Data Observer

Jan Goebel, Markus M. Grabka, Stefan Liebig*, Martin Kroh, David Richter, Carsten Schröder and Jürgen Schupp

The German Socio-Economic Panel (SOEP)

https://doi.org/10.1515/jbnst-2018-0022

1 Introduction

In his 2015 Economic Sciences Nobel lecture, Angus Deaton emphasized key issues for understanding welfare-enhancing policies (see Deaton 2015): First, differences in resources across individuals should be measured not only at specific points in time but also across the life course. Second, to better assess socio-economic outcomes, direct economic measures of well-being should be linked with other measures of well-being developed by other social science branches, such as sociology, demography, and psychology. Third, data observations should be reconciled with lifecycle models to investigate the causal mechanisms behind socio-economic outcomes.

The main goal of the German Socio-Economic Panel (SOEP) is in line with the views and visions expressed in Angus Deaton's lecture: Established in 1984 and located at the German Institute for Economic Research (DIW Berlin), SOEP serves a global research community by providing representative longitudinal data of private households in Germany. The SOEP's primary research interests and the original survey concept was rooted in the multidisciplinary Collaborative Research Center SfB3, "Microanalytic Foundations of Social Policy" at the Universities of

***Corresponding author: Stefan Liebig,** German Institute for Economic Research, Mohrenstr. 58, 10117 Berlin, Germany; and Bielefeld University, Universitätstraße 25, Bielefeld, Germany, E-mail: sliebig@diw.de
Jan Goebel: E-mail: jgoebel@diw.de, **Markus M. Grabka:** E-mail: mgrabka@diw.de, German Institute for Economic Research, Mohrenstr. 58, 10117 Berlin, Germany
Martin Kroh, German Institute for Economic Research, Mohrenstr. 58, 10117 Berlin, Germany; and Bielefeld University, Universitätstraße 25, Bielefeld, Germany, E-mail: martin.kroh@uni-bielefeld.de
David Richter, German Institute for Economic Research, Mohrenstr. 58, 10117 Berlin, Germany, E-mail: drichter@diw.de
Carsten Schröder: E-mail: cschroeder@diw.de, **Jürgen Schupp:** E-mail: jschupp@diw.de, German Institute for Economic Research, Mohrenstr. 58, 10117 Berlin, Germany; and Freie Universität Berlin, Kaiserswerther Str. 16-18, Berlin, Germany

Frankfurt am Main and Mannheim and funded by the German Research Foundation (DFG) from 1983 to 2002. Since 2003, SOEP is part of Germany's research infrastructure under the umbrella of the Leibniz Association (WGL), and funded by the Federal Ministry of Education and Research (BMBF) and state governments (see Schupp 2015).

Nearly 15,000 households and about 30,000 persons participate in the SOEP survey. Guided by the advice of internal and external experts, the SOEP has expanded its scope over the years. As a result, SOEP provides both a broad set of self-reported "objective" variables, such as income, age, gender, education, employment status, or gripping force, and a broad set of self-reported "subjective" variables, such as from satisfaction with life, over fairness and reciprocity perceptions to psychological measurement like the "Big Five."

Running for already 35 years, SOEP gathers information from a spectrum of birth cohorts. As such, it is a valuable empirical basis for researchers to explore long-time societal changes; relationships between early life events on later life outcomes; interdependencies between the individual and the family or household; mechanisms of inter-generational mobility and transmission; accumulation processes of resources; short- and long-term effects of institutional change and policy reforms; speed of convergence between East and West or between migrants and natives. Most notably, SOEP is the only database worldwide in which the political unification of a society that had been divided for 40 years took place during the course of the study: German unification. In sum, SOEP is a comprehensive multi-dimensional database to understanding human behavior and decision making in varying social as well as institutional settings and policy regimes.

To further improve its data, the SOEP team is engaged in own research, research collaborations and joint projects with scholars worldwide, whose discipline-specific expertise adds to the depth and diversity of the SOEP data.

The scientific community acknowledges these activities and features of SOEP, as reflected by a broad international and multidisciplinary research community, an international board of voluntary advisors, and a well-embeddedness in international data-infrastructures like the Cross National Equivalent Files or the Cross-National Data Center in Luxembourg.

