables (with information on the federal states) remain in the edition but they are assigned the missing code -7 "Only available in less restricted edition" if a variable cannot be made accessible in a specific edition. For more information on the missing codes in SOEP-Core, see the chapter *Missing Conventions*.

The least restrictive edition of the data but the one containing the least information is the Teaching Edition. Here, a data distribution contract is required for teaching staff; students only need to sign the data protection declaration, which the contract holder must keep on file. The contract holder is responsible for ensuring strict adherence to data protection. German data protection laws stipulate that a maximum of 50% of all cases in the original dataset may be used for teaching purposes. The Teaching Edition has the same data structure as the International Edition (with the exception of the EU-SILC Clone) but contains half the number of cases in the EU Edition. The Teaching Edition provided to students must be stored in a separate hard drive area to which the user guarantees controlled access. Students may under no circumstances take data home with them or transfer the data to any other device at the university.

5.1.2 Add-ons: Area Types and Planning Regions

In addition to the EU Edition, the SOEP offers additional datasets that can extend the standard file to include municipality size classes (add-on: Area Types) or even spatial planning units (add-on: Planning Regions). Access to these files is more restricted because they provide users with more sensitive information about the respondents.

For the add-on Area Types, a regional data contract is required in addition to the data distribution contract. This requires that the user submit a data protection concept to the SOEP. There is no template for this; users must develop this concept specifically for the workplace in which they want to use the data.

For the add-on Planning Regions, a regional data contract is also necessary, and the SOEP requires that users submit a data protection concept that they have developed themselves. For this add-on, however, the requirements for the data protection concept are significantly higher.

5.1.3 Remote Edition

Further information such as official county codes (KKZ), identifying administrative districts (Landkreise) and urban districts (kreisfreie Städte) can be accessed through remote execution using the Remote Edition (or on site). For this edition, users are required to submit an application to use SOEPremote in addition to the data distribution contract. For the remote execution contract, no separate data protection concept is required, as users will only access the information remotely and no files are transmitted to computers outside of the Research Data Center of the SOEP (RDC SOEP).

To access the Remote Edition, there are two options available:

- SOEPremote execution (e-mail processing)
- SOEPremote access (on-site processing at special workstations)

With SOEPremote execution, users can email their Stata syntax to a remote server, which processes the syntax and returns the results to users by email. With SOEPremote access, users can use IGEL clients at RDC SOEP in Berlin. By using the IGEL clients, onsite users have the advantage of working directly with the Remote Edition instead of having to go through the email procedure. The disadvantage is having to plan and book a visit to the RDC SOEP in Berlin.

5.1.4 Onsite Edition

The Onsite Edition is the edition with all available information. Guests using RDC SOEP IGEL clients (*How to Use SOEP IGEL*) can access the additional information about the municipalities or postal codes of the SOEP households or data from microm GmbH on households' neighborhoods. Users can even analyze geocoded data. To access these data, researchers are first required to sign a data protection agreement, and a complete record is kept of all data access. The concept for providing the geo-coordinates of SOEP households is that the point coordinates are kept separate from the actual survey information throughout the entire process of analysis by data users due to privacy concerns. Researchers therefore never have simultaneous access to the SOEP survey data and the geo-coordinates of SOEP households. The results may only be published in completely anonymous form and are checked before they are transmitted from the secure server to the user.

- To apply to use a guest work station, click here:
- For more information about your workplace at the SOEP Research Data Center see the section *How to Use SOEP IGEL*
- For more information about how to work with SOEP's spatial data see the section *Working with spatial data in R*

Last change: May 12, 2022

5.2 Principles of Data Analysis

All SOEPtutorials can be found on our YouTube Channel

The structure of panel data has three dimensions. First, the respective examination units (n) and a matrix of dependent and independent variables (y,x) are completely analogous to a cross-sectional design. Second, the dimension of time (t), whereby a distinction is made between two data formats for panel data structures - "wide" or "long" (with wide format the variable matrix is indexed with the dimension of time and with long format the respective examination units). Regardless of the selected data format, when using panel data with several survey waves, the data matrices often do not contain complete information due to the panel mortality of individual survey units or because data from new panel members are only collected at a later point in time. In both cases, the term "unbalanced panel data" is used. In contrast, the classical panel data structure, on the other hand, is "balanced", i.e., as many observations of dependent and independent variables are available for all study units as there are waves of data collection. Social science panel data often show a data structure characterized by many investigation units (large n) as well as, in relation to it, few waves and therefore measuring time (small t). When data from a panel study are available, even descriptive forms of data analysis are often of particular interest, since the identification of changes in a variable over time and the corresponding separation of interindividual and intraindividual changes can represent important social facts, particularly in the case of generalizable samples. It is of social scientific interest whether a constant 15% proportion of people whose income is below the poverty risk level is repeatedly found in the same person over time, or whether there was a even balance of increases and decreases in poverty risks and only half of the population was permanently exposed to the risk. The choice of complex analysis methods for panel data depends first and foremost on the respective measurement level of the dependent and independent variables, but also on whether they are time-constant variables (such as gender or migration background) or time-invariant variables. The statistical analysis models of panel data range from structural equation models, various regression models, event analysis, sequence data analysis, latent growth models to causal analyses using matching methods. A particular advantage of panel data is that the chronological sequence of changes can be modelled and calculated and the problem of unobserved heterogeneity, which is often encountered in the social sciences, can be significantly reduced, at least in comparison with cross-sectional data.

5.2.1 Cross-Sectional Data Structure (CS)

Cross-sectional data is a type of data that observes many subjects at the same point in time. Each person is assigned a row in the dataset and is only included once in such a dataset. By merging cross-sectional SOEP data across waves, you obtain a dataset in wide-format.

Row	ID	wave	sex	income
1	1	2015	m	1500
2	2	2015	m	1000
3	6	2015	f	2000
4	8	2015	m	5500

5.2.2 Data Structure in "Wide" Format (wide)

The SOEP data are available with different data structures. In the wide format, a respondent's repeated responses are displayed in a single row and each response in a separate column. Each column represents a variable. We provide four datasets in the wide format: ppath, phrf, hpath, hhrf.

Row	ID	sex	income2015	income2016	income2017
1	1	m	1500	1500	2000
2	2	m	1000	1200	1200
3	6	f	2000	2000	2000
4	8	m	5500	6000	6500

5.2.3 Data Structure in "Long" Format (long)

The long format is a condensed and user-friendly dataset structure for longitudinal section analysis. Here, each person has one line per survey year. This means that you do not have several datasets for the different waves, but one dataset in which all survey waves are represented. A person can appear more than once in such a dataset. In the long format, one line describes a person-year combination.

Row	ID	syear	sex	income
1	1	2015	m	1500
2	1	2016	m	1500
3	1	2017	m	2000
4	2	2015	m	1000
5	2	2016	m	1200
6	2	2017	m	1200
7	6	2015	f	2000
8	6	2016	f	2000
9	6	2017	f	2000

In addition to the classic long format where one row in the dataset describes a person-year combination, there are also datasets that describe a longer period or a whole life, but only appear uniquely in the dataset without a survey year. These data sets can contain longitudinal information, but are constant over time. These time-constant data sets may include, for example, information on biological parents or employment history up to a certain age.

Row	ID	father_id	mother_id
1	20	18	19
2	21	18	19
3	35	34	35
4	36	34	35
5	37	34	35

5.2.4 Data Structure in Spell Format (spell)

In the strict sense of the word, spell data are about time periods with a defined start and end. When handling spell data it is necessary to take potential censoring into account. Censoring denotes that the beginning (left censored) or ending (right censored) of a spell is imprecise because of missing information or the beginning or ending of a spell is outside of the period of observation. It is quite conceivable that a person has only one spell over a given period, such as a male who is full-time employed. For a ten year period, there may be just the one spell "full-time employed". In panel data, the same person would have 10 observations, one per year. A person may have many spells over a time period, and even have overlapping spells, like working part-time and receiving a disability pension. Spell data are useful for looking at stays in a certain state, and transitions in and out of that state.

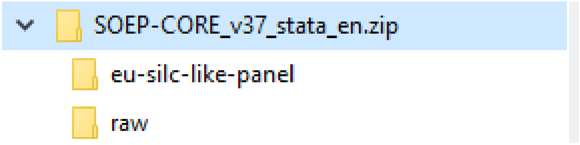
Row	ID	spellnr	spelltype	begin	end	censored
1	1	1	Retired	1983	2007	left and right censored
2	1	2	Housewife/husband	1983	1984	left censored
3	1	3	Housewife/husband	1994	1994	uncensored
4	1	4	Housewife/husband	1998	1998	uncensored
5	2	1	Full-Time Employment	1984	1984	left censored
6	2	2	Full-Time Employment	1985	1985	uncensored

Last change: May 12, 2022

5.3 Data Distribution File

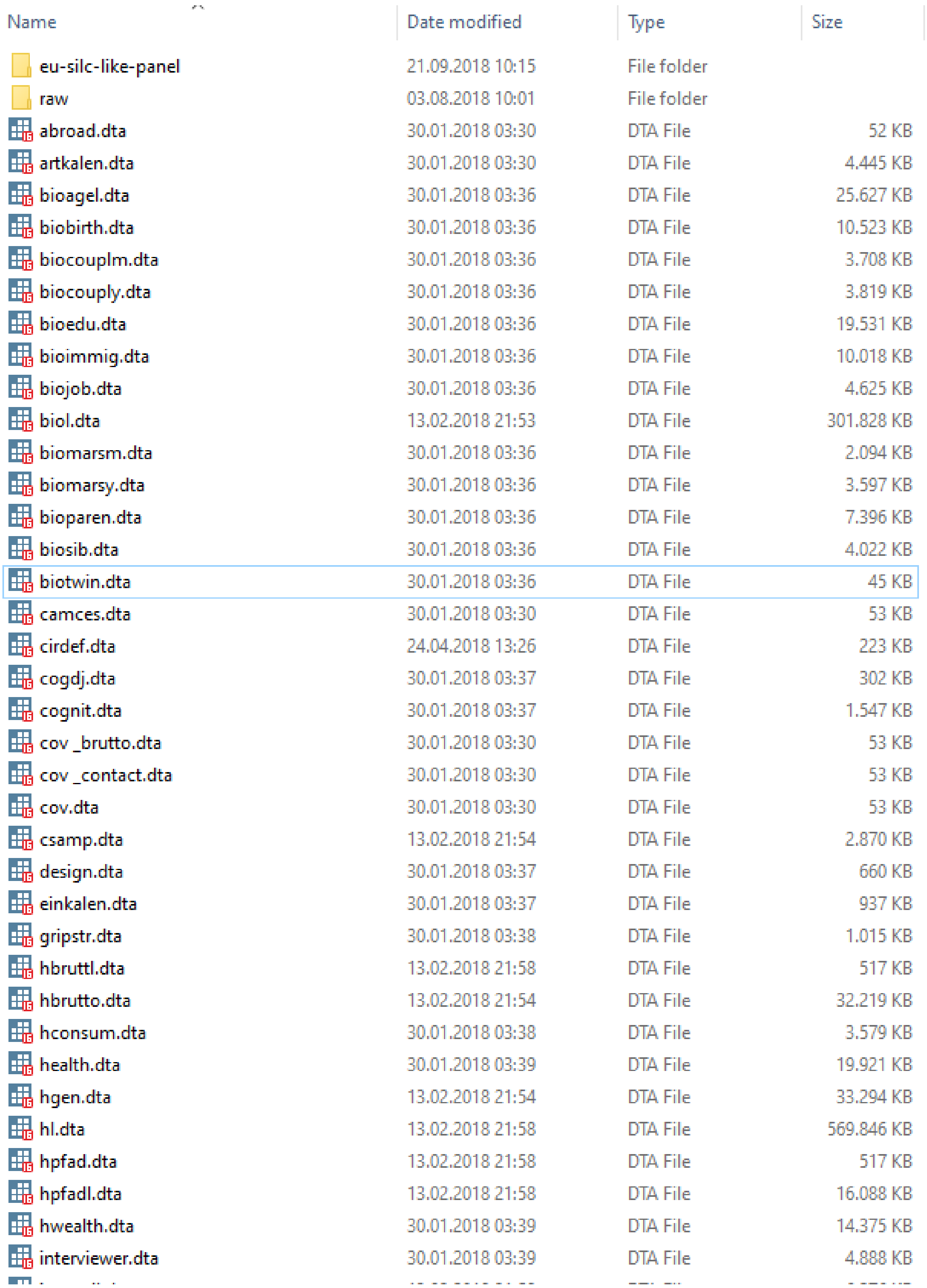
In the SOEP, each survey year is allocated to a data wave, which is abbreviated using the letters of the alphabet. One data wave may be released in several versions, which are displayed in SOEP with a "v" for version and the respective version number. The version number represents the survey years since the beginning of the survey. The SOEP has recently published the 34th version since the survey began in 1984. Within a data wave, updates may be made over time, such as v34.1. If updates have been made, users will be informed through various channels and be asked to order the data again. After ordering the data, the data will be sent to you in a zip file.

Within this zip file you will find various datasets, a "raw" subdirectory and the "eu-silc-like-panel" subdirectory.



The datasets in the top-level folder are a highly compressed and easy-to-analyze version of the SOEP data.

Note: SOEP strongly recommends that users use the top-level folder.



Name	Date modified	Type	Size
eu-silc-like-panel	21.09.2018 10:15	File folder	
raw	03.08.2018 10:01	File folder	
abroad.dta	30.01.2018 03:30	DTA File	52 KB
artkalen.dta	30.01.2018 03:30	DTA File	4.445 KB
bioagel.dta	30.01.2018 03:36	DTA File	25.627 KB
biobirth.dta	30.01.2018 03:36	DTA File	10.523 KB
biocouplm.dta	30.01.2018 03:36	DTA File	3.708 KB
biocouply.dta	30.01.2018 03:36	DTA File	3.819 KB
bioedu.dta	30.01.2018 03:36	DTA File	19.531 KB
bioimmig.dta	30.01.2018 03:36	DTA File	10.018 KB
biojob.dta	30.01.2018 03:36	DTA File	4.625 KB
biol.dta	13.02.2018 21:53	DTA File	301.828 KB
biomarsm.dta	30.01.2018 03:36	DTA File	2.094 KB
biomarsy.dta	30.01.2018 03:36	DTA File	3.597 KB
bioparen.dta	30.01.2018 03:36	DTA File	7.396 KB
biosib.dta	30.01.2018 03:36	DTA File	4.022 KB
biotwin.dta	30.01.2018 03:36	DTA File	45 KB
camces.dta	30.01.2018 03:30	DTA File	53 KB
cirdef.dta	24.04.2018 13:26	DTA File	223 KB
cogdj.dta	30.01.2018 03:37	DTA File	302 KB
cognit.dta	30.01.2018 03:37	DTA File	1.547 KB
cov _brutto.dta	30.01.2018 03:30	DTA File	53 KB
cov _contact.dta	30.01.2018 03:30	DTA File	53 KB
cov.dta	30.01.2018 03:30	DTA File	53 KB
csamp.dta	13.02.2018 21:54	DTA File	2.870 KB
design.dta	30.01.2018 03:37	DTA File	660 KB
einkalen.dta	30.01.2018 03:37	DTA File	937 KB
gripstr.dta	30.01.2018 03:38	DTA File	1.015 KB
hbruttl.dta	13.02.2018 21:58	DTA File	517 KB
hbrutto.dta	13.02.2018 21:54	DTA File	32.219 KB
hconsum.dta	30.01.2018 03:38	DTA File	3.579 KB
health.dta	30.01.2018 03:39	DTA File	19.921 KB
hgen.dta	13.02.2018 21:54	DTA File	33.294 KB
hl.dta	13.02.2018 21:58	DTA File	569.846 KB
hpfad.dta	13.02.2018 21:58	DTA File	517 KB
hpfadl.dta	13.02.2018 21:58	DTA File	16.088 KB
hwealth.dta	30.01.2018 03:39	DTA File	14.375 KB
interviewer.dta	30.01.2018 03:39	DTA File	4.888 KB

The data in SOEP-Core are no longer provided only as wave-specific individual files but are now pooled across all available years (in "long" format). In some cases, variables are harmonized to ensure that they are defined consistently over time. For example, the income information provided up to 2001 is given in euros, and categories are modified over time when versions of the questionnaire have been changed. The longitudinal nature of the data is one of the biggest assets of the SOEP. This is why we provide longitudinal datasets such as PL or HL. The advantage of such a dataset is that longitudinal analyses can be carried out without great effort.

If you need more information about the "long" data structure, see chapter *Data Structure in "Long" Format (long)*.

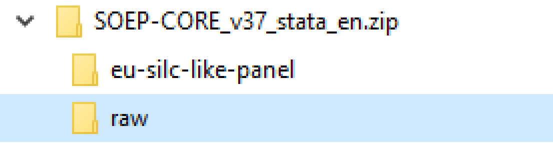
5.3.1 Core Datasets

The datasets in the top-level folder:

Tracking Data	Original Data	Survey Data	Generated Data	Spell Data
ppathl	pl	design	pgen	artkalen
hpathl	hl	exit	hgen	biocouplm
pbrutto	biol	cov_contact	bioagel	biocouply
hbrutto	jugendl		kidlong	biomarsm
hbrutt	plueckel		pequiv	biomarsy
pbr_exit	abroad		biobirth	einkalen
cov_brutto	vpl		bioedu	lifespell
	cov		bioimmig	migspell
			biojob	pbiospe
			bioparen	refugspell
			biopupil	sozkalen
			biosib	
			biotwin	
			camces	
			cogdj	
			cognit	
			cog_refu	
			gripstr	
			hconsum	
			health	
			hwealth	
			interviewer	
			mihinc	
			pflege	
			pkal	
			pwealth	
			timepref	
			trust	

5.3.2 Raw Datasets

In the "raw" directory, you will find all wave-specific datasets that were used to generate the long datasets on the previously presented level.

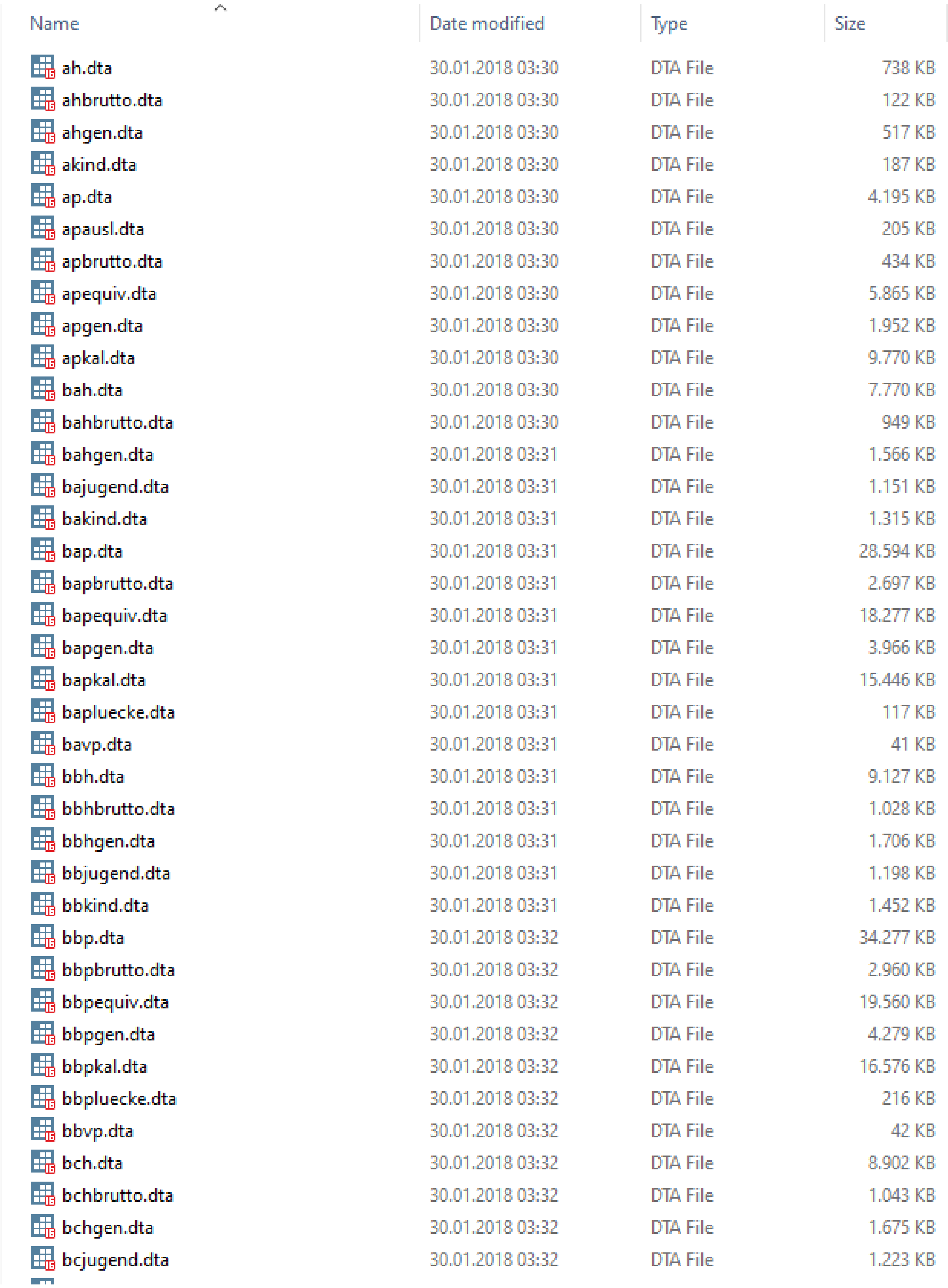


SOEP-CORE_v37_stata_en.zip

eu-silc-like-panel

raw

Attention: Please note that the datasets in the top-level folder are completely sufficient for your data analysis. The datasets used to generate the SOEP-Core data can be found in the raw subdirectory. Detailed information about the raw datasets can be found here Raw **"raw"**



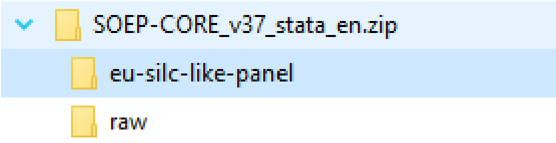
Name	Date modified	Type	Size
ah.dta	30.01.2018 03:30	DTA File	738 KB
ahbrutto.dta	30.01.2018 03:30	DTA File	122 KB
ahgen.dta	30.01.2018 03:30	DTA File	517 KB
akind.dta	30.01.2018 03:30	DTA File	187 KB
ap.dta	30.01.2018 03:30	DTA File	4.195 KB
apausl.dta	30.01.2018 03:30	DTA File	205 KB
apbrutto.dta	30.01.2018 03:30	DTA File	434 KB
apequiv.dta	30.01.2018 03:30	DTA File	5.865 KB
apgen.dta	30.01.2018 03:30	DTA File	1.952 KB
apkal.dta	30.01.2018 03:30	DTA File	9.770 KB
bah.dta	30.01.2018 03:30	DTA File	7.770 KB
bahbrutto.dta	30.01.2018 03:30	DTA File	949 KB
bahgen.dta	30.01.2018 03:31	DTA File	1.566 KB
bajugend.dta	30.01.2018 03:31	DTA File	1.151 KB
bakind.dta	30.01.2018 03:31	DTA File	1.315 KB
bap.dta	30.01.2018 03:31	DTA File	28.594 KB
bapbrutto.dta	30.01.2018 03:31	DTA File	2.697 KB
bapequiv.dta	30.01.2018 03:31	DTA File	18.277 KB
bapgen.dta	30.01.2018 03:31	DTA File	3.966 KB
bapkal.dta	30.01.2018 03:31	DTA File	15.446 KB
bapluecke.dta	30.01.2018 03:31	DTA File	117 KB
bavp.dta	30.01.2018 03:31	DTA File	41 KB
bbh.dta	30.01.2018 03:31	DTA File	9.127 KB
bbhbrutto.dta	30.01.2018 03:31	DTA File	1.028 KB
bbhgen.dta	30.01.2018 03:31	DTA File	1.706 KB
bbjugend.dta	30.01.2018 03:31	DTA File	1.198 KB
bbkind.dta	30.01.2018 03:31	DTA File	1.452 KB
bbp.dta	30.01.2018 03:32	DTA File	34.277 KB
bbpbrutto.dta	30.01.2018 03:32	DTA File	2.960 KB
bbpequiv.dta	30.01.2018 03:32	DTA File	19.560 KB
bbpgen.dta	30.01.2018 03:32	DTA File	4.279 KB
bbpkal.dta	30.01.2018 03:32	DTA File	16.576 KB
bbpluecke.dta	30.01.2018 03:32	DTA File	216 KB
bbvp.dta	30.01.2018 03:32	DTA File	42 KB
bch.dta	30.01.2018 03:32	DTA File	8.902 KB
bchbrutto.dta	30.01.2018 03:32	DTA File	1.043 KB
bchgen.dta	30.01.2018 03:32	DTA File	1.675 KB
bcjugend.dta	30.01.2018 03:32	DTA File	1.223 KB

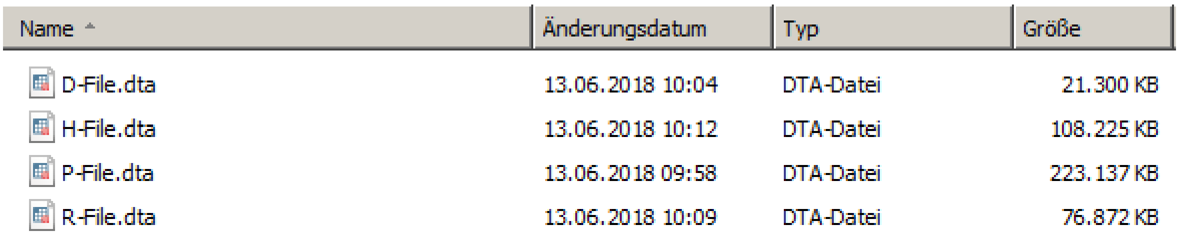
Within this “raw” directory, each wave is identified by letters of the alphabet: the first wave in 1984 is wave “A”, 1985 is wave “B”, and so on. To simplify the notation, the “$” sign is used when referring to all waves of one group of datasets. For example, $H refers to all household-level datasets from AH to now. For each year of SOEP data, there are single data files for households (e.g., $H) as well as for individual respondents (e.g., $P) and children (e.g., $KIND) based on interview information. These observations make up the “net” population, with each of these files containing as many records as interviews could be conducted. Additional data files with a limited number of variables based on the “address log” constitute the “gross” number of households and persons, i.e., all households and their members that were eligible for an interview in any given year. Within the “raw” directory, the datasets are stored on a wave-specific basis and are the basis for generating the majority of the long datasets described above. In addition to these wave-specific datasets, the “RAW” directory also contains additional datasets in cross-sectional format that have not yet been distributed in long format ($SCHOOL, $SCHOOL2, EV, EXIT, $PKALOST and PBR_HHCH).

Tracking Data	Original Data	Survey Data	Generated Data
ppfad	$p	phrf	$pgen
hpfad	$pausl	hhrf	$hgen
$pbrutto	$pluecke	pbr_hhch	$kind
$hbrutto	$h	$biorki	$pequiv
hbrutt$$	$post		$pkal
	$jugend		$pkalost
	$school		
	$school2		
	ev		
	$vp		
	biol		

5.3.3 eu-silc-like-panel

The European Union Statistics on Income and Living Conditions (EU-SILC) contains data from across Europe on individual and household income, household living conditions, individual health, aspects of child care, employment, and self-assessed financial situation. EU-SILC offers both cross-sectional and longitudinal data. The German EU-SILC dataset currently contains only cross-sectional data. The eu-silc-like-panel dataset provided at DIW Berlin offers additional longitudinal information on private households in Germany based on data from the Socio-Economic Panel (SOEP) study since 2005. The eu-silc-like-panel is included in the annual SOEP data release since 2018 and requires a data distribution contract with DIW Berlin. The SOEP data are provided free of charge for scientific research. Researchers can compare all of the information in the dataset with longitudinal data on other European countries that can be obtained from Eurostat upon request.





Name	Änderungsdatum	Typ	Größe
D-File.dta	13.06.2018 10:04	DTA-Datei	21.300 KB
H-File.dta	13.06.2018 10:12	DTA-Datei	108.225 KB
P-File.dta	13.06.2018 09:58	DTA-Datei	223.137 KB
R-File.dta	13.06.2018 10:09	DTA-Datei	76.872 KB

The eu-silc-like-panel includes all of the four EU-SILC sub-datasets: The household register (D-File), the personal register (R-File), personal data (P-File), and household data (H-File). The clone datasets can be combined using the R-File, which includes both the current household and individual identifier. The identifiers in the eu-silc-like-panel are unique and do not vary among the four datasets. Complete documentation on the datasets can be found here: Documentation EU-SILC.

Last change: May 12, 2022

5.4 Datasets SOEP-Core

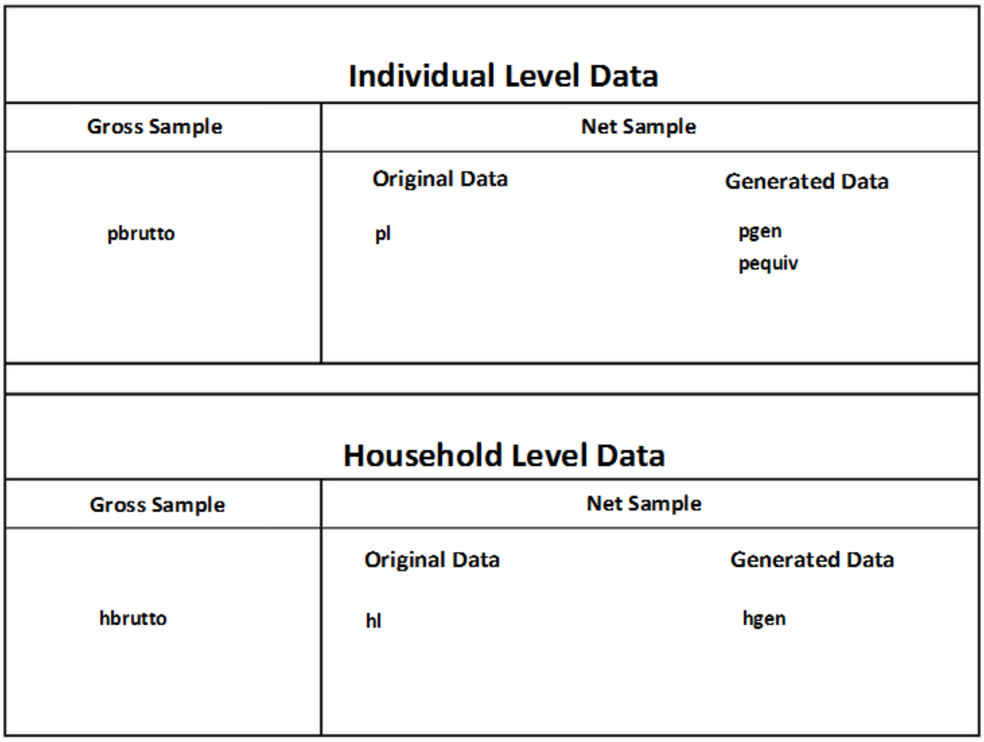
SOEP-Core contains a multitude of different datasets. An overview of the documentation for the different datasets can be found on our website, under Documentation of SOEP-Core .

To get an overview of the data types, a somewhat simplified categorization helps:

There are *Tracking Data* and *Survey Data* files which describe the development of the sample, such that the user knows which individual or household was part of the interviewed sample in any given year. Then there are *Original Data* files, which contain the data from each year's questionnaires without any changes except for very basic consistency checks. To help the user with the data, there also are *Generated Data*. These contain consistently coded variables across all waves with common names, such that the users can easily use this information when combining datasets across waves. The SOEP also provides various data on the respondent's background, called biographical data. Biography data in general can conceptually be separated into biographical data which are unchanging (such as information on parent's education, or data from the Mother-Child Questionnaires) and data which may be updated through changes in a respondent's life (such as new children in the birth biography, or a job change in the job history). Some of the changing data are stored as *Spell Data*. For each spell there is a definition of the spell type, begin point, end point and the censoring status, indicating if a given employment or income spell is censored (left and/or right) or uncensored. One of the biggest assets of the SOEP data is their longitudinal nature, i.e., repeated observations of the same unit (individual or household) over time. That's why we provide longitudinal datasets, such as PL or HL. Finally, there are some files which cannot be easily categorized - some are one-time datasets, some provide information about the interviewers, some about respondents outside of Germany.

There are two datasets which should be the building block of any analysis, as they allow users to define longitudinal populations very easily: PPATHL and HPATHL. HPATHL includes all households which have been interviewed successfully at least once. Similarly, PPATHL contains all individuals who have ever lived in a household that has participated in the SOEP, i.e., that has been captured in HPATHL, including non-respondents and children. Both data files contain one record per household or individual, respectively, with wave-specific variables for each year's survey status. In addition to some time-invariant information (like gender, year of birth, migrant status), these files contain all necessary identifiers to combine other files with PPATHL and HPATHL. Although they provide essential information, PPATHL and HPATHL alone are of little use for actual analyses. The most often used sources for additional information in SOEP-Core are the cross-sectional data files provided in each survey year (or "wave") or the datasets in the "long" Format.

The SOEP datasets can be viewed based on their content classification (Tracking Data, Original Data, Survey Data, Generated Data and Spell Data), the data structure (cross-sectional (cs), wide, long, spell) and also from the respondent's perspective. From the respondent's perspective, datasets can contain gross or net information. In addition, some datasets provide information only at the household level and others provide information at the individual level.



Individual Level Data

Gross Sample	Net Sample	
	Original Data	Generated Data
pbrutto	pl	pgen pequiv

Household Level Data

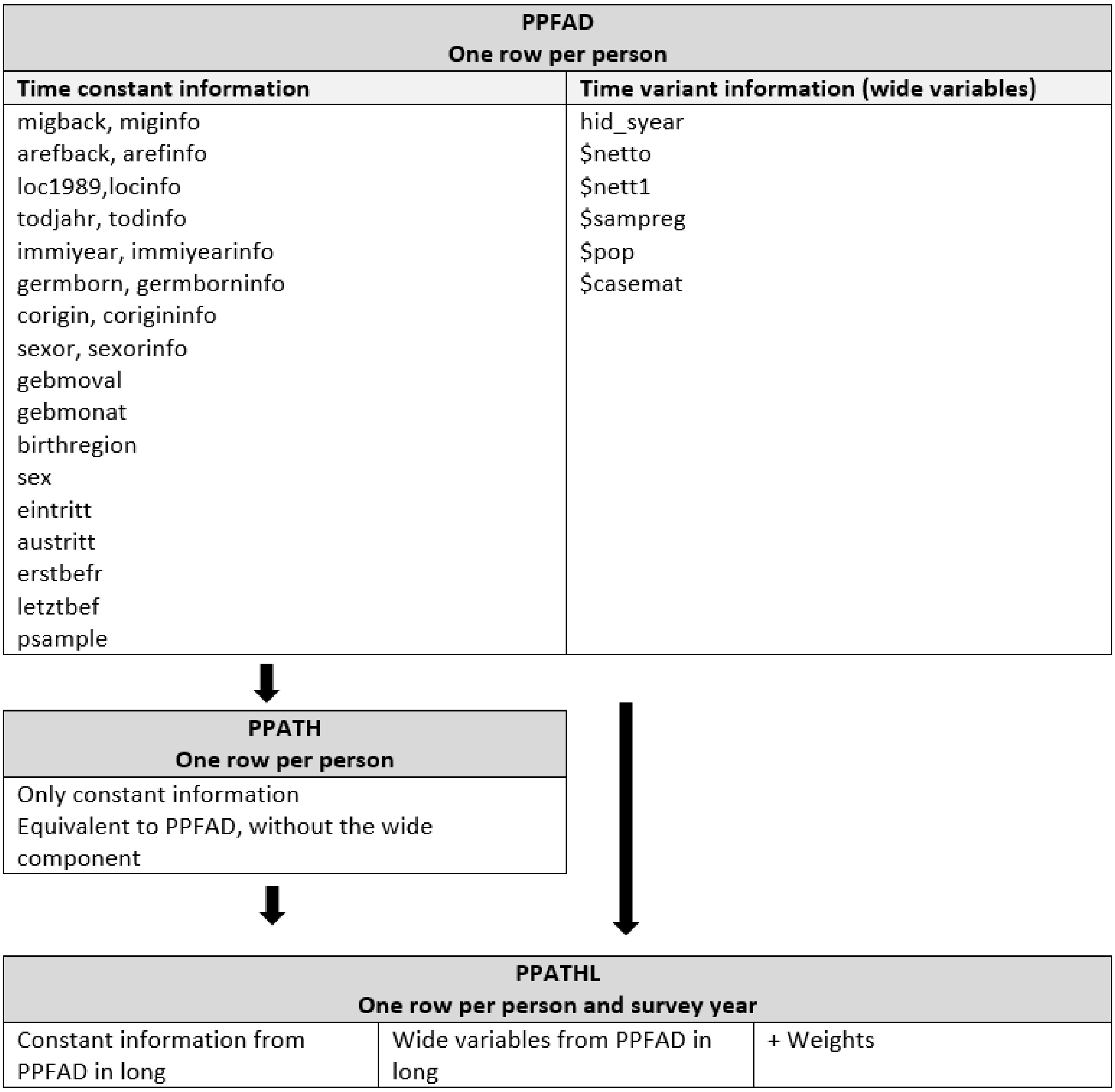
Gross Sample	Net Sample	
	Original Data	Generated Data
hbrutto	hl	hgen

Gross information at household or individual level is provided to users in the datasets hbrutto, hbrutt and pbrutto. Content information collected from household or individual questionnaires, for example, is original data and is stored in HL and PL. The SOEP team generates data from these original data, which are generated from the many SOEP questionnaires. New generated and user-friendly datasets such as pgen are created from the components of PL.

5.4.1 Tracking Data

Tracking data are the basis for linking your research-relevant variables. In addition to various demographic information, tracking data also provide information on how the interview was conducted. These datasets should be understood as initial data that you can use to merge your research-relevant variables via the individual and household numbers.

Dataset	Label	Format	Identifier (ID)	Additional Identifier
ppathl	Individual Tracking File	*long*	pid syear	hid cid parid
hpathl	Household Tracking File	*long*	hid syear	cid
pbrutto	Gross Individual Data	*long*	pid syear	hid cid intid hhnrold
hbrutto	Gross Household Data	*long*	hid syear	cid intid1 intid
hbrutt	Original gross population of a first wave sample	*long*	hid syear	cid
pbr_exit	Cumulated Exit	*long*	pid syear	hid cid hnrold
cov_brutto	Gross Household Data SOEP-COV	*long*	hid syear	tranche



PPFAD One row per person	
Time constant information	**Time variant information (wide variables)**
migback, miginfo arefback, arefinfo loc1989,locinfo todjahr, todinfo immiyear, immiyearinfo germborn, germborninfo corigin, corigininfo sexor, sexorinfo gebmoval gebmonat birthregion sex eintritt austritt erstbefr letztbef psample	hid_syear $netto $nett1 $sampreg $pop $casemat

PPATH One row per person
Only constant information Equivalent to PPFAD, without the wide component

PPATHL One row per person and survey year		
Constant information from PPFAD in long	Wide variables from PPFAD in long	+ Weights

hpathl "Household Tracking File" (long): HPATHL consists of all waves of the raw datasets HPATH and HHRF. For all years since 1984, the HPATHL dataset contains information on all households that have ever participated in the SOEP survey at any point in time. HPATHL is important for the delimitation of the unit of investigation (household), especially in longitudinal analysis. HPATHL is useful particularly for household analysis and can be used for pre-selection of specific households.

ppathl "Individual Tracking File" (long): PPATHL consists of all waves of the raw datasets PPATH and PHRF. For all years since 1984, the PPATHL dataset contains information on all individuals who have ever lived in a SOEP household at the point in time of a survey (i.e., all respondents, but also children under 17 years of age and individuals who have never given an interview). PPATHL is important for the delimitation of the units of investigation (individuals), especially for longitudinal analysis. It contains one record for each individual and year a individual has been a member of a respondent household. It is keyed on pid and syear, the survey year identifier. It contains the Household ID, the unvarying individual characteristics, individual weights, as well as the response status for that individual in each wave.

pbrutto "Gross Individual Data" (long): PBRUTTO consists of all waves of the raw datasets $PBRUTTO. PBRUTTO covers all respondents who were either interviewed for the first time or contacted for the purpose of being interviewed again in a given wave. The dataset provides gross information on all SOEP respondents' interviews as well as their positions in the panel framework.

hbrutto "Gross Household Data" (long): HBRUTTO consists of all waves of the raw datasets $HBRUTTO. HBRUTTO covers all households that were successfully interviewed for the first time in a wave or were contacted for the purpose of being interviewed again. The datasets provide gross information on all SOEP households' interviews as well as their positions in the panel framework.

hbrutt "Original gross population of a first wave sample" (long): The dataset HBRUTT contains demographic information and data on the interviews of all newly surveyed samples in the respective survey year that were successfully interviewed or contacted for the first time. All cross-sectional variables from HBRUTT$$ are used for hbrutt.

pbr_exit "Cumulated Exit" (long): The dataset pbr_exit is a supplement of pbrutto for individual dropouts. Individual dropouts are removed from the original pbrutto population, so that pbrutto covers all current household members. Pbr_exit contains the corresponding register information on individual dropouts from households.

cov_brutto "Gross Household Data SOEP-COV" (long): COV_BRUTTO contains the brutto information of SOEP-CoV study. This datset is associated with the 9 tranches of the SOEP-CoV in 2020, the SOEP-CoV wave in 2021 and COVID-19-special interviews 2020 from the IAB-BAMF-SOEP Survey of Refugees in Germany. More information about the project can be found online:

5.4.2 Original Data

These datasets contain respondents' direct information. The contents of these variables mirror the contents of the survey instruments. By searching the questionnaires, you can determine the exact wording of the question and obtain possible filter guidance.

Dataset	Label	Format	Identifier (ID)	Additional Identifier
pl	Individual questionnaire	*long*	pid, syear	hid, cid, intid
hl	Household questionnaire	*long*	hid, syear	cid, intid
biol	Biographical data	*long*	pid, syear	hid, cid, intid
jugendl	Youth questionnaire for first-time respondents at age 17	*long*	pid, syear	hid, cid, intid
more_docu	Dataset on the Mentoring of Refugees (MORE) Project	*long*	pid, syear	hid, cid
more_local	Dataset on the Mentoring of Refugees (MORE) Project	*long*	pid, syear	hid, cid
plueckel	Follow-up questionnaire	*long*	pid, syear	hid, cid, intid
abroad	Questionnaire for respondents who have moved abroad	*long*	pid, syear	hid, cid
vpl	Deceased individual	*long*	vpid, syear	hid, cid, intid
cov	SOEP-COV questionnaire	*long*	pid, syear	hid, cid, intid

pl "Individual questionnaire" (long): The PL dataset contains all waves of the $P datasets from SOEP-Core. In addition, the PL file includes all variables of all waves of the datasets $POST and $PAUSL. This means that the PL dataset contains all variables from the individual questionnaire for all waves. In addition, the individual-specific data from the IAB-SOEP Migration Survey and IAB-BAMF-SOEP Refugee Survey are integrated into the PL dataset.

Attention: For large datasets we recommend the use of Stata/MP or Stata/SE on a computer with an internal memory of 16GB. Users can still work with the data in Stata/IC or on less powerful computers, but some modifications allow users to work effectively with even the largest datasets while placing low demands on their hard- and software.

```
clear
global data = "\\hume\rdc-prod\complete\soep-core\soep.v35"

* Search all available pl variables on paneldata.org: https://paneldata.org/soep-core/
↪data/pl
describe using "$data/pl.dta"

use pid hid cid syear plh0149-plh0151 using "$data/pl.dta"
```

hl "Household questionnaire" (long): HL contains all waves of the datasets $H from SOEP-Core. This means that the HL dataset includes all questions of the household questionnaire. In addition, the household-specific data from the IAB-SOEP Migration Survey and IAB-BAMF-SOEP Refugee Survey are integrated into the original HL dataset.

biol "Biographical data" (long): BIOL contains cumulated individual-level raw data from the biographical questionnaire and from wave-specific biographical modules of the individual questionnaire. BIOL is intended to be used in addition to the generated biographical files (by advanced users) to complete (or modify) generated biographical variables.

jugendl "Youth questionnaire for first-time respondents at age 17" (long): JUGENDL contains the waves q (2000) up to the current wave of $JUGEND in SOEP-Core. Since 2000 (wave Q), first-time respondents between the age of 16 and 17 have received a separate biographical questionnaire with additional age-group-specific questions, for instance, about their relationship to their parents or about what they do in their free time.

MORE_Docu "Dataset on the Mentoring of Refugees Projekt": A dataset on the Mentoring of Refugees (MORE) project. Carried out in partnership with Start with a Friend (SWAF), this project aimed at bringing refugees and locals together to form friendships. This dataset contains information on German contacts provided to refugees. Information about the federal state of the SWAF location is includes in the EU data edition. More localized information (county or municipality) is only available remotely or on site.

MORE_Local "Dataset on the Mentoring of Refugees Projekt": A new dataset on the Mentoring of Refugees (MORE) project. Carried out in partnership with Start with a Friend (SWAF), this project aimed at bringing refugees and locals together to form friendships. This dataset contains information from the surveys of the locals in the project.

plueckel "Catch-up questionnaire" (long): The PLUECKEL dataset contains all waves of the $PLUECKE datasets in SOEP-Core. Temporary drop-outs ("gaps") can cause problems for longitudinal analyses. This has especially negative consequences for the employment and income data. That is why the SOEP tries to fill in at least some of the key missing information. PLUECKEL is a small questionnaire covering information on the year previous to which the temporary drop-out occurred. It covers questions on job-related changes, employment calendar, income, education, and qualifications.

abroad "Questionnaire for respondents who have moved abroad" (long): With the pilot study "Life outside Germany" in 2008, the longitudinal SOEP study ventured into completely uncharted methodological territory by attempting to locate the addresses of former SOEP respondents who have since moved abroad and to survey these individuals with the help of a specially developed written questionnaire on the reasons for their move. The project was discontinued due to insufficient case numbers in 2014.

vpl "Questionnaire on the deceased individual" (long): The VPL dataset contains all waves of the $VP datasets of SOEP-Core. The VPL file contains information about respondents who lost a relative in the previous year. It provides information about the deceased individual and the respondent who reported the death.

cov "SOEP-COV questionnaire" (long): COV contains the survey content of SOEP-CoV study. This datset is associated with the 9 tranches of the SOEP-CoV in 2020, the SOEP-CoV wave in 2021 and COVID-19-special interviews 2020 from the IAB-BAMF-SOEP Survey of Refugees in Germany. More information about the project can be found online:

5.4.3 Survey Data

These datasets contain information on survey methodologies used in SOEP-Core. The various datasets contain detailed exit information provided by respondents and the household weighting factors that users need for representative analysis.

Dataset	Label	Format	Identifier (ID)	Special Identifier
design	Survey design	*long (time-constant)*	cid	intid
exit	Cumulative drop-outs	*long (time-constant)*	pid	cid, syear
pbr_hhch	PBR_HHCH	*long*	pid, syear	hid, cid, pnralt, pnrneu, hhnrold
cirdef	Randomized survey file	*long (time-constant)*	cid	
cov_contact	Contact Data SOEP-COV	*long*	hid ContactDate	tranche

design "Survey design": The dataset DESIGN provides information on the stratified sampling of the SOEP in the form of two variables. The variable STRAT identifies each of the discrete sampling groups described above. Altogether, the SOEP consists of 40 strata: one stratum in sample A, twenty-seven in sample B, one in sample C, three in sample D, one in sample E, two in sample F, four in sample G, and one in sample H. Each of these strata have unique inclusion probabilities. The variable design contains the inverse of this probability, i.e., the design weight.

exit "Follow-up study [Verbleibstudie]": The dataset EXIT delivers the results from the follow-up study [Verbleibstudie] conducted by Kantar Public (formerly: TNS Infratest) in 2008/2009. This study has been used to identify reasons for (demographic) dropouts. Deceased individuals identified through the follow-up study are included in the corresponding variables in PPATH/L [todjahr, todinfo].

pbr_hhch "PBR_HHCH": The dataset pbr_hhch is a subfile of pbrutto that was used from 1984 to 2009 to identify individuals from households that underwent split-offs in subsamples A-H.

cirdef "Randomized survey file": This dataset includes randomized groups of original sample households [rgroup] for selection of representative shares across all subsamples with full representation of any cross-sectional and longitudinal information (variables) at all levels (case, households, individuals, spells) for the entire SOEP population across all waves.

cov_contact "Contact Data SOEP-COV": COV_CONTACT contains the contact information of SOEP-CoV study. This datset is associated with the 9 tranches of the SOEP-CoV in 2020, the SOEP-CoV wave in 2021 and COVID-19-special interviews 2020 from the IAB-BAMF-SOEP Survey of Refugees in Germany. More information about the project can be found online:

5.4.4 Generated Data

The SOEP team has prepared these datasets for easy use and subjects them to additional plausibility checks and quality controls prior to data release. In most cases, they consist of several variables and different survey instruments and are described in the documentation provided. As a result, these datasets cannot be assigned 1:1 to a single survey instrument.

Dataset	Label	Format	Identifier (ID)	Additional Identifier
pgen	Generated individual data	*long*	pid, syear	hid, cid, pgpartnr

continues on next page

Table 1 – continued from previous page

Dataset	Label	Format	Identifier (ID)	Additional Identifier
hgen	Generated household data	*long*	hid, syear	cid
bioagel	Generated biographical information	*long*	pid, syear, persnre	hid, cid,
biopupil	Generated biographical information	*long*	pid, syear	hid, cid
kidlong	Data on children	*long*	pid, syear	hid, cid
pequiv	Cross National Equivalent File	*long*	pid, syear	hid, cid
biobirth	Generated biographical information	*wide*	pid	cid, kidpnr01-kidpnr15
bioedu	Generated biographical information	*long (time-constant)*	pid	cid
bioimmig	Generated biographical information	*long*	pid, syear	hid, cid
biojob	Generated biographical information	*long (time-constant)*	pid	cid
bioparen	Generated biographical information	*long (time-constant)*	pid	cid, fnr, mnr
bioregion	Generated biographical information	*long (time-constant)*	pid, syear	cid
bioresidrefinG	Generated biographical information	*wide*	pid	
biosib	Generated biographical information	*wide*	pid	cid, sibpnr1-sibpnr11
biotwin	Generated biographical information	*wide*	pid	cid, pnrtwin, pnrtrip, pnrquad
camces	Highest educational qualification, migrants sample M1 and M2	*long*	pid	hid, syear, cid
cogdj	Data on cognitive tests (Youth)	*long*	pid	syear, cid
cognit	Data on cognitive potential	*long*	pid	syear, cid, intid
cog_refu	Data on cognitive tests (Refugees)	*CS*	pid	syear, cid, hid
gripstr	Grip Strength Measures	*long*	pid, syear	cid, intid
hconsum	Household Consumption Module	*CS*	hid	syear, cid
health	Data on health indicators	*long*	pid, syear	cid
hwealth	Wealth module	*long*	hid, syear	cid
interviewer	Data on the SOEP interviewer	*long*	intid, syear	cid
mihinc	Multiple imputed data on monthly household income	*long*	hid, syear	cid

continues on next page

Table 1 – continued from previous page

Dataset	Label	Format	Identifier (ID)	Additional Identifier
pflege	Persons needing care within the household	*long*	pid, syear	cid
pkal	Individual calendar	*long*	pid, syear	hid, cid
pwealth	Wealth module	*long*	pid, syear	hid
timepref	Experiment on time preferences	*CS*	pid	hid, syear, cid
trust	Experiment on trust	*long*	pid	hid, syear, cid

pgen "Generated individual data" (long): PGEN contains all waves of the $PGEN datasets in SOEP-Core. The PGEN-file contains user-friendly data on the individual level that are consolidated from different sources. The plausibility is validated longitudinally in many respects, making the data superior to those in PL in most situations. The file contains one row for each individual (pid is unique) with a completed individual or youth questionnaire.

hgen "Generated household data" (long): HGEN contains all waves of the $HGEN datasets in SOEP-Core. In order to minimize computational effort for the user, the SOEP provides yearly status variables on the household level. The HGEN data provide a set of time-invariant variables generated from the SOEP household questionnaire. They only include households that participated in the respective year.

bioagel "Generated biographical information" (long): The BIOAGEL data files are generated using information collected in the "Mother & Child" and "Parent" questionnaires. BIOAGEL is now provided in one dataset.

biopupil "Generated biographical information" (long): The BIOPUPIL data files are generated using information collected in the "Pre-Teen" and "Early-Youth" questionnaires. BIOPUPIL is provided in one dataset.

bioregion "Generated biographical information" (long): A dataset on places in Germany that are of biographical importance to respondents (place of birth, first place of residence). Information about the federal state of these important places is includes in the EU data edition. More localized information (county or municipality) is only available remotely or on site.

bioresidrefinG "Generated biographical information" (long): A dataset on refugees' place(s) of residence in Germany (Wohnorthistorie). Information about the federal state in which refugees reside is included in the EU data edition. More localized information (county or municipality) is only available remotely or on site.

kidlong "Data on children" (long): The variables stored in the KIDLONG file are based on the information collected annually and contained in the wave-specific $KIND files. The relevant information is not provided by children themselves but is obtained from answers to questions in the household questionnaire provided by the respondent within the household (usually the head of the household). This data is reaggregated at the individual level and stored as child-specific entries in the file $KIND.

pequiv "Cross-National Equivalent File" (long): PEQUIV contains all waves of the $PEQUIV datasets in SOEP-Core. The PEQUV-File is based on the Cross-National Equivalent File (CNEF) with extended income information for the SOEP. This file comprises not only the aggregated income figures from CNEF but also additional separate income components.

pkal "Individual calendar" (long): PKAL contains all waves of the $PKAL datasets in SOEP-Core. The PKAL datasets contain calendar variables from the individual questionnaire. The dataset includes the individual's employment or educational status on a monthly basis as well as the individual' income status.

biobirth "Generated biographical information" (wide): The file BIOBIRTH provides information on fertility histories of adult respondents in the SOEP. Up to 2014 (version 30, wave BD), the data were stored in two separate files: BIOBIRTH containing female fertility histories, and BIOBRTHM providing male fertility histories. Fertility histories in BIOBIRTH provide information on every woman (as well as every man with panel entry since 2001) who has ever completed at least one SOEP interview.

Outdated bioedu "Generated biographical information" (long time constant): The SOEP contains a broad range of variables on early childhood education and care, educational participation, educational degrees, and related topics. The BIOEDU dataset is designed to provide ready-made variables on educational transitions and related topics for use in longitudinal analysis.

bioimmig "Generated biographical information" (long): The variables contained in BIOIMMIG relate to foreigners in (and migrants to) Germany. Questions deal with the desire to return to the home country, the presence of relatives in the home country, reasons for coming to Germany, and conditions upon initial arrival in Germany.

Outdated biojob "Generated biographical information" (long time constant): The purpose of BIOJOB is to provide a file that offers the user convenient access to biographical information on past job activities. BIOJOB consists of generated variables as well as plain questionnaire information. Up to now, all but two variables in BIOJOB are time-invariant. Information on occupational changes and on the age at the most recent change of occupation refer to the date of the respondent's biography interview.

Outdated bioparen "Generated biographical information" (long time constant): The dataset BIOPAREN contains biographical entries on the parents' and respondent's background. The information in BIOPAREN is obtained from two sources: from proxy entries by children on their parents in the biography questionnaire and youth questionnaire, and from direct entries by parents when the respondent lives in the same household as the parents. Please note that BIOPAREN focuses on the social parent. Biological parent identifiers can be found in BIOBIRTH.

Outdated biosib "Generated biographical information" (wide): BIOSIB provides information on siblings living within SOEP households. The dataset contains the individual identifiers of all siblings in a SOEP household. It includes information on the individual sibling's sex, year of birth, number of siblings, position in birth order, and relationship between siblings.

Outdated biotwin "Generated biographical information" (wide): The file BIOTWIN contains all twins that were ever identified within the SOEP. To be classified as a twin, a individual is required to have exactly the same age as his or her sibling (year & month of birth), have a relationship to the head of the household that indicates that he or her and a second individuals are siblings, and have the same mother (as far as a pointer to the mother is available). Furthermore, it is not only twins that are recorded in the BIOTWIN dataset, but also triplets or quadruple siblings.

camces "Highest educational qualification, migrant samples M1 and M2" (CS): The CAMCES-File provides information about computer-assisted measurement and coding of educational qualifications in surveys.

cogdj "Data from cognitive tests (Youth)" (CS): In SOEP 2006, a separate questionnaire with cognitive tests for adolescents was used for the first time: "Lust auf DJ". The acronym "DJ" stands for "Denksport und Jugend" (mind sports and youth)", but it was named for its more common association with "disc jockey". The questionnaire "Lust auf DJ" was created for all respondents aged 16-17.

cognit "Data on cognitive potential" (long): In the 2006 survey year, for the first time, short cognitive tests were carried out with a subsample of the SOEP. The goal was to employ a robust set of instruments that could be administered easily by trained interviewers within just a few minutes. COGNIT06 provides the aggregated sum scores (total values for three time packages, so-called "parcels" of 30, 60 and 90 seconds).

cog_refu "Data on cognitive tests (refugees)" (CS): The dataset contains sum scores for two competence measurements (previous school knowledge and basic cognitive skills) of youths born in 2000, 2003 and 2005 surveyed in 2017.

gripstr "Measures of grip strength (left and right hand)" (long): The data on grip strength from the survey year 2012 is now included in the GRIPSTR dataset.

hconsum "HH consumption module" (CS)": We were faced with three methodological challenges in generating the final consumption data. First, due to the design of the consumption module, inconsistent answers arose between the amounts give for monthly and annual consumption. Second, there was the common problem of missing data, here in particular item nonresponse. And third, consumption data are usually blurred by heaping. For researchers who do not want their consumption variables to include changes from all steps of data preparation, the new dataset "HCONSUM" contains not only the prepared consumption variables but also flag variables providing researchers the opportunity to select individual solutions.

health "Data on health indicators" (long): Starting in 2002, the SOEP health module in the individual questionnaire has been revised and replicated at two-year intervals. In the HEALTH file, users find, for instance, the generated variables on height and weight with imputation flags and a user-friendly longitudinal checked generated variable for Body Mass Index (BMI).

hwealth "Wealth module" (long): The generated SOEP wealth data is stored in two separate data files called PWEALTH for information at the individual level and HWEALTH for correspondingly aggregated data at the household level. HWEALTH contains all information on the household level; it is purely the result of aggregating the individual-level information in PWEALTH. However, for all individuals with valid household-level information who did not respond to the individual questionnaire (partial unit non-response), imputations have been carried out and the results are included in HWEALTH.

interviewer "Data on the SOEP interviewer" (long): The SOEP aims not only to collect high-quality data on the living conditions and well-being of households, but also to provide a valuable empirical source for survey research. The INTERVIEWER file provides users with easy access to all available longitudinal information on the SOEP interviewers.

mihinc "Multiple imputed data on monthly household income" (long): The dataset MIHINC contains the complete imputation results and is available separately. To be compatible with methods for analyzing multiply imputed data, MIHINC is constructed in the "stacked" or MIM data format. It contains the following variables: HHNRAKT, SVYYEAR, MJ, MI, IHINC and IMPFLAG. Since 1995 for every survey household in all survey years, there are ten imputed values for current household income.

pflege "Persons needing care within the household" (long): Since wave B (1985), the SOEP household questionnaire includes questions on household members in need of care. In order to support individual-level analysis, this information has been restructured and is stored in the cumulative file PFLEGE.

pwealth "Wealth module" (long): For the first time in 2002, the individual questionnaire included a special module focusing on wealth. It included questions on seven different wealth components: owner-occupied property (including debt), other property (including debt), financial assets, private pensions (including life insurance and building savings contracts), business assets, tangible assets, and consumer credit. The generated SOEP wealth data are stored in two separate data files called PWEALTH for information at the individual level and HWEALTH for correspondingly aggregated data at the household level. Wealth-related variable names in the file PWEALTH consist of six digits. The first digit tells the user which wealth component is referred to, and the second to sixth digits provide more detailed information about possible filter information, the personal share, the gross amount, and the amount of any outstanding debt. In principle, a digit is coded "1" if a given variable does indeed contain this specific piece of information and "0" otherwise. The wealth information in the SOEP questionnaire is surveyed at the individual level and thus also imputed or edited at the individual level (although checked against household information for consistency).

timepref "Experiment on time preferences" (CS): Following the behavioral experiment on trust and trustworthiness carried out in the 2003, 2004, and 2005 SOEP surveys, the experiment "time preferences" was run in 2006. In this experiment on economic behavior, respondents were asked to decide how they would want to receive €200 in prize money: if they would want to receive it immediately by check or if they would want to wait and receive a larger amount later, that is, with interest.

trust "Experiment on trust" (long): The economic behavior experiment on trust and trustworthiness from survey years 2003, 2004, and 2005 served to measure trust based on an investment game, a one-off game for two players who interact anonymously. The first player receives a credit of ten points and can overwrite any number of points of the second player. Each overwritten point is doubled. The second player also receives a credit of ten points. After receiving the (doubled) points from the first player, the second player decides how much of her own credit she will transfer to the first player (zero to ten points). As with the first transfer, the recipient's points are doubled. After the decision of the second player, the game ends and the other players are paid (one point corresponds to one euro, the total is paid by check a few days later). The trust dataset thus contains the information from all three waves in which the behavioral experiment was conducted.

5.4.5 Spell Data

Spell, duration, and event history data are used frequently in the social sciences. In the strict sense of the word, spell data are about time periods with a defined start and end. General information about the data structure of spell data can be found in the chapter *Data Structure in Spell Format (spell)*

Working with spell data:

Working with spell data (pdf):

Working with spell data (do-files):

How to generate spell data from data in wide format: Based on the migration biographies in the IAB-SOEP Migration Sample:

Generating spell data:

Dataset	Label	Format	Identifier (ID)	Additional Identifier
artkalen	Spell data from the activity calendar	*spell*	pid	cid
biocouplm	Generated biographical information	*spell*	pid	cid, coupid
biocouply	Generated biographical information	*spell*	pid	cid
biomarsm	Generated biographical information	*spell*	pid	cid
biomarsy	Generated biographical information	*spell*	pid	cid
einkalen	[deprecated] Spell data on income	*spell*	pid	cid
lifespell	Spell Information on the pre- and post-survey history of SOEP respondents	*spell*	pid	cid
migspell	Migration history	*spell*	pid	cid
pbiospe	Generated biographical information	*spell*	pid	cid
refugspell	Migration history	*spell*	pid	cid
sozkalen	[deprecated] Spell data on social benefits	*spell*	hid, cid	

artkalen "Spell data from the activity calendar" (long): The ARTKALEN contains spells (monthly) for events starting in January 1983. This is in contrast to PBIOSPE, where spells were in yearly durations, and events previous to 1983 were included. The information on activity status is collected on a monthly basis in the yearly individual questionnaire and stored in the file ARTKALEN.

biocouplm "Generated biographical information" (long): With the BIOCOUPLM the SOEP provides consistent and continuous partnership histories for nearly all adult respondents. BIOCOUPLM is built on the prospective information at the time of each interview. The relationsship histories are collected on a monthly basis from all adult SOEP participants since their entry into the SOEP.

biocouply "Generated biographical information" (long): With the BIOCOUPLY, the SOEP provides consistent and continuous partnership histories for nearly all adult respondents. BIOCOUPLY is built on retrospective and prospective information at the time of each interview. The relationship histories are provided on an annual basis.

biomarsm "Generated biographical information" (long): With BIOMARSM the SOEP provides consistent and continuous marital histories for nearly all adult respondents. BIOMARSM is built on the prospective information at the time of each interview. The martial histories are collected on a monthly basis from all adult SOEP participants since their entry into the SOEP.

biomarsy "Generated biographical information" (long): With BIOMARSY the SOEP provides consistent and continuous marital histories for nearly all adult respondents. BIOMARSY is built on retrospective and prospective information at the time of each interview. The marital histories are provided on an annual basis.

`Outdated` einkalen "Spell data on income" (long) The income calendar is used to gain information about sources of income throughout the year. The respondent checks off for each month all appropriate sources of income.

lifespell "Spell information on the pre- and post-survey history of SOEP respondents" The SOEP team regularly conducts follow-up studies to relocate attritors. These studies draw on official register data and allow us to determine whether a individual is still living in Germany, is deceased, or has moved abroad since the last SOEP interview. The information is combined in a spell file LIFESPELL. This dataset reports all available information on the pre- and post-survey history of all individuals who have ever been a member of a SOEP household.

`Outdated` migspell "Migration history" (long): MIGSPELL is derived from the migration biographies, which are collected from each new respondent of the IAB-SOEP migration samples M1 and M2. It contains data on moves by foreign-born migrants as well as on stays abroad by German-born respondents.

pbiospe "Generated biographical information" (long): The spell file PBIOSPE is based on the information on activity status over the life course, which is collected as a matrix from every respondent who completes the biographical questionnaire. The observations start at the age of 15 and end at the current age (up to age 65). To update ongoing employment information in PBIOSPE, information from the yearly individual questionnaire is also used.

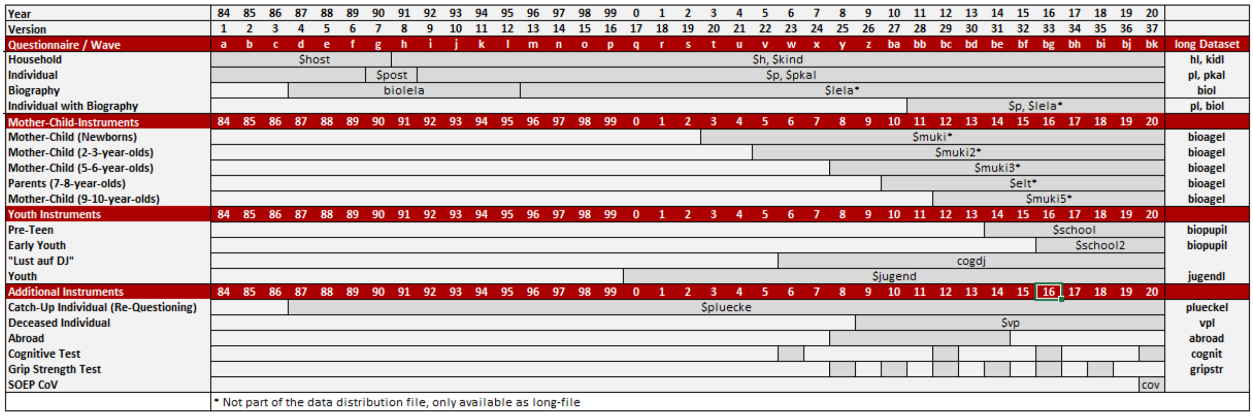
`Outdated` refugspell "Migration history" (long): For migration biographies in the refugee samples, we created the spell dataset REFUGSPELL. The variables in MIGSPELL and REFUGSPELL are derived from different instruments and only partially overlap. The data structure allows the dataset to be linked with MIGSPELL if desired.

`Outdated` sozkalen "Spell data on social benefits": The file SOZKALEN provides spell data on households receiving social assistance, defining the beginning, end, and censoring status of any period of receiving 3 different types of assistance. This file is set up using information from the calendar that is collected for the previous year (between 1992-2000). Thus, it contains information on a monthly basis.

Last change: May 12, 2022

5.5 Data Processing

The following overview shows which datasets form the basis of each questionnaire, and the respective data processing process, from the questionnaire to the wave-specific datasets, to the prepared "long" datasets. Please note that not all datasets are based on questionnaires but that many have been prepared meticulously by our staff. The table therefore does not show the full range of datasets available.



Year	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	
Questionnaire / Wave	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	ba	bb	bc	bd	be	bf	bg	bh	bi	bj	bk	long Dataset
Household	$host							$h, $kind																														hl, kidl
Individual	▨	▨	▨	▨	▨	▨	$post		$p, $pkal																													pl, pkal
Biography				biolela									$lela*																									biol
Individual with Biography																												$p, $lela*										pl, biol
Mother-Child-Instruments	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Mother-Child (Newborns)																				$muki*																		bioagel
Mother-Child (2-3-year-olds)																						$muki2*																bioagel
Mother-Child (5-6-year-olds)																									$muki3*													bioagel
Parents (7-8-year-olds)																											$elt*											bioagel
Mother-Child (9-10-year-olds)																													$muki5*									bioagel
Youth Instruments	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Pre-Teen																															$school							biopupil
Early Youth																																	$school2					biopupil
"Lust auf DJ"																							cogdj															
Youth																	$jugend																					jugendl
Additional Instruments	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Catch-Up Individual (Re-Questioning)				$pluecke																																		plueckel
Deceased Individual																										$vp												vpl
Abroad																									▨	▨	▨	▨	▨	▨	▨	▨						abroad
Cognitive Test																							▨						▨				▨				▨	cognit
Grip Strength Test																									▨		▨		▨		▨		▨		▨			gripstr
SOEP CoV																																					cov	
	* Not part of the data distribution file, only available as long-file																																					

In addition to the classic SOEP survey instruments, there are also a large number of sample-specific questionnaires whose information flows into other unlisted raw datasets (e.g. $pausl, $post, $pkalost etc.). The chapter *Sample-Specific Questionnaires* explains why such special survey instruments exist, how they become raw datasets and in which long datasets these variables can be found.

Last change: May 12, 2022

5.6 Dataset Identifiers

Because of the overall data structure with data on different observational levels, any analysis requires the combination of data using matching or merging procedures. These merging procedures need identifiers such that a combination of datasets becomes feasible. The central individual identifier across time is **pid**, which is fixed over time (and of course datasets). Since a person might change the household in which he or she lives at any point in time, yearly household identifiers called **hid** are necessary, facilitating matching depending on the dataset used. Finally, each individual (respondents as well as children) can be traced back to be a member of or a split-off from an original household from the very first wave. This household's ID, which is fixed no matter how often a person changes households over time, is called **cid**. In addition, respondents in long data can be differentiated by survey year. The **syear** variable can be used to identify a respondent's survey year. The SOEP provides additional identifiers in the various datasets in order to identify respondents and to provide further possibilities for merging datasets. A excerpt of these additional identifiers can be found here:

Please note that these are not all identifier variables. The name of the identifier variable can change depending on the dataset used.

- **parid** "Unchanging Individual identifier of Partner (PID)"
- **pgpartnr** "Individual Identifier of Partner"
- **coupid** "Couple Identifier"
- **intid** "Interviewer Identifier"
- **intid1** "Identifier of First Interviewer"
- **vpid** "Individual Identifier of Deceased Indivdiual"
- **mnr** "Individual Identifier Mother"
- **fnr** "Individual Identifier Father"
- **kidpnr01-kidpnr19** "Individual Identifier nth Child"
- **sibpnr1-sibpnr11** "Individual Identifier, nth Sibling"
- **pnrtwin** "Individual Identifier 2nd Sibling"
- **pnrtrip** "Individual Identifier 3rd Sibling"
- **pnrquad** "Individual Identifier 4th Sibling"

5.6.1 Partner Identifier

Partner identification (parid and pgpartnr)

Partner indicators (parid from ppathl and pgpartnr from pgen) have the purpose of defining couples in SOEP households and thus to make possible analyses on the dyadic level. Persons without spouse and (cohabitating) partner receive a missing code "-2" (=does not apply). The assignment of the partner ID within households is based on four sources of information: A question in the person-file, that asks (unmarried) respondents to identify their partner in the household, the household matrix reported by the head of household at the beginning of the interview (stell from pbrutto), the partnership biography in the lifehistory calendar reported by new respondents, and self-reports on marital status and life events, such as marriage, move in with partner, separation, etc. In unclear cases, due to tempo- ral non-response for instance, we also consider longitudinal information from previous and prospective waves. Moreover, parid is self-consistent between two individuals. For analyses of partner relationships, this information can be used to link all persons with their respective partners, and all information on both partners can also be stored in a common dataset. parid includes all persons that have ever participated in the SOEP. pgpartnr from the pgen dataset contains the same information, but is restricted to the pgen population and includes only persons with persons interview.

Monthly Couple Identifier (coupid)

The COUPID can only be found in the biocouplm dataset, because partnerships can change many times within a year. Multiple partnerships within a year can only be recorded correctly if the partnership is assigned to the exact month. So if a partnership for example ends in March, its joint COUPID also ends and a new one begins in June or so.

Family Identifier

Individual Identifier nth Child (kidpnr01-kidpnr19)

kidpnr01-kidpnr19 in the data set biobirth contains invariable individual identifier of biological children[nn] of the respondent (for the first child up to the 19th child), given it is identifiable in the SOEP. The sequence of children within biobirth is recorded with regards to the birth order in terms of age of the children. The order ranks from the oldest child specified under kidpnr01 to the youngest child. If the age is missing it is listed in the first record (kidpnr01), and in subsequent records following kidpnr01 if more than one child's personal identifier remains missing. kidpnr[nn] is "-1" if a child was reported in the birth biography who could not be assigned to a SOEP household (Children outside the parental household). If no child could be identified in the household context or in the birth biography, the code "-2" is assigned.

Individual Identifier Mother/Father (mnr fnr)

The personal ID of the parents (fnr and mnr) from bioparen is generated in three steps:

- The parents of the respondent are identified by the relationship to the head of the household (stell in pbrutto). Ideally, the children's parents are identified at the time of the first survey of the child. Furthermore, the social parents and not necessarily the biological parents are identified.
- The parents of the respondent are identified via the mother's ID as well as the mother's partner ID in $$kind. By using these variables the "oldest" parents are identified. Ideally, these are the parents at the time the child is 17 years old (one year before the first survey).
- The biological mother-ID and father -ID of the respondent can be identified in biobirth.

As bioparen aims at identifying the social parents that live in the household when the child is surveyed, the steps above are carried out in the hierarchy 1-3 with step 1 having the highest priority. If one is interested in only biological parents, please have a look at the information in biobirth

Individual Identifier, nth Sibling (sibpnr1-sibpnr11)

The variables provide the never changing person IDs for the siblings of the individual identified by PID. The sibling relationship is generated from the parent information in biobirth and bioparen. Two persons are defined as siblings if they report both, the same mother and father, only the same mother, or only the same father. This information on the sibling relationship is stored in sibdef1-sibdef11. In the case of inconsistent information on parents in biobirth and bioparen, bioparen was assigned the lowest priority. Please note, that bioparen uses a social definition of parenthood based on cohabitation. In contrast, biosib contains both biological (biobirth) and social siblings with a higher priority on biological relations.

Individual Identifier Twins, Triplets, Quadruplets, (pnrtwin, pnrtrip, pnrtquad)

The ids pnrtwin, pnrtrip and pnrtquad from biotwin contain all twins that were ever identified within the SOEP. pnrtwin and – in rare cases if available – pnrtrip or pnrtquad contain the individual identifier of second, and third or fourth sibling in the group. This means that every case in the data set consists of a group of twins (or triplets or quadruplets). The code “-2” is assigned to pnrtrip and/or pnrtquad if a third or fourth twin sibling doesn’t exist. PERSNR and PNRTWIN however should always contain valid codes.

Individual Identifier of Deceased Indivdiual (vpid)

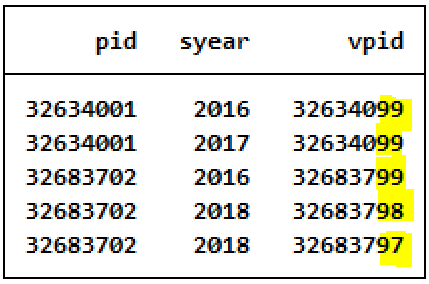
vpid in the vpl data set contains the individual identifier of a deceased indivdiual and is difficult to interpret because

- SOEP respondents in a household may provide information on several deceased persons. These deceased persons may or may not have participated in SOEP.
- Non-SOEP respondents provide information about one (or even more) deceased SOEP person(s).

So the following scenarios can occur:

	Deceased person in SOEP	Deceased person NOT in SOEP
Respondent is interviewee in SOEP	PID, VPID	PID, -/-
Respondent is NOT interviewee in SOEP	-/-, VPID	-/-, -/-

This means a person with a pid=0 has not been part of the SOEP, but has completed a deceased person questionnaire for a deceased person from a SOEP household (vpid has a SOEP ID). When individuals with SOEP ID (pid) report a deceased person who was not part of SOEP, special vpids are assigned:



```
      pid   syear       vpid

 32634001    2016   32634099
 32634001    2017   32634099
 32683702    2016   32683799
 32683702    2018   32683798
 32683702    2018   32683797
```

90s numbers in vpid are assinged to “non-Soep participants”. E.g. if a mother is deceased, but she did not participate in the Soep. In this example a No-SOEP-Person died in 2016, another No-SOEP-Person in 2017. Different 90s (e.g. 98/99) are only assigned if more than one No-SOEP-Person died in a year. When a SOEP person dies and is reported by another person, the vpid is the pid of the deceased respondent.

5.6.2 Interviewer Identifier

Interviewer ID (intid and intid1)

Intid and intid1 are fixed IDs over time to identify interviewers across years, households and questionnaires within datasets. The interviewer ID is used to identify the respective interviewer of different respondents. Unlike most other datasets in the SOEP the interviewer dataset has no PID or HID to identify the observations, but you can merge the interviewer information to other datasets using the intid. Due to changing IDs in the SOEP raw data in this and past versions of the interviewer data set it may happen that the intid of an interviewer changes over time. This can happen at most once per interviewer and is unfortunately not flagged. When the interviewer is replaced or when the interviewer changes over time, intid1 references to the first interviewer who conducted an interview in the survey household.

Last change: May 12, 2022

5.7 Versioning and Harmonization

In some cases, variables in long format with the same content but collected in different ways need to be harmonized to ensure that they remain consistent and comparable over time. Starting with SOEP Core v.34, SOEP offers versioning and harmonization solutions for such variables in all *Original Data* in long format. These versions and harmonizations are recognizable in the variable name. The "_v" suffix indicates possible differences in a variable. Harmonization suggestions generated by SOEP from the different versions of these variables can be recognized with the "_h" suffix. In general, particular caution is required when using variables marked "_v" or "_h":

1.) Differences in Response Options

Variables are versioned and harmonized because the response options have changed over time.

2.) Differences in Coding of Response Options

Variables are versioned and harmonized because the coding of the response options has changed over time. Since the values of certain response options can change, it is not possible to easily integrate the various wave-specific variables into a variable in long format. The variable must be appropriately harmonized to be useable.

3.) Content Differences in the Questions.

Variables are versioned and harmonized because the questions were asked differently in different years, but the content belongs together. If the content or wording of the question changes, the wave-specific variables cannot easily be integrated into a long variable.

4.) Changes in Question Type.

Variables are versioned and harmonized because the questions were asked differently in different years, for example, as a question with multiple response options and later as a question with a single response option. A possible multiple answer in certain years makes it difficult to easily integrate the wave-specific variables into a variable in long format.

5.) Euro Harmonization

Variables are versioned and harmonized because they are metric and were surveyed as DM amounts before the introduction of the euro. For the long version of the variable, metric variables based on different currencies in different years are harmonized as euro amounts.

6.) Differences in Metric Variables

Variables are versioned and harmonized if they contain a year and were provided in the wave-specific raw data with different numbers of digits. The years are standardized and presented in the harmonized version with four digits. In addition, possible problems with decimal digits in metric variables from the raw datasets are corrected for the long-format variable.

7.) Different Respondents

Variables are versioned and harmonized when different groups of respondents have received different survey instruments and the variables have not been integrated into the wave-specific raw datasets. Special samples or a specific filtering in the questionnaire can lead to certain groups of people receiving different questions that belong together in terms of content. Such different variables are harmonized in the long version of the variable.

A more detailed explanation of the versioning and harmonization concept can be found in the section *Working with harmonized Variables*

Last change: May 12, 2022

5.8 Missing Conventions

Survey variables might be missing, that is, lacking a valid code or value, for different reasons. In the SOEP, negative values are not valid for any variable, but are used instead to code different reasons for missing information. There are two possible origins of missing values: the respondent's answer or the survey design. In the first case, the respondent may refuse to answer or not know an answer or may report invalid values. In the second case, the interview design may exclude respondents with certain characteristics from some questions (e.g., men will never be asked if they are pregnant). The following codes are used:

Code	Label
-1	No answer / don't know
-2	Does not apply
-3	Implausible value
-4	Inadmissable multiple response
-5	Not included in this version of the questionnaire
-6	Version of questionnaire with modified filtering
-7	Only available in less restricted edition
-8	Question not part of the survey program this year[1]

[1] *Only applicable to datasets in long format.*

A person might decline to answer a question. This occurs mainly with sensitive questions (e.g., income-related questions) and when respondents simply do not know the answer. In such cases, the missing code is "-1" for "no answer / don't know". Note that the SOEP does not distinguish between a refusal to answer and a true "don't know". Information may be missing when a question is not asked because it is not relevant to a specific person, e.g., owner-occupiers will not be asked about the amount of rent they pay. In such cases, the question "does not apply" to this person, and the variable receives a code of "-2". Sometimes invalid answers occur when respondents fill out a PAPI interview themselves or the interviewer mistypes an answer (e.g., working hours over 168 per week). In such cases, multiple checks are carried out, and if the inconsistency remains, the variable is recoded "-3 Implausible value". Some questions contain multiple answer possibilities and respondents are asked to pick one answer. In the SOEP PAPI questionnaires, respondents sometimes ignore this request and give more than one answer (e.g., "very good" and "good" when asked about their current health status). In such cases, if the correct answer cannot be determined from the questionnaire itself, the code "-4 Invalid Multiple Answers" is assigned to this variable. With the extensions to the SOEP in recent years, entirely new samples have been added to SOEP-core. In these samples, questions are sometimes left out completely, e.g., to shorten the questionnaire or because the focus of the sample is different (as is the case with SOEP-related studies). In such cases, the variable will be set to "-5 Not included in this version of the questionnaire" for an entire subsample. With the use of CAPI, recent developments include an "integrated" individual questionnaire, i.e., the biography part and the "regular" part of the questionnaire are combined into one questionnaire. Some of the questions in the biography part are repeated in the regular part. Whereas the respondent will answer the same question twice the PAPI mode, the CAPI allows the respondent to filter around the question if it has already been asked. These cases are very rare, but if they occur, they receive a code "-6 Version of questionnaire with modified filtering". SOEP-Core offers a variety of different *editions* of the data. Due to data protection regulations, some variables of these *editions* may not be made accessible. Variables with increased restrictions are for example variables that provide federal state level information.

Because the variable may not be made accessible in a specific edition the federal state level information still remains in the data but they are assigned a missingcode “-7 Only available in less restricted edition”.

Last change: May 12, 2022

CHAPTER

SIX

WORKING WITH SOEP DATA

The following exercises are taken from our SOEP Campus Workshops, a service especially for young scholars in the disciplines of sociology, economics, and psychology. Here we provide introductions to the use of the SOEP data.

In order to familiarize yourself with the SOEP data as best as possible as a new user, you should first familiarize yourself with the tracking data.

Tracking data are the basis for linking your research-relevant variables. In addition to various demographic information, tracking data also provide information on how the interview was conducted. These datasets should be understood as initial data that you can use to merge your research-relevant variables via the individual and household numbers.

6.1 Working with Tracking Data (PPATHL)

For all years since 1984, the PPATHL dataset contains information on all persons who have ever lived in a SOEP household when a survey was conducted (i.e., all adult respondents as well as children under 17 years of age and household members who have never given an interview). PPATHL is important in distinguishing research units (persons), especially for longitudinal analysis.

Time-constant information on individuals:

- Never Changing Person ID (pid)
- ID Household (hid)
- Gender (sex), year of birth (gebjahr), year of death if applicable (todjahr)
- Migration Background (migback)
- Sample Member (psample)
- Year Moved to Germany (immiyear)
- Country Born In (corigin)

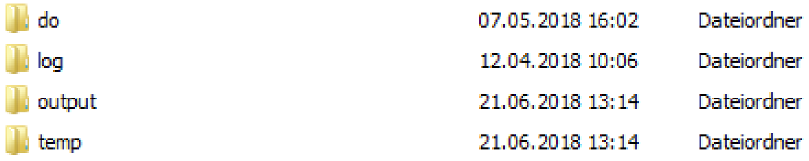
Time-varying information from individuals:

- Survey Year (syear)
- Survey Status (netto)
- Sample Membership (pop)
- Survey Region in 1989 (East or West Germany) (loc1989)

The dataset is explained in more detail in the following documentation:

Dokumentation PPATHL:

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/log
- H:/material/exercises/output
- H:/material/exercises/temp

These are used to store your script, log files, datasets and temporary datasets. Open an empty do-file and define your paths with globals:

```
***************************************************
* Set relative paths to the working directory
***************************************************
global AVZ         "H:/material/exercises"
global MY_IN_PATH "//hume/rdc-prod/distribution/soep-core/soep.v37/eu/Stata/"
global MY_DO_FILES "$AVZ/do/"
global MY_LOG_OUT "$AVZ/log/"
global MY_OUT_DATA "$AVZ/output/"
global MY_OUT_TEMP "$AVZ/temp/"
```

> **Attention:** Please note that until version 33 (v33), PPATH was called PPFAD. The following exercises are done with version 37 (v37).

The global „AVZ“ defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to the data you ordered.

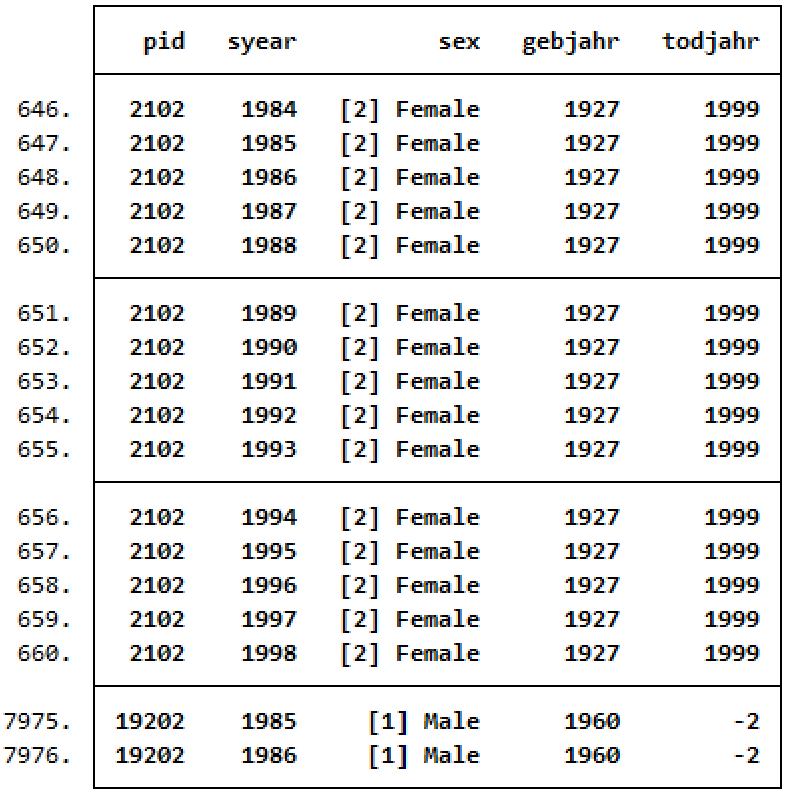
Based on the data in PPATHL, answer the following questions:

1. Look at the two people with the Person IDs (pid) 2102 and 19202

a) What is their gender? When were they born and when (if applicable) did they die?

Open the PPATHL dataset. Search the dataset for variables that describe survey year, sex, year of birth and year of death. Display the information from the variables for individuals 2102 and 19202.

```
use "${MY_IN_PATH}ppathl.dta", clear
list pid syear sex gebjahr todjahr if pid == 2102 | pid == 19202
```



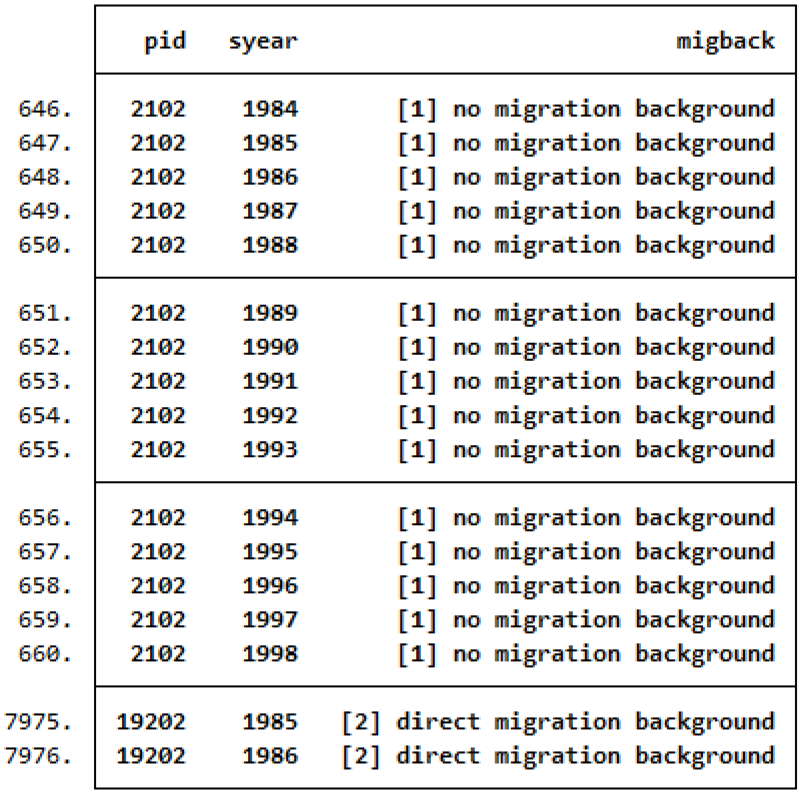
```
        +-------------------------------------------------+
        |   pid   syear          sex   gebjahr   todjahr  |
        |-------------------------------------------------|
  646.  |  2102    1984   [2] Female      1927      1999  |
  647.  |  2102    1985   [2] Female      1927      1999  |
  648.  |  2102    1986   [2] Female      1927      1999  |
  649.  |  2102    1987   [2] Female      1927      1999  |
  650.  |  2102    1988   [2] Female      1927      1999  |
        |-------------------------------------------------|
  651.  |  2102    1989   [2] Female      1927      1999  |
  652.  |  2102    1990   [2] Female      1927      1999  |
  653.  |  2102    1991   [2] Female      1927      1999  |
  654.  |  2102    1992   [2] Female      1927      1999  |
  655.  |  2102    1993   [2] Female      1927      1999  |
        |-------------------------------------------------|
  656.  |  2102    1994   [2] Female      1927      1999  |
  657.  |  2102    1995   [2] Female      1927      1999  |
  658.  |  2102    1996   [2] Female      1927      1999  |
  659.  |  2102    1997   [2] Female      1927      1999  |
  660.  |  2102    1998   [2] Female      1927      1999  |
        |-------------------------------------------------|
 7975.  | 19202    1985     [1] Male      1960        -2  |
 7976.  | 19202    1986     [1] Male      1960        -2  |
        +-------------------------------------------------+
```

Individual 2102 is female, was born in 1927 and died in 1999. She has participated annually since 1984 until 1998. Individual 19202 is male, was born in 1960 and participated twice, in 1985 and 1986. The value "-2" for the variable year of death (todjahr) stands for "Does not apply". For more information on the values, see the Missing Conventions.

b) Were these people and their parents born in Germany?

In the dataset, search for a variable that describes the migration background and the survey year. Display the information from the variables for indivIduals 2102 and 19202.

```
list pid syear migback if pid == 2102 | pid == 19202
```



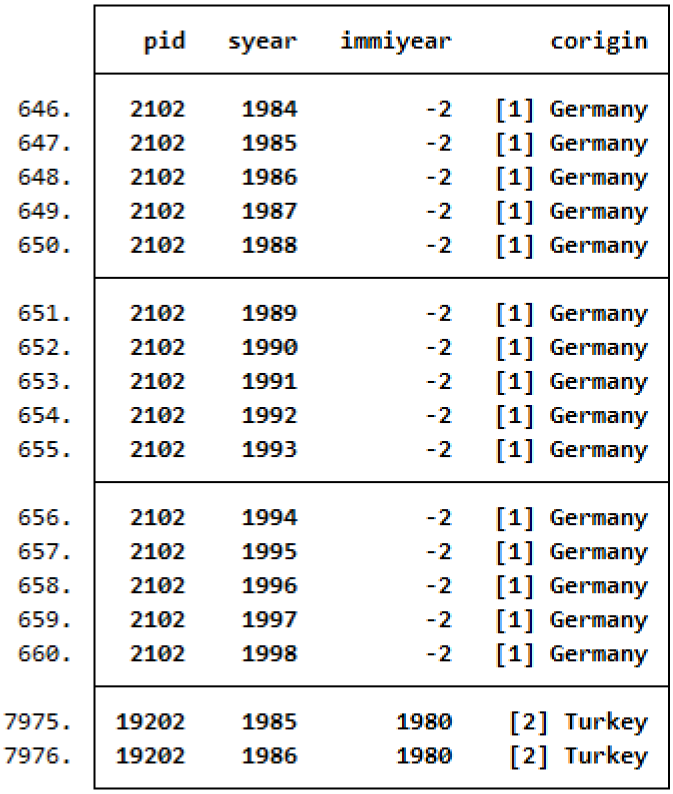
```
        pid   syear                          migback

 646.   2102    1984        [1] no migration background
 647.   2102    1985        [1] no migration background
 648.   2102    1986        [1] no migration background
 649.   2102    1987        [1] no migration background
 650.   2102    1988        [1] no migration background

 651.   2102    1989        [1] no migration background
 652.   2102    1990        [1] no migration background
 653.   2102    1991        [1] no migration background
 654.   2102    1992        [1] no migration background
 655.   2102    1993        [1] no migration background

 656.   2102    1994        [1] no migration background
 657.   2102    1995        [1] no migration background
 658.   2102    1996        [1] no migration background
 659.   2102    1997        [1] no migration background
 660.   2102    1998        [1] no migration background

7975.  19202    1985    [2] direct migration background
7976.  19202    1986    [2] direct migration background
```

Individual 2102 has no migration background. Individual 19202 has a direct migration background which means that he was not born in Germany.

c) If they immigrated to Germany, in which year and from what country?