This paper provides an update to, and complements, the earlier description of SOEP by Wagner et al. (2007). It is organized as follows. Section 2 provides the basic features of the SOEP – from the basic sampling strategy to the structure of the released data. Section 3 outlines initiatives to enrich the SOEP data with auxiliary datasets. Section 4 is about data access and user support, Section 5 about main scientific contributions, and Section 6 concludes.

2 Basic features

2.1 Sampling and weighting

The target population to be represented by the SOEP is Germany's resident population. The two initial samples in 1984 include private households with a German national as household head (Sample A, n = 4,528) and, oversampled, households with a Greek, Italian, Spanish, Turkish, or Yugoslavian household head (Sample B, n = 1,393).

To maintain the cross-sectional representativeness in the presence of influx to the underlying target population, the enlargement samples are integrated (see Online Appendix Table A1 for an overview of all samples, and Kroh et al. 2015 for further details). One enlargement sample is the East German sample integrated shortly before German unification in 1990 (Sample C, n = 2,179), making SOEP unique among other household panel surveys worldwide. Other enlargement samples are migrant samples, typically integrated after periods of increased gross influx. This applies to the immigration of large numbers of ethnic Germans following the collapse of the Soviet Union (Sample D, n = 531, integrated in 1994/5), the immigration of EU citizens from Central and Eastern Europe after freedom of movement was implemented in the EU (Sample M1, n = 2,732, integrated in 2013, and Sample M2, n = 1,096, integrated in 2015), and the recent immigration of refugees from the Middle East, in particular Syria (Sample M3/4, n = 3,554, integrated in 2016 and Sample M5, n = 1,555, integrated in 2017).[1] Enlargement samples of immigrants of SOEP aim at consecutively covering all immigration years, i. e. the target population of the initial Sample B is immigration until 1983, Sample D covers subsequent immigration years 1984 to 1994, Sample M1 1995 to 2010, Sample M2 2011 to 2013, Sample M3/4 2013 to 2015, and finally Sample M5 immigration in 2016.

Panel attrition is another issue challenging representativeness and a reasonable sample size. Refreshment samples of the residential population of Germany are a means to address both issues. These refreshment samples either have the form of a general population sample or as a boost sample, the latter focusing on specific population subgroups that are the focus of the research community or policy makers. General population refreshments were integrated in 1998 (Sample E, n = 1,056), in 2000 (Sample F, n = 6,043), in 2006 (Sample H, n = 1,506), in 2011 (Sample J, n = 3,136), in 2012 (Sample K, n = 1,526), and in 2017 (Sample N, n = 2,314).

1 See Online Appendix Table A2 for further details on the migration boost samples.

Boost samples focused on high-income households in 2002 (Sample G, n = 1,224), families with newborn and young children in 2010 (Sample L1, n = 2,074), and low income/large families/single parents in 2010/11 (Sample L2, n = 2,500, Sample L3, n = 924).

Two SOEP subsamples that originate from other studies have also been successfully integrated. The L-samples originate from the Families in Germany (FiD) project. FiD started in 2010 in order to enrich the available stock of data on single parents, low income families, large families with many and, in particular, young children (see Schröder et al. 2013). FiD data are the foundation underlying the evaluation of family policy measures in Germany. In 2014, FiD was fully integrated into SOEP-Core version 31 (https://doi.org/10.5684/soep.v31). Sample N originates from data for the Programme for the International Assessment of Adult Competencies Longitudinal (PIAAC-L).[2] Examining the basic skills necessary for adults to participate successfully in society and working life, PIAAC-L is the world's first internationally comparable longitudinal study on competencies and their significance across the life course (see Rammstedt et al. 2017).

SOEP only uses random probability samples. General population samples typically draw on a nation-wide two-stage stratified sampling procedure (Sample A, C, E, F, H, J, K). First, nation-wide sample points are sampled by federal state and municipality size. To secure efficient face-to-face interviewing, the number of regional sample points ranges between 125 and 985 per sample. Second, within each sample point, households are sampled in a random walk procedure. In refreshment samples F, J, and K, a second stratification stage distinguishes natives and migrants. For special populations, SOEP relied on two types of sampling frames: (a) screening in large telephone surveys of the fieldwork organization KANTAR Public (Infratest Dimap) for Samples D, G, L2, and L3; and (b) register data. Sample L1 draws on local population registers, Samples B, M3, M4, and M5 on Central Register of Foreigners held by the Federal Office for Migration and Refugees. Finally, Samples M1 and M2 draw on employment and transfer registers of the Federal Employment Agency. Sample M1 and M2, the IAB-SOEP-Migration Studies, are a joint project of the Institute for Employment Research (IAB) and SOEP and can be accessed as a single sample by the Research Data Center of SOEP as well as the Research Data Center of the IAB. Samples M3, M4, and M5, the IAB-BAMF-SOEP Refugee Survey, can be accessed as a stand-alone file by the three partner institutes.

Random sampling with known selection probabilities allows constructing design weights, subsample-specific wave-1 cross-sectional weights, and the

2 PIAAC-L is a cooperative project of GESIS, The Leibniz Institute for Educational Trajectories (LiFBi) and the Socio-Economic Panel.

integration of new subsamples into a single sampling frame (Rendtel 1995; Schonlau et al. 2013). Wave-1 cross-sectional weights rely on three sources of information: properties of the sampling design of the gross sample, estimated non-response probabilities in the gross-sample of contacted households, and post-stratification of the net sample of participating households to official population margins. Information on the sampling designs contains the data-file "design." Wave-1 cross-sectional weights are available at the level of households as well as individuals.[3]

Longitudinal weights at the household and individual level represent the estimated non-response probabilities between two consecutive waves. The product of cross-sectional weights at t and longitudinal weights between t and $t+1$ is the basis of cross-sectional weighting in $t+1$. Further, in waves $t+1$ and following, cross-sectional weights are post-stratified to official population margins of that year.

SOEP uses a very general concept of following rules of households in the longitudinal perspective (Kroh et al. 2008). That is, if members of an originally sampled household leave the household, for instance, because of a divorce or children forming their own household, both the original as well as the split-household are interviewed, remaining part of the integrated weighting scheme. The comprehensive following rules, which cover all persons who (even temporarily) lived in SOEP households, is a comparative advantage of SOEP compared to other household panel surveys, as they allow users to track various forms of household dynamics and their implications at the household and individual level (Schonlau et al. 2011).[4]

2.2 Fieldwork, survey modes, and questionnaires

SOEP is centered on the analysis of the life course with objective and subjective indicators of well-being. Core topics are household demography and population, education and qualification, occupation and employment, earnings and working-time, housing and rent, physical and mental health, as well as subjective indicators on attitudes, values and personality (see Richter et al. 2017). Annual questions on migration and integration, with extended information for all non-German

3 For instance, the variable "*q*hhrfe" in data-file "hhrf" is the wave-1 cross-sectional weight of households of Sample *E*, that were first interviewed in wave *Q* (at the person level: variable "q*p*hrfe" in data-file "*p*hrf").

4 The annually updated "Documentation of Sample Sizes and Panel Attrition" (Kroh et al. 2018) provides further details on sampling design, sample size, non-response, weighting procedures, and cross-references to the subsample specific research notes.

subsamples, are also incorporated. SOEP also entails variables that link the participating households with information from the actual fieldwork, meaning that data users can measure, for example, interviewer and sampling mode effects.

SOEP uses several modes of data collection with face-to-face interviewing as the default. Originally, part of the data was recorded by an interviewer using paper-and-pencil interviews (PAPI mode). Since 1998, answers are also recorded using computer-assisted personal interviews (CAPI mode), a mode that eases the data transfer into an electronic format, an important asset, especially with the extensions of the panel starting in 2000. In the future, new modes will be integrated as they develop. We are currently testing the implementation of computer-assisted web interview (CAWI) in Samples L2 and L3. However, it is not anticipated that CAWI will replace either CAPI or PAPI; rather it will serve as an alternative that the respondents may use, much like self-administered mailed questionnaires. Thus, personal contact between respondent and interviewer will remain the core interview concept of the SOEP survey.

Information is collected on the *household*- and *individual*-levels, a feature that expands the research potential compared to a single-level approach. As an example, SOEP surveys both individual and household incomes, thus facilitating research on within-household bargaining or checks of internal consistency of the data. The surveyed information usually covers the current situation (e. g., family composition or satisfaction with life), but in some contexts it includes the past (e. g., job changes and employment biographies) and the future (e. g., expected life satisfaction in 5 years, and chance of re-employment). In addition, questionnaires on early life include retrospective questions on youth, childhood, and early childhood.