Search the dataset for a variable that describes the country of birth, the year of moving to Germany and the survey year. Display the information from the variables on individuals 2102 and 19202.

```
list pid syear immiyear corigin if pid == 2102 | pid == 19202
```



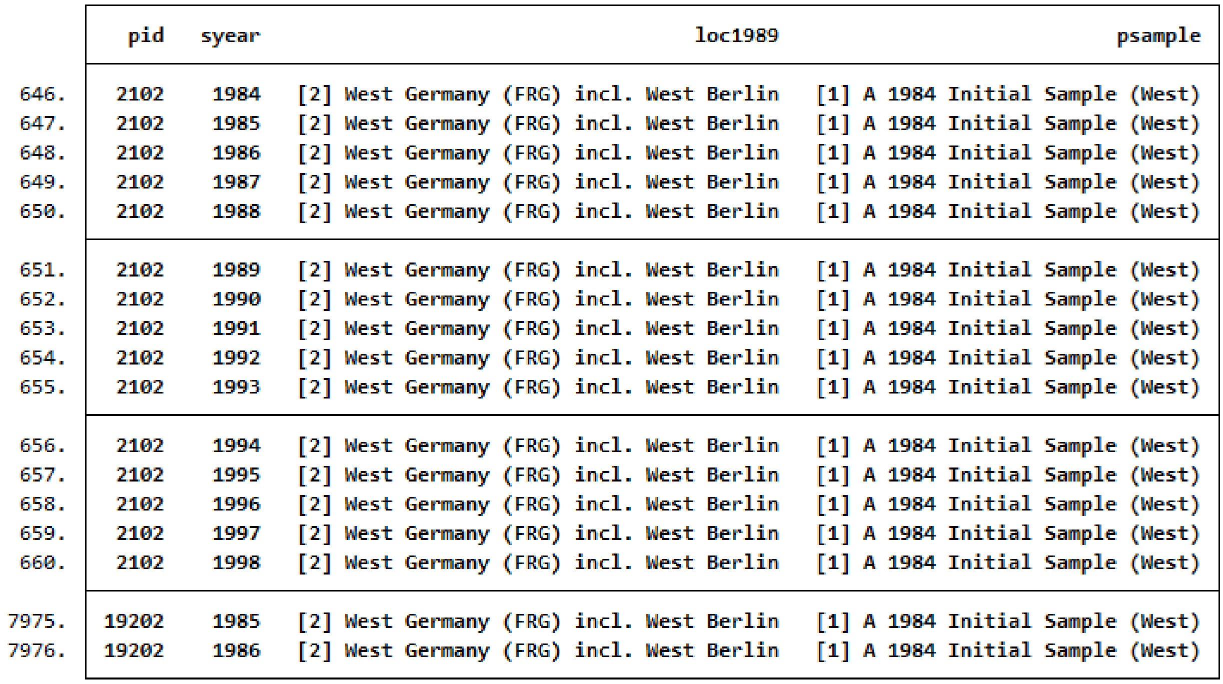
```
      +----------------------------------------+
      |   pid   syear   immiyear       corigin |
      |----------------------------------------|
646.  |  2102    1984         -2   [1] Germany |
647.  |  2102    1985         -2   [1] Germany |
648.  |  2102    1986         -2   [1] Germany |
649.  |  2102    1987         -2   [1] Germany |
650.  |  2102    1988         -2   [1] Germany |
      |----------------------------------------|
651.  |  2102    1989         -2   [1] Germany |
652.  |  2102    1990         -2   [1] Germany |
653.  |  2102    1991         -2   [1] Germany |
654.  |  2102    1992         -2   [1] Germany |
655.  |  2102    1993         -2   [1] Germany |
      |----------------------------------------|
656.  |  2102    1994         -2   [1] Germany |
657.  |  2102    1995         -2   [1] Germany |
658.  |  2102    1996         -2   [1] Germany |
659.  |  2102    1997         -2   [1] Germany |
660.  |  2102    1998         -2   [1] Germany |
      |----------------------------------------|
7975. | 19202    1985       1980    [2] Turkey |
7976. | 19202    1986       1980    [2] Turkey |
      +----------------------------------------+
```

Individual 2102 is born in Germany and has therefore no immigration year. Individual 19202 immigrated from Turkey in 1980.

d) Are these people from East or West Germany?

Search the dataset for a variable that tells whether respondents are from the East or West, the survey year and sample. Display the information from the variables for individuals 2102 and 19202.

```
list pid syear loc1989 psample if pid == 2102 | pid == 19202
```



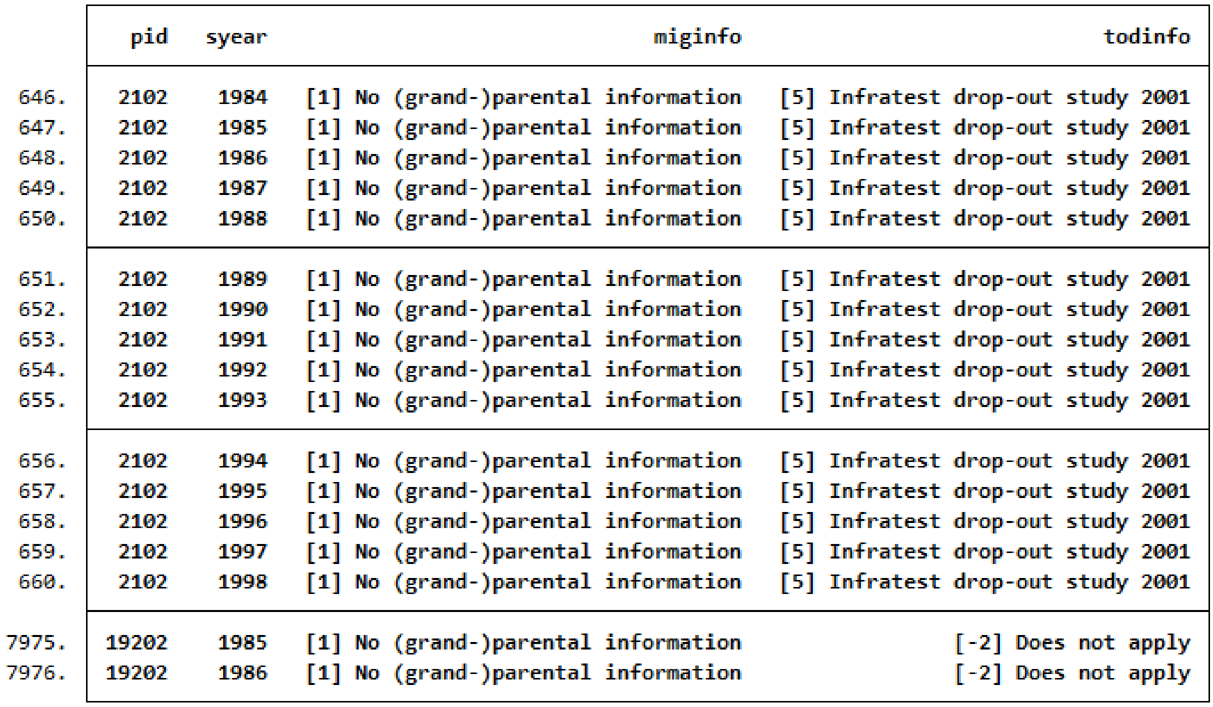
```
         pid   syear                             loc1989                        psample
  
  646.   2102   1984   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  647.   2102   1985   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  648.   2102   1986   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  649.   2102   1987   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  650.   2102   1988   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  
  651.   2102   1989   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  652.   2102   1990   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  653.   2102   1991   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  654.   2102   1992   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  655.   2102   1993   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  
  656.   2102   1994   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  657.   2102   1995   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  658.   2102   1996   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  659.   2102   1997   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  660.   2102   1998   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
  
 7975.  19202   1985   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
 7976.  19202   1986   [2] West Germany (FRG) incl. West Berlin   [1] A 1984 Initial Sample (West)
```

The variable loc1989 shows where the individual lived in 1989. Individuals 2102 and 19202 lived in West Germany in 1989 and, accordingly, were from Sample A (West).

e) What sources provide the information on the migration background and year of death

Search the data set for variables that give you the sources of information for year of death and migration background. The variable miginfo contains the information about the usage of (grand-)parents' migration history in the SOEP. The variable todinfo gives the source of the information for all persons who have been identified as deceased over the course of SOEP. Display the information from the variables for individuals 2102 and 19202.

```
list pid syear miginfo todinfo if pid == 2102 | pid == 19202
```



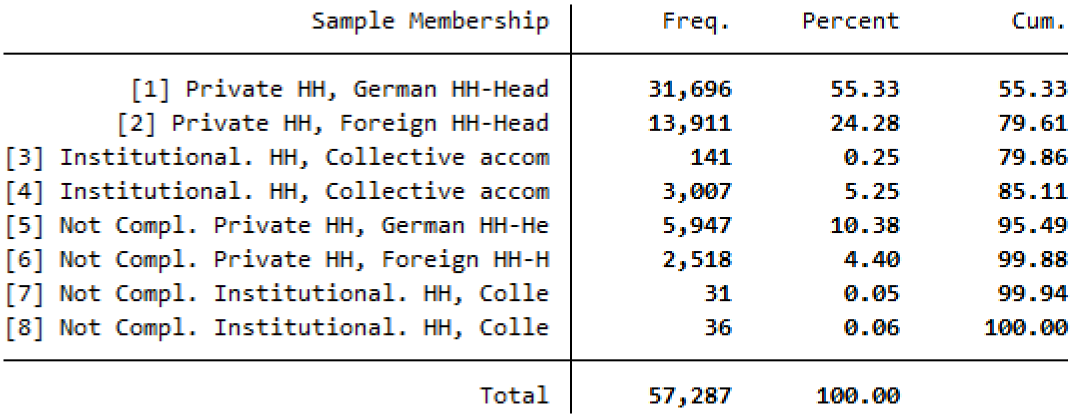
```
       +--------------------------------------------------------------------------------------+
       |   pid   syear                          miginfo                              todinfo |
       |--------------------------------------------------------------------------------------|
 646.  |  2102    1984   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 647.  |  2102    1985   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 648.  |  2102    1986   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 649.  |  2102    1987   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 650.  |  2102    1988   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
       |--------------------------------------------------------------------------------------|
 651.  |  2102    1989   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 652.  |  2102    1990   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 653.  |  2102    1991   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 654.  |  2102    1992   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 655.  |  2102    1993   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
       |--------------------------------------------------------------------------------------|
 656.  |  2102    1994   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 657.  |  2102    1995   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 658.  |  2102    1996   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 659.  |  2102    1997   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
 660.  |  2102    1998   [1] No (grand-)parental information   [5] Infratest drop-out study 2001 |
       |--------------------------------------------------------------------------------------|
7975.  | 19202    1985   [1] No (grand-)parental information                  [-2] Does not apply |
7976.  | 19202    1986   [1] No (grand-)parental information                  [-2] Does not apply |
       +--------------------------------------------------------------------------------------+
```

The information on the migration background for both individulas come from the respondents themselves. No further indicators are provided. For individual 2102, the information for the year of death comes from an Infratest Follow-Up Study of drop-outs in 2001. For individual 19202 the year of death is not provided.

2. How many people lived in a private household that was interviewed in 2016 and completed the individual questionnaire?

Search the dataset for variable that describe the population in the 2016 survey year. Display the characteristics of the population variable.

```
tab pop if syear==2016
```

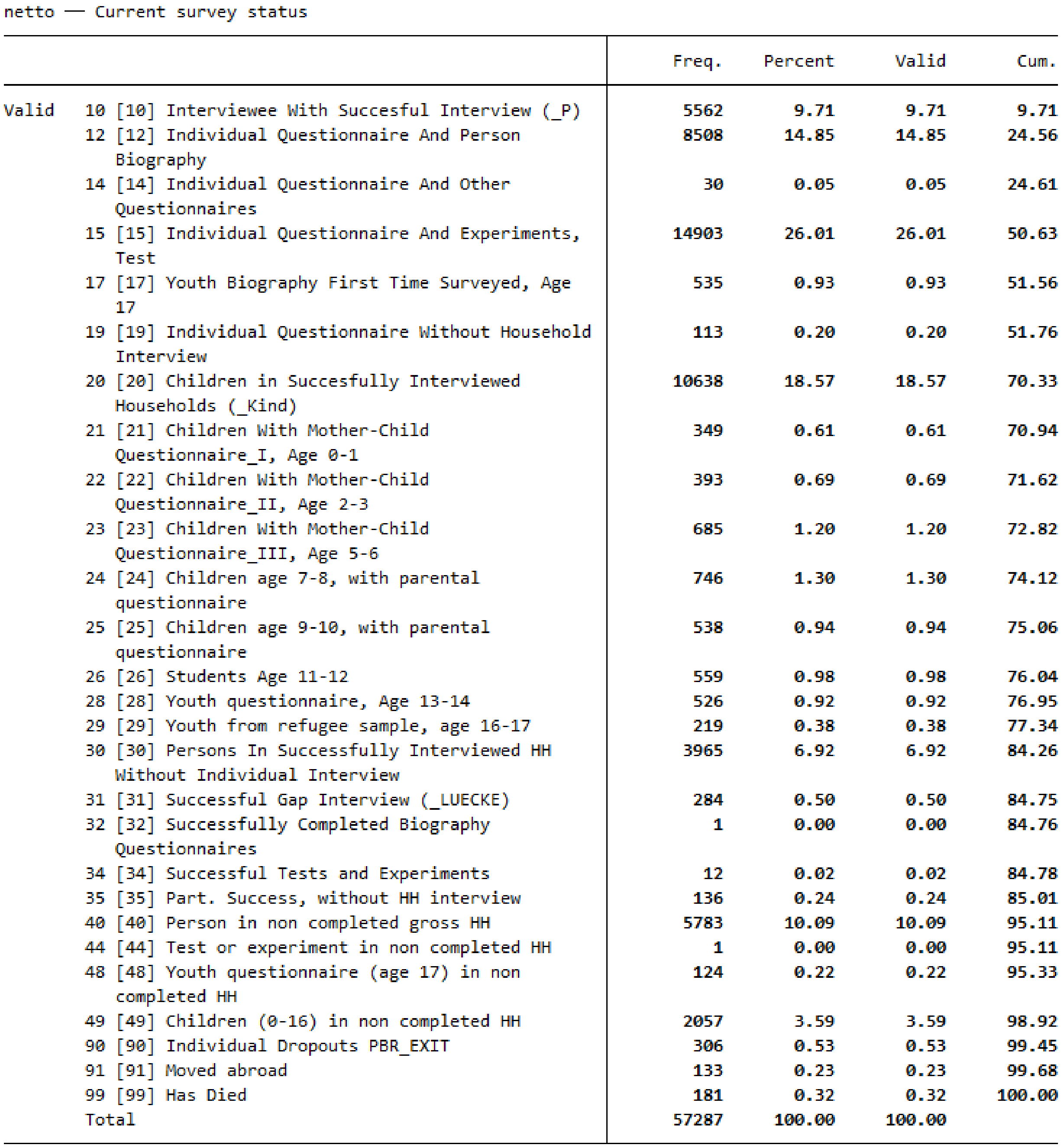


```
                 Sample Membership |      Freq.     Percent        Cum.
-----------------------------------+-----------------------------------
          [1] Private HH, German HH-Head |     31,696       55.33       55.33
         [2] Private HH, Foreign HH-Head |     13,911       24.28       79.61
[3] Institutional. HH, Collective accom |        141        0.25       79.86
[4] Institutional. HH, Collective accom |      3,007        5.25       85.11
[5] Not Compl. Private HH, German HH-He |      5,947       10.38       95.49
[6] Not Compl. Private HH, Foreign HH-H |      2,518        4.40       99.88
[7] Not Compl. Institutional. HH, Colle |         31        0.05       99.94
[8] Not Compl. Institutional. HH, Colle |         36        0.06      100.00
-----------------------------------+-----------------------------------
                             Total |     57,287      100.00
```

Values 1 and 2 are relevant to answer the question because they describe private households with completed interview.

Search the dataset for variable that describe the survey status in the 2016 survey year. Display the characteristics of the survey status:

```
fre netto if syear==2016
```

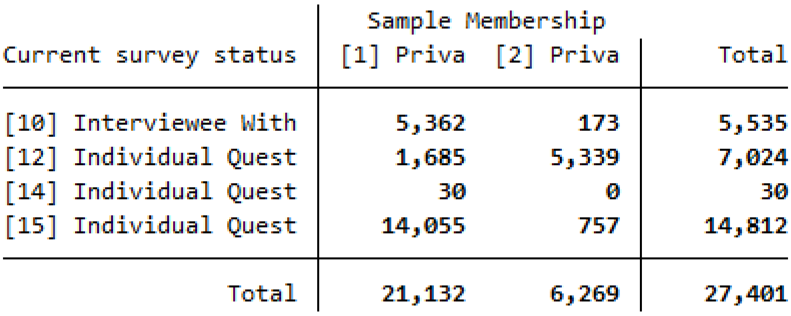


```
netto — Current survey status

                                                               Freq.   Percent     Valid      Cum.

Valid   10 [10] Interviewee With Succesful Interview (_P)      5562      9.71      9.71      9.71
        12 [12] Individual Questionnaire And Person            8508     14.85     14.85     24.56
           Biography
        14 [14] Individual Questionnaire And Other               30      0.05      0.05     24.61
           Questionnaires
        15 [15] Individual Questionnaire And Experiments,     14903     26.01     26.01     50.63
           Test
        17 [17] Youth Biography First Time Surveyed, Age        535      0.93      0.93     51.56
           17
        19 [19] Individual Questionnaire Without Household      113      0.20      0.20     51.76
           Interview
        20 [20] Children in Succesfully Interviewed           10638     18.57     18.57     70.33
           Households (_Kind)
        21 [21] Children With Mother-Child                      349      0.61      0.61     70.94
           Questionnaire_I, Age 0-1
        22 [22] Children With Mother-Child                      393      0.69      0.69     71.62
           Questionnaire_II, Age 2-3
        23 [23] Children With Mother-Child                      685      1.20      1.20     72.82
           Questionnaire_III, Age 5-6
        24 [24] Children age 7-8, with parental                 746      1.30      1.30     74.12
           questionnaire
        25 [25] Children age 9-10, with parental                538      0.94      0.94     75.06
           questionnaire
        26 [26] Students Age 11-12                              559      0.98      0.98     76.04
        28 [28] Youth questionnaire, Age 13-14                  526      0.92      0.92     76.95
        29 [29] Youth from refugee sample, age 16-17            219      0.38      0.38     77.34
        30 [30] Persons In Successfully Interviewed HH         3965      6.92      6.92     84.26
           Without Individual Interview
        31 [31] Successful Gap Interview (_LUECKE)              284      0.50      0.50     84.75
        32 [32] Successfully Completed Biography                  1      0.00      0.00     84.76
           Questionnaires
        34 [34] Successful Tests and Experiments                 12      0.02      0.02     84.78
        35 [35] Part. Success, without HH interview             136      0.24      0.24     85.01
        40 [40] Person in non completed gross HH               5783     10.09     10.09     95.11
        44 [44] Test or experiment in non completed HH            1      0.00      0.00     95.11
        48 [48] Youth questionnaire (age 17) in non             124      0.22      0.22     95.33
           completed HH
        49 [49] Children (0-16) in non completed HH            2057      3.59      3.59     98.92
        90 [90] Individual Dropouts PBR_EXIT                    306      0.53      0.53     99.45
        91 [91] Moved abroad                                    133      0.23      0.23     99.68
        99 [99] Has Died                                        181      0.32      0.32    100.00
        Total                                                 57287    100.00    100.00
```

Respondents with survey status between 10 and 15 or survey status 19 completed the individual questionnaire. These are all individuals 18 years and older.

Cross-tabulate the variables netto and pop with an appropriate restricting condition to answer the question.

```
tab netto pop if ((netto>=10 & netto<=15) | netto==19) & (pop==1 | pop==2) & (syear==2016)
```



```
                      |   Sample Membership
Current survey status | [1] Priva  [2] Priva |     Total
----------------------+----------------------+----------
[10] Interviewee With |     5,362        173 |     5,535 
[12] Individual Quest |     1,685      5,339 |     7,024 
[14] Individual Quest |        30          0 |        30 
[15] Individual Quest |    14,055        757 |    14,812 
----------------------+----------------------+----------
                Total |    21,132      6,269 |    27,401 
```

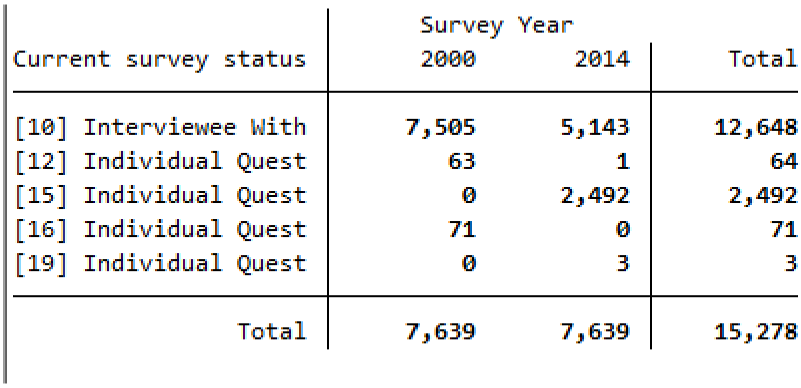
In 2016, a total of 27,401 respondents completed the individual questionnaire for Sample Membership 1 and 2.

3. PPATHL allows you to see which populations can be viewed from a longitudinal perspective:

a) How many people who answered the individual questionnaire in 2000 also took part in the survey in 2014?

Generate a variable and limit the survey status to individuals who answered an individual questionnaire in 2000 and 2014. Note that the values 10,12,13,14,15,16,18,19 of the netto variable mean realized interviews. Sort the dataset by the variable pid and generate a second variable to calculate the sum of the first generated variable. Display the characteristics of the survey status under the condition that the individual questionnaire has been answered.

```
gen v1 = 1 if (netto>=10 & netto<=19  & syear==2000) | (netto>=10 & netto<=19  & syear==2014)
bysort pid : egen v2 = sum(v1) if netto>=10 & netto<=19
tab netto syear if v2==2 & (syear==2014| syear==2000)
```



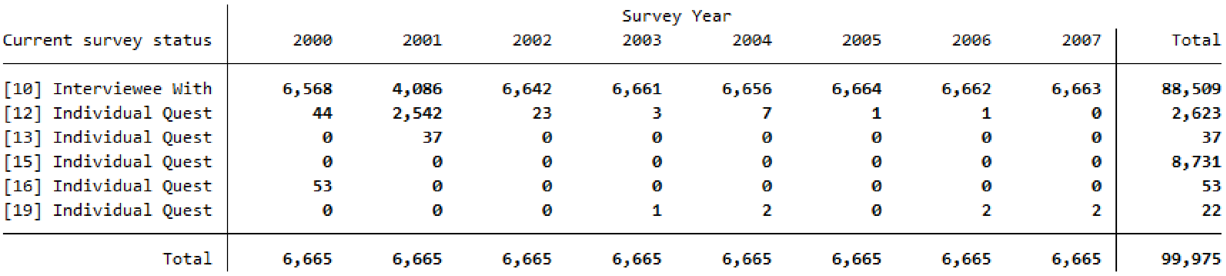
```
                      |      Survey Year
Current survey status |      2000       2014 |     Total
----------------------+----------------------+----------
[10] Interviewee With |     7,505      5,143 |    12,648 
[12] Individual Quest |        63          1 |        64 
[15] Individual Quest |         0      2,492 |     2,492 
[16] Individual Quest |        71          0 |        71 
[19] Individual Quest |         0          3 |         3 
----------------------+----------------------+----------
                Total |     7,639      7,639 |    15,278 
```

A total of 7,639 respondents completed the individual questionnaire in 2000 and 2014.

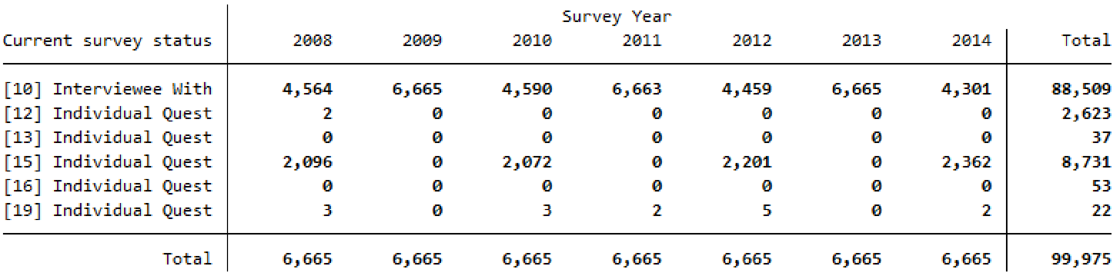
b) How many people answered the individual questionnaire every year from 2000 to 2014?

Generate a variable that counts the number of waves of completed individual interviews and limit it to the years 2000 until 2014. If the generated variable takes the value 15, a person has completed a personal interview 15 years in a row. Display the survey status and the survey year with the newly created variable.

```
egen h1 = count(syear) if netto>=10 & netto<=19 & syear>=2000 & syear<=2014, by(pid)
tab netto syear if h1==15 & syear>=2000 & syear<=2014
```



```
                                                     Survey Year
Current survey status       2000       2001       2002       2003       2004       2005       2006       2007       Total

[10] Interviewee With      6,568      4,086      6,642      6,661      6,656      6,664      6,662      6,663      88,509
[12] Individual Quest         44      2,542         23          3          7          1          1          0       2,623
[13] Individual Quest          0         37          0          0          0          0          0          0          37
[15] Individual Quest          0          0          0          0          0          0          0          0       8,731
[16] Individual Quest         53          0          0          0          0          0          0          0          53
[19] Individual Quest          0          0          0          1          2          0          2          2          22

                Total      6,665      6,665      6,665      6,665      6,665      6,665      6,665      6,665      99,975
```



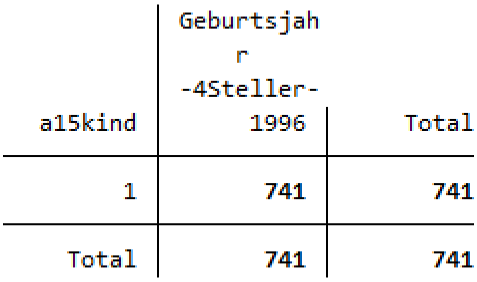
```
                                                Survey Year
Current survey status       2008       2009       2010       2011       2012       2013       2014       Total

[10] Interviewee With      4,564      6,665      4,590      6,663      4,459      6,665      4,301      88,509
[12] Individual Quest          2          0          0          0          0          0          0       2,623
[13] Individual Quest          0          0          0          0          0          0          0          37
[15] Individual Quest      2,096          0      2,072          0      2,201          0      2,362       8,731
[16] Individual Quest          0          0          0          0          0          0          0          53
[19] Individual Quest          3          0          3          2          5          0          2          22

                Total      6,665      6,665      6,665      6,665      6,665      6,665      6,665      99,975
```

A total of 6,665 people completed the individual questionnaire every year from 2000-2014.

c) How many people who turned 15 in 2011 and spent at least part of their childhood in a SOEP household took part in the survey in 2016?

Generate a variable with people who turned 15 in 2011 and had lived in a survey household as a child. The age of the respondent can be determined with the year of birth, and you can limit children using the net code. Display the new generated variable and the year of birth.

```
gen a15kind = 1 if 2011-gebjahr==15 & netto>=20 & netto<30 & syear==2011
tab a15kind gebjahr
```

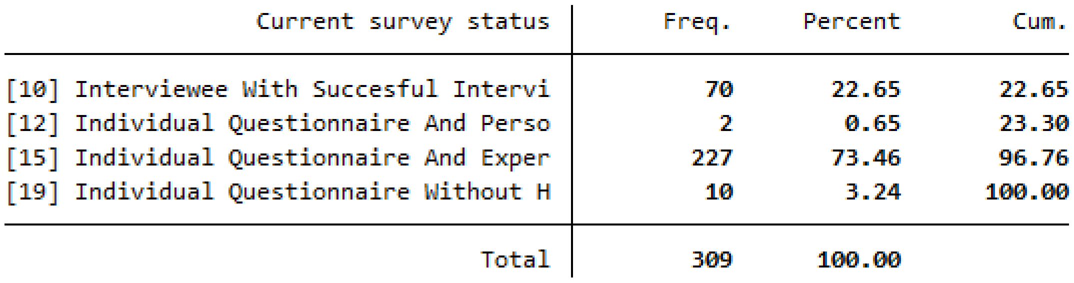


```
           Geburtsjah
               r
           -4Steller-
   a15kind       1996       Total

         1        741         741

     Total        741         741
```

A total of 741 people were 15 years old in 2011 and lived as children in a survey household.

To find out if these 741 people filled out a person questionnaire in 2016, we generate a second variable that fills up the value of one person for all remaining available years. Limit the net code and survey year to narrow down the cases appropriately.

```
bysort pid : egen a1 = max(a15kind)
tab netto if a1==1 & netto>=10 & netto<20 & syear==2016
```



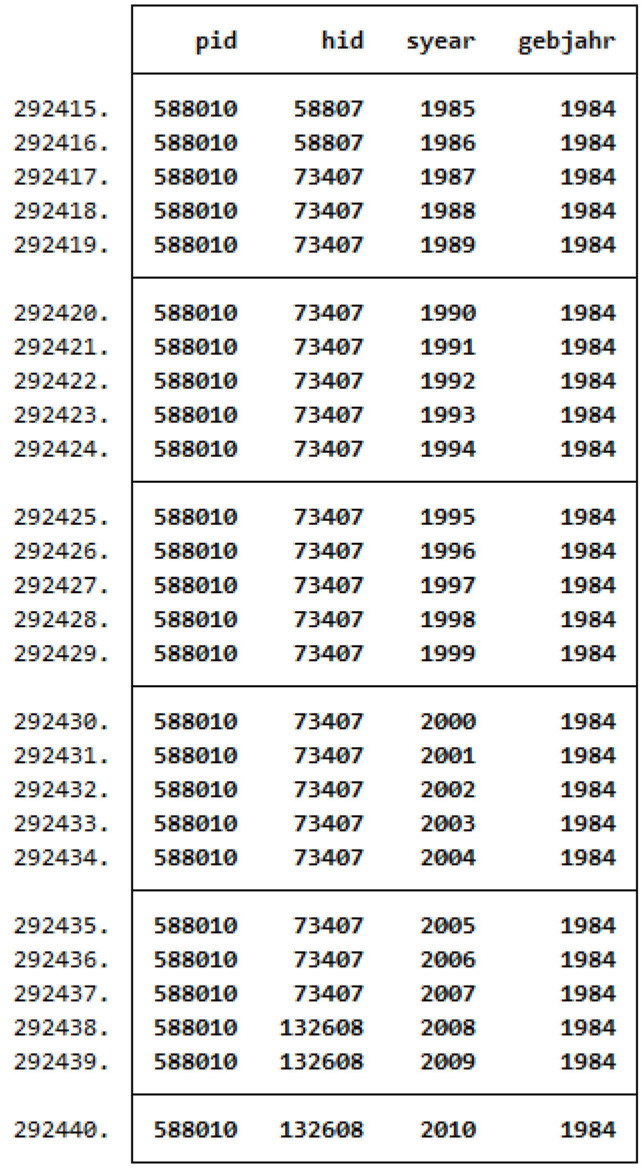
```
              Current survey status |      Freq.     Percent        Cum.
------------------------------------+-----------------------------------
[10] Interviewee With Succesful Intervi |         70       22.65       22.65
[12] Individual Questionnaire And Perso |          2        0.65       23.30
[15] Individual Questionnaire And Exper |        227       73.46       96.76
[19] Individual Questionnaire Without H |         10        3.24      100.00
------------------------------------+-----------------------------------
                              Total |        309      100.00
```

A total of 309 people who were 15 years old at the time of the survey and had been part of a survey household as a child in 2011 completed an individual interview in 2016.

d) The individual with pid=588010 was born in 1984 in a panel household and was still part of the sample in 2009. The individual changed households twice during this time. In which years?

To identify how often and when a individual changed households, you must display all available household numbers in PPATHL for individual 588010.

```
list pid hid syear gebjahr if pid==588010
```



```
          +----------------------------------+
          |    pid      hid   syear   gebjahr |
          |----------------------------------|
 292415.  | 588010    58807    1985      1984 |
 292416.  | 588010    58807    1986      1984 |
 292417.  | 588010    73407    1987      1984 |
 292418.  | 588010    73407    1988      1984 |
 292419.  | 588010    73407    1989      1984 |
          |----------------------------------|
 292420.  | 588010    73407    1990      1984 |
 292421.  | 588010    73407    1991      1984 |
 292422.  | 588010    73407    1992      1984 |
 292423.  | 588010    73407    1993      1984 |
 292424.  | 588010    73407    1994      1984 |
          |----------------------------------|
 292425.  | 588010    73407    1995      1984 |
 292426.  | 588010    73407    1996      1984 |
 292427.  | 588010    73407    1997      1984 |
 292428.  | 588010    73407    1998      1984 |
 292429.  | 588010    73407    1999      1984 |
          |----------------------------------|
 292430.  | 588010    73407    2000      1984 |
 292431.  | 588010    73407    2001      1984 |
 292432.  | 588010    73407    2002      1984 |
 292433.  | 588010    73407    2003      1984 |
 292434.  | 588010    73407    2004      1984 |
          |----------------------------------|
 292435.  | 588010    73407    2005      1984 |
 292436.  | 588010    73407    2006      1984 |
 292437.  | 588010    73407    2007      1984 |
 292438.  | 588010   132608    2008      1984 |
 292439.  | 588010   132608    2009      1984 |
          |----------------------------------|
 292440.  | 588010   132608    2010      1984 |
          +----------------------------------+
```

Individual 588010 has participated in the survey since 1985 as part of household 58807. From 1987 to 2007 the individual was in household 73407, from 2008 on, the individual was in household 132608.

Last change: Sep 26, 2022

To this tracking data, users should merge variables they want to analyze. For putting together cross-sectional data sets, these exercises are helpful:

6.2 Generating a Cross-Sectional Dataset

This example involves generating a dataset to analyze health satisfaction determinants in 2008, and you can either use the Paneldata.org syntax generator or write your own syntax file to perform this task. You can search for the variable names in Paneldata.org (or use the variables below directly).

1. Generate a cross-sectional dataset for the year 2008, which should contain all persons with the following characteristics:

- Respondents in 2008 **"ynetto"**
- Lived in a private household in 2008 **"ypop"**

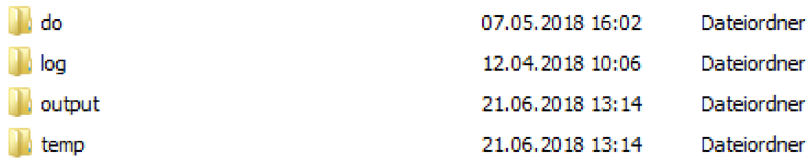
The dataset should contain the following variables of interest.

- satisfaction with health **"yp0101"**
- smoking currently yes/no **"yp10601"**
- current employment status **"emplst08"**
- monthly household net income **"hinc08"**

In addition, the dataset should contain the following additional information for a 2008 cross-sectional analysis (these variables are automatically generated by paneldata.org):

- current cross-section weighting factor **"yphrf"**
- personal number **"persnr"**
- original household number **"hhnr"**
- current household number **"yhhnr"**
- sample affiliation **"psample"**
- gender **"sex"**
- year of birth **"gebjahr"**

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store commands, log files, datasets, and temporary datasets. Open an empty do file and define your created paths with globals:

```
************************************************
* Set relative paths to the working directory
************************************************
global AVZ         "H:/material/exercises"
global MY_IN_PATH "//hume/rdc-prod/distribution/soep-core/soep.v37/eu/Stata/"
global MY_DO_FILES "$AVZ/do/"
global MY_LOG_OUT "$AVZ/log/"
global MY_OUT_DATA "$AVZ/output/"
global MY_OUT_TEMP "$AVZ/temp/"
```

The global “AVZ” defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to your data.

Use ppath as the source file together with the required variables. Keep all cases with completed interviews. In addition, your dataset should only contain respondents who can make a statement on the content of the question. For example, you can use the net code to identify and remove children from your dataset.

```
* * * PFAD * * *

use hhnr persnr sex gebjahr psample yhhnr ynetto ypop using "${MY_IN_PATH}ppfad.dta"


* * * BALANCED VS UNBALANCED * * *

keep if ( (ynetto >= 10 & ynetto < 20) )


* * * PRIATVE VS ALL HOUSEHOLDS * * *

keep if ( (ypop == 1 | ypop == 2) )


* * * SORT PFAD * * *

sort persnr
save "${MY_OUT_TEMP}ppfad.dta", replace
clear
```

Attention: Please note that since version 34 (v34), PPFAD can be found in the subdirectory “Raw” of the data distribution file. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFAD.

Save the modified data temporarily. Now link your dataset with the weights of the SOEP and save your dataset as a master file.

```
* * * HRF * * *

use "${MY_IN_PATH}phrf.dta"
sort persnr
save "${MY_OUT_TEMP}hrf.dta", replace
clear
```

(continues on next page)

(continued from previous page)

```


* * * CREATE MASTER * * *

use "${MY_OUT_TEMP}ppfad.dta"
merge 1:1 persnr using "${MY_OUT_TEMP}hrf.dta"
drop if _merge == 2
drop _merge
sort persnr
save "${MY_OUT_TEMP}master.dta", replace
clear
```

Now prepare the content variables. Search for the content variables you are looking for from the various datasets and temporarily save the datasets you have created.

```
* * * READ DATA * * *

use hinc08 yhhnr using "${MY_IN_PATH}yhgen.dta"
sort yhhnr
save "${MY_OUT_TEMP}yhgen.dta", replace
clear


use yp10601 yhhnr yp0101 persnr using "${MY_IN_PATH}yp.dta"
sort persnr
save "${MY_OUT_TEMP}yp.dta", replace
clear


use emplst08 yhhnr persnr using "${MY_IN_PATH}ypgen.dta"
sort persnr
save "${MY_OUT_TEMP}ypgen.dta", replace
clear
```

Link the datasets you have created to your master file and save for analysis.

```
* * * MERGE DATA * * *

use   "${MY_OUT_TEMP}master.dta"

sort yhhnr
merge yhhnr using "${MY_OUT_TEMP}yhgen.dta"
drop if _merge == 2
drop _merge

sort persnr
merge persnr using "${MY_OUT_TEMP}yp.dta"
drop if _merge == 2
drop _merge

sort persnr
merge persnr using "${MY_OUT_TEMP}ypgen.dta"
```

(continues on next page)

(continued from previous page)

```
drop if _merge == 2
drop _merge


* * * DONE * * *

save "${MY_OUT_DATA}my_dataset.dta", replace
desc
```

You have successfully created a cross-sectional dataset for the year 2008.

2. Encode missing values into system missings (STATA)!

In SOEP, the missing codes of variables are described in detail with the values -1 to -8. To learn more about missing codes, see the section *Missing Conventions*. For content analysis, it is not always necessary to differentiate missing codes. Therefore you should be able to convert missing codes:

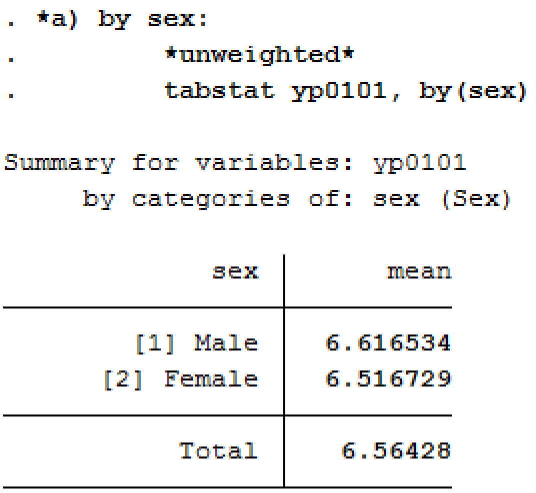
```
use "$MY_OUT_DATA\my_dataset.dta", clear


********************************************************************************
*** Exercise 2) ***
* Encode missing values into missing values in system missings (STATA)!
********************************************************************************

* mvdecode = Change missing values to numeric values and vice versa
        mvdecode _all, mv(-1=. \ -2=.t \ -3=.x \ -5=.y \ -8=.z)
```

Open the dataset for your analysis and summarize all missing codes.

3. How does average health satisfaction differ a) by gender

Satisfaction was measured on a scale of 1 to 10. To compare average satisfaction with health between women and men, you should display the mean value for both genders.

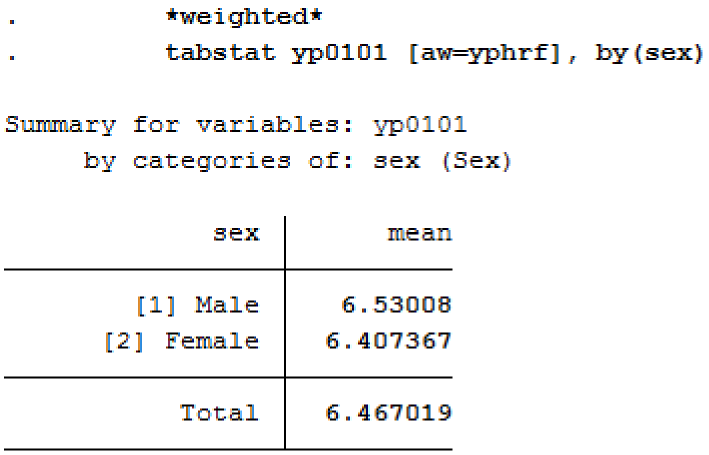
```
        *unweighted*
        tabstat yp0101, by(sex)
```



```
. *a) by sex:
.          *unweighted*
.          tabstat yp0101, by(sex)

Summary for variables: yp0101
     by categories of: sex (Sex)

          sex |      mean
--------------+----------
     [1] Male |  6.616534
   [2] Female |  6.516729
--------------+----------
        Total |   6.56428
-------------------------
```

Since you have previously added the SOEP weighting factors to the dataset for your analysis, you should use the weighting for a representative analysis.

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(sex)
```

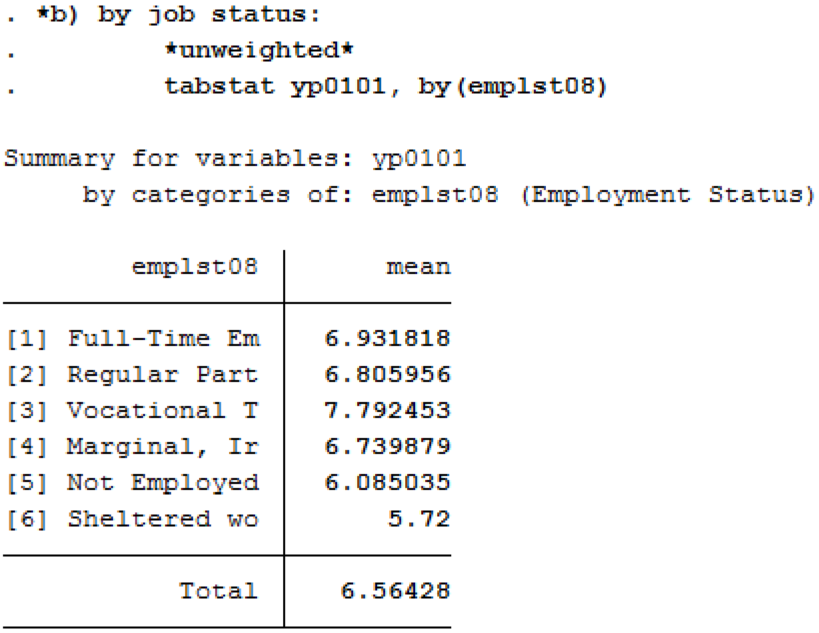


```
.           *weighted*
.           tabstat yp0101 [aw=yphrf], by(sex)

Summary for variables: yp0101
     by categories of: sex (Sex)

              sex |      mean
------------------+----------
         [1] Male |   6.53008
       [2] Female |  6.407367
------------------+----------
            Total |  6.467019
------------------+----------
```

b) Employment status

Now proceed in a similar way when comparing satisfaction with health and employment status. Compare the mean values again:

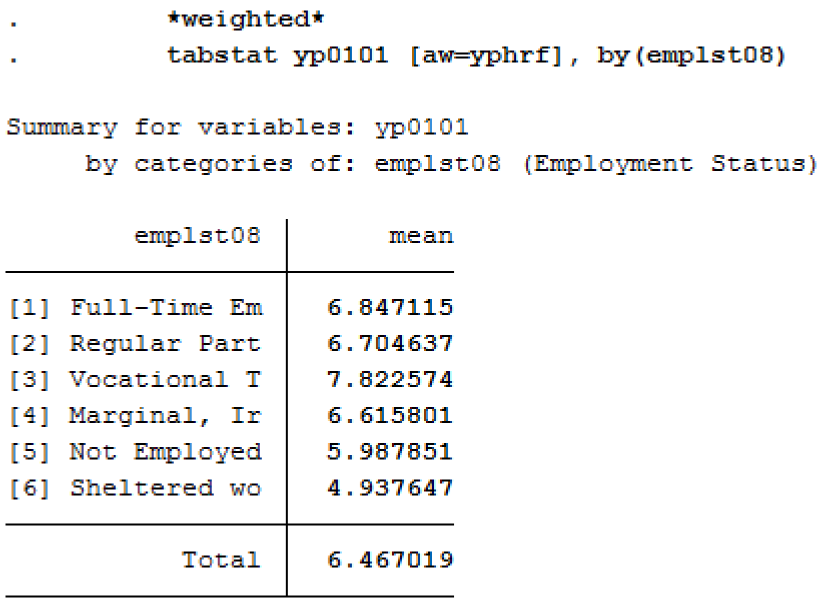
```
*b) by job status:
        *unweighted*
        tabstat yp0101, by(emplst08)
```



```
. *b) by job status:
.           *unweighted*
.           tabstat yp0101, by(emplst08)

Summary for variables: yp0101
     by categories of: emplst08 (Employment Status)

         emplst08 |      mean
------------------+----------
[1] Full-Time Em  |  6.931818
[2] Regular Part  |  6.805956
[3] Vocational T  |  7.792453
[4] Marginal, Ir  |  6.739879
[5] Not Employed  |  6.085035
[6] Sheltered wo  |      5.72
------------------+----------
            Total |   6.56428
------------------+----------
```

Since you have previously added the SOEP weighting factors to the dataset for your analysis, you should use the weighting for a representative analysis.

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(emplst08)
```



```
.           *weighted*
.           tabstat yp0101 [aw=yphrf], by(emplst08)

Summary for variables: yp0101
     by categories of: emplst08 (Employment Status)

        emplst08 |      mean
-----------------+----------
[1] Full-Time Em |  6.847115
[2] Regular Part |  6.704637
[3] Vocational T |  7.822574
[4] Marginal, Ir |  6.615801
[5] Not Employed |  5.987851
[6] Sheltered wo |  4.937647
-----------------+----------
           Total |  6.467019
----------------------------
```

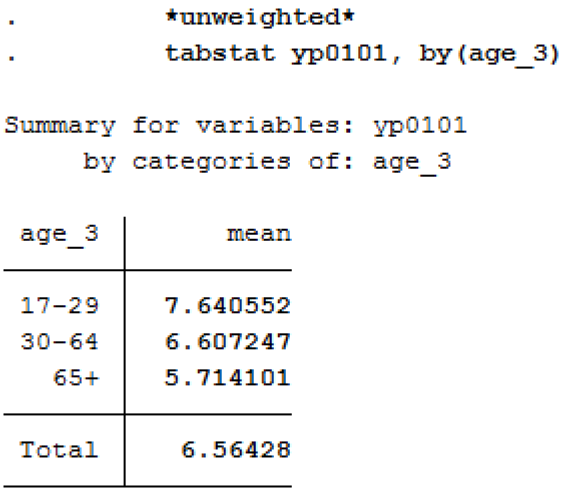
c) Age

Since you do not have a variable that represents age, you must generate a suitable age variable using the birth year variable. The year of birth is metric and should be categorized for analysis. Define categories for your age variable and assign suitable labels.

```
*c) by age in 2008 (<30, 30-64, 65+)

        gen age=2008-gebjahr
        gen age_3=age
        recode age_3 (17/29=1) (30/64=2) (65/120=3)
        label define age_3 1 "17-29" 2 "30-64" 3 "65+"
        label values age_3 age_3
```

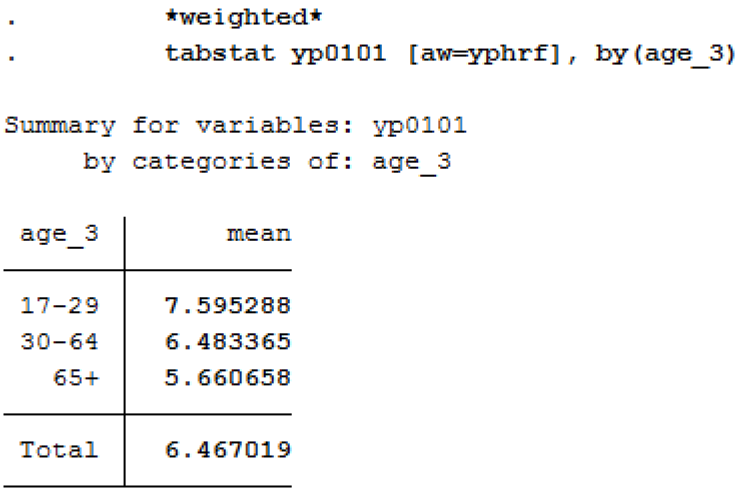
Create a mean value comparison with your age variable and health satisfaction in weighted and unweighted form.

```
        *unweighted*
        tabstat yp0101, by(age_3)
```



```
.          *unweighted*
.          tabstat yp0101, by(age_3)

Summary for variables: yp0101
     by categories of: age_3

   age_3 |      mean
---------+----------
   17-29 |  7.640552
   30-64 |  6.607247
     65+ |  5.714101
---------+----------
   Total |   6.56428
--------------------
```

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(age_3)
```



```
.          *weighted*
.          tabstat yp0101  [aw=yphrf], by(age_3)

Summary for variables: yp0101
     by categories of: age_3

   age_3 |      mean
---------+----------
   17-29 |  7.595288
   30-64 |  6.483365
     65+ |  5.660658
---------+----------
   Total |  6.467019
--------------------
```

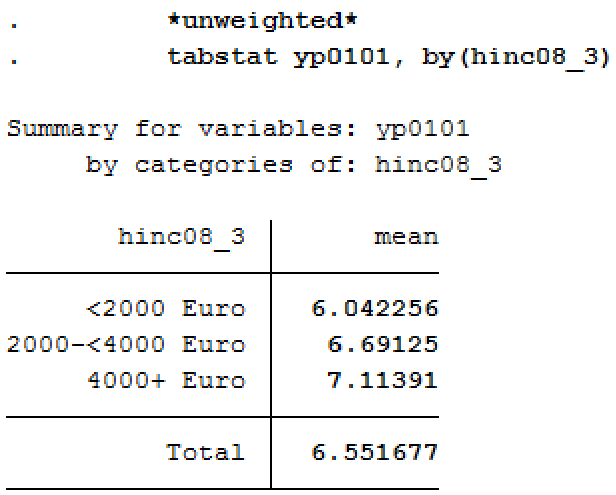
d) Income

As with age, generate a categorized version of income for household net income:

```
*d) by monthly houshold net income (-1.999, 2.000-3.999, 4000+ Euro)
        gen hinc08_3 = hinc08
        recode hinc08_3 (0/1999=1) (2000/3999=2) (4000/99999=3)
        label define hinc08_3 1 "<2000 Euro" 2 "2000-<4000 Euro" 3 "4000+ Euro"
        label values hinc08_3 hinc08_3
```

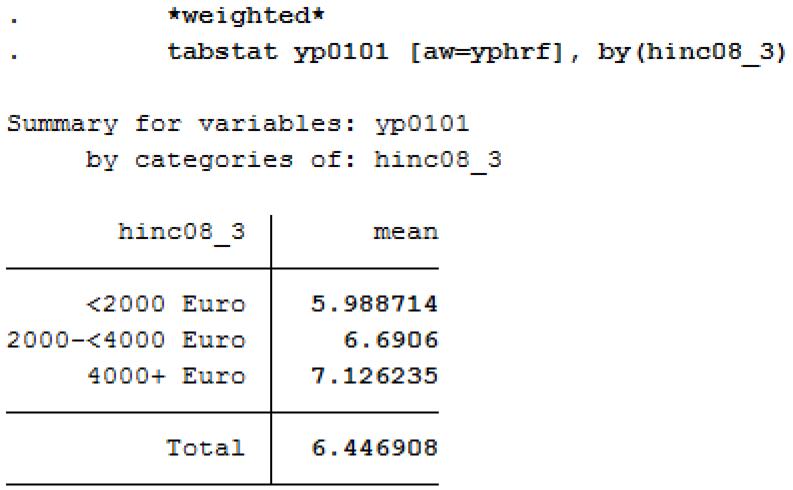
Display the mean values in weighted and unweighted form:

```
        *unweighted*
        tabstat yp0101, by(hinc08_3)
```



```
.         *unweighted*
.         tabstat yp0101, by(hinc08_3)

Summary for variables: yp0101
     by categories of: hinc08_3

        hinc08_3 |      mean
-----------------+----------
      <2000 Euro |  6.042256
 2000-<4000 Euro |   6.69125
      4000+ Euro |   7.11391
-----------------+----------
           Total |  6.551677
----------------------------
```

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(hinc08_3)
```

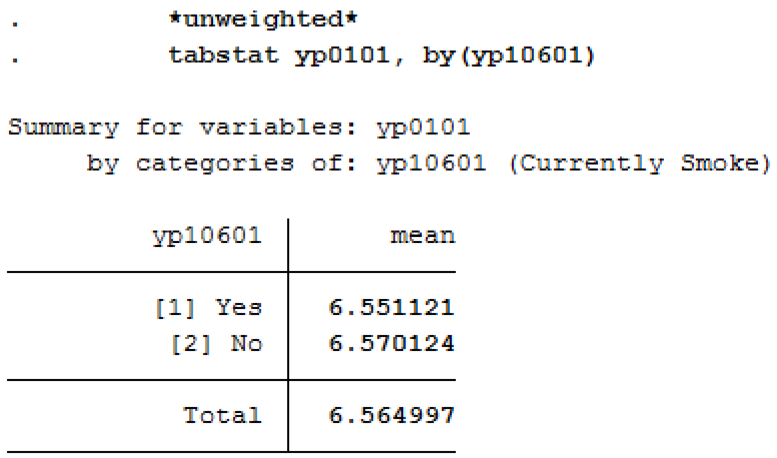


```
.         *weighted*
.         tabstat yp0101 [aw=yphrf], by(hinc08_3)

Summary for variables: yp0101
     by categories of: hinc08_3

        hinc08_3 |      mean
-----------------+----------
      <2000 Euro |  5.988714
 2000-<4000 Euro |    6.6906
      4000+ Euro |  7.126235
-----------------+----------
           Total |  6.446908
----------------------------
```

e) Smoking

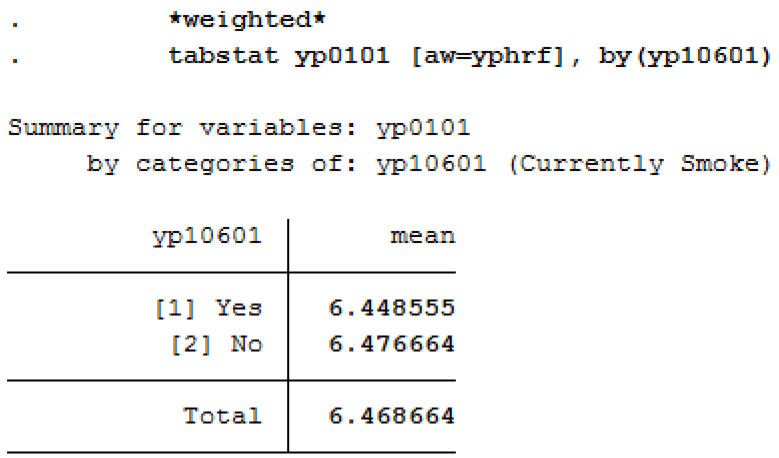
Since this variable is nominal, adjustments to this variable are not necessary. Display average satisfaction with health for smokers and non-smokers in weighted and unweighted form:

```
*e) by smoking yes/no

        *unweighted*
        tabstat yp0101, by(yp10601)
```



```
.         *unweighted*
.         tabstat yp0101, by(yp10601)

Summary for variables: yp0101
     by categories of: yp10601 (Currently Smoke)

     yp10601 |      mean
-------------+----------
     [1] Yes |  6.551121
      [2] No |  6.570124
-------------+----------
       Total |  6.564997
------------------------
```

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(yp10601)
```



```
.         *weighted*
.         tabstat yp0101 [aw=yphrf], by(yp10601)

Summary for variables: yp0101
     by categories of: yp10601 (Currently Smoke)

     yp10601 |      mean
-------------+----------
     [1] Yes |  6.448555
      [2] No |  6.476664
-------------+----------
       Total |  6.468664
------------------------
```

Last change: Jul 20, 2022

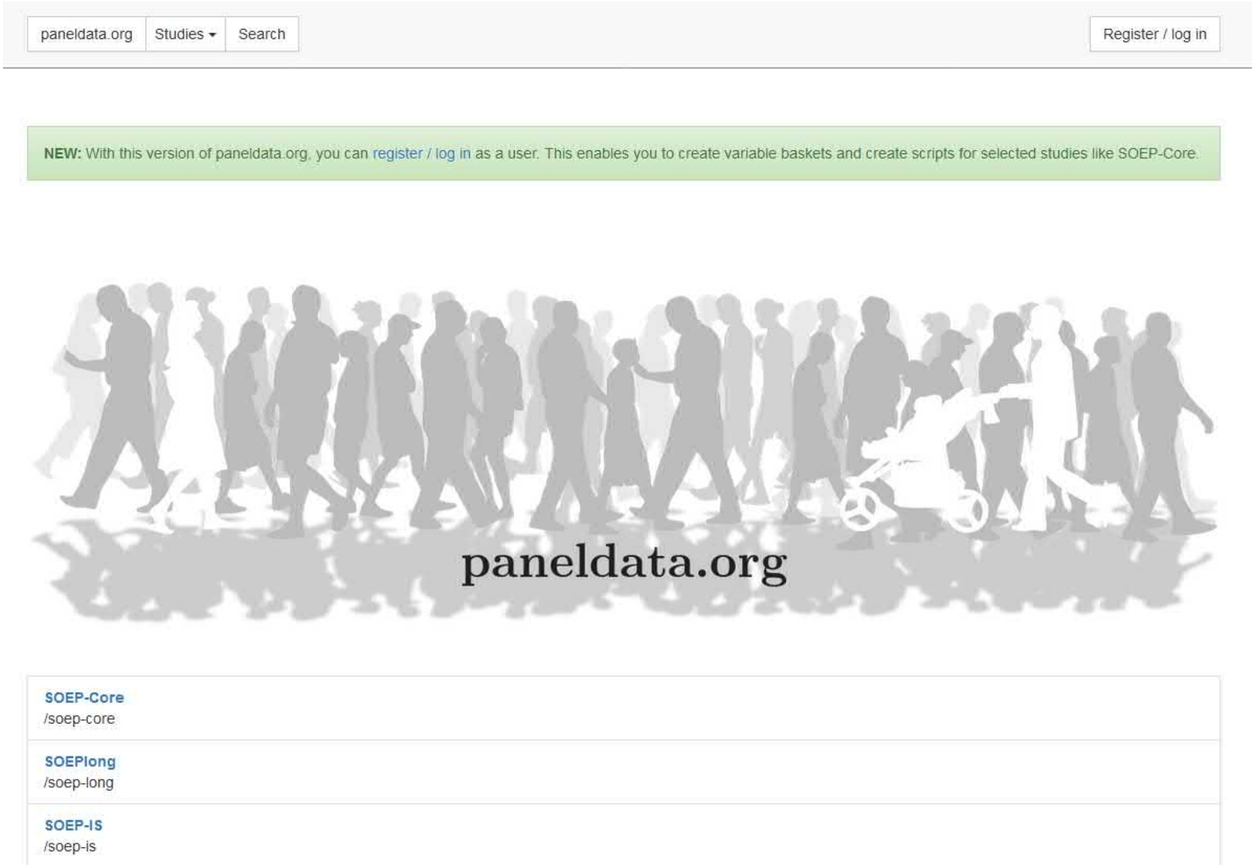
6.3 Syntax Generator on paneldata.org

Paneldata allows registered users to collect and save variables relevant to their research in a variable basket. These variables can be simply written into a single dataset with the script generator. The script generator helps you with data management and can save valuable working time.

Open Paneldata

For our experienced users, we have temporarily equipped the old soepinfo with the current data so that the variable basket function and the script generator can also be used there.

Open soepinfo



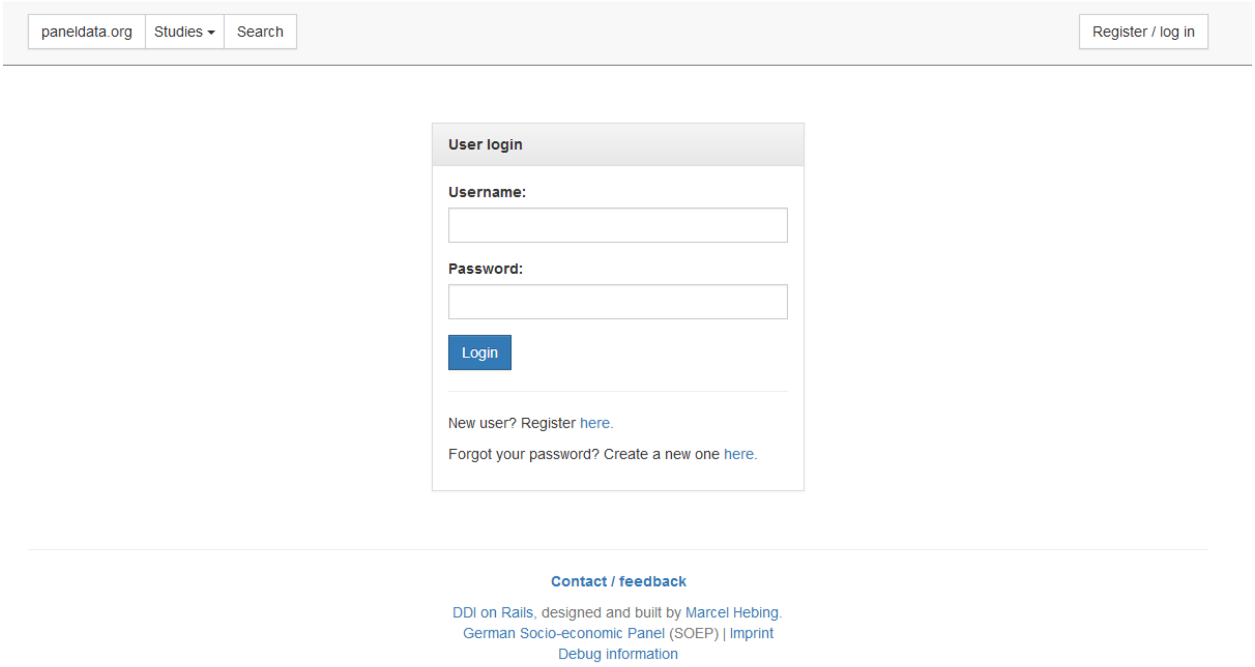
NEW: With this version of paneldata.org, you can register / log in as a user. This enables you to create variable baskets and create scripts for selected studies like SOEP-Core.

SOEP-Core
/soep-core

SOEPlong
/soep-long

SOEP-IS
/soep-is

Click on “Register / login” to log in to paneldata.org.



User login

Username:

Password:

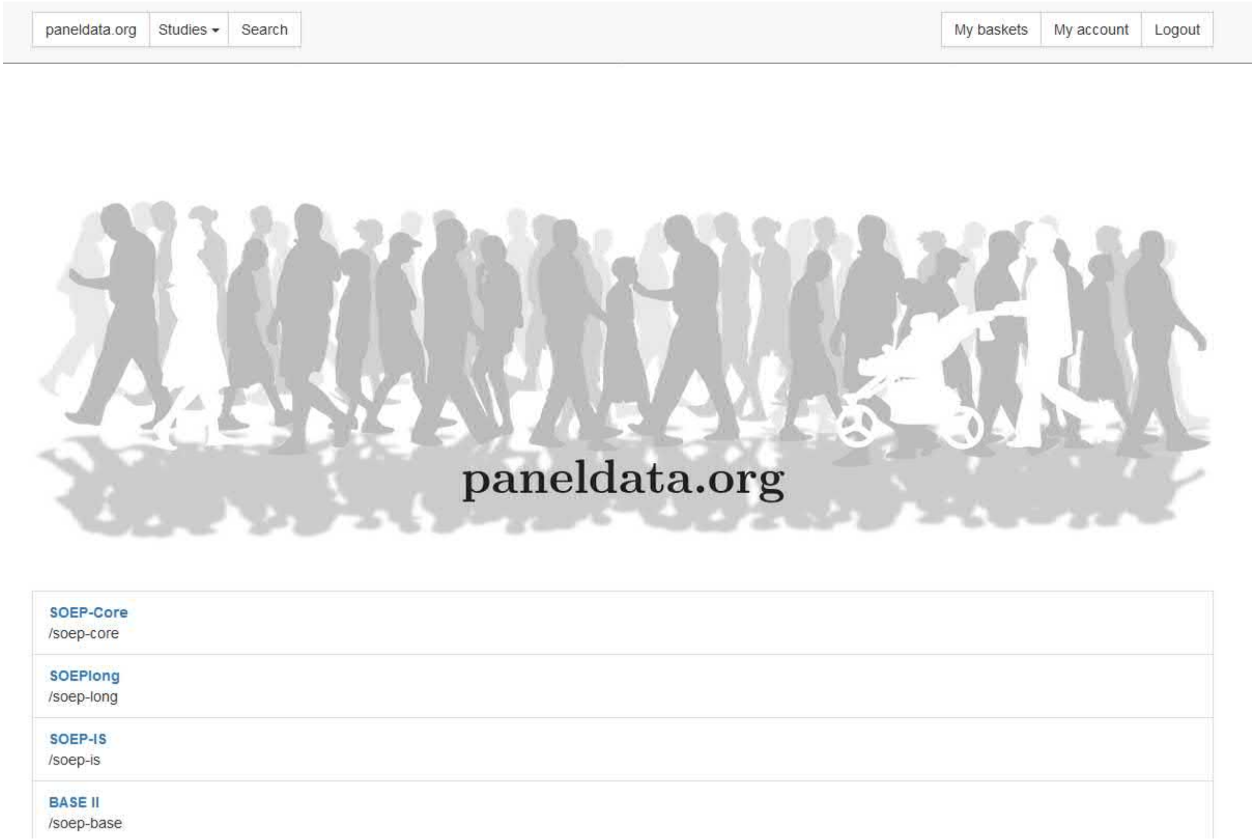
New user? Register here.

Forgot your password? Create a new one here.

Contact / feedback

DDI on Rails, designed and built by Marcel Hebing.
German Socio-economic Panel (SOEP) | Imprint
Debug information

If you have already registered, go to "user login". As a new user, you can register under "register here". Once you have logged in, you have access to the variable basket and the syntax generator.



SOEP-Core
/soep-core

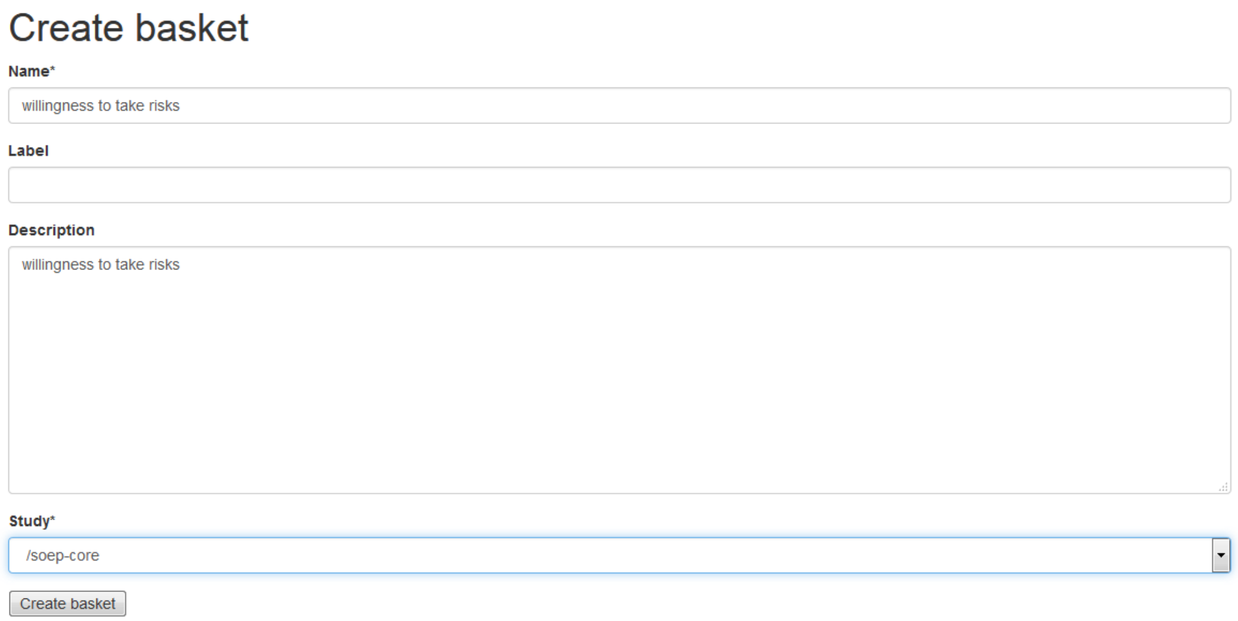
SOEPlong
/soep-long

SOEP-IS
/soep-is

BASE II
/soep-base

To access the activated functions, click on the button "my baskets". You will be taken to your personal workspace on paneldata.org.

"My baskets" displays your variable baskets. If you click on "create basket", you can create a new basket.


Create basket

Name*: willingness to take risks

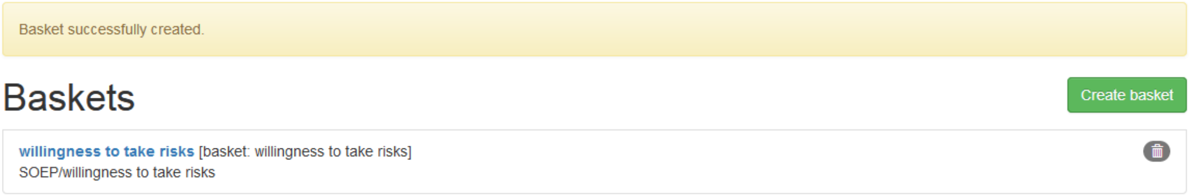
Label:

Description: willingness to take risks

Study*: /soep-core

When creating the basket, first define the name of the variable basket. The name must be lower case to be accepted by

Paneldata. Optionally, you can assign a label and enter a description. Finally, you select the study that you want to use as a database for your research. Now click on “Create basket” and your newly created variable basket appears in the interface.

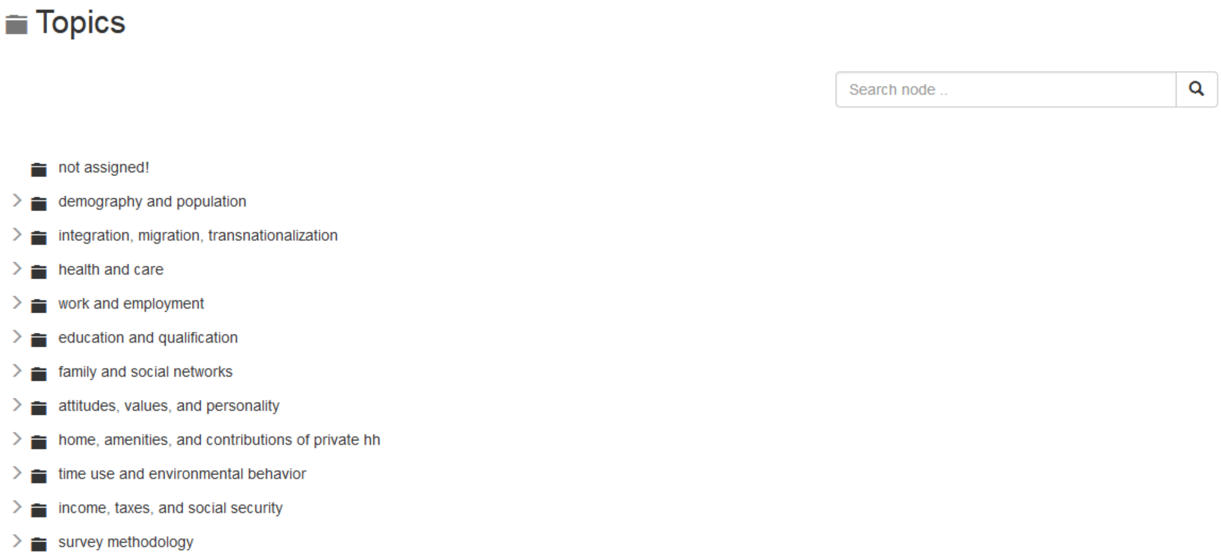


Basket successfully created.

Baskets

willingness to take risks [basket: willingness to take risks]
SOEP/willingness to take risks

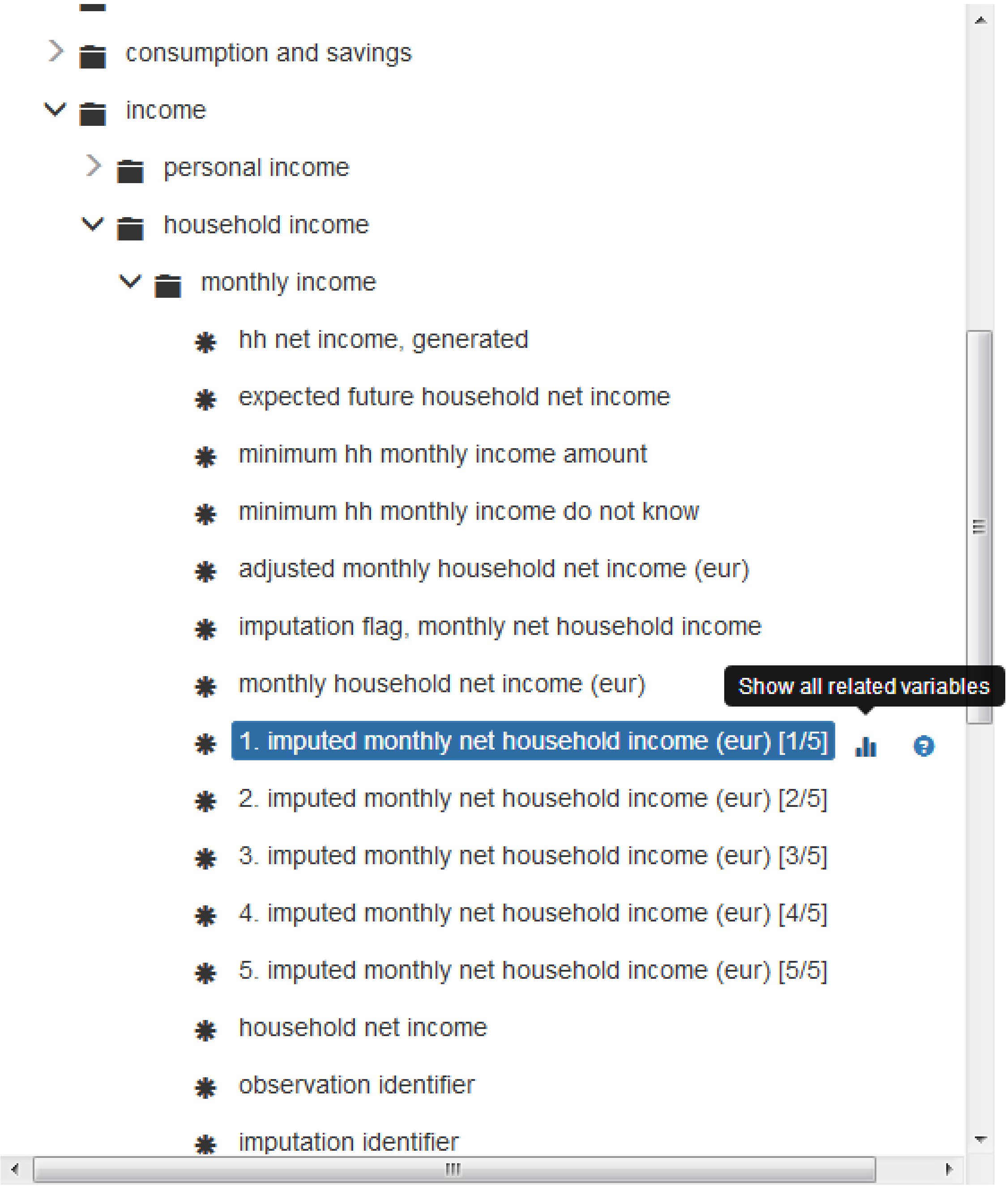
Now search for the relevant variables on paneldata.org and add them to your individual basket. For example, you are interested in monthly net household income. If you do not know the variable name, you can find the overarching concept using the topic search. Click on “paneldata.org” to get to the main page. Select the study SOEP-Core and click on “topics” at the top of the page.



Topics

- not assigned!
- demography and population
- integration, migration, transnationalization
- health and care
- work and employment
- education and qualification
- family and social networks
- attitudes, values, and personality
- home, amenities, and contributions of private hh
- time use and environmental behavior
- income, taxes, and social security
- survey methodology

Check the different topics for income-relevant concepts and select “income, taxes, and social security”.



consumption and savings

income

personal income

household income

monthly income

hh net income, generated

expected future household net income

minimum hh monthly income amount

minimum hh monthly income do not know

adjusted monthly household net income (eur)

imputation flag, monthly net household income

monthly household net income (eur)

1. imputed monthly net household income (eur) [1/5]

2. imputed monthly net household income (eur) [2/5]

3. imputed monthly net household income (eur) [3/5]

4. imputed monthly net household income (eur) [4/5]

5. imputed monthly net household income (eur) [5/5]

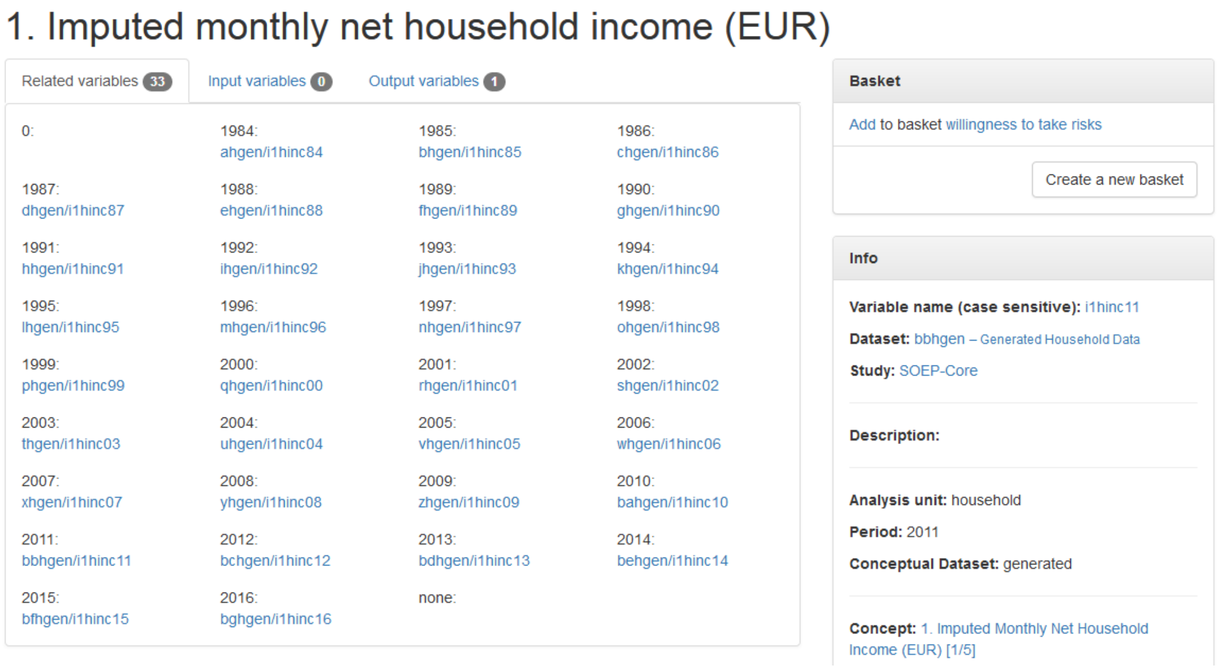
household net income

observation identifier

imputation identifier

Browse the topic list and you will reach the sub-topic “income” –> “household income” –> “monthly income”. There you will find the variables you are looking for. Click on “show all related variables” and you will see the history of variables.

Select the variable of your desired study SOEP-Core and you will reach the variable overview with important information about the variable. In the variable overview, you should make sure that the variable also meets your requirements.



1. Imputed monthly net household income (EUR)

0:	1984: ahgen/i1hinc84	1985: bhgen/i1hinc85	1986: chgen/i1hinc86
1987: dhgen/i1hinc87	1988: ehgen/i1hinc88	1989: fhgen/i1hinc89	1990: ghgen/i1hinc90
1991: hhgen/i1hinc91	1992: ihgen/i1hinc92	1993: jhgen/i1hinc93	1994: khgen/i1hinc94
1995: lhgen/i1hinc95	1996: mhgen/i1hinc96	1997: nhgen/i1hinc97	1998: ohgen/i1hinc98
1999: phgen/i1hinc99	2000: qhgen/i1hinc00	2001: rhgen/i1hinc01	2002: shgen/i1hinc02
2003: thgen/i1hinc03	2004: uhgen/i1hinc04	2005: vhgen/i1hinc05	2006: whgen/i1hinc06
2007: xhgen/i1hinc07	2008: yhgen/i1hinc08	2009: zhgen/i1hinc09	2010: bahgen/i1hinc10
2011: bbhgen/i1hinc11	2012: bchgen/i1hinc12	2013: bdhgen/i1hinc13	2014: behgen/i1hinc14
2015: bfhgen/i1hinc15	2016: bghgen/i1hinc16	none:	

Basket

Add to basket willingness to take risks

Create a new basket

Info

Variable name (case sensitive): i1hinc11

Dataset: bbhgen – Generated Household Data

Study: SOEP-Core

Description:

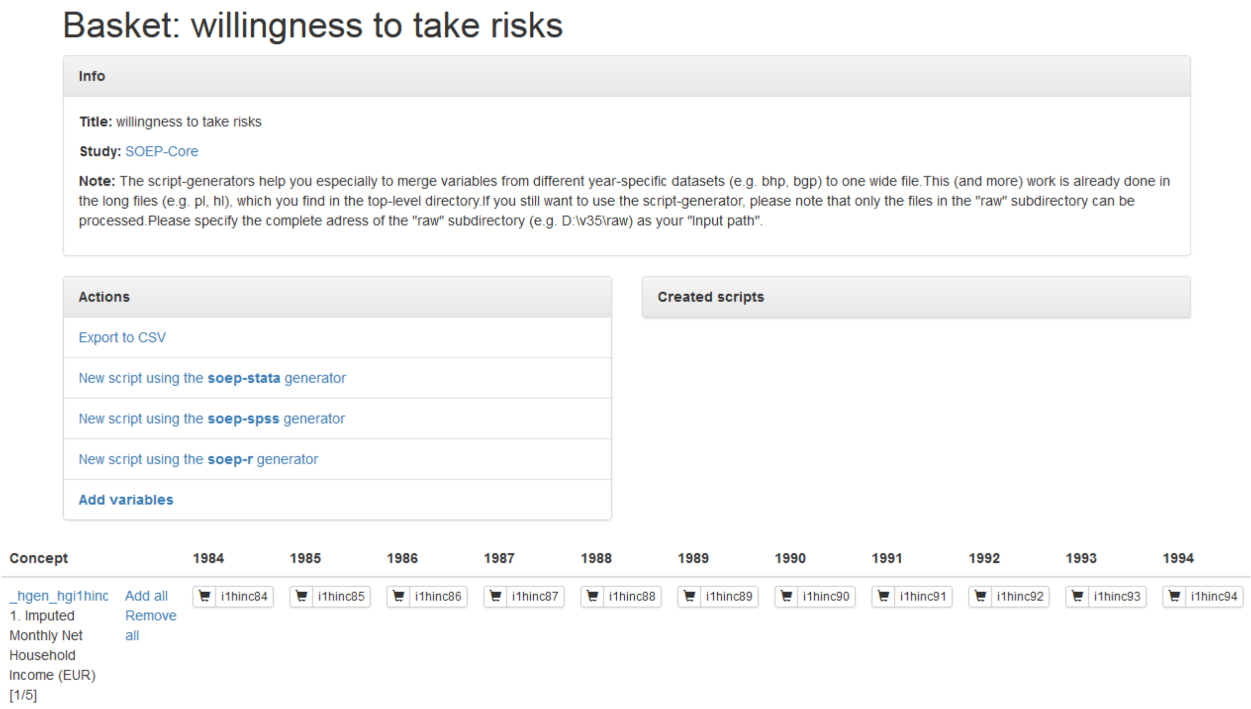
Analysis unit: household

Period: 2011

Conceptual Dataset: generated

Concept: 1. Imputed Monthly Net Household Income (EUR) [1/5]

When logged in, the basket area appears in the overview of variables. Your baskets are listed there. If you want to add the variable to a basket, click on "add to basket". If the variable is already in the basket and you want to remove it, select "remove from basket". If you want to create a new basket within the overview of variables, click on "create a new basket" and your variable will automatically be placed in the new basket. You can access the basket overview by clicking on the name of your basket in the "basket" section. Alternatively, you can click on the button "my baskets" and you will also return to the basket overview.



Basket: willingness to take risks

Info

Title: willingness to take risks

Study: SOEP-Core

Note: The script-generators help you especially to merge variables from different year-specific datasets (e.g. bhp, bgp) to one wide file.This (and more) work is already done in the long files (e.g. pl, hl), which you find in the top-level directory.If you still want to use the script-generator, please note that only the files in the "raw" subdirectory can be processed.Please specify the complete adress of the "raw" subdirectory (e.g. D:\v35\raw) as your "Input path".

Actions

Export to CSV

New script using the **soep-stata** generator

New script using the **soep-spss** generator

New script using the **soep-r** generator

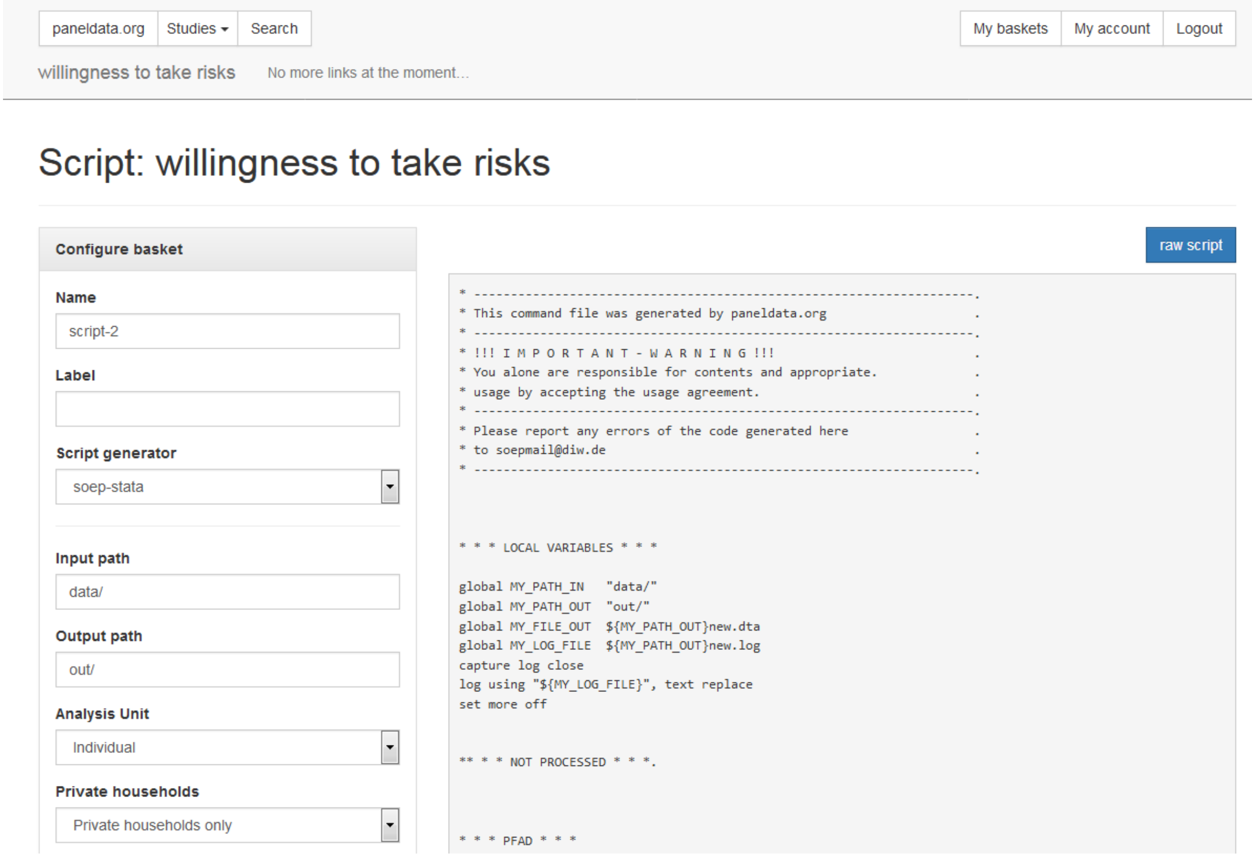
Add variables

Created scripts

Concept		1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
_hgen_hgi1hinc 1. Imputed Monthly Net Household Income (EUR) [1/5]	Add all Remove all	i1hinc84	i1hinc85	i1hinc86	i1hinc87	i1hinc88	i1hinc89	i1hinc90	i1hinc91	i1hinc92	i1hinc93	i1hinc94

Click on the basket with your added variable and you will get an overview of all variables in your basket. With "add all", you add the variables of all survey waves and the shopping cart is highlighted in green. If you are interested in a specific survey period, you can select the wave-specific variables by clicking on the shopping cart. Click on "remove all" to remove the variable from your basket. Furthermore, you can export your chosen variables to a CSV-File (Comma Seperated Value-File) and, for example, import them in STATA.

Once you have filled your basket and selected the desired survey waves, you can merge all variables into one dataset. To do this, click on "new script using the soep-xxx generator" in the "actions" area. You can choose between different statistical programs.



Script: willingness to take risks

Configure basket

Name: script-2

Label:

Script generator: soep-stata

Input path: data/

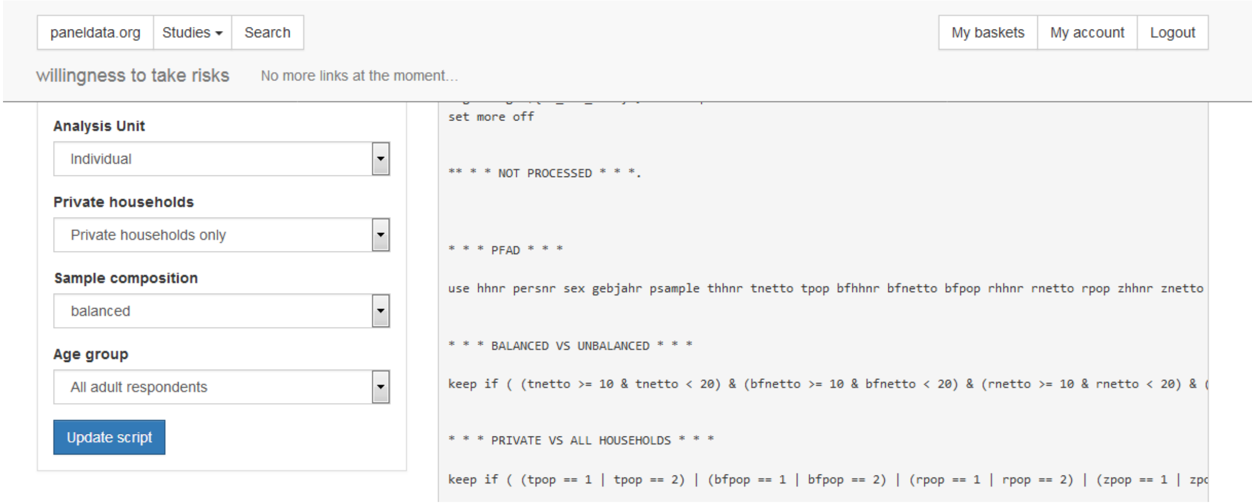
Output path: out/

Analysis Unit: Individual

Private households: Private households only

```
* --------------------------------------------------------------------.
* This command file was generated by paneldata.org                    .
* --------------------------------------------------------------------.
* !!! I M P O R T A N T - W A R N I N G !!!                           .
* You alone are responsible for contents and appropriate.             .
* usage by accepting the usage agreement.                             .
* --------------------------------------------------------------------.
* Please report any errors of the code generated here                 .
* to soepmail@diw.de                                                  .
* --------------------------------------------------------------------.



* * * LOCAL VARIABLES * * *

global MY_PATH_IN   "data/"
global MY_PATH_OUT  "out/"
global MY_FILE_OUT  ${MY_PATH_OUT}new.dta
global MY_LOG_FILE  ${MY_PATH_OUT}new.log
capture log close
log using "${MY_LOG_FILE}", text replace
set more off


** * * NOT PROCESSED * * *.



* * * PFAD * * *
```

In the script generator, you can create a script that matches your preferred variables. Specify the name of your script, select the statistics program you are using. Enter the (local) addresses of the folders in which the data is to be found or the result (and temporary intermediate results) is to be written in the two path fields ("input" and "output" path).



Analysis Unit: Individual

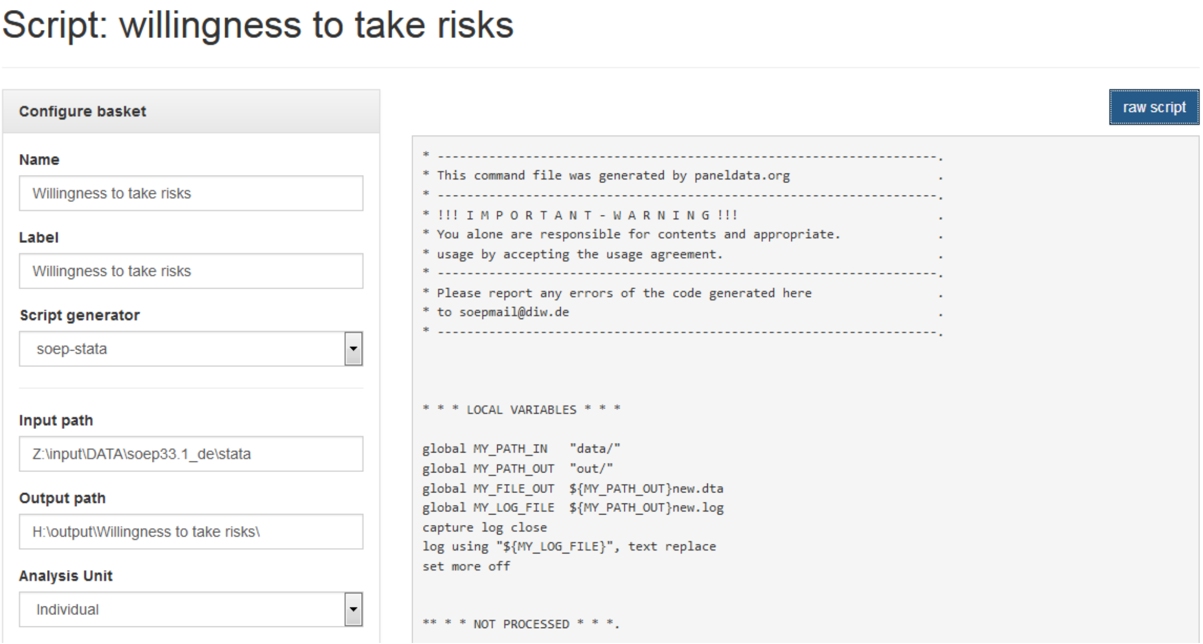
Private households: Private households only

Sample composition: balanced

Age group: All adult respondents

```
set more off


** * * NOT PROCESSED * * *.



* * * PFAD * * *

use hhnr persnr sex gebjahr psample thhnr tnetto tpop bfhhnr bfnetto bfpop rhhnr rnetto rpop zhhnr znetto


* * * BALANCED VS UNBALANCED * * *

keep if ( (tnetto >= 10 & tnetto < 20) & (bfnetto >= 10 & bfnetto < 20) & (rnetto >= 10 & rnetto < 20) & (


* * * PRIVATE VS ALL HOUSEHOLDS * * *

keep if ( (tpop == 1 | tpop == 2) | (bfpop == 1 | bfpop == 2) | (rpop == 1 | rpop == 2) | (zpop == 1 | zpc
```

In the "analysis unit" section, you decide whether all persons are considered individually within the household ("individual") or whether you are interested solely in the household as a whole ("household"). With "sample composition" you can choose between "balanced" and "unbalanced". If you select "balanced", you will receive a dataset without missing codes. The respondents provided information on all variables. For more information about balanced and unbalanced datasets, see the section *Principles of Data Analysis*. Under "age group", you can limit the respondents. When you are satisfied with your settings, click on "Update Script" and your script will be created.



Script: willingness to take risks

Configure basket

Name: Willingness to take risks

Label: Willingness to take risks

Script generator: soep-stata

Input path: Z:\input\DATA\soep33.1_de\stata

Output path: H:\output\Willingness to take risks\

Analysis Unit: Individual

```
* ---------------------------------------------------------------------.
* This command file was generated by paneldata.org                    .
* ---------------------------------------------------------------------.
* !!! I M P O R T A N T - W A R N I N G !!!                           .
* You alone are responsible for contents and appropriate.             .
* usage by accepting the usage agreement.                             .
* ---------------------------------------------------------------------.
* Please report any errors of the code generated here                 .
* to soepmail@diw.de                                                  .
* ---------------------------------------------------------------------.



* * * LOCAL VARIABLES * * *

global MY_PATH_IN   "data/"
global MY_PATH_OUT  "out/"
global MY_FILE_OUT  ${MY_PATH_OUT}new.dta
global MY_LOG_FILE  ${MY_PATH_OUT}new.log
capture log close
log using "${MY_LOG_FILE}", text replace
set more off


** * * NOT PROCESSED * * *.
```

If you click on the "raw script" button, the script is displayed in text form. Copy it to your statistical software. The result dataset has the name new(.dta, .sav). If you want to change it, you have to do it in the script. Execute the script with your statistical software and you will receive your dataset with all your chosen variables.

Last change: Sep 26, 2022

The SOEP team recommends using the long data sets. How to merge longitudinal datasets and what user service these datasets provide can be seen in these exercises:

6.4 Generating a Longitudinal Dataset

This example focuses on generating a dataset to analyze determinants of health satisfaction. You can either use the syntax generator in paneldata.org or write a syntax file yourself. You can search for variable names in Paneldata.org.

In the previous examples, you created an exercise path with four subfolders as well as corresponding globals in the STATA do-file. You can use the same folders and globals for this exercise.

1.Generate an unbalanced panel dataset for the years 2006 to 2008 using paneldata.org if you wish. The dataset should contain all respondents in private households:

The data set should contain the following variables of interest:

- health satisfaction **"wp0101" "xp0101" "yp0101"**
- currently smoking yes/no **"wp9301" "yp10601"**
- current employment status **"emplst06" "emplst07" "emplst08"**
- monthly household net income **"hinc06" "hinc07" "hinc08"**

In addition, the dataset should include the following additional information for analysis from 2006 to 2008:

- cross-sectional weighting factors for all relevant years **"wphrf" "xphrf" "yphrf"**
- individual identifier **"persnr"**
- original household number **"hhnr"**
- household number for all relevant years **"whhnr" "xhhnr" "yhhnr"**
- sample membership **"psample"**
- sex **"sex"**
- year of birth **"gebjahr"**
- population membership **"wpop" "xpop" "ypop"**

If you need detailed instructions on how the script generator works in paneldata.org, you can find them in the chapter *Syntax Generator on paneldata.org*.

If you would like to assemble your dataset yourself, you can do this with the datasets you have assembled. From the previous exercise with tracking data, you may already have an idea where to get most of the variables.

Since we want to have an unbalanced panel set, the $netto variable for the years 2006 to 2008 must also be used. In addition, our analysis must limit population membership, as we are only interested in household respondents.

Tip: If a dataset is created from several variables of different datasets, it is worth sorting the individual identifier before saving the individual data sets in order to be able to merge the data sets more easily afterwards.

1.1. Create a Master File

Use ppfad as the source file together with the required variables that you may have already found in Paneldata.org or identified from the variable label in the dataset. Note that only variables from the years to be analyzed should be used.

```

use hhnr persnr sex gebjahr psample xhhnr xnetto xpop yhhnr ynetto ypop whhnr wnetto wpop using "${MY_PATH_IN}ppfad.dta"

```

Since we want to obtain an unbalanced data set, i.e., individuals who have completed an individual questionnaire at least once within the last three years, you must restrict the variable $netto (survey status). Also, we only want to analyze private households, so we need a further restriction of the $pop (sample membership) variable.

```

keep if ( (xnetto >= 10 & xnetto < 20) | (ynetto >= 10 & ynetto < 20) | (wnetto >= 10 & wnetto < 20) )


* * * PRIVATE VS ALL HOUSEHOLDS * * *

keep if ( (xpop == 1 | xpop == 2) | (ypop == 1 | ypop == 2) | (wpop == 1 | wpop == 2) )

```

Then we sort the persnr (individual identifier) in the datasets and save it.

```

sort persnr
save "${MY_PATH_OUT}ppfad.dta", replace
clear

```

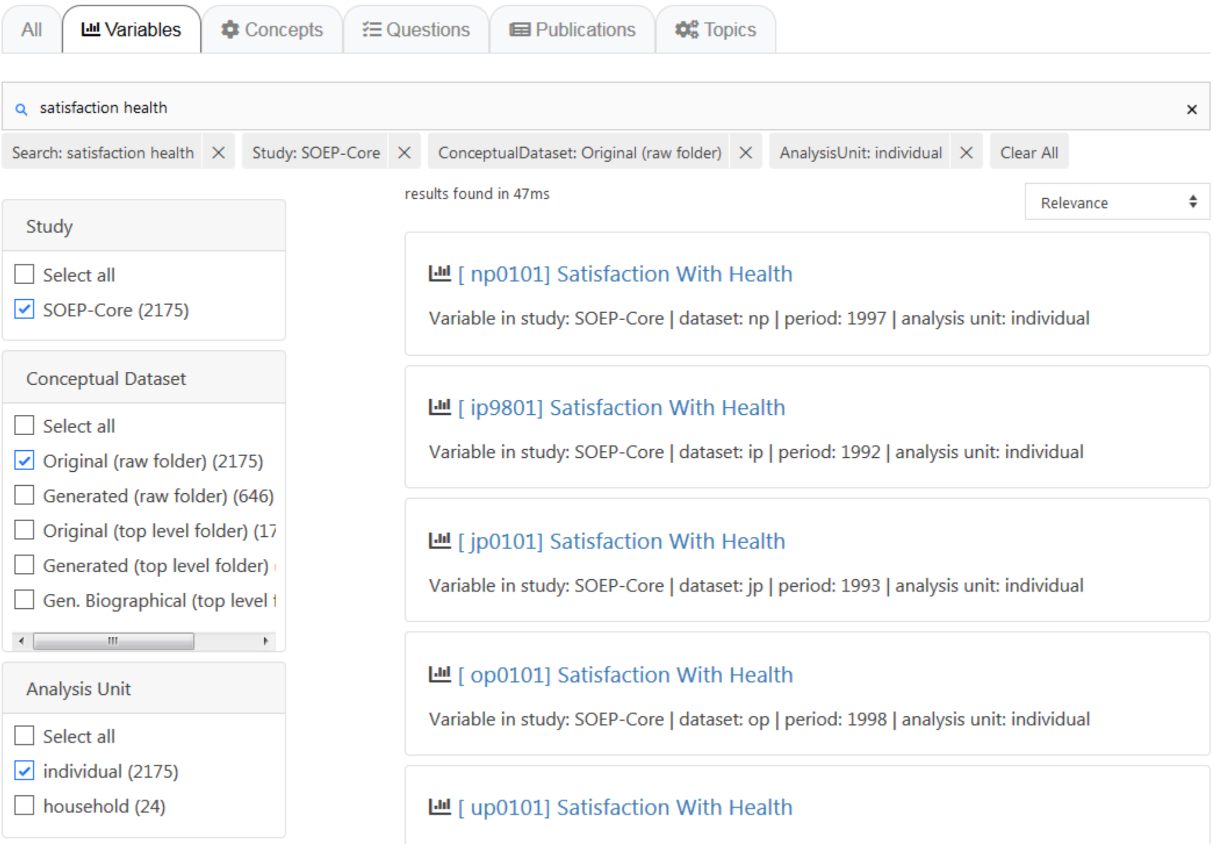
What is still missing is the cross-sectional weighting factor and the variables of interest for the analysis. To apply the weighting factors to the dataset, open the weighting dataset for the person-level phrf, sort it, and save it again.

```

use persnr wphrf xphrf yphrf using "${MY_PATH_IN}phrf.dta"
sort persnr
save "${MY_PATH_OUT}phrf.dta", replace
clear

```

Now we come to the content variables. In order not to have to click through all of the datasets in the data release, it is recommended that the label be entered for the variable of interest from paneldata.org.

Use the filter to narrow your search. Select our main study SOEP-Core, the search type “variable”, the conceptual dataset “Original (raw folder)”, analysis unit and the corresponding year. Once you have clicked on the year of interest, a variable history is displayed. You can use this to see which years the variable was collected and what the variable is called.

Example: Variable Label “Satisfaction Health”



satisfaction health

Search: satisfaction health ✕ | Study: SOEP-Core ✕ | ConceptualDataset: Original (raw folder) ✕ | AnalysisUnit: individual ✕

results found in 47ms

Study

□ Select all
☑ SOEP-Core (2175)

Conceptual Dataset

□ Select all
☑ Original (raw folder) (2175)
□ Generated (raw folder) (646)
□ Original (top level folder) (17
□ Generated (top level folder)
□ Gen. Biographical (top level

Analysis Unit

□ Select all
☑ individual (2175)
□ household (24)

[np0101] Satisfaction With Health

Variable in study: SOEP-Core | dataset: np | period: 1997 | analysis unit: individual

[ip9801] Satisfaction With Health

Variable in study: SOEP-Core | dataset: ip | period: 1992 | analysis unit: individual

[jp0101] Satisfaction With Health

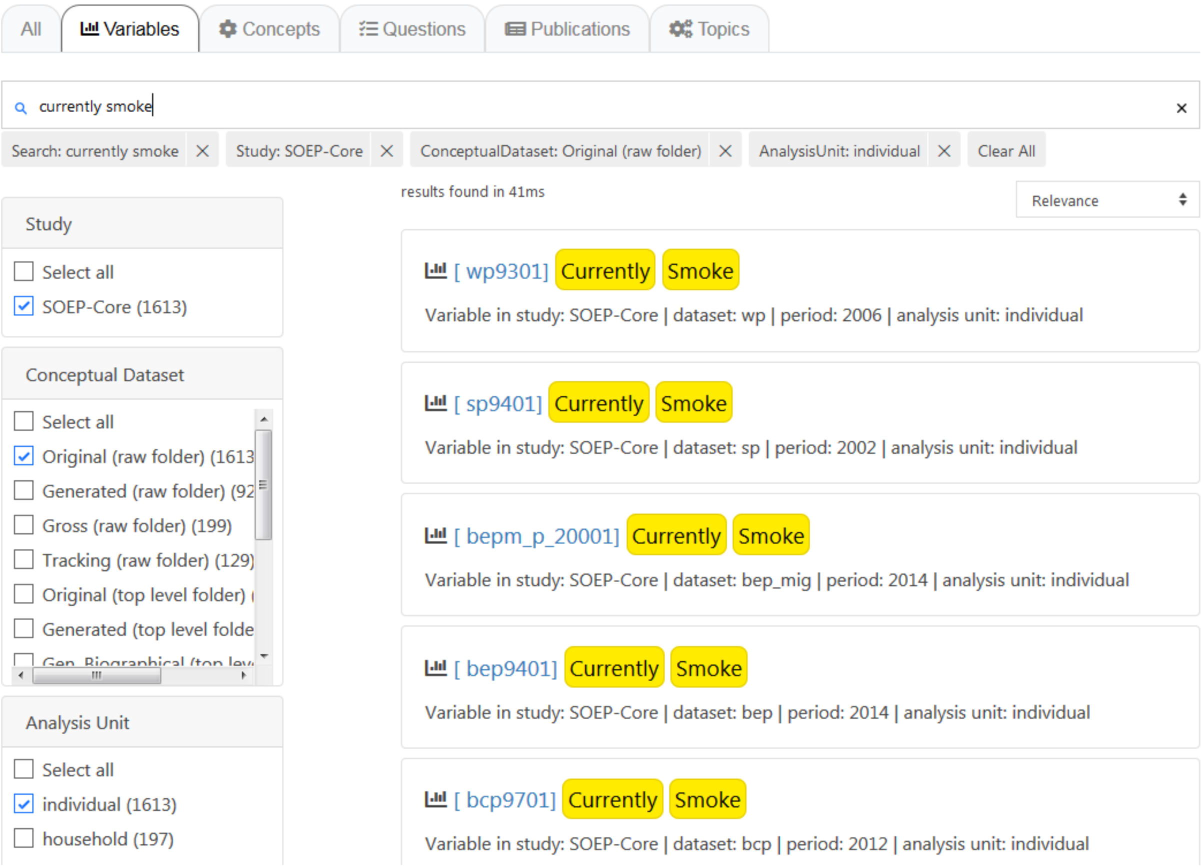
Variable in study: SOEP-Core | dataset: jp | period: 1993 | analysis unit: individual

[op0101] Satisfaction With Health

Variable in study: SOEP-Core | dataset: op | period: 1998 | analysis unit: individual

[up0101] Satisfaction With Health

Example: Variable Label “currently smoking yes/no”



currently smoke

Search: currently smoke ✕ | Study: SOEP-Core ✕ | ConceptualDataset: Original (raw folder) ✕ | AnalysisUnit: individual ✕ | Clear All

results found in 41ms

Relevance

Study

□ Select all
☑ SOEP-Core (1613)

Conceptual Dataset

□ Select all
☑ Original (raw folder) (1613
□ Generated (raw folder) (92
□ Gross (raw folder) (199)
□ Tracking (raw folder) (129)
□ Original (top level folder) (
□ Generated (top level folde
□ Gen. Biographical (top lev

Analysis Unit

□ Select all
☑ individual (1613)
□ household (197)

[wp9301] Currently Smoke
Variable in study: SOEP-Core | dataset: wp | period: 2006 | analysis unit: individual

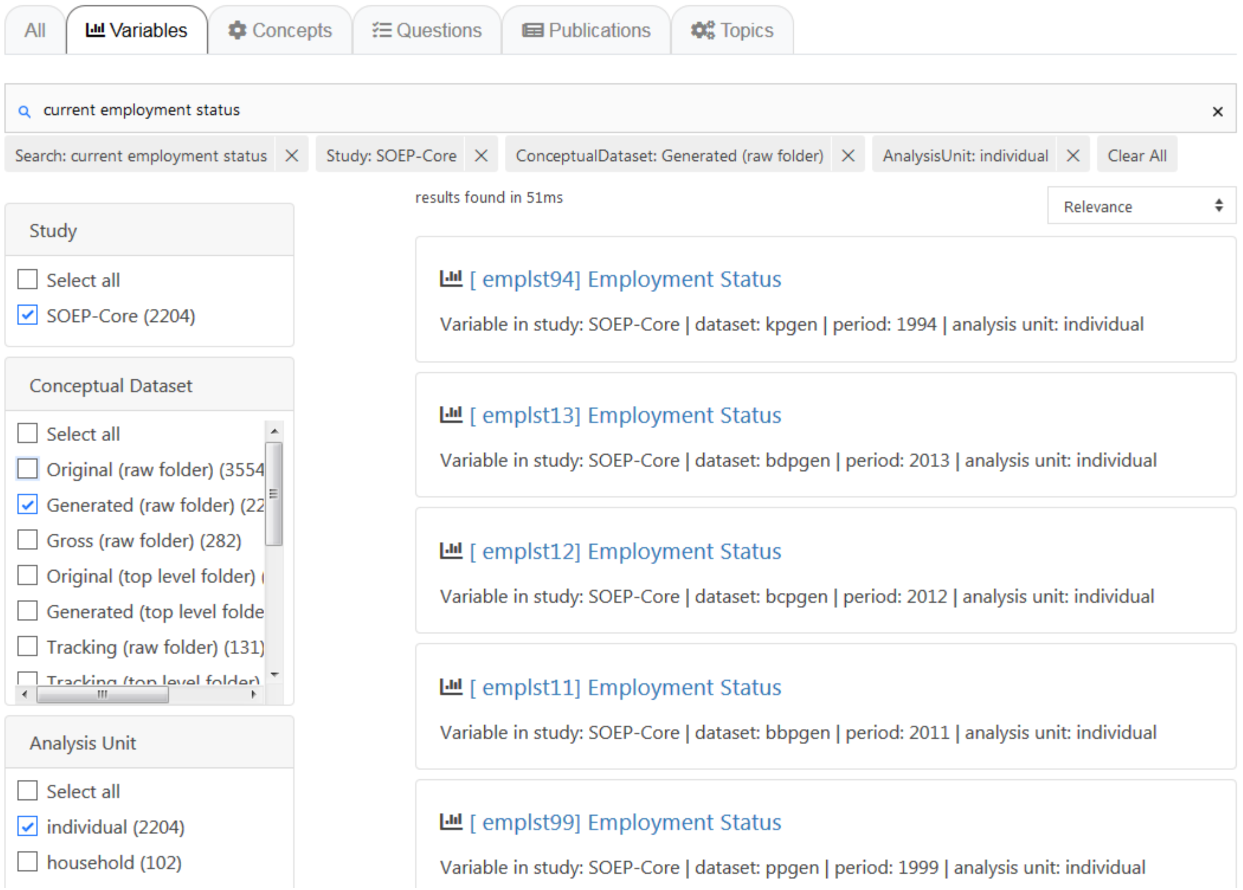
[sp9401] Currently Smoke
Variable in study: SOEP-Core | dataset: sp | period: 2002 | analysis unit: individual

[bepm_p_20001] Currently Smoke
Variable in study: SOEP-Core | dataset: bep_mig | period: 2014 | analysis unit: individual

[bep9401] Currently Smoke
Variable in study: SOEP-Core | dataset: bep | period: 2014 | analysis unit: individual

[bcp9701] Currently Smoke
Variable in study: SOEP-Core | dataset: bcp | period: 2012 | analysis unit: individual

Example: Variable Label “current employment status”



current employment status

Search: current employment status × Study: SOEP-Core × ConceptualDataset: Generated (raw folder) × AnalysisUnit: individual ×

results found in 51ms

Study

□ Select all
☑ SOEP-Core (2204)

Conceptual Dataset

□ Select all
□ Original (raw folder) (3554
☑ Generated (raw folder) (22
□ Gross (raw folder) (282)
□ Original (top level folder) (
□ Generated (top level folde
□ Tracking (raw folder) (131)
□ Tracking (top level folder)

Analysis Unit

□ Select all
☑ individual (2204)
□ household (102)

[emplst94] Employment Status
Variable in study: SOEP-Core | dataset: kpgen | period: 1994 | analysis unit: individual

[emplst13] Employment Status
Variable in study: SOEP-Core | dataset: bdpgen | period: 2013 | analysis unit: individual

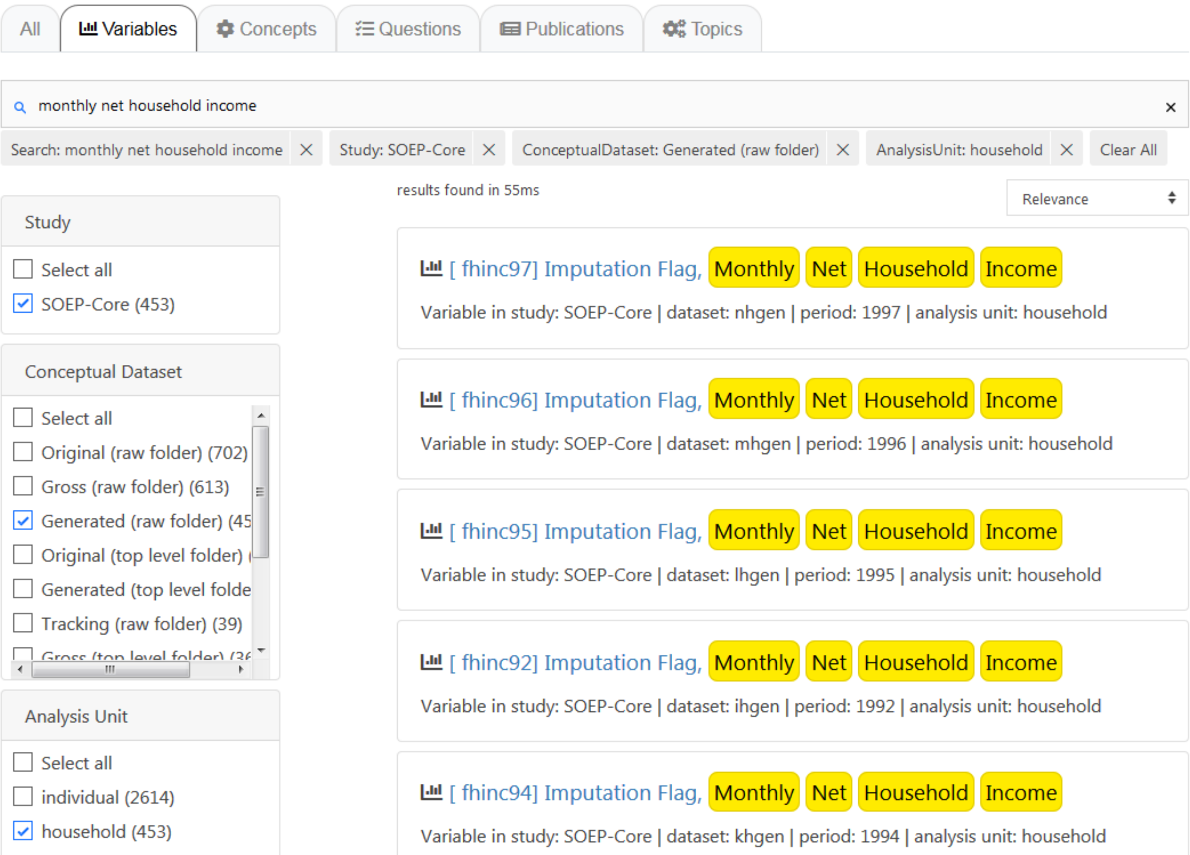
[emplst12] Employment Status
Variable in study: SOEP-Core | dataset: bcpgen | period: 2012 | analysis unit: individual

[emplst11] Employment Status
Variable in study: SOEP-Core | dataset: bbpgen | period: 2011 | analysis unit: individual

[emplst99] Employment Status
Variable in study: SOEP-Core | dataset: ppgen | period: 1999 | analysis unit: individual

Example: Variable Label "monthly net household income"

]



All | Variables | Concepts | Questions | Publications | Topics

monthly net household income ×

Search: monthly net household income × | Study: SOEP-Core × | ConceptualDataset: Generated (raw folder) × | AnalysisUnit: household × | Clear All

Study
□ Select all
☑ SOEP-Core (453)

Conceptual Dataset
□ Select all
□ Original (raw folder) (702)
□ Gross (raw folder) (613)
☑ Generated (raw folder) (45
□ Original (top level folder)
□ Generated (top level folde
□ Tracking (raw folder) (39)
□ Gross (top level folder) (36

Analysis Unit
□ Select all
□ individual (2614)
☑ household (453)

results found in 55ms

Relevance

[fhinc97] Imputation Flag, Monthly Net Household Income
Variable in study: SOEP-Core | dataset: nhgen | period: 1997 | analysis unit: household

[fhinc96] Imputation Flag, Monthly Net Household Income
Variable in study: SOEP-Core | dataset: mhgen | period: 1996 | analysis unit: household

[fhinc95] Imputation Flag, Monthly Net Household Income
Variable in study: SOEP-Core | dataset: lhgen | period: 1995 | analysis unit: household

[fhinc92] Imputation Flag, Monthly Net Household Income
Variable in study: SOEP-Core | dataset: ihgen | period: 1992 | analysis unit: household

[fhinc94] Imputation Flag, Monthly Net Household Income
Variable in study: SOEP-Core | dataset: khgen | period: 1994 | analysis unit: household

To merge the data, you can either use the script generator on paneldata.org or write the syntax manually into a do-file.

We now have all the information we need to create a master file. As already mentioned with **TIP!**, it is recommended to save the datasets sorted by the persnr (individual identifier) before merging.

```
use persnr wp0101 wp9301 using "${MY_PATH_IN}wp.dta"
sort persnr
save "${MY_PATH_OUT}wp.dta", replace
clear

* * * Persons 2007 * * *
use persnr xp0101 using "${MY_PATH_IN}xp.dta"
sort persnr
save "${MY_PATH_OUT}xp.dta", replace
clear

* * * Persons 2008 * * *
use persnr yp0101 yp10601 using "${MY_PATH_IN}yp.dta"
sort persnr
save "${MY_PATH_OUT}yp.dta", replace
clear

```

With the help of a unique identifier, which is either the household ($hhnr) or individual identifier (persnr), you can now merge all datasets or individual variables to ppfad. Which identifier to use when depends on the unit of analysis. Since we are on the individual level, our indicator is persnr (individual identifier).

We load the dataset ppfad and merge our datasets or variables to ppfad.

```

merge 1:1 persnr using "${MY_PATH_OUT}phrf.dta", keep(master match) nogen


* merge data from $p.dta
merge 1:1 persnr using "${MY_PATH_IN}/wp.dta", keepus(wp0101 wp9301)  keep(master match) nogen // health & smoking
merge 1:1 persnr using "${MY_PATH_IN}/xp.dta", keepus(xp0101)                    keep(master match) nogen // health
merge 1:1 persnr using "${MY_PATH_IN}/yp.dta", keepus(yp0101 yp10601) keep(master match) nogen // health & smoking

* merge data from $pgen.dta
local y = 6
foreach wave in w x y {
        merge 1:1 persnr using "${MY_PATH_IN}/`wave'pgen.dta", keepus(emplst0`y')nogen keep(master match)
        local y = `y' + 1
}

* merge data from $hgen.dta
local y = 6
foreach wave in w x y {
        merge m:1 `wave'hhnr using "${MY_PATH_IN}/`wave'hgen.dta", keepus(hinc0`y') nogen keep(master match)
        local y = `y' + 1
}

```

2. Encode missing values in system failings (STATA)!

After the master file has been created with all required information, the missing values, which can take between -1 to -8 in SOEP, must be recoded to missings. This step is important for converting a wide-format data set to a long format.

```
********************************************************************************
*** Task 2) ***
* Encode missing values in systemmissings (STATA)!
********************************************************************************

        mvdecode _all, mv(-1=. \ -2=.t \ -3=.x \ -5=.y \ -8=.z)
```

3. The data set is in "wide" format, i.e., additional years are displayed as additional variables (columns). For many analyses, it makes sense to convert datasets into the "long" format. In long format, additional years are displayed as additional lines. If the dataset covers three years, as in this example, there are three lines for each person. Convert the data set to long format using the STATA command reshape.!

Since these are cross-sectional variables, it can be assumed that each variable has at least one wave abbreviation, which makes the variable unique. Conversely, this means that the variables must be renamed before the reshape command.

Before renaming all original variables (e.g., from $P data sets) it must be checked whether the question and the answer categories were the same in all years (you can also look up the exact wording of the question in the corresponding questionnaire). If changes are made, the variables may have to be recoded.

```
*Check if original variable have changed over time
        tab1 wp0101 xp0101 yp0101
        tab1 wp9301 yp10601
        /*additionally check questionaires for exact wording*/
```

How you rename the variables is largely up to you. However, you should ensure that the name remains consistent over time and that the variable only differs according to the year (variable name + four-digit year suffix, e.g., zufr2006, zufr2007, zufr2008). You can rename the variables either manually, line by line, or for advanced users using a loop.

Example of manual renaming:

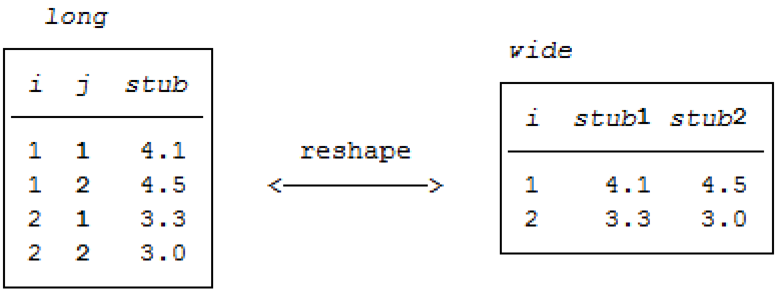
```
*rename time-variant variables
*with examples how to use loops (but can also be done "manually")
        rename wp9301 smoke2006
        rename yp10601 smoke2008
        rename wp0101 health2006
        rename xp0101 health2007
        rename yp0101 health2008
        ...
```

Example of a loop:

```
        foreach  x in 6 7 8 {
                rename hinc0`x' hinc200`x'
                rename emplst0`x' emplst200`x'
                }


        local y=2006
        foreach w in w x y {
                rename `w'hhnr hhnrakt`y'
                rename `w'netto netto`y'
                rename `w'pop pop`y'
                rename `w'phrf phrf`y'
                local y=`y'+1
                }
```

3.1. The reshape command

Now that we have made all relevant preparations, you can start to convert the dataset. If you want to convert a dataset, you can do this in both directions:



i	j	stub
1	1	4.1
1	2	4.5
2	1	3.3
2	2	3.0

i	stub1	stub2
1	4.1	4.5
2	3.3	3.0

In our case, we reshape from wide to long. This means that a new variable name must be assigned for the year of the survey (j). The variable is then generated automatically. Currently, each person is assigned a line in Stata.

persnr	hhnr	wave	sex	smoke2006	smoke2008
12345	123	x	m	yes	yes
54321	211	x	m	no	no

```
*reshape dataset to long-format
        reshape long health smoke emplst hinc netto pop hhnrakt phrf, i(persnr) j(year)
        bys persnr: gen waves=_N                /*additional information: count number of waves per person*/
        tab waves
```

After the reshape command, you have one line per year for each person:

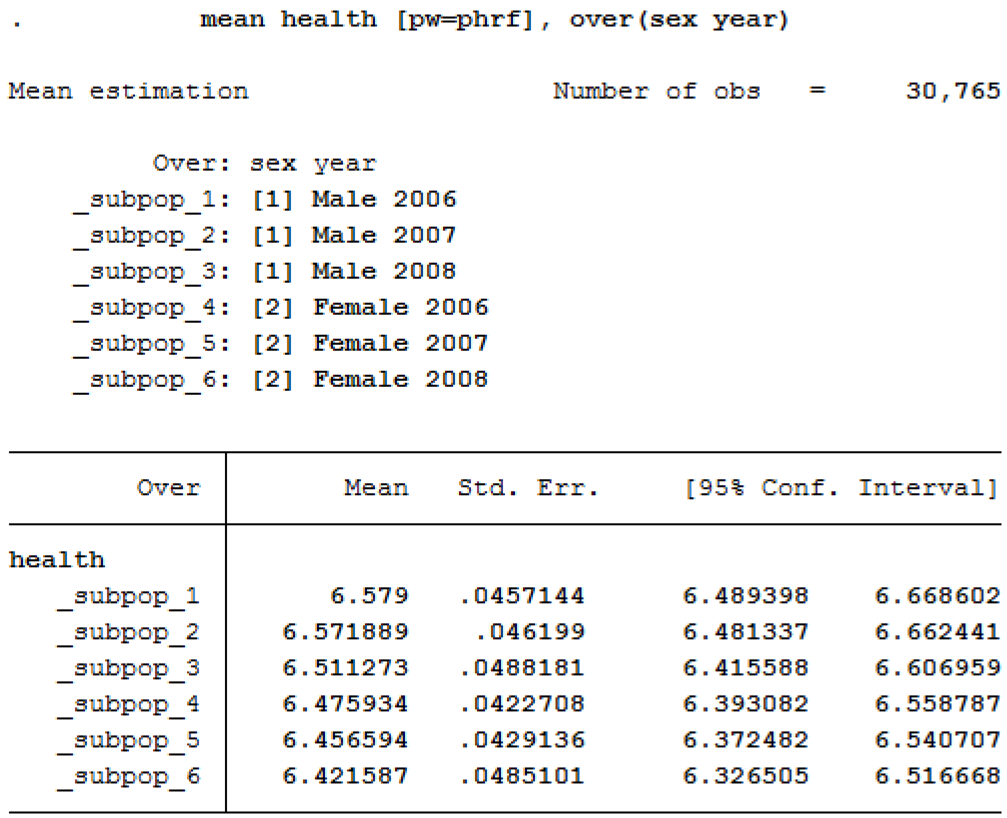
persnr	hhnr	wave	year	sex	smoke
12345	123	x	2006	m	yes
12345	123	y	2007	m	.
12345	123	z	2008	m	yes

4. Perform analyses based on the data. Try to answer the following questions:

a. Has men's and women's average satisfaction with health changed over the three years?

Satisfaction with health was measured on a scale from 1 to 10, with a value of 10 representing the highest possible level of satisfaction. To compare the average satisfaction with health between women and men, you should display the mean value for both sexes. The mean value is displayed weighted here.

```
*a) Has the average satisfaction with men's health and women changed
*   over the three years?

          mean health [pw=phrf], over(sex year)
```



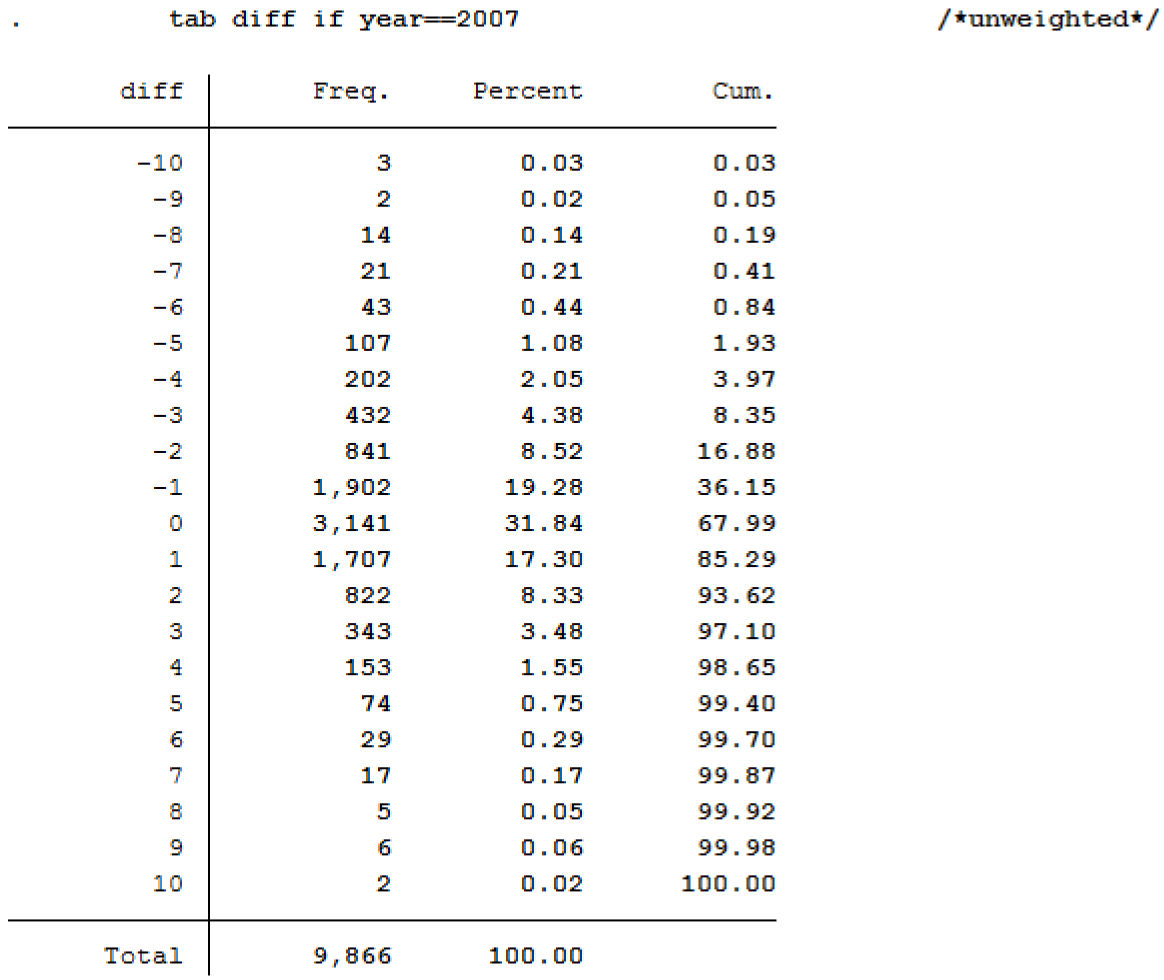
```
.           mean health [pw=phrf], over(sex year)

Mean estimation                        Number of obs    =     30,765

          Over: sex year
     _subpop_1: [1] Male 2006
     _subpop_2: [1] Male 2007
     _subpop_3: [1] Male 2008
     _subpop_4: [2] Female 2006
     _subpop_5: [2] Female 2007
     _subpop_6: [2] Female 2008

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
health       |
   _subpop_1 |      6.579   .0457144      6.489398    6.668602
   _subpop_2 |   6.571889    .046199      6.481337    6.662441
   _subpop_3 |   6.511273   .0488181      6.415588    6.606959
   _subpop_4 |   6.475934   .0422708      6.393082    6.558787
   _subpop_5 |   6.456594   .0429136      6.372482    6.540707
   _subpop_6 |   6.421587   .0485101      6.326505    6.516668
--------------------------------------------------------------
```

The output shows the average values for men and women for all three years. The first three values show men's average satisfaction with health between 2006 and 2008, while the last three values show women's average satisfaction with health.

b. What is the proportion of people for whom health satisfaction has increased from 2006 to 2007?

To answer this question, the difference between 2006 and 2007 should be displayed. You should make sure that the analysis is conducted only within one persnr (individual identifier) and only for satisfaction in the following year.

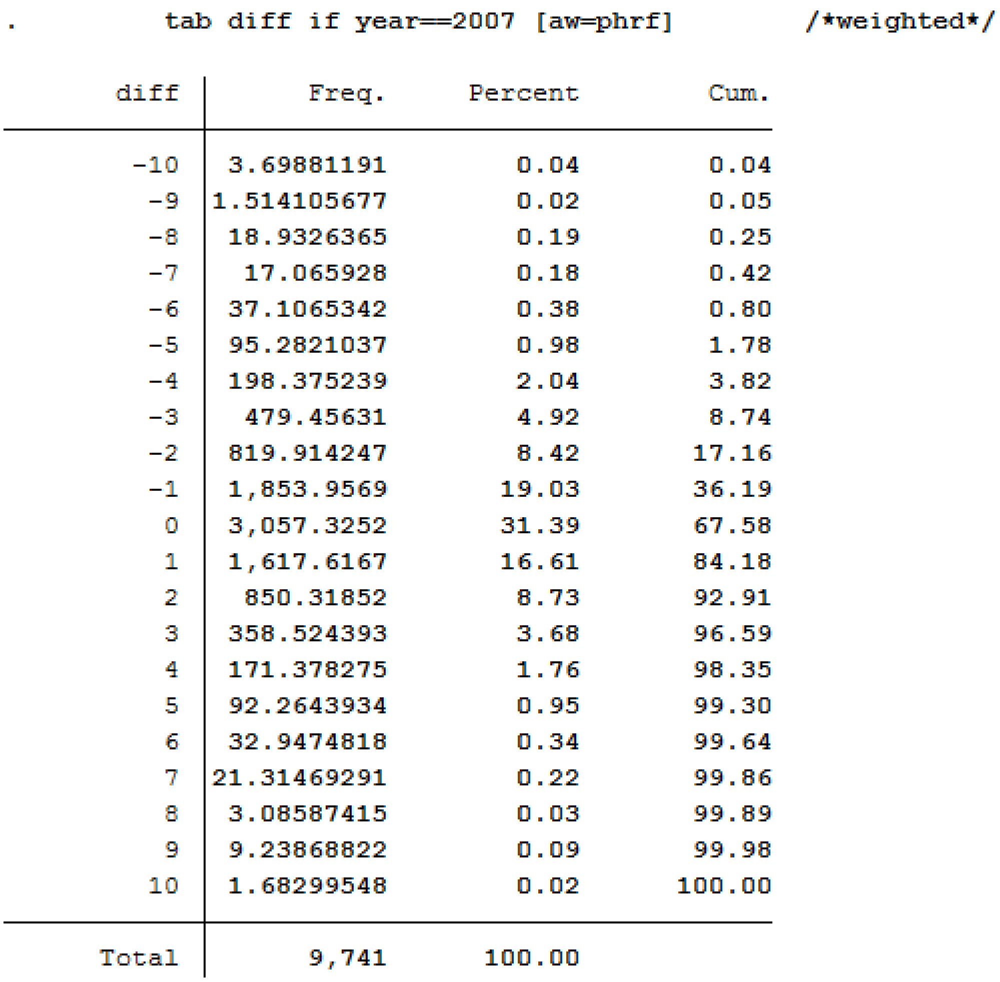
```
*b) What is the proportion of people for whom health satisfaction has increased
*   from 2006 to 2007??
        sort persnr year
        gen diff=health-health[_n-1] if persnr==persnr[_n-1] & year==year[_n-1]+1
        tab diff if year==2007                                /*unweighted*/
```



```
.          tab diff if year==2007                                /*unweighted*/

       diff |      Freq.     Percent        Cum.
------------+-----------------------------------
        -10 |          3        0.03        0.03
         -9 |          2        0.02        0.05
         -8 |         14        0.14        0.19
         -7 |         21        0.21        0.41
         -6 |         43        0.44        0.84
         -5 |        107        1.08        1.93
         -4 |        202        2.05        3.97
         -3 |        432        4.38        8.35
         -2 |        841        8.52       16.88
         -1 |      1,902       19.28       36.15
          0 |      3,141       31.84       67.99
          1 |      1,707       17.30       85.29
          2 |        822        8.33       93.62
          3 |        343        3.48       97.10
          4 |        153        1.55       98.65
          5 |         74        0.75       99.40
          6 |         29        0.29       99.70
          7 |         17        0.17       99.87
          8 |          5        0.05       99.92
          9 |          6        0.06       99.98
         10 |          2        0.02      100.00
------------+-----------------------------------
      Total |      9,866      100.00
```

Since you have previously added the SOEP weighting factors to the dataset for your analysis, you should use the weighting for a representative analysis.

```
        tab diff if year==2007 [aw=phrf]        /*weighted*/
```



```
.         tab diff if year==2007 [aw=phrf]        /*weighted*/

       diff |      Freq.     Percent        Cum.
------------+-----------------------------------
        -10 |  3.69881191        0.04        0.04
         -9 |1.514105677        0.02        0.05
         -8 |  18.9326365        0.19        0.25
         -7 |   17.065928        0.18        0.42
         -6 |  37.1065342        0.38        0.80
         -5 |  95.2821037        0.98        1.78
         -4 |  198.375239        2.04        3.82
         -3 |   479.45631        4.92        8.74
         -2 |  819.914247        8.42       17.16
         -1 |  1,853.9569       19.03       36.19
          0 |  3,057.3252       31.39       67.58
          1 |  1,617.6167       16.61       84.18
          2 |   850.31852        8.73       92.91
          3 |  358.524393        3.68       96.59
          4 |  171.378275        1.76       98.35
          5 |  92.2643934        0.95       99.30
          6 |  32.9474818        0.34       99.64
          7 |21.31469291        0.22       99.86
          8 |  3.08587415        0.03       99.89
          9 |  9.23868822        0.09       99.98
         10 |  1.68299548        0.02      100.00
------------+-----------------------------------
      Total |       9,741      100.00
```

The values less than 0 show a deterioration in health satisfaction. The value 0 means constant health satisfaction, and all values above 0 show a positive change in satisfaction with their health. With a value of 10, it can be assumed that these people were interviewed for the first time in 2007 or 2008.

c. In what direction and how much has satisfaction with health changed from 2006 to 2008 among people who quit smoking after 2006?

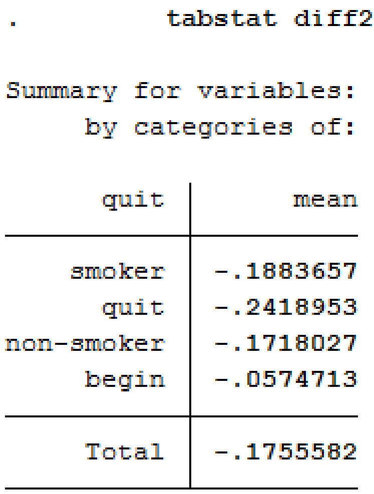
The procedure is similar to the previous question, except that the element “smoke yes/no” is added.

```
*c) In what direction and how much has satisfaction with
*   health changed from 2006 to 2008 among people who quit smoking after 2006?

        gen diff2=health-health[_n-2] if persnr==persnr[_n-2] & year==year[_n-2]+2 & year==2008
        gen quit=.
        replace quit=0 if smoke==1 & smoke[_n-2]==1 & persnr==persnr[_n-2] & year==year[_n-2]+2 & year==2008
        replace quit=1 if smoke==2 & smoke[_n-2]==1 & persnr==persnr[_n-2] & year==year[_n-2]+2 & year==2008
        replace quit=2 if smoke==2 & smoke[_n-2]==2 & persnr==persnr[_n-2] & year==year[_n-2]+2 & year==2008
```

(continues on next page)

(continued from previous page)

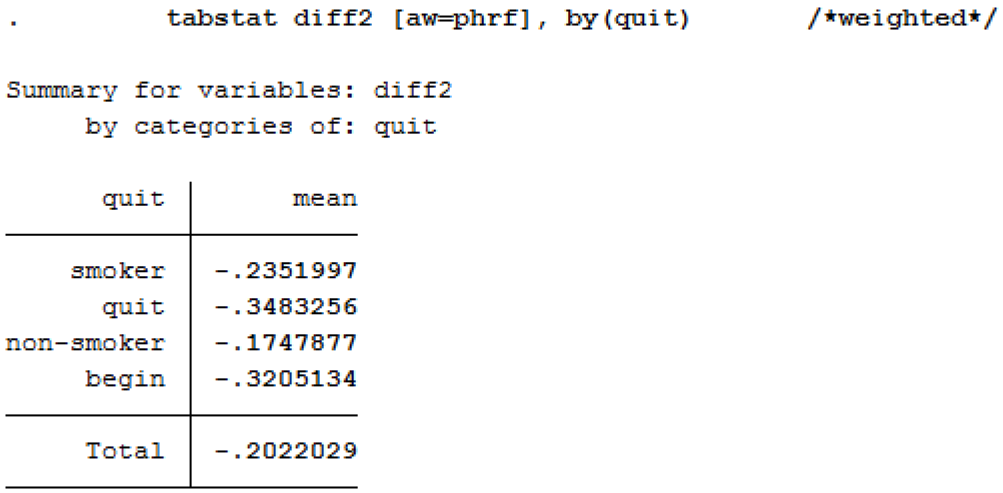
```
        replace quit=3 if smoke==1 & smoke[_n-2]==2 & persnr==persnr[_n-2] & year==year[_n-2]+2 & year==2008
        label define quit 0 "smoker" 1 "quit" 2 "non-smoker" 3 "begin"
        label values quit quit
        tabstat diff2, by(quit)
```



```
.         tabstat diff2, by(quit)

Summary for variables: diff2
     by categories of: quit

      quit |      mean
-----------+----------
    smoker | -.1883657
      quit | -.2418953
non-smoker | -.1718027
     begin | -.0574713
-----------+----------
     Total | -.1755582
----------------------
```

To obtain a weighted mean value, address the analysis weight after the generated variable.

```
        tabstat diff2 [aw=phrf], by(quit)        /*weighted*/
```



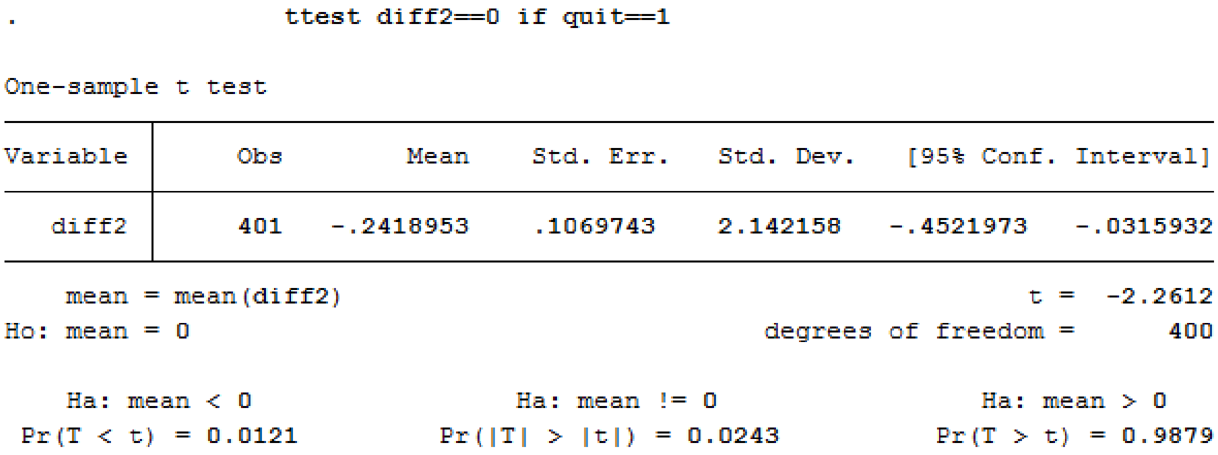
```
.         tabstat diff2 [aw=phrf], by(quit)        /*weighted*/

Summary for variables: diff2
     by categories of: quit

      quit |      mean
-----------+----------
    smoker | -.2351997
      quit | -.3483256
non-smoker | -.1747877
     begin | -.3205134
-----------+----------
     Total | -.2022029
----------------------
```

This illustration shows the mean of the health variable under the condition of the quit variable that we generated beforehand. With a mean of -0.24 (weighted -0.35), the biggest change in health satisfaction is seen in people who quit smoking after 2006. For example, if a person smoked in 2006 and indicated a satisfaction value of 8, the person indicates a satisfaction value of 7.76 after he/she stopped smoking in 2008. So you can assume that when a person stops smoking, their perceived health state deteriorates. Now we have to test if the assumption is correct.

d. Does quitting smoking make your health worse? To what extent could the result of the analysis "stop smoking" be distorted?

In order to establish a connection between health satisfaction and stopping smoking, one should use the t-test or to be more specific, the one-sample t-test. It checks whether the mean value of a sample deviates significantly from a known expected value (specified in the null hypothesis).

```
*d) Does quitting smoking make your health worse? To what extent can the
*   result of the analysis "Stop smoking" be distorted?

        * Notes: So far we have not tested whether the difference is statistically significant
                ttest diff2==0 if quit==1
```



```
.                 ttest diff2==0 if quit==1

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   diff2 |     401   -.2418953    .1069743    2.142158   -.4521973   -.0315932
------------------------------------------------------------------------------
    mean = mean(diff2)                                            t =  -2.2612
Ho: mean = 0                                     degrees of freedom =      400

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0121         Pr(|T| > |t|) = 0.0243          Pr(T > t) = 0.9879
```

H0 Hypothesis: If one stops smoking, it has no effect on health.

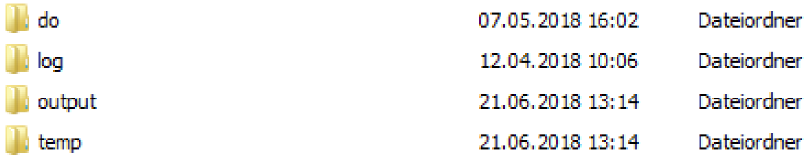
For this test we assume a 95% probability. What we want to check now is whether the H0 hypothesis can be rejected or not. If you look at the output of the test, you first see the mean value of 1 (quit smoking) of the variable quit. The last line of the output shows the significance level. If it falls below the value 0.05, one can speak of a statistically significant result. In our example, the null hypothesis can be discarded because its value is less than 0.05 percent. So quitting smoking has a significant impact on a person's perceived health.

Last change: Sep 26, 2022

6.5 Working with harmonized Variables

This exercise shows you how to work effectively with versioned and harmonized SOEP variables. Please note that the new SOEP versioning and harmonizing concept has only been available since SOEP-Core v34 and only applies to the original SOEP-Core data in long format.

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

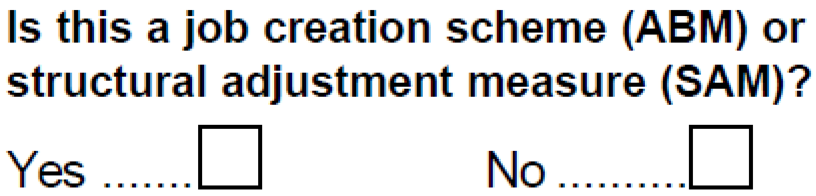
These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
************************************************
* Set relative paths to the working directory
************************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-gen\consolidated\soep-long\soep.v34"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to your ordered data.

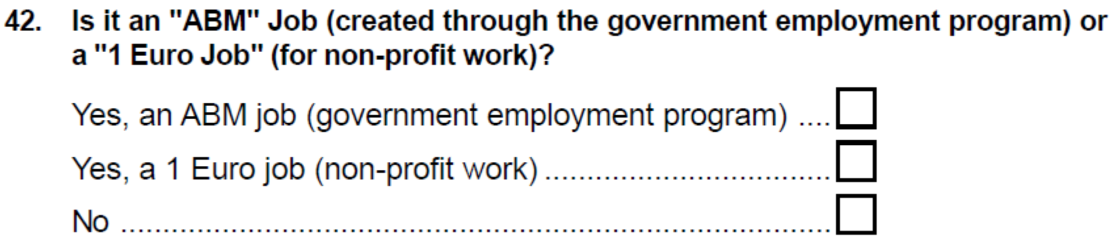
1.) Differences in Response Options

Variables are versioned and harmonized because the response options have changed over time.



Is this a job creation scheme (ABM) or structural adjustment measure (SAM)?

Yes .... □ No .... □



42. Is it an "ABM" Job (created through the government employment program) or a "1 Euro Job" (for non-profit work)?

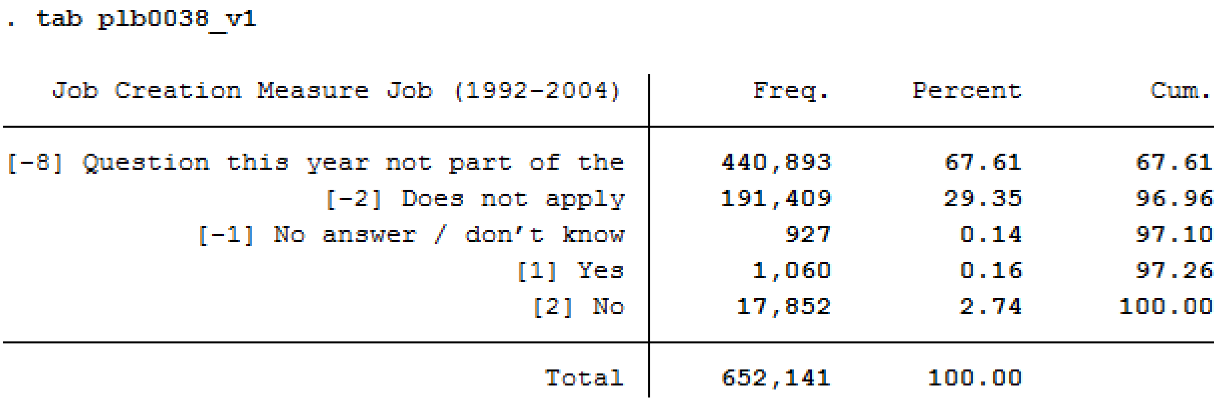
Yes, an ABM job (government employment program) .... □

Yes, a 1 Euro job (non-profit work) .... □

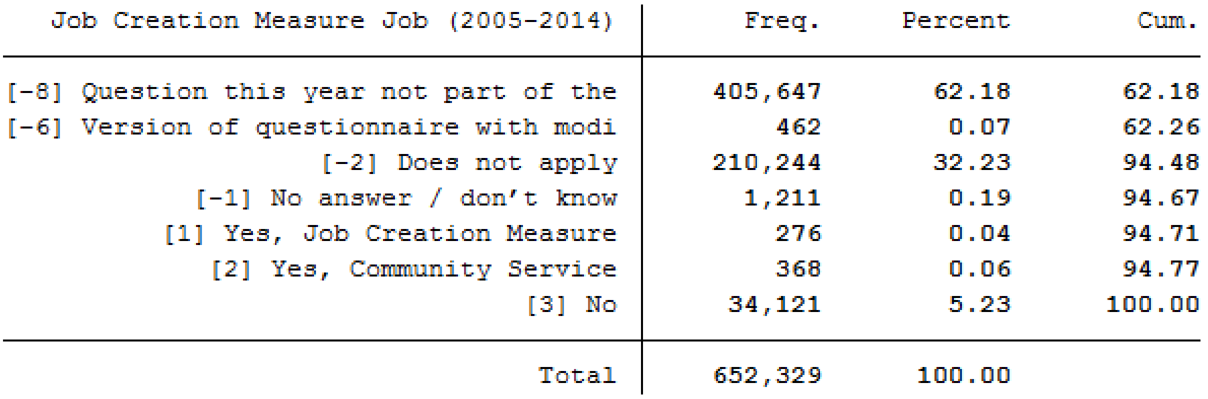
No .... □

The variable plb0038_v1 was obtained from a simple yes/no question between 1992 and 2004. Since 2005, new response options have been added. The individual questionnaires from 2004 and 2005 show these differences. Through the versioning of the variable plb0038, this difference is recognizable to the data user when tabulating the variable. The variable label also shows the beginning and end of the period in which the question was asked differently.

```
use "$MY_IN_PATH\pl.dta"
tab plb0038_v1
tab plb0038_v2
```



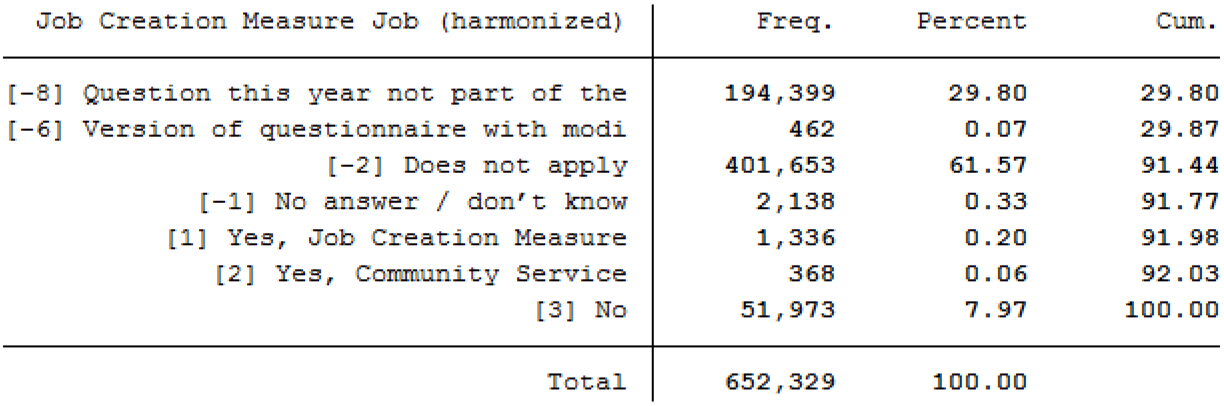
```
. tab plb0038_v1

   Job Creation Measure Job (1992-2004) |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[-8] Question this year not part of the |    440,893       67.61       67.61
                    [-2] Does not apply |    191,409       29.35       96.96
              [-1] No answer / don't know |        927        0.14       97.10
                               [1] Yes |      1,060        0.16       97.26
                                [2] No |     17,852        2.74      100.00
----------------------------------------+-----------------------------------
                                 Total |    652,141      100.00
```



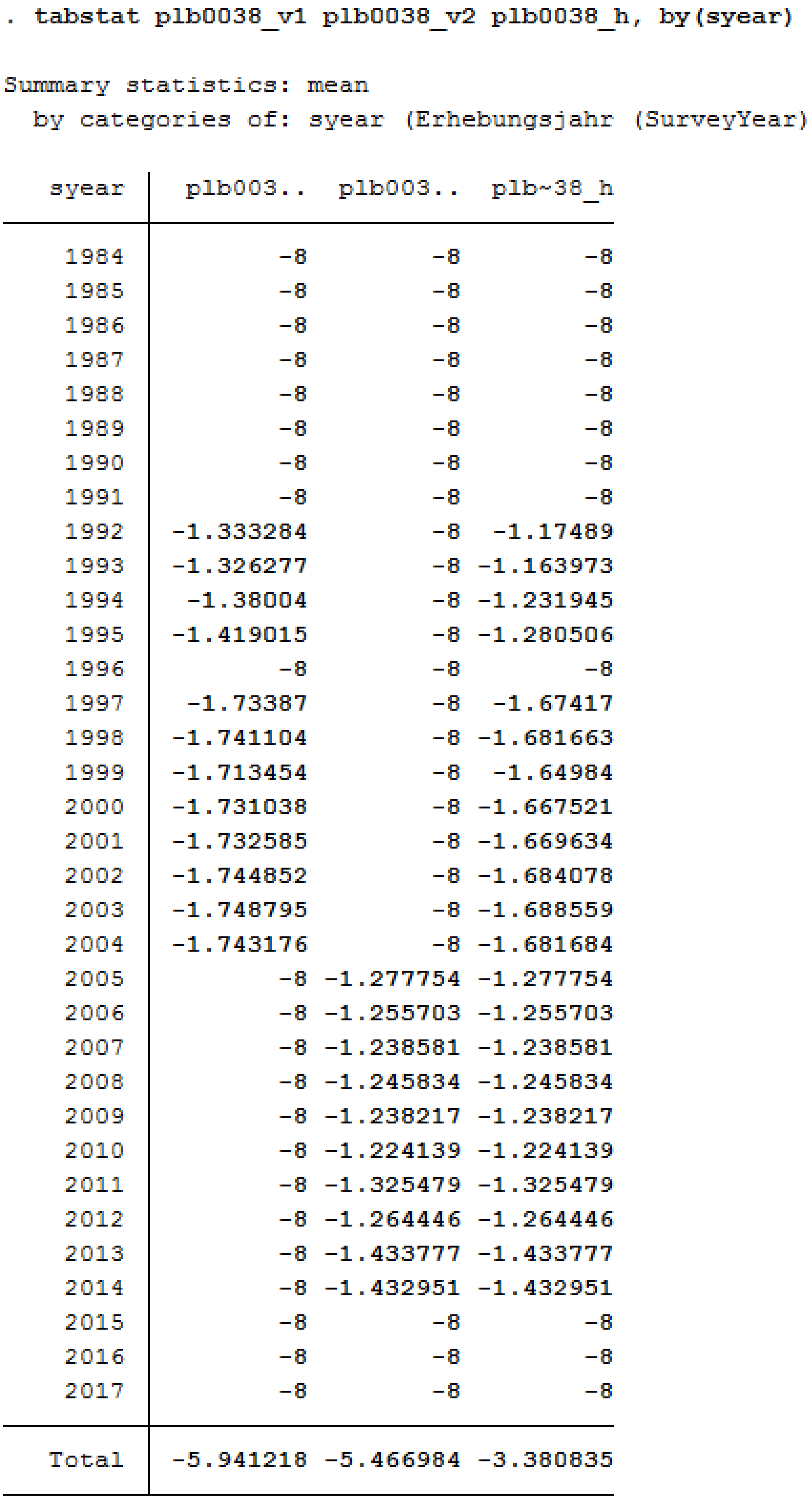
```
   Job Creation Measure Job (2005-2014) |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[-8] Question this year not part of the |    405,647       62.18       62.18
[-6] Version of questionnaire with modi |        462        0.07       62.26
                    [-2] Does not apply |    210,244       32.23       94.48
              [-1] No answer / don't know |      1,211        0.19       94.67
            [1] Yes, Job Creation Measure |        276        0.04       94.71
              [2] Yes, Community Service |        368        0.06       94.77
                                [3] No |     34,121        5.23      100.00
----------------------------------------+-----------------------------------
                                 Total |    652,329      100.00
```

The variable plb0038_v1 is recoded during the harmonization process and written into a new variable, plb0038_h, together with plb0038_v2. The harmonized version of the variable should cover the survey period from 1992 to 2014 and should be usable.

```
tab plb0038_h
tabstat plb0038_v1 plb0038_v2 plb0038_h, by(syear)
```



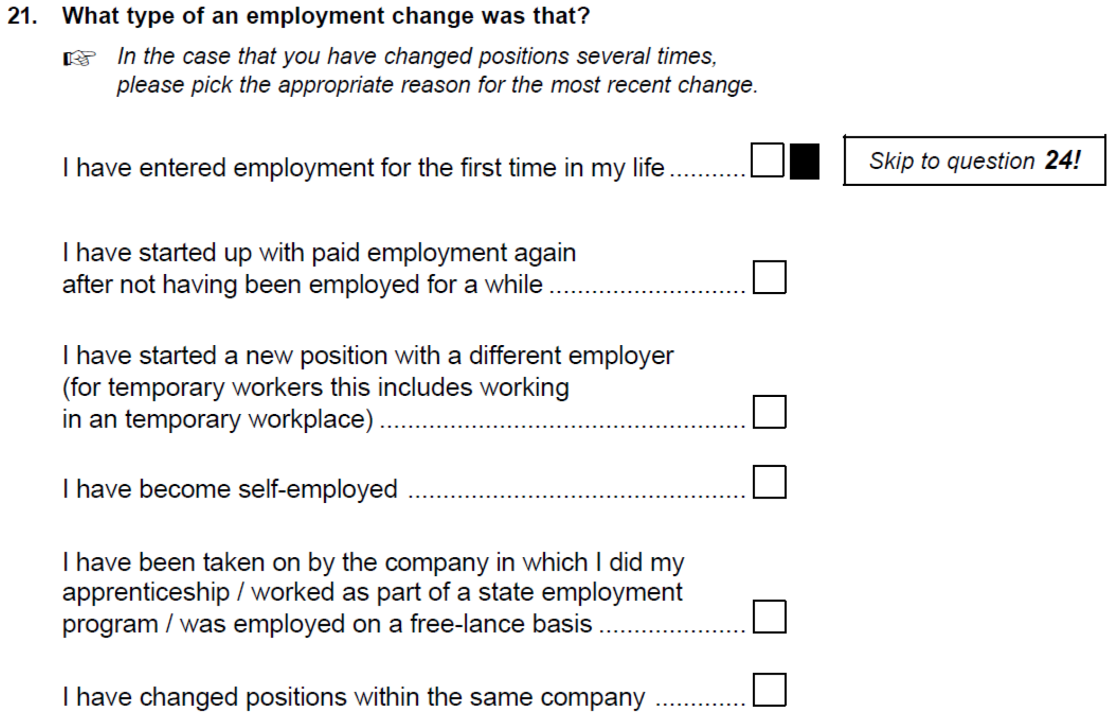
```
 Job Creation Measure Job (harmonized) |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
[-8] Question this year not part of the |    194,399       29.80       29.80
[-6] Version of questionnaire with modi |        462        0.07       29.87
                    [-2] Does not apply |    401,653       61.57       91.44
            [-1] No answer / don't know |      2,138        0.33       91.77
              [1] Yes, Job Creation Measure |      1,336        0.20       91.98
              [2] Yes, Community Service |        368        0.06       92.03
                                 [3] No |     51,973        7.97      100.00
---------------------------------------+-----------------------------------
                                  Total |    652,329      100.00
```



```
. tabstat plb0038_v1 plb0038_v2 plb0038_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear)

   syear |  plb003..  plb003..  plb~38_h
---------+------------------------------
    1984 |        -8        -8        -8
    1985 |        -8        -8        -8
    1986 |        -8        -8        -8
    1987 |        -8        -8        -8
    1988 |        -8        -8        -8
    1989 |        -8        -8        -8
    1990 |        -8        -8        -8
    1991 |        -8        -8        -8
    1992 | -1.333284        -8  -1.17489
    1993 | -1.326277        -8 -1.163973
    1994 |  -1.38004        -8 -1.231945
    1995 | -1.419015        -8 -1.280506
    1996 |        -8        -8        -8
    1997 |  -1.73387        -8  -1.67417
    1998 | -1.741104        -8 -1.681663
    1999 | -1.713454        -8  -1.64984
    2000 | -1.731038        -8 -1.667521
    2001 | -1.732585        -8 -1.669634
    2002 | -1.744852        -8 -1.684078
    2003 | -1.748795        -8 -1.688559
    2004 | -1.743176        -8 -1.681684
    2005 |        -8 -1.277754 -1.277754
    2006 |        -8 -1.255703 -1.255703
    2007 |        -8 -1.238581 -1.238581
    2008 |        -8 -1.245834 -1.245834
    2009 |        -8 -1.238217 -1.238217
    2010 |        -8 -1.224139 -1.224139
    2011 |        -8 -1.325479 -1.325479
    2012 |        -8 -1.264446 -1.264446
    2013 |        -8 -1.433777 -1.433777
    2014 |        -8 -1.432951 -1.432951
    2015 |        -8        -8        -8
    2016 |        -8        -8        -8
    2017 |        -8        -8        -8
---------+------------------------------
   Total | -5.941218 -5.466984 -3.380835
----------------------------------------
```

2.) Differences in Coding of Response Options

Variables are versioned and harmonized because the coding of the response options has changed over time. Since the values of certain response options can change, the various wave-specific variables cannot be integrated easily into a variable in long format. The variable must be appropriately harmonized to be useable.



21. What type of an employment change was that?

☞ *In the case that you have changed positions several times, please pick the appropriate reason for the most recent change.*

I have entered employment for the first time in my life .... □ ■ *Skip to question* ***24!***

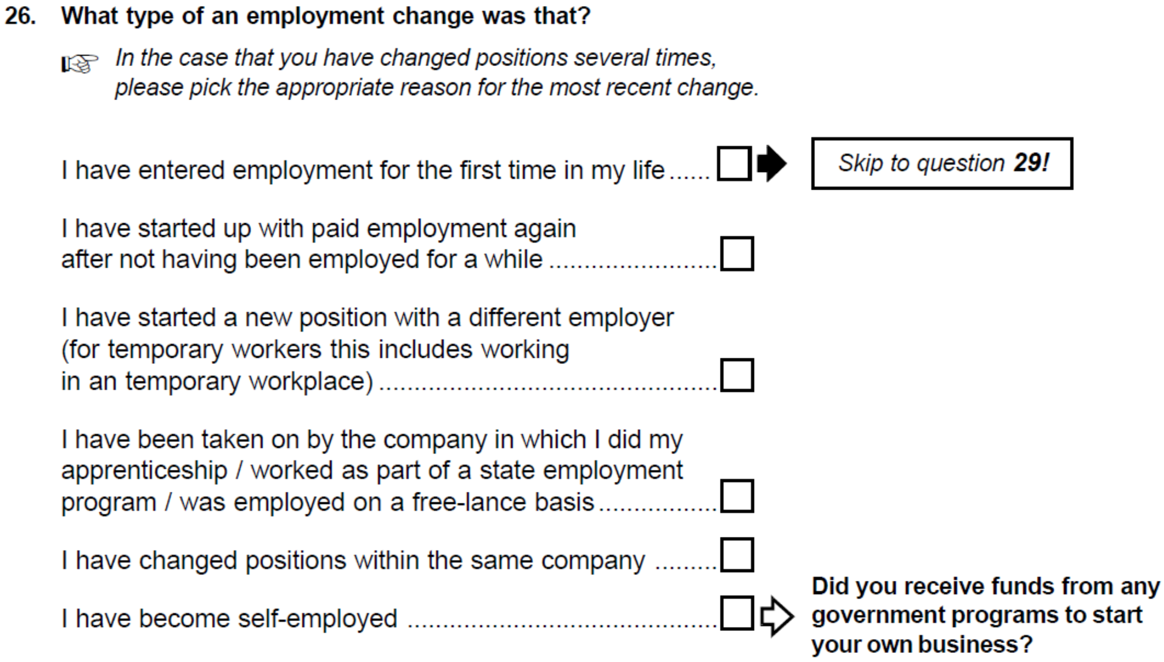
I have started up with paid employment again
after not having been employed for a while .... □

I have started a new position with a different employer
(for temporary workers this includes working
in an temporary workplace) .... □

I have become self-employed .... □

I have been taken on by the company in which I did my
apprenticeship / worked as part of a state employment
program / was employed on a free-lance basis .... □

I have changed positions within the same company .... □



26. What type of an employment change was that?

☞ *In the case that you have changed positions several times, please pick the appropriate reason for the most recent change.*

I have entered employment for the first time in my life ...... □ ➡ *Skip to question* ***29!***

I have started up with paid employment again
after not having been employed for a while .... □

I have started a new position with a different employer
(for temporary workers this includes working
in an temporary workplace) .... □

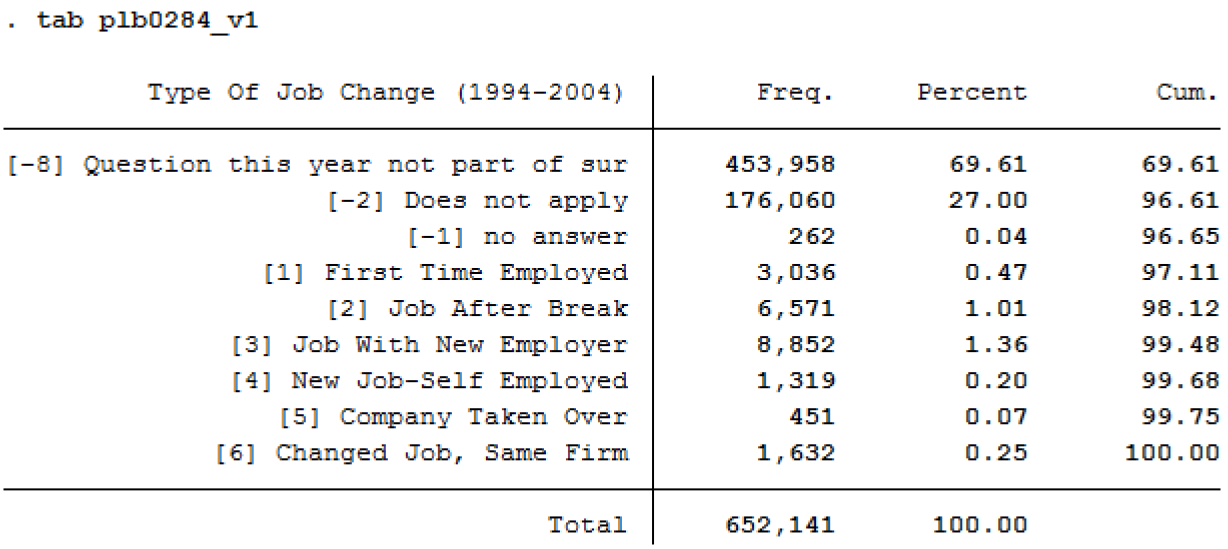
I have been taken on by the company in which I did my
apprenticeship / worked as part of a state employment
program / was employed on a free-lance basis .... □

I have changed positions within the same company .... □

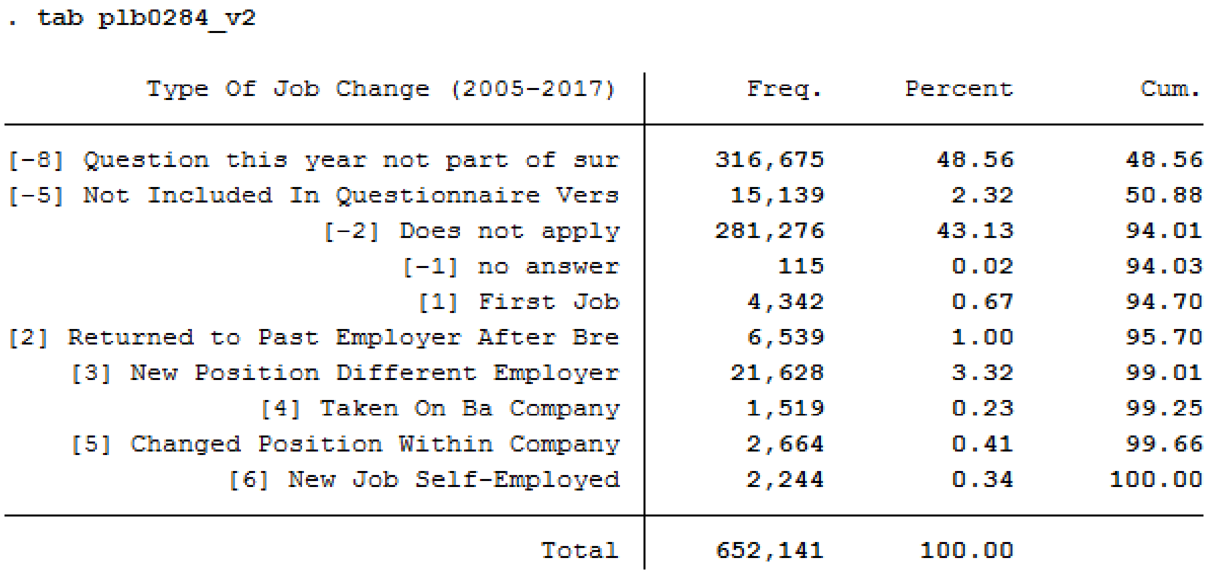
I have become self-employed .... □ ⇨ **Did you receive funds from any government programs to start your own business?**

From 1994 to 2004, the question about “job change” was asked in the individual questionnaire as a category question with six response options. The order of the response options changed in 2005.

```
tab plb0284_v1
tab plb0284_v2
```



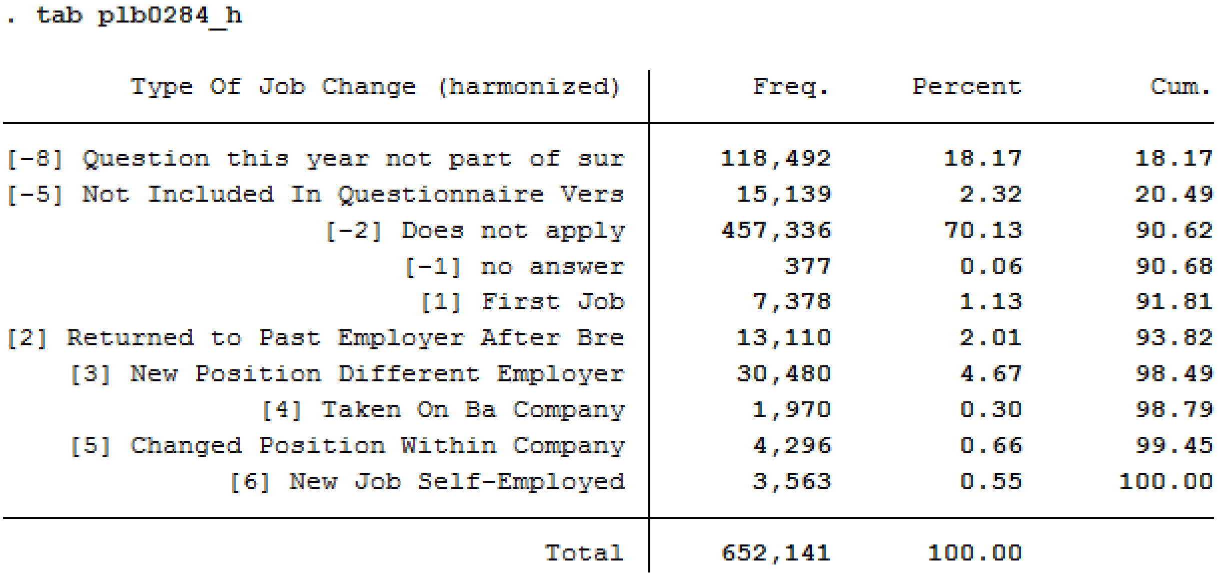
```
. tab plb0284_v1

            Type Of Job Change (1994-2004) |      Freq.     Percent        Cum.
-------------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    453,958       69.61       69.61
                    [-2] Does not apply |    176,060       27.00       96.61
                         [-1] no answer |        262        0.04       96.65
                [1] First Time Employed |      3,036        0.47       97.11
                    [2] Job After Break |      6,571        1.01       98.12
              [3] Job With New Employer |      8,852        1.36       99.48
              [4] New Job-Self Employed |      1,319        0.20       99.68
                 [5] Company Taken Over |        451        0.07       99.75
             [6] Changed Job, Same Firm |      1,632        0.25      100.00
-------------------------------------------+-----------------------------------
                                  Total |    652,141      100.00
```



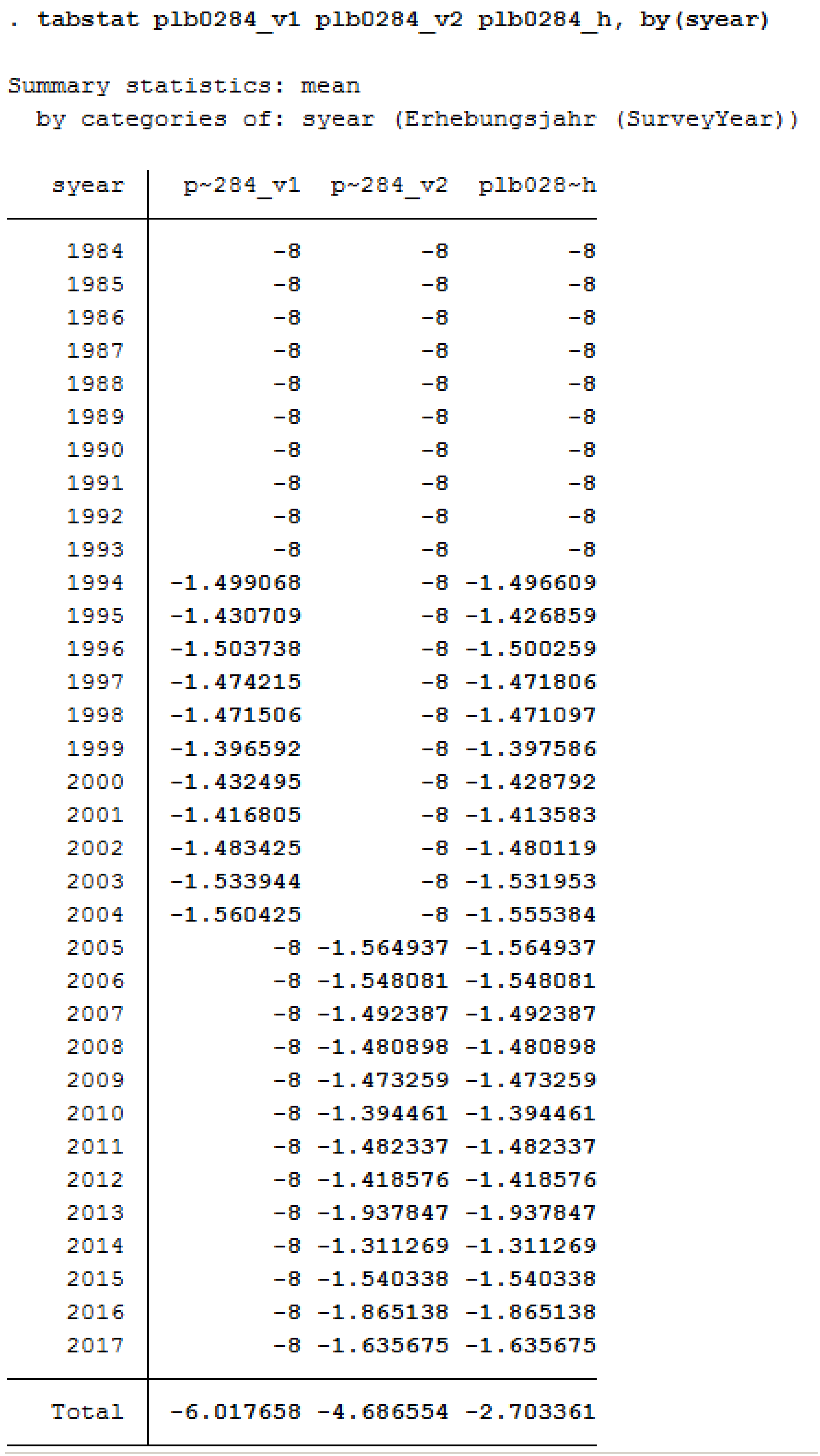
```
. tab plb0284_v2

            Type Of Job Change (2005-2017) |      Freq.     Percent        Cum.
-------------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    316,675       48.56       48.56
[-5] Not Included In Questionnaire Vers |     15,139        2.32       50.88
                    [-2] Does not apply |    281,276       43.13       94.01
                         [-1] no answer |        115        0.02       94.03
                          [1] First Job |      4,342        0.67       94.70
[2] Returned to Past Employer After Bre |      6,539        1.00       95.70
    [3] New Position Different Employer |     21,628        3.32       99.01
                [4] Taken On Ba Company |      1,519        0.23       99.25
    [5] Changed Position Within Company |      2,664        0.41       99.66
               [6] New Job Self-Employed |      2,244        0.34      100.00
-------------------------------------------+-----------------------------------
                                  Total |    652,141      100.00
```

In addition to the different order of the response options, the coding order also changed. The data are stored in the wave-specific “raw” datasets with different coding and are contained in the variables plb0284_v1 and plb0284_v2. To use the variable for all survey years, it is necessary to harmonize the different versions. The variable plb0284_v1 is recoded (recode (1=1)(2=2)(3=3)(4=6)(5=4)(6=5)) and then written together with plb0284_v2 as plb0284_h. The new variable plb0284_h is created by the harmonization process.

```
tab plb0284_h
tabstat plb0284_v1 plb0284_v2 plb0284_h, by(syear)
```



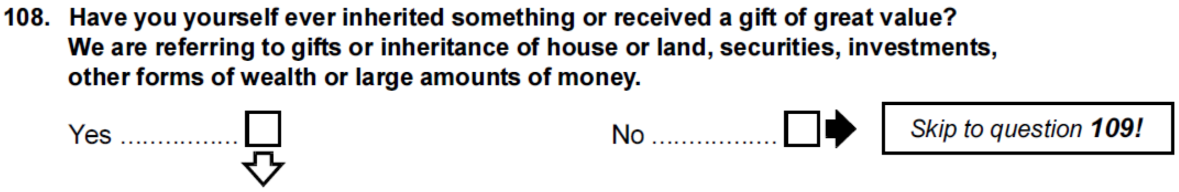
```
. tab plb0284_h

        Type Of Job Change (harmonized) |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    118,492       18.17       18.17
[-5] Not Included In Questionnaire Vers |     15,139        2.32       20.49
                    [-2] Does not apply |    457,336       70.13       90.62
                         [-1] no answer |        377        0.06       90.68
                          [1] First Job |      7,378        1.13       91.81
[2] Returned to Past Employer After Bre |     13,110        2.01       93.82
    [3] New Position Different Employer |     30,480        4.67       98.49
                [4] Taken On Ba Company |      1,970        0.30       98.79
    [5] Changed Position Within Company |      4,296        0.66       99.45
              [6] New Job Self-Employed |      3,563        0.55      100.00
----------------------------------------+-----------------------------------
                                  Total |    652,141      100.00
```



```
. tabstat plb0284_v1 plb0284_v2 plb0284_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear))

   syear |  p~284_v1  p~284_v2  plb028~h
---------+------------------------------
    1984 |        -8        -8        -8
    1985 |        -8        -8        -8
    1986 |        -8        -8        -8
    1987 |        -8        -8        -8
    1988 |        -8        -8        -8
    1989 |        -8        -8        -8
    1990 |        -8        -8        -8
    1991 |        -8        -8        -8
    1992 |        -8        -8        -8
    1993 |        -8        -8        -8
    1994 | -1.499068        -8 -1.496609
    1995 | -1.430709        -8 -1.426859
    1996 | -1.503738        -8 -1.500259
    1997 | -1.474215        -8 -1.471806
    1998 | -1.471506        -8 -1.471097
    1999 | -1.396592        -8 -1.397586
    2000 | -1.432495        -8 -1.428792
    2001 | -1.416805        -8 -1.413583
    2002 | -1.483425        -8 -1.480119
    2003 | -1.533944        -8 -1.531953
    2004 | -1.560425        -8 -1.555384
    2005 |        -8 -1.564937 -1.564937
    2006 |        -8 -1.548081 -1.548081
    2007 |        -8 -1.492387 -1.492387
    2008 |        -8 -1.480898 -1.480898
    2009 |        -8 -1.473259 -1.473259
    2010 |        -8 -1.394461 -1.394461
    2011 |        -8 -1.482337 -1.482337
    2012 |        -8 -1.418576 -1.418576
    2013 |        -8 -1.937847 -1.937847
    2014 |        -8 -1.311269 -1.311269
    2015 |        -8 -1.540338 -1.540338
    2016 |        -8 -1.865138 -1.865138
    2017 |        -8 -1.635675 -1.635675
---------+------------------------------
   Total | -6.017658 -4.686554 -2.703361
----------------------------------------
```

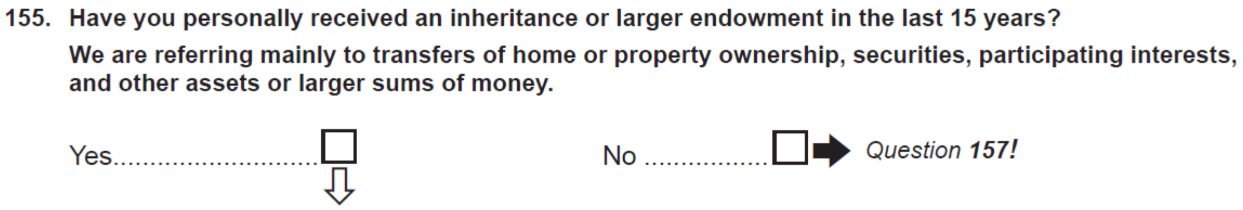
3.) Content Differences in the Questions.

Variables are versioned when questions were asked differently in different years but the content belongs together. If the content or wording of the question changes, the wave-specific variables cannot easily be integrated into a long variable.



108. Have you yourself ever inherited something or received a gift of great value?
We are referring to gifts or inheritance of house or land, securities, investments, other forms of wealth or large amounts of money.

Yes .... □ ⇩ No .... □ ➡ *Skip to question* ***109!***

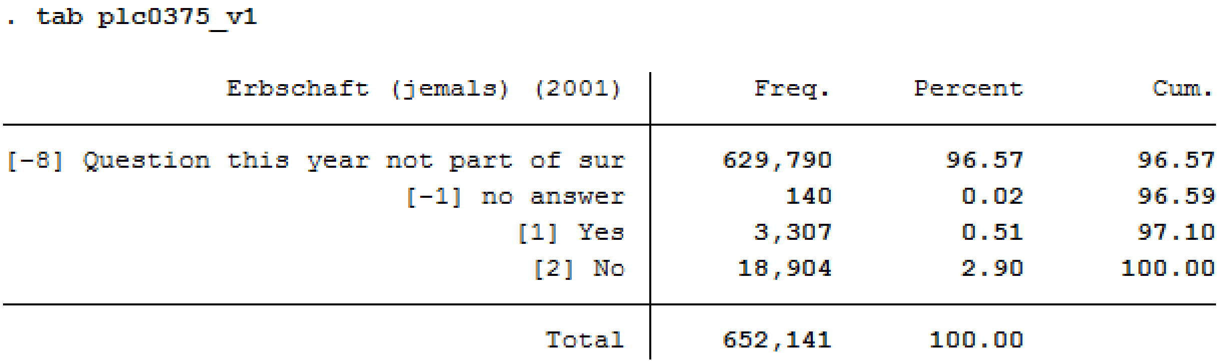


155. Have you personally received an inheritance or larger endowment in the last 15 years?
We are referring mainly to transfers of home or property ownership, securities, participating interests, and other assets or larger sums of money.

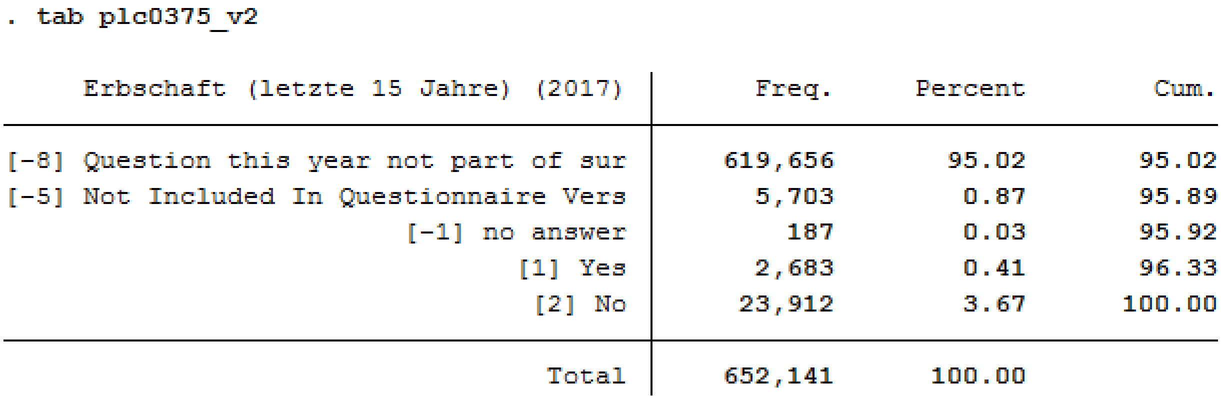
Yes............................ □ ⇩ No .... □ ➡ *Question* ***157!***

In the 2001 individual questionnaire, respondents were asked whether they had ever received an inheritance. In 2017, this question was worded differently: respondents were asked whether they had received an inheritance in the last 15 years. The questions are similar but cover different time periods. Therefore, the variable is not harmonized but made available as versioned variables. Data users have to decide whether or not to use the variables in the same way.