The central survey instruments are a household questionnaire, responded to by the household head, and an individual questionnaire, which each adult household member is supposed to answer. Furthermore, beginning with the 1997 wave, there are wave-specific "$lela" data-files (life course) containing the retrospective biography information as collected in the respective year.

A rather stable set of core questions is asked every year, enhanced by topical modules, and perennial rotating modules on topics such as wealth, neighborhood, family, and networks (see Table A3 in the Online Appendix; all questionnaires are also available online via the website of the SOEP Research Data Center (http://www.diw.de/soep-questionnaires)). Over time, the surveyed topics have expanded. Since 2000, there is a specific youth questionnaire for household members turning 17 in the survey year; it covers topics such as the situation at home, relationship to parents and friends, as well as job aspirations. Since 2001, psychological and health questions as well as age-specific questionnaires are integrated in SOEP. Since 2003 parents are asked about their young children and also a questionnaire was added for infants and very young children born

during the current or previous survey year. Since then, four additional questionnaires have been added for children in different age groups. In 2012, parents were asked about their children turning 10 during the current survey year for the first time.

Although SOEP predominantly relies on questionnaire-based surveying, alternative and innovative forms of data collection are also used. For instance, experiments from behavioral economics were used to measure trust, time preferences, and risk aversion of respondents; measures of hand grip strength provide a non-invasive health indicator, while qualitative interviewing of respondents from special populations were conducted by SOEP-researchers.

2.3 Data quality management

High data quality and confidence in its provision are crucial for establishing a scientific data infrastructure. Beyond the usual test routines to check data plausibility and consistency that take place after data collection, the SOEP team undertakes various efforts to ensure data quality. In fact, there are a number of institutional safeguards.

To assure high cooperation rates, not only was a close collaboration with the field institute established, but the SOEP scientific staff also conducts interviewer workshops. With monitoring mechanisms, we verify the correct work of the interviewer, like routine controls of the electronic completion of the questionnaire and eventual control questionnaires. Via a website, we also provide (potential) respondents with information about the study, selected scientific publications, reports, and newspaper articles related to "their" data.

The SOEP staff undertakes the generation and integration of the cross-sectional data in a panel structure and the provision of user-friendly ("generated") variables. This ensures that the experienced researchers who guide the data generation process can ensure that the generated data are in line with scientific demands, along with proper data documentation and guidance to users.

In contrast to data provided by official statistics or registers, SOEP's credo is "data from scientists for scientists." Beyond the data privacy act, no other legal restrictions limit the data access for scientific purposes. Of course, this also implies that *all* data has to be accessible to external researchers[5] as *early* as possible. In other words, all users have the possibility to compare the original raw data from the field with the generated data from the SOEP group. In case of

5 Of course without the access to directly individual-related information and the possible restriction in the mode of access. For example only remote access for more information.

doubt, each researcher can generate his own version of a generated or imputed variable, even if this makes a research stay at DIW Berlin necessary: sensitive data can only be accessed in-house.

In very low number of cases, the very strict plausibility checks led to the deletion of a given response by the respondents. If this is the case, the respective number is changed to the special code "-3," thus ensuring that this editing is visible to all researchers. The unchanged original files are also accessible from within the SOEP Research Data Center.

Since 2011, each new published version of SOEP data is given a Digital Object Identifier (DOI), which is a unique persistent identifier. This ensures that not only can the researcher clearly specify and cite the exact version of SOEP used in their research, but also that each user can verify that the SOEP data used is exactly the version published under the DOI by using the simultaneously published checksum of each dataset within the DOI. In other words, it is not possible to publish a changed dataset unnoticed using the same DOI, as each new SOEP version and each bug fix needs a unique new persistent identifier (for example, the bug fix of version 33 is doi:10.5684/soep.v33.1).

2.4 Structuring of datasets

The SOEP started with a basic cross-sectional data structure, SOEP-Core. SOEP-Core contains four types of datasets:

1. Files documenting the development of the sample, for example whether persons or households have been part of the interviewed sample/population in a given year.
2. Files providing generated intertemporal consistently named and coded variables to ease the combination of datasets across waves.
3. Files providing the originally surveyed data (except for very basic consistency checks). These files allow users to cross check generated and original variables as well as to construct their own variables, if needed.
4. Files providing respondents' biographies prior to their first survey year.

To reduce the complexity and ease the accessibility of the data, SOEP introduced a new data format in 2012 with pooled data over all available years: the so-called SOEPlong format. SOEPlong comprises various harmonized variables analyzable without any further data work by the user. For example, income information gathered prior to 2001 has been converted to Euros. Other categories are modified as needed to reflect changes in the questionnaire. Of course, all modifications are documented and all modified variables are also provided in their original form.

2.5 SOEP-IS – a user-tailored questionnaire and experimental module

The SOEP Innovation Sample (SOEP-IS, http://www.diw.de/soep-is), established in 2012, offers researchers the opportunity to collect data tailored for their particular research question (see Richter/Schupp 2015). In addition to containing a relatively short set of core questions, SOEP-IS incorporates user-designed survey and experimental modules (e. g., on the effects of life events on attitudes, personality, and behavior).

Every year, interested researchers can propose their projects to SOEP Survey Management. The deadline is November 30. After a few days pre evaluation, viable projects are invited to submit a formal proposal by December 31. Accepted projects are included in SOEP-IS at no additional cost to the applicant. In case of experiments, additional funding is required for financing monetary incentives. Applicants benefit from an embargo period of 12 months. Thereafter, the data is available to the entire SOEP user community as part of the regular data provisions. The SOEP-IS modules page (http://www.diw.de/soep-is-mod) provides an overview of all previous modules.

3 Augmenting the basic SOEP data

3.1 SOEP related studies

Several studies in Germany have incorporated SOEP questions to validate their results with representative sample data ("SOEP as Reference Data", see Siedler et al. 2009). Although these SOEP-Related Studies (SOEP-RS) are externally funded, they are designed and implemented in close cooperation with the SOEP team and follow the SOEP structure. This makes it possible to link the SOEP-RS datasets with either the original SOEP questionnaire (SOEP-Core) or with the SOEP-IS questionnaires, thus making it possible to analyze the data jointly. Examples of SOEP-RS studies include:

1. BASE-II – Berlin Aging Study II: BASE-II, an extension and expansion of the Berlin Aging Study (BASE), complements the analysis of cognitive development across the lifespan by including socio-economic and biological factors, such as living conditions, health, and genetic preconditions (see Bertram et al. 2014). It entails about 2,200 respondents.
2. Bonn Intervention Study (BIP) and Bremen Initiative to Foster Early Childhood Development (BRISE; http://brise-bremen.de/wissenschaft/):

BIP and BRISE provide an intervention for children and new born babies that tries to emulate a Random Control Treatment.

3. Early childhood education and care quality in the Socio-Economic Panel (K^2ID-SOEP). K^2ID-SOEP has collected new data on the quality of ECEC institutions that are attended by children below school age who are sample members of SOEP (see Spieß et al. 2018).

3.2 Linking SOEP with additional variables and data sets

3.2.1 Adding contextual or spatial information

There are various possibilities for users to augment SOEP data with complementary information i.e., users can link SOEP with context variables. For example, SOEP provides the occupational classification of all employees following the so-called International Standard Classification of Occupations (ISCO). The ISCO code allows researchers to link the SOEP with further information on job-specific tasks. The day of the respondents' interview can be used to link media information or data from, for example, Google trends, that describe the time period around the interview.

SOEP also offers diverse possibilities for spatial (http://www.diw.de/soep-regiondata) analyses. With the anonymized regional information on the residences of SOEP respondents, it is possible to link SOEP cases with spatial indicators at various spatial levels: federal states, spatial planning regions, counties, municipalities, and postal codes. Since 2000, the use of the exact geo-location is possible within a specialized secured setting at the SOEP Research Data Center (Goebel/Pauer 2014). Utilizing the geocoded information offers new possibilities to combine SOEP data with Big Data, which is growing rapidly due to cost-reducing technological advances (Internet of things, mobile devices, readers, wireless sensor networks, etc.).

Information on the federal state is contained in the standard data set. For the use of small-scale coded geographical data, a research stay at the SOEP Research Data Center located at DIW Berlin is mandatory.

3.2.2 Linkage with administrative and other survey data

It is argued that large sample sizes and low measurement error are comparative advantages of administrative over survey data, while the latter usually entail a much richer variable spectrum. Thus, combining administrative and survey data is a promising innovation.