```
tab plc0375_v1
tab plc0375_v2
```



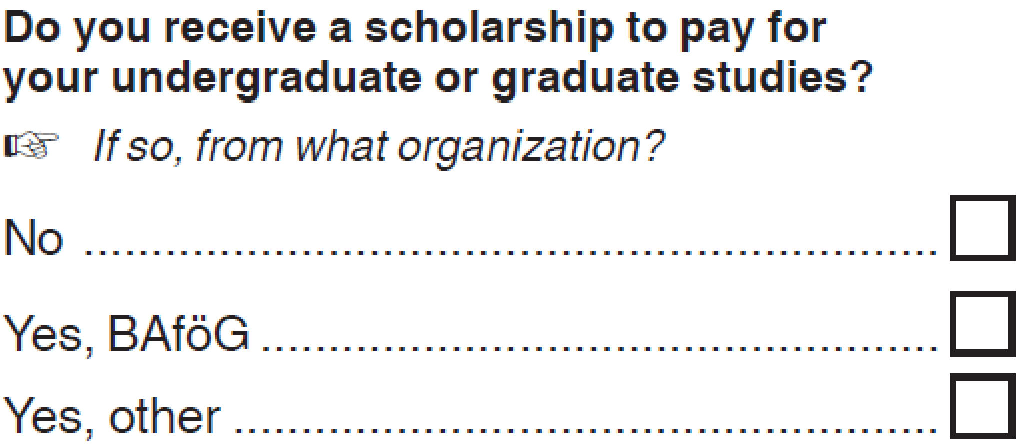
```
. tab plc0375_v1

         Erbschaft (jemals) (2001) |      Freq.     Percent        Cum.
-----------------------------------+-----------------------------------
[-8] Question this year not part of sur |    629,790       96.57       96.57
                    [-1] no answer |        140        0.02       96.59
                          [1] Yes |      3,307        0.51       97.10
                           [2] No |     18,904        2.90      100.00
-----------------------------------+-----------------------------------
                            Total |    652,141      100.00
```



```
. tab plc0375_v2

 Erbschaft (letzte 15 Jahre) (2017) |      Freq.     Percent        Cum.
------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    619,656       95.02       95.02
[-5] Not Included In Questionnaire Vers |      5,703        0.87       95.89
                     [-1] no answer |        187        0.03       95.92
                           [1] Yes |      2,683        0.41       96.33
                            [2] No |     23,912        3.67      100.00
------------------------------------+-----------------------------------
                             Total |    652,141      100.00
```

4.) Change of Question Type.

Variables are versioned and harmonized when questions were asked differently in different years, for example, first as a question with multiple response options and later as a question with a single response option. The possibility to provide multiple answers in certain years makes it difficult to integrate the wave-specific variables into a variable in long format.



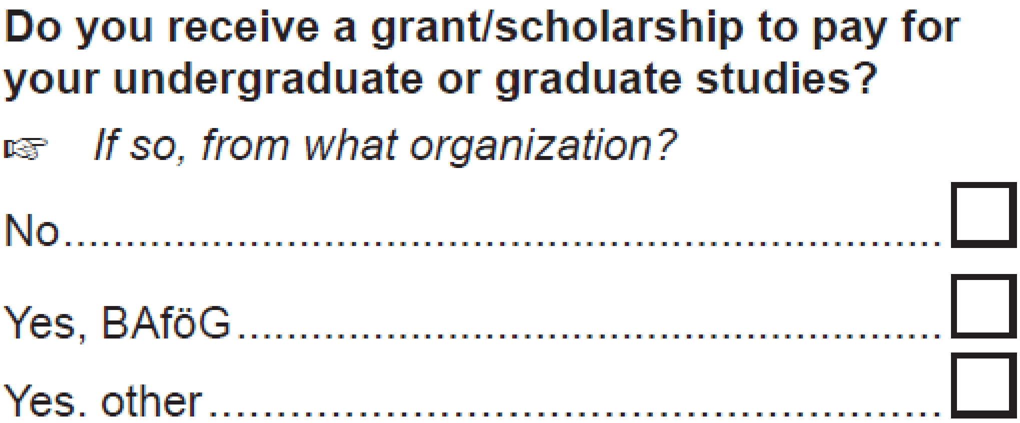
Do you receive a scholarship to pay for your undergraduate or graduate studies?

☞ *If so, from what organization?*

No .... □

Yes, BAföG .... □

Yes, other .... □



Do you receive a grant/scholarship to pay for your undergraduate or graduate studies?

☞ *If so, from what organization?*

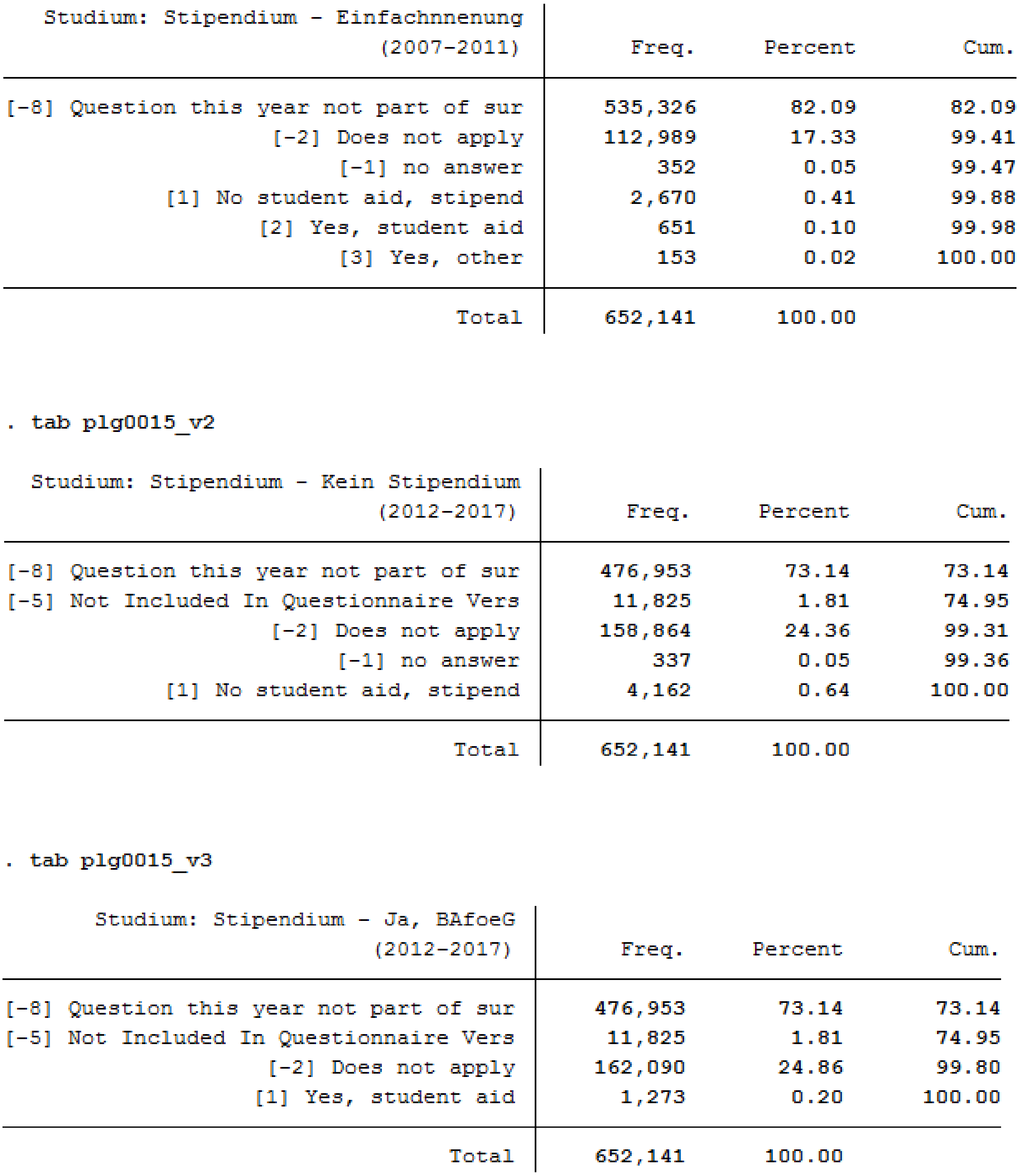
No... □

Yes, BAföG... □

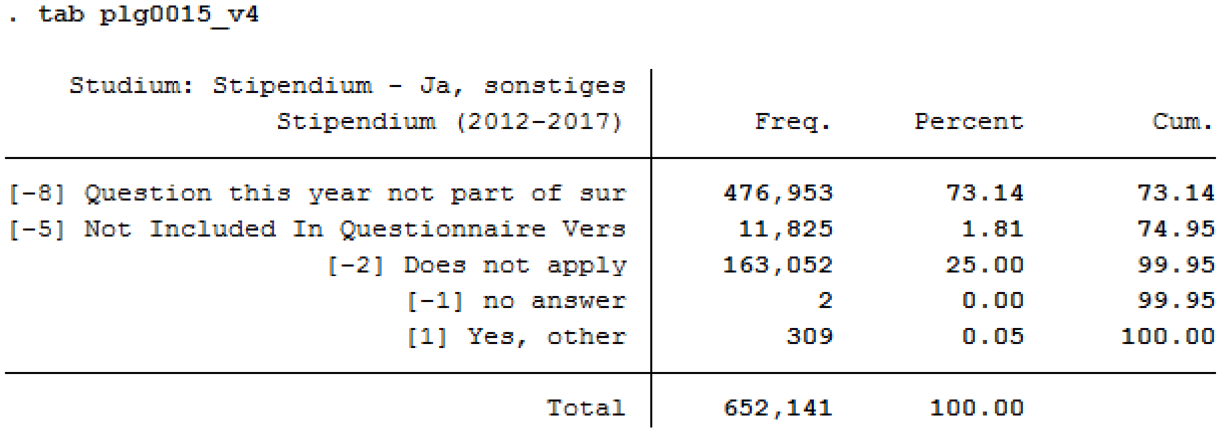
Yes. other... □

When comparing the question on scholarships in the individual questionnaires from 2011 and 2012, it appears that there should be no differences in the variables. Nevertheless, the two questions seem to have been asked differently and stored differently in the raw datasets. This results in several versioned variables.

```
tab plg0015_v1
tab plg0015_v2
tab plg0015_v3
tab plg0015_v4
```



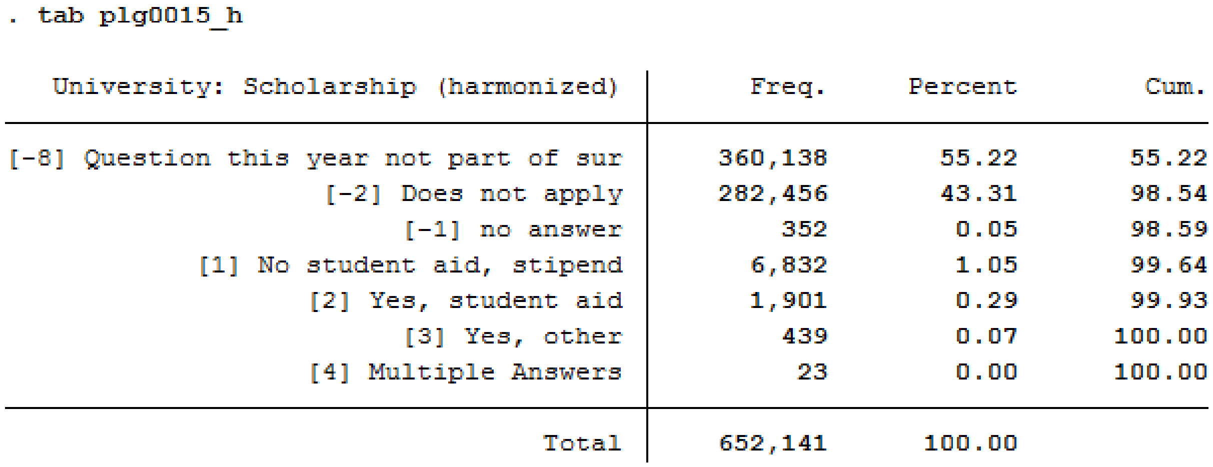
```
   Studium: Stipendium - Einfachnnenung |
                            (2007-2011) |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    535,326       82.09       82.09
                    [-2] Does not apply |    112,989       17.33       99.41
                         [-1] no answer |        352        0.05       99.47
           [1] No student aid, stipend |      2,670        0.41       99.88
                  [2] Yes, student aid |        651        0.10       99.98
                        [3] Yes, other |        153        0.02      100.00
----------------------------------------+-----------------------------------
                                  Total |    652,141      100.00


. tab plg0015_v2

  Studium: Stipendium - Kein Stipendium |
                            (2012-2017) |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    476,953       73.14       73.14
[-5] Not Included In Questionnaire Vers |     11,825        1.81       74.95
                    [-2] Does not apply |    158,864       24.36       99.31
                         [-1] no answer |        337        0.05       99.36
           [1] No student aid, stipend |      4,162        0.64      100.00
----------------------------------------+-----------------------------------
                                  Total |    652,141      100.00


. tab plg0015_v3

       Studium: Stipendium - Ja, BAfoeG |
                            (2012-2017) |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    476,953       73.14       73.14
[-5] Not Included In Questionnaire Vers |     11,825        1.81       74.95
                    [-2] Does not apply |    162,090       24.86       99.80
                  [1] Yes, student aid |      1,273        0.20      100.00
----------------------------------------+-----------------------------------
                                  Total |    652,141      100.00
```



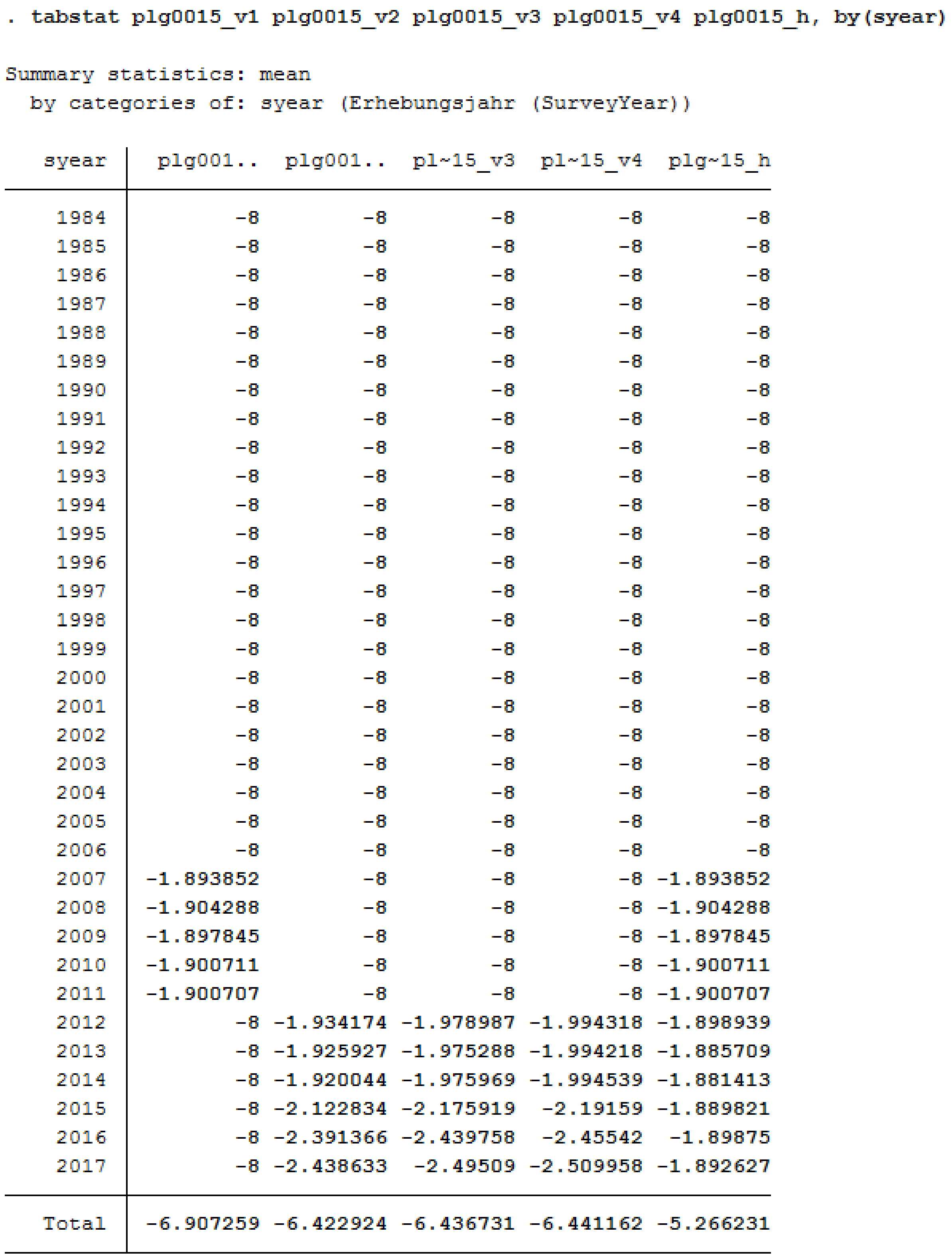
```
. tab plg0015_v4

     Studium: Stipendium - Ja, sonstiges |
                   Stipendium (2012-2017) |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    476,953       73.14       73.14
[-5] Not Included In Questionnaire Vers |     11,825        1.81       74.95
                    [-2] Does not apply |    163,052       25.00       99.95
                         [-1] no answer |          2        0.00       99.95
                         [1] Yes, other |        309        0.05      100.00
-----------------------------------------+-----------------------------------
                                   Total |    652,141      100.00
```

As you can see, the variable was asked from 2007 to 2011 as a category question with three response options. As a result, respondents could only give one answer. Since 2012, the question has used binary items. It is quite possible that a respondent gave more than one answer. The harmonized version of the variable integrates the binary items from plg0015_v2, plg0015_v3, and plg0015_v4 into the harmonized version plg0015_h. The coding of the variable plg0015_v1 is used as the generation framework. In addition, the harmonization proposal takes into account the problematic multiple answers with the value four.

```
plg0015_h
tabstat plg0015_v1 plg0015_v2 plg0015_v3 plg0015_v4 plg0015_h, by(syear)
```



```
. tab plg0015_h

    University: Scholarship (harmonized) |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
[-8] Question this year not part of sur |    360,138       55.22       55.22
                    [-2] Does not apply |    282,456       43.31       98.54
                         [-1] no answer |        352        0.05       98.59
              [1] No student aid, stipend |      6,832        1.05       99.64
                   [2] Yes, student aid |      1,901        0.29       99.93
                         [3] Yes, other |        439        0.07      100.00
                   [4] Multiple Answers |         23        0.00      100.00
-----------------------------------------+-----------------------------------
                                   Total |    652,141      100.00
```

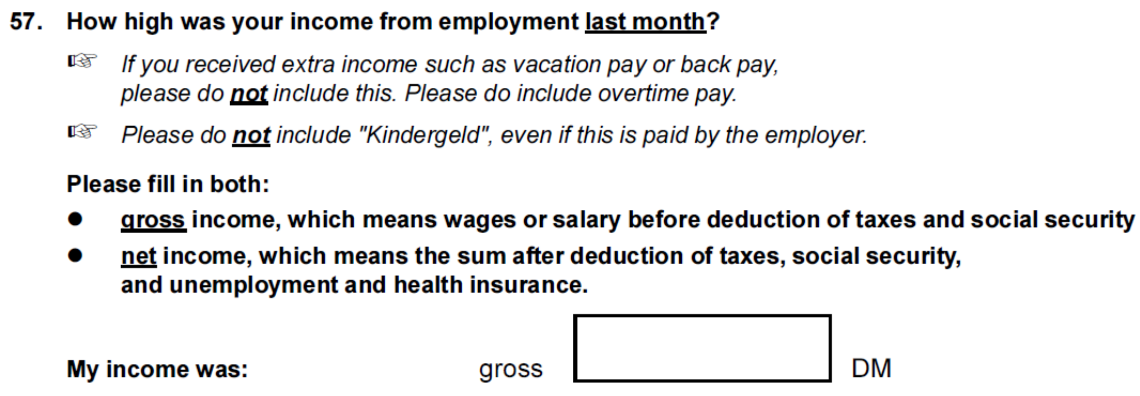


```
. tabstat plg0015_v1 plg0015_v2 plg0015_v3 plg0015_v4 plg0015_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear))

   syear |  plg001..  plg001..  pl~15_v3  pl~15_v4  plg~15_h
---------+--------------------------------------------------
    1984 |        -8        -8        -8        -8        -8
    1985 |        -8        -8        -8        -8        -8
    1986 |        -8        -8        -8        -8        -8
    1987 |        -8        -8        -8        -8        -8
    1988 |        -8        -8        -8        -8        -8
    1989 |        -8        -8        -8        -8        -8
    1990 |        -8        -8        -8        -8        -8
    1991 |        -8        -8        -8        -8        -8
    1992 |        -8        -8        -8        -8        -8
    1993 |        -8        -8        -8        -8        -8
    1994 |        -8        -8        -8        -8        -8
    1995 |        -8        -8        -8        -8        -8
    1996 |        -8        -8        -8        -8        -8
    1997 |        -8        -8        -8        -8        -8
    1998 |        -8        -8        -8        -8        -8
    1999 |        -8        -8        -8        -8        -8
    2000 |        -8        -8        -8        -8        -8
    2001 |        -8        -8        -8        -8        -8
    2002 |        -8        -8        -8        -8        -8
    2003 |        -8        -8        -8        -8        -8
    2004 |        -8        -8        -8        -8        -8
    2005 |        -8        -8        -8        -8        -8
    2006 |        -8        -8        -8        -8        -8
    2007 | -1.893852        -8        -8        -8 -1.893852
    2008 | -1.904288        -8        -8        -8 -1.904288
    2009 | -1.897845        -8        -8        -8 -1.897845
    2010 | -1.900711        -8        -8        -8 -1.900711
    2011 | -1.900707        -8        -8        -8 -1.900707
    2012 |        -8 -1.934174 -1.978987 -1.994318 -1.898939
    2013 |        -8 -1.925927 -1.975288 -1.994218 -1.885709
    2014 |        -8 -1.920044 -1.975969 -1.994539 -1.881413
    2015 |        -8 -2.122834 -2.175919  -2.19159 -1.889821
    2016 |        -8 -2.391366 -2.439758  -2.45542  -1.89875
    2017 |        -8 -2.438633  -2.49509 -2.509958 -1.892627
---------+--------------------------------------------------
   Total | -6.907259 -6.422924 -6.436731 -6.441162 -5.266231
------------------------------------------------------------
```

5.) Euro harmonisation

Variables are versioned and harmonized because they are metric and were asked as DM amounts before the introduction of the euro. For the long version of the variable, metric variables based on different currencies in different years are harmonized as euro amounts.

Most of the variables harmonized in the long datasets are amounts of money. Before the introduction of the euro, such information was collected in DM.



57. How high was your income from employment last month?

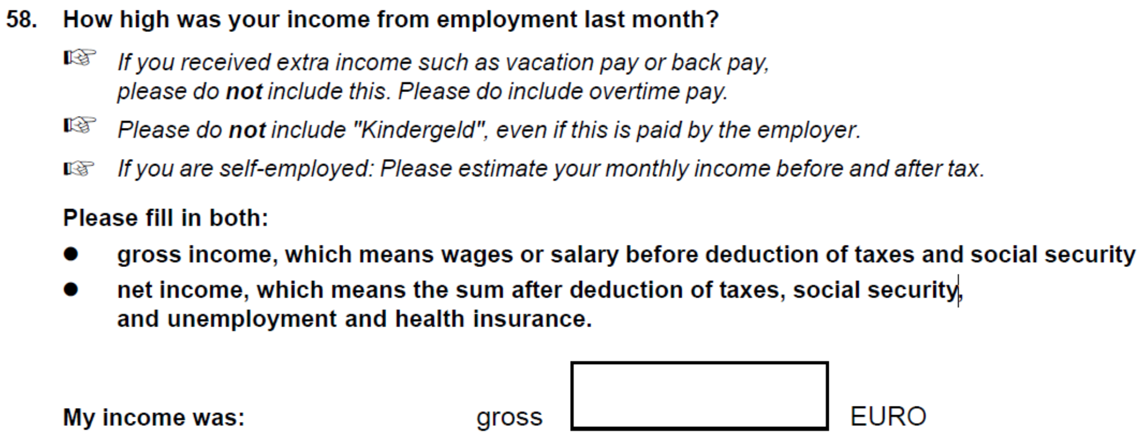
☞ *If you received extra income such as vacation pay or back pay, please do **not** include this. Please do include overtime pay.*

☞ *Please do **not** include "Kindergeld", even if this is paid by the employer.*

Please fill in both:

- **gross income, which means wages or salary before deduction of taxes and social security**
- **net income, which means the sum after deduction of taxes, social security, and unemployment and health insurance.**

My income was: gross ☐ DM



58. How high was your income from employment last month?

☞ *If you received extra income such as vacation pay or back pay, please do **not** include this. Please do include overtime pay.*

☞ *Please do **not** include "Kindergeld", even if this is paid by the employer.*

☞ *If you are self-employed: Please estimate your monthly income before and after tax.*

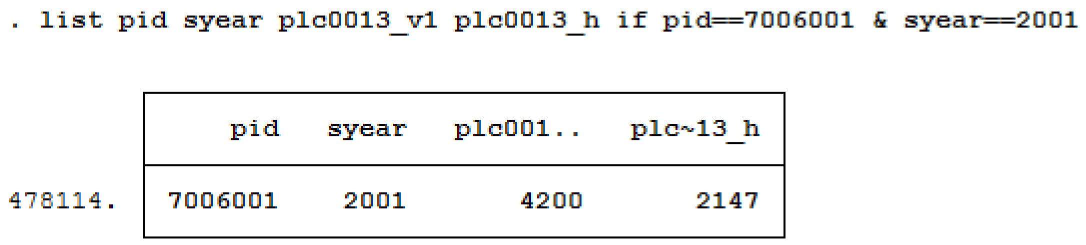
Please fill in both:

- **gross income, which means wages or salary before deduction of taxes and social security**
- **net income, which means the sum after deduction of taxes, social security, and unemployment and health insurance.**

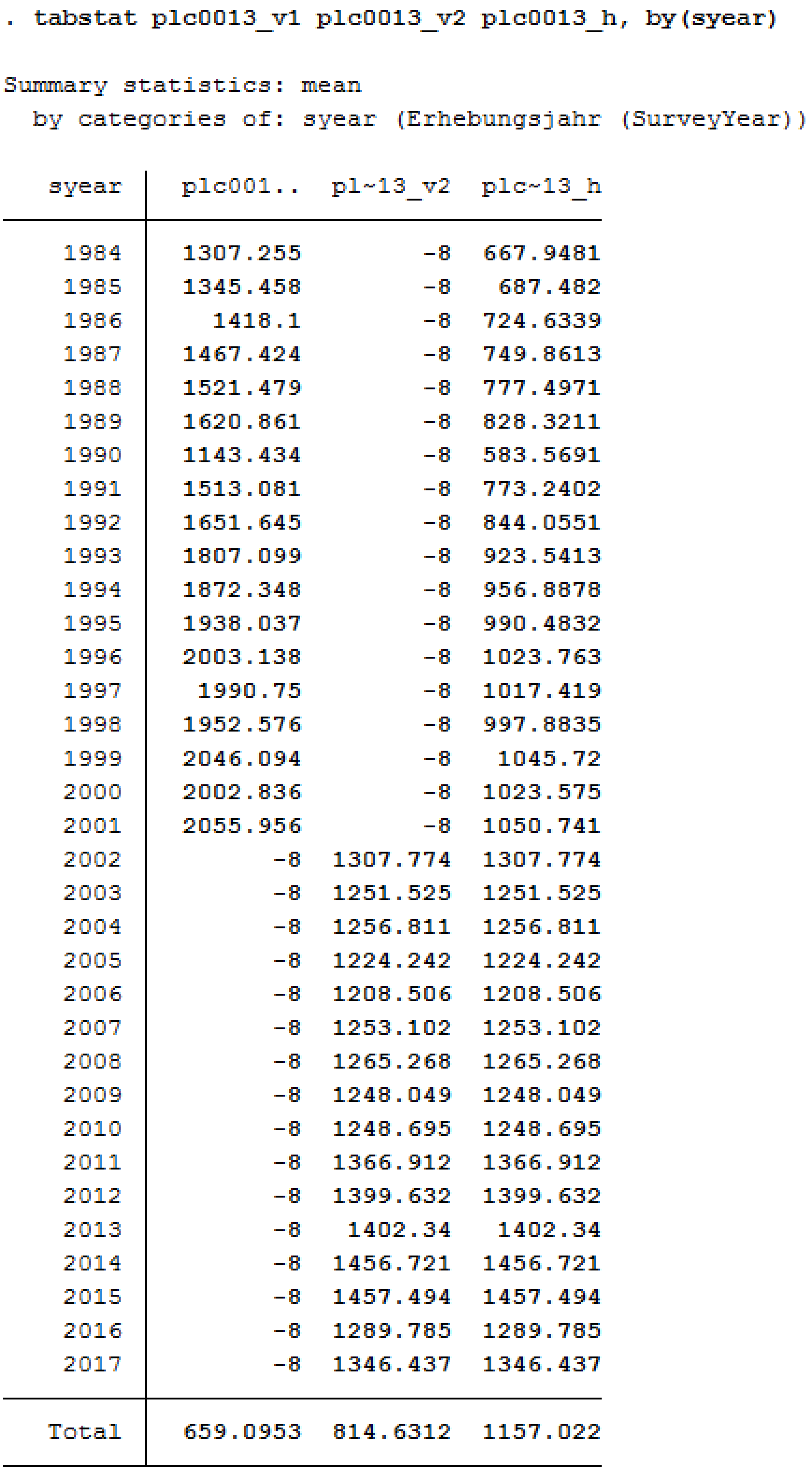
My income was: gross ☐ EURO

Euro harmonisation involves DM amounts being multiplied by the exchange rate so that the harmonized version of the variable represents euro amounts.

```
list pid syear plc0013_v1 plc0013_h if pid==7006001 & syear==2001
tabstat plc0013_v1 plc0013_v2 plc0013_h, by(syear)
```



```
. list pid syear plc0013_v1 plc0013_h if pid==7006001 & syear==2001

        +---------------------------------------+
        |     pid   syear   plc001..   plc~13_h |
        |---------------------------------------|
478114. | 7006001    2001       4200       2147 |
        +---------------------------------------+
```



```
. tabstat plc0013_v1 plc0013_v2 plc0013_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear))

   syear |  plc001..  pl~13_v2  plc~13_h
---------+------------------------------
    1984 |  1307.255        -8  667.9481
    1985 |  1345.458        -8   687.482
    1986 |    1418.1        -8  724.6339
    1987 |  1467.424        -8  749.8613
    1988 |  1521.479        -8  777.4971
    1989 |  1620.861        -8  828.3211
    1990 |  1143.434        -8  583.5691
    1991 |  1513.081        -8  773.2402
    1992 |  1651.645        -8  844.0551
    1993 |  1807.099        -8  923.5413
    1994 |  1872.348        -8  956.8878
    1995 |  1938.037        -8  990.4832
    1996 |  2003.138        -8  1023.763
    1997 |   1990.75        -8  1017.419
    1998 |  1952.576        -8  997.8835
    1999 |  2046.094        -8   1045.72
    2000 |  2002.836        -8  1023.575
    2001 |  2055.956        -8  1050.741
    2002 |        -8  1307.774  1307.774
    2003 |        -8  1251.525  1251.525
    2004 |        -8  1256.811  1256.811
    2005 |        -8  1224.242  1224.242
    2006 |        -8  1208.506  1208.506
    2007 |        -8  1253.102  1253.102
    2008 |        -8  1265.268  1265.268
    2009 |        -8  1248.049  1248.049
    2010 |        -8  1248.695  1248.695
    2011 |        -8  1366.912  1366.912
    2012 |        -8  1399.632  1399.632
    2013 |        -8   1402.34   1402.34
    2014 |        -8  1456.721  1456.721
    2015 |        -8  1457.494  1457.494
    2016 |        -8  1289.785  1289.785
    2017 |        -8  1346.437  1346.437
---------+------------------------------
   Total |  659.0953  814.6312  1157.022
----------------------------------------
```

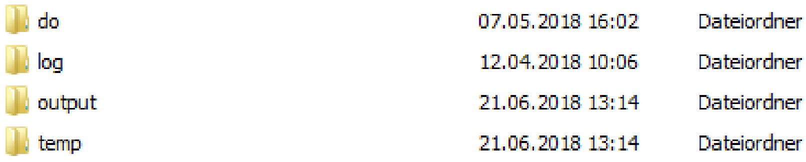
Last change: Sep 26, 2022

To get an idea of the analysis potential offered by the SOEP, we recommend the following exercises to our users:

6.6 Longitudinal Data Analysis

Simple cross-sectional analyses show that married people have higher life satisfaction than singles. You want to check this on the basis of longitudinal analysis with the SOEP.

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define the paths you created with globals:

```
***********************************************
* Set some useful commands
***********************************************
version 13
clear all
set more off
**increase buffer size
set scrollbufsize 2000000
**now restart stata!

***********************************************
* Set relative paths to the working directory
***********************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\distribution\soep-long\soep.v33.1\stata_en\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global “AVZ” defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to your ordered data.

Create a master file that uses the important variables from ppathl.

You should always add some variables from PPATHL to your dataset by default. Download the following information from PPATHL:

- Individual identifier **"pid"**

- Household identifier **"pid"**
- Survey year **"syear"**
- The net variable with information on the interview type **"netto"**
- The weighting variable **"phrf"**
- The gender of the person **"sex"**
- The migration background **"migback"**

```
*-------------------------------------------------------------------------------
*** Step 1) Start with basic information from PPFADL ***

use pid hid syear netto phrf migback sex using ${MY_IN_PATH}\ppfadl.dta
```

> **Attention:** Please note that since version 34 (v34), PPFADL has been renamed PPATHL. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFADL.

Search for matching variables and add them to your dataset

To perform your analysis, you need different SOEP variables. The SOEP offers various options for a variable search:

- Search the questionnaires for useful variables. (for more information, see the section *Variable Search with Questionnaires*)
- Find a suitable variable via the topic list of paneldata.org (for more information, see the section *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information, see the section *Variable Search with paneldata.org*)
- Use the documentation provided on the generated variables (for more information, see the section *Documentation on Generated Data*)

In this case, you need the variables **"pgfamstd"** (martial status) and **"plh0182"** (life satisfaction).

```
*-------------------------------------------------------------------------------
*** Step 2) Add the relavant variables: here: family status and life satisfaction ***
merge 1:1 pid syear using ${MY_IN_PATH}\pgen, keepusing(pgfamstd) keep(1 3) nogen

                // merges family status from pgen
                // Documentation for PGEN can be found here
                // http://panel.gsoep.de/soep-docs/surveypapers/diw_ssp0307.pdf)


*describe using pl (directory)
                // for checking out variable names without opening the dataset

merge 1:1 pid syear using ${MY_IN_PATH}\pl, keepusing(plh0182) keep(1 3) nogen
                // merges life satisfaction from pl

save $MY_OUT_DATA\ppfad.dta, replace
```

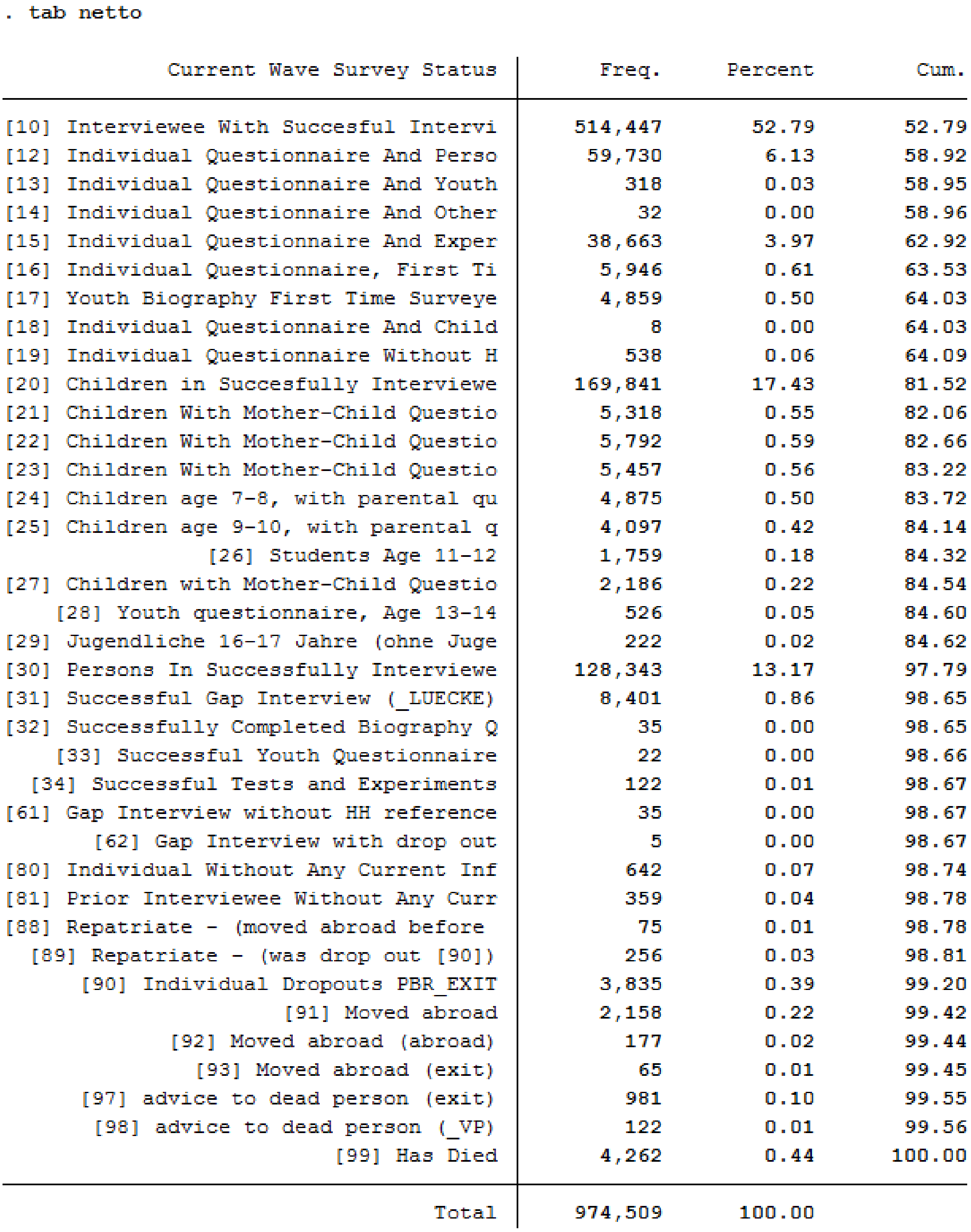
Clean and inspect the data

Recode all missing values with commas to period decimal format.

```
*--------------------------------------------------------------------------------
*** Step 3) Clean and inspect the data
mvdecode _all, mv(-8/-1)
```

Since you are interested in individual characteristics in your analysis: Delete all measurements that are not based on successful individual interviews.

```
tab netto
drop if netto>19
```



```
. tab netto

             Current Wave Survey Status |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
[10] Interviewee With Succesful Intervi |    514,447       52.79       52.79
[12] Individual Questionnaire And Perso |     59,730        6.13       58.92
[13] Individual Questionnaire And Youth |        318        0.03       58.95
[14] Individual Questionnaire And Other |         32        0.00       58.96
[15] Individual Questionnaire And Exper |     38,663        3.97       62.92
[16] Individual Questionnaire, First Ti |      5,946        0.61       63.53
[17] Youth Biography First Time Surveye |      4,859        0.50       64.03
[18] Individual Questionnaire And Child |          8        0.00       64.03
[19] Individual Questionnaire Without H |        538        0.06       64.09
[20] Children in Succesfully Interviewe |    169,841       17.43       81.52
[21] Children With Mother-Child Questio |      5,318        0.55       82.06
[22] Children With Mother-Child Questio |      5,792        0.59       82.66
[23] Children With Mother-Child Questio |      5,457        0.56       83.22
[24] Children age 7-8, with parental qu |      4,875        0.50       83.72
[25] Children age 9-10, with parental q |      4,097        0.42       84.14
                [26] Students Age 11-12 |      1,759        0.18       84.32
[27] Children with Mother-Child Questio |      2,186        0.22       84.54
    [28] Youth questionnaire, Age 13-14 |        526        0.05       84.60
[29] Jugendliche 16-17 Jahre (ohne Juge |        222        0.02       84.62
[30] Persons In Successfully Interviewe |    128,343       13.17       97.79
[31] Successful Gap Interview (_LUECKE) |      8,401        0.86       98.65
[32] Successfully Completed Biography Q |         35        0.00       98.65
    [33] Successful Youth Questionnaire |         22        0.00       98.66
  [34] Successful Tests and Experiments |        122        0.01       98.67
[61] Gap Interview without HH reference |         35        0.00       98.67
       [62] Gap Interview with drop out |          5        0.00       98.67
[80] Individual Without Any Current Inf |        642        0.07       98.74
[81] Prior Interviewee Without Any Curr |        359        0.04       98.78
[88] Repatriate - (moved abroad before  |         75        0.01       98.78
  [89] Repatriate - (was drop out [90]) |        256        0.03       98.81
      [90] Individual Dropouts PBR_EXIT |      3,835        0.39       99.20
                   [91] Moved abroad |      2,158        0.22       99.42
          [92] Moved abroad (abroad) |        177        0.02       99.44
            [93] Moved abroad (exit) |         65        0.01       99.45
      [97] advice to dead person (exit) |        981        0.10       99.55
       [98] advice to dead person (_VP) |        122        0.01       99.56
                          [99] Has Died |      4,262        0.44      100.00
----------------------------------------+-----------------------------------
                                  Total |    974,509      100.00
```

How many people contribute measurements and what is the proportion of people contributing at least 10 measurements?

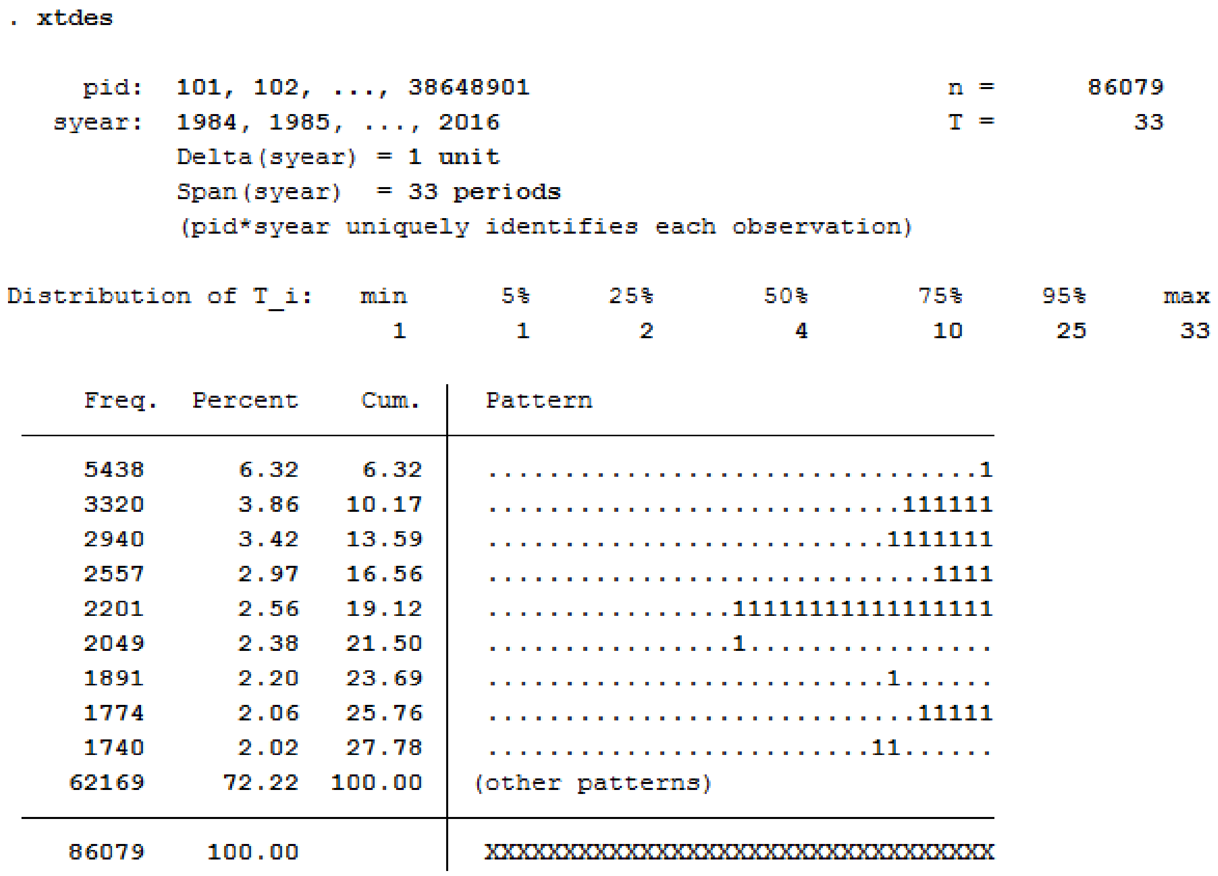
Define the dataset as a panel dataset.

```
**define the dataset as panel data
```

(continues on next page)

(continued from previous page)

```
xtset pid syear
xtdes
```

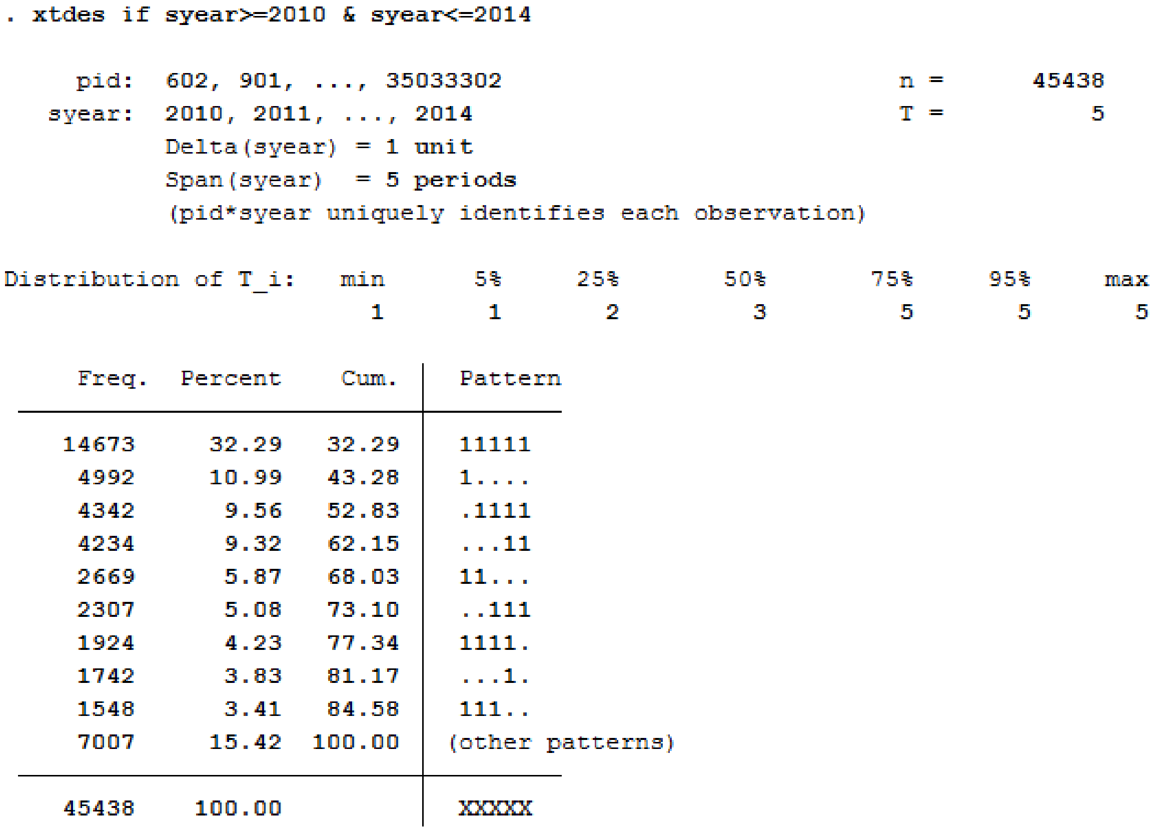


```
. xtdes

     pid:  101, 102, ..., 38648901                          n =      86079
   syear:  1984, 1985, ..., 2016                            T =         33
           Delta(syear) = 1 unit
           Span(syear)  = 33 periods
           (pid*syear uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         4        10      25      33

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------------------------------
     5438      6.32    6.32 |  ................................1
     3320      3.86   10.17 |  ...........................111111
     2940      3.42   13.59 |  ..........................1111111
     2557      2.97   16.56 |  .............................1111
     2201      2.56   19.12 |  ................11111111111111111
     2049      2.38   21.50 |  ................1................
     1891      2.20   23.69 |  ..........................1......
     1774      2.06   25.76 |  ............................11111
     1740      2.02   27.78 |  .........................11......
    62169     72.22  100.00 | (other patterns)
 ---------------------------+-----------------------------------
    86079    100.00         |  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

86,079 respondents have contributed information in waves a (1984) to bg (2016) and 75% of the 86,079 respondents have provided information for at least 10 waves.

How many people took part in the survey in 2010 and contributed to continuous measurements up to 2014?

```
xtdes if syear>=2010 & syear<=2014
```



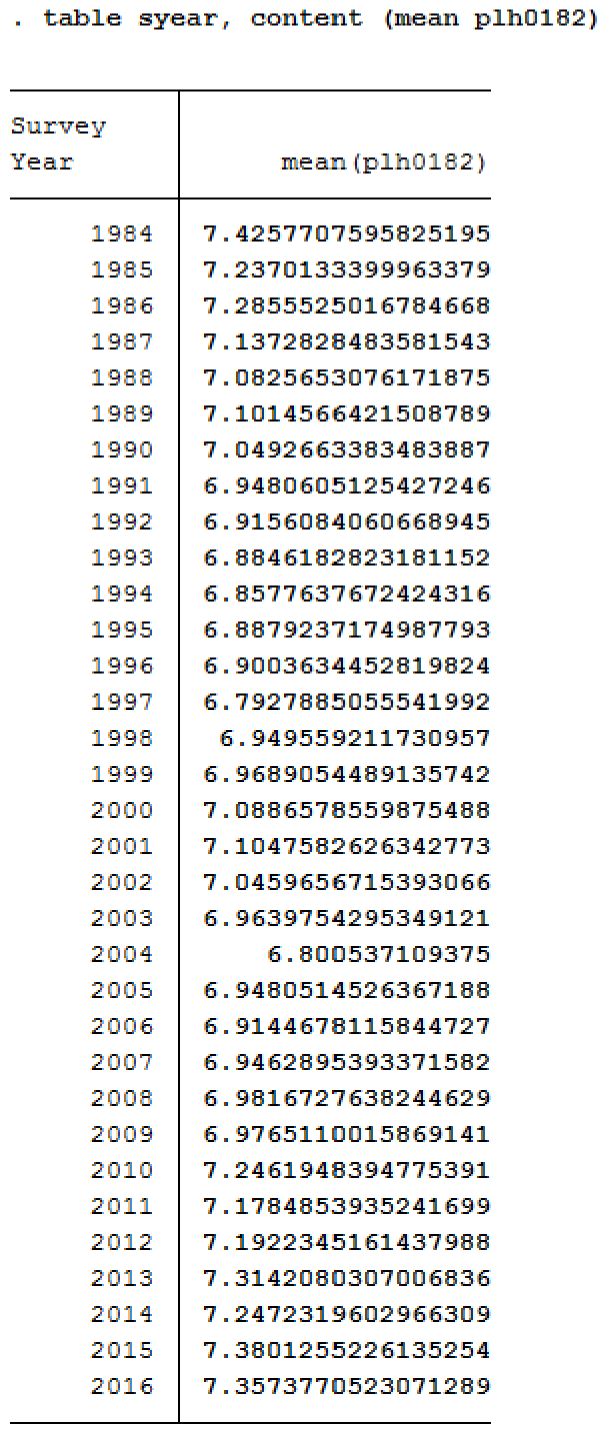
```
. xtdes if syear>=2010 & syear<=2014

     pid:  602, 901, ..., 35033302                          n =      45438
   syear:  2010, 2011, ..., 2014                            T =          5
           Delta(syear) = 1 unit
           Span(syear)  = 5 periods
           (pid*syear uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         3         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
    14673     32.29   32.29 |  11111
     4992     10.99   43.28 |  1....
     4342      9.56   52.83 |  .1111
     4234      9.32   62.15 |  ...11
     2669      5.87   68.03 |  11...
     2307      5.08   73.10 |  ..111
     1924      4.23   77.34 |  1111.
     1742      3.83   81.17 |  ...1.
     1548      3.41   84.58 |  111..
     7007     15.42  100.00 | (other patterns)
 ---------------------------+---------
    45438    100.00         |  XXXXX
```

14,673 respondents provided continuous information from 2010 to 2014.

Univariate inspection & analysis

How does the mean of life satisfaction change over time?

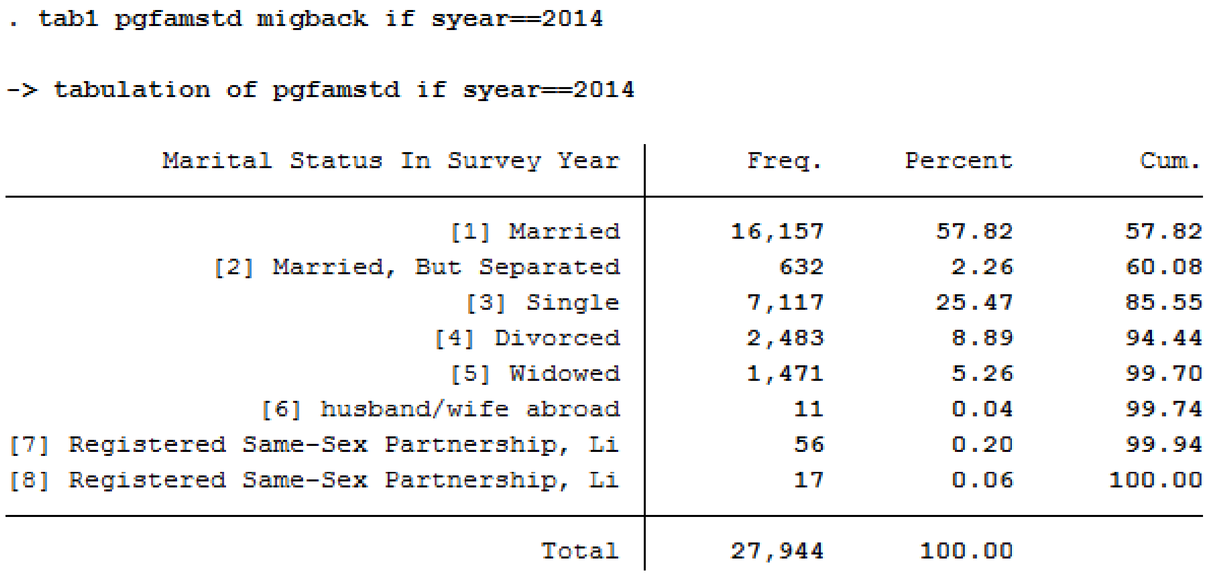
```
*-------------------------------------------------------------------------------
*** Step 4) univariate inspection & analysis
table syear, content (mean plh0182)
```



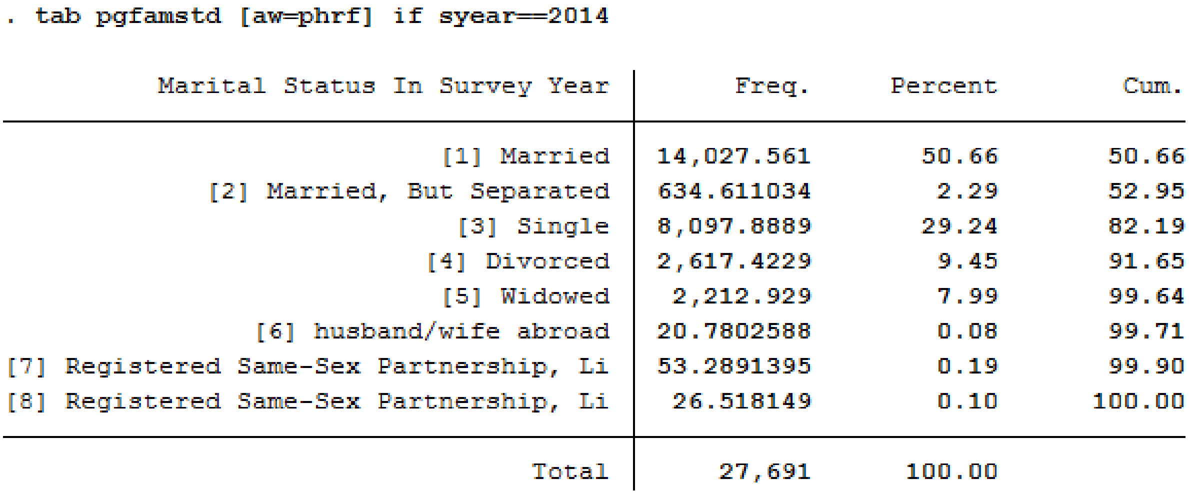
```
. table syear, content (mean plh0182)

Survey
Year   |       mean(plh0182)
-------+--------------------
  1984 |  7.4257707595825195
  1985 |  7.2370133399963379
  1986 |  7.2855525016784668
  1987 |  7.1372828483581543
  1988 |  7.0825653076171875
  1989 |  7.1014566421508789
  1990 |  7.0492663383483887
  1991 |  6.9480605125427246
  1992 |  6.9156084060668945
  1993 |  6.8846182823181152
  1994 |  6.8577637672424316
  1995 |  6.8879237174987793
  1996 |  6.9003634452819824
  1997 |  6.7927885055541992
  1998 |   6.949559211730957
  1999 |  6.9689054489135742
  2000 |  7.0886578559875488
  2001 |  7.1047582626342773
  2002 |  7.0459656715393066
  2003 |  6.9639754295349121
  2004 |      6.800537109375
  2005 |  6.9480514526367188
  2006 |  6.9144678115844727
  2007 |  6.9462895393371582
  2008 |  6.9816727638244629
  2009 |  6.9765110015869141
  2010 |  7.2461948394775391
  2011 |  7.1784853935241699
  2012 |  7.1922345161437988
  2013 |  7.3142080307006836
  2014 |  7.2472319602966309
  2015 |  7.3801255226135254
  2016 |  7.3573770523071289
```

What proportion of people are a) married in 2014 or b) have a migration background? Compare weighted with unweighted frequency tables: Who is overrepresented in SOEP?

```
tab1 pgfamstd migback if syear==2014
tab pgfamstd [aw=phrf] if syear==2014
tab migback [aw=phrf] if syear==2014
```

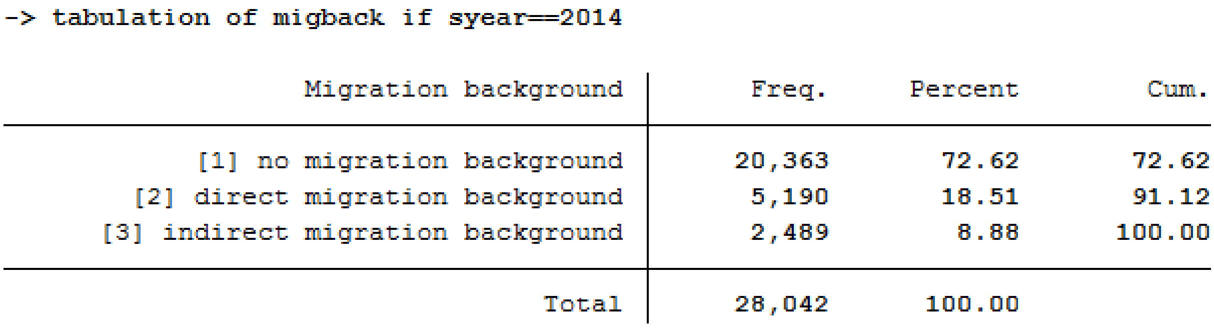


```
. tab1 pgfamstd migback if syear==2014

-> tabulation of pgfamstd if syear==2014

           Marital Status In Survey Year |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
                             [1] Married |     16,157       57.82       57.82
               [2] Married, But Separated |        632        2.26       60.08
                              [3] Single |      7,117       25.47       85.55
                            [4] Divorced |      2,483        8.89       94.44
                             [5] Widowed |      1,471        5.26       99.70
                    [6] husband/wife abroad |         11        0.04       99.74
[7] Registered Same-Sex Partnership, Li |         56        0.20       99.94
[8] Registered Same-Sex Partnership, Li |         17        0.06      100.00
-----------------------------------------+-----------------------------------
                                   Total |     27,944      100.00
```

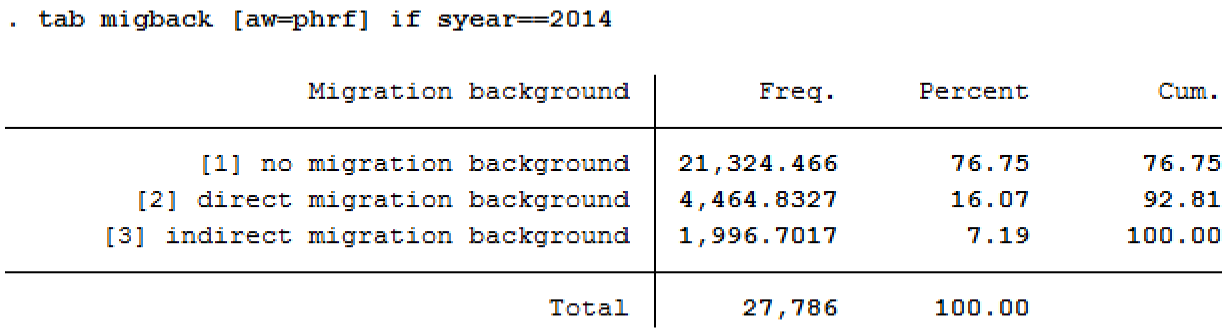


```
. tab pgfamstd [aw=phrf] if syear==2014

           Marital Status In Survey Year |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
                             [1] Married | 14,027.561       50.66       50.66
               [2] Married, But Separated | 634.611034        2.29       52.95
                              [3] Single | 8,097.8889       29.24       82.19
                            [4] Divorced | 2,617.4229        9.45       91.65
                             [5] Widowed |  2,212.929        7.99       99.64
                    [6] husband/wife abroad | 20.7802588        0.08       99.71
[7] Registered Same-Sex Partnership, Li | 53.2891395        0.19       99.90
[8] Registered Same-Sex Partnership, Li |  26.518149        0.10      100.00
-----------------------------------------+-----------------------------------
                                   Total |     27,691      100.00
```

The data show that married people are overrepresented in the SOEP and single people are underrepresented. The weighting makes it representative again for Germany.



```
-> tabulation of migback if syear==2014

             Migration background |      Freq.     Percent        Cum.
----------------------------------+-----------------------------------
       [1] no migration background |     20,363       72.62       72.62
   [2] direct migration background |      5,190       18.51       91.12
 [3] indirect migration background |      2,489        8.88      100.00
----------------------------------+-----------------------------------
                             Total |     28,042      100.00
```

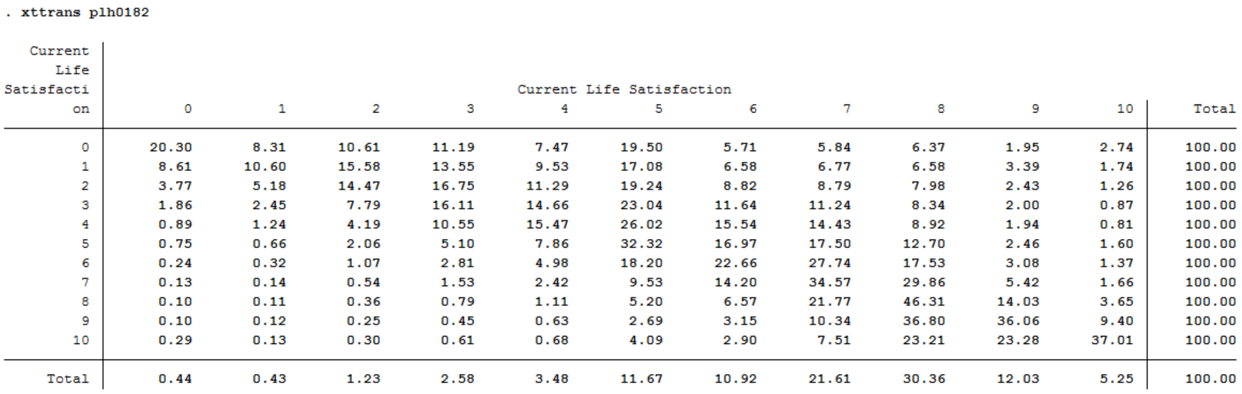


```
. tab migback [aw=phrf] if syear==2014

           Migration background |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
        [1] no migration background | 21,324.466       76.75       76.75
    [2] direct migration background | 4,464.8327       16.07       92.81
  [3] indirect migration background | 1,996.7017        7.19      100.00
--------------------------------+-----------------------------------
                          Total |     27,786      100.00
```

In the SOEP sample, respondents with a direct or indirect migration background are overrepresented.

How many of those persons who reported a life satisfaction scale value of 7 in one survey year also indicated the scale value of 7 in the following survey year?

```
xttrans plh0182
```

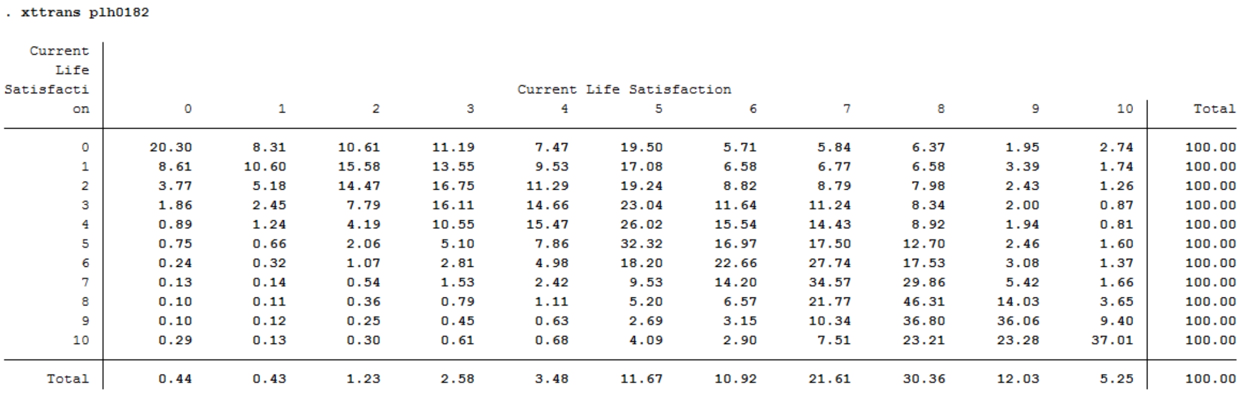


```
. xttrans plh0182

   Current |
      Life |
Satisfacti |                                 Current Life Satisfaction
        on |         0          1          2          3          4          5          6          7          8          9         10 |     Total
-----------+--------------------------------------------------------------------------------------------------------------------------+----------
         0 |     20.30       8.31      10.61      11.19       7.47      19.50       5.71       5.84       6.37       1.95       2.74 |    100.00
         1 |      8.61      10.60      15.58      13.55       9.53      17.08       6.58       6.77       6.58       3.39       1.74 |    100.00
         2 |      3.77       5.18      14.47      16.75      11.29      19.24       8.82       8.79       7.98       2.43       1.26 |    100.00
         3 |      1.86       2.45       7.79      16.11      14.66      23.04      11.64      11.24       8.34       2.00       0.87 |    100.00
         4 |      0.89       1.24       4.19      10.55      15.47      26.02      15.54      14.43       8.92       1.94       0.81 |    100.00
         5 |      0.75       0.66       2.06       5.10       7.86      32.32      16.97      17.50      12.70       2.46       1.60 |    100.00
         6 |      0.24       0.32       1.07       2.81       4.98      18.20      22.66      27.74      17.53       3.08       1.37 |    100.00
         7 |      0.13       0.14       0.54       1.53       2.42       9.53      14.20      34.57      29.86       5.42       1.66 |    100.00
         8 |      0.10       0.11       0.36       0.79       1.11       5.20       6.57      21.77      46.31      14.03       3.65 |    100.00
         9 |      0.10       0.12       0.25       0.45       0.63       2.69       3.15      10.34      36.80      36.06       9.40 |    100.00
        10 |      0.29       0.13       0.30       0.61       0.68       4.09       2.90       7.51      23.21      23.28      37.01 |    100.00
-----------+--------------------------------------------------------------------------------------------------------------------------+----------
     Total |      0.44       0.43       1.23       2.58       3.48      11.67      10.92      21.61      30.36      12.03       5.25 |    100.00
```

34.57% of the respondents who reported a life satisfaction of 7 again reported a value of 7 in the following year.

Is it more likely that a highly dissatisfied person (value: 0) will be less dissatisfied the following year or that a very satisfied (value: 10) person will be less satisfied the following year?

```
xttrans plh0182
```

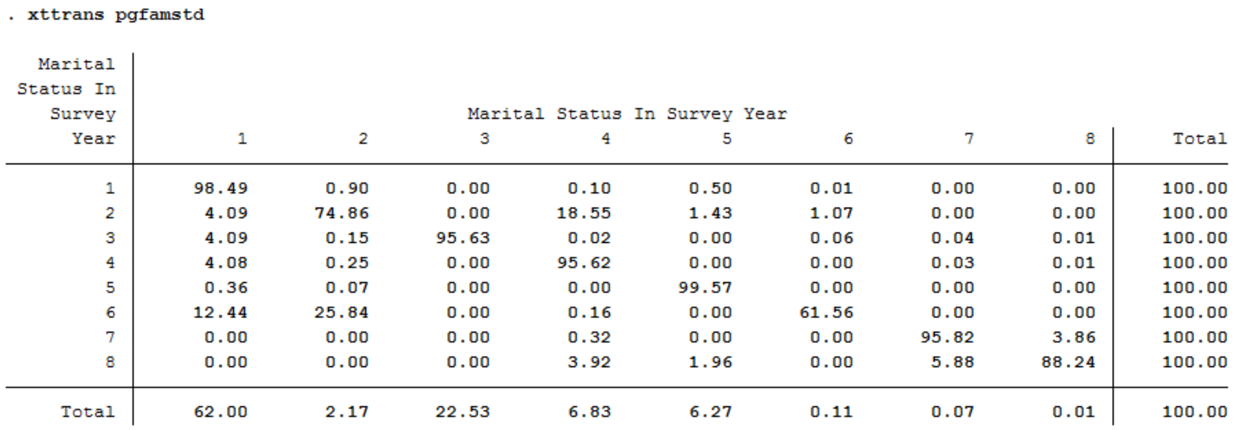


```
. xttrans plh0182

   Current |
      Life |
Satisfacti |                                 Current Life Satisfaction
        on |         0          1          2          3          4          5          6          7          8          9         10 |     Total
-----------+--------------------------------------------------------------------------------------------------------------------------+----------
         0 |     20.30       8.31      10.61      11.19       7.47      19.50       5.71       5.84       6.37       1.95       2.74 |    100.00
         1 |      8.61      10.60      15.58      13.55       9.53      17.08       6.58       6.77       6.58       3.39       1.74 |    100.00
         2 |      3.77       5.18      14.47      16.75      11.29      19.24       8.82       8.79       7.98       2.43       1.26 |    100.00
         3 |      1.86       2.45       7.79      16.11      14.66      23.04      11.64      11.24       8.34       2.00       0.87 |    100.00
         4 |      0.89       1.24       4.19      10.55      15.47      26.02      15.54      14.43       8.92       1.94       0.81 |    100.00
         5 |      0.75       0.66       2.06       5.10       7.86      32.32      16.97      17.50      12.70       2.46       1.60 |    100.00
         6 |      0.24       0.32       1.07       2.81       4.98      18.20      22.66      27.74      17.53       3.08       1.37 |    100.00
         7 |      0.13       0.14       0.54       1.53       2.42       9.53      14.20      34.57      29.86       5.42       1.66 |    100.00
         8 |      0.10       0.11       0.36       0.79       1.11       5.20       6.57      21.77      46.31      14.03       3.65 |    100.00
         9 |      0.10       0.12       0.25       0.45       0.63       2.69       3.15      10.34      36.80      36.06       9.40 |    100.00
        10 |      0.29       0.13       0.30       0.61       0.68       4.09       2.90       7.51      23.21      23.28      37.01 |    100.00
-----------+--------------------------------------------------------------------------------------------------------------------------+----------
     Total |      0.44       0.43       1.23       2.58       3.48      11.67      10.92      21.61      30.36      12.03       5.25 |    100.00
```

The rows reflect the initial values, and the columns reflect the final values. Around 20% of those who were completely dissatisfied (value: 0) in the base year remained completely dissatisfied in the following year. About 80% of these completely dissatisfied people from the base year were more satisfied in the following year. Of the completely satisfied persons (value: 10), about 37% remained just as satisfied in the following year, but 63% became less satisfied. It is more likely that a completely dissatisfied person will become more satisfied in the following year than that a completely satisfied person will become less satisfied.

Which transitions in marital status can be observed particularly frequently in the data?

```
xttrans pgfamstd
```



```
. xttrans pgfamstd

   Marital |
 Status In |
    Survey |                       Marital Status In Survey Year
      Year |         1          2          3          4          5          6          7          8 |     Total
-----------+----------------------------------------------------------------------------------------+----------
         1 |     98.49       0.90       0.00       0.10       0.50       0.01       0.00       0.00 |    100.00
         2 |      4.09      74.86       0.00      18.55       1.43       1.07       0.00       0.00 |    100.00
         3 |      4.09       0.15      95.63       0.02       0.00       0.06       0.04       0.01 |    100.00
         4 |      4.08       0.25       0.00      95.62       0.00       0.00       0.03       0.01 |    100.00
         5 |      0.36       0.07       0.00       0.00      99.57       0.00       0.00       0.00 |    100.00
         6 |     12.44      25.84       0.00       0.16       0.00      61.56       0.00       0.00 |    100.00
         7 |      0.00       0.00       0.00       0.32       0.00       0.00      95.82       3.86 |    100.00
         8 |      0.00       0.00       0.00       3.92       1.96       0.00       5.88      88.24 |    100.00
-----------+----------------------------------------------------------------------------------------+----------
     Total |     62.00       2.17      22.53       6.83       6.27       0.11       0.07       0.01 |    100.00
```

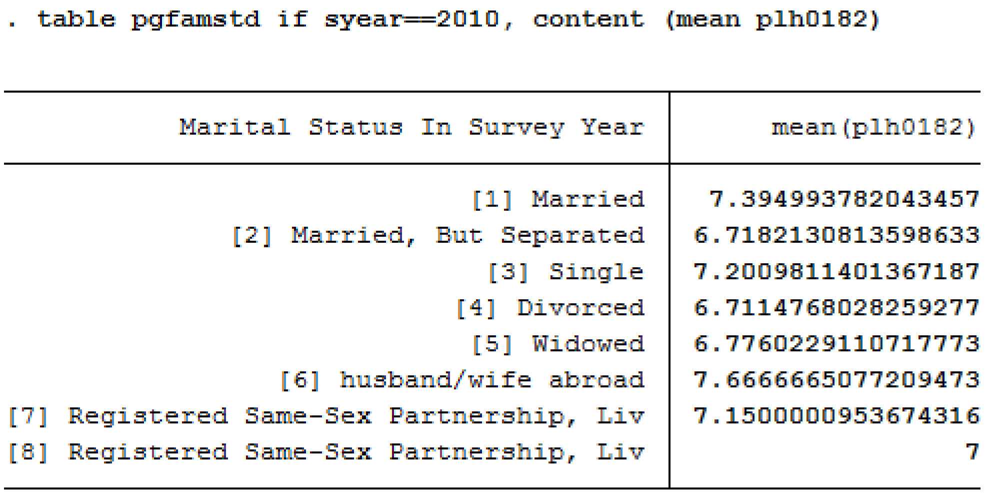
Survey respondents who were married but separated in the base year and reported divorce as their family status in the following year can be observed particularly frequently. (about 19%).

Simple cross sectional analyses

You now want to find the correlation between marital status and life satisfaction. Is there an effect of marriage on life satisfaction? And if so, is it a sustained effect?

First, calculate the correlation between family status and life satisfaction in from a cross-sectional perspective for 2010: Are married people happier than singles?

```
*-------------------------------------------------------------------------------
*** Step 5)simple cross-sectional analyses
table pgfamstd if syear==2010, content (mean plh0182)
```



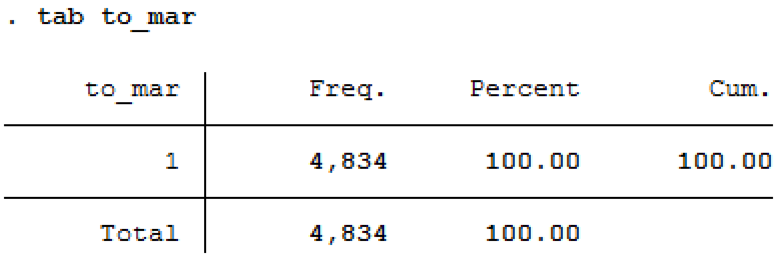
```
. table pgfamstd if syear==2010, content (mean plh0182)

---------------------------------------------------------------
                 Marital Status In Survey Year |   mean(plh0182)
-----------------------------------------------+---------------
                                   [1] Married | 7.394993782043457
                   [2] Married, But Separated | 6.718213081359863
                                    [3] Single | 7.200981140136718 7
                                  [4] Divorced | 6.711476802825927 7
                                   [5] Widowed | 6.776022911071777 3
                        [6] husband/wife abroad | 7.666666507720947 3
[7] Registered Same-Sex Partnership, Liv | 7.150000095367431 6
[8] Registered Same-Sex Partnership, Liv |                 7
---------------------------------------------------------------
```

At first glance, married couples seem happier than singles.

Now generate a variable that indicates a transition from “single” to “married”.

How many such transitions can you find in the data?

```
***perform longitudinal analysis
**define event: transition to marriage
generate to_mar=1 if pgfamstd==1 & l.pgfamstd==3
tab to_mar
```

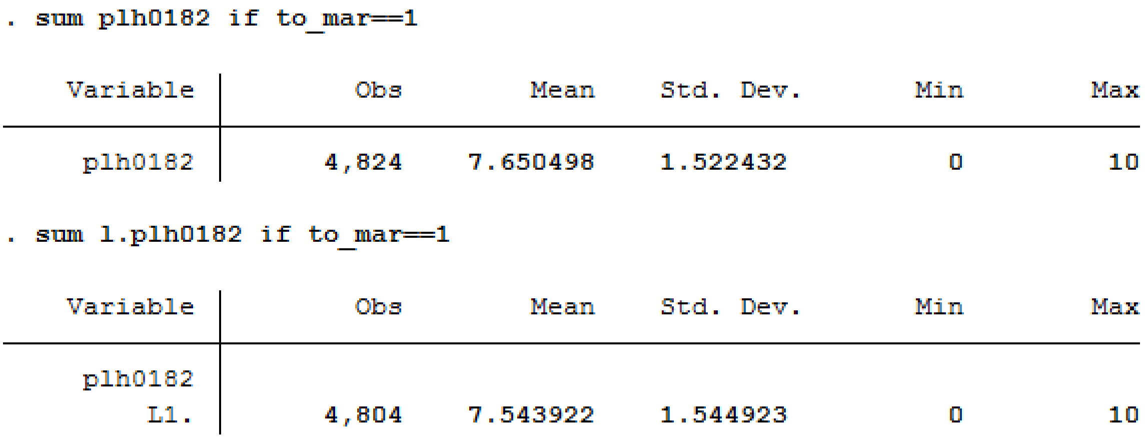


```
. tab to_mar

     to_mar |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      4,834      100.00      100.00
------------+-----------------------------------
      Total |      4,834      100.00
```

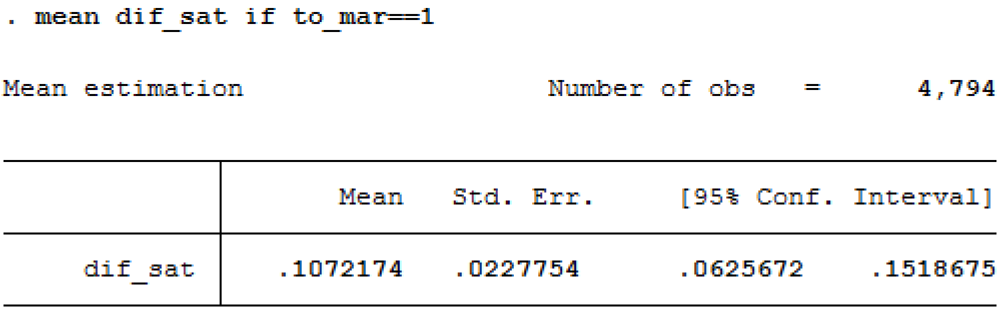
A total of 4,834 people can be observed changing status from single to married.

What is the average level of life satisfaction immediately after the transition to marriage (i.e., in the first survey in which the transition can be observed) and how high is life satisfaction immediately before the transition to marriage?

```
**standard way of life-event analysis
sum plh0182 if to_mar==1
sum l.plh0182 if to_mar==1

**alternative way
generate dif_sat= plh0182- l.plh0182
mean dif_sat if to_mar==1
```



```
. sum plh0182 if to_mar==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     plh0182 |      4,824    7.650498    1.522432          0         10

. sum l.plh0182 if to_mar==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     plh0182 |
         L1. |      4,804    7.543922    1.544923          0         10
```

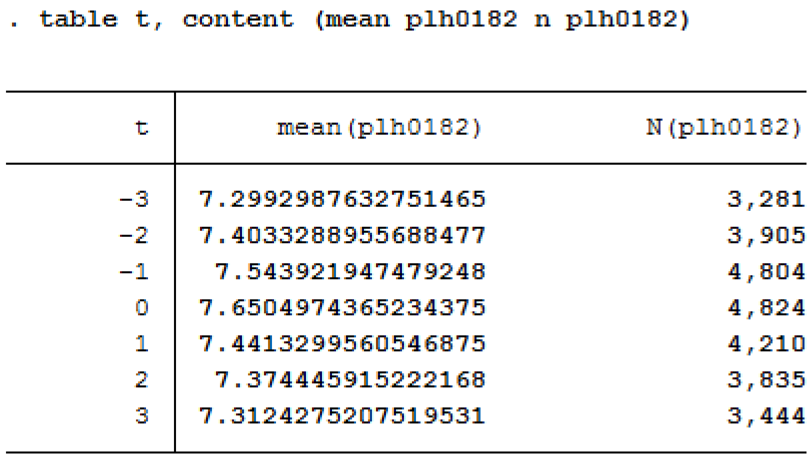


```
. mean dif_sat if to_mar==1

Mean estimation                   Number of obs     =      4,794

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     dif_sat |   .1072174   .0227754      .0625672    .1518675
--------------------------------------------------------------
```

Before the transition to marriage, the average life satisfaction of the respondents is 7.54. In the following year, that is, after the transition to marriage, the average life satisfaction of the respondents is 7.65. It can be seen that with the transition to marriage, average life satisfaction rises slightly by 0.11.

Map the complete satisfaction history around the "marriage entry" event [3 years before; 3 years after].

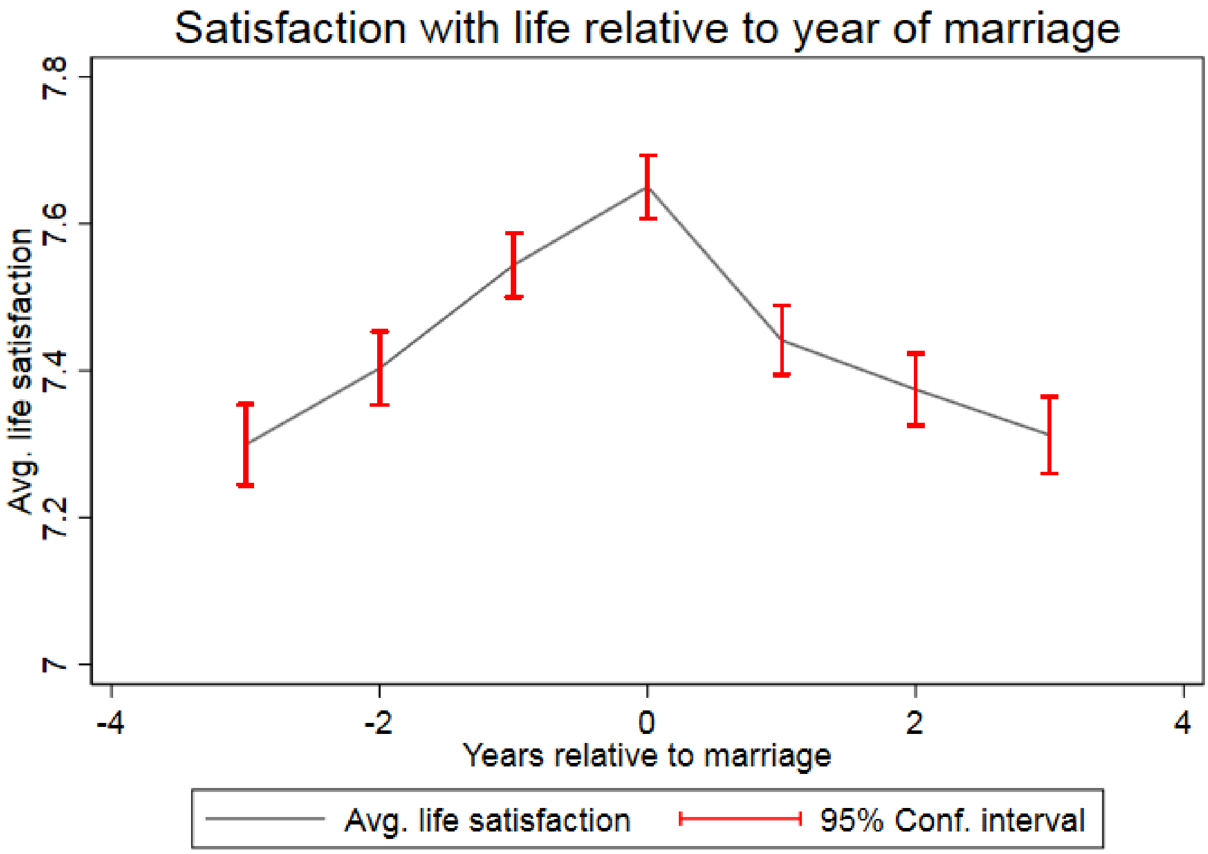
```
**preparing illustration of trajectory
generate t=0 if to_mar==1 & l.to_mar~=1 &l2.to_mar~=1 & l3.to_mar~=1 & l4.to_mar~=1 & l5.to_mar~=1 & l6.to_mar~=1 & l7.to_mar~=1 & l8.to_mar~=1 & l9.to_mar~=1 & l10.to_mar~=1 & l11.to_mar~=1 & l12.to_mar~=1 & l13.to_mar~=1 & l14.to_mar~=1
replace t=1 if l.t==0
replace t=2 if l2.t==0
replace t=3 if l3.t==0
replace t=-1 if f.t==0
replace t=-2 if f2.t==0
replace t=-3 if f3.t==0

table t, content (mean plh0182 n plh0182)
```



```
. table t, content (mean plh0182 n plh0182)

          t        mean(plh0182)      N(plh0182)

         -3   7.2992987632751465           3,281
         -2   7.4033288955688477           3,905
         -1    7.543921947479248           4,804
          0   7.6504974365234375           4,824
          1   7.4413299560546875           4,210
          2    7.374445915222168           3,835
          3   7.3124275207519531           3,444
```

Choose a suitable presentation for your results and let Stata create a graphic.

```
** Preparing graph of event analysis
sort t
cap drop meanplh0182
by t: egen meanplh0182 = mean(plh0182)

cap drop upper
gen upper = .
forval i = -3/3{
        su plh0182 if t == `i'
        replace upper = r(mean) + 1.96 * r(sd)/sqrt(r(N)) if t == `i'
}

cap drop lower
gen lower = .
forval i = -3/3{
        su plh0182 if t == `i'
        replace lower = r(mean) - 1.96 * r(sd)/sqrt(r(N)) if t == `i'
}

twoway (line meanplh0182 t) (rcap upper lower t, lcolor("red")) , title("Satisfaction with life relative to year of marriage") legend(label(1 "Avg. life satisfaction") label(2 "95% Conf. interval")) scheme(s1mono) xtitle("Years relative to marriage") ytitle("Avg. life satisfaction")
```



The graph shows that a positive effect on life satisfaction can be observed when family status changes from single to married. In the following years of the existing marriage, life satisfaction decreases again and approaches the initial satisfaction before the marriage.

Last change: Sep 26, 2022

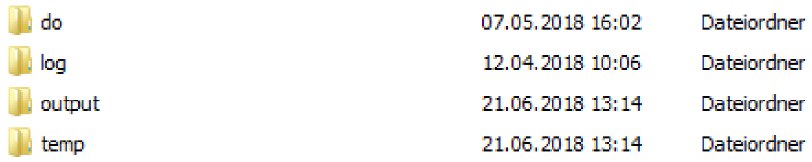
6.7 Working with Migration Data (BIOIMMIG)

With its migration and refugee samples, SOEP provides a wide range of information on people with a history of migration or forced migration.

In the BIOIMMIG dataset, you will find relevant information on the history of migration or forced migration, including refugees' and migrants' motives for leaving their country of origin, their living conditions upon arrival in Germany, as well as information in edited form on any relatives in the country of origin and the desire to return to the country of origin. For more information about this dataset and a list of the variables it contains, see: BIOIMMIG Documentation.

In the following, we will use this record and other information from the SOEP to create a status variable that you can use to distinguish whether or not people with a migration background also have a background of forced migration, that is, whether migrants are also refugees.

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store commands, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
***************************************************
* Set relative paths to the working directory
***************************************************
global AVZ          "H:/material/exercises"
global MY_IN_PATH "//hume/rdc-prod/distribution/soep-core/soep.v37/eu/Stata/"
global MY_DO_FILES "$AVZ/do/"
global MY_LOG_OUT "$AVZ/log/"
global MY_OUT_DATA "$AVZ/output/"
global MY_OUT_TEMP "$AVZ/temp/"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to the data you ordered.

Task 1: Preparation of BIOIMMIG

a) Which variable contains information about the status of each person when they immigrated to Germany?

Open the record or browse the documentation and search for a variable describing the immigration status. The biimgrp variable from the BIOIMMIG data set is the appropriate variable.