One SOEP record linkage project is IAB-SOEP-MIG (see Eisnecker et al. 2017), funded by the Leibniz Association. It combines two IAB-SOEP Migration Samples with Integrated Employment Biographies (IEB) provided by the Institute for Employment Research. The basic motivation of the project is to gather, for the first time, in-depth information on the labor-market integration processes of migrants in Germany. In the first wave 2013, permission to link data was requested of a random sample of about 2,700 SOEP migration households. In the second subsample, started in 2015, about 1,100 migration households were asked. Permission to link survey data and register data was given by about 50 % of the samples.

A second project is the Linked Pension Insurance study (SOEP-RV), funded by the Research Network Old-Age Provision (Forschungsnetzwerk Alterssicherung). Starting in 2018, for those persons who have provided written consent, social-security biographies will be record-linked with their unique social security number to SOEP. The biographies encompass both the active and retirement phases, containing monthly-level information on earnings, social security status, etc. By combining the precise long-running information in the biographies and SOEP's broad variable spectrum, SOEP-RV opens new research avenues regarding the estimation of lifetime incomes, the cross-validation of survey and administrative data, as well as the role of social policies on individuals' life courses. The aim is to make the SOEP-RV dataset available in the research data centers of SOEP and the German Pension Insurance (FDZ-RV) starting in 2020.

A third record-linkage project is the Linked Employer–Employee Study (SOEP-LEE, see Weinhardt et al. 2017). In 2012/13, a survey of about 1,700 German employers was conducted, with establishments sampled based on address information provided by employed SOEP respondents. The information obtained from both surveys is linked in order to create a linked employer–employee data set concerning organizational context and individual outcomes. SOEP-LEE enriches the SOEP data with supplemental data about the workplace and the employees' working conditions, thus permitting researchers to investigate organizational impact on social inequalities and the individual development of the life course. SOEP-LEE is available at the SOEP Research Data Center and the Research Data Center for Business and Organization Data at DIW Berlin.

3.2.3 Analyzing SOEP with survey data around the world

SOEP data also form the German part of comparative international data infrastructures. One key infrastructure is the Cross National Equivalence File (CNEF), located at Ohio State University, that provides harmonized cross-national variables for panel surveys from Australia, Canada, Germany, Great Britain, Korea,

Russia, Switzerland and the United States. A reference dataset cross-links each of the individual studies, thus facilitating cross-national comparisons.

SOEP also contributes to the Cross-National Data Center in Luxembourg. The Center acquires cross-sectional household micro data on income, wealth, employment, and demographics from many high- and middle-income countries. After harmonization to enable cross-national comparisons, these data are made publicly available in two databases: the Luxembourg Income Study Database (LIS) and the Luxembourg Wealth Study Database (LWS). Today, LIS provides data for 36 countries and LWS for 15 countries.

In sum, the CNEF and LIS/LWS frameworks enable researchers to conduct cross-national comparative research without requiring the substantial harmonization efforts that would be required if the original national datasets were used.

4 Data access and user support

The SOEP scientific use file with anonymized microdata is made available free of charge to universities and research institutes for research and teaching purposes around the world in various data formats. Interested users must sign a user contract (http://www.diw.de/soep-contractmanagement) and, after approval, the data can be downloaded from the website via a secure data transfer system.

SOEP offers different forms of user support:

1. SOEP hotline (soepmail@diw.de) with instant services related to user-contracts, data distribution, as well as support for general and specialized questions on data structure and data analysis.
2. Paneldata.org (https://paneldata.org/) is SOEP's documentation system that provides basic information on each variable. Item-correspondence tables indicate changes in variable names and/or value labels across time. A script-generator produces syntax files for standard software programs to combine and generate datasets.
3. SOEP-in-Residence (https://www.diw.de/en/diw_02.c.222617.en/soep_in_residence.html) provides SOEP users the opportunity for research stays at SOEP at DIW Berlin to discuss data matters and research projects with the SOEP team. Since 2017, European researchers can apply for visitation grants via the EU's InGRID-2 project (http://www.inclusivegrowth.eu/visiting-grants).
4. SOEPcampus (http://www.diw.de/soepcampus) is a modular training program that familiarizes SOEP users with the data. Campus events take place at the SOEP offices in Berlin, at German universities, or as pre-conference workshops (in cooperation with other household-panels).

5 Science impact

It is a challenge to describe all of the scientific contributions made possible with the SOEP in a comprehensive manner, as SOEP provides the empirical basis for researchers in disciplines as diverse as economics and sociology; political sciences and psychology; demography and gerontology; transportation, architecture and city planning; nutrition and dietetics; as well as genetics and neuro science.

As of the end of 2017, SOEP had about 3,500 users worldwide, with about 50 percent resident in Germany. As seen in Figure 1, the yearly number of SOEP-based publications amounts to between 300 and 400 annually, thereof about 25 percent in (S)SCI journals. SOEP-data is used in various internationally recognized studies, including OECD reports on the development of income inequality in OECD countries (see OECD 2008, 2011, 2015), or Education at a Glance. SOEP is also an integral database for official government reports in Germany, including the German Federal Government's 5th Report on Poverty and Wealth (Federal Ministry of Labour and Social Affairs 2017) and the report of the German Federal Government on Wellbeing in Germany (Federal Press Office 2016).

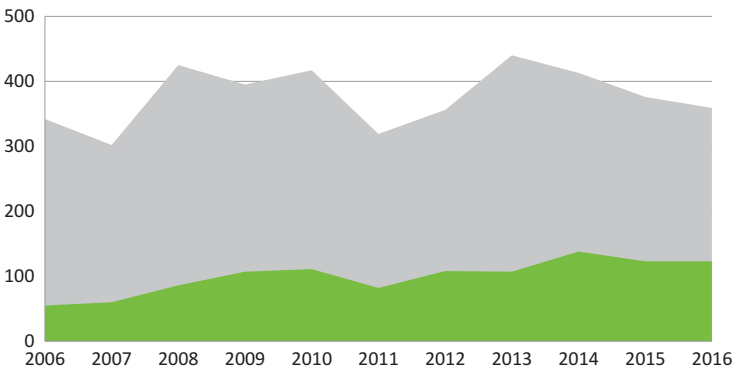

Figure 1: Number of publications using SOEP data.
Note. Grey area: number of publications with SOEP (based) data; green: thereof in (S)SCI journals.

Major areas of SOEP-based research include:

1. Research on the life course. SOEP's long time-series allow the construction of family and individual life-course profiles – from birth to death and across generations. Life course research seeks to understand how events early in life (e. g., welfare receipt or early human capital investment) affect outcomes later in life (e. g., career or life expectancy), controlling for observed and unobserved characteristics.

2. Research on inequalities and mobility. SOEP provides time-series of multiple measures of individual and household economic well-being (various types of income and wealth). As an example, since the start, annual disposable household income measures to Canberra Group standards, derived by aggregating over income source reports of household members, are routinely produced. With individual-level information on real and financial wealth positions, SOEP is an exceptional data source to study the within-household wealth accumulation processes.
3. Psychological outcomes and attitudes. Self-assessments of satisfaction in various life domains together with personality and risk attitudes pave the way to a better understanding of behavioral aspects in household decision making and of the role for "hard" factors, like income, for subjective well-being. Collecting such psychological items was innovative and remains a distinctive feature of SOEP.
4. Migration. From the beginning, SOEP has given high priority to the adequate coverage of specific groups by oversampling, starting with "guest workers" in the 1980s. The data allow a comprehensive scientific assessment of the integration processes in various domains from social networks to employment.
5. Transition to a unified Germany: The SOEP is the only database worldwide in which political unification of a society that had been divided for 40 years took place during the course of the study. In June 1990, soon after the fall of the Wall, the first wave of the East sample was collected – one month ***before*** the currency, economic, and social union. The processes of transformation and adaptation initiated by the fall of the Wall have still not concluded and it will still take generations for living conditions to reach parity in East and West Germany. Many publications address those challenges.

6 Concluding remarks

SOEP sets new national and international standards in the conception, design, implementation, and user-friendly preparation and distribution of household panel data and related data. It strives to lead the field internationally in the quality, originality, significance, and rigor of its work. Together with its international counterparts, SOEP provides not just an indispensable empirical foundation to describe longitudinal phenomena and relationships, but also a better understanding of socioeconomic processes and behavior, thereby better informing policy makers.

Acknowledgments: We thank Sandra Bohmann, Michaela Engelmann, Adam Lederer, and Uta Rahmann for their efforts and support.

Funding: The SOEP is funded by the Federal Ministry for Education and Research (BMBF) and the state governments under the umbrella of the Leibniz Association.

References

Bertram, L., A. Böckenhoff, I. Demuth, S. Düzel, R. Eckardt, S.-C. Li, U. Lindenberger, G. Pawelec, T. Siedler, G.G. Wagner, E. Steinhagen-Thiessen (2014), Cohort Profile: The Berlin Aging Study II (BASE-II). International Journal of Epidemiology 43: 703–712.