```
*** Exercise 1 ******************************************************************

/*
a)        Which variable contains information about the status of each person when they immigrated to Germany?
*/

* Immigration status is stored in the variable biimgrp.

use $MY_IN_PATH\bioimmig.dta, clear
```

b) Identify this variable in the BIOIMMIG dataset and retrieve it from the dataset together with the person number and survey year.

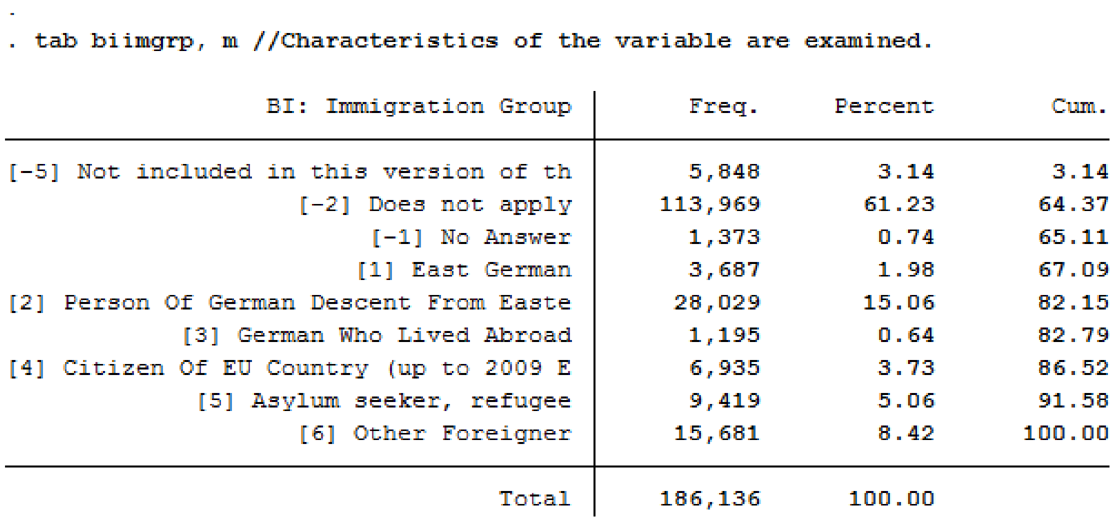
Open your dataset with just the required variables to maintain clarity for your analysis.

```
/*
b)        Identify this variable in the BIOIMMIG dataset and load it from the data set, together with the individual indentifier and the survey year.
*/

use persnr syear biimgrp using $MY_IN_PATH\bioimmig.dta, clear
```

c) What are the values of this variable?

Familiarize yourself with your variable and check the coding and case numbers.

```
/*
c)        What are the values of this variable?
*/

tab biimgrp, m //Characteristics of the variable are examined.
```



```
.
. tab biimgrp, m //Characteristics of the variable are examined.

                     BI: Immigration Group |      Freq.     Percent        Cum.
-------------------------------------------+-----------------------------------
 [-5] Not included in this version of th   |      5,848        3.14        3.14
                     [-2] Does not apply   |    113,969       61.23       64.37
                          [-1] No Answer   |      1,373        0.74       65.11
                         [1] East German   |      3,687        1.98       67.09
 [2] Person Of German Descent From Easte   |     28,029       15.06       82.15
              [3] German Who Lived Abroad  |      1,195        0.64       82.79
 [4] Citizen Of EU Country (up to 2009 E   |      6,935        3.73       86.52
                [5] Asylum seeker, refugee |      9,419        5.06       91.58
                     [6] Other Foreigner   |     15,681        8.42      100.00
-------------------------------------------+-----------------------------------
                                     Total |    186,136      100.00
```

d) On the basis of this variable, generate the variable "escape", which only distinguishes among three groups:

- 0 = Cases where no information is available
- 1 = All persons without escape background
- 2 = Asylum seekers / refugees

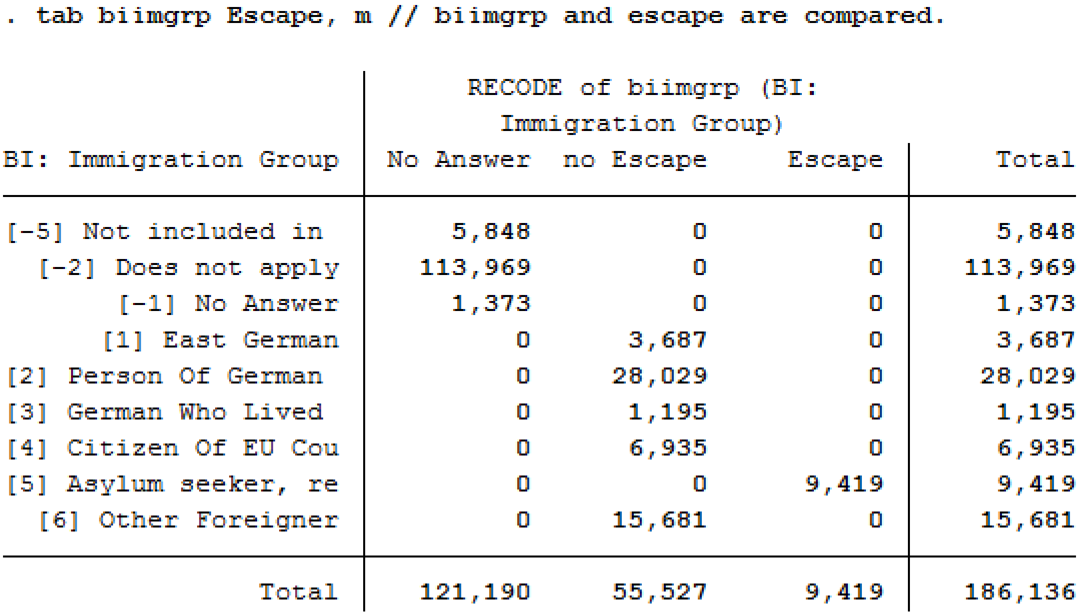
After you have familiarized yourself with the variable, recode it to fit your project. Then check the case numbers of your generated variable with the source variable.

```
/*
d)        On the basis of this variable, generate the variable "Escape", which only distinguishes between three groups:
    0 = Cases where no information is available
    1 = All persons without escape background
    2 = Asylum seekers / refugees
*/

```

(continues on next page)

(continued from previous page)

```
recode biimgrp (-5 -2 -1 = 0 "No Answer") (1 2 3 4 6 = 1 "no Escape") (5 = 2 "Escape"), gen(Escape)
tab biimgrp Escape, m // biimgrp and escape are compared.
```



```
. tab biimgrp Escape, m // biimgrp and escape are compared.

                      |      RECODE of biimgrp (BI:
                      |        Immigration Group)
BI: Immigration Group | No Answer  no Escape     Escape |     Total
----------------------+---------------------------------+----------
[-5] Not included in  |     5,848          0          0 |     5,848
  [-2] Does not apply |   113,969          0          0 |   113,969
       [-1] No Answer |     1,373          0          0 |     1,373
     [1] East German  |         0      3,687          0 |     3,687
[2] Person Of German  |         0     28,029          0 |    28,029
[3] German Who Lived  |         0      1,195          0 |     1,195
[4] Citizen Of EU Cou |         0      6,935          0 |     6,935
[5] Asylum seeker, re |         0          0      9,419 |     9,419
  [6] Other Foreigner |         0     15,681          0 |    15,681
----------------------+---------------------------------+----------
                Total |   121,190     55,527      9,419 |   186,136
```

e) It may be that initially there is no information on the immigration status but this will change one year later. Limit the dataset to the last observation available on the respective person, since this gives you the most comprehensive information.

```
e)      It may happen that tinitially there is no information on the status of
*   immigration, but this will change in a later year. Limit the data record to
*   the last observation that is available for the respective person, since this
*   way the specification with the most information content is used.
*/

bysort persnr: egen syear_max = max(syear) //A variable is created, which shows the last existing yearly observation
keep if syear_max == syear //Annual observations which are not the last observation are deleted.
```

f) Save the generated data temporarily on your personal drive.

```
f)      Save the generated data record on your personal drive temporarily
*/

save $MY_OUT_TEMP\biimgrp.dta, replace
```

Task 2: Add basic variables from PPATH and weights

Attention: Please note that since version 34 (v34), PPFAD can be found in the subdirectory "Raw" of the data distribution file. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFAD.

a) Load the following information from PPATH:

- Permanent Indivdiual Identifier **"persnr"**
- Household Identifier **"hhnr"** and the current household number **"bghhnr"**
- The net variable with information about the interview type **"bgnetto"**
- The gender of the person **"sex"**
- The year of birth **"gebjahr"**
- Variables on the migration background **"migback"**, **"germborn"**, **"corigin"**, **"immiyear"**
- Information about the survey status: **"psample"**

If you want to familiarize yourself with the PPATH dataset, see the section *Working with Tracking Data (PPATHL)*.

```
/*
a)        Use the following information from PPFAD:
  - Unchanging Person ID „persnr“
  - Household number "hhnr" and the current household number "bghhnr".
  - the net variable with information about the interview type "bgnetto".
  - the gender of the person "sex"
  - the year of birth "semester"
  - Variables on the migration background "migback", "germborn" "corigin" "immiyear"
  - Information about the survey status: "bgnetto" and "psample".
*/

use persnr hhnr bghhnr bgnetto psample sex gebjahr germborn corigin immiyear migback ␣
↪using $MY_IN_PATH\ppfad.dta, clear
```

b) Merge the previously generated data using the individual identifier.

If you don't understand how to create your own cross-sectional dataset, see the chapter *Generating a Cross-Sectional Dataset*.

```
/*
b)        Merge the previously generated data set using the person number.
*/

merge 1:1 persnr using $MY_OUT_TEMP\biimgrp.dta, nogen
```

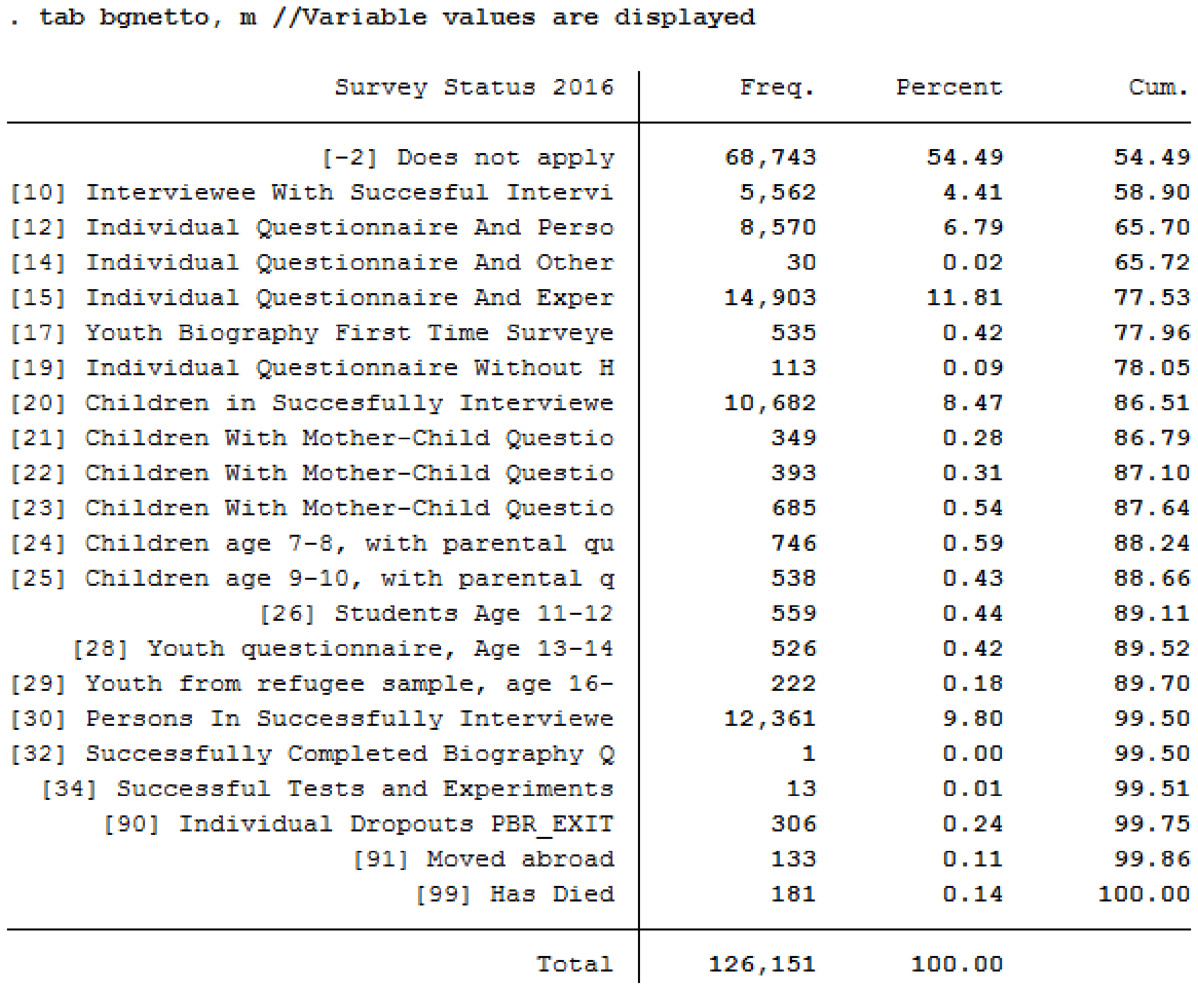
c) Add the corresponding individual extrapolation factors to the data.

```
c)        Add the corresponding data using the individual identifier.
*/

merge 1:1 persnr using $MY_IN_PATH\phrf.dta, keepus(bgphrf) nogen
```

d) Only keep respondents who completed a youth or individual questionnaire in 2016.

For example, to exclude children who have not provided immigration status information, use the net code from PPATH. Only keep individuals who completed an individual or youth interview.

```
/*
d)        Only keep respondents who completed a youth or individual questionnaire in 2016.
*/

tab bgnetto, m //Variable values are displayed

keep if inrange(bgnetto, 10, 19) // People who have a code between 10 and 19 will be kept.
```



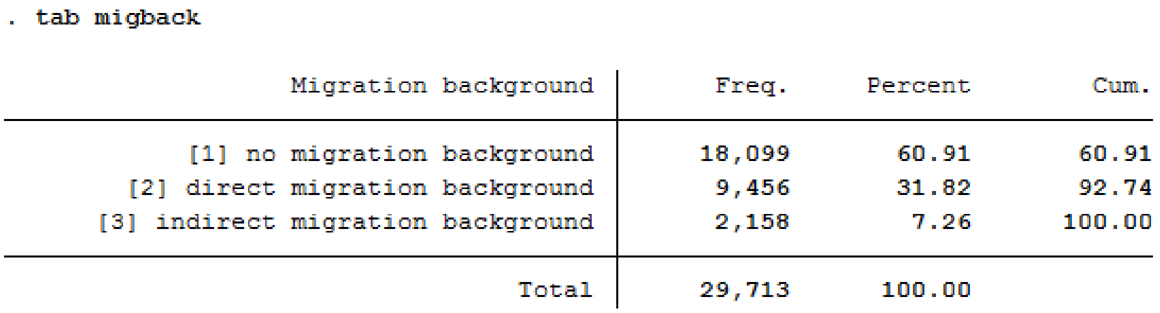
```
. tab bgnetto, m //Variable values are displayed

                   Survey Status 2016 |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
                   [-2] Does not apply |     68,743       54.49       54.49
[10] Interviewee With Succesful Intervi |      5,562        4.41       58.90
[12] Individual Questionnaire And Perso |      8,570        6.79       65.70
[14] Individual Questionnaire And Other |         30        0.02       65.72
[15] Individual Questionnaire And Exper |     14,903       11.81       77.53
[17] Youth Biography First Time Surveye |        535        0.42       77.96
[19] Individual Questionnaire Without H |        113        0.09       78.05
[20] Children in Succesfully Interviewe |     10,682        8.47       86.51
[21] Children With Mother-Child Questio |        349        0.28       86.79
[22] Children With Mother-Child Questio |        393        0.31       87.10
[23] Children With Mother-Child Questio |        685        0.54       87.64
[24] Children age 7-8, with parental qu |        746        0.59       88.24
[25] Children age 9-10, with parental q |        538        0.43       88.66
               [26] Students Age 11-12 |        559        0.44       89.11
     [28] Youth questionnaire, Age 13-14 |        526        0.42       89.52
[29] Youth from refugee sample, age 16- |        222        0.18       89.70
[30] Persons In Successfully Interviewe |     12,361        9.80       99.50
[32] Successfully Completed Biography Q |          1        0.00       99.50
  [34] Successful Tests and Experiments |         13        0.01       99.51
      [90] Individual Dropouts PBR_EXIT |        306        0.24       99.75
                      [91] Moved abroad |        133        0.11       99.86
                          [99] Has Died |        181        0.14      100.00
--------------------------------------+-----------------------------------
                                  Total |    126,151      100.00
```

Task 3: Generate a status variable with the following categories:.

- No migration background
- Migrant, 2nd generation
- Migrant, no information
- Migrant, not refugee
- Migrant, refugee

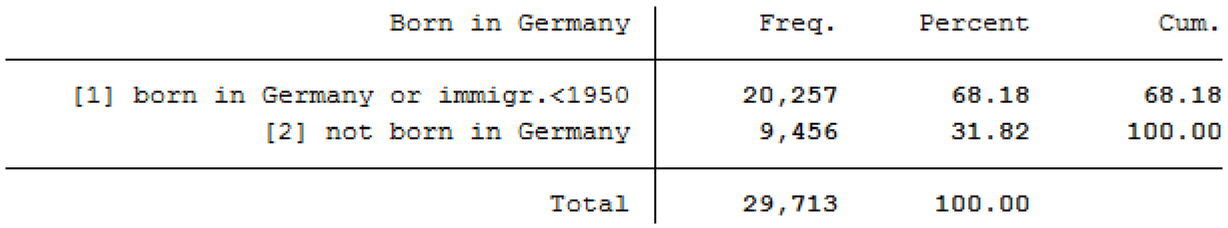
To generate this status variable, check the contents of the existing migration variables from PPATH (migback germborn).

```
/*
Generate a status variable with the following categories:
*/

tab migback
```



```
. tab migback

           Migration background |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
      [1] no migration background |     18,099       60.91       60.91
  [2] direct migration background |      9,456       31.82       92.74
[3] indirect migration background |      2,158        7.26      100.00
--------------------------------+-----------------------------------
                          Total |     29,713      100.00
```

```
tab germborn
```



```
                   Born in Germany |      Freq.     Percent        Cum.
-----------------------------------+-----------------------------------
[1] born in Germany or immigr.<1950 |     20,257       68.18       68.18
            [2] not born in Germany |      9,456       31.82      100.00
-----------------------------------+-----------------------------------
                             Total |     29,713      100.00
```

Use the migration variables from PPATH (migback, germborn) and link this information with your previously generated refugee variable to build the described status variable from Task 3.

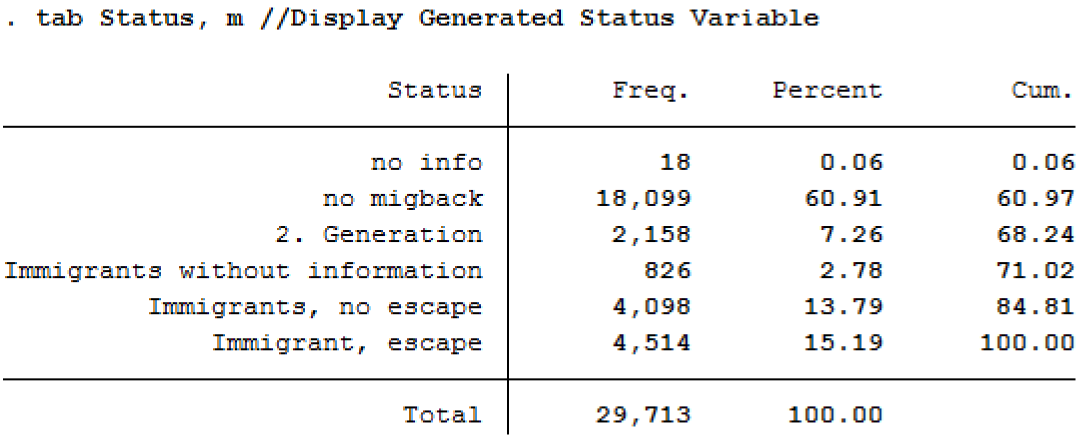
```
gen Status = 0 // All persons will first receive the missing code for "no info".
replace Status = 1 if migback == 1 & germborn == 1 // "no migback"
replace Status = 2 if migback == 3                  // "2nd generation" (2nd generation migrants born by definition in Germany, therefore "& germborn == 1" here unnecessary
replace Status = 3 if germborn == 2 & Escape == 0  // "Immigrants without information"
replace Status = 4 if germborn == 2 & Escape == 1  // "Immigrants, no escape"
replace Status = 5 if germborn == 2 & Escape == 2  // "Immigrant, escape"

label def Statuslbl 0"no info" 1"no migback" 2"2. Generation" 3"Immigrants without information"  4"Immigrants, no escape" 5"Immigrant, escape"
label val Status Statuslbl // Values of the status veriable receive label
```

Task 4: Content analysis:

a) How many refugees (foreign-born with refugee/asylum status) are now in your file?

Look at your status variable previously generated in task 3 to answer the question.

```
*** Exercise 4 ******************************************************************

/*
a)        How many refugees (foreign-born with refugee/asylum status) are now in your file?
*/

tab Status, m //Display Generated Status Variable
```



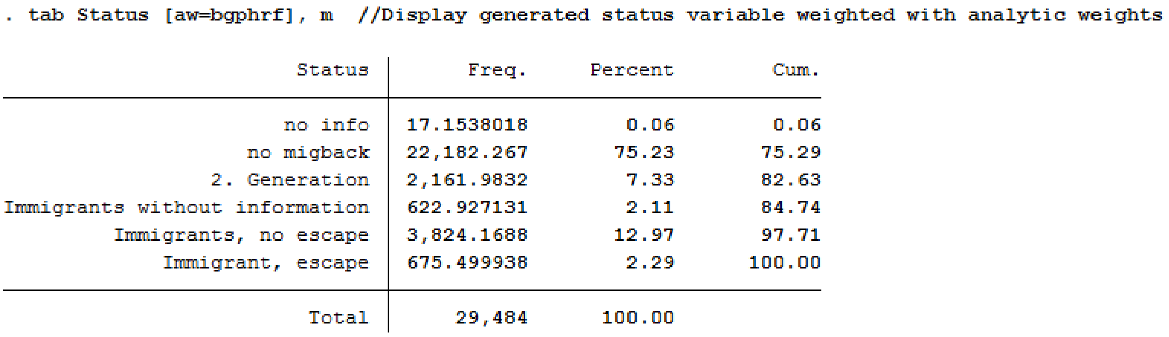
```
. tab Status, m //Display Generated Status Variable

                         Status |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
                        no info |         18        0.06        0.06
                     no migback |     18,099       60.91       60.97
                  2. Generation |      2,158        7.26       68.24
Immigrants without information |        826        2.78       71.02
          Immigrants, no escape |      4,098       13.79       84.81
              Immigrant, escape |      4,514       15.19      100.00
--------------------------------+-----------------------------------
                          Total |     29,713      100.00
```

All 4,514 respondents who received the value 5 for the generated status variable have a direct migration background (migback==2), were not born in Germany (germborn==2), and fled their country of origin (flight==2 and biimgrp==5).

b) How many are there if you take the individual extrapolation factors into account? Interpret the results.

Look at the status variable generated in task 3 to answer the question.

```
/*
b)        How many are there if you take the individual extrapolation factors into account? Interpret the results.
*/

tab Status [aw=bgphrf], m  //Display generated status variable weighted with analytic weights
```



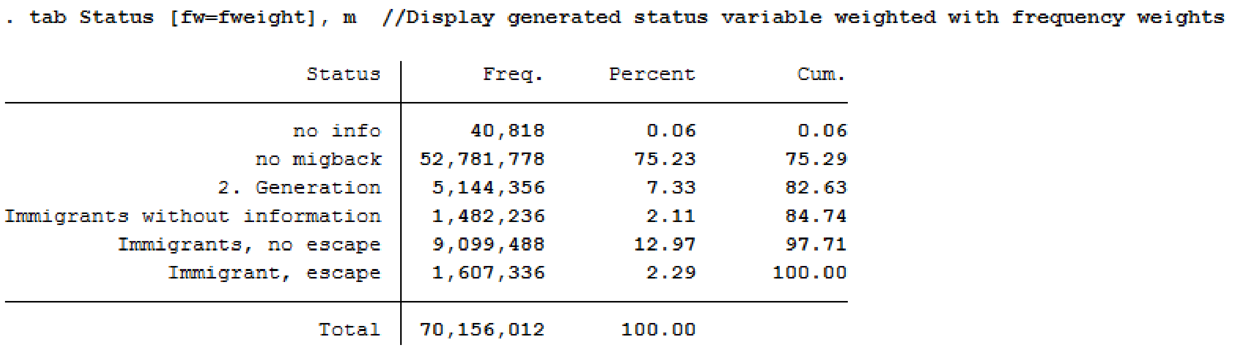
```
. tab Status [aw=bgphrf], m  //Display generated status variable weighted with analytic weights

                         Status |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
                        no info | 17.1538018        0.06        0.06
                     no migback | 22,182.267       75.23       75.29
                  2. Generation | 2,161.9832        7.33       82.63
Immigrants without information | 622.927131        2.11       84.74
          Immigrants, no escape | 3,824.1688       12.97       97.71
              Immigrant, escape | 675.499938        2.29      100.00
--------------------------------+-----------------------------------
                          Total |     29,484      100.00
```

After weighting, there are approximately 675 refugees in the dataset. The weighting thus corrected the number of refugees downwards.

c) How many persons are represented in the sample, taking the extrapolation factors into account?

To use frequency weights in STATA, integer weights are required. Create an integer frequency weight from the weighting factor provided so that you can make representative statements. Then take a look at the new results.

```
/*
c)        How many persons are represented when the sample taking the extrapolation factors into account?
*/

gen fweight = round(bgphrf) //Frequency weights for stata require integer weight
tab Status [fw=fweight], m  //Display generated status variable weighted with frequency weights
```



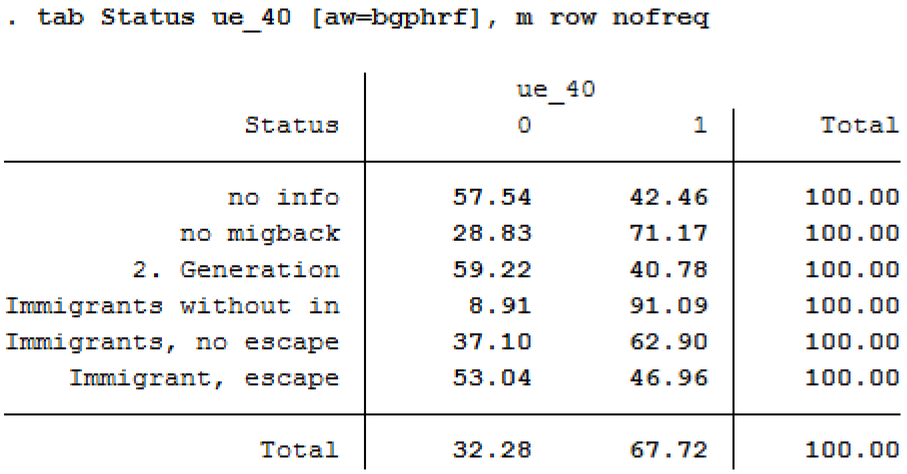
```
. tab Status [fw=fweight], m  //Display generated status variable weighted with frequency weights

                     Status |      Freq.     Percent        Cum.
----------------------------+-----------------------------------
                    no info |     40,818        0.06        0.06
                 no migback | 52,781,778       75.23       75.29
              2. Generation |  5,144,356        7.33       82.63
Immigrants without information |  1,482,236        2.11       84.74
      Immigrants, no escape |  9,099,488       12.97       97.71
          Immigrant, escape |  1,607,336        2.29      100.00
----------------------------+-----------------------------------
                      Total | 70,156,012      100.00
```

Around 1,600,000 people are represented.

d) What is the proportion of people over 40 years of age among the refugees?

Since the data in this exercise come from the wave “bg”, we are currently in the survey year 2016; if you need a description of the wave designations, please refer to the chapter Label. To generate a suitable age variable, you can use the year of birth (year of birth). If we look at the survey year 2016, all persons born in 1976 or earlier were over 40 years old. Generate a suitable age variable and look at the proportion of refugees over 40 years of age in weighted form:

```
/*
d)        What is the proportion of people over 40 years of age among the refugees?
*/

gen ue_40 = 0
replace ue_40 = 1 if gebjahr <= 1976 // Persons receive proficiency 1 if they were born before 1975.

tab Status ue_40 [aw=bgphrf], m row nofreq
```



```
. tab Status ue_40 [aw=bgphrf], m row nofreq

                      |        ue_40
               Status |         0          1 |     Total
----------------------+----------------------+----------
              no info |     57.54      42.46 |    100.00
           no migback |     28.83      71.17 |    100.00
        2. Generation |     59.22      40.78 |    100.00
Immigrants without in |      8.91      91.09 |    100.00
Immigrants, no escape |     37.10      62.90 |    100.00
    Immigrant, escape |     53.04      46.96 |    100.00
----------------------+----------------------+----------
                Total |     32.28      67.72 |    100.00
```

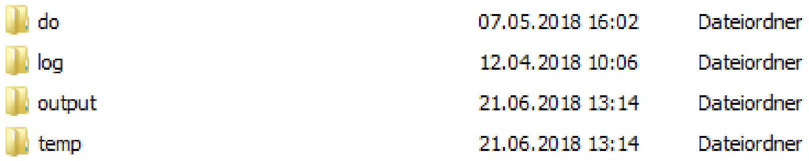
The proportion of refugees over 40 years of age is about 47%.

Last change: Sep 26, 2022

6.8 Fixed Effects Estimation

Let's say you want to find out whether certain variables relevant to the labor market, such as work experience or time in education, influence a person's hourly wage. Other variables such as gender or marital status should also be taken into account. You decide to use the SOEP data to set up a fixed effects estimation model.

Create a path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
*************************************************
* Set relative paths to the working directory
*************************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\distribution\soep-long\soep.v33.1\stata_en\"
global MY_DO_FILES "$AVZ\do\"
```

(continues on next page)

(continued from previous page)

```
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global “AVZ” defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to your data.

a) Generate your own SOEPwage.dta dataset. The dataset should contain information on gross monthly wage, marital status, and other personal characteristics.

To perform your analysis, you need different SOEP variables. The SOEP offers various options for a variable search:

- Search the questionnaires for useful variables. (For more information, see the section *Variable Search with Questionnaires*)
- Find a suitable variable in the topic list at paneldata.org (for more information, see the section *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information, see the section *Variable Search with paneldata.org*)
- Use the documentation provided for the generated variables (for more information, seethe section *Documentation on Generated Data*)

Use the various important variables of the ppfadl.dta dataset as your start file. Your source file should contain the following variables:

- Individual identifier **"pid"**
- Survey year **"syear"**
- Birth Year **"gebjahr"**
- The net variable with information on the interview type **"netto"**
- The weighting variable **"phrf"**
- The gender of the person **"sex"**
- Sample membership **"pop"**

```
use pid syear sex gebjahr netto pop phrf using "${MY_IN_PATH}/ppfadl.dta", clear
```

Attention: Please note that since version 34 (v34), PPFADL has been renamed PPATHL. The following ecxercises are done with version 33.1 (v33.1), where the tracking file was named PPFADL.

Apply the necessary content variables to your starting dataset. You need the following variables for your analysis:

- Employment status plb0022_h
- Current gross income in euros **"pglabgro"**
- Actual weekly working hours **"pgtatzeit"**
- Full-time work experience **"pgexpft"**
- Years of education or training **"pgbilzeit"**
- Marital status in survey year **"pgfamstd"**

```
merge 1:1 pid syear using "${MY_IN_PATH}/pl.dta", keepus(plb0022) keep(master match)␣
↪nogen
merge 1:1 pid syear using "${MY_IN_PATH}/pgen.dta", keepus(pglabgro pgtatzeit pgexpft␣
↪pgbilzeit pgfamstd) keep(master match) nogen
```

Only keep people who have completed an interview and who live in a private household.

```
* Only select people with completed interviews
keep if inrange(netto, 10, 19)

* Only private households
keep if pop==1 | pop==2
```

Since you are only interested in the period from 2012 to 2016, remove all survey information that does not fall within this period. To finish, save your dataset.

```
* Period from 2012 to 2016
keep if syear>=2012 & syear<=2016
```

Exercise 1: Prepare your dataset

a) Load your created SOEPWage.dta dataset. It contains information on gross monthly wage, marital status, and other personal characteristics.

```
*** Exercise 1: Prepare your dataset
* a) Load data set
use "${MY_OUT_DATA}/SOEPWage.dta", clear
```

b) Recode all missing values in systemmissings (.)

```
* b) Recode Missings
mvdecode _all, mv(-8/-1 = .)
```

For more information about the missing codes for SOEP data, see the chapter *Missing Conventions*

c) Generate the variables "hourly wage" (gross monthly wage/4.33*working time) for persons who have earned at least 1 euro and have worked at least one hour, "Married vs. Unmarried" and age.

```
* c) Generate Variables
gen wage = pglabgro/(4.33*pgtatzeit) if pglabgro>=1 & pgtatzeit>=1

gen married = 1 if pgfamstd==1 | pgfamstd==6 | pgfamstd==7 | pgfamstd==8
replace married = 0 if inrange(pgfamstd, 2, 5)

gen age = syear - gebjahr
```

d) Adjust the variable "hourly wage" from outlier values by setting values smaller than the first percentile to the same value. Set values greater than 3 times the 99th percentile to 3*99th percentile. Then generate the variable lwage = log(wage).

```
* d) Adjust wage variable
sum wage, detail
replace wage = 1/3*r(p1) if wage<1/3*r(p1)
replace wage = 3*r(p99) if wage>3*r(p99) & wage<.

```

(continues on next page)

(continued from previous page)

```
gen lwage = log(wage)
label variable lwage "Log hourly wage"

save "${MY_OUT_DATA}/SOEPWage_temp.dta", replace
```

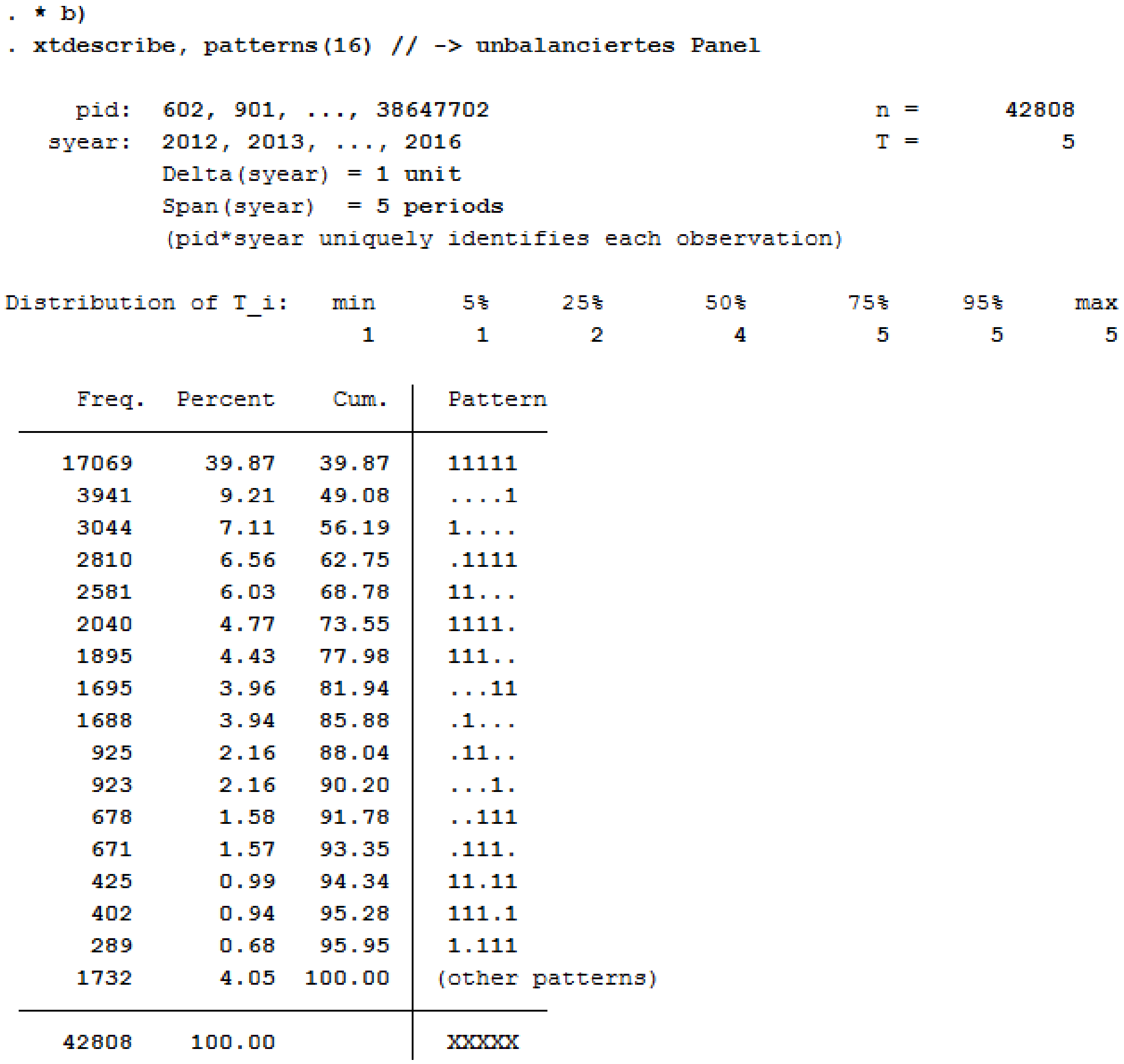
Exercise 2: Descriptive statistics

a) Define the dataset as a panel dataset.

```
*** Exercise 2: Descriptive statistics
* a)
xtset pid syear // Declaring data as panel data
```

b) What percentage of people participated in all five waves (xtdescribe)

```
* b)
xtdescribe, patterns(16) // -> unbalanced panel
```

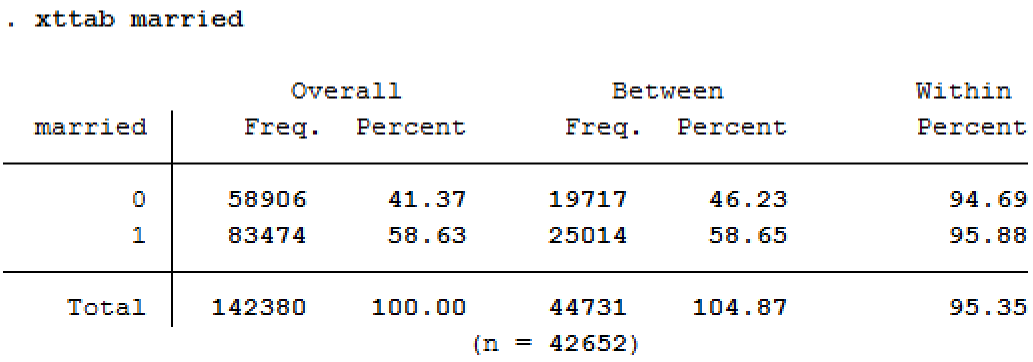


```
. * b)
. xtdescribe, patterns(16) // -> unbalanciertes Panel

     pid:  602, 901, ..., 38647702                          n =      42808
   syear:  2012, 2013, ..., 2016                            T =          5
           Delta(syear) = 1 unit
           Span(syear)  = 5 periods
           (pid*syear uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         4         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
    17069     39.87   39.87 |  11111
     3941      9.21   49.08 |  ....1
     3044      7.11   56.19 |  1....
     2810      6.56   62.75 |  .1111
     2581      6.03   68.78 |  11...
     2040      4.77   73.55 |  1111.
     1895      4.43   77.98 |  111..
     1695      3.96   81.94 |  ...11
     1688      3.94   85.88 |  .1...
      925      2.16   88.04 |  .11..
      923      2.16   90.20 |  ...1.
      678      1.58   91.78 |  ..111
      671      1.57   93.35 |  .111.
      425      0.99   94.34 |  11.11
      402      0.94   95.28 |  111.1
      289      0.68   95.95 |  1.111
     1732      4.05  100.00 | (other patterns)
 ---------------------------+---------
    42808    100.00         |  XXXXX
```

42808 respondents have contributed information within waves bc (2012) - bg (2016) and about 40% (17069) of the 42808 respondents have provided information for all waves.

c) Describe the variable "Married" with xttab and xttrans. Take a look at some individual wage (pid=30320901, pid=30932501, pid==3101602, pid==3101801) developments with xtline.

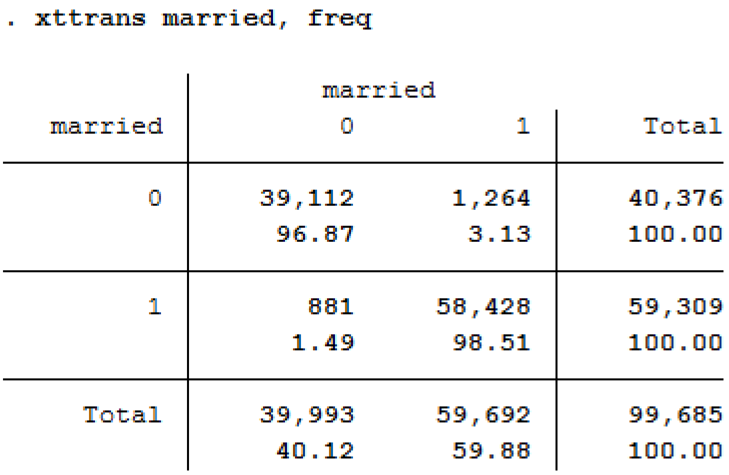
```
* c)
* Stability of the relationship status
xttab married
```



```
. xttab married

                  Overall             Between            Within
married |    Freq.  Percent      Freq.  Percent         Percent
--------+---------------------------------------------------------
      0 |    58906    41.37      19717    46.23           94.69
      1 |    83474    58.63      25014    58.65           95.88
--------+---------------------------------------------------------
  Total |   142380   100.00      44731   104.87           95.35
                                 (n = 42652)
```

You can observe 41.37 percent of person-year observations with "married==no". Within the period from 2012 to 2016, 19717 people responded at least once that they were not married. During the same period, 25014 persons reported at least once that they were married. Those who were not married for at least one year responded with "married==no" in 94.69% of the observations, whereas those who were married at least once responded in 95.88 percent of the observations with "married==yes". This indicates very stable response behavior.

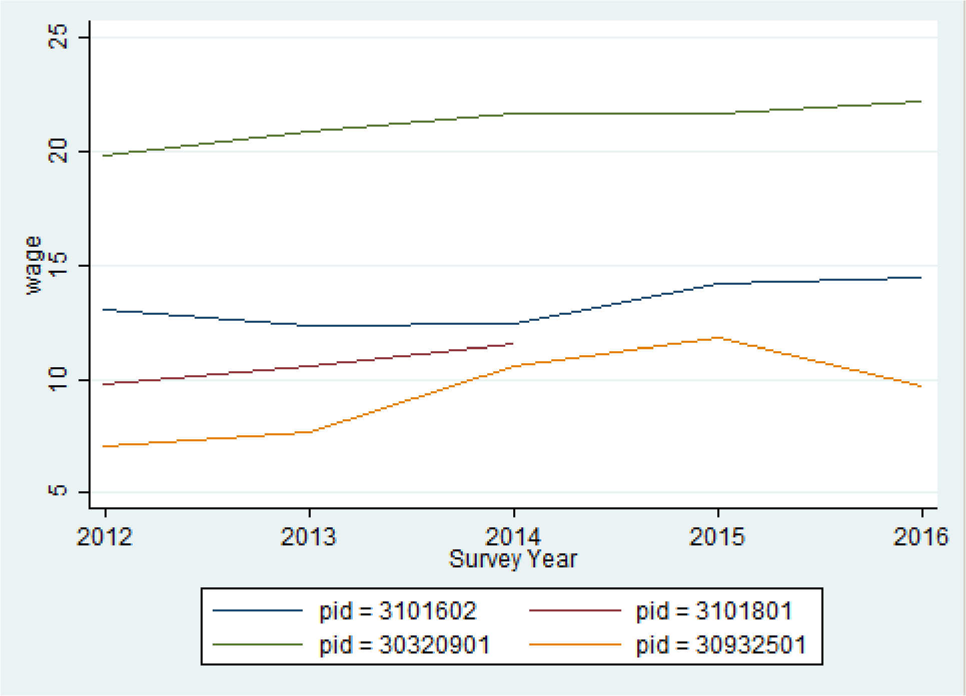
```
* Transition probabilities
xttrans married, freq
```



```
. xttrans married, freq

           |        married
   married |         0          1 |     Total
-----------+----------------------+----------
         0 |    39,112      1,264 |    40,376
           |     96.87       3.13 |    100.00
-----------+----------------------+----------
         1 |       881     58,428 |    59,309
           |      1.49      98.51 |    100.00
-----------+----------------------+----------
     Total |    39,993     59,692 |    99,685
           |     40.12      59.88 |    100.00
```

96.87 percent of the person-year observations with "married==no" are still not married in the next period. 98.51 percent of the persons who are married indicate that they will also be married in the following period. This is evidence of stable response behavior.

```
* Individual sequences of "wage"
xtline wage if pid==30320901 | pid==30932501 | pid==3101602 | pid==3101801, overlay
```

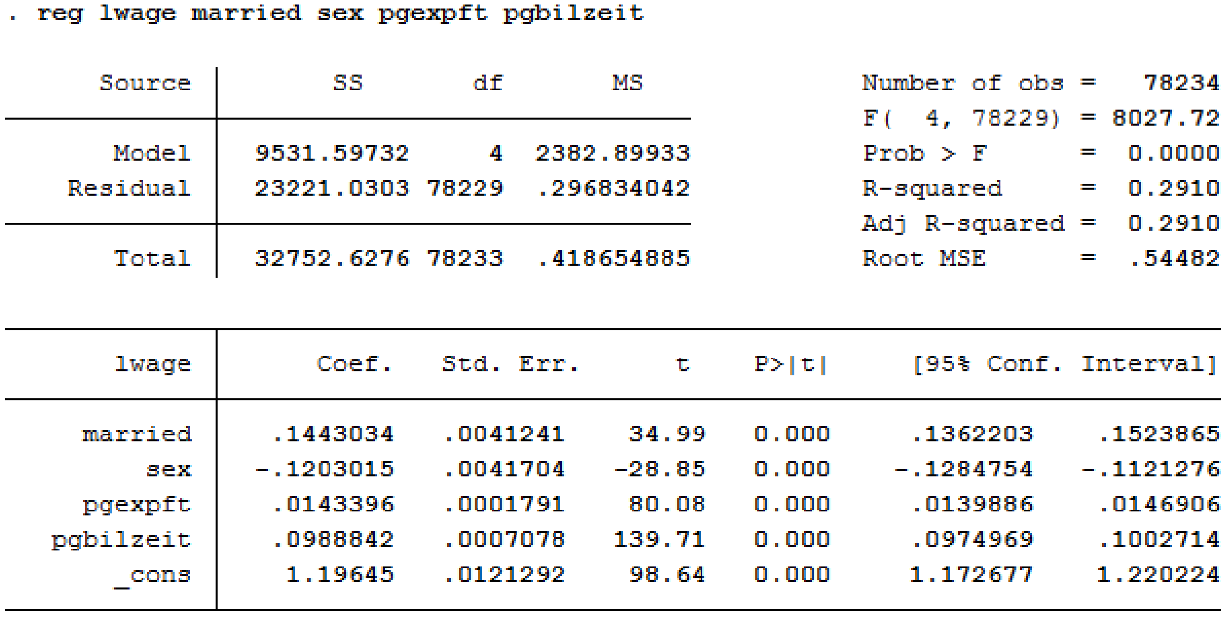


The graphic shows a comparison of the hourly wage for four different respondents.

Exercise 3: Pooled OLS Regression

a) Execute a pooled OLS regression with "log hourly wage" as dependent variable and "married", "gender", "work experience" and "training time" as independent variables. Interpret the coefficients for "married", "gender" and "length of training". Why are these not causal effects?

```
*** Exercise 3: Pooled OLS Regression
* a) Pooled OLS
reg lwage married sex pgexpft pgbilzeit
```

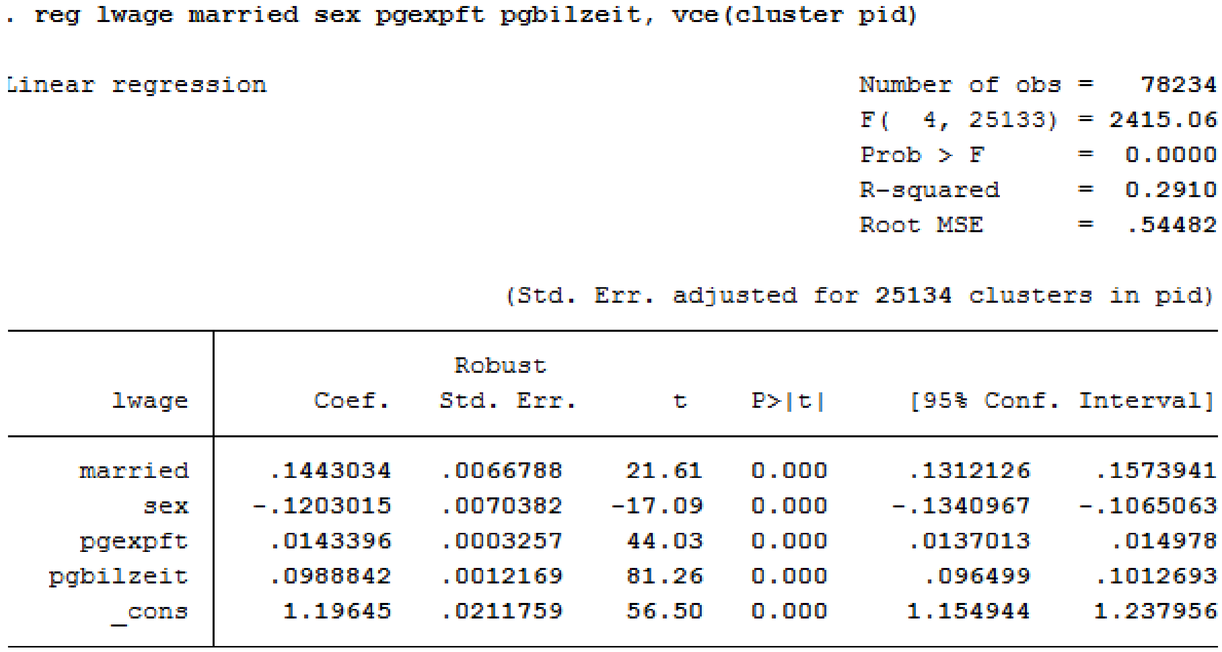


```
. reg lwage married sex pgexpft pgbilzeit

      Source |       SS       df       MS              Number of obs =   78234
-------------+------------------------------           F(  4, 78229) = 8027.72
       Model |  9531.59732     4  2382.89933           Prob > F      =  0.0000
    Residual |  23221.0303 78229  .296834042           R-squared     =  0.2910
-------------+------------------------------           Adj R-squared =  0.2910
       Total |  32752.6276 78233  .418654885           Root MSE      =  .54482

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .1443034   .0041241    34.99   0.000     .1362203    .1523865
         sex |  -.1203015   .0041704   -28.85   0.000    -.1284754   -.1121276
     pgexpft |   .0143396   .0001791    80.08   0.000     .0139886    .0146906
   pgbilzeit |   .0988842   .0007078   139.71   0.000     .0974969    .1002714
       _cons |    1.19645   .0121292    98.64   0.000     1.172677    1.220224
------------------------------------------------------------------------------
```

The variables married, sex, and pgbilzeit most likely correlate with other disregarded/unobserved variables that have an effect on the wage. For example, women more often work in occupations with lower wages.

b) Run the regression again with the option "vce(cluster persnr)" to get clustered standard errors. How do the standard errors of the coefficients change?

```
* b) Pooled OLS with cluster standard errors
reg lwage married sex pgexpft pgbilzeit, vce(cluster pid)
```



```
. reg lwage married sex pgexpft pgbilzeit, vce(cluster pid)

Linear regression                                      Number of obs =   78234
                                                       F(  4, 25133) = 2415.06
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2910
                                                       Root MSE      =  .54482

                               (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .1443034   .0066788    21.61   0.000     .1312126    .1573941
         sex |  -.1203015   .0070382   -17.09   0.000    -.1340967   -.1065063
     pgexpft |   .0143396   .0003257    44.03   0.000     .0137013     .014978
   pgbilzeit |   .0988842   .0012169    81.26   0.000      .096499    .1012693
       _cons |    1.19645   .0211759    56.50   0.000     1.154944    1.237956
------------------------------------------------------------------------------
```

The standard errors are getting bigger.

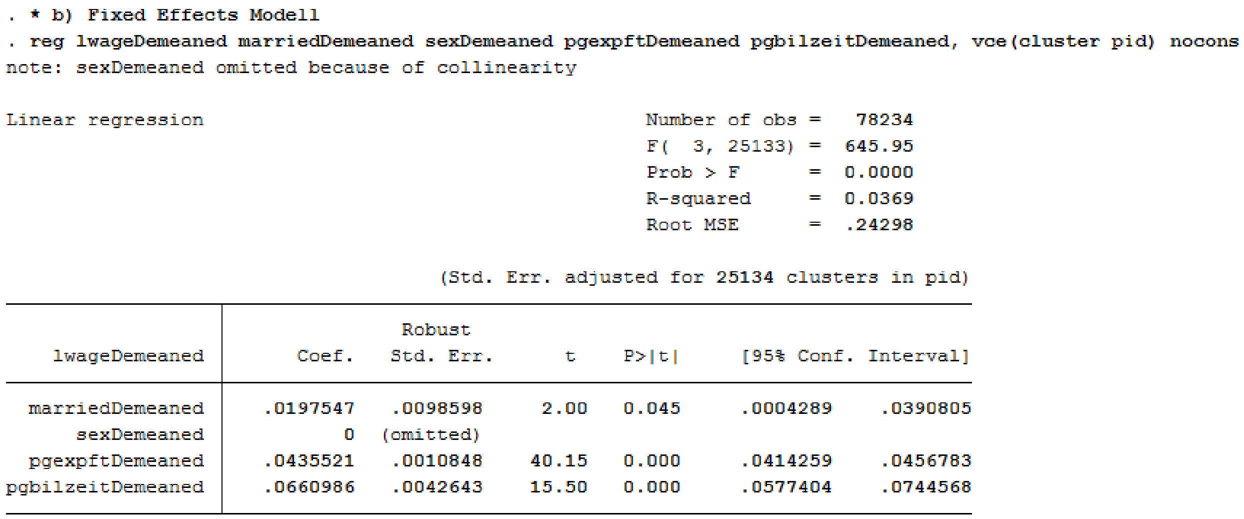
Exercise 4: Fixed Effects

a) Subtract the person-specific mean value from each variable of the model. Use the "egen" function. Ideally you should also use a loop.

```
*** Exercise 4: Fixed Effects
* a) Subtract person-specific averages

gen sample = 1
foreach var in lwage married sex pgexpft pgbilzeit {

        bysort pid: egen `var'Mean = mean(`var')
        replace `var'Mean = . if `var'==.
        gen `var'Demeaned =  `var' - `var'Mean
        replace sample = 0 if `var'==.
}
bysort pid (sample): replace sample = sample[1]
```

b) Estimate the fixed effects model with the previously generated variables. Why isn't a coefficient estimated for "gender"? How do the coefficients change compared to the pooled OLS estimate? Is the effect of "married" now causally interpretable?

```
reg lwageDemeaned marriedDemeaned sexDemeaned pgexpftDemeaned pgbilzeitDemeaned, vce(cluster pid) nocons
```

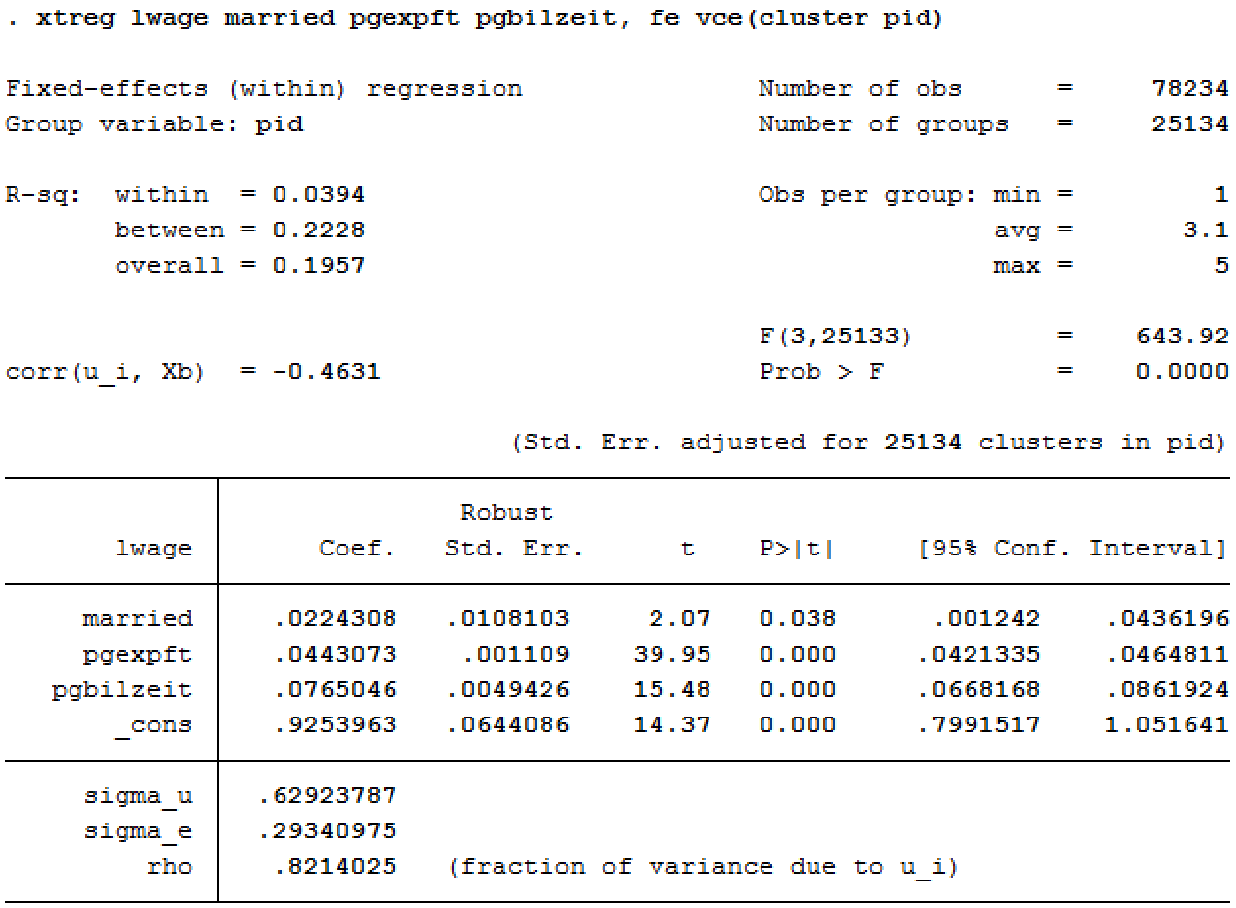


```
. * b) Fixed Effects Modell
. reg lwageDemeaned marriedDemeaned sexDemeaned pgexpftDemeaned pgbilzeitDemeaned, vce(cluster pid) nocons
note: sexDemeaned omitted because of collinearity

Linear regression                               Number of obs =     78234
                                                F(  3, 25133) =    645.95
                                                Prob > F      =    0.0000
                                                R-squared     =    0.0369
                                                Root MSE      =    .24298

                                (Std. Err. adjusted for 25134 clusters in pid)

                                   Robust
    lwageDemeaned       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

  marriedDemeaned    .0197547   .0098598     2.00   0.045     .0004289    .0390805
      sexDemeaned           0  (omitted)
  pgexpftDemeaned    .0435521   .0010848    40.15   0.000     .0414259    .0456783
pgbilzeitDemeaned    .0660986   .0042643    15.50   0.000     .0577404    .0744568
```

No coefficient was estimated for gender because gender was stable over time for all observations. The coefficient of married is now significant at the 5% level!

c) Now estimate the fixed effects model using the command "xtreg lwage married sex pgexpft pgbilzeit, fe". What do you notice about the coefficients compared to task 4 b)? And with the standard errors?

```
* c) xtreg, fe
xtreg lwage married pgexpft pgbilzeit, fe vce(cluster pid)
```

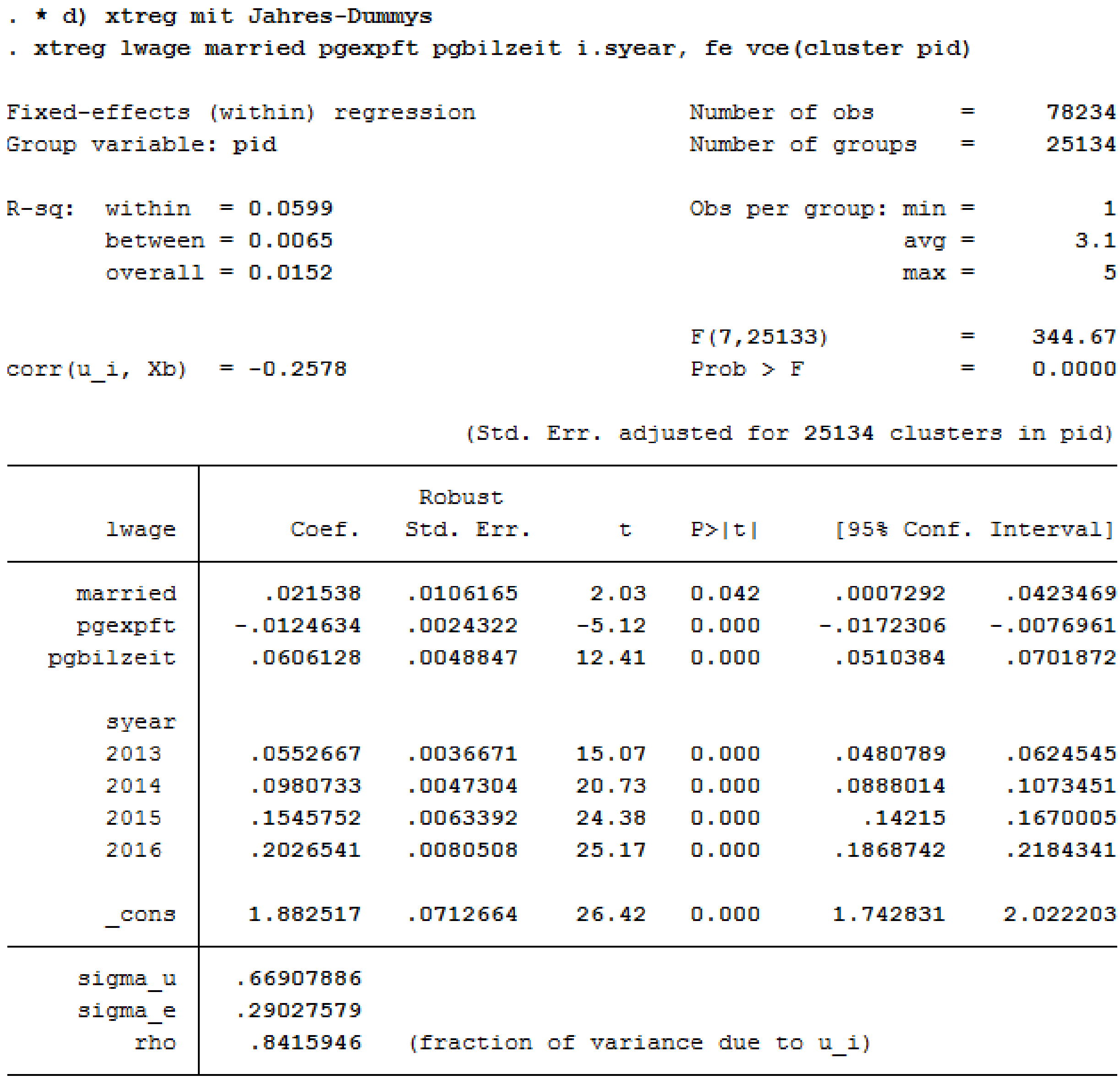


```
. xtreg lwage married pgexpft pgbilzeit, fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs      =     78234
Group variable: pid                             Number of groups   =     25134

R-sq:  within  = 0.0394                         Obs per group: min =         1
       between = 0.2228                                        avg =       3.1
       overall = 0.1957                                        max =         5

                                                F(3,25133)         =    643.92
corr(u_i, Xb)  = -0.4631                        Prob > F           =    0.0000

                              (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0224308   .0108103     2.07   0.038      .001242    .0436196
     pgexpft |   .0443073    .001109    39.95   0.000     .0421335    .0464811
   pgbilzeit |   .0765046   .0049426    15.48   0.000     .0668168    .0861924
       _cons |   .9253963   .0644086    14.37   0.000     .7991517    1.051641
-------------+----------------------------------------------------------------
     sigma_u |  .62923787
     sigma_e |  .29340975
         rho |   .8214025   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The coefficients are not identical to 4 b) and the standard errors become larger because model b) does not take into account the estimation of mean values in the standard errors.

d) Now add dummy variables for the years (i.syear). What happens to the effect of "labor market experience"?

```
* d) xtreg with dummy
xtreg lwage married pgexpft pgbilzeit i.syear, fe vce(cluster pid)
```

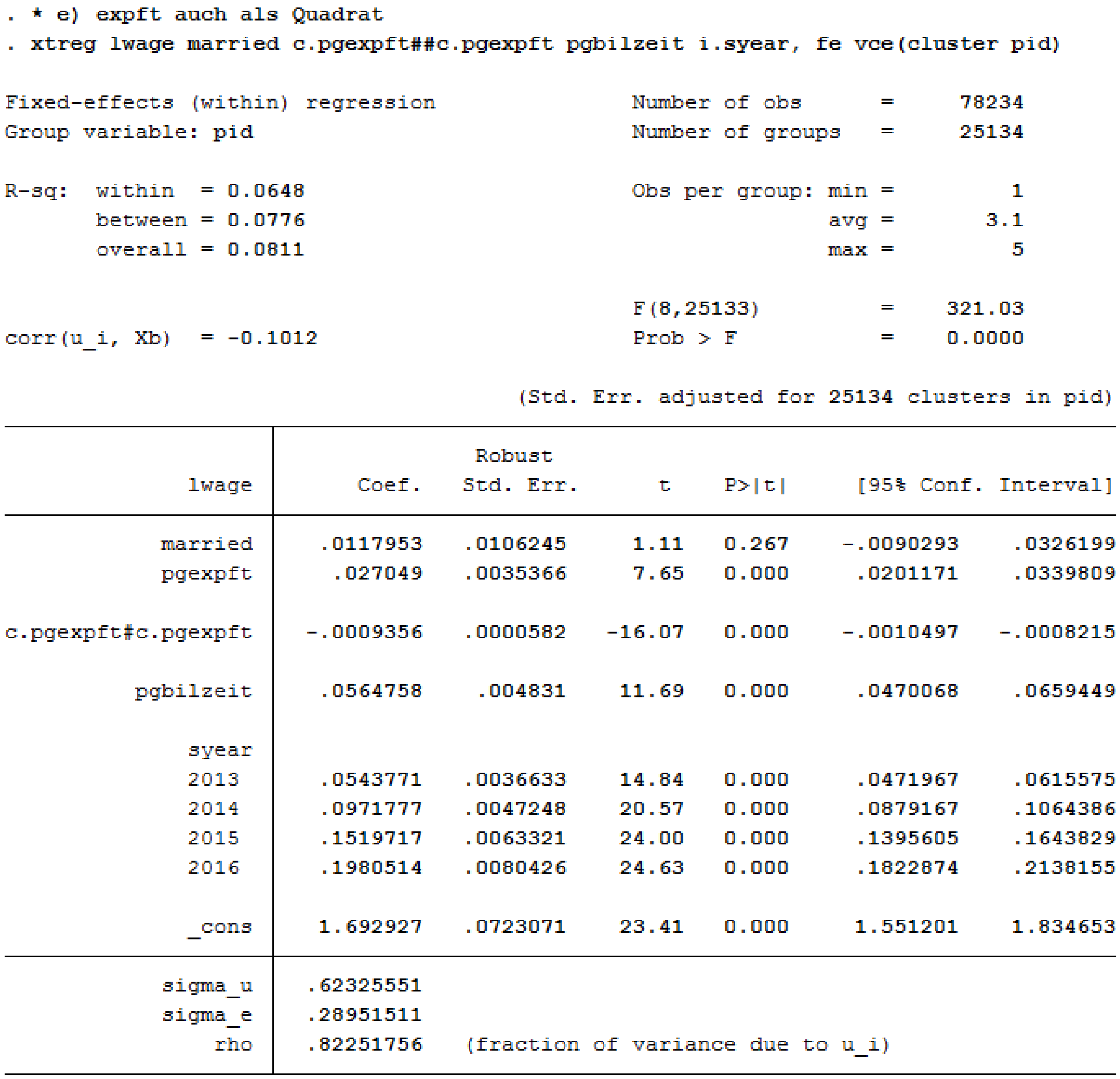


```
. * d) xtreg mit Jahres-Dummys
. xtreg lwage married pgexpft pgbilzeit i.syear, fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs     =     78234
Group variable: pid                             Number of groups  =     25134

R-sq:  within  = 0.0599                         Obs per group: min =         1
       between = 0.0065                                        avg =       3.1
       overall = 0.0152                                        max =         5

                                                F(7,25133)        =    344.67
corr(u_i, Xb)  = -0.2578                        Prob > F          =    0.0000

                             (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |    .021538   .0106165     2.03   0.042     .0007292    .0423469
     pgexpft |  -.0124634   .0024322    -5.12   0.000    -.0172306   -.0076961
   pgbilzeit |   .0606128   .0048847    12.41   0.000     .0510384    .0701872
             |
       syear |
       2013  |   .0552667   .0036671    15.07   0.000     .0480789    .0624545
       2014  |   .0980733   .0047304    20.73   0.000     .0888014    .1073451
       2015  |   .1545752   .0063392    24.38   0.000       .14215    .1670005
       2016  |   .2026541   .0080508    25.17   0.000     .1868742    .2184341
             |
       _cons |   1.882517   .0712664    26.42   0.000     1.742831    2.022203
-------------+----------------------------------------------------------------
     sigma_u |  .66907886
     sigma_e |  .29027579
         rho |   .8415946   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Effects on the variables remain significant. The model could possibly be specified on a case-by-case basis. The Mincer equation is based on (potential) labor market experience squared.

e) Now you can also square labor market experience into the model. To what extent does the effect of labor market experience change compared to task 5d)?

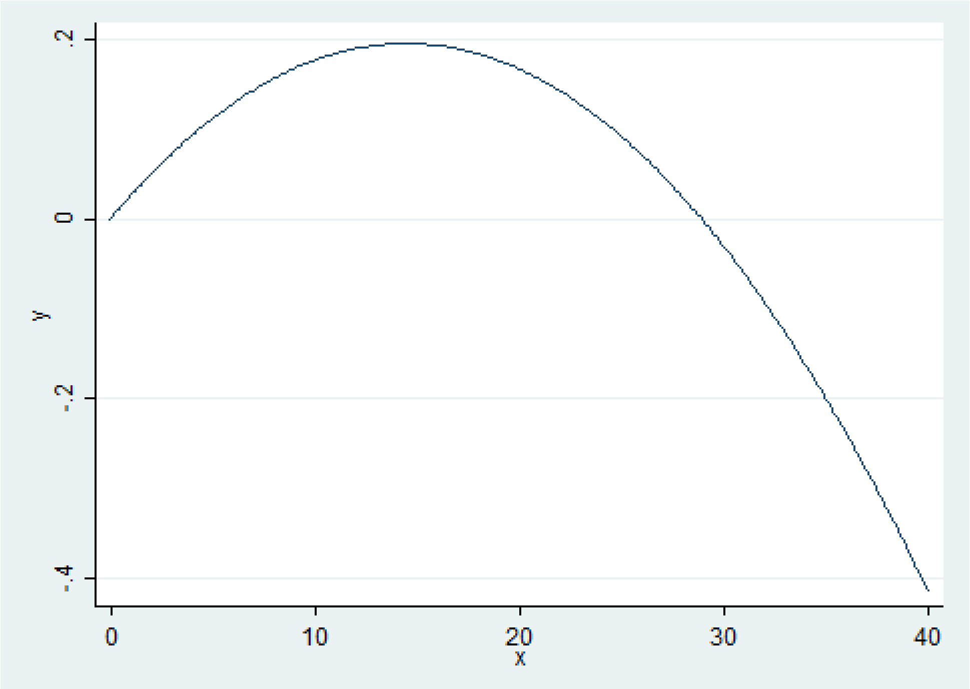
```
* e) expft squared
xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear, fe vce(cluster pid)
```



```
. * e) expft auch als Quadrat
. xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear, fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs     =     78234
Group variable: pid                             Number of groups  =     25134

R-sq:  within  = 0.0648                         Obs per group: min =         1
       between = 0.0776                                        avg =       3.1
       overall = 0.0811                                        max =         5

                                                F(8,25133)        =    321.03
corr(u_i, Xb)  = -0.1012                        Prob > F          =    0.0000

                              (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------------
                   |               Robust
             lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
           married |   .0117953   .0106245     1.11   0.267    -.0090293    .0326199
           pgexpft |    .027049   .0035366     7.65   0.000     .0201171    .0339809
                   |
c.pgexpft#c.pgexpft |  -.0009356   .0000582   -16.07   0.000    -.0010497   -.0008215
                   |
         pgbilzeit |   .0564758    .004831    11.69   0.000     .0470068    .0659449
                   |
             syear |
             2013  |   .0543771   .0036633    14.84   0.000     .0471967    .0615575
             2014  |   .0971777   .0047248    20.57   0.000     .0879167    .1064386
             2015  |   .1519717   .0063321    24.00   0.000     .1395605    .1643829
             2016  |   .1980514   .0080426    24.63   0.000     .1822874    .2138155
                   |
             _cons |   1.692927   .0723071    23.41   0.000     1.551201    1.834653
-------------------+----------------------------------------------------------------
           sigma_u |  .62325551
           sigma_e |  .28951511
               rho |  .82251756   (fraction of variance due to u_i)
------------------------------------------------------------------------------------
```

The coefficients of pgexpft and pgexpft^2 remain significant, whereas the coefficient for married is no longer significant.

```
graph twoway (func y = _b[pgexpft]*x + _b[c.pgexpft#c.pgexpft]*x*x, range(0 40))
```



The graph shows that the effects of the labor market experience decrease after approximately 15 years of professional experience.

f) Now estimate the model from task 5e) with longitudinal section weights. Why is the number of cases now significantly smaller? Why could the coefficient of "pgbilzeit" have changed?

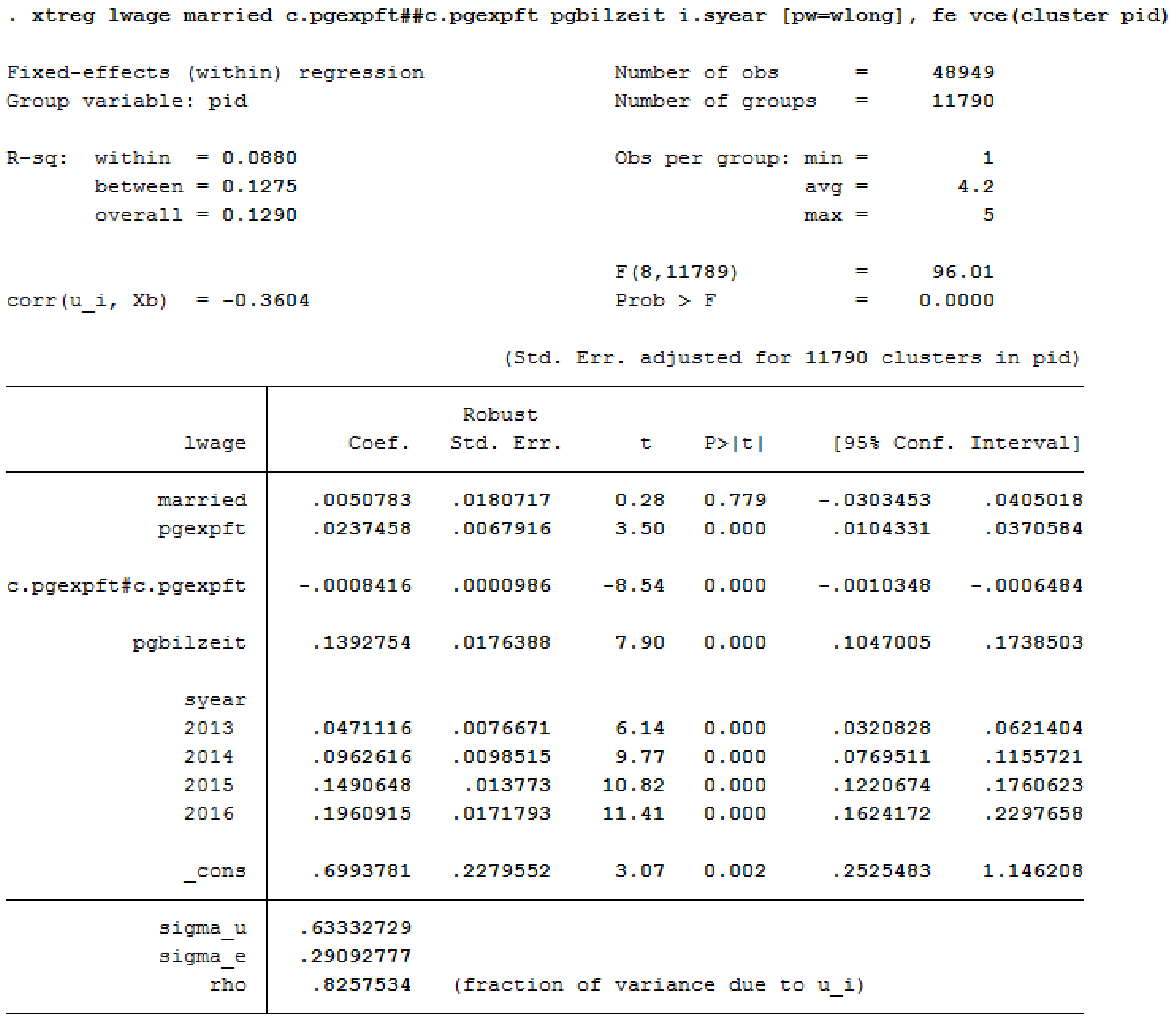
Tip: Create your own longitudinal person weights, e.g., longitudinal person weight from wave A to wave D. Take the starting wave cross-sectional weight (aphrf) and multiply through by each following wave staying factor, as in the following example: gen adphrf=aphrf*bpbleib*cpbleib*dpbleib

Since you are looking at the period 2012-2016, you must create a suitable longitudinal weight. To do this, use the phrf dataset from the RAW subdirectory. Apply the required variables to your analysis dataset and generate your period-related longitudinal section weight. To understand the structure of the data distribution file and the location of the different datasets, visit the section *Data Distribution File*. For more information about the weighting datasets and other survey datasets, see the section *Survey Data*.

```
* f) Fixed Effects weighted
global MY_IN_PATH2 "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
rename pid persnr
merge m:1 persnr using "${MY_IN_PATH2}/phrf.dta", nogen keep(master match) keepus(bcphrf bdpbleib bepbleib bfpbleib bgpbleib)
gen wlong = bcphrf*bdpbleib*bepbleib*bfpbleib*bgpbleib
label variable wlong "Weighting BC-BG"
rename persnr pid
```

Now estimate the model from 5e) and use the created weight.

```
xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear [pw=wlong], fe vce(cluster pid)
```



```
. xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear [pw=wlong], fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs     =     48949
Group variable: pid                             Number of groups  =     11790

R-sq:  within  = 0.0880                         Obs per group: min =         1
       between = 0.1275                                        avg =       4.2
       overall = 0.1290                                        max =         5

                                                F(8,11789)        =     96.01
corr(u_i, Xb)  = -0.3604                        Prob > F          =    0.0000

                                   (Std. Err. adjusted for 11790 clusters in pid)
------------------------------------------------------------------------------------
                    |               Robust
              lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
            married |   .0050783   .0180717     0.28   0.779    -.0303453    .0405018
            pgexpft |   .0237458   .0067916     3.50   0.000     .0104331    .0370584
                    |
c.pgexpft#c.pgexpft |  -.0008416   .0000986    -8.54   0.000    -.0010348   -.0006484
                    |
          pgbilzeit |   .1392754   .0176388     7.90   0.000     .1047005    .1738503
                    |
              syear |
              2013  |   .0471116   .0076671     6.14   0.000     .0320828    .0621404
              2014  |   .0962616   .0098515     9.77   0.000     .0769511    .1155721
              2015  |   .1490648    .013773    10.82   0.000     .1220674    .1760623
              2016  |   .1960915   .0171793    11.41   0.000     .1624172    .2297658
                    |
              _cons |   .6993781   .2279552     3.07   0.002     .2525483    1.146208
--------------------+---------------------------------------------------------------
            sigma_u |  .63332729
            sigma_e |  .29092777
                rho |   .8257534   (fraction of variance due to u_i)
------------------------------------------------------------------------------------
```

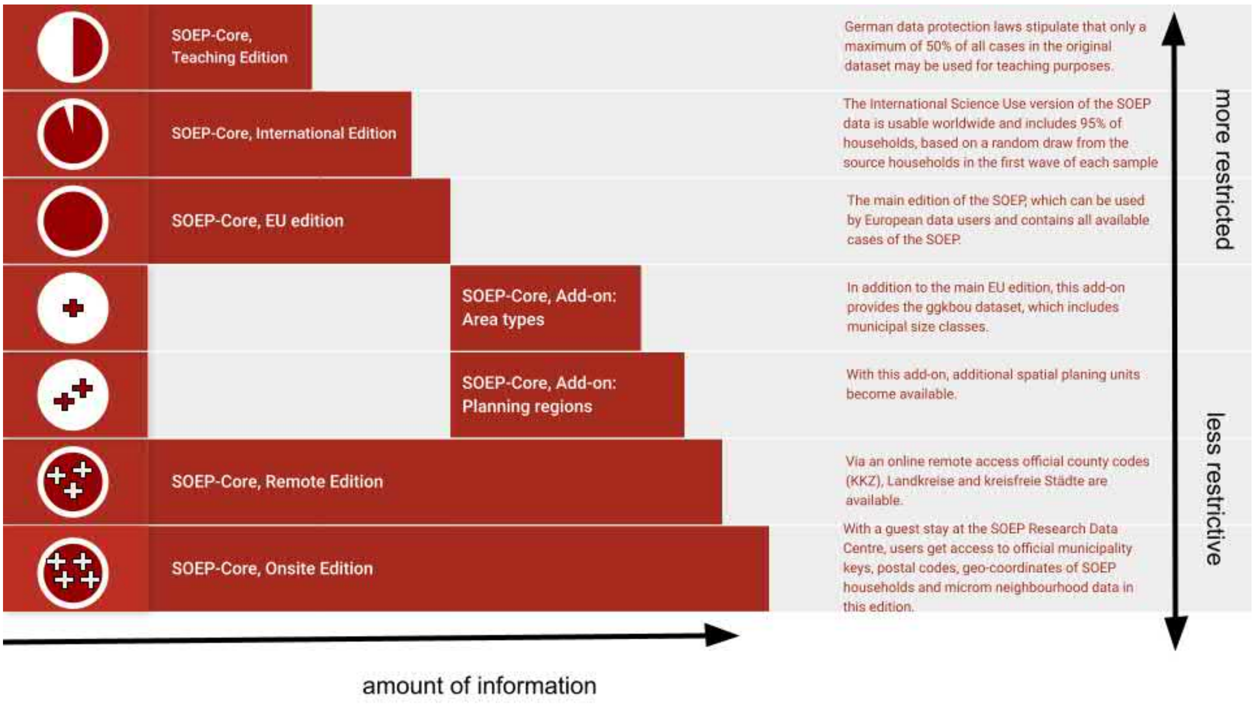
The number of observations is now much smaller. The effect of pgbilzeit is greater than before. Pgbilzeit has a lower effect in the wlong==0 group, where the return is different for each additional educational year. People in the wlong===0 group may not get the returns on additional education they expected on the local labor market and may therefore move -> higher dropout probability.

Last change: Sep 26, 2022

In order to gain the best possible insight into how to work with the various regional data on the SOEP, we recommend the following exercises:

6.9 Working with SOEP Regional Data

SOEP offers diverse possibilities for regional and spatial analysis. With the anonymized regional information on SOEP respondents' (households' and individuals') place of residence, it is possible to link numerous regional indicators on the levels of the federal states (Bundesländer), spatial planning regions, districts, and postal codes with the data on the SOEP households. However, specific security provisions must be made due to the sensitivity of the data under data protection law. Accordingly, data users are not allowed to give any information in their analyses that could indicate, for instance, the city or district in which respondents reside. The data nevertheless provide valuable background information for regional analysis.



SOEP-Core, Teaching Edition — German data protection laws stipulate that only a maximum of 50% of all cases in the original dataset may be used for teaching purposes.

SOEP-Core, International Edition — The International Science Use version of the SOEP data is usable worldwide and includes 95% of households, based on a random draw from the source households in the first wave of each sample

SOEP-Core, EU edition — The main edition of the SOEP, which can be used by European data users and contains all available cases of the SOEP.

SOEP-Core, Add-on: Area types — In addition to the main EU edition, this add-on provides the ggkbou dataset, which includes municipal size classes.

SOEP-Core, Add-on: Planning regions — With this add-on, additional spatial planing units become available.

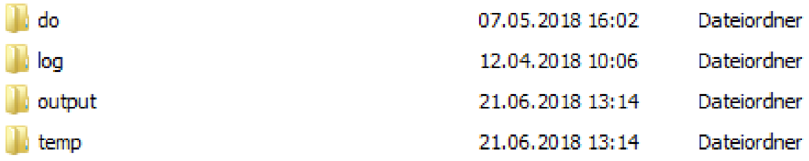
SOEP-Core, Remote Edition — Via an online remote access official county codes (KKZ), Landkreise and kreisfreie Städte are available.

SOEP-Core, Onsite Edition — With a guest stay at the SOEP Research Data Centre, users get access to official municipality keys, postal codes, geo-coordinates of SOEP households and microm neighbourhood data in this edition.

For more information and to access the data, see **Regional Data**

Assume that for your research project, you want to measure current (2016) urban-rural differences in the population. You are particularly interested in the differences in interest in politics and the different satisfaction variables provided by the SOEP. You also want to take into account demographic differences in gender and age. To be able to evaluate the potential of the data for your project, you first need an overview. For regional analysis, for example, the municipal size classes from the regional data are suitable.

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do

- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
************************************************
* Set relative paths to the working directory
************************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
global region "\\hume\soep-region\DATA\soep33_de\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global “AVZ” defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to the data you ordered.

a) Prepare a dataset for cross-sectional analysis covering the survey year 2016 (wave bg).

To perform your analysis, you need different SOEP variables. The SOEP offers various options for a variable search:

- Search the questionnaires for useful variables (for more information, see the section *Variable Search with Questionnaires*)
- Find a suitable variable in the topic list on paneldata.org (for more information, see the section *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information, see the section *Variable Search with paneldata.org*)
- Use the documentation provided by the generated variables (for more information, see the section *Documentation on Generated Data*)

Your source file should contain the following variables:

- Permanent Individual ID **"persnr"**
- Original Household Number **"hhnr"**
- Current Wave Household Number **"bghhnr"**
- The Sex of the Person **"sex"**
- Year of Birth **"gebjahr"**
- Survey Status 2016 **"bgnetto"**
- Sample Membership 2016 **"bgpop"**
- Weighting Factor 2016 **"bgphrf"**
- Satisfaction With Health **"bgp0101"**
- Satisfaction With Sleep **"bgp0102"**
- Satisfaction With Work **"bgp0103"**
- Satisfaction With Housework **"bgp0104"**

- Satisfaction With Household Income **"bgp0105"**
- Satisfaction With Personal Income **"bgp0106"**
- Satisfaction With Dwelling **"bgp0107"**
- Satisfaction With Amount Of Leisure Time **"bgp0108"**
- Satisfaction With Child Care **"bgp0109"**
- Satisfaction With Family Life **"bgp0110"**
- Satisfaction With Social Life **"bgp0111"**
- Satisfaction with Democracy **"bgp0112"**
- Political Interest **"bgp143"**
- Current Sample Region **"bgsampreg"**
- Federal State **"bgbula"**
- Spatial Category by BBSR **"bgregtyp"**
- Municipal Class Sizes “ggk”

Use the key variables from the ppath.dta dataset as your starting file.

```
use hhnr persnr bghhnr sex gebjahr bgnetto bgpop using ${MY_IN_PATH}\ppfad.dta, clear
```

Attention: Please note that since version 34 (v34), PPFAD can be found in the subdirectory “Raw” of the data distribution file. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFAD.

Keep people who completed a questionnaire in 2016 and lived in a private household.

```
* Keep people who completed a questionnaire in 2016 and live in a private household
keep if bghhnr>0 & inrange(bgnetto, 10, 29) & inlist(bgpop, 1, 2)
keep hhnr persnr bghhnr sex gebjahr bgnetto bgpop
merge 1:1  persnr using ${MY_IN_PATH}\phrf.dta, keep(match master) keepusing (bgphrf) nogenerate
tempfile ppfad
save `ppfad'
```

Prepare the different datasets bgp, bghbrutto, regionl

```
* Prepare dataset bgp
use ${MY_IN_PATH}\bgp.dta, replace
keep persnr hhnr bghhnr bgp01* bgp143
tempfile bgp
save `bgp'

* Prepare dataset bghbrutto
use ${MY_IN_PATH}\bghbrutto.dta, replace
keep hhnr bghhnr bgsampreg bgbula bgregtyp
tempfile bghbrutto
save `bghbrutto'

* Prepare dataset regionl
```

(continues on next page)

(continued from previous page)

```
use ${region}\regionl_v33.dta, replace
keep if syear==2016
keep syear hhnr hhnrakt ggk
rename hhnrakt bghhnr
tempfile regionl
save `regionl'
```

Merge all datasets.

```
* Merge all datasets
use `ppfad'
merge 1:1 persnr using `bgp', keep(match master) nogenerate
merge m:1 bghhnr hhnr using `regionl', keep(match master) nogenerate
merge m:1 bghhnr hhnr using `bghbrutto', keep(match master) nogenerate
```

Recode negative values as missings.

```
* Recode negative values into missings
mvdecode sex gebjahr bgp01* bgp143,mv(-5/-1)
```

Categorize the municipal class sizes from the SOEP regional dataset.

```
* Categorize community class size
gen ggk_cat=.
replace ggk_cat=-1 if ggk==-1
replace ggk_cat=1 if ggk==1 | ggk==2
replace ggk_cat=2 if ggk==3
replace ggk_cat=3 if ggk==4 | ggk==5
replace ggk_cat=4 if ggk>5 & ggk<=7

lab var ggk_cat "Community Size categorised"
lab def ggk_cat -1 "No information" 1 "<=5000" 2 "5001 - 20000" 3 "20001 - 100000" ///
4 ">100000"
lab val ggk_cat ggk_cat
```

Generate an age variable.

```
* Generate age variable
gen alter= 2016-gebjahr if gebjahr > 0
gen alter_cat=1 if alter<=20
replace alter_cat=2 if alter>20 & alter<=30
replace alter_cat=3 if alter>30 & alter<=65
replace alter_cat=4 if alter>65 & alter<=120

lab var alter "age"
lab var alter_cat "age categorized"
lab def alter_cat 1 "<=20" 2 "21-30" 3 "31-65" 4 ">65"
lab val alter_cat alter_cat
```

Categorize a federal states variable.

```
* Categorize federal states
gen bgbula_cat=.
```

(continues on next page)

(continued from previous page)

```
* Schleswig-Holstein + Hamburg
replace bgbula_cat=1 if bgbula==1 | bgbula==2
* Lower Saxony + Bremen
replace bgbula_cat=2 if bgbula==3 | bgbula==4
* Mecklenburg Western Pomerania + Brandenburg
replace bgbula_cat=3 if bgbula==13 | bgbula==12
* Saarland + Rhineland Palatinate
replace bgbula_cat=4 if bgbula==7 | bgbula==10
* Northrhine-Westphalia
replace bgbula_cat=5 if bgbula==5
* Hesse
replace bgbula_cat=6 if bgbula==6
* Baden-Württemberg
replace bgbula_cat=7 if bgbula==8
* Bavaria
replace bgbula_cat=8 if bgbula==9
* Berlin
replace bgbula_cat=9 if bgbula==11
* Saxony
replace bgbula_cat=10 if bgbula==14
* Saxony-Anhalt
replace bgbula_cat=11 if bgbula==15
* Thuringia
replace bgbula_cat=12 if bgbula==16

lab var bgbula_cat "Federal states categorized"
lab def bgbula_cat 1 "Schleswig-Holstein/Hamburg" 2 "Lower Saxony/Bremen" 3 "Mecklenburg Western Pomerania/Brandenburg" ///
4 "Saarland/Rhineland Palatinate" 5 "Northrhine-Westphalia" 6 "Hesse" ///
7 "Baden-Wuerttenberg" 8 "Bavaria" 9 "Berlin" 10 "Saxony" 11 "Saxony-Anhalt" 12 "Thuringia"
lab val bgbula_cat bgbula_cat
drop bgbula
rename bgbula_cat bgbula
```

Put the variables in your preferred order and save your dataset.