Deaton, A. (2015), Nobel Prize Lecture by Angus Deaton. Available under:http://www.nobelprize.org/mediaplayer/index.php?id = 2585

Eisnecker, Ph., K. Erhardt, M. Kroh, P. Trübswetter (2017), The Request for Record Linkage in the IAB -SOEP Migration Sample. SOEP Survey Papers 291, C.

Federal Ministry of Labour and Social Affairs (2017), The German Federal Government's 5th Report on Poverty and Wealth. Bonn.

Federal Press Office (2016), Government Report on wellbeing in Germany. Berlin.

Goebel, J., B. Pauer (2014), Datenschutzkonzept zur Nutzung von SOEPgeo im Forschungsdatenzentrum SOEP am DIW Berlin. Zeitschrift für amtliche Statistik Berlin-Brandenburg 3: 42–47.

Kroh, M., R. Pischner, M. Spiess, G. Wagner (2008), On the Treatment of Non-Original Sample Members in the German Household Panel Study (SOEP): Tracing, Weighting, and Frequencies. Methoden, Daten, Analysen. Zeitschrift Für Empirische Sozialforschung 2: 179–198.

Kroh, M., R. Siegers, S. Kühne (2015), Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP) in: J. Schupp, C. Wolf (Hrsg.), Nonresponse Bias: Qualitätssicherung sozialwissenschaftlicher Umfragen. Wiesbaden, Springer VS Verlag.

Kroh, M., S. Kühne, R. Siegers, V. Belcheva (2018), SOEP-Core – Documentation of Sample Sizes and Panel Attrition (1984 until 2016). SOEP Survey Paper 480, Series C.

OECD (2008), Growing unequal? Income distribution and poverty in OECD countries. Paris.

OECD (2011), Divided We Stand: Why Inequality Keeps Rising. Paris.

OECD (2015), In It Together: Why Less Inequality Benefits All. Paris.

Rammstedt, B., S. Martin, A. Zabal, C. Carstensen, J. Schupp (2017), The PIAAC Longitudinal Study in Germany: rationale and design. Large-Scale Assessments in Education 5: 4.

Rendtel, U. (1995), Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität. Campus Verlag, Frankfurt/New York.

Richter, D., J. Rohrer, M. Metzing, W. Nestler, M. Weinhardt, J. Schupp (2017), SOEP Scales Manual. SOEP Survey Papers 423, Series C.

Richter, D., J. Schupp (2015), The SOEP Innovation Sample (SOEP IS). Schmollers Jahrbuch 135 (3): 389–399.

Schonlau, M., M. Kroh, N. Watson (2013), The Implementation of Cross-Sectional Weights in Household Panel Surveys. Statistics Surveys 7: 37–57.

Schonlau, M., N. Watson, M. Kroh (2011), Household Survey Panels: How Much Do following Rules Affect Sample Size? Survey Research Methods 5: 53–61.

Schröder, M., R. Siegers, C.K. Spieß (2013), "Familien in Deutschland" – FiD. Schmollers Jahrbuch 133(4): 595–606.

Schupp, J. (2015), Forty Years of Social Reporting and Quality of Life Reserach in Germany. 107–126 in: G. Trommsdorff, W.R. Assmann (Eds.), A Look Back and Prospects for the Future, Forschung fördern. Am Beispiel von Lebensqualität im Kulturkontext. Konstanz und München, UVK.

Siedler, T., J. Schupp, C.K. Spieß, G.G. Wagner (2009), The German Socio-Economic Panel (SOEP) as Reference Data Set. Schmollers Jahrbuch 129(2): 367–374.

Spieß, C.K., Schober, P.S., Stahl, J.F. (2018), Early Childhood Education and Care Quality in the Socio-Economic Panel (SOEP) – the K2ID-SOEP Study. Journal of Economics and Statistics (online first), doi: https://doi.org/10.1515/jbnst-2018-0001.

Wagner, G.G., J.R. Frick, J. 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.

Weinhardt, M., A. Meyermann, S. Liebig, J. Schupp (2017), The Linked Employer–Employee Study of the Socio-Economic Panel (SOEP-LEE): Content, Design and Research Potential. Journal of Economics and Statistics 237(5): 457–467.