```
* Order demography and identifiers first
order persnr hhnr bghhnr syear sex gebjahr alter alter_cat bgsampreg bgbula ggk ///
ggk_cat bgregtyp

save ${MY_OUT_DATA}\zeit_online.dta, replace
```

b) You want to get an initial overview of regional differences in satisfaction with various aspects of life. Use the variable bgsampreg and cross-stabilize the variable with all satisfaction variables to identify differences between East and West Germany, display the absolute and relative frequencies.

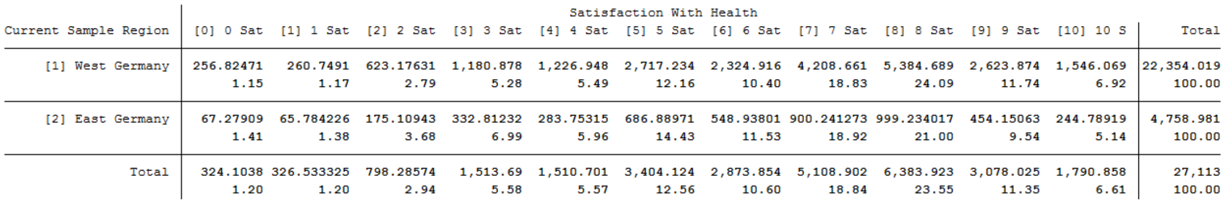
To save the tables, save them in a log file.

```
********************************************************************************
capture log close
log using "${MY_LOG_OUT}\satisfaction.log", replace

```

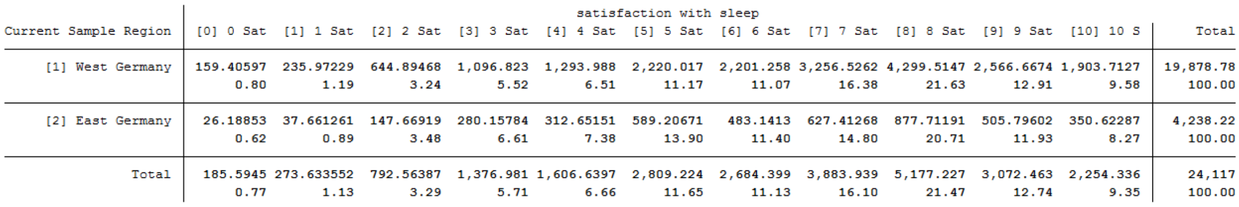
(continues on next page)

(continued from previous page)

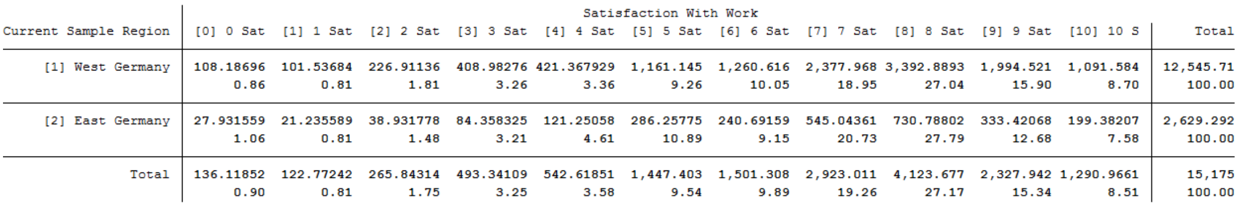
```
* Life satisfaction

local varlist bgp0101 bgp0102 bgp0103 bgp0104 bgp0105 bgp0106 bgp0107 bgp0108 ///
bgp0109 bgp0110 bgp0111 bgp0112
foreach x of local varlist {
tab bgsampreg `x' [aw= bgphrf] , row
}
```



```
                                            Satisfaction With Health
Current Sample Region   [0] 0 Sat  [1] 1 Sat  [2] 2 Sat  [3] 3 Sat  [4] 4 Sat  [5] 5 Sat  [6] 6 Sat  [7] 7 Sat  [8] 8 Sat  [9] 9 Sat  [10] 10 S      Total
-----------------------------------------------------------------------------------------------------------------------------------------------------------
    [1] West Germany    256.82471   260.7491  623.17631  1,180.878  1,226.948  2,717.234  2,324.916  4,208.661  5,384.689  2,623.874  1,546.069 22,354.019
                             1.15       1.17       2.79       5.28       5.49      12.16      10.40      18.83      24.09      11.74       6.92     100.00
-----------------------------------------------------------------------------------------------------------------------------------------------------------
    [2] East Germany     67.27909  65.784226  175.10943  332.81232  283.75315  686.88971  548.93801 900.241273 999.234017  454.15063  244.78919  4,758.981
                             1.41       1.38       3.68       6.99       5.96      14.43      11.53      18.92      21.00       9.54       5.14     100.00
-----------------------------------------------------------------------------------------------------------------------------------------------------------
               Total     324.1038 326.533325  798.28574   1,513.69  1,510.701  3,404.124  2,873.854  5,108.902  6,383.923  3,078.025  1,790.858     27,113
                             1.20       1.20       2.94       5.58       5.57      12.56      10.60      18.84      23.55      11.35       6.61     100.00
```



```
                                            satisfaction with sleep
Current Sample Region   [0] 0 Sat  [1] 1 Sat  [2] 2 Sat  [3] 3 Sat  [4] 4 Sat  [5] 5 Sat  [6] 6 Sat  [7] 7 Sat  [8] 8 Sat  [9] 9 Sat  [10] 10 S      Total
-----------------------------------------------------------------------------------------------------------------------------------------------------------
    [1] West Germany    159.40597  235.97229  644.89468  1,096.823  1,293.988  2,220.017  2,201.258 3,256.5262 4,299.5147 2,566.6674 1,903.7127  19,878.78
                             0.80       1.19       3.24       5.52       6.51      11.17      11.07      16.38      21.63      12.91       9.58     100.00
-----------------------------------------------------------------------------------------------------------------------------------------------------------
    [2] East Germany     26.18853  37.661261  147.66919  280.15784  312.65151  589.20671   483.1413  627.41268  877.71191  505.79602  350.62287   4,238.22
                             0.62       0.89       3.48       6.61       7.38      13.90      11.40      14.80      20.71      11.93       8.27     100.00
-----------------------------------------------------------------------------------------------------------------------------------------------------------
               Total     185.5945 273.633552  792.56387  1,376.981 1,606.6397  2,809.224  2,684.399  3,883.939  5,177.227  3,072.463  2,254.336     24,117
                             0.77       1.13       3.29       5.71       6.66      11.65      11.13      16.10      21.47      12.74       9.35     100.00
```

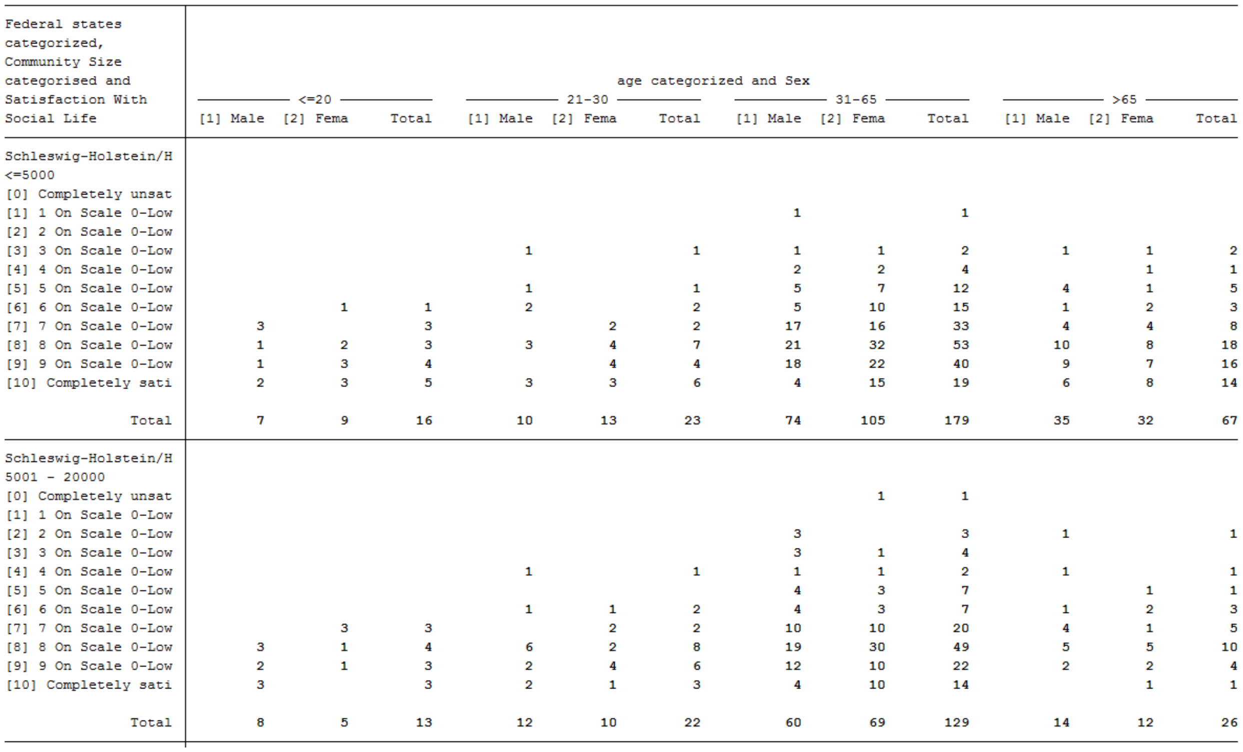


```
                                            Satisfaction With Work
Current Sample Region   [0] 0 Sat  [1] 1 Sat  [2] 2 Sat  [3] 3 Sat  [4] 4 Sat  [5] 5 Sat  [6] 6 Sat  [7] 7 Sat  [8] 8 Sat  [9] 9 Sat  [10] 10 S      Total
-----------------------------------------------------------------------------------------------------------------------------------------------------------
    [1] West Germany    108.18696  101.53684  226.91136  408.98276 421.367929  1,161.145  1,260.616  2,377.968 3,392.8893  1,994.521  1,091.584  12,545.71
                             0.86       0.81       1.81       3.26       3.36       9.26      10.05      18.95      27.04      15.90       8.70     100.00
-----------------------------------------------------------------------------------------------------------------------------------------------------------
    [2] East Germany    27.931559  21.235589  38.931778  84.358325  121.25058  286.25775  240.69159  545.04361  730.78802  333.42068  199.38207  2,629.292
                             1.06       0.81       1.48       3.21       4.61      10.89       9.15      20.73      27.79      12.68       7.58     100.00
-----------------------------------------------------------------------------------------------------------------------------------------------------------
               Total    136.11852  122.77242  265.84314  493.34109  542.61851  1,447.403  1,501.308  2,923.011  4,123.677  2,327.942 1,290.9661     15,175
                             0.90       0.81       1.75       3.25       3.58       9.54       9.89      19.26      27.17      15.34       8.51     100.00
```

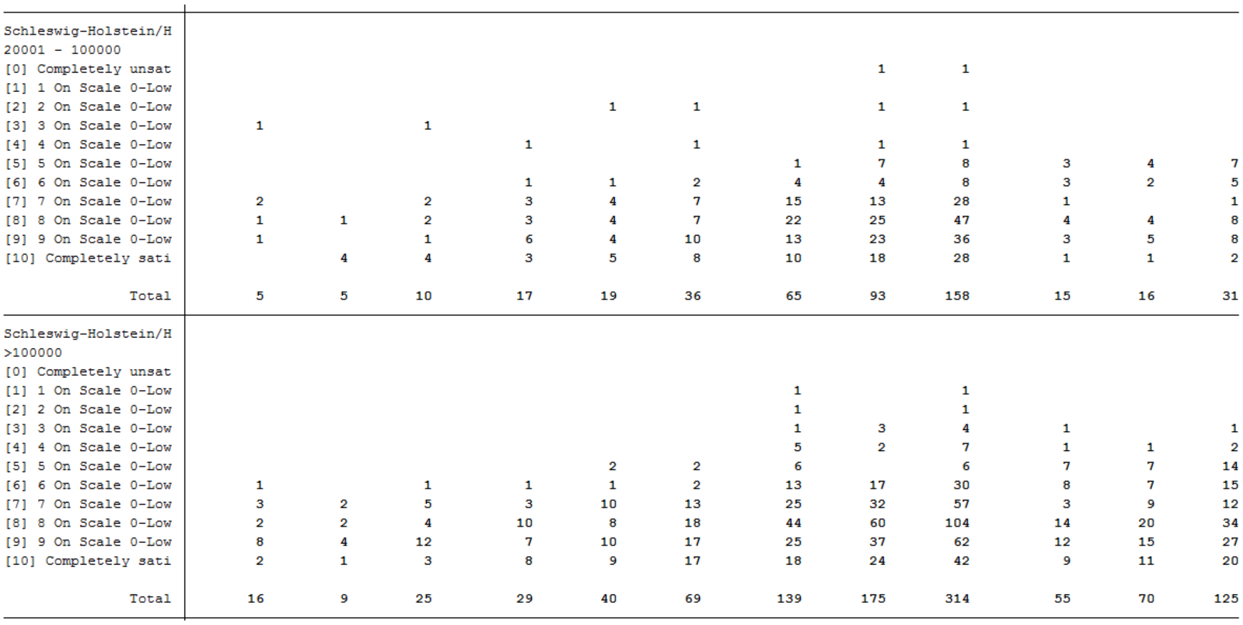
To view all tables, look at your generated log file.

c) Now take a closer look at satisfaction with various aspects of life with the help of SOEP regional data. Use the municipal size classes. Create a table showing satisfaction with different aspects of life and highlighting differences by sex, age, municipal size class, and federal state.

```
foreach x of local varlist {
* Tabulation of satisfaction by municipal size class and federal state
table `x' sex alter_cat, by(bgbula ggk_cat) contents(freq) column row stubwidth(20) cellwidth(8) csepwidth(2) nomissing
* Tabulation of satisfaction by municipal size class
table `x' sex alter_cat, by(ggk_cat) contents(freq) column row stubwidth(20) cellwidth(8) csepwidth(2) nomissing
* Tabulation of satisfaction by federal state
table `x' sex alter_cat, by(bgbula) contents(freq) column row stubwidth(20) cellwidth(8) csepwidth(2) nomissing
}
```



```
Federal states       |
categorized,         |
Community Size       |
categorised and      |                                                  age categorized and Sex
Satisfaction With    |  ----------- <=20 -----------    ----------- 21-30 ----------    ---------- 31-65 -----------    ------------ >65 -----------
Social Life          |  [1] Male  [2] Fema     Total    [1] Male  [2] Fema     Total    [1] Male  [2] Fema     Total    [1] Male  [2] Fema     Total
---------------------+--------------------------------------------------------------------------------------------------------------------------------
Schleswig-Holstein/H |
<=5000               |
[0] Completely unsat |
[1] 1 On Scale 0-Low |                                                                       1                   1
[2] 2 On Scale 0-Low |
[3] 3 On Scale 0-Low |                                       1                   1           1         1         2           1         1         2
[4] 4 On Scale 0-Low |                                                                       2         2         4                     1         1
[5] 5 On Scale 0-Low |                                       1                   1           5         7        12           4         1         5
[6] 6 On Scale 0-Low |                 1         1           2                   2           5        10        15           1         2         3
[7] 7 On Scale 0-Low |       3                   3                     2         2          17        16        33           4         4         8
[8] 8 On Scale 0-Low |       1         2         3           3         4         7          21        32        53          10         8        18
[9] 9 On Scale 0-Low |       1         3         4                     4         4          18        22        40           9         7        16
[10] Completely sati |       2         3         5           3         3         6           4        15        19           6         8        14
                     |
               Total |       7         9        16          10        13        23          74       105       179          35        32        67
---------------------+--------------------------------------------------------------------------------------------------------------------------------
Schleswig-Holstein/H |
5001 - 20000         |
[0] Completely unsat |                                                                                 1         1
[1] 1 On Scale 0-Low |
[2] 2 On Scale 0-Low |                                                                       3                   3           1                   1
[3] 3 On Scale 0-Low |                                                                       3         1         4
[4] 4 On Scale 0-Low |                                       1                   1           1         1         2           1                   1
[5] 5 On Scale 0-Low |                                                                       4         3         7                     1         1
[6] 6 On Scale 0-Low |                                       1         1         2           4         3         7           1         2         3
[7] 7 On Scale 0-Low |                 3         3                     2         2          10        10        20           4         1         5
[8] 8 On Scale 0-Low |       3         1         4           6         2         8          19        30        49           5         5        10
[9] 9 On Scale 0-Low |       2         1         3           2         4         6          12        10        22           2         2         4
[10] Completely sati |       3                   3           2         1         3           4        10        14                     1         1
                     |
               Total |       8         5        13          12        10        22          60        69       129          14        12        26
---------------------+--------------------------------------------------------------------------------------------------------------------------------
```



```
---------------------+--------------------------------------------------------------------------------------------------------------------------------
Schleswig-Holstein/H |
20001 - 100000       |
[0] Completely unsat |                                                                                 1         1
[1] 1 On Scale 0-Low |
[2] 2 On Scale 0-Low |                                                 1         1                     1         1
[3] 3 On Scale 0-Low |       1                   1
[4] 4 On Scale 0-Low |                                       1                   1                     1         1
[5] 5 On Scale 0-Low |                                                                       1         7         8           3         4         7
[6] 6 On Scale 0-Low |                                       1         1         2           4         4         8           3         2         5
[7] 7 On Scale 0-Low |       2                   2           3         4         7          15        13        28           1                   1
[8] 8 On Scale 0-Low |       1         1         2           3         4         7          22        25        47           4         4         8
[9] 9 On Scale 0-Low |       1                   1           6         4        10          13        23        36           3         5         8
[10] Completely sati |                 4         4           3         5         8          10        18        28           1         1         2
                     |
               Total |       5         5        10          17        19        36          65        93       158          15        16        31
---------------------+--------------------------------------------------------------------------------------------------------------------------------
Schleswig-Holstein/H |
>100000              |
[0] Completely unsat |
[1] 1 On Scale 0-Low |                                                                       1                   1
[2] 2 On Scale 0-Low |                                                                       1                   1
[3] 3 On Scale 0-Low |                                                                       1         3         4           1                   1
[4] 4 On Scale 0-Low |                                                                       5         2         7           1         1         2
[5] 5 On Scale 0-Low |                                                 2         2           6                   6           7         7        14
[6] 6 On Scale 0-Low |       1                   1           1         1         2          13        17        30           8         7        15
[7] 7 On Scale 0-Low |       3         2         5           3        10        13          25        32        57           3         9        12
[8] 8 On Scale 0-Low |       2         2         4          10         8        18          44        60       104          14        20        34
[9] 9 On Scale 0-Low |       8         4        12           7        10        17          25        37        62          12        15        27
[10] Completely sati |       2         1         3           8         9        17          18        24        42           9        11        20
                     |
               Total |      16         9        25          29        40        69         139       175       314          55        70       125
---------------------+--------------------------------------------------------------------------------------------------------------------------------
```

To view all tables, look at your generated log file. As you can see, SOEP regional data can be used to analyze variables at the lowest regional levels.

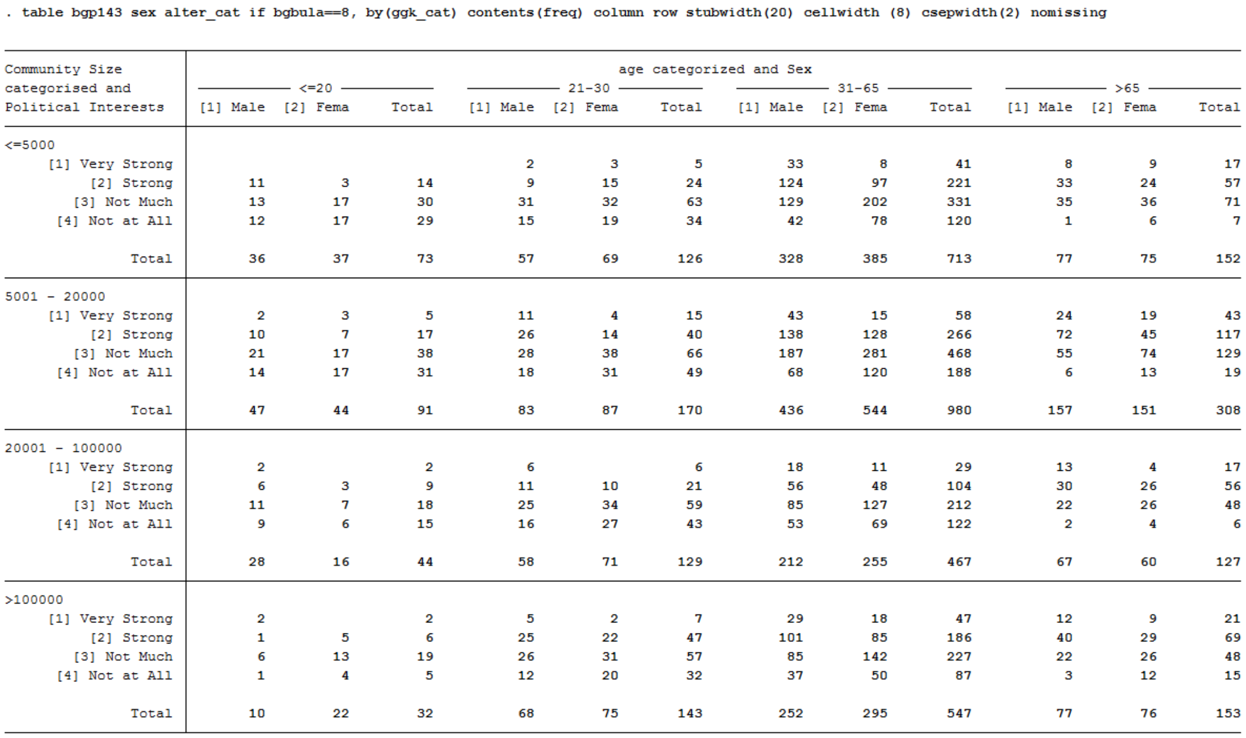
d) Create a table that shows political interest differentiated by age, sex, and municipal size class in Bavaria

```
********************************************************************************
capture log close
```

(continues on next page)

(continued from previous page)

```
log using "${MY_LOG_OUT}\political_interest.log", replace

* Political interest
* Tabulation of political interest by municipal size class for Bavaria
table bgp143 sex alter_cat if bgbula==8, by(ggk_cat) contents(freq) column row␣
↪stubwidth(20) cellwidth (8) csepwidth(2) nomissing
```



```
. table bgp143 sex alter_cat if bgbula==8, by(ggk_cat) contents(freq) column row stubwidth(20) cellwidth (8) csepwidth(2) nomissing

---------------------------------------------------------------------------------------------------------------------------------------------
Community Size       |                                                  age categorized and Sex
categorised and      |  ----------- <=20 -----------  ----------- 21-30 ----------  ----------- 31-65 ----------  ------------ >65 -----------
Political Interests  |  [1] Male  [2] Fema     Total  [1] Male  [2] Fema     Total  [1] Male  [2] Fema     Total  [1] Male  [2] Fema     Total
---------------------+-----------------------------------------------------------------------------------------------------------------------
<=5000               |
     [1] Very Strong |                                       2         3         5        33         8        41         8         9        17
          [2] Strong |        11         3        14         9        15        24       124        97       221        33        24        57
        [3] Not Much |        13        17        30        31        32        63       129       202       331        35        36        71
      [4] Not at All |        12        17        29        15        19        34        42        78       120         1         6         7
                     |
               Total |        36        37        73        57        69       126       328       385       713        77        75       152
---------------------+-----------------------------------------------------------------------------------------------------------------------
5001 - 20000         |
     [1] Very Strong |         2         3         5        11         4        15        43        15        58        24        19        43
          [2] Strong |        10         7        17        26        14        40       138       128       266        72        45       117
        [3] Not Much |        21        17        38        28        38        66       187       281       468        55        74       129
      [4] Not at All |        14        17        31        18        31        49        68       120       188         6        13        19
                     |
               Total |        47        44        91        83        87       170       436       544       980       157       151       308
---------------------+-----------------------------------------------------------------------------------------------------------------------
20001 - 100000       |
     [1] Very Strong |         2                   2         6                   6        18        11        29        13         4        17
          [2] Strong |         6         3         9        11        10        21        56        48       104        30        26        56
        [3] Not Much |        11         7        18        25        34        59        85       127       212        22        26        48
      [4] Not at All |         9         6        15        16        27        43        53        69       122         2         4         6
                     |
               Total |        28        16        44        58        71       129       212       255       467        67        60       127
---------------------+-----------------------------------------------------------------------------------------------------------------------
>100000              |
     [1] Very Strong |         2                   2         5         2         7        29        18        47        12         9        21
          [2] Strong |         1         5         6        25        22        47       101        85       186        40        29        69
        [3] Not Much |         6        13        19        26        31        57        85       142       227        22        26        48
      [4] Not at All |         1         4         5        12        20        32        37        50        87         3        12        15
                     |
               Total |        10        22        32        68        75       143       252       295       547        77        76       153
---------------------------------------------------------------------------------------------------------------------------------------------
```

As you have seen here, the SOEP offers a wide range of possibilities for regional analysis. It is possible to allocate a multitude of regional indicators at the level of federal states, regional planning regions, districts, and postal codes.

Last change: Sep 26, 2022

6.10 Working with spatial data in R

6.10.1 Prerequisites

Data Access

To work with SOEP's spatial data you need to have a data distribution contract together with an application for stationary access to a terminal in the Research Data Center (RDC). The forms can be downloaded from here. Having data access granted, you can use the IGEL workstations in the SOEP RDC to get access to the corresponding server infrastructures MORAN and HAUSER. The MORAN server hosts the coordinates of the SOEP households together with some fake coordinates that cannot be distinguished. On this server you can, for example, compute distances or nearest points as shown later in this tutorial. On the HAUSER server you are provided your computational outcome form the MORAN server but without the coordinates. The coordinates of the SOEP households **must never** be available together with SOEP survey data. More information on how to use the SOEP IGEL technology can be found here.

R

When working with SOEP's spatial data at the RDC using SOEP IGEL we provide you with the software you need. If you want to work with spatial test data before, you need a certain software setup. To be able to work yourself through the following examples you need to have R and RStudio installed. Further, for working the spatial data in R, you need to have GEOS, GDAL, and PROJ installed, too. When running a Windows OS, simply install the Rtools for Windows from here. MacOS and Linux users are referred to the website of the `sf` package for installation instructions. You can find them here.

If you are not yet familiar with R we recommend the book *R for Data Science* by Wickham and Grolemund which is available here for free. Besides that, there two freely available books which are helpful to get you started with spatial data in R:

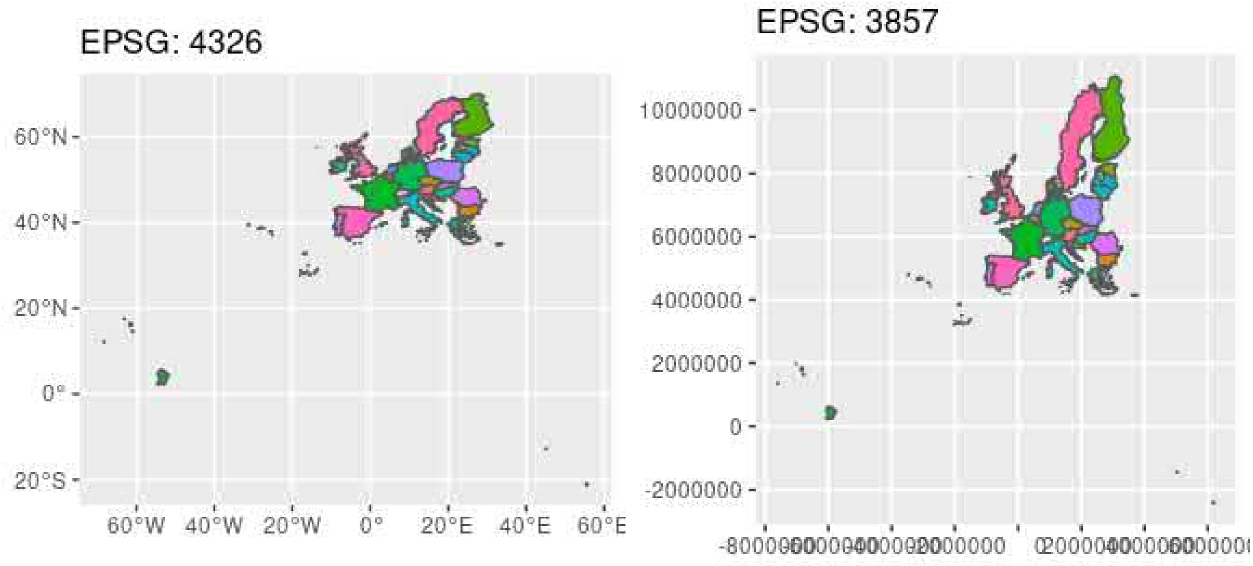
- *Spatial Data Science* by Pebesma and Bivand which is available here
- *Geocomputation with R* by Lovelace, Nowosad, and Muenchow which is available here

Coordinate Reference Systems

Any location of an object can be characterised by its coordinates. Besides a coordinate the altitude of an object might be of interest. So locations can be given in 2-dimensional or 3-dimensional space. A 2-dimensional coordinate referenced to two orthogonal axes (say x and y) can be given by either a pair of values (x,y) or by an angle (with respect to one of the axes) and a distance. The pair of coordinates (x,y) is referred to as cartesian coordinates and the latter as polar or spherical coordinates. In a 3-dimensional world, where you define a location on a globe, things become a little more complicated, because earth is an ellipsoid rather than a perfect sphere. In defining a location in a 3-dimensional coordinate system, the *coordinate reference system* (CRS) is very important. The CRS combines the information about the coordinate system itself and a geodetic datum. The geodetic datum describes how the location of the coordinate system is related to earth. So different CRS will provide different values of coordinates for the same location. Finally, the projection method provides information on how to flatten objects that are on a round surface so they can be viewed on a flat surface. The CRS is often summarised by an EPSG-code (European Petroleum Survey Group Geodesy) allowing for convenient transformations between different CRS. For more details we refer you to Chapter 2 of *Spatial Data Science* or Chapter 6 of *Geocomputation with R*. The most frequently used CRS are displayed in the following table.

EPSG code	Note	PRJ4-String	Projection
3857	WGS 84 / Pseudo-Mercator	+proj=merc +a=6378137 +b=6378137 +lat_ts=0 +lon_0=0 +x_0=0 +y_0=0 +k=1 +units=m +nadgrids=@null +wktext +no_defs +type=crs	Popular Visualisation Pseudo Mercator
4326	WGS 84	+proj=longlat +datum=WGS84 +no_defs +type=crs	(null)
25832	ETRS89 / UTM zone 32N	+proj=utm +zone=32 +ellps=GRS80 +units=m +no_defs +type=crs	Transverse Mercator
25833	ETRS89 / UTM zone 33N	+proj=utm +zone=33 +ellps=GRS80 +units=m +no_defs +type=crs	Transverse Mercator
32632	WGS 84 / UTM zone 32N	+proj=utm +zone=32 +datum=WGS84 +units=m +no_defs +type=crs	Transverse Mercator
32633	WGS 84 / UTM zone 33N	+proj=utm +zone=33 +datum=WGS84 +units=m +no_defs +type=crs	Transverse Mercator

The differences in the maps output can be seen in the following figure. The maps show Europe (provided by https://data.opendatasoft.com) in two different CRS.



The CRS of the fake SOEP households is defined by the EPSG code 32632. The CRS usually used by Google Maps / OpenStreetMap is defined by EPSG code 4326. To check you can use for example https://coordinates-converter.com/.

If you already know about all of the prerequisites and are familiar with R, the sf package as well as the tidyverse, you can directly switch to the complete example provided in section *Complete Example*. If you are not familiar with all of this, we recommend you go through the following text first.

6.10.2 Reading data

Spatial data comes in a variety of formats, ranging from ASCII-files to more sophisticated data structures like shapefiles or XML-dialects. Fortunately the R package sf provides functions to conveniently read most of the data formats. Besides that, the package and its use is well documented here. The package can easily be installed and loaded by using the following lines of code in R.

```
install.packages('sf') # installing the package
library(sf) # loading the package

# further packages
install.packages('tidyverse') # a selection of packages
library(tidyverse) # (mainly using dplyr, ggplot2)
install.packages('here') # for working with .RProj and paths
library(here)
```

Further packages that are useful when working with R are the tidyverse (a selection of several packages) and here (makes working in projects easier). A complete list of packages used to create this page is provided in Appendix at the end of this page.

When working with spatial data in the SOEP we provide you with fake coordinates of the SOEP households in a shapefile format. This format is easy to read using the sf package in R and provides you with all the details you need. The data you bring along is likely to be in a different format. Formats that can be read using `st_read` are listed in the GDAL Documentation. If you bring along data in other formats consider the haven package for SAS, SPSS, or Stata formats, documented here. Non-spatial data formats however, mostly include no information on the CRS. Thus you need to know and assign it to the data. An example when reading data from spreadsheets is provided below.

Shapefiles

Shapefiles consist of at least three files, namely

- `.shp` containing the geometries for example for a point (address) or a polygon (state)
- `.dbf` containing the attribute data for each geometry, for exampe, the address data (street, zip code, city) or names of the states
- `.shx` containing the indices linking geometries and attribute data

There are usually some more files provided containing further information. One example of shapefiles is provided by the Federal Agency for Cartography and Geodesy (BKG, GeoBasis-DE / BKG 2019). The BKG provides public available data on administrative areas in Germany. The dataset `VG250` includes theses administrative units of the hierarchical levels from the country down to municipalities and is available here. Unzipping the folder will provide you with the shapefiles and the documentation of the data. Because the data contains information on different levels (Germany, its states, district, municipalities) these are stored in different layers. To check which information is avaialable in the data, use `st_layers`. If you do not explicitly provide a layer, `st_read` will take the first available layer. The function returns the layer names, the according type of geometry, the number of features (geometries) and fields (variables). Reading the German Federal States you need to pick the layer `VG250_LAN`.

```
# layers in the data
st_layers(here('Daten', 'vg250_ebenen_0101'))
```

```
## Driver: ESRI Shapefile
## Available layers:
##    layer_name geometry_type features fields
## 1   VG250_GEM       Polygon    11135     26
## 2   VG250_KRS       Polygon      431     26
## 3   VG250_LAN       Polygon       35     26
## 4    VG250_LI   Line String    35483      4
## 5    VG250_PK         Point    11003     13
## 6   VG250_RBZ       Polygon       20     26
## 7   VG250_STA       Polygon       11     26
## 8   VG250_VWG       Polygon     4688     26
## 9    VG_DATEN            NA       35      8
## 10   VG_WERTE            NA       32      4
```

```
# read the shapefile for Federal States
States <- st_read(here('Daten', 'vg250_ebenen_0101'), layer = 'VG250_LAN', quiet = TRUE)
```

Taking a look at a the main variables `GEN` (Geographical name), `BEZ` (Designation), and `geometry` (a multipolygon), we are provided with information displayed differently from usual `data.frame` or `tibble` like data. The 'header' tells you it's a collection of simple features having 16 features (rows) and 2 fields (variables, excluding the geometry column). The type of geomtry is a multipolygon (set of polygons) in a two-dimensional space (XY). The bounding box (`bbox`) provides information on the bottom left and top right corner of the box bounding the geometries. Finally, you get information on the coordinate reference system (CRS), summarized by EPSG-Code 25832.

```
# look at the data
States %>%
  filter(GF == 4) %>% # use land with structure only
  select(GEN, BEZ, geometry) # select necessary variables
```

```
## Simple feature collection with 16 features and 2 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 280371.1 ymin: 5235856 xmax: 921292.4 ymax: 6101444
## projected CRS:  ETRS89 / UTM zone 32N
## First 10 features:
##                   GEN                     BEZ                       geometry
## 1   Schleswig-Holstein                    Land MULTIPOLYGON (((464810.7 61...
## 2              Hamburg Freie und Hansestadt MULTIPOLYGON (((578219 5954...
## 3        Niedersachsen                    Land MULTIPOLYGON (((479451.1 59...
## 4               Bremen     Freie Hansestadt MULTIPOLYGON (((466930.3 58...
## 5  Nordrhein-Westfalen                    Land MULTIPOLYGON (((477607.3 58...
## 6               Hessen                    Land MULTIPOLYGON (((534242 5721...
## 7      Rheinland-Pfalz                    Land MULTIPOLYGON (((416304.5 56...
## 8    Baden-Württemberg                    Land MULTIPOLYGON (((546771.2 55...
## 9               Bayern            Freistaat MULTIPOLYGON (((609387.6 52...
## 10            Saarland                    Land MULTIPOLYGON (((359841 5499...
```

SOEP Regional Data

Also the fake coordinates for the SOEP households are provided in a shapefile format. The data contains an ID-variable (ID) and the survey year (erhebj) together with a point geometry. The ID provided is not related to any of the SOEP's hid or pid. It is simply a number ranging from 1 to the corresponding number of rows in the data set. Households can thus only be identified, merged, or selected using their coordinate. Moreover the data also contain the SOEP's initial sample. Thus containing regional information on respondents and non-respondents. Because of data protection regulations, the SOEP survey data must not be used together with the household's coordinates. This is why the environment where you are allowed to work with the household's coordinates is strictly separated. Unlike the VG250 data the SOEP data has the EPSG-code 32632.

```
SOEP <- st_read(here('Daten', 'soep_v29'))
```

```
## Reading layer `soep_hh_korr_nur_fakes_utm32-v29' from data source `/media/H/Projekte/
↪Adhoc_Analysen/GeodatenBSP/Daten/soep_v29' using driver `ESRI Shapefile'
## Simple feature collection with 209734 features and 2 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: 288107 ymin: 5250939 xmax: 888152 ymax: 6055336
## projected CRS:  WGS 84 / UTM zone 32N
```

```
SOEP
```

```
## Simple feature collection with 209734 features and 2 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: 288107 ymin: 5250939 xmax: 888152 ymax: 6055336
## projected CRS:  WGS 84 / UTM zone 32N
## First 10 features:
```

(continues on next page)

(continued from previous page)

```
##    erhebj ID               geometry
## 1    2000  1 POINT (355929 5665928)
## 2    2000  2 POINT (800240 5821852)
## 3    2000  3 POINT (794464 5831012)
## 4    2000  4 POINT (478831 5431318)
## 5    2000  5 POINT (799318 5825023)
## 6    2000  6 POINT (794752 5833726)
## 7    2000  7 POINT (799149 5824960)
## 8    2000  8 POINT (791594 5827687)
## 9    2000  9 POINT (791657 5827422)
## 10   2000 10 POINT (792280 5826345)
```

Spreadsheets

When reading spatial data from spreadsheets the data is read by different drivers. The example data is contained in an Excel spreadsheet with three variables `X` (longitude), `Y` (latitude), and `Object`. Using the `st_read` function the csv-driver from GDAL is chosen automatically, reading 4 columns however. Thus we get rid of the last column first.

```
POI <- st_read(here('Daten', 'POIs_Berlin.csv'))
```

```
## Reading layer `POIs_Berlin' from data source `/media/H/Projekte/Adhoc_Analysen/
↪GeodatenBSP/Daten/POIs_Berlin.csv' using driver `CSV'
```

```
POI
```

```
##                    X               Y                    Objekt field_4
## 1  13.3776978296166 52.5162788298363         Brandenburger Tor    <NA>
## 2  13.3693019102667 52.5250274843424              Hauptbahnhof    <NA>
## 3  13.3886070846104 52.5121727411876                DIW Berlin    <NA>
## 4  13.4589278721678  52.496825784817              Mulecule Man    <NA>
## 5  13.2887256012712 52.5580119710438           Flughafen Tegel    <NA>
## 6  13.4005469632413 52.4792779435797          Tempelhofer Feld    <NA>
## 7  13.4484578382279 52.4482123689564          Hufeisensiedlung    <NA>
## 8   13.281110837709 52.4473851451014                 FU Berlin    <NA>
## 9  13.1725456943567 52.4300072503147                   Wannsee    <NA>
## 10 13.2128459704725 52.5411476127425         Zitadelle Spandau    <NA>
## 11 13.5727300494297 52.4438163134271          Schloss Köpenick    <NA>
## 12 13.2795658406826 52.5080182126806 Zentraler Omnibus Bahnhof    <NA>
```

```
POI <- POI[, -4]
```

Because this data contains latitude and longitude already, we can simply transform it to a spatial dataset using `st_as_sf` and the CRS has to be assigned to it.

```
POI <- st_as_sf(POI, coords = c("X", "Y"), crs = 4326)
POI
```

```
## Simple feature collection with 12 features and 1 field
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: 13.17255 ymin: 52.43001 xmax: 13.57273 ymax: 52.55801
```

(continues on next page)

(continued from previous page)

```
## geographic CRS: WGS 84
## First 10 features:
##                Objekt                  geometry
## 1  Brandenburger Tor  POINT (13.3777 52.51628)
## 2       Hauptbahnhof  POINT (13.3693 52.52503)
## 3         DIW Berlin POINT (13.38861 52.51217)
## 4       Mulecule Man POINT (13.45893 52.49683)
## 5     Flughafen Tegel POINT (13.28873 52.55801)
## 6    Tempelhofer Feld POINT (13.40055 52.47928)
## 7    Hufeisensiedlung POINT (13.44846 52.44821)
## 8          FU Berlin POINT (13.28111 52.44739)
## 9            Wannsee POINT (13.17255 52.43001)
## 10 Zitadelle Spandau POINT (13.21285 52.54115)
```

6.10.3 Transformations

To be able to work with all three data sets (States, SOEP, and POI) we have to make sure all have the same CRS. Using st_transform we can reproject the data sets to have the same EPSG-code (common_crs).

```
common_crs <- 25832

SOEP <- st_transform(SOEP, crs = common_crs)
SOEP <- SOEP[SOEP$erhebj == 2005, ] # only use 2005 SOEP data

POI <- st_transform(POI, crs = common_crs)

# States <- st_transform(States, crs = common_crs) aleady correct crs
```

Sometimes it might be easier for you to work in other programs and you wish to have latitude and longitude data. In this case you can use st_coordinates to transform point geometries into (x,y) coordinates. The below example first transforms the coordinates for the households 1 to 5 in the SOEP data into latitude and longitude data and then creates the (x,y) coordinates from the point geometry.

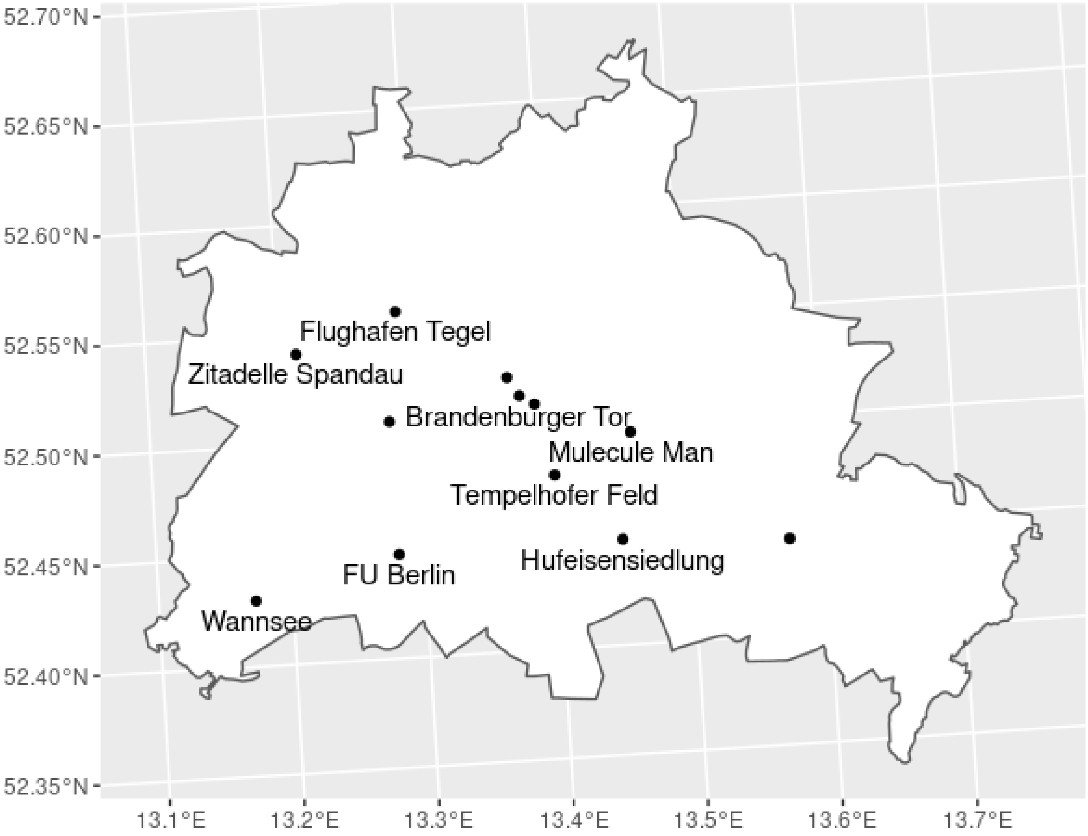
```
soep5_lat_lon <-  st_transform(SOEP[1:5,], 4326)
st_coordinates(soep5_lat_lon)
```

```
##          X        Y
## 1  6.94659 51.12465
## 2 13.36828 52.48428
## 3 13.40601 52.39064
## 4  8.70727 49.03966
## 5 13.43205 52.48811
```

6.10.4 Plotting Spatial Data

Besides being able to look up coordinates or objects in google maps or OpenStreetMap, we can use the ggplot2 package included in the tidyverse package for displaying the data. The package provides the geom_sf function to easily plot the data.

```
# subsetting and plotting the data
States %>% # use the states data
  filter(GEN == 'Berlin') %>% # filter for Berlin only
  ggplot() + # plot base
  geom_sf(fill = 'white') + # add the polygon for Berlin and fill it whtie
  geom_sf(data = POI) + # add the POI data
  geom_sf_text(data = POI, aes(label = Objekt),
               nudge_y = -1000,
               check_overlap = TRUE) + # overlapping labels will not be displayed
  xlab('') + ylab('')
```



6.10.5 Frequently Used Operations

This section will provide an overview of some frequently used operations when working with spatial data. The dataset SOEP contains some fake coordinates of household addresses we will work with throughout the examples.

Finding Households in a Specified Area

Suppose we are interested in identifying all the SOEP households located in Berlin. The corresponding polygon for Berlin is provided in the States dataset. Because we are interested in Berlin only, we save the polygon in an own object BE. The function st_contains identifies the row-index in the SOEP dataset that fall within the polygon BE and returns a list (soep_in_berlin). For checks along the way we can look at plots of the data.

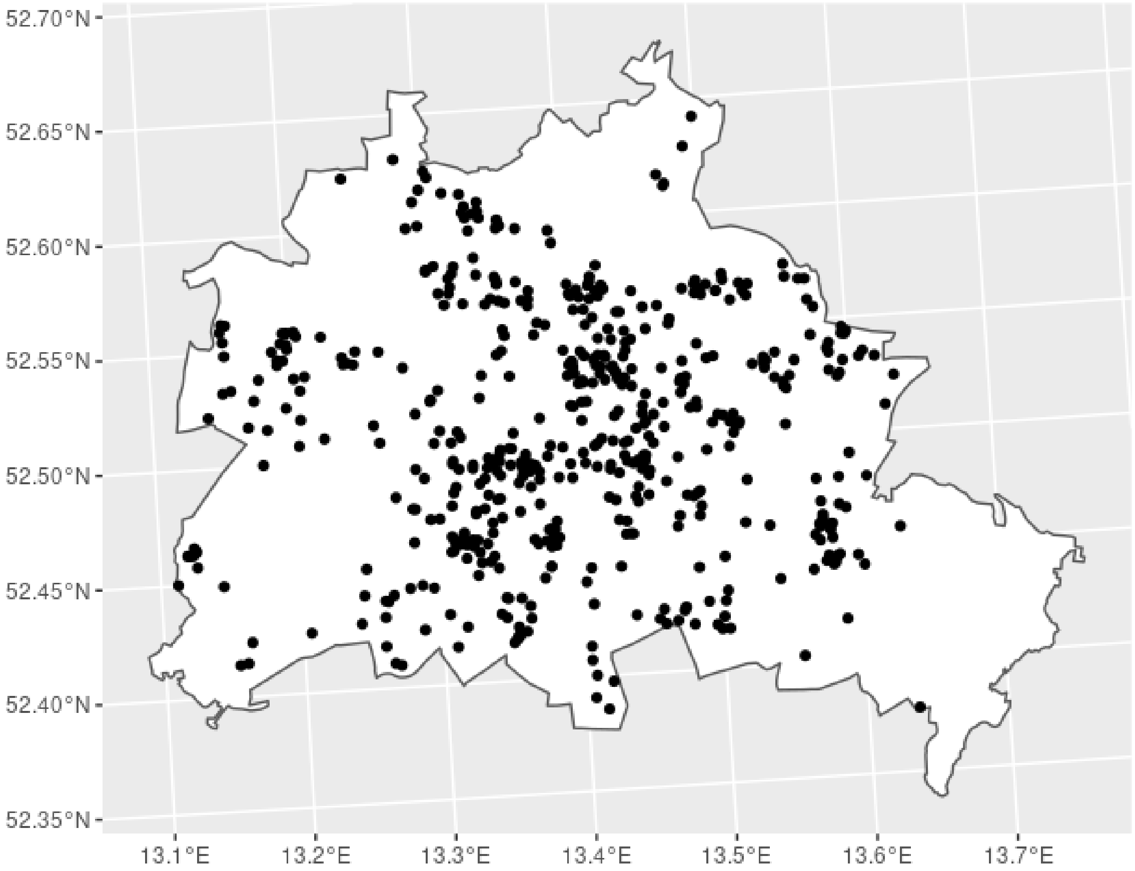
```
BE <- States[States$GEN == 'Berlin', ]
BE %>% select(GEN, BEZ)
```

```
## Simple feature collection with 1 feature and 2 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 777974.1 ymin: 5808837 xmax: 823510.5 ymax: 5845580
## projected CRS:  ETRS89 / UTM zone 32N
##       GEN  BEZ                      geometry
## 11 Berlin Land MULTIPOLYGON (((802831.7 58...
```

```
soep_in_berlin <- st_contains(BE, SOEP)
soep_in_berlin
```

```
## Sparse geometry binary predicate list of length 1, where the predicate was `contains'
##  1: 2, 3, 5, 6, 8, 9, 10, 11, 12, 13, ...
```

```
SOEP_BE <- SOEP[unlist(soep_in_berlin), ]

BE %>%
  ggplot() + geom_sf(fill = 'white') + geom_sf(data = SOEP_BE)
```



Distances / Areas

To compute distances the function `st_distance` can be provided a single dataset with n rows providing a n x n matrix of distances of the geometries contained in the data (`dist_m`). The unit of the distance returned depends on the CRS. In the example provided below the distances of the POIs to the location of the DIW are given in meters.

Providing the function a second dataset of m rows will create a n x m distance matrix (`dist_soep_poi`). The created object is a matrix with 529 rows (SOEP households in Berlin) and 12 columns (POIs in Berlin).

When computing distances on large data sets it might be helpful to subset the data, because the distance of a household in Munich might be irrelevant to a research question focusing on Berlin or distances smaller than 5000m. According the the CRS, consider specifying the `which` argument.

Areas can be computed for (multi-)polygons. The function `st_area` provides the corresponding information.

```
# distances between the POIs
dist_m <- st_distance(POI)
rownames(dist_m) <- POI$Objekt
colnames(dist_m) <- POI$Objekt

# distance of POIs to DIW Berlin
dist_m['DIW Berlin', ]
```

```
## Units: [m]
##         Brandenburger Tor                Hauptbahnhof                  DIW Berlin
```

(continues on next page)

(continued from previous page)

```
##                   870.8119                 1941.2940                    0.0000
##               Mulecule Man           Flughafen Tegel          Tempelhofer Feld
##                  5074.7951                 8488.2154                 3751.7696
##           Hufeisensiedlung                 FU Berlin                   Wannsee
##                  8202.7627                10269.1064                17307.5100
##          Zitadelle Spandau          Schloss Köpenick Zentraler Omnibus Bahnhof
##                 12364.7837                14651.8215                 7422.6480
```

```
# distance between each household and each POI
dist_soep_poi <- st_distance(SOEP_BE, POI)
dim(dist_soep_poi)
```

```
## [1] 529  12
```

```
# save distances in an object
DIST <- as_tibble(dist_soep_poi)
# add names
names(DIST) <- str_c('distance_to_', str_remove(POI$Objekt, ' '))

# attach distances to data
SOEP_BE <- bind_cols(SOEP_BE, DIST)


# area covered by Berlin
st_area(BE)
```

```
## 893060962 [m^2]
```

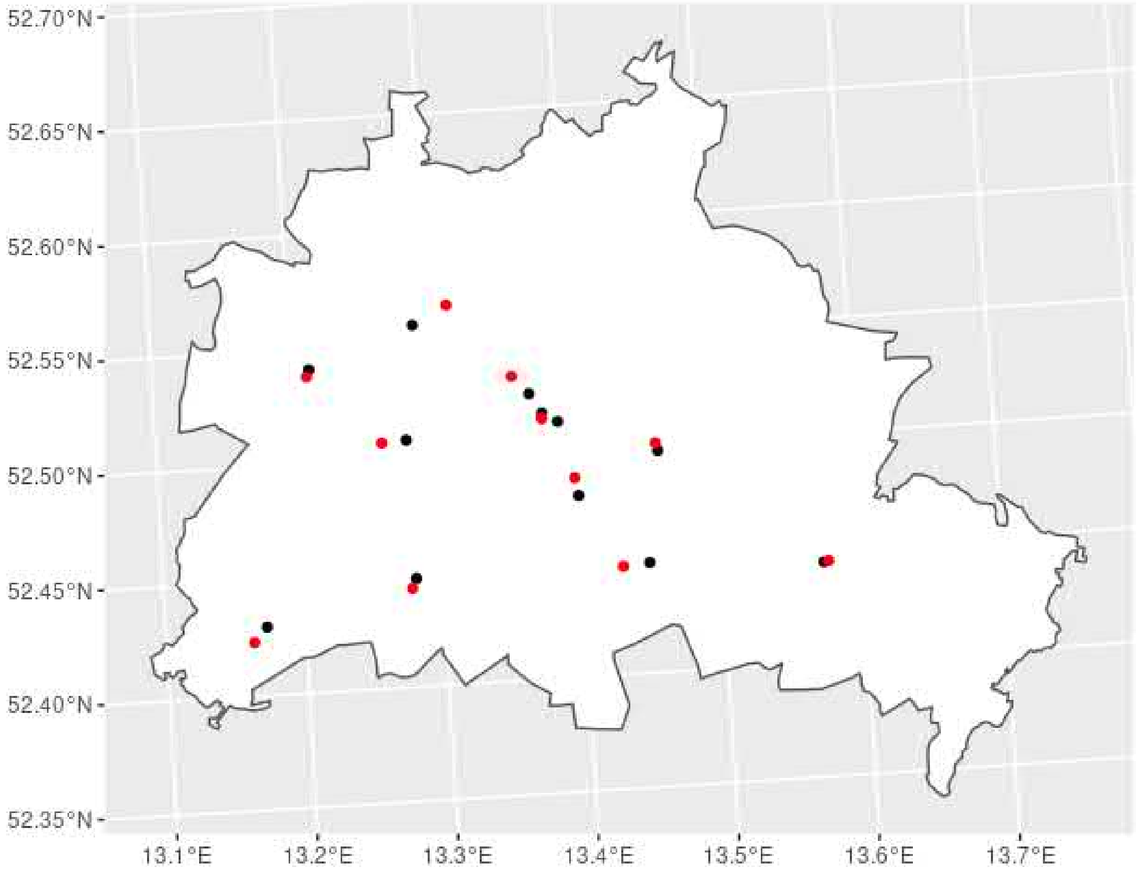
Nearest Point

To find the point or feature closest to another one the function `st_nearest_feature` will return a vector with indices of the nearest feature. In the example below we are looking for the households living closest to the POIs in Berlin. In the second step we compute the corresponding distances between the POI and the household.

```
nearest_hh <- st_nearest_feature(POI, SOEP_BE)

diag(st_distance(POI, SOEP_BE[nearest_hh, ]))
```

```
## Units: [m]
##  [1]  256.5810 1195.8833  789.6812  392.5787 1913.8130  891.4404 1295.8621
##  [8]  520.3830  971.0417  364.4532  250.2833 1194.7832
```

```
BE %>% # polygon for Berlin
  ggplot() + geom_sf(fill = 'white') + # plot the Berlin polygon
  geom_sf(data = POI) + # add the POIs
  geom_sf(data = SOEP_BE[nearest_hh, ], col = 'red') # add the nearest household
```



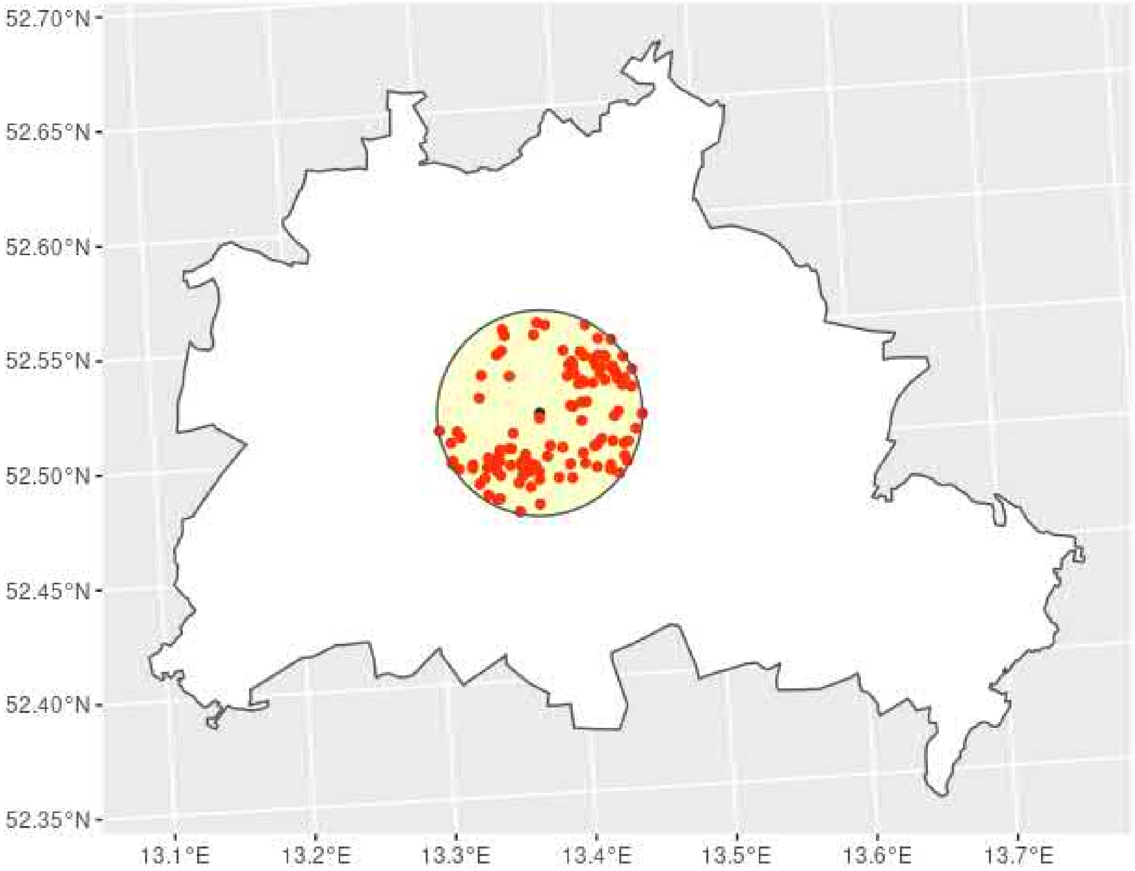
Within Distance

If your interest is about which households live within a certain distance to a specific point, `st_is_within_distance` lets you lookup geometries in a given distance (argument `dist`) and returns a list. The below example looks up households in a 5km distance of the Brandenburger Tor. The plot shows the 5km radius area in yellow, the location of the Brandenburger Tor (black dot) and than households within the distance (red dots).

```
r_5000 <- st_is_within_distance(POI[POI$Objekt == 'Brandenburger Tor',],
                                SOEP_BE,
                                dist = 5000)
r_5000
```

```
## Sparse geometry binary predicate list of length 1, where the predicate was `is_within_
↪distance'
##  1: 1, 3, 6, 12, 21, 25, 28, 39, 54, 55, ...
```

```
BE %>% # polygon for Berlin
  ggplot() + geom_sf(fill = 'white') + # plot the Berlin polygon
  geom_sf(data = POI[POI$Objekt == 'Brandenburger Tor',]) + # add the POI
  geom_sf(data = SOEP_BE[unlist(r_5000), ], col = 'red') + # add the nearest household
  geom_sf(data = st_buffer(POI[POI$Objekt == 'Brandenburger Tor',], dist = 5000),
          alpha = 0.2, fill = 'yellow')
```



The question cal also be asked the other way around: How many POIs are within a 5km radius of the SOEP households? This way the function `st_is_within_distance` returns a list of length equal to the number of SOEP households in Berlin (529). For each household the (row) index for the POI is given. To get the number of POIs in the 5km radius, we can simply ask for the length (the number of row indices) of each list-component. To get the according distances see section *Distances / Areas*

```
poi_5000 <- st_is_within_distance(SOEP_BE, POI, dist = 5000)
poi_5000
```

```
## Sparse geometry binary predicate list of length 529, where the predicate was `is_
↪within_distance'
## first 10 elements:
##  1: 1, 2, 3, 6
##  2: (empty)
##  3: 1, 3, 4, 6, 7
##  4: 5
##  5: 2, 5, 12
##  6: 1, 2, 12
##  7: 10
##  8: 5, 10, 12
##  9: 8
##  10: 5, 10, 12
```

```
N_POI <- as_tibble(sapply(poi_5000, length))
names(N_POI) <- 'n_poi_in_5km'

SOEP_BE <- bind_cols(SOEP_BE, N_POI)
```

Spatial joins

When you are used to working with SOEP data you will have probably merged / joined data sets using the identifying variables (`cid`, `hid`, `pid`) and the survey year (`syear`) before. When you are working with spatial data you will have to choose one of the geometry predicate function provided by the `sf` package. The default is a left join of the two data sets using `st_intersects` as the geometry predicate function for joining. You can however change this, for example, to join the nearest features, see section *Nearest Point*. In our example here, we join the nearest SOEP household to each of the points of interest. The geometry column here provides the coordinates from the POI data set.

```
NEAR <- st_join(POI, SOEP, join = st_nearest_feature)
NEAR
```

```
## Simple feature collection with 12 features and 3 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: 783630.5 ymin: 5817059 xmax: 810721.7 ymax: 5831749
## projected CRS:  ETRS89 / UTM zone 32N
## First 10 features:
##               Objekt erhebj    ID                 geometry
## 1  Brandenburger Tor   2005 75759 POINT (796986.8 5827473)
## 2       Hauptbahnhof   2005 69076 POINT (796358.6 5828411)
## 3         DIW Berlin   2005 75759 POINT (797754.3 5827062)
## 4        Mulecule Man   2005 75646 POINT (802628.5 5825649)
## 5     Flughafen Tegel   2005 73266 POINT (790677.7 5831749)
## 6    Tempelhofer Feld   2005 69820 POINT (798787.2 5823455)
## 7    Hufeisensiedlung   2005 69477 POINT (802251.6 5820202)
## 8           FU Berlin   2005 67568 POINT (790892.2 5819422)
## 9             Wannsee   2005 67923 POINT (783630.5 5817059)
## 10 Zitadelle Spandau   2005 68827 POINT (785646.9 5829572)
```

Export Results

To export your results you can use `st_write` to create a `.csv` file. When exporting your results pleace check the requirements here.

```
path_export <- paste0('/home/', Sys.info()['user'], '/transfer/export/',  Sys.Date())

if(!file.exists(path_export)){
  dir.create(path_export, recursive = TRUE)
}

st_write(SOEP_BE, file.path(path_export, 'Output_SOEP_BE.csv'),
         append = FALSE, overwrite = TRUE)

README <- tibble(name = names(SOEP_BE)[-grep('geometry', names(SOEP_BE))],
                 description = c('erhebj',
                                 'ID',
                                 'distance (in meters) of household to Brandenburger Tor
↪',
```

(continues on next page)

(continued from previous page)

```
                                 'distance (in meters) of household to Hauptbahnhof',
                                 'distance (in meters) of household to DIW-Berlin',
                                 'distance (in meters) of household to Mulecule Man',
                                 'distance (in meters) of household to Flughafen Tegel',
                                 'distance (in meters) of household to Tempelhofer Feld',
                                 'distance (in meters) of household to Hufeisensiedlung',
                                 'distance (in meters) of household to FU Berlin',
                                 'distance (in meters) of household to Wannsee',
                                 'distance (in meters) of household to Zitadelle Spandau
↪',
                                 'distance (in meters) of household to Schloss Köpenick',
                                 'distance (in meters) of household to Zentraler Omnibus␣
↪Bahnhof',
                                 'number of POIs within 5 km radius of household'))

write.csv(README, file.path(path_export, 'Output_SOEP_BE.csv'),
         row.names = FALSE)
```

6.10.6 Complete Example

Suppose you want to know which households of the SOEP from survey year 2011 live within a distance of 5000m to the following points of interest (POI):

- Brandenburger Tor
- Zitadelle Spandau
- Wannsee

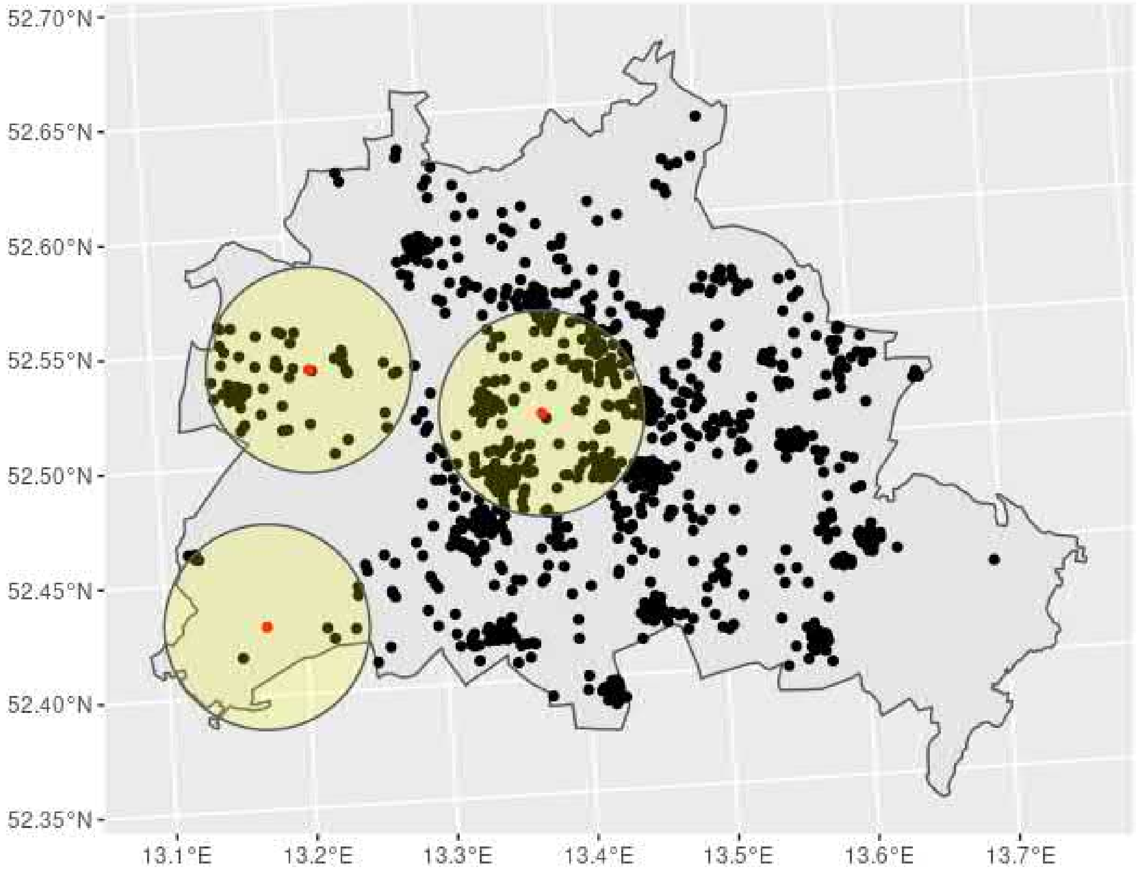
Besides that, you want to know how far their distance to the corresponding POI is and which household lives closest to the corresponding POI. After computing the informations need you want to export the results for further use on the HAUSER server.

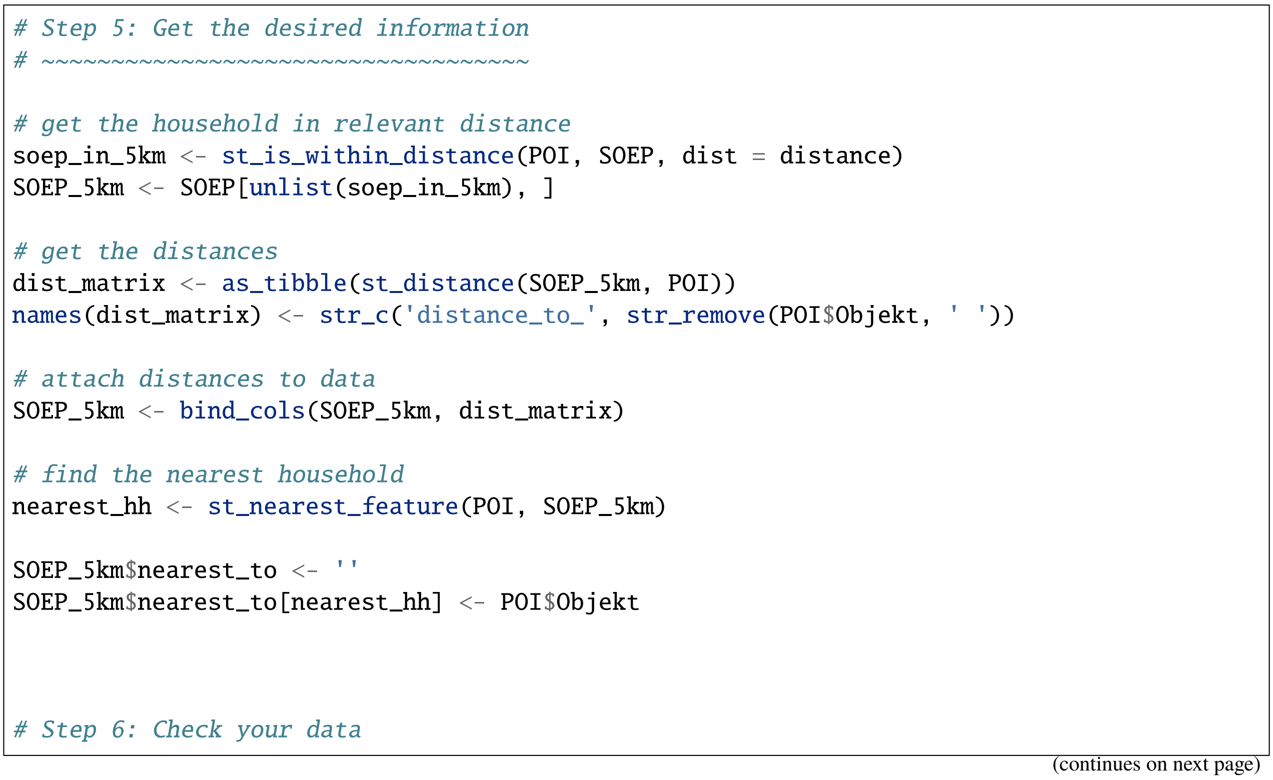
```
# Global stuff
# ~~~~~~~~~~~~

# packages
library(here)
library(sf)
library(tidyverse)

# global values
survey_year <- 2011
distance <- 5000 # meter
common_crs <- 25832



# Step 1: read the data
# ~~~~~~~~~~~~~~~~~~~~~

# read polygons for Federal States
States <- st_read(here('Daten', 'vg250_ebenen_0101'), layer = 'VG250_LAN', quiet = TRUE)
```

(continues on next page)

(continued from previous page)

```
# read SOEP data
SOEP <- st_read(here('Daten', 'soep_v29'), quiet = TRUE)

# read POI data
POI <- st_read(here('Daten', 'POIs_Berlin.csv'), quiet = TRUE)



# Step 2: Transform data
# ~~~~~~~~~~~~~~~~~~~~~~

# transform SOEP the data
SOEP <- st_transform(SOEP, crs = common_crs)

# transform POIs
POI <- POI[, -4]

POI <- st_as_sf(POI, coords = c("X", "Y"), crs = 4326)
POI <- st_transform(POI, crs = common_crs)



# Step 3: Subset the data
# ~~~~~~~~~~~~~~~~~~~~~~~

# Berlin only
BE <- States %>%
  filter(GEN == 'Berlin')

# interesting POIs only
POI <- POI %>%
  filter(Objekt %in% c("Brandenburger Tor", "Zitadelle Spandau", "Wannsee"))

# SOEP survey year: 2011 only
SOEP <- SOEP %>%
  filter(erhebj == survey_year)

# SOEP housholds in Berlin only
SOEP <- SOEP[unlist(st_contains(BE, SOEP)), ]



# Step 4: Plot the data
# ~~~~~~~~~~~~~~~~~~~~~

BE %>%
  ggplot() + geom_sf() +
  geom_sf(data = SOEP) +
  geom_sf(data = POI, col = 'red') +
  geom_sf(data = st_buffer(POI, dist = distance),
          alpha = 0.2, fill = 'yellow')
```



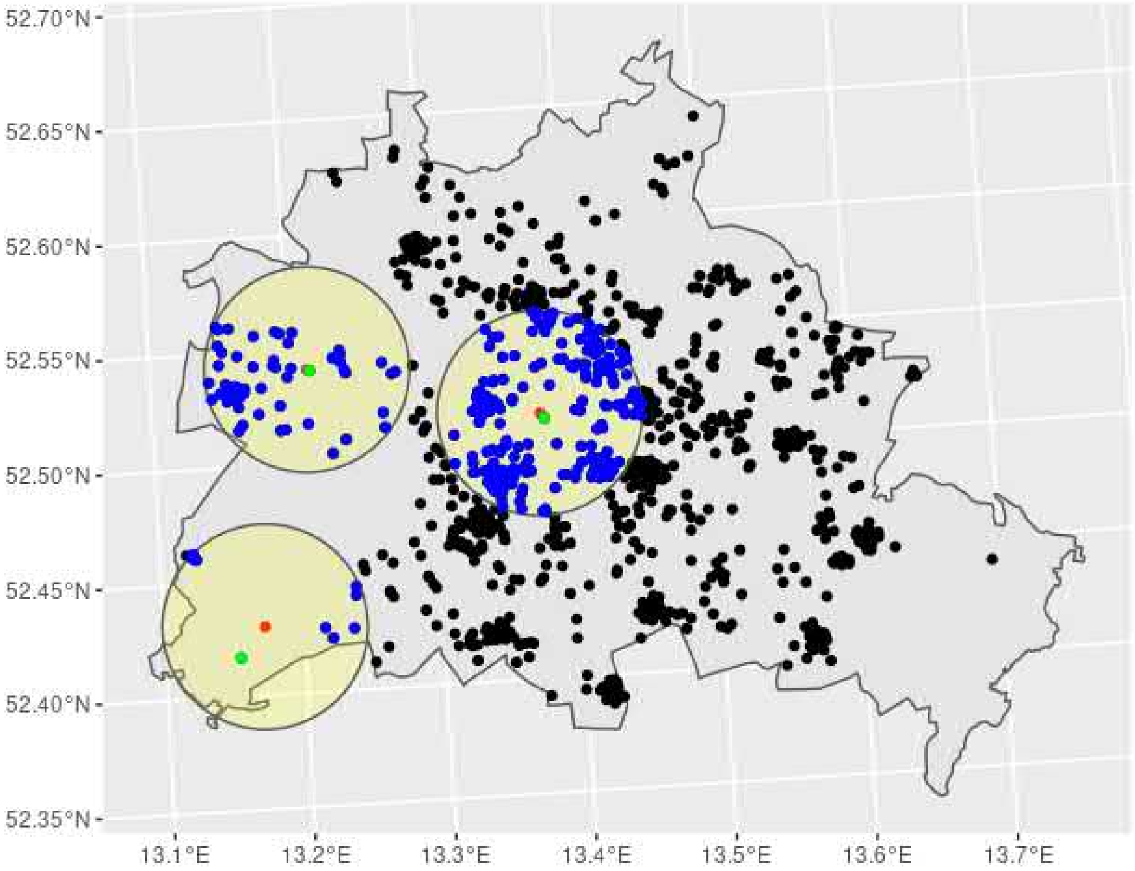


```
# Step 5: Get the desired information
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# get the household in relevant distance
soep_in_5km <- st_is_within_distance(POI, SOEP, dist = distance)
SOEP_5km <- SOEP[unlist(soep_in_5km), ]

# get the distances
dist_matrix <- as_tibble(st_distance(SOEP_5km, POI))
names(dist_matrix) <- str_c('distance_to_', str_remove(POI$Objekt, ' '))

# attach distances to data
SOEP_5km <- bind_cols(SOEP_5km, dist_matrix)

# find the nearest household
nearest_hh <- st_nearest_feature(POI, SOEP_5km)

SOEP_5km$nearest_to <- ''
SOEP_5km$nearest_to[nearest_hh] <- POI$Objekt



# Step 6: Check your data
```

(continues on next page)

(continued from previous page)

```
# ~~~~~~~~~~~~~~~~~~~~~~~

BE %>%
  ggplot() + geom_sf() +
  geom_sf(data = SOEP) +
  geom_sf(data = POI, col = 'red') +
  geom_sf(data = st_buffer(POI, dist = distance),
          alpha = 0.2, fill = 'yellow') +
  geom_sf(data = SOEP_5km, col = 'blue') +
  geom_sf(data = SOEP_5km[SOEP_5km$nearest_to != '', ], col = 'green')
```



```
# Step 7: Export your data
# ~~~~~~~~~~~~~~~~~~~~~~~~
path_export <- paste0('/home/', Sys.info()['user'], '/transfer/export/',  Sys.Date())

if(!file.exists(path_export)){
  dir.create(path_export, recursive = TRUE)
}

st_write(SOEP_5km, file.path(path_export, 'Output.csv'),
         append = FALSE, overwrite = TRUE)
```

```
## Deleting layer `Output' using driver `CSV'
## Writing layer `Output' to data source `/home/hwsteinhauer/transfer/export/2021-01-27/
↪Output.csv' using driver `CSV'
## Updating existing layer Output
## Writing 356 features with 6 fields and geometry type Point.
```

```
README <- tibble(name = names(SOEP_5km)[-grep('geometry', names(SOEP_5km))],
                 description = c('erhebj',
                                 'ID',
                                 'distance (in meters) of household to Brandenburger Tor
↪',
                                 'distance (in meters) of household to Wannsee',
                                 'distance (in meters) of household to Zitadelle Spandau
↪',
                                 'household nearest to the POI'))

write.csv(README, file.path(path_export, 'README.csv'),
         row.names = FALSE)
```

6.10.7 Appendix

Session info - Platform

```
##  setting  value
##  version  R version 4.0.3 (2020-10-10)
##  os       Ubuntu 20.04.1 LTS
##  system   x86_64, linux-gnu
##  ui       X11
##  language (EN)
##  collate  de_DE.UTF-8
##  ctype    de_DE.UTF-8
##  tz       Europe/Berlin
##  date     2021-01-27
```

Session info - Packages

```
##  package     * version date       lib source
##  bookdown    * 0.21    2020-10-13 [1] CRAN (R 4.0.3)
##  dplyr       * 1.0.2   2020-08-18 [1] CRAN (R 4.0.2)
##  forcats     * 0.5.0   2020-03-01 [1] CRAN (R 4.0.0)
##  ggplot2     * 3.3.2   2020-06-19 [1] CRAN (R 4.0.1)
##  gridExtra   * 2.3     2017-09-09 [1] CRAN (R 4.0.0)
##  here        * 1.0.0   2020-11-15 [1] CRAN (R 4.0.3)
##  kableExtra  * 1.3.1   2020-10-22 [1] CRAN (R 4.0.3)
##  knitr       * 1.30    2020-09-22 [1] CRAN (R 4.0.2)
##  pacman      * 0.5.1   2019-03-11 [1] CRAN (R 4.0.2)
##  purrr       * 0.3.4   2020-04-17 [1] CRAN (R 4.0.0)
##  readr       * 1.4.0   2020-10-05 [1] CRAN (R 4.0.3)
##  rgdal       * 1.5-19  2021-01-05 [1] CRAN (R 4.0.3)
##  sessioninfo * 1.1.1   2018-11-05 [1] CRAN (R 4.0.0)
##  sf          * 0.9-7   2021-01-06 [1] CRAN (R 4.0.3)
```

(continues on next page)

(continued from previous page)

```
##  sp          * 1.4-4   2020-10-07 [1] CRAN (R 4.0.3)
##  stringr     * 1.4.0   2019-02-10 [1] CRAN (R 4.0.0)
##  tibble      * 3.0.4   2020-10-12 [1] CRAN (R 4.0.3)
##  tidyr       * 1.1.2   2020-08-27 [1] CRAN (R 4.0.2)
##  tidyverse   * 1.3.0   2019-11-21 [1] CRAN (R 4.0.0)
##
## [1] /home/hwsteinhauer/R/x86_64-pc-linux-gnu-library/4.0
## [2] /usr/local/lib/R/site-library
## [3] /usr/lib/R/site-library
## [4] /usr/lib/R/library
```

Section author: Hans Walter Steinhauer hwsteinhauer@diw.de

Last updated: 2021-02-22

6.11 How to Use SOEP IGEL

6.11.1 IGEL Workstation

IGEL refers to a computer terminal workstation for access to SOEP data:

- The terminal allows data to be entered and displayed.
- The IGEL is a so-called thin client, a computer with little computing power, which only provides a terminal to a server.
- This thin client at the SOEP guest workstation/FDZ is from the manufacturer Igel Technology, where IGEL stands for "**I** ntelligente **G** esamtlösung in der Mikro **el** ektronik".

Account

Access to SOEP data can only be provided in compliance with high security standards to protect respondents' confidentiality and maintain their trust in the survey. The data are also provided solely for scientific research purposes, that is, they are only made available to members of the scientific community. Researchers can therefore only access SOEP data after they have signed a data distribution contract with DIW Berlin. The same rules apply to the secure guest workstations at RDC SOEP and at other secure data access points. Since IGEL terminals also provide access to small-scale regional data, users have to sign additional contracts for these data.

All IGEL users must sign a data distribution contract with the DIW Berlin: Application for a Data Distribution Contract.

6.11.2 Logging in

Turn on the computer and the following screen should appear on the monitor. (see *figure 1*)

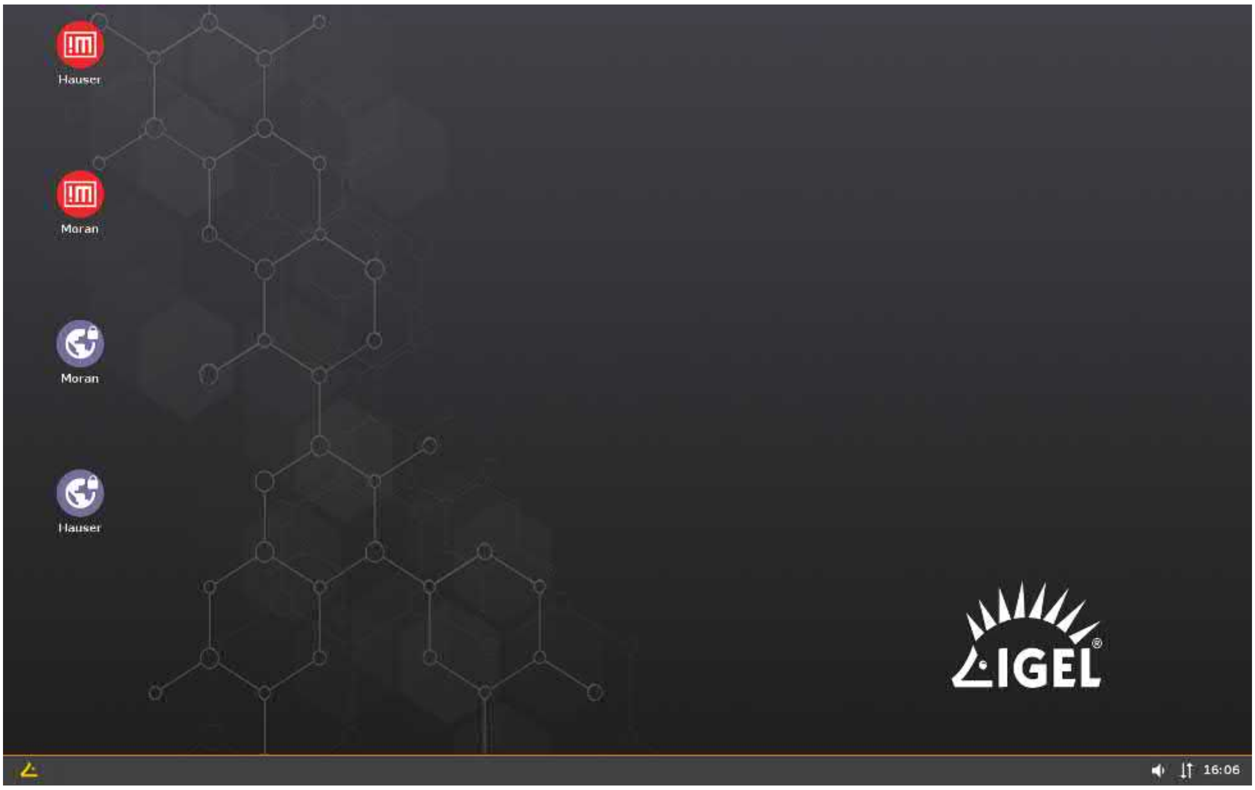
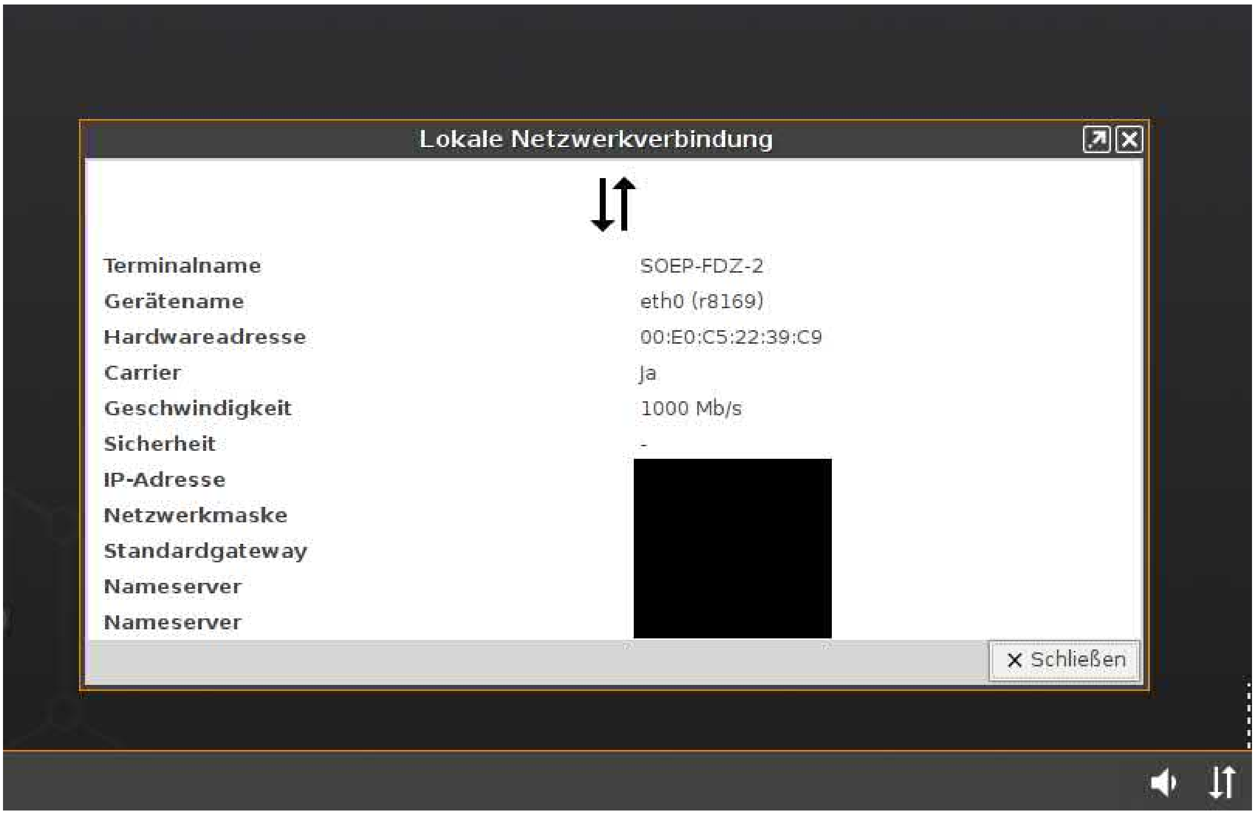


Fig. 1: Figure 1: IGEL start screen

At the bottom right, you should see the icon for an existing network connection appear: Two arrows, one pointing up and one pointing down.

Click on the arrow icon to see the terminal name and the existing network connection . See *figure 2*.



Terminalname	SOEP-FDZ-2
Gerätename	eth0 (r8169)
Hardwareadresse	00:E0:C5:22:39:C9
Carrier	Ja
Geschwindigkeit	1000 Mb/s
Sicherheit	-
IP-Adresse	[redacted]
Netzwerkmaske	[redacted]
Standardgateway	[redacted]
Nameserver	[redacted]
Nameserver	[redacted]

Fig. 2: Figure 2: Connection with LAN available

For each available server, two icons are displayed on the start screen at the top left, a red one and a blue one with the same name. See *figure 3*.

The following two servers are currently available:

1. HAUSER: Access to the SOEP survey data, including connection to small-scale regional indicators (WITHOUT coordinates).
2. MORAN: Access to the coordinates of SOEP households, but without survey data.

Access is only possible from RDC SOEP guest stations at DIW Berlin

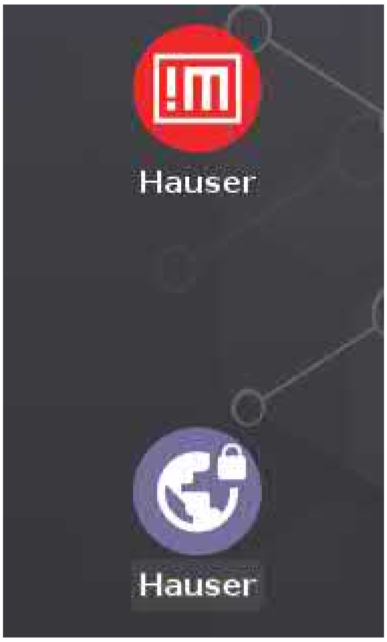
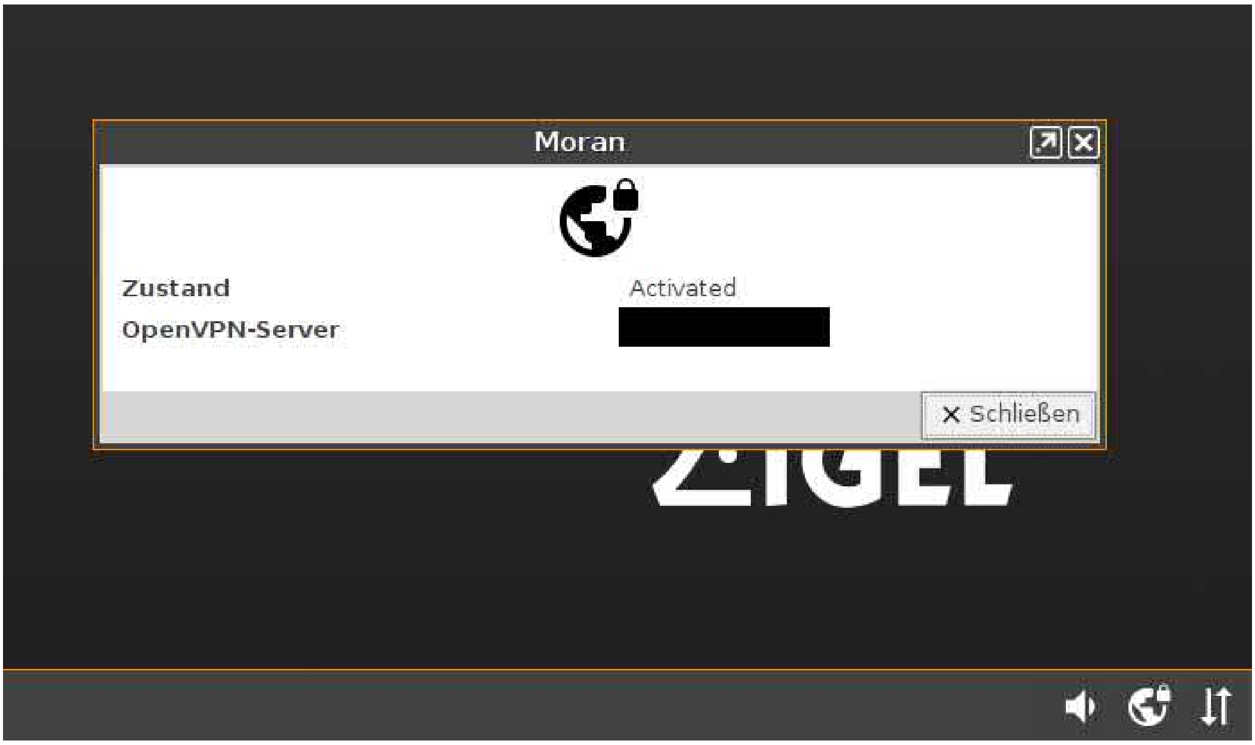


Fig. 3: Figure 3: Icons to connect with the SOEP server

Blue Icon: To connect to one of the two servers at RDC SOEP, first establish an open VPN connection by clicking on the blue icon for the server you would like to connect to. The icon in the lower right corner should then display the existing VPN connection. By clicking once on this icon, you can see the server's IP address . See *figure 4*

Red Icon: Once you have established the VPN connection to the SOEP server, click once on the red icon to start your session. The server's login window should appear, see *figure 5*. Enter the user name and password provided to you by RDC SOEP.

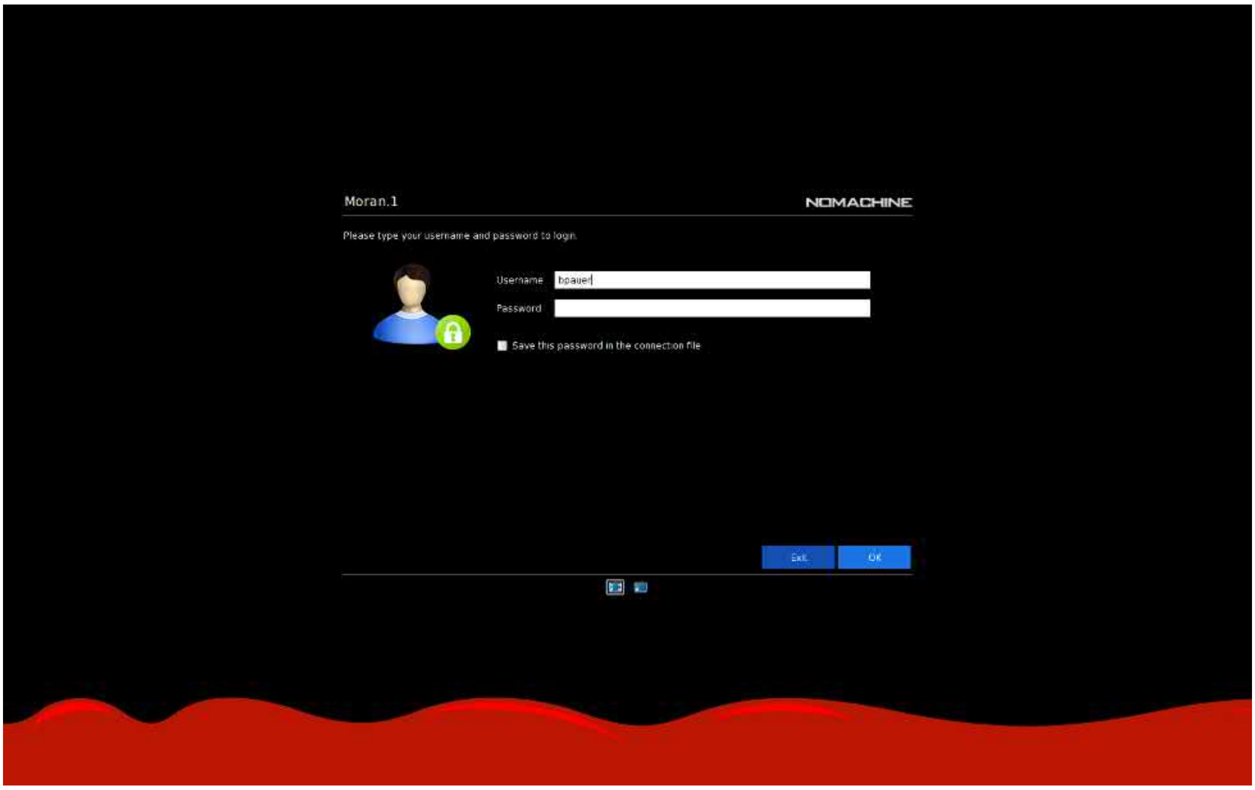


Moran

Zustand Activated

OpenVPN-Server [redacted]

Fig. 4: Figure 4: Open VPN connection established



Moran.1

Please type your username and password to login.

Username bpauer

Password

□ Save this password in the connection file

Fig. 5: Figure 5: Login to the SOEP Server

6.11.3 Working with SOEP DATA

Starting programs

- After you have logged in, a blank desktop will appear with a menu bar at the top.
- In general, programs can be started by clicking on “activities” and then either by clicking on the icon or by typing the name of the program into the search field.
- Users should inform the RDC SOEP team in advance about any additional ados in Stata or packages in R. These will be installed after checking.
- **Start Stata**: Unfortunately, there is no automatic start icon for Stata, so you have to do the following:
  1. Click on activities
  2. Enter “terminal” in the search window
  3. Start either “Terminal” or “XTerm”.
  4. Enter the command “xstata-mp” into the terminal that has now appeared, and press the return key. Stata should now appear.
- The following table shows which programs are installed and available for use on each server:

Program	HAUSER	MORAN
Stata	Yes	-/-
R/RStudio	Yes	Yes
QGIS	-/-	Yes
grass	-/-	Yes
PostGis	-/-	Yes
LibreOffice	Yes	Yes
Emacs	Yes	Yes
gnome-text-editor	Yes	Yes
Nautilus (File manager)	Yes	Yes

Using SOEP data and your own data

- The latest version of the SOEP data is available at the following address directory path:

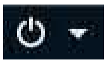
  HAUSER `~/soep-data/` or `/import/SOEP-Regio/data/`

  MORAN `~/soep-data/` or `/import/SOEP-GIS/data`

- You can store your own data and scripts in your personal home directory. `~/work/`

Logging out

- Use the icon in the upper right corner



- click on your username and on logout.

6.11.4 Importing Scripts or External Data

- You can send these data to the RDC team before your stay. Send it to SOEP. Please use the following website: cs-soep.diw.de
- Before you send us your files (only data files, text files and tables), please put all files into a zip archive and name it as **user-YYYY-MM-DD.zip** (mustermann-2020-12-24). Please do not send ados, binaries or r packages in the zip file, ados or r packages will be installed centrally by the SOEP team.
- As receiver for the data and scripts please use soepmail@diw.de.
- Before you come to us, please send us the data to import early (2 days in advance) enough so that we have enough time to install it.
- You will be able to find and use yout imported data here: `/home/USER/transfer/import/`
- You can read, write and save in your personal directory: `/home/USER/work/`

Attention:

Because disk space is limited, we had to introduce the concept of quotas:

- each user gets 10 GB of disk space
- to display there is the quota command
- the data remains on the server until the end of the project duration.
- after the end of the project, the data is taken from the server and archived for 10 years.
- it is possible to upload the data to the server again later with sufficient preparation time

6.11.5 Instructions for exporting from Hauser to user

From a secure guest workstation at the SOEP Research Data Center, users can analyze SOEP data in combination with small-scale regional data. However, to provide users with this sensitive information, we have to carry out additional protective measures of both a technical and organizational nature. At a guest workstation at the SOEP Research Data Center, you work on a thin client from which you cannot export any data on your own. Below we describe how you can obtain the results of your analyses after they have been checked for anonymity.

How can I take my results with me?

In your transfer folder, you will find an import folder (containing your external data that have been imported into the system) and an export folder.

1. Create on the server 'Hauser' below directory '~/tranfer/export/' a new subdirectory with a name as the current date in ISO-format:

mkdir /home/USER/transfer/export/yyyy-mm-dd

Eg.: You are user Jane Doe on the server 'Hauser' and today is February 29, 2021

> *jdoe@Hauser mkdir /home/jdoe/transfer/export/2021-02-29*
>
> *We know there wasn't a February 29th in 2021, but that's just a format example*

2. This folder should contain the following

1. The results that you definitely need to take with you (for formal criteria, see below)
2. A README file (as a .txt file, Word file, or Libre Office file) in which you briefly describe each file in the export folder

3. Please make sure that the README file is readable and that **line breaks** are used

3. Check your files

- Before you make a request for an export please check your data structure with the `OutputControl` command.
- Execute this command in the export folder you want to export.
- This command is used to check whether the formal requirements of the files are met (more information in the chapter "Formal criteria for exporting files").

Change to the folder you want to export

```
cd /home/USER/transfer/export/yyyy-mm-dd
```

In the Terminal, enter the following command: `OutputControl`

Check the `Control_output_USER_yyyy_mm_dd.txt` output in the folder Control in your export folder.

- If you want to make a request for an export, the control file should not contain any warnings.
- If you have any questions, please contact the SOEP hotline.

4. When your folder is complete, please send an e-mail to soepmail@diw.de with your export request

Before you submit an export request, please check that your export is complete and ensure that the following criteria have been met:

> **Attention:** Please read the following rules carefully. If you break the rules, you will not receive your export files.

Formal criteria for exporting files:

- Microdata sets at household or personal level will NOT be exported.
- Only the outputs of analysis (tables, figures), syntax files, and log files will be exported:
- **Tables:**
  - must be stored in the file format **.csv**
  - the maximum number of text files and tables is **200**
- **Figures:**
  - must be saved in one of the following file formats: **.png, .svg, .jpg, .tiff, .eps, .pdf**
- **Text files (scripts or log files):**
  - must be stored in one of the following file formats: **.txt, .tex, .do, .r, .pdf, .log, .md**
  - may have a maximum of **25,000 lines** (a command to count these from a terminal for all .log files in a directory is wc -l *.log)
  - **the maximum number of text files and tables is 200**
    - Please make sure that the files are readable and that **line breaks** are used
- Please note that **no special characters or spaces** are used in file names. Please check if the files are really **readable** after creating.
- An export request can only be made **once a week**

Criteria for exporting results:

- In principle, the results cannot allow any conclusions to be drawn as to which spatial planning region (or smaller-scale geographic unit) a household or individual was or is part of.

- No regional information (e.g., municipality code, district code, zip code . . .) may be listed (e.g., using the list command in Stata) together with identifiers (e.g., individual ID number, household ID number)
- When creating tables and figures, the minimum cell population must be kept at 10 if region-specific characteristics are used.

Additional notes on export:

- Since the export has to be checked manually, checking can take up to two or more weeks, of course depending on the number of files to be checked.
- The export link sent to you will only be available for a specified period of time (at least two weeks).
- To open the export link, use your guest access password.

6.11.6 Data transfer from Moran to Hauser

From the three servers of "SOEPgeo", or the SOEP Research Data Center's guest network, SOEP users can analyze geocoded data for scientific purposes on site at the SOEP Research Data Center. Researchers are first required to sign a data protection agreement, and a complete record is kept of all data access. The concept is to keep the geo-coordinates of SOEP households separate from the actual survey information throughout the entire process of analysis by data users. Only the coordinates and the survey year are needed generate topic-related indicators in a geographic information system (GIS; grass, qgis, and postgis are installed on Moran) or in the statistical package R, and no further information on either the household or household members. SOEP Research Data Center staff transfer indicators generated by users in a GIS. This prevents any possibility of users accessing the data. The key component of the data protection concept is that SOEP households' geo-coordinates are kept separate from the survey information:

- At no time do data users have simultaneous access to coordinates and survey data
- Data users can only generate topic-related indicators on Moran, where the SOEP survey data are not accessible.
- Data users can only analyze the topic-related indicators on Hauser, where the SOEP-household coordinates are not accessible.
- Topic-related indicators that were generated based on household coordinates may only be analyzed on the Hauser server and may not be exported from there.

The data user therefore has no simultaneous access to the SOEP survey data and the geo-coordinates of SOEP households. The results (exported Hauser results) may only be published in completely anonymous form.

How do I initiate data transfer from Moran to Hauser?

Attention: Please read the following rules carefully. If you break the rules, the data transfer cannot be executed.

Steps to initiate data export by the SOEP Research Data Center:

1. Create a subdirectory in the export folder on Moran with the export date: mkdir /home/USER/transfer/export/yyyy-mm-dd
2. This folder should contain both the dataset to be exported and a corresponding *README.csv*:
   - dataset with generated indicators and ID (see below)
   - README.csv (see below)
3. Send an e-mail to soepmail@diw.de with the following information:
   - What input dataset was used for the coordinates? To ensure correct data transfer, we need to know what version of the data was used (e.g., v35)

  - What is the export file format? (.rds, .shp, .csv are permitted) (to save in dataframe in rds format please use `saveRDS()`)
  - What are the unique identifiers for the dataset? (e.g., ID & syear)

Formal criteria for data transfer:

The following criteria apply to exports:

- The *README.csv* is a two-column .csv table
  - $name: column containing the variable names of the indicators to be exported (e.g., distance)
  - $description: short description of the respective variable (e.g., distance in meters to the next flood point for household i in year t (for the flood in 2002))
- The following applies to the dataset containing the indicators to be exported
  - Dataset must have the column/variable ID from the input dataset used
  - Permissible file formats: rds, shp, csv
  - Dataset otherwise only contains the indicators described in the README.csv file

Additional notes on data transfer:

- After the data transfer has taken place, the output (datasets, transfer scripts) will be stored in your transfer folder on Hauser, in a subdirectory of your import folder that is identified by date (/home/USER/transfer/import/fromMoran/yyyy-mm-dd)

Section author: Jan Goebel <jgoebel@diw.de>

Last change: Sep 26, 2022

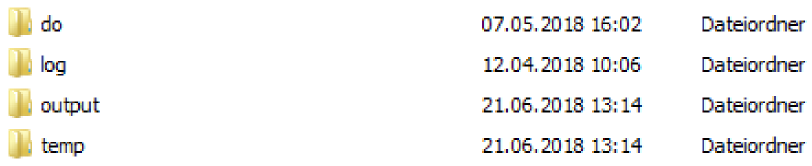
If you want to import the SOEP data as csv files with an older version of Stata, this exercise will help you.

6.12 Working with SOEP data in csv format

SOEP offers the data in statistical program specific file formats (e.g.: Stata .dta) and also as comma-separated values FIle (csv). With these csvs you can read the non-formatted information directly into a statistical program of your choice.

This example shows how to open SOEP data of data version v.36 in csv format with an old Stata version (12) and how to prepare the data in an efficient way.

Create an exercise path with four subfolders:



do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define the paths you created with globals:

```
**************************************************
* Set relative paths to the working directory
**************************************************
global AVZ          "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\distribution\soep-core\soep.v35\csv"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to your ordered data.

For the following script to work, the global "MY_IN_PATH" must contain the folder path to the SOEP csv files of all datasets. The csv files for each data set should always consist of three csvs. If we want to import and prepare the dataset jugendl in csv format, we need the following csv Files:

- jugendl.csv
- jugendl_variables.csv
- jugendl_values.csv

In the SOEP, the csv of each data set contains the variables as columns and their numerical values. Variables and Values csvs contain the variable labels and the value labels for the data set. First some packages for Stata have to be installed so that the process can start.

```
* Import and Labeling of SOEP csv-Files
clear
set more off

* Load ados
capture which adolist
if _rc==111{
        ssc install adolist
}
quietly adolist list
local allAdos `r(names)'
foreach package in fre labutil2 chardef labundef saveascii useold {
        if !regexm("`r(names)'", " `package' ") {
                display as result "Paket " as error  "`package'" as result " wird versucht über SSC-Server zu installieren"
                ssc install `package'
        }
}
```

Once the packages are installed, you will need to define the following functions to be able to label your dataset later. We define the function soeplabelsvars for linking the variables to the variable labels.

```
* Assign German variable labels from *_variables.csv
capture program drop soeplabelsvars
program soeplabelsvars
```

(continues on next page)

(continued from previous page)

```
version 12
syntax , varlabels(string)
  preserve
    insheet using "`varlabels'", clear names
        putmata varLab = (variable label_de) ,replace
    restore
    foreach variable of varlist * {
        label variable `variable' ""
    }

    mata: st_local("n", strofreal(rows(varLab)))
    forvalues i = 1/`n' {
        mata: st_local("varName",varLab[`i',1])
        mata: st_local("varLabel",varLab[`i',2])
        capture confirm variable  `varName'
        if !_rc {
            di "Variable: `varName' mit -`varLabel'- gelabelt"
            label variable `varName' "`varLabel'"
        }
        else di "Variable " as error "`varName'" as result " nicht vorhanden"
    }
end
```

The soeplabelvals function links the information in the data set with valuelabels.

```
* Assign German value labels from *_values.csv
capture program drop soeplabelsvals
program soeplabelsvals
version 12
syntax , vallabels(string)
  quietly label drop _all
    quietly labundef , detach
    preserve
    insheet using "`vallabels'", clear names
      quietly tostring value, replace
      putmata valLab = (variable value label_de) ,replace
      quietly levelsof variable, local(variables) clean
    restore
    foreach variable in `variables' {
            di "------------"
            di "Variable `variable' wird gelabelt"
      mata:  valLabVar= select(valLab, valLab[.,1]:=="`variable'")
            mata: st_vlmodify("`variable'",  strtoreal(valLabVar[.,2]) ,
↪valLabVar[.,3])
            capture confirm variable  `variable'
            if !_rc label value `variable' `variable' , nofix
      else di "Variable " as error "`variable'" as result " nicht vorhanden"
    }
end
```

After both functions have been loaded we can define in a local the dataset we want to import and prepare as csv. We load the csv via the insheet command. Then we use the defined functions and use the variables.csv and values.csv provided by SOEP to label the data.

```
* import and label dataset
local dataset = "jugendl"
insheet using "$MY_IN_PATH/`dataset'.csv", clear names
soeplabelsvars, varlabels("$MY_IN_PATH/`dataset'_variables.csv")
soeplabelsvals, vallabels("$MY_IN_PATH/`dataset'_values.csv")
```

Congratulations you should now have a fully labeled dataset!

Last change: Sep 26, 2022

CHAPTER

SEVEN

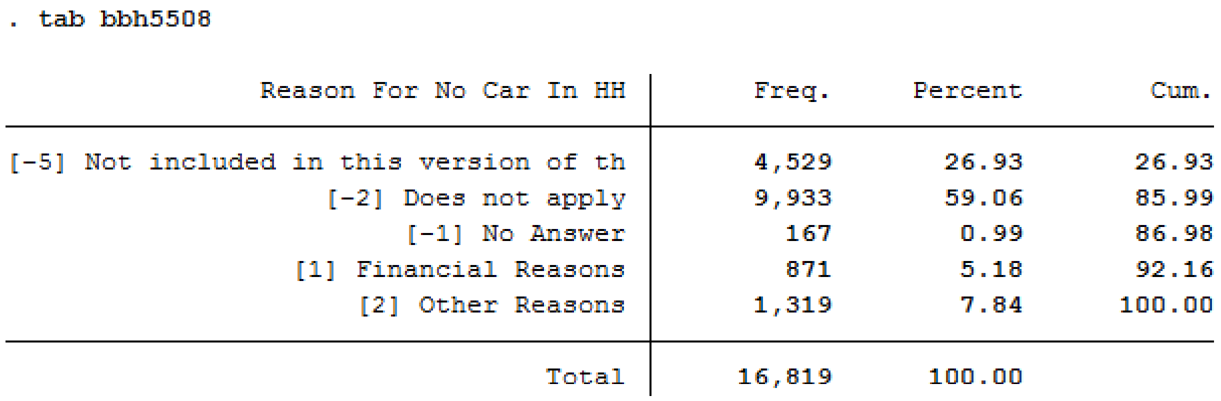
WORKING WITH SOEP DOCUMENTATION

7.1 Variable Search with Questionnaires

If you come across a variable in the dataset whose variable content is unclear, you should always check whether there is a suitable questionnaire for the dataset. Under *Original Core Data* you can see whether the datasets correspond to a survey instrument. The related questionnaires can be found here:

Questionnaires

Example: Working on a research project, you come across the variable bjh_16_04 with the German label "Auto: Gründe" (Car: Reasons) and the English label "Reason for No Car in Household"



```
. tab bbh5508

              Reason For No Car In HH |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
[-5] Not included in this version of th |      4,529       26.93       26.93
                    [-2] Does not apply |      9,933       59.06       85.99
                         [-1] No Answer |        167        0.99       86.98
                  [1] Financial Reasons |        871        5.18       92.16
                      [2] Other Reasons |      1,319        7.84      100.00
--------------------------------------+-----------------------------------
                                  Total |     16,819      100.00
```

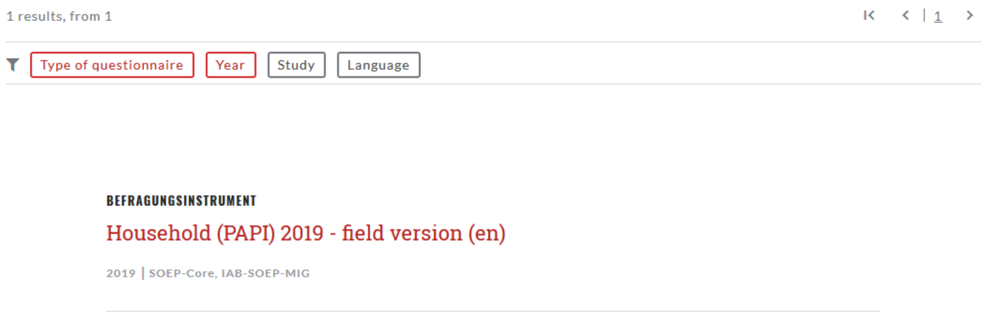
Unfortunately, it is unclear what exactly this variable represents. You should refer to the questionnaires for the complete question and possible filter instructions.

Example Variable:

bjh_16_04: Wave "bj" (Survey Year 2011); household questionnaire ("h"), question number 16, item 4

Open Questionnaires

The variable "bjh_16_04" can be found in the questionnaires for 2019. Select the survey year and questionnaire by using the filter "Year" and "Type of Questionnaire" and download the household questionnaire.



1 results, from 1

Type of questionnaire | Year | Study | Language

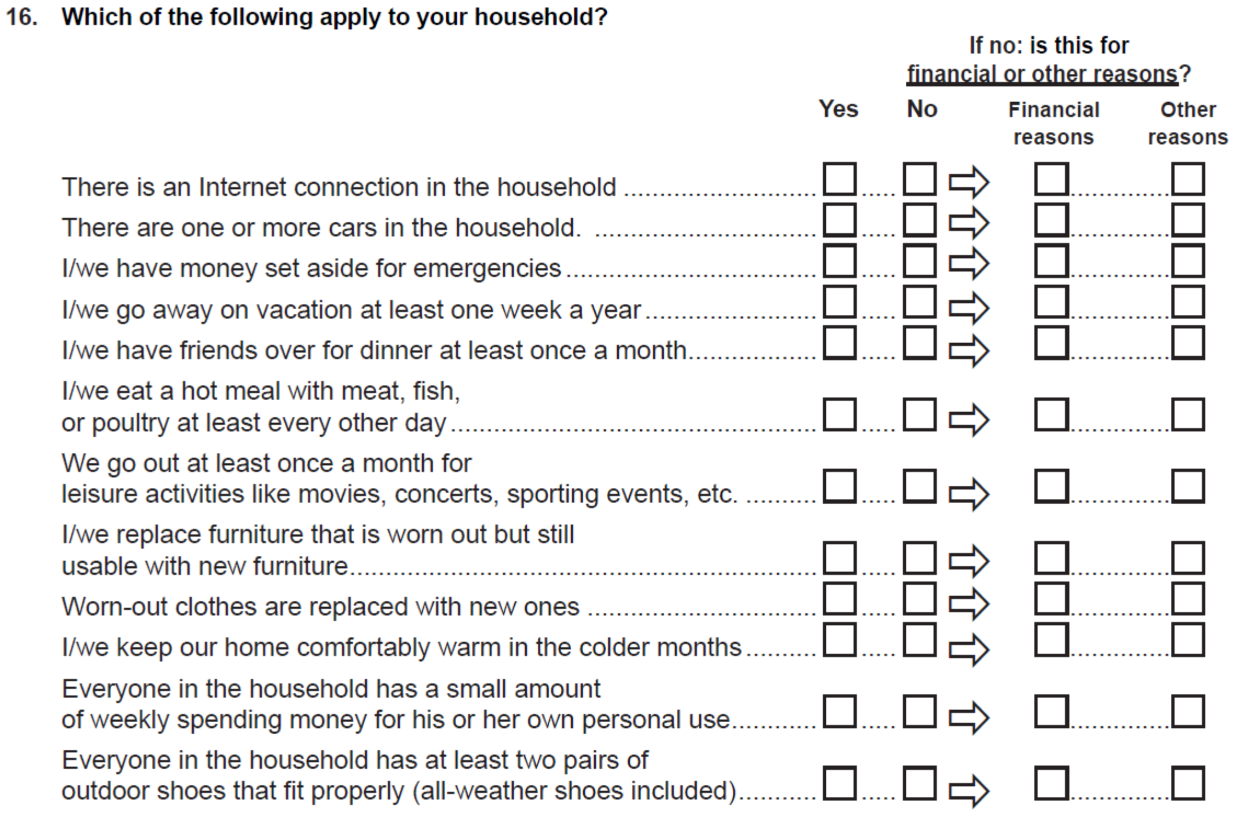
BEFRAGUNGSINSTRUMENT

Household (PAPI) 2019 - field version (en)

2019 | SOEP-Core, IAB-SOEP-MIG

Search the variable “bjh_16_04” in the questionnaire.

Since you are already in the correct questionnaire, you must now search for question 16.



16. Which of the following apply to your household?

				If no: is this for financial or other reasons?	
	Yes	No		Financial reasons	Other reasons
There is an Internet connection in the household	□	□	⇨	□	□
There are one or more cars in the household.	□	□	⇨	□	□
I/we have money set aside for emergencies	□	□	⇨	□	□
I/we go away on vacation at least one week a year	□	□	⇨	□	□
I/we have friends over for dinner at least once a month	□	□	⇨	□	□
I/we eat a hot meal with meat, fish, or poultry at least every other day	□	□	⇨	□	□
We go out at least once a month for leisure activities like movies, concerts, sporting events, etc.	□	□	⇨	□	□
I/we replace furniture that is worn out but still usable with new furniture	□	□	⇨	□	□
Worn-out clothes are replaced with new ones	□	□	⇨	□	□
I/we keep our home comfortably warm in the colder months	□	□	⇨	□	□
Everyone in the household has a small amount of weekly spending money for his or her own personal use	□	□	⇨	□	□
Everyone in the household has at least two pairs of outdoor shoes that fit properly (all-weather shoes included)	□	□	⇨	□	□

To understand which information the variable “bjh_16_04” contains, you have to deal with the question. For each answer category, respondents should indicate whether or not the shown items apply to the household. If the item does not apply, respondents must answer an additional question about the reasons. Both questions should be understood as separate variables.

E.g. the variable “bjh_16_01” indicates whether an internet connection is available in the household. The reasons why there is no internet in the household can be found in the variable “bjh_16_02”. The variable “bjh_16_03” shows whether a car is present in the household and the variable “bjh_16_04” shows reasons why no car is present in the household. By looking into the questionnaire, the variable is now easier to understand. The variable “bjh_16_04” only contains people who do not have a car in their household and shows the reasons given.

Last change: May 12, 2022

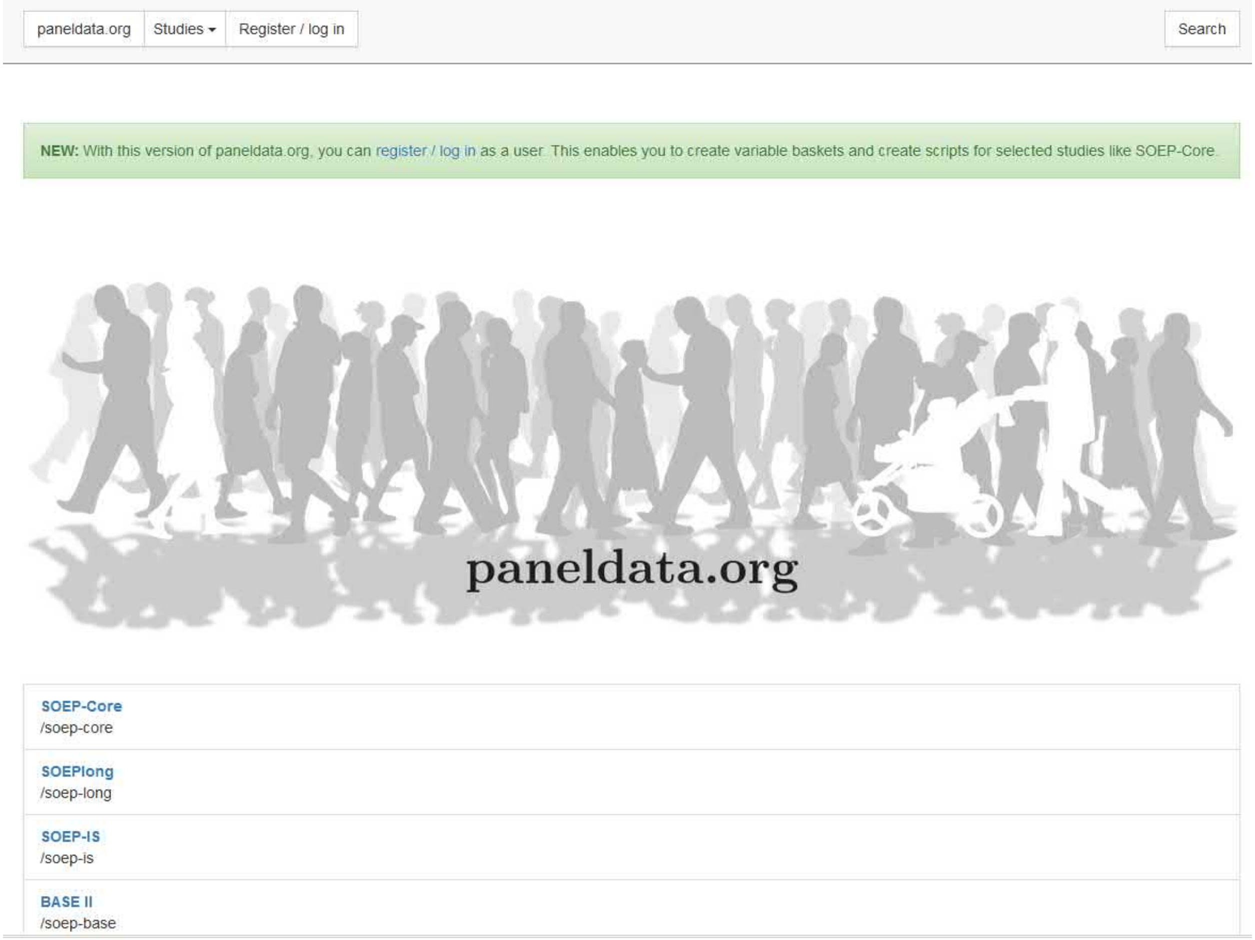
7.2 Variable Search with paneldata.org

Paneldata.org also allows you to search for variables and to find more information about generated variables. It offers comprehensive frequency counts, chronologies of variables, cross-study variable linkage via concepts, a syntax generator, and a topic list for content search in the SOEP.

Example Variable:

bbh5508: Wave “bb” (Survey Year 2011); household questionnaire (“h”), question number 55, item 8

Open Paneldata



NEW: With this version of paneldata.org, you can register / log in as a user. This enables you to create variable baskets and create scripts for selected studies like SOEP-Core.

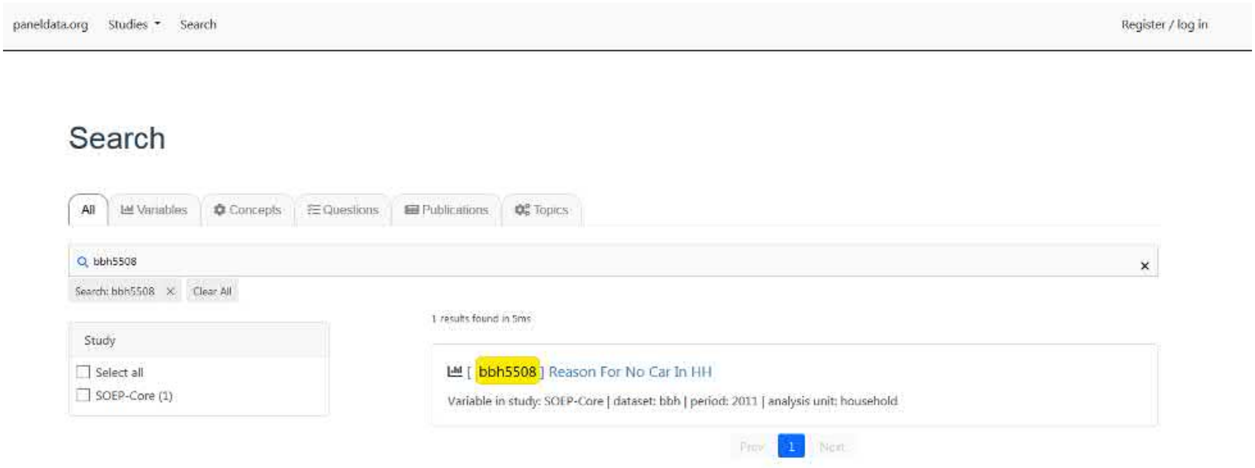
SOEP-Core
/soep-core

SOEPlong
/soep-long

SOEP-IS
/soep-is

BASE II
/soep-base

Select the study SOEP-Core. The SOEP-Core overview contains important general information about the study, e.g., data access, survey method, questionnaires, themes, terms for missing codes, all available datasets in the study and metadata-based questionnaires. To search for a variable, a dataset, or a publication, simply enter the desired search term in the search bar.



Search

bbh5508

Search: bbh5508 ×

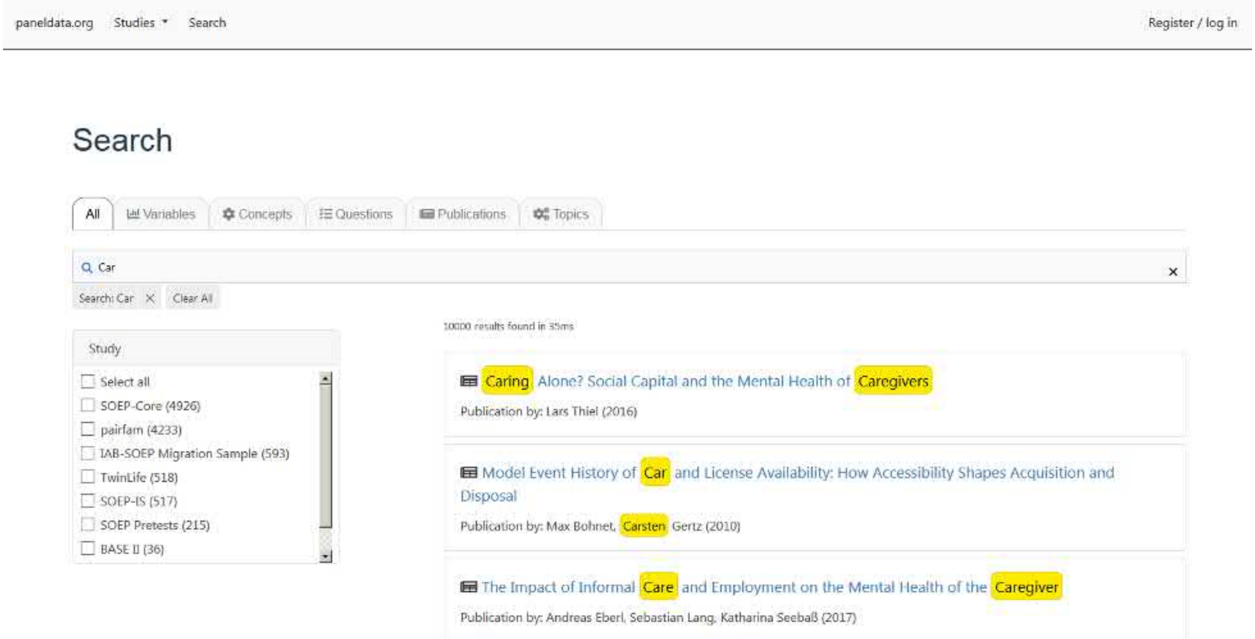
Study

□ Select all
□ SOEP-Core (1)

1 results found in 5ms

[bbh5508] Reason For No Car In HH
Variable in study: SOEP-Core | dataset: bbh | period: 2011 | analysis unit: household

To obtain the desired results, you will need to input specific information. The results window displays all search results. You will see that the variable "bbh5508" originates from SOEP-Core data and can be found in the dataset "bbh" (survey year 2011). If your search is not so specific, you can also search by keywords. We are still interested in the topic "car".



Search

Car

Search: Car ×

Study

□ Select all
□ SOEP-Core (4926)
□ pairfam (4233)
□ IAB-SOEP Migration Sample (593)
□ TwinLife (518)
□ SOEP-IS (517)
□ SOEP Pretests (215)
□ BASE II (36)

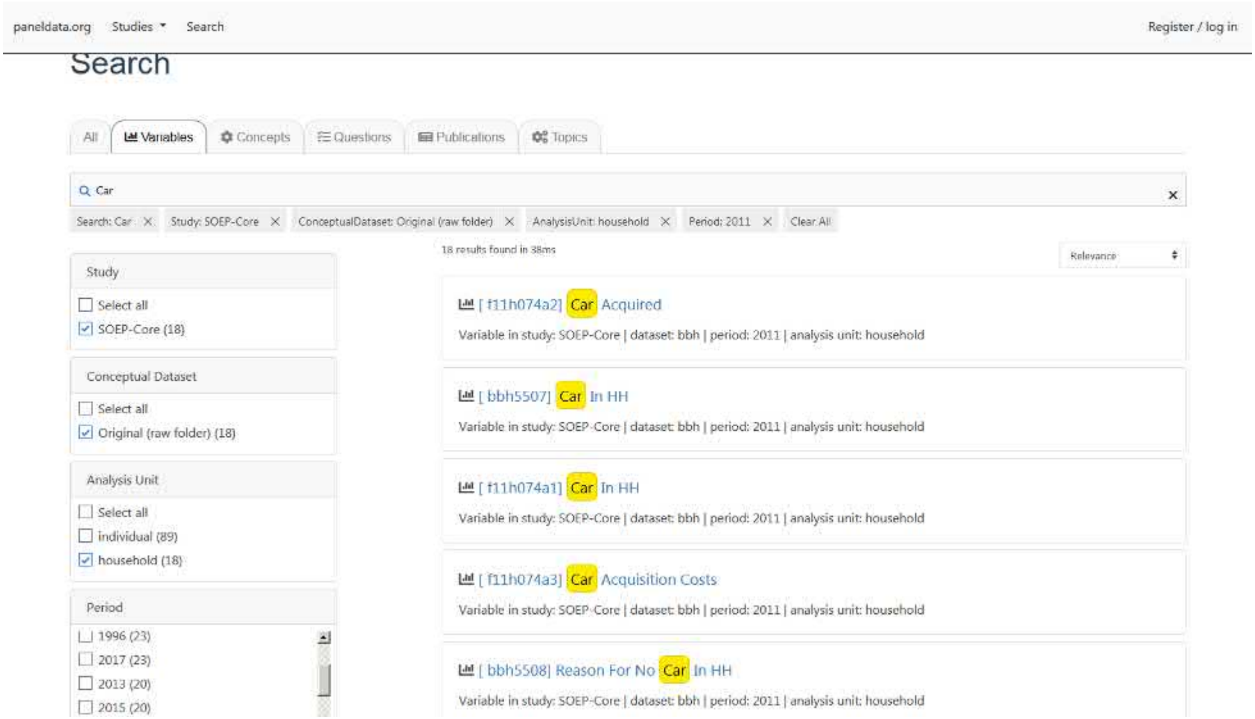
10000 results found in 35ms

Caring Alone? Social Capital and the Mental Health of Caregivers
Publication by: Lars Thiel (2016)

Model Event History of Car and License Availability: How Accessibility Shapes Acquisition and Disposal
Publication by: Max Bohnet, Carsten Gertz (2010)

The Impact of Informal Care and Employment on the Mental Health of the Caregiver
Publication by: Andreas Eberl, Sebastian Lang, Katharina Seebaß (2017)

To better limit the 10000 results, the filter options on the left and on the top should be used. We are looking for variables from the "SOEP-Core" datasets. The search results should be limited with the filter options. Which survey years are of interest to me, do I want to work with original data or generated data? For more information about the different datasets in SOEP-Core, see the section *Data Distribution File*. Should the variable I am looking for be at household level or at individual level?



Search

All | Variables | Concepts | Questions | Publications | Topics

Car ×

Search: Car × Study: SOEP-Core × ConceptualDataset: Original (raw folder) × AnalysisUnit: household × Period: 2011 × Clear All

18 results found in 38ms

Relevance

Study
□ Select all
☑ SOEP-Core (18)

Conceptual Dataset
□ Select all
☑ Original (raw folder) (18)

Analysis Unit
□ Select all
□ individual (89)
☑ household (18)

Period
□ 1996 (23)
□ 2017 (23)
□ 2013 (20)
□ 2015 (20)

[f11h074a2] Car Acquired
Variable in study: SOEP-Core | dataset: bbh | period: 2011 | analysis unit: household

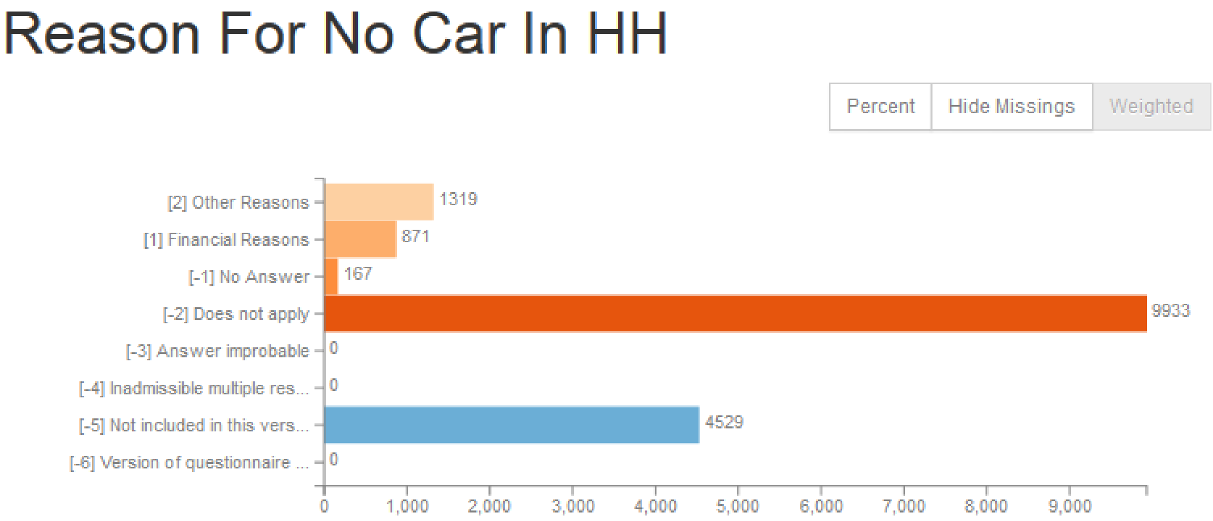
[bbh5507] Car In HH
Variable in study: SOEP-Core | dataset: bbh | period: 2011 | analysis unit: household

[f11h074a1] Car In HH
Variable in study: SOEP-Core | dataset: bbh | period: 2011 | analysis unit: household

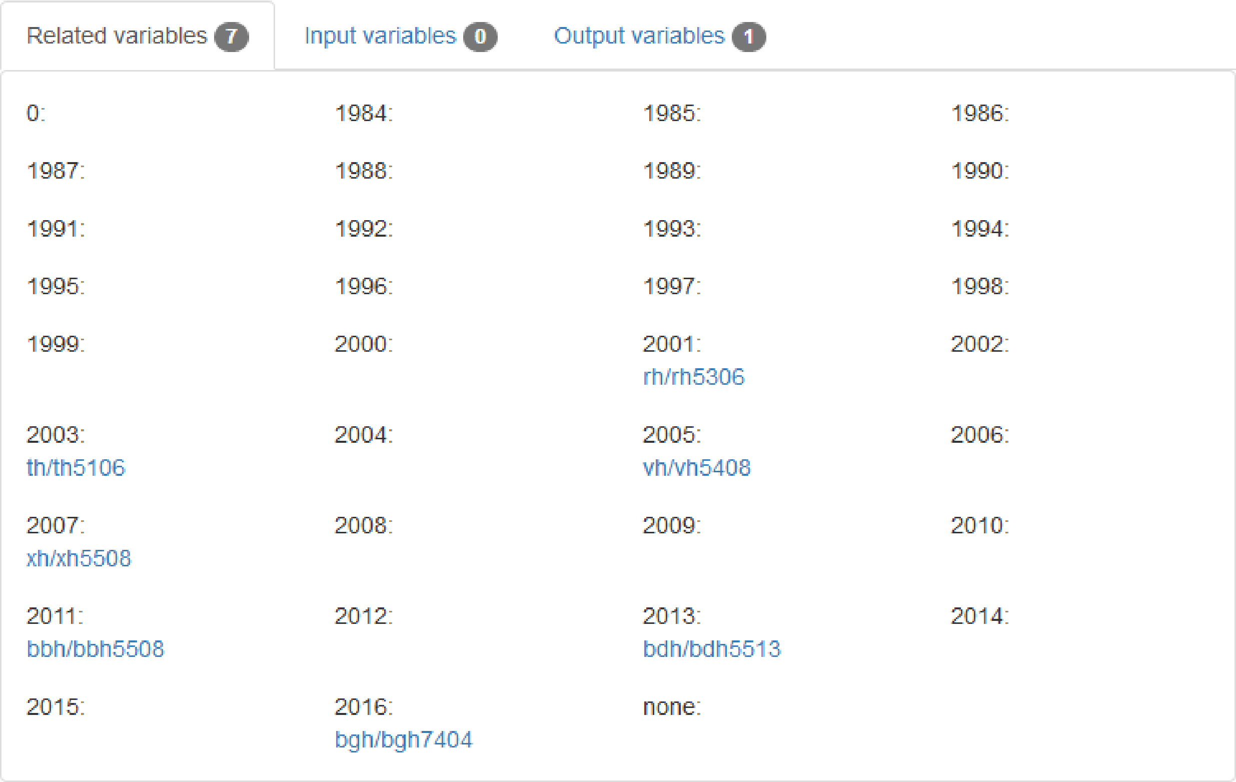
[f11h074a3] Car Acquisition Costs
Variable in study: SOEP-Core | dataset: bbh | period: 2011 | analysis unit: household

[bbh5508] Reason For No Car In HH
Variable in study: SOEP-Core | dataset: bbh | period: 2011 | analysis unit: household

By filtering, the search result is limited to 18 hits, which also shows the variable we searched for. If you click on the variable “bbh5508”, you will find additional information about the variable.

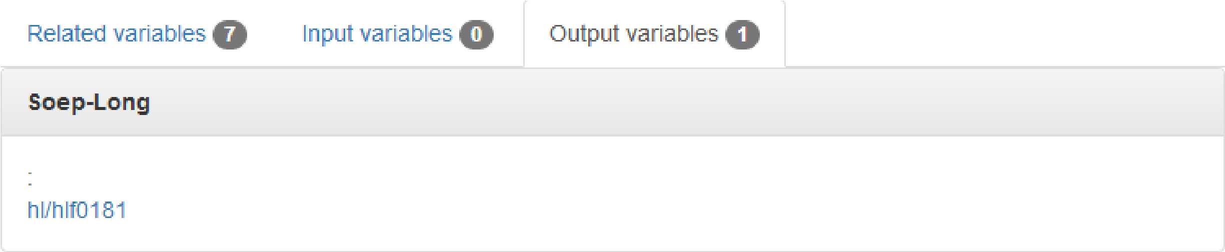


First you see the weighted absolute frequencies for the variable. It is possible to remove the missing codes from the analysis and/or to display the relative frequencies. Even without opening the dataset, paneldata obgain a good overview of the frequencies of a variable.



0:	1984:	1985:	1986:
1987:	1988:	1989:	1990:
1991:	1992:	1993:	1994:
1995:	1996:	1997:	1998:
1999:	2000:	2001: rh/rh5306	2002:
2003: th/th5106	2004:	2005: vh/vh5408	2006:
2007: xh/xh5508	2008:	2009:	2010:
2011: bbh/bbh5508	2012:	2013: bdh/bdh5513	2014:
2015:	2016: bgh/bgh7404	none:	

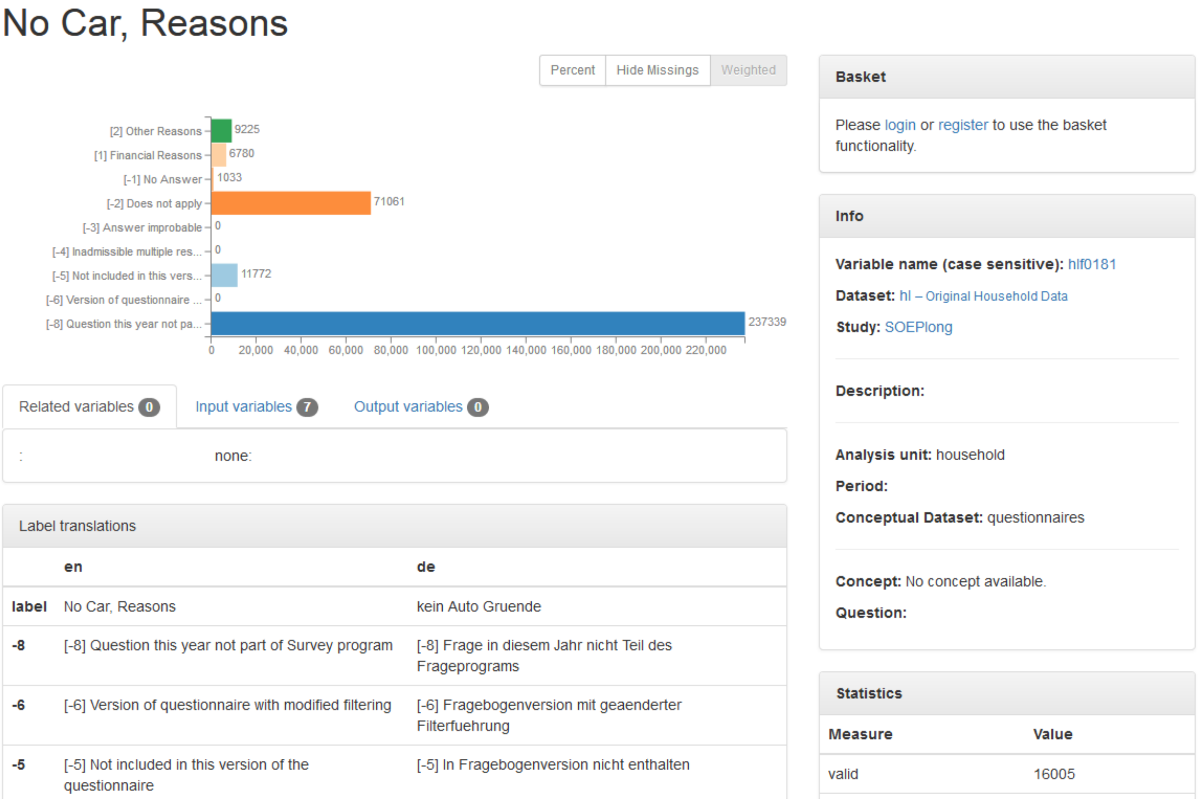
In the Related Variables section you will also find the chronology of the variable you are looking for. The sample variable was collected in 2001, 2003, 2005, 2007, 2011, 2013. Below the survey year, the name of the variable in the respective year is displayed and can be clicked to access the respective variable page. You can see at a glance when the variable was measured, how often it was measured, and what its name is in the respective survey year.



Soep-Long

:
hl/hlf0181

In addition, by clicking on "Output variables", you will find a variable forwarding you to the variable in "long" format. For a more detailed understanding of the long format, read the section *Data Structure in "Long" Format (long)*.



No Car, Reasons

: none:

Label translations

	en	de
label	No Car, Reasons	kein Auto Gruende
-8	[-8] Question this year not part of Survey program	[-8] Frage in diesem Jahr nicht Teil des Frageprograms
-6	[-6] Version of questionnaire with modified filtering	[-6] Fragebogenversion mit geaenderter Filterfuehrung
-5	[-5] Not included in this version of the questionnaire	[-5] In Fragebogenversion nicht enthalten

Basket

Please login or register to use the basket functionality.

Info

Variable name (case sensitive): hlf0181

Dataset: hl – Original Household Data

Study: SOEPlong

Description:

Analysis unit: household

Period:

Conceptual Dataset: questionnaires

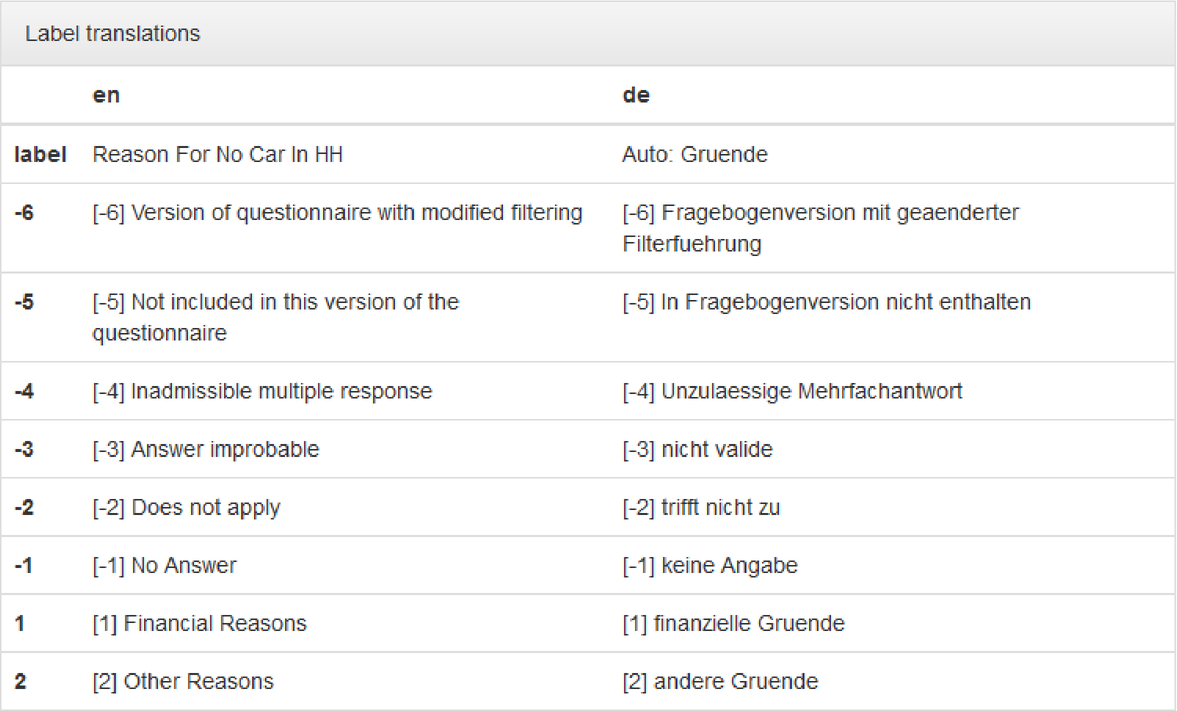
Concept: No concept available.

Question:

Statistics

Measure	Value
valid	16005

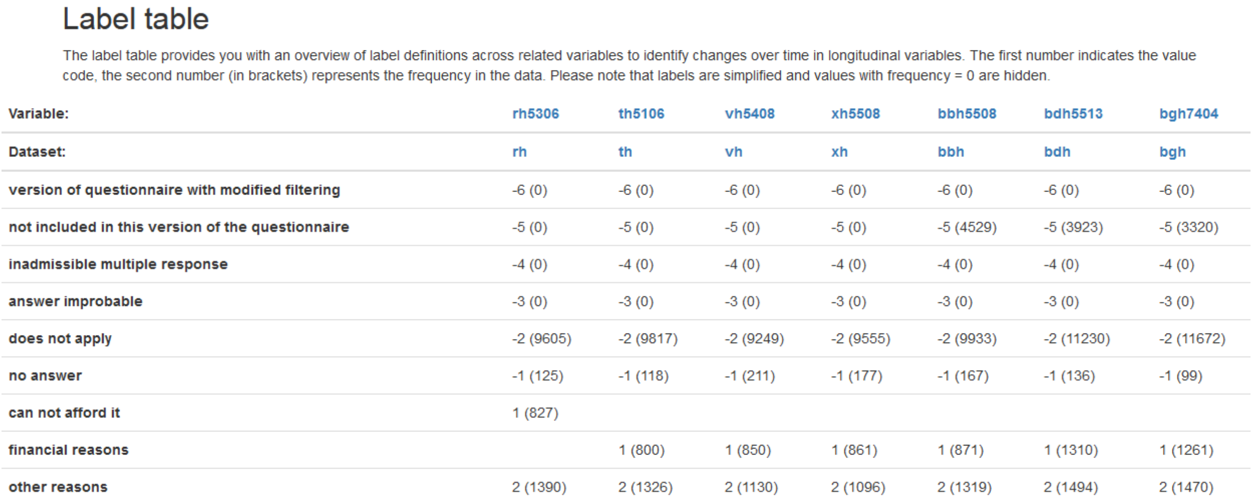
As soon as you click on the "long" variable, you will get to the variable overview for this variable in the long format. The overview of variables does not differ. It can be seen that our example variable "bbh5508" can also be found in long format in the dataset "hl" with the variable "hlf0181".



Label translations

	en	de
label	Reason For No Car In HH	Auto: Gruende
-6	[-6] Version of questionnaire with modified filtering	[-6] Fragebogenversion mit geaenderter Filterfuehrung
-5	[-5] Not included in this version of the questionnaire	[-5] In Fragebogenversion nicht enthalten
-4	[-4] Inadmissible multiple response	[-4] Unzulaessige Mehrfachantwort
-3	[-3] Answer improbable	[-3] nicht valide
-2	[-2] Does not apply	[-2] trifft nicht zu
-1	[-1] No Answer	[-1] keine Angabe
1	[1] Financial Reasons	[1] finanzielle Gruende
2	[2] Other Reasons	[2] andere Gruende

The field “Label translations” shows the value labels of the variables in German and English. In addition, all missing codes used in SOEP are listed and explained.



Label table

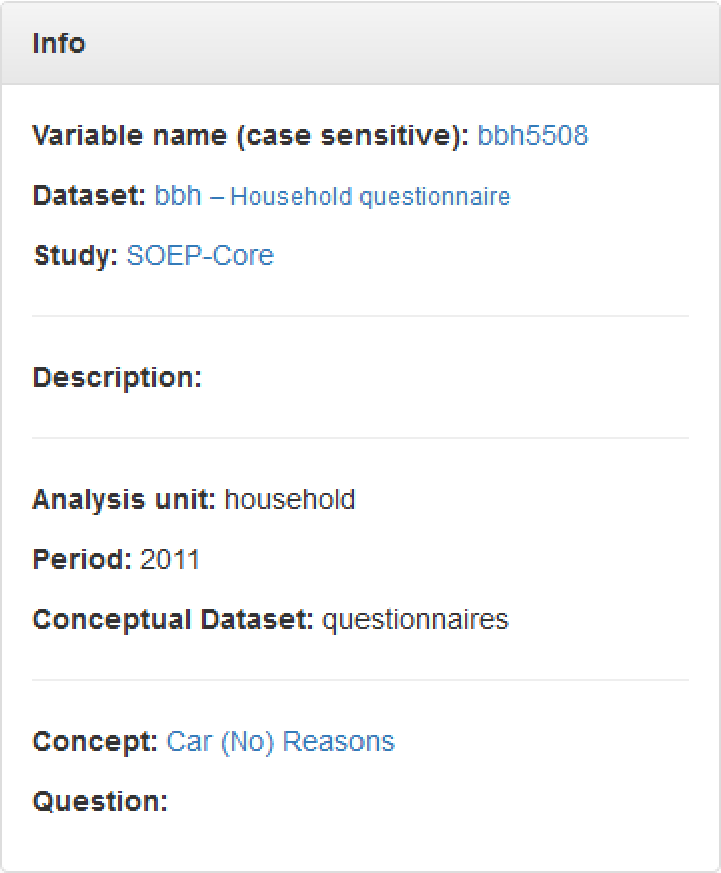
The label table provides you with an overview of label definitions across related variables to identify changes over time in longitudinal variables. The first number indicates the value code, the second number (in brackets) represents the frequency in the data. Please note that labels are simplified and values with frequency = 0 are hidden.

Variable:	rh5306	th5106	vh5408	xh5508	bbh5508	bdh5513	bgh7404
Dataset:	rh	th	vh	xh	bbh	bdh	bgh
version of questionnaire with modified filtering	-6 (0)	-6 (0)	-6 (0)	-6 (0)	-6 (0)	-6 (0)	-6 (0)
not included in this version of the questionnaire	-5 (0)	-5 (0)	-5 (0)	-5 (0)	-5 (4529)	-5 (3923)	-5 (3320)
inadmissible multiple response	-4 (0)	-4 (0)	-4 (0)	-4 (0)	-4 (0)	-4 (0)	-4 (0)
answer improbable	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)
does not apply	-2 (9605)	-2 (9817)	-2 (9249)	-2 (9555)	-2 (9933)	-2 (11230)	-2 (11672)
no answer	-1 (125)	-1 (118)	-1 (211)	-1 (177)	-1 (167)	-1 (136)	-1 (99)
can not afford it	1 (827)						
financial reasons		1 (800)	1 (850)	1 (861)	1 (871)	1 (1310)	1 (1261)
other reasons	2 (1390)	2 (1326)	2 (1130)	2 (1096)	2 (1319)	2 (1494)	2 (1470)

The Label window shows you the absolute frequencies of the variable at different times of data collection. This makes it possible to identify initial trends in how response behavior has changed over a period of time. The assigned value code is output for each possible characteristic value and the absolute frequencies are displayed in parentheses.

In our example, we see that for the variable “th5106” 800 respondents in the wave “t” (2003) stated “financial reasons” as the reason for not having a car in the household. For our variable “bbh5508” in survey year 2011 (wave “bb”), there were 871 respondents.

Paneldata.org is an excellent way to get an first overview of certain variables.



Info

Variable name (case sensitive): bbh5508

Dataset: bbh – Household questionnaire

Study: SOEP-Core

Description:

Analysis unit: household

Period: 2011

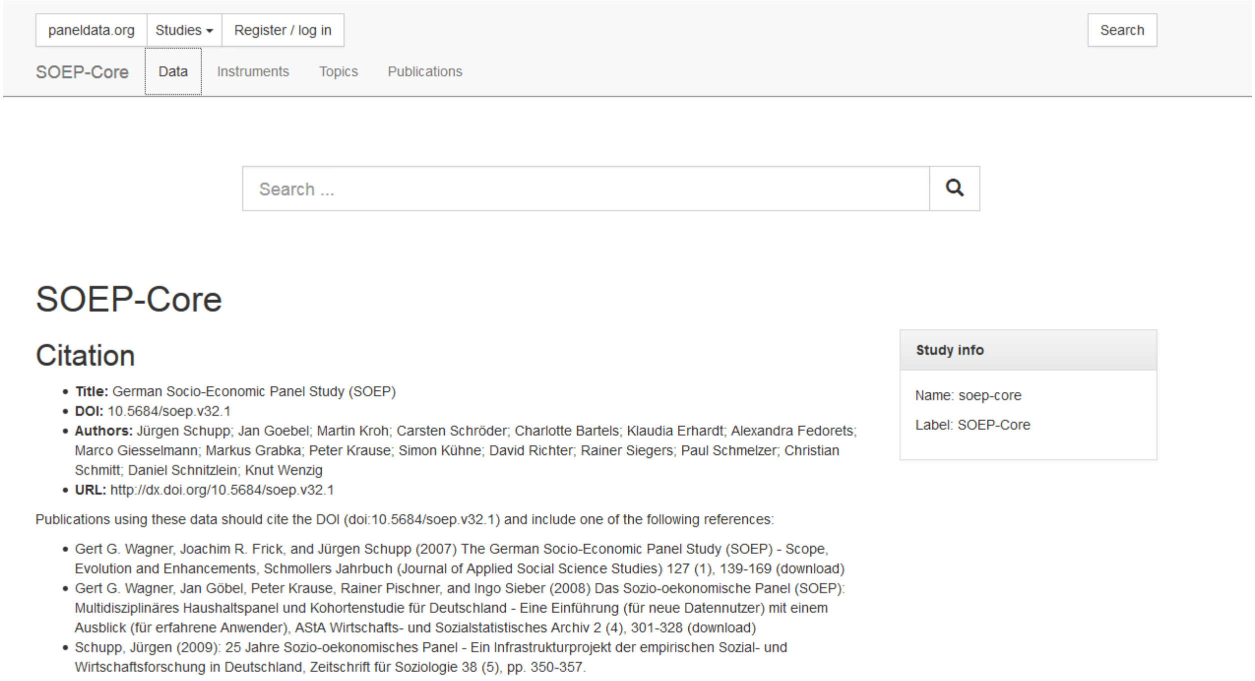
Conceptual Dataset: questionnaires

Concept: Car (No) Reasons

Question:

The information box on the right-hand side provides an overview of all relevant information about the variable and the dataset. In addition to basic information, you will find information about what kind of variable you are looking for under "Conceptual Dataset". In our example "bbh5508", you can see that variables with a "Conceptual Dataset: org/net" describe original variables that are assigned to a questionnaire. Generated variables are "Conceptual Dataset: gen". For an overview of the different dataset types in SOEP-Core, see the section *Data Distribution File*.

In addition to searching for keywords or using the various filter settings, you can also find what you are looking for directly in the data set search. Open paneldata.org, click on the study SOEP-Core and select the menu option "data".



SOEP-Core

Citation

- **Title:** German Socio-Economic Panel Study (SOEP)
- **DOI:** 10.5684/soep.v32.1
- **Authors:** Jürgen Schupp; Jan Goebel; Martin Kroh; Carsten Schröder; Charlotte Bartels; Klaudia Erhardt; Alexandra Fedorets; Marco Giesselmann; Markus Grabka; Peter Krause; Simon Kühne; David Richter; Rainer Siegers; Paul Schmelzer; Christian Schmitt; Daniel Schnitzlein; Knut Wenzig
- **URL:** http://dx.doi.org/10.5684/soep.v32.1

Publications using these data should cite the DOI (doi:10.5684/soep.v32.1) and include one of the following references:

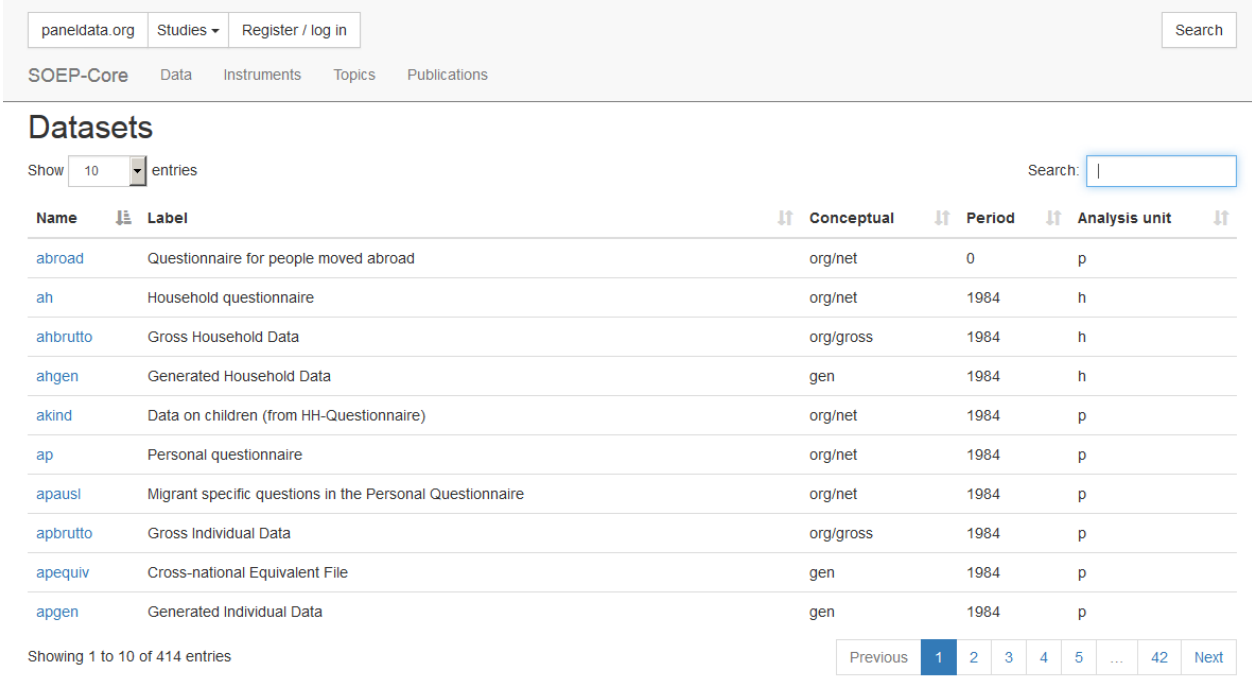
- Gert G. Wagner, Joachim R. Frick, and Jürgen Schupp (2007) The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements, Schmollers Jahrbuch (Journal of Applied Social Science Studies) 127 (1), 139-169 (download)
- Gert G. Wagner, Jan Göbel, Peter Krause, Rainer Pischner, and Ingo Sieber (2008) Das Sozio-oekonomische Panel (SOEP): Multidisziplinäres Haushaltspanel und Kohortenstudie für Deutschland - Eine Einführung (für neue Datennutzer) mit einem Ausblick (für erfahrene Anwender), AStA Wirtschafts- und Sozialstatistisches Archiv 2 (4), 301-328 (download)
- Schupp, Jürgen (2009): 25 Jahre Sozio-oekonomisches Panel - Ein Infrastrukturprojekt der empirischen Sozial- und Wirtschaftsforschung in Deutschland, Zeitschrift für Soziologie 38 (5), pp. 350-357.

Study info

Name: soep-core

Label: SOEP-Core

Now you will be directed to an overview that shows you all datasets contained in SOEP-Core.



Datasets

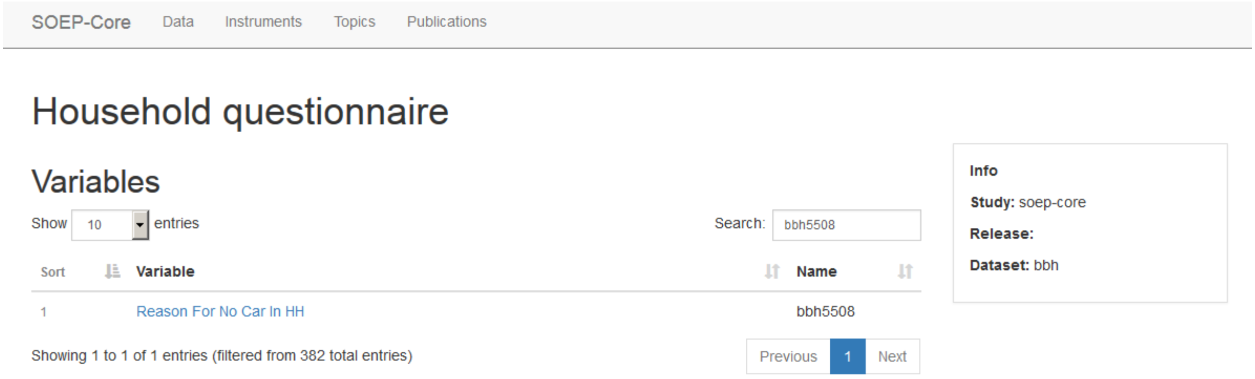
Show 10 entries

Search:

Name	Label	Conceptual	Period	Analysis unit
abroad	Questionnaire for people moved abroad	org/net	0	p
ah	Household questionnaire	org/net	1984	h
ahbrutto	Gross Household Data	org/gross	1984	h
ahgen	Generated Household Data	gen	1984	h
akind	Data on children (from HH-Questionnaire)	org/net	1984	p
ap	Personal questionnaire	org/net	1984	p
apausl	Migrant specific questions in the Personal Questionnaire	org/net	1984	p
apbrutto	Gross Individual Data	org/gross	1984	p
apequiv	Cross-national Equivalent File	gen	1984	p
apgen	Generated Individual Data	gen	1984	p

Showing 1 to 10 of 414 entries

Enter the dataset you are looking for (“bbh”) in the search bar at the top right and click on the dataset. You will be directed to an overview that shows you all variables from the “bbh” dataset.



SOEP-Core Data Instruments Topics Publications

Household questionnaire

Variables

Show 10 entries

Search: bbh5508

Sort	Variable	Name
1	Reason For No Car In HH	bbh5508

Showing 1 to 1 of 1 entries (filtered from 382 total entries)

Previous 1 Next

Info

Study: soep-core

Release:

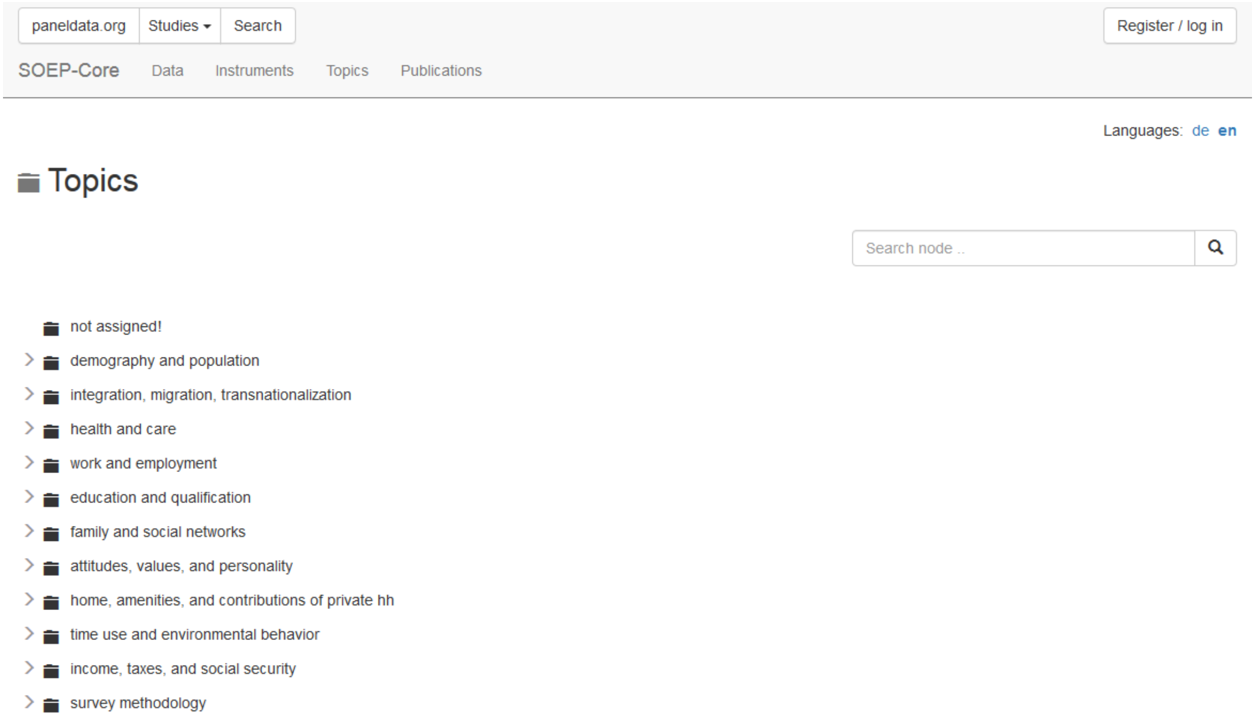
Dataset: bbh

Now enter the variable you are looking for in the search bar at the top right and click on the variable of interest. You will be directed to the variable overview, where you will find detailed information on the variable. Paneldata.org offers a variety of search options to fit the user's search needs.

7.3 Topic Search with paneldata.org

To provide an overview of the various topics in the SOEP, the variables have been grouped together on paneldata.org by topic. If you are looking for your research variables and do not want to check all datasets or questionnaires, the topic search on paneldata.org may help.

Open Paneldata and select the main study SOEP-Core. The upper navigation bar leads you to the Topics area. Click on Topics and look at the list of variables.



paneldata.org Studies ▾ Search Register / log in

SOEP-Core Data Instruments Topics Publications

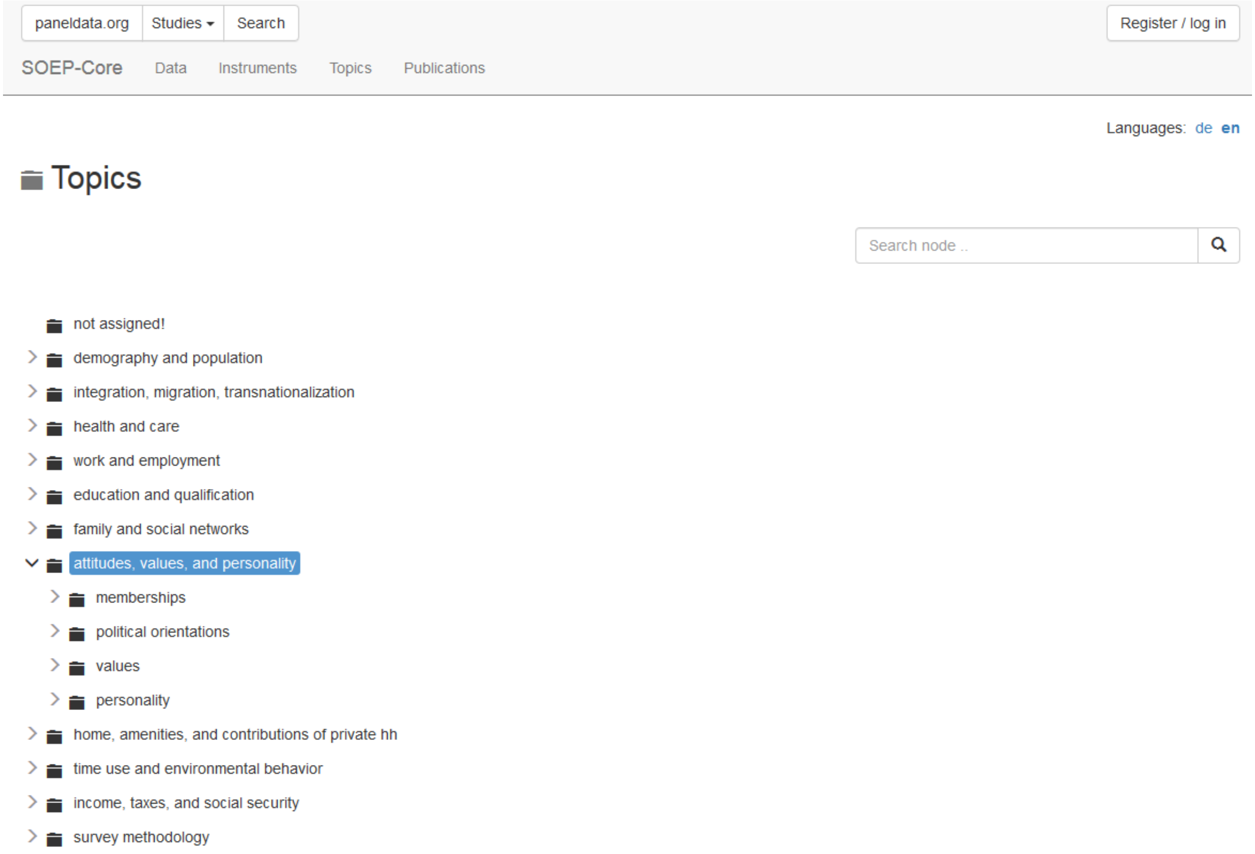
Languages: de **en**

Topics

Search node ..

- not assigned!
- demography and population
- integration, migration, transnationalization
- health and care
- work and employment
- education and qualification
- family and social networks
- attitudes, values, and personality
- home, amenities, and contributions of private hh
- time use and environmental behavior
- income, taxes, and social security
- survey methodology

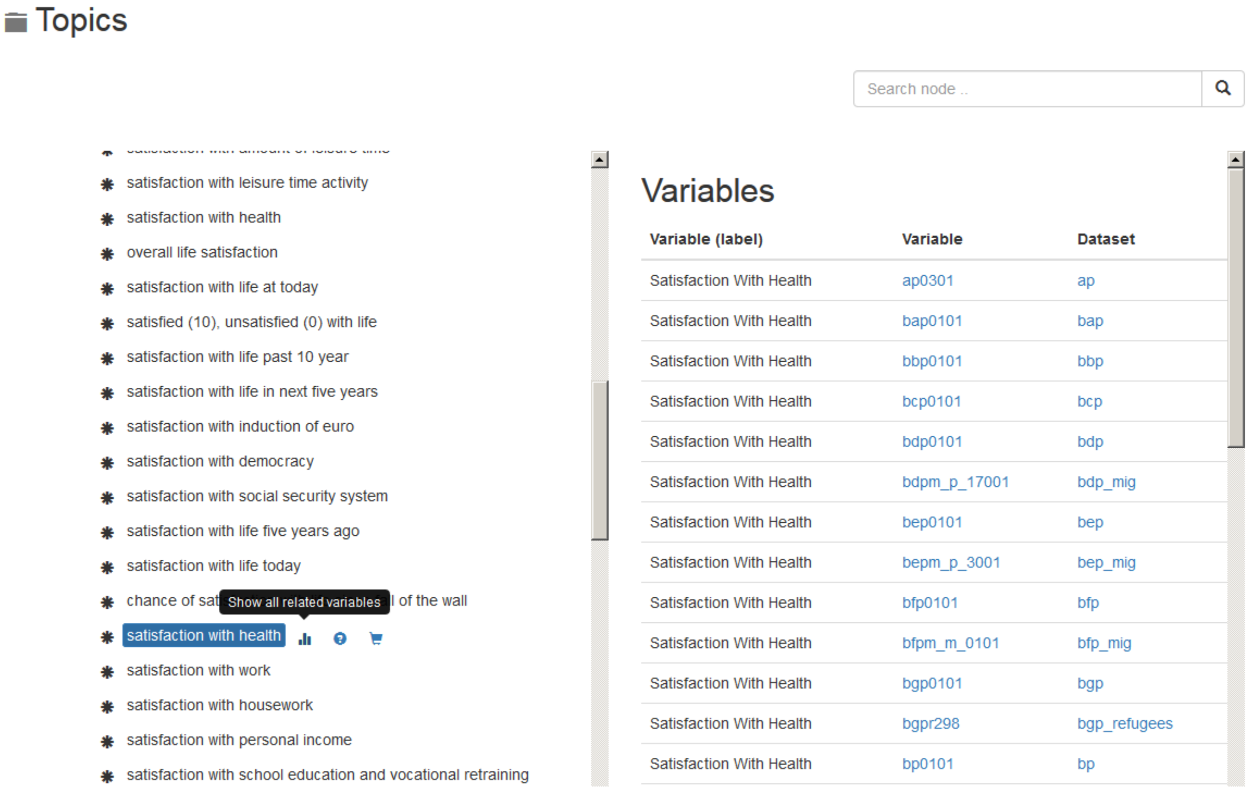
Select a topic that corresponds to your research interest, and a more detailed list of sub-topics will appear under the main topic heading.



Topics

- not assigned!
- demography and population
- integration, migration, transnationalization
- health and care
- work and employment
- education and qualification
- family and social networks
- attitudes, values, and personality
  - memberships
  - political orientations
  - values
  - personality
- home, amenities, and contributions of private hh
- time use and environmental behavior
- income, taxes, and social security
- survey methodology

For example, if you are interested in different types of satisfaction, click on the topic "attitudes, values, and personality". Underneath it, you will find the sub-topic "personality". Suppose you are interested in health satisfaction. If you have found a suitable sub-topic, click on "show all the related variables". All variables that fall under this topic will be displayed.



Topics

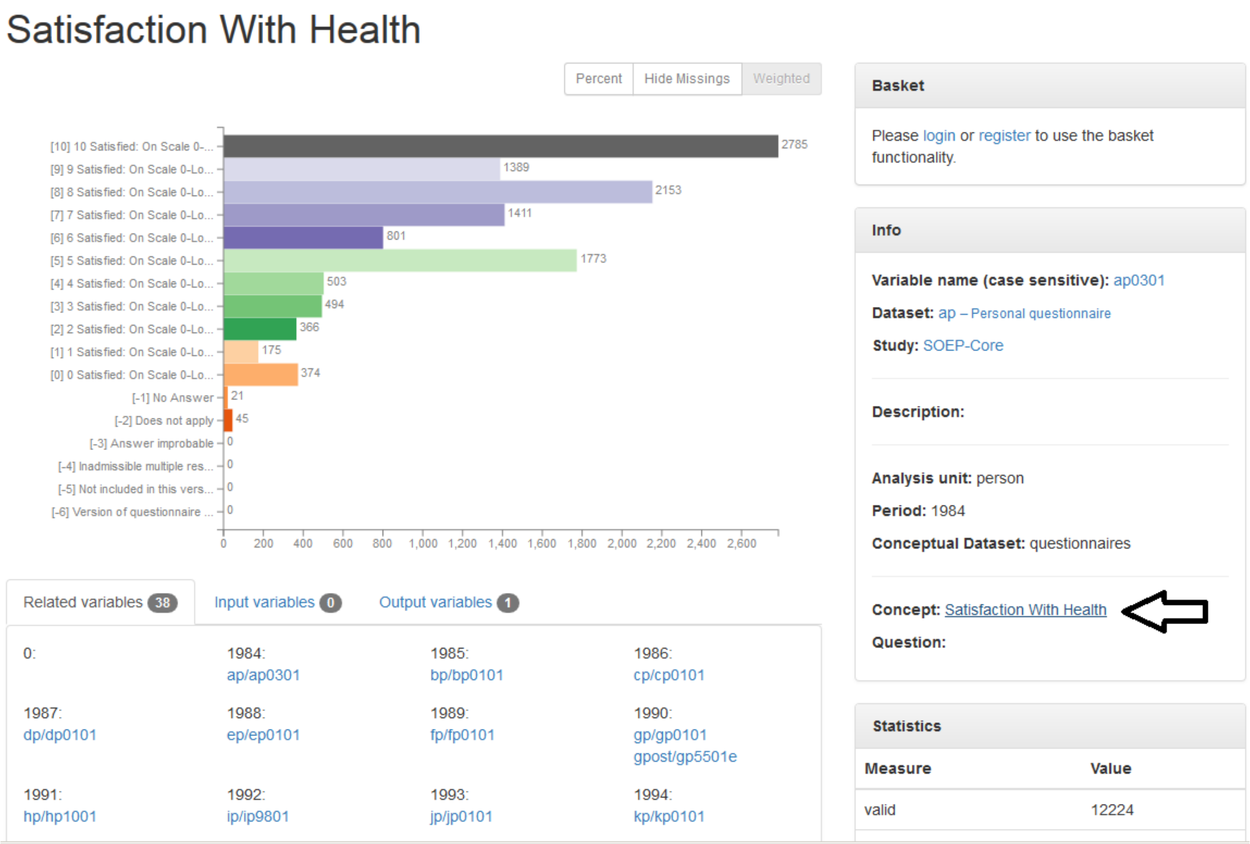
Search node ..

- satisfaction with amount of leisure time
- satisfaction with leisure time activity
- satisfaction with health
- overall life satisfaction
- satisfaction with life at today
- satisfied (10), unsatisfied (0) with life
- satisfaction with life past 10 year
- satisfaction with life in next five years
- satisfaction with induction of euro
- satisfaction with democracy
- satisfaction with social security system
- satisfaction with life five years ago
- satisfaction with life today
- chance of sat Show all related variables l of the wall
- satisfaction with health
- satisfaction with work
- satisfaction with housework
- satisfaction with personal income
- satisfaction with school education and vocational retraining

Variables

Variable (label)	Variable	Dataset
Satisfaction With Health	ap0301	ap
Satisfaction With Health	bap0101	bap
Satisfaction With Health	bbp0101	bbp
Satisfaction With Health	bcp0101	bcp
Satisfaction With Health	bdp0101	bdp
Satisfaction With Health	bdpm_p_17001	bdp_mig
Satisfaction With Health	bep0101	bep
Satisfaction With Health	bepm_p_3001	bep_mig
Satisfaction With Health	bfp0101	bfp
Satisfaction With Health	bfpm_m_0101	bfp_mig
Satisfaction With Health	bgp0101	bgp
Satisfaction With Health	bgpr298	bgp_refugees
Satisfaction With Health	bp0101	bp

The paneldata topic list has three possible functions for each sub-topic. You can display all variables that belong to a sub-topic. In the future, paneldata will also display the texts of the questions from the SOEP questionnaires in which the variables in that sub-topic appear. Paneldata also allows you to keep variables from a sub-topic in a variable basket. The chapter *Syntax Generator on paneldata.org* explains in detail how to use the basket in your research and what possibilities this offers. Click on one of the variables to see the variable overview.


Percent
Hide Missings
Weighted
[10] 10 Satisfied: On Scale 0-...
[9] 9 Satisfied: On Scale 0-Lo...
[8] 8 Satisfied: On Scale 0-Lo...
[7] 7 Satisfied: On Scale 0-Lo...
[6] 6 Satisfied: On Scale 0-Lo...
[5] 5 Satisfied: On Scale 0-Lo...
[4] 4 Satisfied: On Scale 0-Lo...
[3] 3 Satisfied: On Scale 0-Lo...
[2] 2 Satisfied: On Scale 0-Lo...
[1] 1 Satisfied: On Scale 0-Lo...
[0] 0 Satisfied: On Scale 0-Lo...
[-1] No Answer
[-2] Does not apply
[-3] Answer improbable
[-4] Inadmissible multiple res...
[-5] Not included in this vers...
[-6] Version of questionnaire ...
2785
1389
2153
1411
801
1773
503
494
366
175
374
21
45
0
0
0
0
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,600
Related variables 38
Input variables 0
Output variables 1

Satisfaction With Health

0:	1984: ap/ap0301	1985: bp/bp0101	1986: cp/cp0101
1987: dp/dp0101	1988: ep/ep0101	1989: fp/fp0101	1990: gp/gp0101 gpost/gp5501e
1991: hp/hp1001	1992: ip/ip9801	1993: jp/jp0101	1994: kp/kp0101

Basket

Please login or register to use the basket functionality.

Info

Variable name (case sensitive): ap0301

Dataset: ap – Personal questionnaire

Study: SOEP-Core

Description:

Analysis unit: person

Period: 1984

Conceptual Dataset: questionnaires

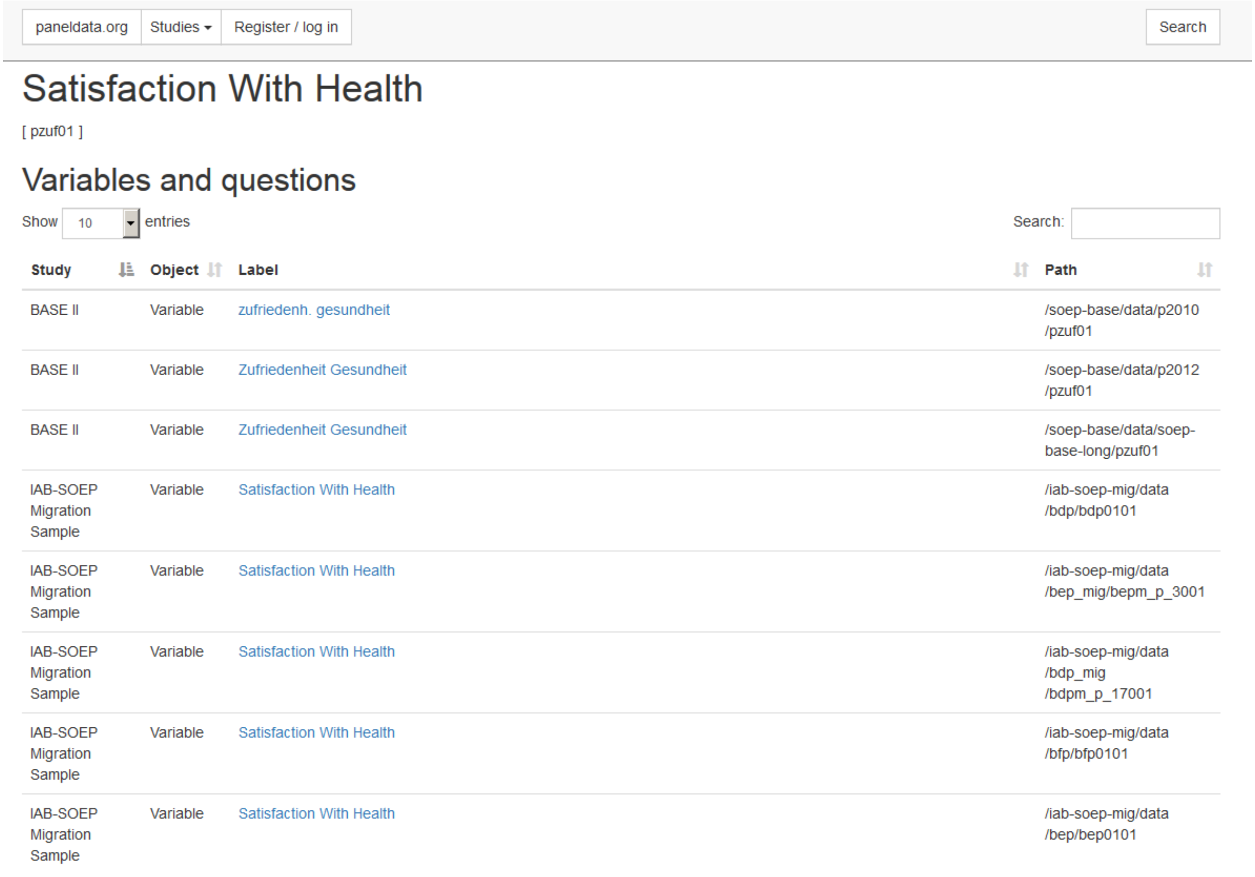
Concept: Satisfaction With Health

Question:

Statistics

Measure	Value
valid	12224

If you click on the concept of a variable, you will get to the concept overview. Concepts in SOEP are used to link variables with the same content. The concepts can even be used to link variables with the same content across studies.



Satisfaction With Health

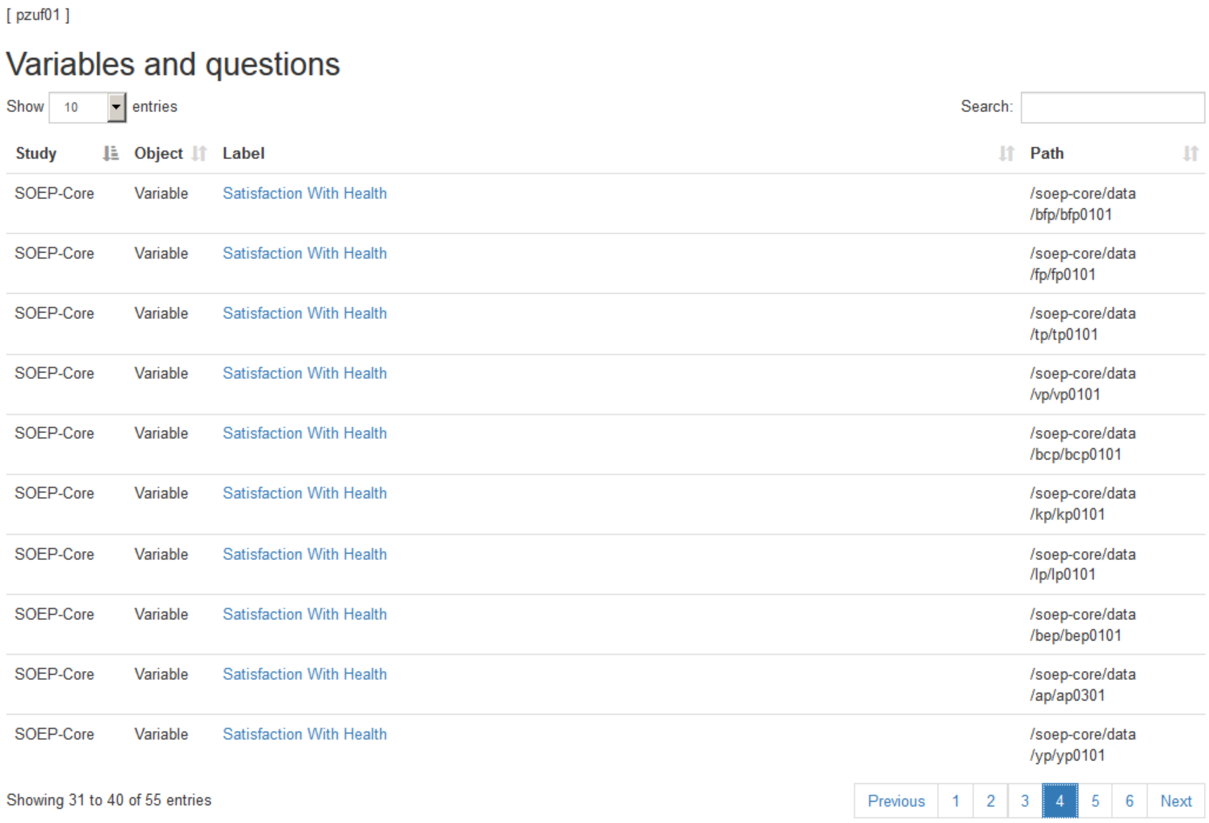
[pzuf01]

Variables and questions

Show 10 entries Search:

Study	Object	Label	Path
BASE II	Variable	zufriedenh. gesundheit	/soep-base/data/p2010 /pzuf01
BASE II	Variable	Zufriedenheit Gesundheit	/soep-base/data/p2012 /pzuf01
BASE II	Variable	Zufriedenheit Gesundheit	/soep-base/data/soep-base-long/pzuf01
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bdp/bdp0101
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bep_mig/bepm_p_3001
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bdp_mig /bdpm_p_17001
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bfp/bfp0101
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bep/bep0101

The concept overview displays the study- and wave-specific variables with this concept. The concept allows you to determine whether the variable you are looking for is also available and comparable across studies. In the column “Study” you can see which studies have the same variable linked by concept. The label of the respective variable is also displayed in the “Label” column. The column “path” shows the wave name of the variable. By clicking on the label, you will get to the overview of variables with all of the relevant information. The “Object” column in the concept overview shows you the type of information displayed.



[pzuf01]

Variables and questions

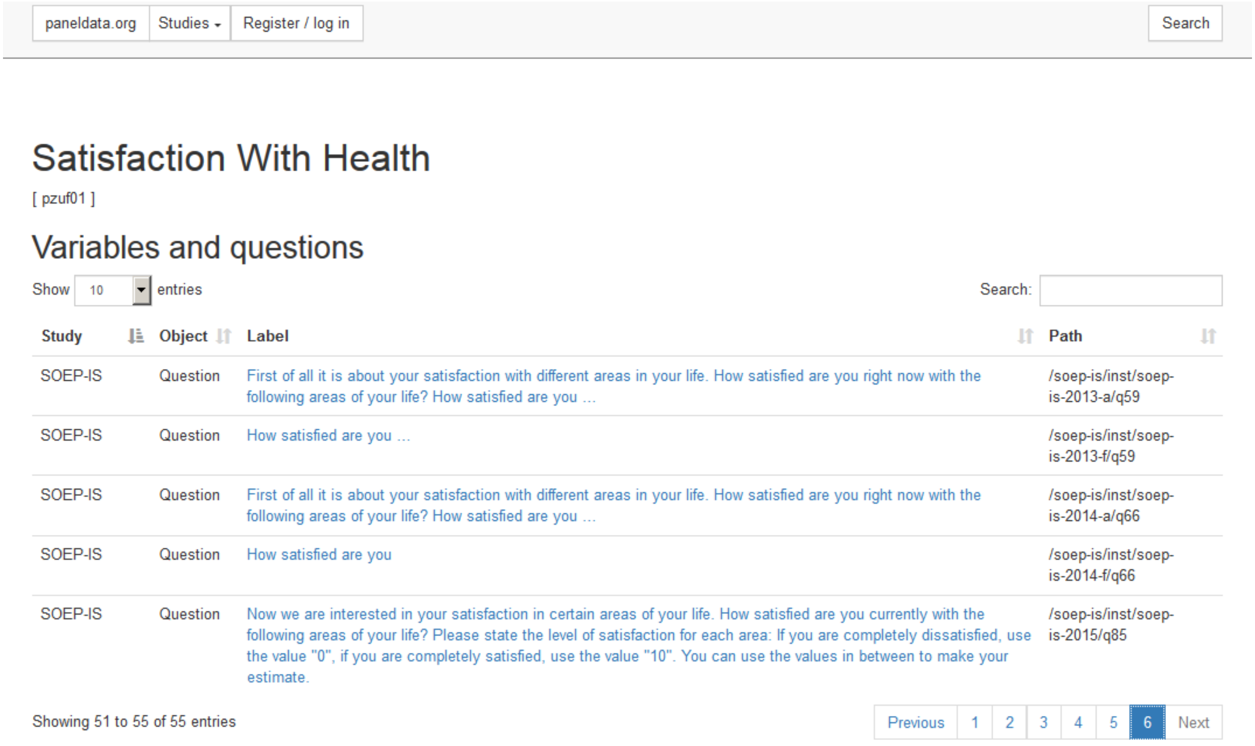
Show 10 entries

Search:

Study	Object	Label	Path
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /bfp/bfp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /fp/fp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /tp/tp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /vp/vp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /bcp/bcp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /kp/kp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /lp/lp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /bep/bep0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /ap/ap0301
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /yp/yp0101

Showing 31 to 40 of 55 entries

In addition to the variables linked by concept, you can find the relevant questions in the concept overview. Questions are displayed in the “Object” column with question. Without having to open the questionnaire, you can read the question and identify possible differences. Click on the desired question and you will be taken to the question display.



Satisfaction With Health

[pzuf01]

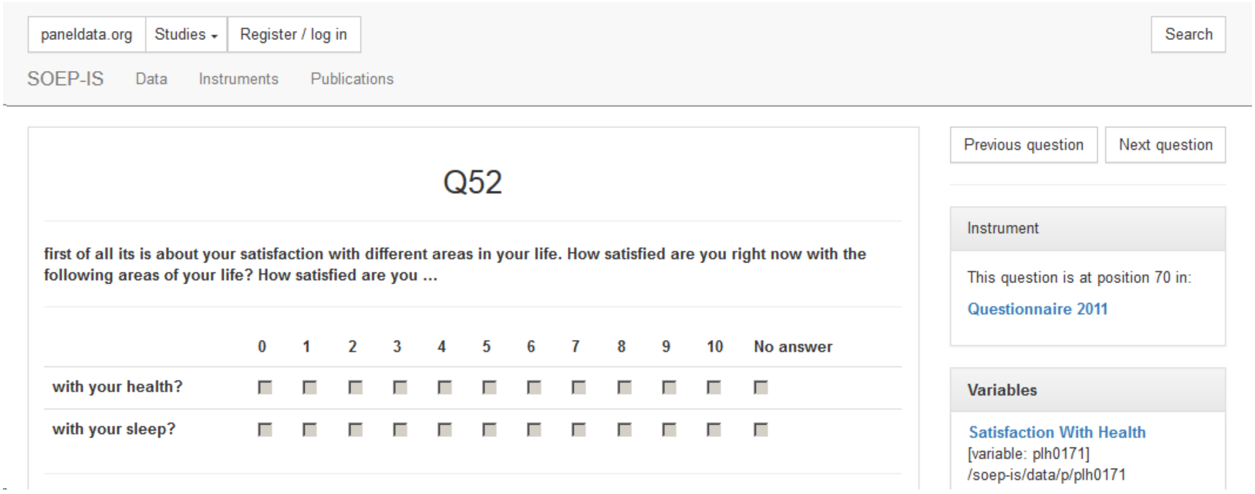
Variables and questions

Show 10 entries

Search:

Study	Object	Label	Path
SOEP-IS	Question	First of all it is about your satisfaction with different areas in your life. How satisfied are you right now with the following areas of your life? How satisfied are you …	/soep-is/inst/soep-is-2013-a/q59
SOEP-IS	Question	How satisfied are you …	/soep-is/inst/soep-is-2013-f/q59
SOEP-IS	Question	First of all it is about your satisfaction with different areas in your life. How satisfied are you right now with the following areas of your life? How satisfied are you …	/soep-is/inst/soep-is-2014-a/q66
SOEP-IS	Question	How satisfied are you	/soep-is/inst/soep-is-2014-f/q66
SOEP-IS	Question	Now we are interested in your satisfaction in certain areas of your life. How satisfied are you currently with the following areas of your life? Please state the level of satisfaction for each area: If you are completely dissatisfied, use the value "0", if you are completely satisfied, use the value "10". You can use the values in between to make your estimate.	/soep-is/inst/soep-is-2015/q85

Showing 51 to 55 of 55 entries



Q52

first of all its is about your satisfaction with different areas in your life. How satisfied are you right now with the following areas of your life? How satisfied are you …

	0	1	2	3	4	5	6	7	8	9	10	No answer
with your health?	□	□	□	□	□	□	□	□	□	□	□	□
with your sleep?	□	□	□	□	□	□	□	□	□	□	□	□

Instrument

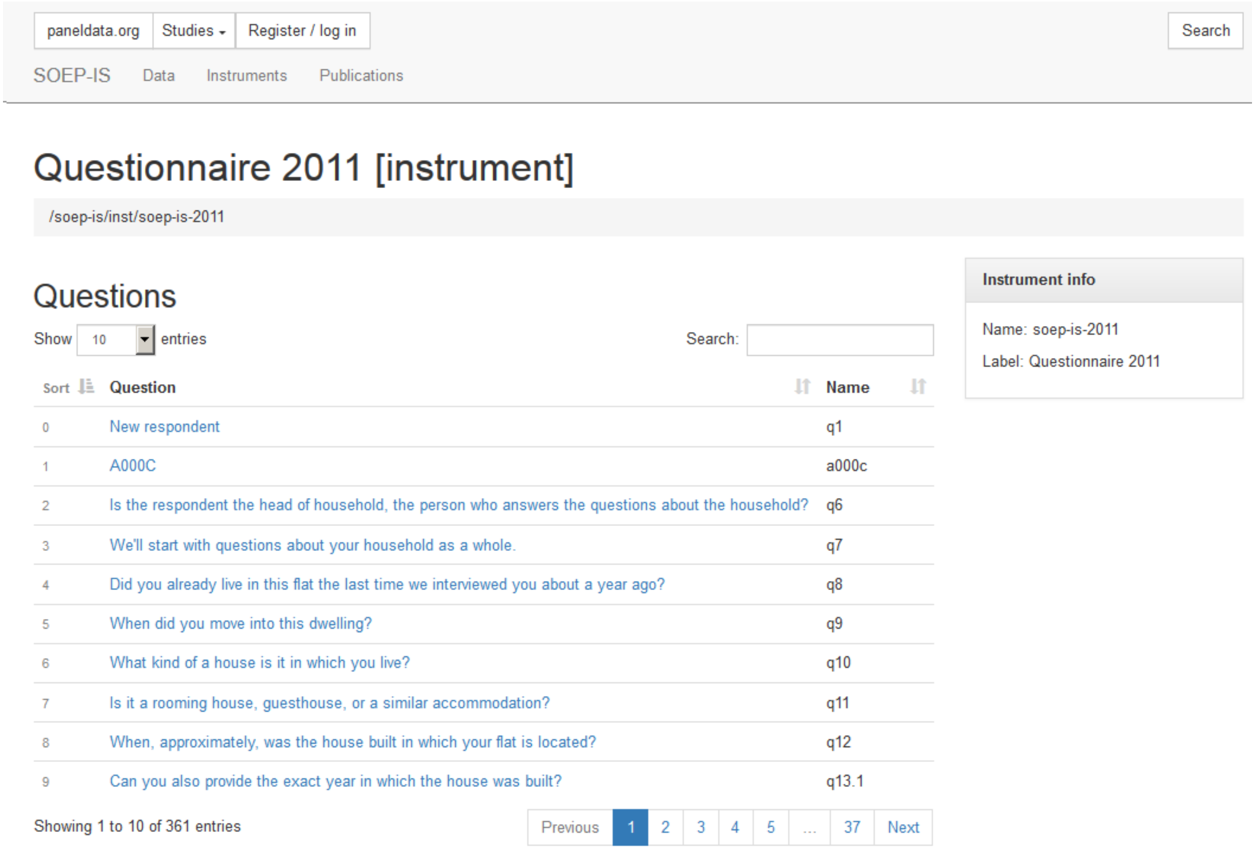
This question is at position 70 in:

Questionnaire 2011

Variables

Satisfaction With Health
[variable: plh0171]
/soep-is/data/p/plh0171

Attention: To find out the exact wording of the question and possible filter structures, a variable search in the questionnaires is necessary. The question display in Paneldata only provides a quick overview. In the question overview, you can navigate through the questionnaire using the “next question” and “previous question” buttons. The “Instrument” section shows the position of the question in the questionnaire, the survey year, and links to the metadata-based survey instrument. Click on the survey instrument “Questionnaire 2011”.



Questionnaire 2011 [instrument]

/soep-is/inst/soep-is-2011

Questions

Show 10 entries

Search:

Sort	Question	Name
0	New respondent	q1
1	A000C	a000c
2	Is the respondent the head of household, the person who answers the questions about the household?	q6
3	We'll start with questions about your household as a whole.	q7
4	Did you already live in this flat the last time we interviewed you about a year ago?	q8
5	When did you move into this dwelling?	q9
6	What kind of a house is it in which you live?	q10
7	Is it a rooming house, guesthouse, or a similar accommodation?	q11
8	When, approximately, was the house built in which your flat is located?	q12
9	Can you also provide the exact year in which the house was built?	q13.1

Showing 1 to 10 of 361 entries

Instrument info

Name: soep-is-2011

Label: Questionnaire 2011

The survey instrument used in the SOEP-IS study in 2011 is now displayed. You can navigate through the questionnaire in this overview. The search bar allows you to search for research-relevant terms. Click on the question to access the question display.

Last change: May 12, 2022

7.4 Documentation on Generated Data

SOEP-Core contains a wide range of generated variables and datasets. To facilitate data use, we generate a large number of variables in the process of data preparation and release them with the SOEP-Core data. To make the generation process transparent to users, we provide comprehensive documentation on the numerous generated datasets and variables. For an overview, see our Documentation on Generated Data

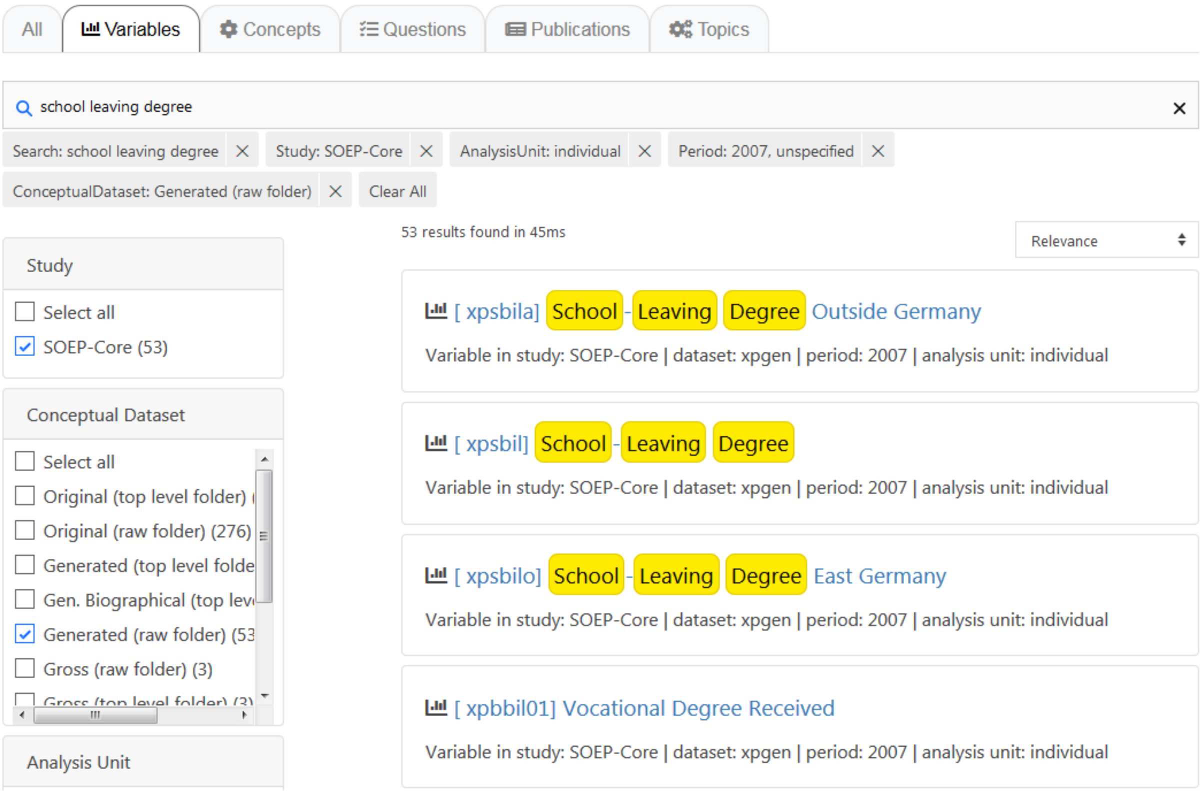
Example: A number of frequently used variables are provided in SOEP as "generated variables" (e.g., the datasets $PGEN and $HGEN). These variables are checked for consistency across waves. The documentation can be used to answer the following questions:

a) Which variable gives the highest school-leaving certificate attained by individuals surveyed in 2007?

To search for the variable that provides this information, open Paneldata , click on the search button and the tab "Variables", then enter "school leaving degree" in the search bar. Specify your search by adjusting the filter settings as follows:

- study: soep-core
- Conceptual dataset: Generated (raw folder)

- analysis unit: individual
- period: 2007



school leaving degree ×

Search: school leaving degree × | Study: SOEP-Core × | AnalysisUnit: individual × | Period: 2007, unspecified ×

ConceptualDataset: Generated (raw folder) × | Clear All

53 results found in 45ms

Relevance

Study

□ Select all
☑ SOEP-Core (53)

Conceptual Dataset

□ Select all
□ Original (top level folder)
□ Original (raw folder) (276)
□ Generated (top level folde
□ Gen. Biographical (top lev
☑ Generated (raw folder) (53
□ Gross (raw folder) (3)
□ Gross (top level folder) (3)

Analysis Unit

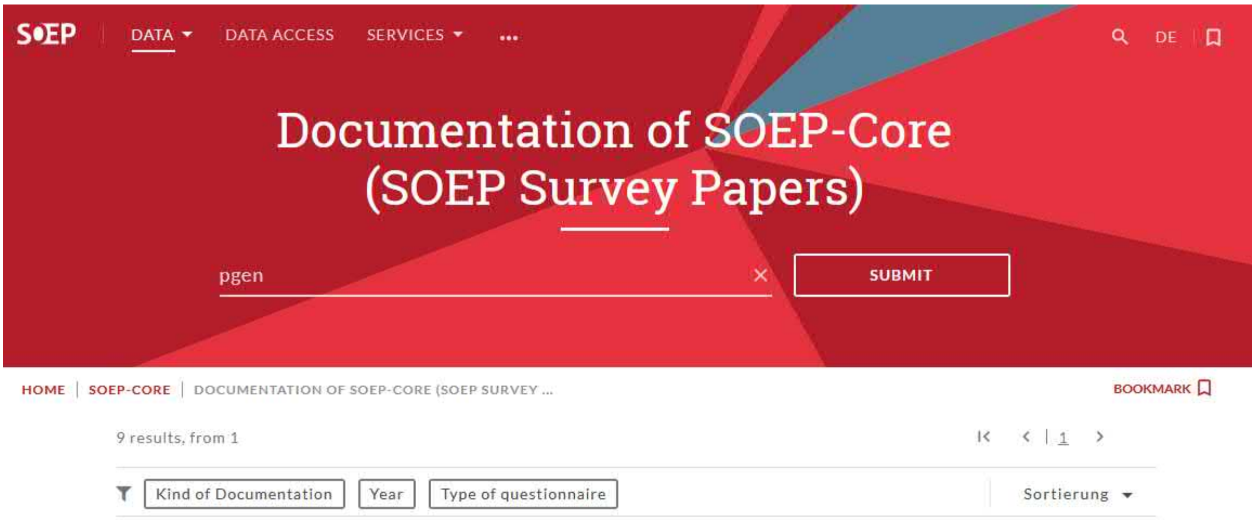
[xpsbila] School - Leaving Degree Outside Germany
Variable in study: SOEP-Core | dataset: xpgen | period: 2007 | analysis unit: individual

[xpsbil] School - Leaving Degree
Variable in study: SOEP-Core | dataset: xpgen | period: 2007 | analysis unit: individual

[xpsbilo] School - Leaving Degree East Germany
Variable in study: SOEP-Core | dataset: xpgen | period: 2007 | analysis unit: individual

[xpbbil01] Vocational Degree Received
Variable in study: SOEP-Core | dataset: xpgen | period: 2007 | analysis unit: individual

All variables could contain the information you are looking for. Since almost all variables in the search result come from the generated "xpgen" dataset, the documentation for the $pgen dataset should be used. Visit the Documentation of SOEP-Core Page and enter the search term pgen in the search field.



Documentation of SOEP-Core (SOEP Survey Papers)

pgen ×

SUBMIT

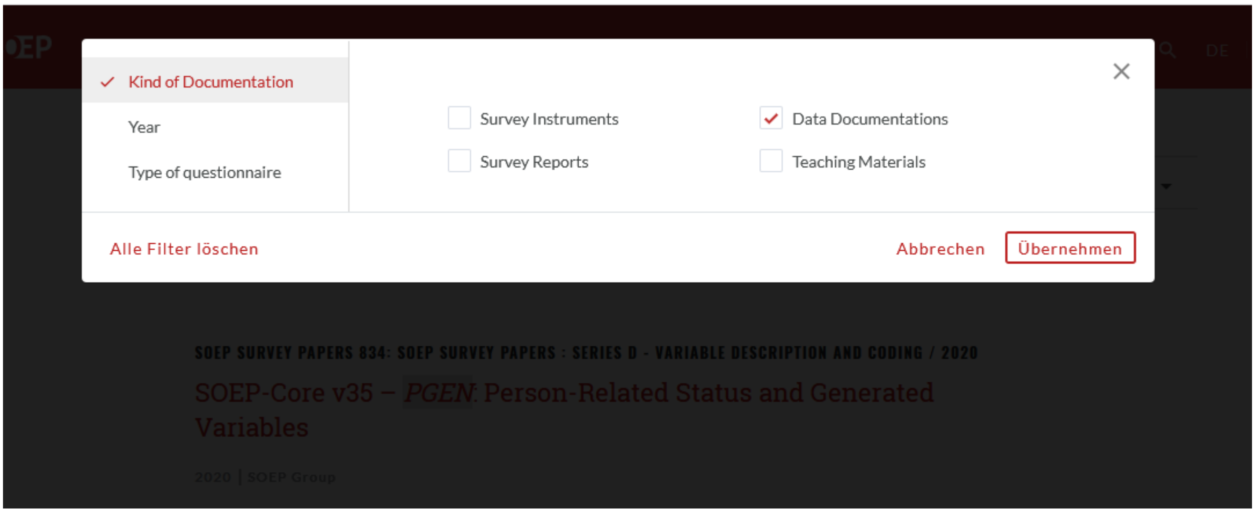
HOME | SOEP-CORE | DOCUMENTATION OF SOEP-CORE (SOEP SURVEY ... BOOKMARK

9 results, from 1

1

Kind of Documentation | Year | Type of questionnaire | Sortierung

Alternatively, you can also use the filters and select "Data Documentations":



✓ Kind of Documentation

Year

Type of questionnaire

□ Survey Instruments

☑ Data Documentations

□ Survey Reports

□ Teaching Materials

Alle Filter löschen

Abbrechen

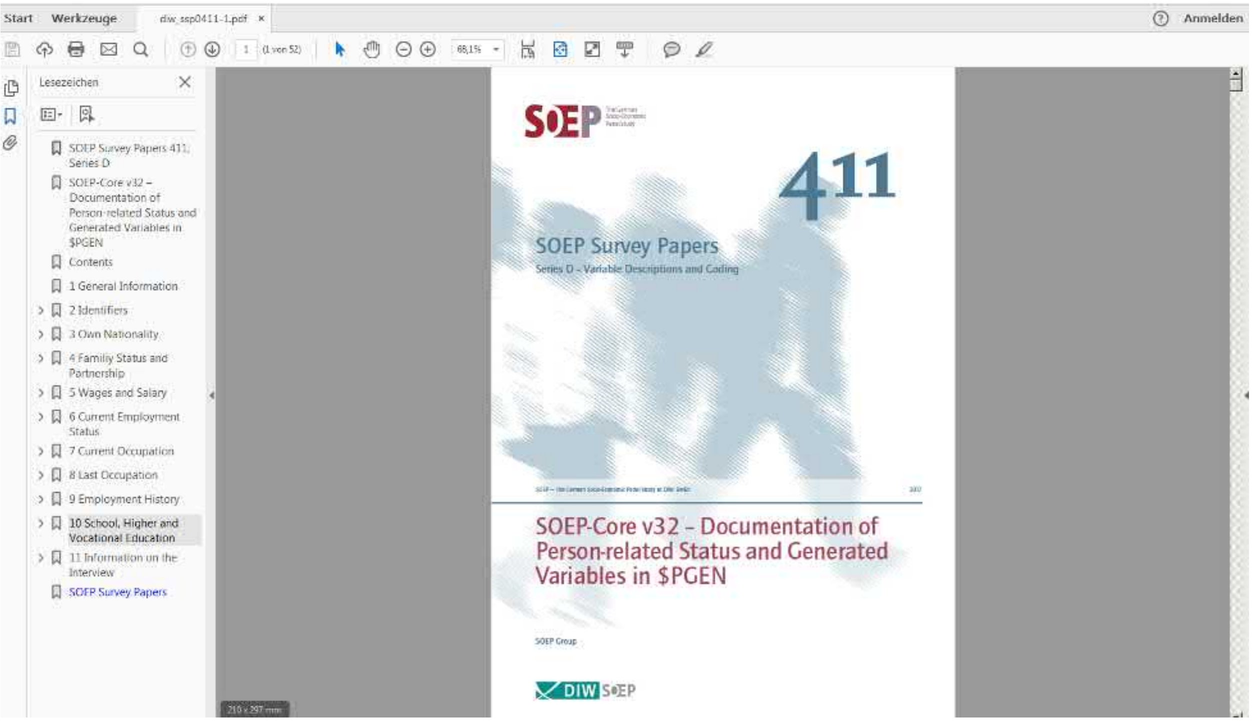
Übernehmen

SOEP SURVEY PAPERS 834: SOEP SURVEY PAPERS : SERIES D - VARIABLE DESCRIPTION AND CODING / 2020

SOEP-Core v35 – *PGEN*: Person-Related Status and Generated Variables

2020 | SOEP Group

Now select the documentation of the required version of pgen



SOEP Survey Papers 411: Series D

SOEP-Core v32 – Documentation of Person-related Status and Generated Variables in $PGEN

Contents

1 General Information

2 Identifiers

3 Own Nationality

4 Family Status and Partnership

5 Wages and Salary

6 Current Employment Status

7 Current Occupation

8 Last Occupation

9 Employment History

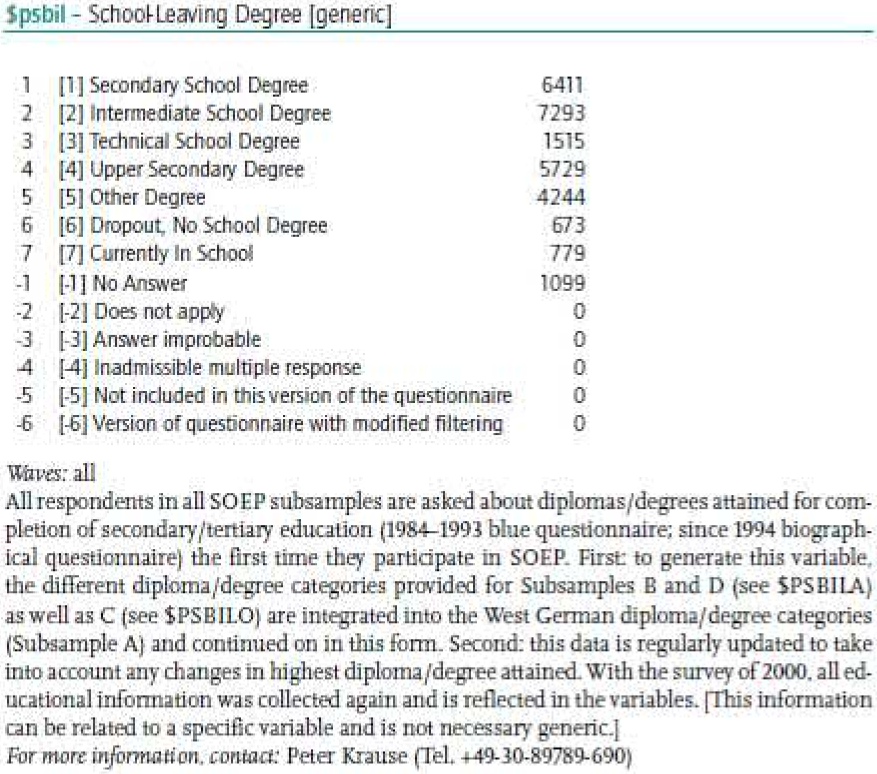
10 School, Higher and Vocational Education

11 Information on the Interview

SOEP Survey Papers

SOEP-Core v32 – Documentation of Person-related Status and Generated Variables in $PGEN

The table of contents on the left gives you a classification of the dataset by topics. To find the variable you are looking for, select topic area 10.



$psbil - School-Leaving Degree [generic]

1	[1] Secondary School Degree	6411
2	[2] Intermediate School Degree	7293
3	[3] Technical School Degree	1515
4	[4] Upper Secondary Degree	5729
5	[5] Other Degree	4244
6	[6] Dropout, No School Degree	673
7	[7] Currently In School	779
-1	[-1] No Answer	1099
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0

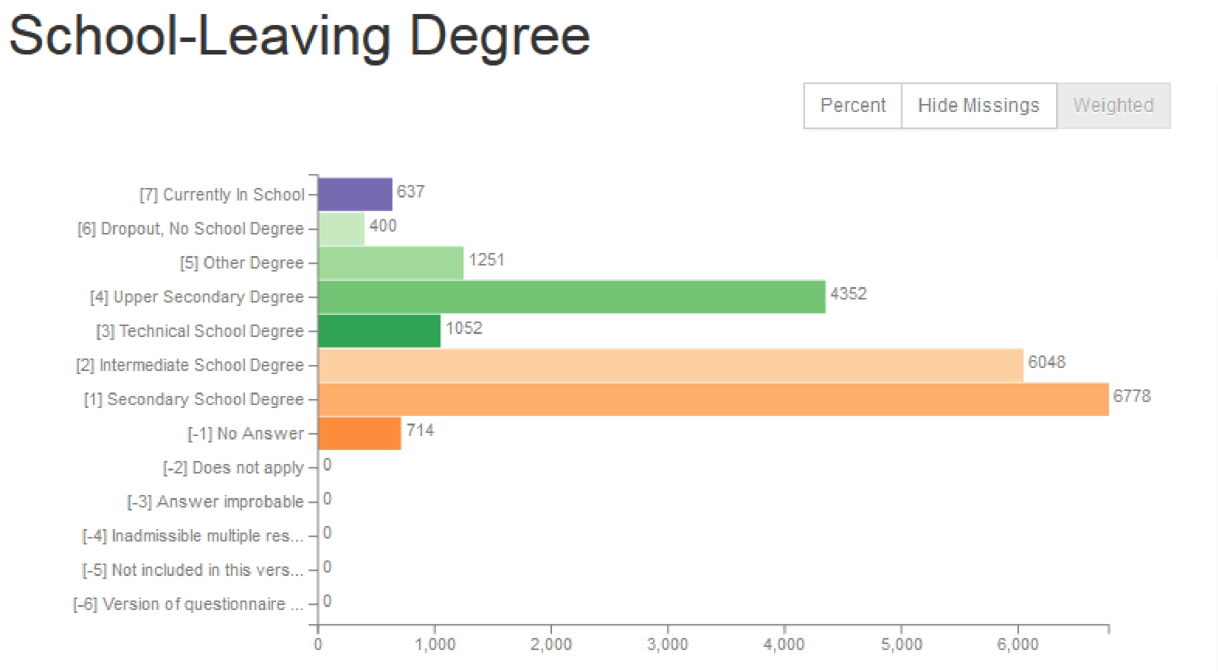
Waves: all
All respondents in all SOEP subsamples are asked about diplomas/degrees attained for completion of secondary/tertiary education (1984–1993 blue questionnaire; since 1994 biographical questionnaire) the first time they participate in SOEP. First: to generate this variable, the different diploma/degree categories provided for Subsamples B and D (see $PSBILA) as well as C (see $PSBILO) are integrated into the West German diploma/degree categories (Subsample A) and continued on in this form. Second: this data is regularly updated to take into account any changes in highest diploma/degree attained. With the survey of 2000, all educational information was collected again and is reflected in the variables. [This information can be related to a specific variable and is not necessary generic.]
For more information, contact: Peter Krause (Tel. +49-30-89789-690)

After a few searches, you will find the variable you are looking for. The documentation provides useful information about the generated variable: it comes from the biography questionnaire, which was introduced in 1994 and is administered only once per respondent. The documentation also explains the two additional variables $psbila and $psbilo in more detail: the $psbil variable is updated regularly to take into account possible changes in the respondent's highest school-leaving certificate. For this reason, the generated variable is useful in providing the most up-to-date information on completed secondary schooling.

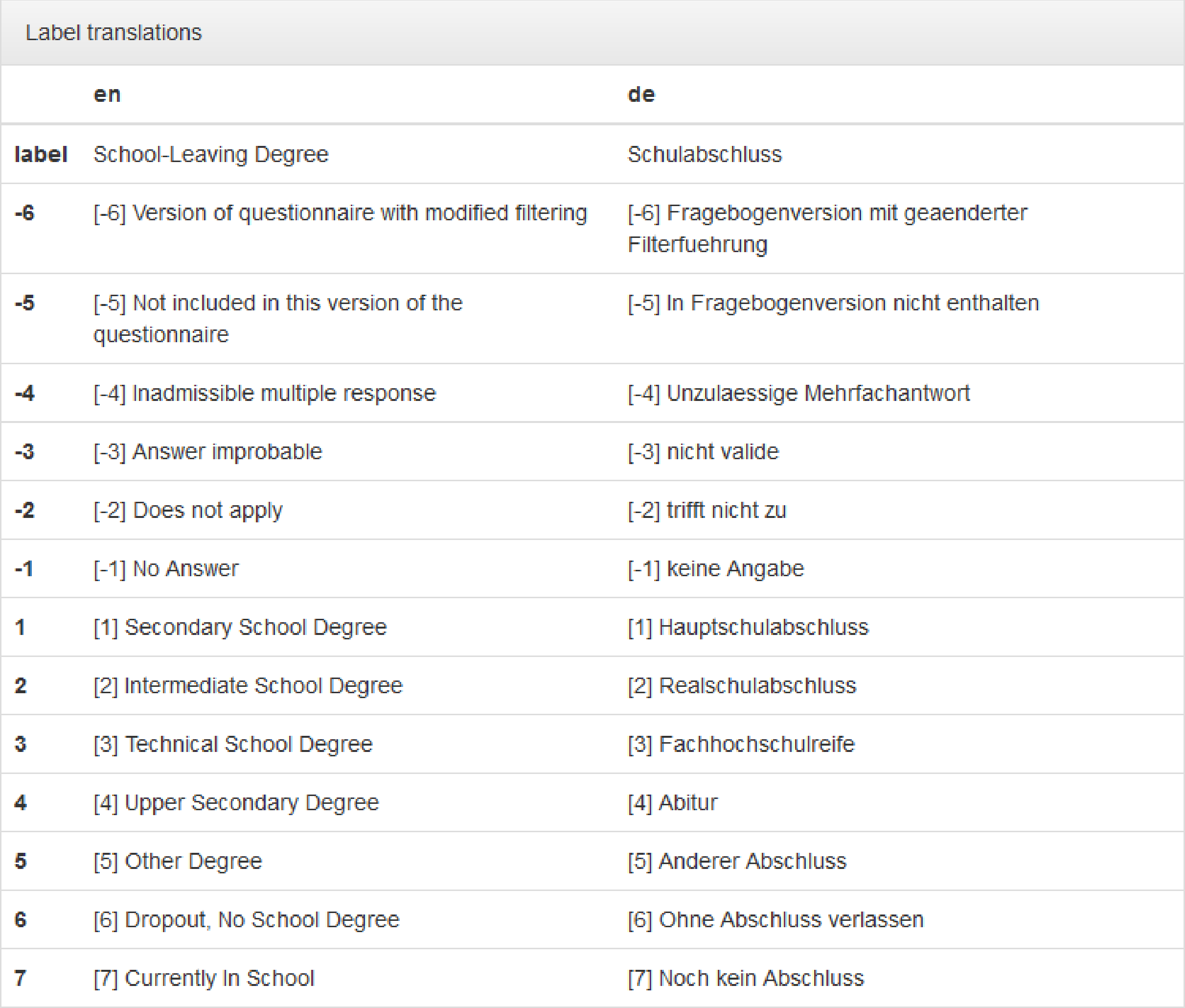
The variable we are looking for is xpsbil and describes the highest degree in certificate attained by individuals surveyed since 2007.

b) What values do individuals with an upper secondary school-leaving certificate (Abitur) have for this variable??

Since you now know the variable you are looking for, you can use the extensive functions of paneldata.org in addition to the information from the documentation. If you search for the variable "xpsbil" in paneldata.org and click on it, the frequency counts are displayed.



In addition to the absolute and relative frequencies, you can also read the value codes of specific response categories. A translation of the answer categories can be found in the "Label translations" section:



Label translations

	en	de
label	School-Leaving Degree	Schulabschluss
-6	[-6] Version of questionnaire with modified filtering	[-6] Fragebogenversion mit geaenderter Filterfuehrung
-5	[-5] Not included in this version of the questionnaire	[-5] In Fragebogenversion nicht enthalten
-4	[-4] Inadmissible multiple response	[-4] Unzulaessige Mehrfachantwort
-3	[-3] Answer improbable	[-3] nicht valide
-2	[-2] Does not apply	[-2] trifft nicht zu
-1	[-1] No Answer	[-1] keine Angabe
1	[1] Secondary School Degree	[1] Hauptschulabschluss
2	[2] Intermediate School Degree	[2] Realschulabschluss
3	[3] Technical School Degree	[3] Fachhochschulreife
4	[4] Upper Secondary Degree	[4] Abitur
5	[5] Other Degree	[5] Anderer Abschluss
6	[6] Dropout, No School Degree	[6] Ohne Abschluss verlassen
7	[7] Currently In School	[7] Noch kein Abschluss

You can answer the question without opening the data. In the 2007 survey year, the variable “xpsbil” with the value code “4” describes the response category “upper secondary school-leaving certificate (Abitur)”.

Last change: May 12, 2022

7.5 Working with SOEPhelp

For data users, the SOEP provides assistance in two analysis programs. One is a stata.ado and the other an R package. The application simply has to be installed in the respective program and helpful information on the desired data sets or variables can be provided. In the following, the installation and use of SOEPhelp is explained:

7.5.1 Working with SOEPhelp in R

SOEPhelp is an R package providing help pages for SOEP-Core data sets (top-level folder only) and their variables. Starting with version v35 SOEPhelp is available for R users, too. Using meta data and the documentation available for SOEP-Core data, this package displays the information on data sets and variables in the default R help. Wherever available the package provides labels in English and in German language. The basic information used to create the help pages is taken from the meta data available from the Public Core Documentation on git. Help pages are **not** available for data sets in the raw and EU-SILC Clone folders.

Installation

If you are working on a Windows OS, you need to install the Rtools (get them here) first. The installation can take a little while according to your CPU (between 3 minutes up to over an hour), because the package contains more than 14000 help pages. The most recent version (SOEPhelp_0.37.1.tar.gz) has been build using R 4.2.0.

```
install.packages("http://companion.soep.de/SOEPhelpR/SOEPhelp_0.37.1.tar.gz",
                 repos = NULL, type = 'source', quiet = TRUE)
```

Usage

Load the package into your library and read the main page carefully.

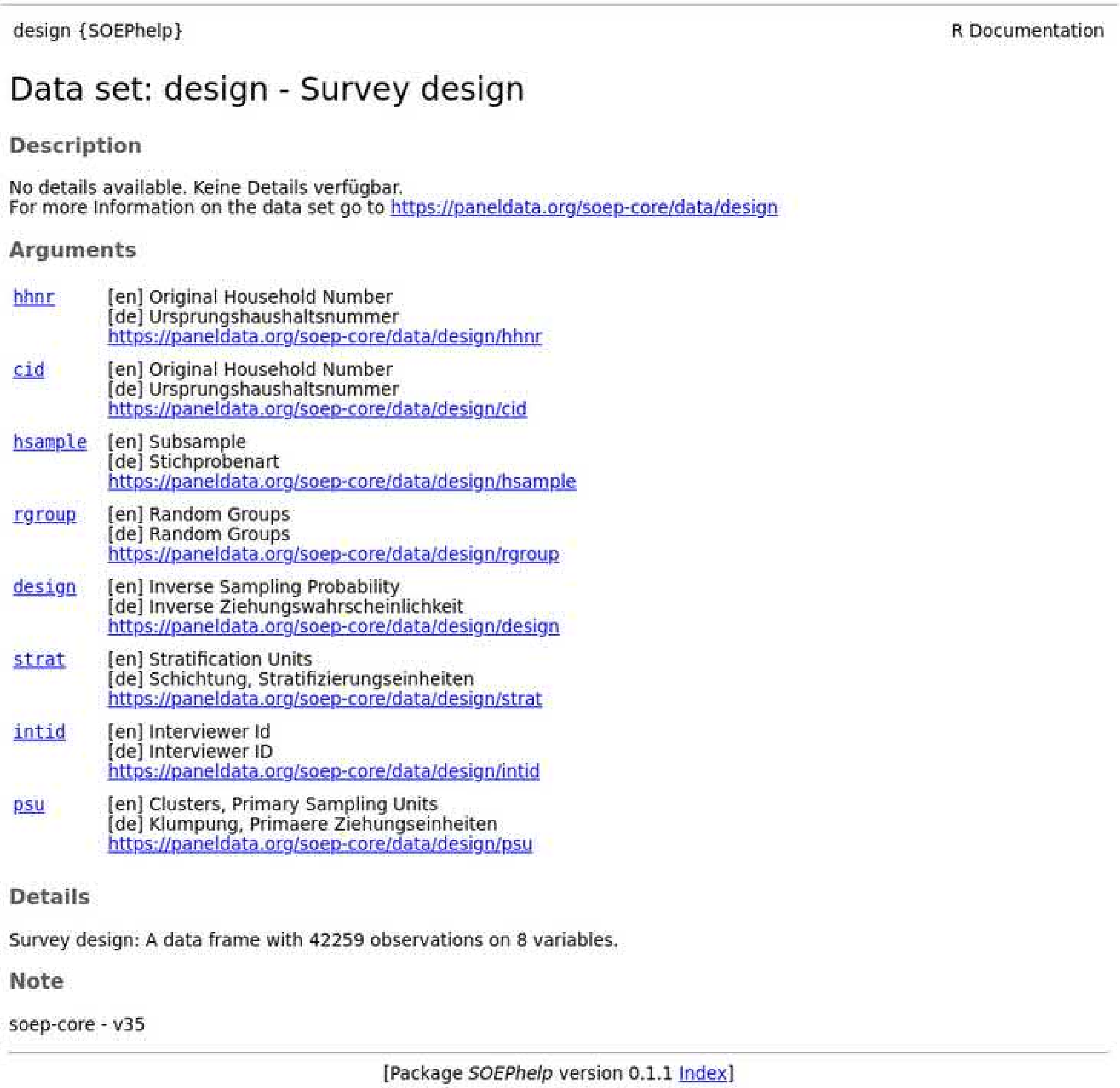
```
library(SOEPhelp)
?SOEPhelp
```

You can get to the help pages by using familiar R help functions like `?` or `help()` as well as `??`.

Example for a dataset

```
?design
```

Asking for the help page of the design data set in R will open the following help page. The title provides you with the basic information that the help page belongs to a data set, its name, and a brief information on what the data set contains. This is followed by the description section providing (if available) further description on the data set and a link to the paneldata.org page. In the arguments section you will find the list of variables for the data set. The variable names are linked to their help pages and the variable labels are given in English and German (if available), together with the link to the paneldata.org page. The details section gives information on the number of observations and variables. Finally, the notes section refers to the version of the SOEP-Core.



design {SOEPhelp}   R Documentation

Data set: design - Survey design

Description

No details available. Keine Details verfügbar.
For more Information on the data set go to https://paneldata.org/soep-core/data/design

Arguments

hhnr	[en] Original Household Number [de] Ursprungshaushaltsnummer https://paneldata.org/soep-core/data/design/hhnr
cid	[en] Original Household Number [de] Ursprungshaushaltsnummer https://paneldata.org/soep-core/data/design/cid
hsample	[en] Subsample [de] Stichprobenart https://paneldata.org/soep-core/data/design/hsample
rgroup	[en] Random Groups [de] Random Groups https://paneldata.org/soep-core/data/design/rgroup
design	[en] Inverse Sampling Probability [de] Inverse Ziehungswahrscheinlichkeit https://paneldata.org/soep-core/data/design/design
strat	[en] Stratification Units [de] Schichtung, Stratifizierungseinheiten https://paneldata.org/soep-core/data/design/strat
intid	[en] Interviewer Id [de] Interviewer ID https://paneldata.org/soep-core/data/design/intid
psu	[en] Clusters, Primary Sampling Units [de] Klumpung, Primaere Ziehungseinheiten https://paneldata.org/soep-core/data/design/psu

Details

Survey design: A data frame with 42259 observations on 8 variables.

Note

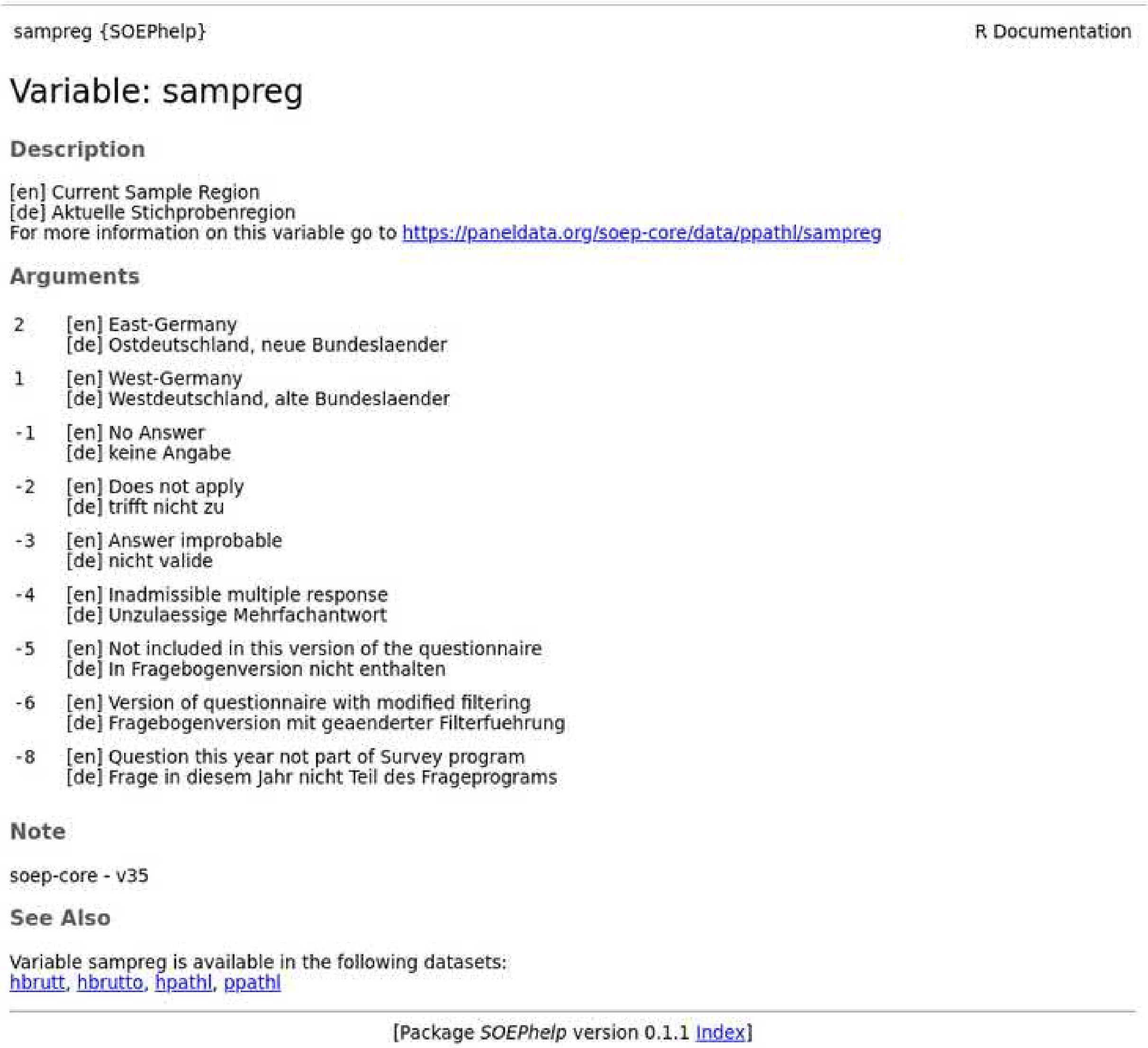
soep-core - v35

[Package *SOEPhelp* version 0.1.1 Index]

Example for a variable

```
?sampreg
```

Asking for the help page of the sampreg variable in R will open the following help page. The title provides you with the basic information that the help page belongs to a variable and its name. This is followed by the description section providing (if available) the variable label in English and German as well as a link to the paneldata.org page. In the arguments section you will find the corresponding values of the variable and their value labels. The following notes section refers to the version of the SOEP-Core. Finally, the see also section lists data sets which contain this variable.



sampreg {SOEPhelp} R Documentation

Variable: sampreg

Description

[en] Current Sample Region
[de] Aktuelle Stichprobenregion
For more information on this variable go to https://paneldata.org/soep-core/data/ppathl/sampreg

Arguments

2	[en] East-Germany [de] Ostdeutschland, neue Bundeslaender
1	[en] West-Germany [de] Westdeutschland, alte Bundeslaender
-1	[en] No Answer [de] keine Angabe
-2	[en] Does not apply [de] trifft nicht zu
-3	[en] Answer improbable [de] nicht valide
-4	[en] Inadmissible multiple response [de] Unzulaessige Mehrfachantwort
-5	[en] Not included in this version of the questionnaire [de] In Fragebogenversion nicht enthalten
-6	[en] Version of questionnaire with modified filtering [de] Fragebogenversion mit geaenderter Filterfuehrung
-8	[en] Question this year not part of Survey program [de] Frage in diesem Jahr nicht Teil des Frageprograms

Note

soep-core - v35

See Also

Variable sampreg is available in the following datasets:
hbrutt, hbrutto, hpathl, ppathl

[Package *SOEPhelp* version 0.1.1 Index]

7.5.2 Working with SOEPhelp in STATA

Attention: The following tool is available starting with Version v34 (Wave bh) and Stata Version 12.

The SOEP data contain a wide array of useful additional information. SOEPhelp is a stata.ado that displays documentation on the dataset at hand. It displays information such as variable histories directly in your Stata window.

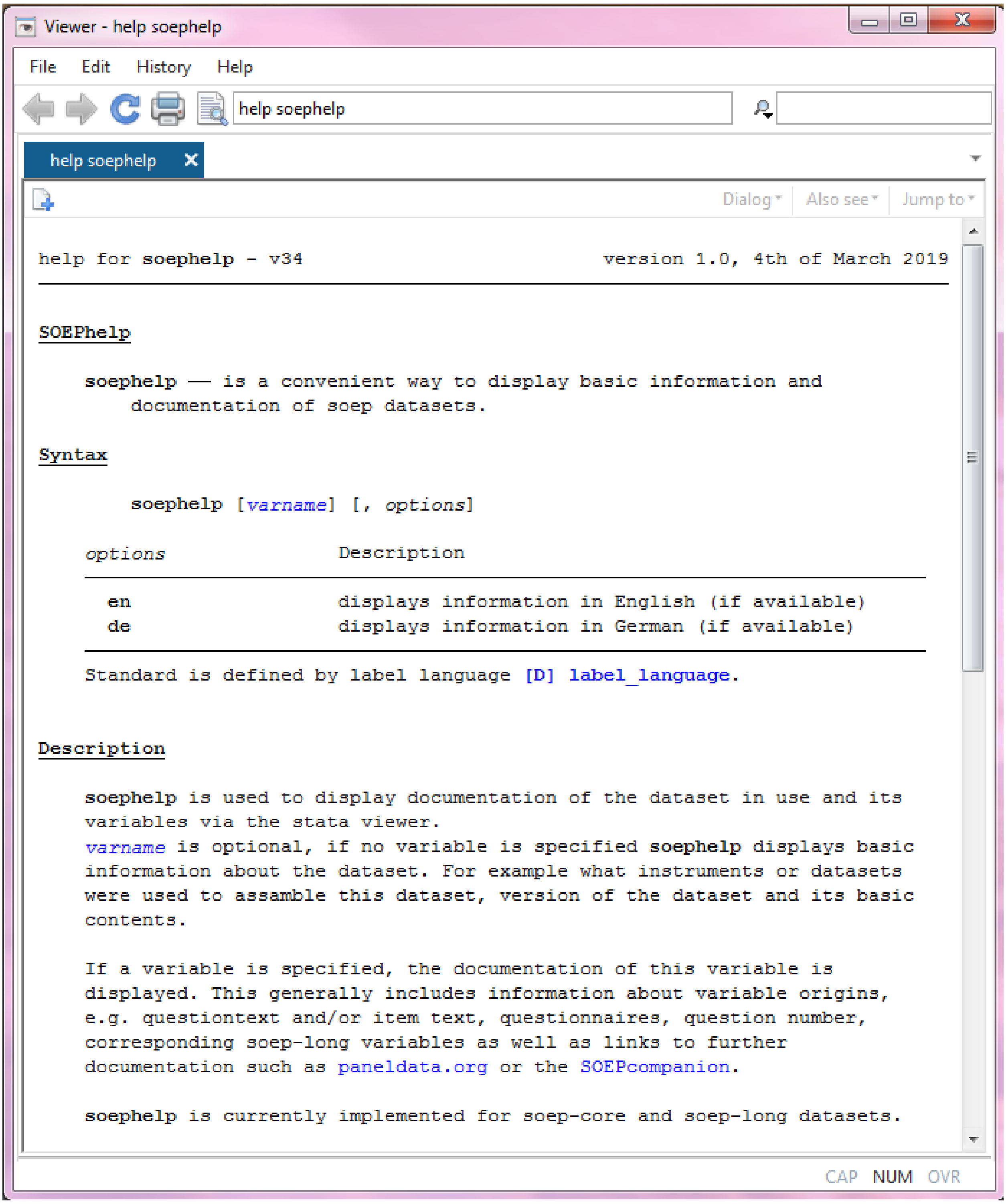
Installation

Open Stata and enter the following command:

```
net install soephelp, replace from(http://companion.soep.de/SOEPhelp/)
```

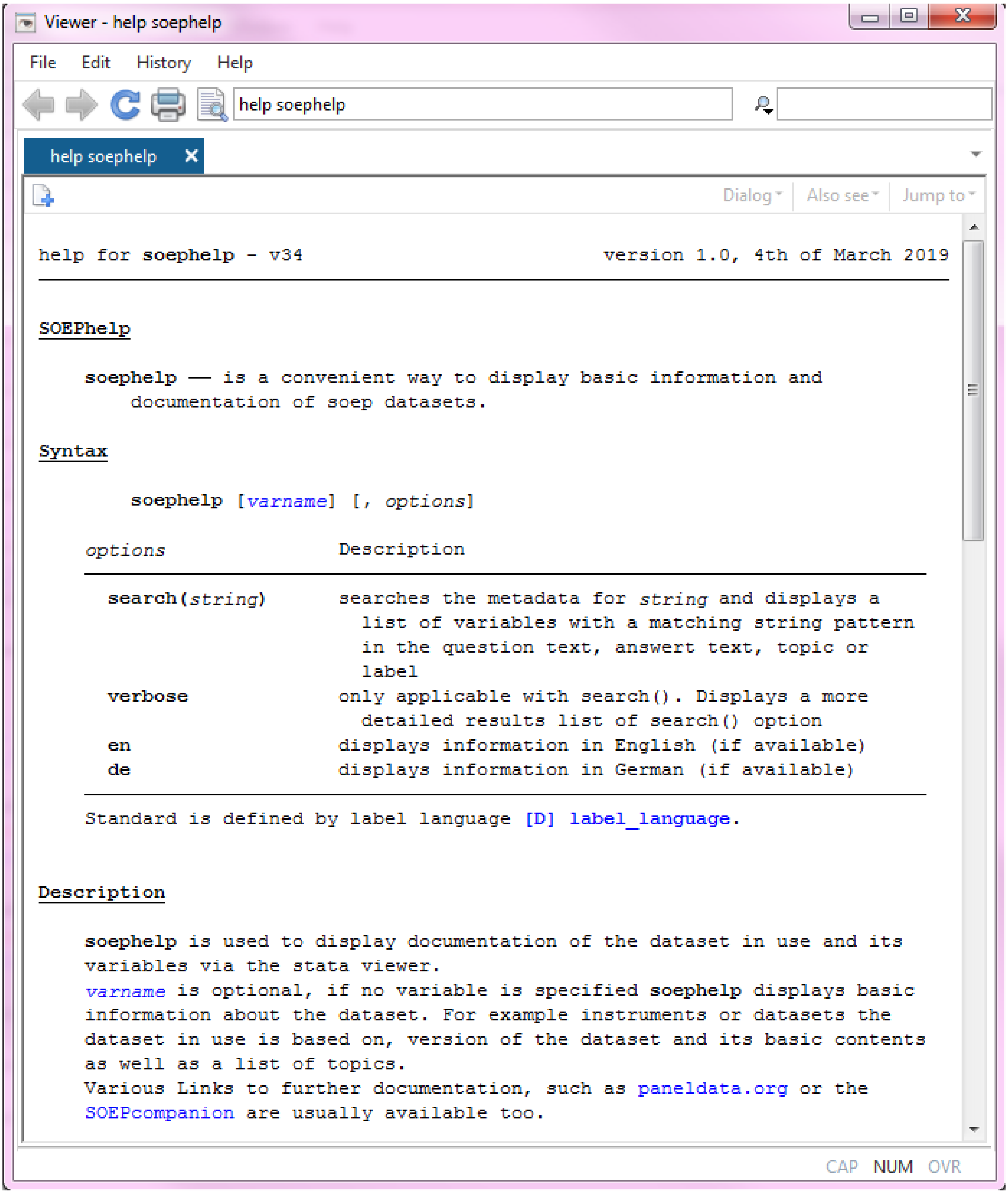
The following commands are provided by .ado:

For a general introduction to SOEPhelp, type in the command `help soephelp`. Here you will find a detailed explanation of the Stata.ado and the different ways to use it. The .ado is available in German and English.



```
help for soephelp - v34                              version 1.0, 4th of March 2019
-----------------------------------------------------------------------------------

SOEPhelp

    soephelp -- is a convenient way to display basic information and
        documentation of soep datasets.

Syntax

        soephelp [varname] [, options]

    options                     Description
    -------------------------------------------------------------------------
      en                        displays information in English (if available)
      de                        displays information in German (if available)
    -------------------------------------------------------------------------
    Standard is defined by label language [D] label_language.


Description

    soephelp is used to display documentation of the dataset in use and its
    variables via the stata viewer.
    varname is optional, if no variable is specified soephelp displays basic
    information about the dataset. For example what instruments or datasets
    were used to assamble this dataset, version of the dataset and its basic
    contents.

    If a variable is specified, the documentation of this variable is
    displayed. This generally includes information about variable origins,
    e.g. questiontext and/or item text, questionnaires, question number,
    corresponding soep-long variables as well as links to further
    documentation such as paneldata.org or the SOEPcompanion.

    soephelp is currently implemented for soep-core and soep-long datasets.
```

With the command `soephelp`, using wave specific datasets (subdirectory *raw*), you receive a basic description of the dataset as well as a list of samples contained in it, including the instruments corresponding to the sample. By clicking on the provided links, you will get to the respective questionnaires or to the dataset on paneldata.



```
help for soephelp - v34                          version 1.0, 4th of March 2019
```

SOEPhelp

    soephelp — is a convenient way to display basic information and
        documentation of soep datasets.

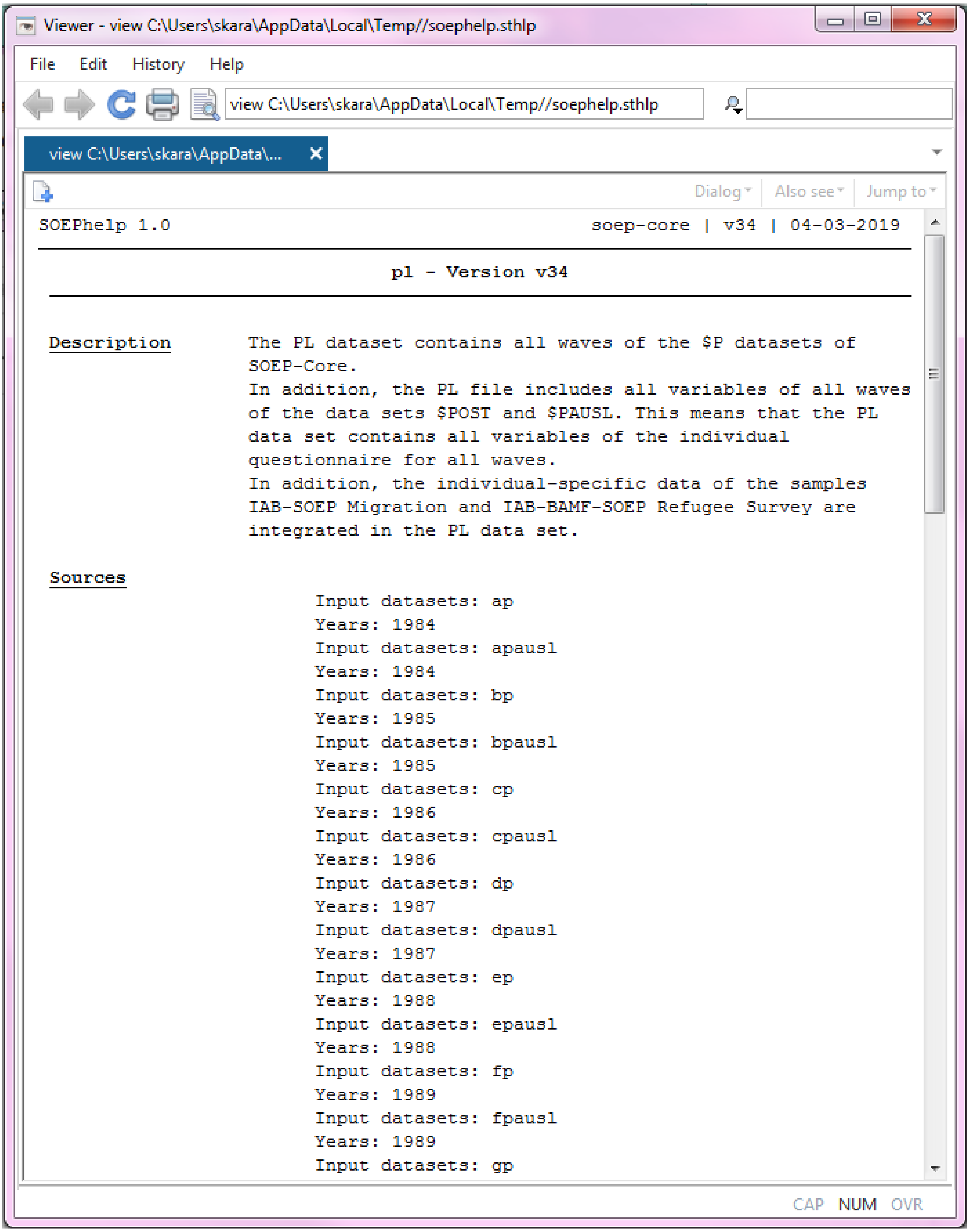
Syntax

```
        soephelp [varname] [, options]

    options                Description
    -----------------------------------------------------------------------
      search(string)       searches the metadata for string and displays a
                             list of variables with a matching string pattern
                             in the question text, answert text, topic or
                             label
      verbose              only applicable with search(). Displays a more
                             detailed results list of search() option
      en                   displays information in English (if available)
      de                   displays information in German (if available)
    -----------------------------------------------------------------------
    Standard is defined by label language [D] label_language.
```

Description

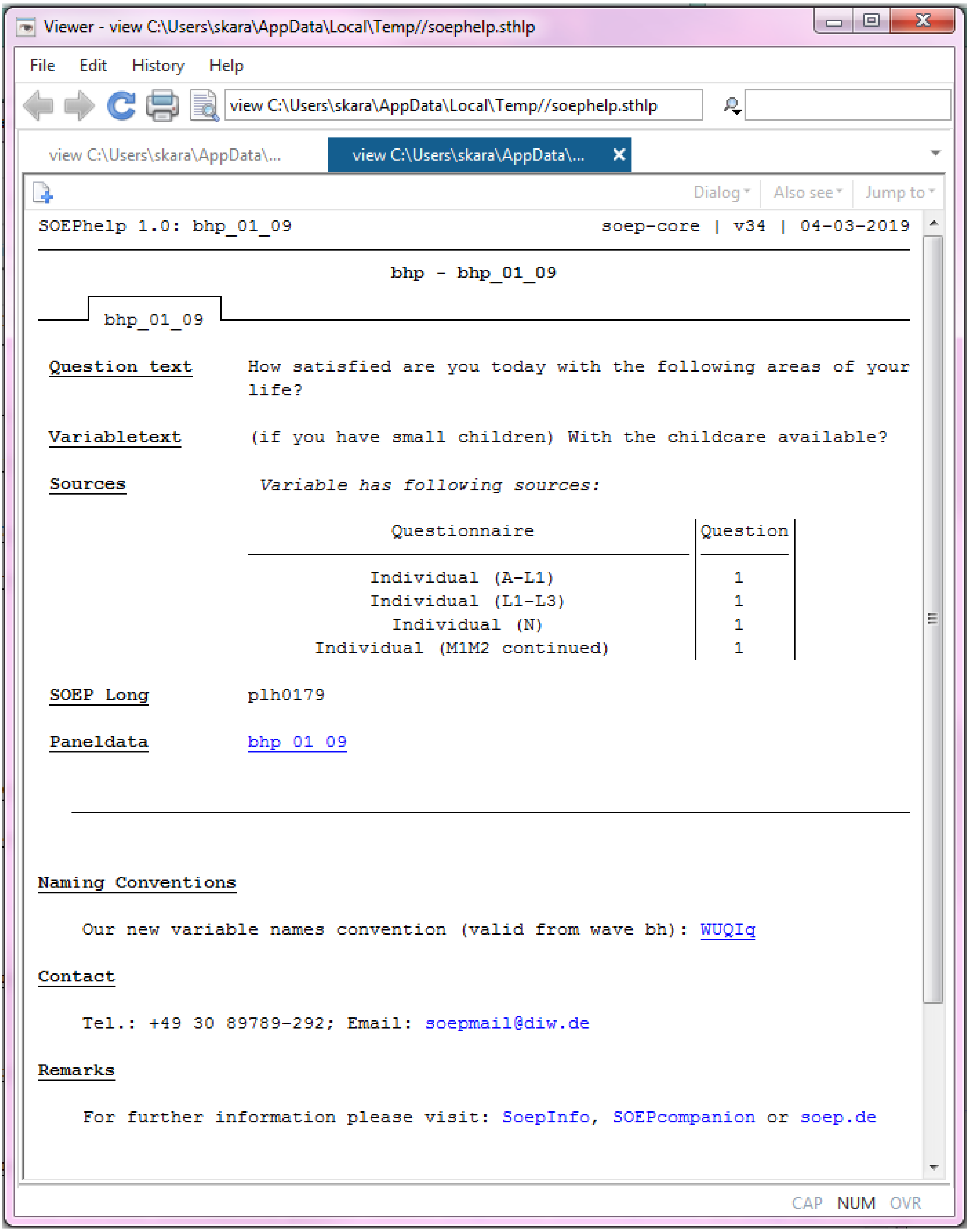
    soephelp is used to display documentation of the dataset in use and its variables via the stata viewer.
    varname is optional, if no variable is specified soephelp displays basic information about the dataset. For example instruments or datasets the dataset in use is based on, version of the dataset and its basic contents as well as a list of topics.
    Various Links to further documentation, such as paneldata.org or the SOEPcompanion are usually available too.

Using `soephelp` in longitudinal datasets, you also receive a basic description as well as a list of wave-specific datasets that are used to generate the longitudinal version.


Viewer - view C:\Users\skara\AppData\Local\Temp//soephelp.sthlp
File Edit History Help
view C:\Users\skara\AppData\Local\Temp//soephelp.sthlp
view C:\Users\skara\AppData\...
Dialog
Also see
Jump to
CAP NUM OVR

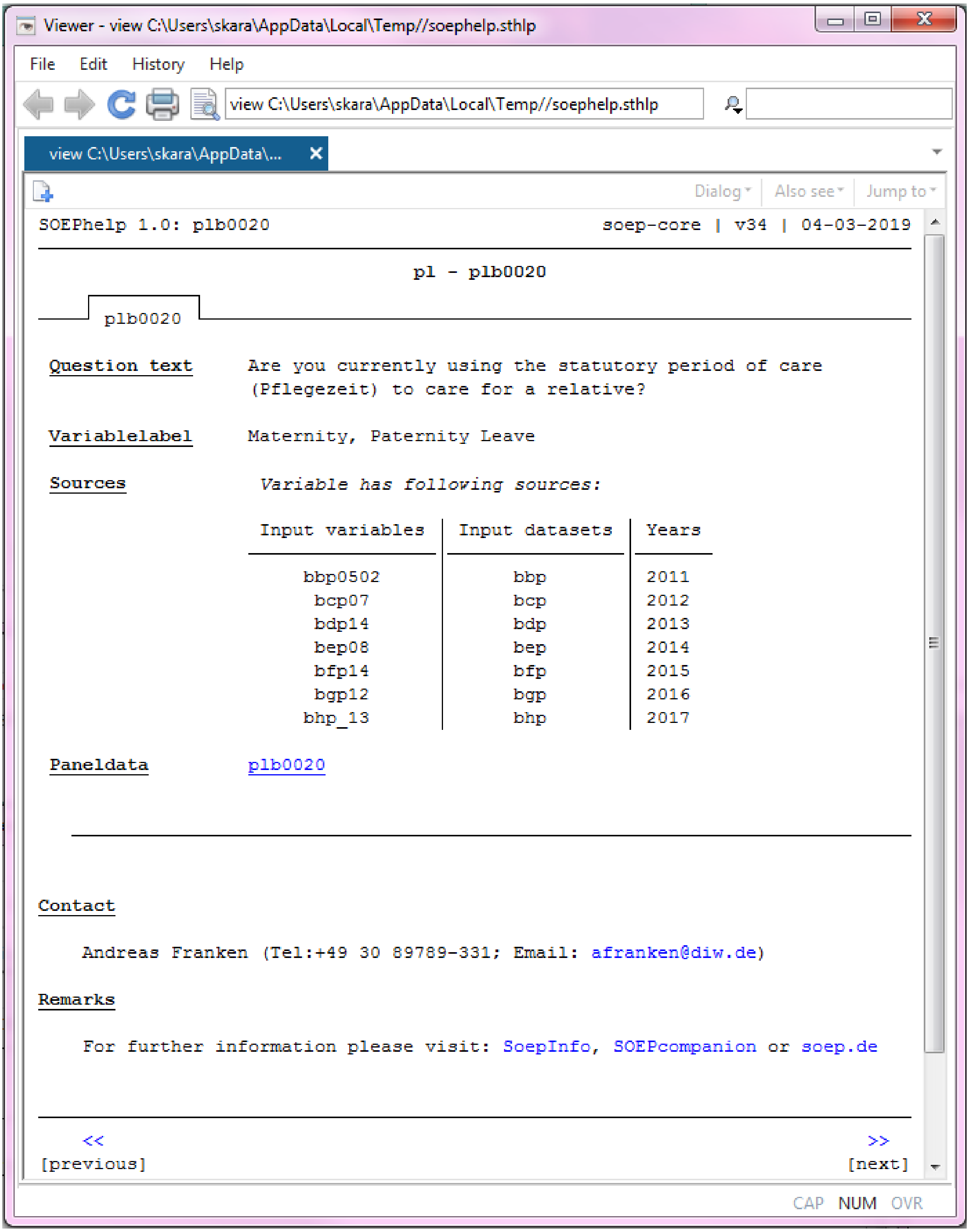
```
SOEPhelp 1.0                                  soep-core | v34 | 04-03-2019

                              pl - Version v34

 Description      The PL dataset contains all waves of the $P datasets of
                  SOEP-Core.
                  In addition, the PL file includes all variables of all waves
                  of the data sets $POST and $PAUSL. This means that the PL
                  data set contains all variables of the individual
                  questionnaire for all waves.
                  In addition, the individual-specific data of the samples
                  IAB-SOEP Migration and IAB-BAMF-SOEP Refugee Survey are
                  integrated in the PL data set.

 Sources
                        Input datasets: ap
                        Years: 1984
                        Input datasets: apausl
                        Years: 1984
                        Input datasets: bp
                        Years: 1985
                        Input datasets: bpausl
                        Years: 1985
                        Input datasets: cp
                        Years: 1986
                        Input datasets: cpausl
                        Years: 1986
                        Input datasets: dp
                        Years: 1987
                        Input datasets: dpausl
                        Years: 1987
                        Input datasets: ep
                        Years: 1988
                        Input datasets: epausl
                        Years: 1988
                        Input datasets: fp
                        Years: 1989
                        Input datasets: fpausl
                        Years: 1989
                        Input datasets: gp
```

If you enter the command `soephelp <variable>` in a wave-specific data set, you will get detailed information about the variable in question. The question asked in the questionnaire is displayed as well as the samples and instruments in which the question was asked. Additionally, the command offers the corresponding long variable as well as the link of the displayed variable to the documentation at paneldata.org.


Viewer - view C:\Users\skara\AppData\Local\Temp//soephelp.sthlp
File Edit History Help
view C:\Users\skara\AppData\Local\Temp//soephelp.sthlp
view C:\Users\skara\AppData\...
view C:\Users\skara\AppData\...
Dialog
Also see
Jump to
CAP NUM OVR

```
SOEPhelp 1.0: bhp_01_09                        soep-core | v34 | 04-03-2019
---------------------------------------------------------------------------
                          bhp - bhp_01_09

      bhp_01_09

 Question text      How satisfied are you today with the following areas of your
                    life?

 Variabletext       (if you have small children) With the childcare available?

 Sources             Variable has following sources:

                                 Questionnaire                 | Question
                    ----------------------------------------   |---------
                              Individual (A-L1)                |   1
                              Individual (L1-L3)               |   1
                                Individual (N)                 |   1
                         Individual (M1M2 continued)           |   1

 SOEP Long          plh0179

 Paneldata          bhp 01 09

   ------------------------------------------------------------------------

Naming Conventions

    Our new variable names convention (valid from wave bh): WUQIq

Contact

    Tel.: +49 30 89789-292; Email: soepmail@diw.de

Remarks

    For further information please visit: SoepInfo, SOEPcompanion or soep.de
```

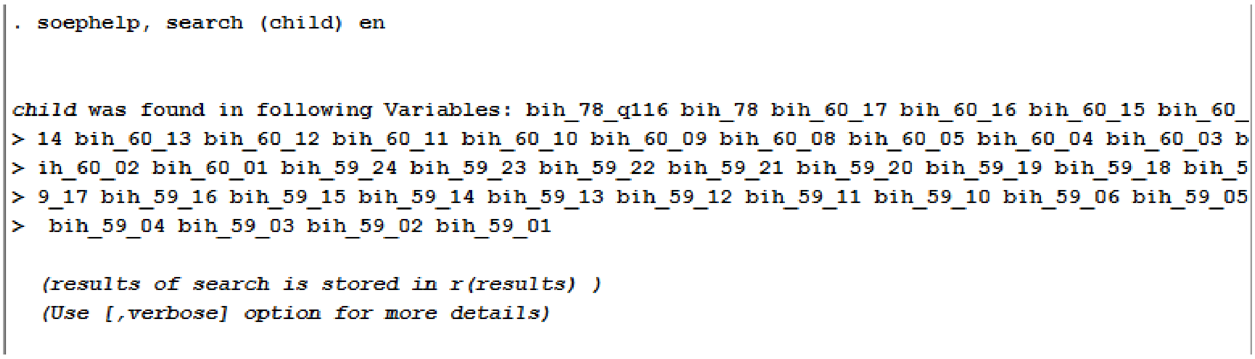
Conversely, with long data, you receive the wave-specific input variables and datasets used to generate the long-variable.


Viewer - view C:\Users\skara\AppData\Local\Temp//soephelp.sthlp
File
Edit
History
Help
view C:\Users\skara\AppData\Local\Temp//soephelp.sthlp
view C:\Users\skara\AppData\...
Dialog
Also see
Jump to
CAP
NUM
OVR

```
SOEPhelp 1.0: plb0020                          soep-core | v34 | 04-03-2019

                               pl - plb0020

     plb0020

 Question text     Are you currently using the statutory period of care
                   (Pflegezeit) to care for a relative?

 Variablelabel     Maternity, Paternity Leave

 Sources            Variable has following sources:

                    Input variables | Input datasets  | Years

                        bbp0502     |      bbp        | 2011
                         bcp07      |      bcp        | 2012
                         bdp14      |      bdp        | 2013
                         bep08      |      bep        | 2014
                         bfp14      |      bfp        | 2015
                         bgp12      |      bgp        | 2016
                        bhp_13      |      bhp        | 2017

 Paneldata         plb0020


Contact

    Andreas Franken (Tel:+49 30 89789-331; Email: afranken@diw.de)

Remarks

    For further information please visit: SoepInfo, SOEPcompanion or soep.de


    <<                                                                    >>
[previous]                                                            [next]
```

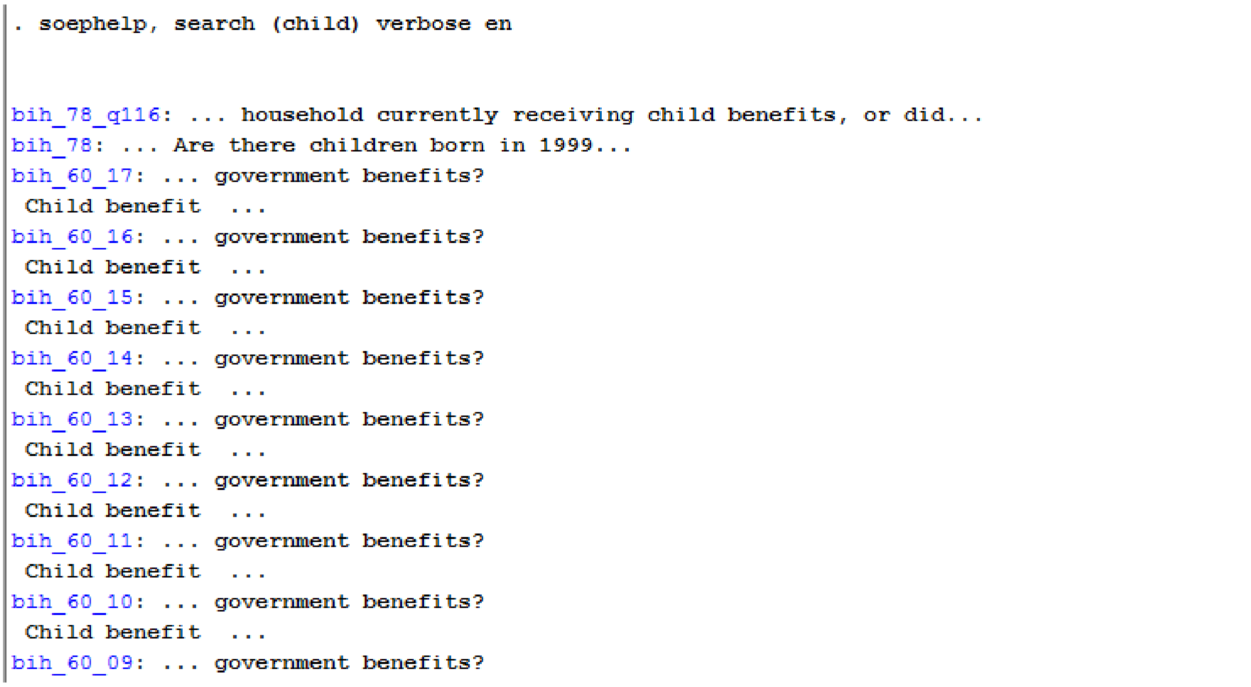
Since our recent wave (v35) a new stata command option is being introduced. With `soephelp, search (string)` you reveive a list of variables that contain the respective word or label you are looking for.

For example, you are interestet in variables regarding children in a household. With `soephelp, search (child)` you are able to see all variables having the word child in their label.



```
. soephelp, search (child) en


child was found in following Variables: bih_78_q116 bih_78 bih_60_17 bih_60_16 bih_60_15 bih_60_
> 14 bih_60_13 bih_60_12 bih_60_11 bih_60_10 bih_60_09 bih_60_08 bih_60_05 bih_60_04 bih_60_03 b
> ih_60_02 bih_60_01 bih_59_24 bih_59_23 bih_59_22 bih_59_21 bih_59_20 bih_59_19 bih_59_18 bih_5
> 9_17 bih_59_16 bih_59_15 bih_59_14 bih_59_13 bih_59_12 bih_59_11 bih_59_10 bih_59_06 bih_59_05
>  bih_59_04 bih_59_03 bih_59_02 bih_59_01

  (results of search is stored in r(results) )
  (Use [,verbose] option for more details)
```

To receive more details on the list of variables, use `soephelp, search (child) verbose`. Now you have the possibility to click on a variable and a new window opens up with details on the variable, like the question asked in the latest questionnaire, the question`s source, the long or core variable, depending on the data format.



```
. soephelp, search (child) verbose en


bih_78_q116: ... household currently receiving child benefits, or did...
bih_78: ... Are there children born in 1999...
bih_60_17: ... government benefits?
 Child benefit  ...
bih_60_16: ... government benefits?
 Child benefit  ...
bih_60_15: ... government benefits?
 Child benefit  ...
bih_60_14: ... government benefits?
 Child benefit  ...
bih_60_13: ... government benefits?
 Child benefit  ...
bih_60_12: ... government benefits?
 Child benefit  ...
bih_60_11: ... government benefits?
 Child benefit  ...
bih_60_10: ... government benefits?
 Child benefit  ...
bih_60_09: ... government benefits?
```

To use this option in english, add `en` at the end of the option. For example, `soephelp, search (child) verbose en`.

SOEPhelp is directly linked to the SOEPcompanion.

Contributions

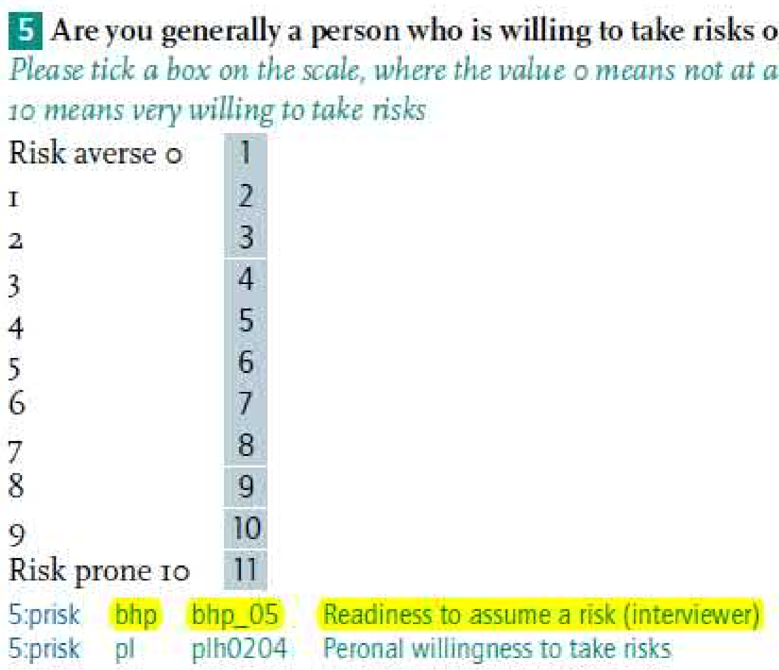
Contributions of all sorts are very welcome. Issues and requests can be reported to:

> Hans Walter Steinhauer for R and Marvin Petrenz for STATA

Last change: Jun 23, 2022

7.6 Working with Metadata-based Questionnaires

Metadata-based questionnaires make it considerably easier to find the variables of interest from the perspective of the questionnaire. Each of the generated PDFs reflects a questionnaire. With the help of these documents the user learns which questions have been asked in the respective sample and in which sequence. In addition, the documents make it clear what the question variable is called and which dataset it can be found in. The example shows question 5 from the individual questionnaire of SOEP-Core, which can be found in the data set bhp under the variable name bhp_05.



5 Are you generally a person who is willing to take risks or do you try to avoid taking risks?

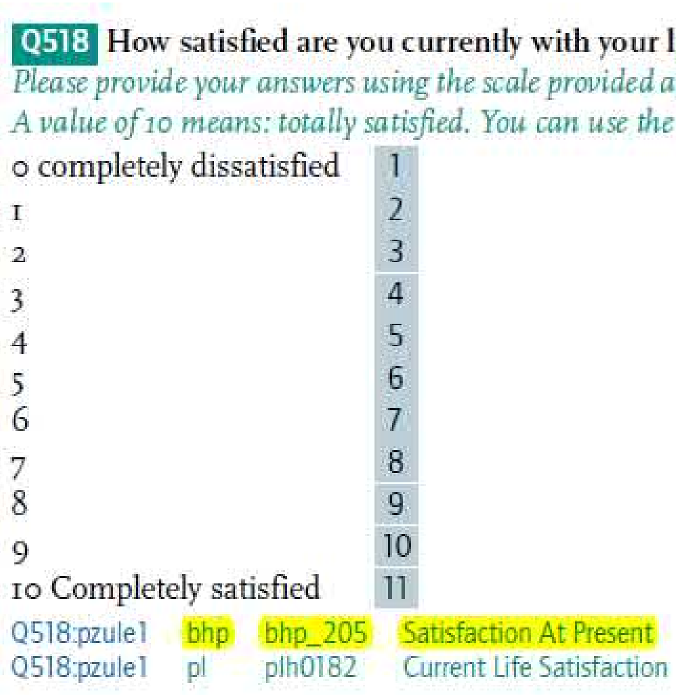
Please tick a box on the scale, where the value 0 means not at all willing to take risks and the value 10 means very willing to take risks

Risk averse 0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10
Risk prone 10	11

5:prisk	bhp	bhp_05	Readiness to assume a risk (interviewer)
5:prisk	pl	plh0204	Peronal willingness to take risks

1. Example: Integrated Variable

Let's say you're interested in finding out about refugees' general life satisfaction. Search the questionnaire to find which refugees were surveyed for a second time in 2017. You'll find what you're looking for under question Q518. Below the question is the information on the name of the variable and the dataset where it is found.



Q518 How satisfied are you currently with your life in general?

Please provide your answers using the scale provided again. A value of 0 means: totally dissatisfied. A value of 10 means: totally satisfied. You can use the in-between ratings to tailor your response.

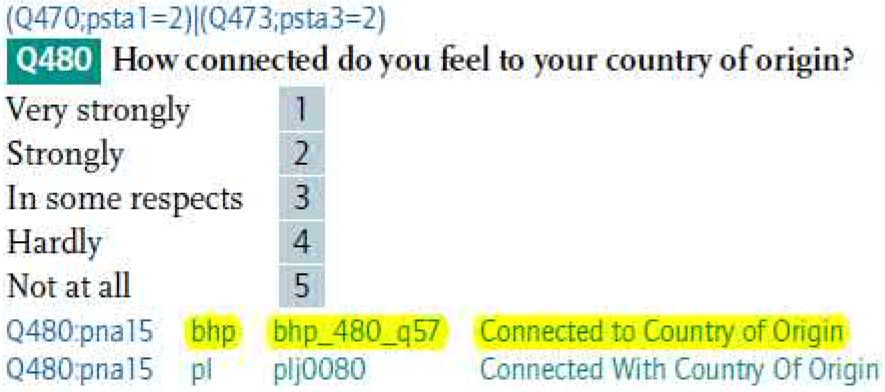
0 completely dissatisfied	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10
10 Completely satisfied	11

Q518:pzule1	bhp	bhp_205	Satisfaction At Present
Q518:pzule1	pl	plh0182	Current Life Satisfaction

The general satisfaction with life can be found in the dataset bhp under the name bhp_205.

2. Example: Additional Variable

Let's say you're interested in finding out about countries or origin. You want to know specifically how connected respondents feel to their country of origin. You'll find the question in the questionniare given to refugees participating in the survey for the second time or more under question number Q480.



(Q470;psta1=2)|(Q473;psta3=2)

Q480 How connected do you feel to your country of origin?

Very strongly	1
Strongly	2
In some respects	3
Hardly	4
Not at all	5

Q480:pna15	bhp	bhp_480_q57	Connected to Country of Origin
Q480:pna15	pl	plj0080	Connected With Country Of Origin

The information on the question is stored in the data file bhp under the name bhp_480_q57. The name indicates that the question is not in the samples A-M2 because it has the suffix _q57. This does not preclude the question from being further down the integration hierarchy in questionnaires.

Last change: May 12, 2022

CHAPTER EIGHT

CONTACT INFORMATION

The first version of the SOEPcompanion (formerly Desktop Companion) was published as a PDF document by John P. Haisken-DeNew and Joachim R. Frick in September 1996. It was originally intended to give novice users a broad introduction in understanding the SOEP, its structure, depth, and research potential. The Desktop Companion was updated several times between 1996 and 2005. The first major change came in 2014, when Jan Goebel and Mathis Schröder decided to shorten the Desktop Companion to its most important content and make it web-based.

The new, completely edited version of the SOEPcompanion (formerly Desktop Companion) has a strong focus on the use of the SOEP-Core data from the perspective of a data user who has received our most recent data release from the SOEP Research Data Center. This new version is not only a web-based documentation, we also offer it as a download.

Address: SOEP, DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany

Homepage: http://www.diw.de/soep

E-Mail: soepmail@diw.de

SOEPhotline: +49 30 89789-292

Developers: Selin Kara, Stefan Zimmermann