Lüthen, Holger et al.

Working Paper

SOEP-RV: Linking German Socio-Economic Panel data to pension records

SOEPpapers on Multidisciplinary Panel Data Research, No. 1137

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Lüthen, Holger et al. (2021) : SOEP-RV: Linking German Socio-Economic Panel data to pension records, SOEPpapers on Multidisciplinary Panel Data Research, No. 1137, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/243181

1137 2021

SOEP papers
on Multidisciplinary Panel Data Research

SOEP-RV: Linking German Socio-Economic Panel data to pension records

Holger Lüthen, Carsten Schröder, Markus M. Grabka, Jan Goebel, Tatjana Mika, Daniel Brüggmann, Sebastian Ellert, Hannah Penz

DIW SOEP

SOEPpapers on Multidisciplinary Panel Data Research at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at **http://www.diw.de/soeppapers**

ISSN: 1864-6689 (online)

German Socio-Economic Panel (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: soeppapers@diw.de

SOEP-RV: Linking German Socio-Economic Panel data to pension records

Holger Lüthen[♣], Carsten Schröder[♠], Markus M. Grabka[♦], Jan Goebel[♦], Tatjana Mika[•], Daniel Brüggmann[•], Sebastian Ellert[•], Hannah Penz[♦]

Abstract: The aim of the project SOEP-RV is to link data from participants in the German Socio-Economic Panel (SOEP) survey to their individual Deutsche Rentenversicherung (German Pension Insurance) records. For all SOEP respondents who give explicit consent to record linkage, SOEP-RV creates a linked dataset that combines the comprehensive multi-topic SOEP data with detailed cross-sectional and longitudinal data on social security pension records covering the individual's entire insurance history. This article provides an overview of the record linkage project, highlights potentials for analysis of the linked data, compares key SOEP and pension insurance variables, and suggests a re-weighting procedure that corrects for selectivity. It concludes with details on the process of obtaining the data for scientific use.

Keywords: record linkage, SOEP, SOEP-RV, pension records, consent

JEL codes: C89, C18

Acknowledgement: We thank the Forschungsnetzwerk Alterssicherung (FNA), contract number 0640-FNA-P-2016-12, for financial support. We thank Deborah Anne Bowen for her careful editing of the paper.

♣ Federal Ministry for Economic Affairs and Energy.
♠ SOEP at DIW Berlin and Freie Universitaet Berlin.
♦ SOEP at DIW Berlin.
• Deutsche Rentenversicherung Bund, Research Data Centre of German Pension Insurance.

1 Introduction

Record linkage is a method for precisely matching microdata from different sources with the goal of expanding the potential of the data for research (e.g., Schnell 2014). Data linkage offers several benefits: In addition to broadening the range of variables and the observed temporal horizon, it provides opportunities for cross-validation of information and reduces the time burden on respondents. Potential complications include higher requirements for data anonymization and protection (see Künn 2015) and the need to obtain explicit consent from the individuals, households, or companies surveyed.

This paper describes the data linkage project SOEP-RV, which is being conducted by the German Socio-Economic Panel (SOEP) in partnership with the Research Data Centre of the German Pension Insurance (FDZ-RV). A shorter version of this paper is published in the Data Observer section of the Journal of Economics and Statistics (see Lüthen et al. 2021). The aim of SOEP-RV is to link the SOEP data to administrative pension records. We do so by obtaining the social security numbers of consenting SOEP respondents from the statutory pension insurance based on the individual's name, birth date, and place of birth, and using these—in adherence to the highest data security standards—for one-to-one linkage of the SOEP and pension insurance data.

The SOEP survey, established in 1984, is a multi-topic household panel study providing individual- and household-level information (see Goebel et al. 2019). Through the use of the unique identifiers assigned to SOEP respondents, SOEP data can be matched to the individuals' pension records. Of particular interest is the information on pension stocks (Rentenbestand, RTBN) and insurance accounts (Versicherungskontenstichprobe, VSKT). RTBN is a cross-sectional dataset that provides detailed information on retirees' pension accounts. VSKT is a longitudinal dataset in spell form that is comprised of an individual's insurance history from the age of 14 to 67. The linkage increases the potential for analysis of all of these datasets—for research on pensions and long-term inequality but in many other areas as well.

SOEP's main advantage is the broad set of variables it provides for the resident population of Germany,[1] both at the individual and household level, including individual relationships within and between surveyed households. SOEP also enables analysis of small-N populations through the inclusion of special samples of migrants, refugees, top-wealth individuals, and other population groups. Administrative data, on the

[1] The institutionalized population is an exception.

other hand, provide comprehensive social security information virtually without measurement error on a monthly level.

SOEP-RV expands the research potential of SOEP data in several respects: First, it broadens the range of variables available for analysis. For a number of specific pension types, the administrative data clearly exceed the SOEP's level of detail. Second, SOEP-RV extends the SOEP's biographical information beyond the time of the initial survey and provides supplementary information that can be used to fill in gaps that occurred due to nonresponse (Frick and Grabka 2005) or to correct for recall bias (Bound et al. 2001). The pension records add biographical social security information starting earliest at the age of 14 for all SOEP respondents, including those who are new to the SOEP. Importantly, most individuals who are currently exempt from mandatory insurance (such as civil servants and the self-employed) have one or more previous periods in their biography that were relevant to the pension insurance (e.g., periods of military service or enrollment in higher education). Since pension records contain this information, these data offer an enhancement to the SOEP data. Third, the linked data allow cross-validation of information in both datasets. Fourth, the administrative insurance biographies are set up as spell data, whereas the SOEP data (with the exception of the retrospective biographies) provide measurements at specific points during the year. SOEP-RV also expands the potential for research with the administrative data: The individual-level information in the administrative data complements the SOEP's detailed information on family and household relationships. In sum, SOEP-RV is especially useful for describing and explaining the employment, pension, and income biographies of individuals and households. Furthermore, it allows the quantification of (lifetime) income—at the individual and household level—while taking into account earned and pension income as well as other types of income, including capital income and government benefits, without the need to rely on strong modelling assumptions.[2] Analogous possibilities arise for the measurement of wealth according to asset types, including pension entitlements, as well as debts.

Against this backdrop, the SOEP-RV data provide, amongst others, an important basis for the empirical analysis of two highly socially relevant topics: population aging and the stability of social security systems. In recent decades, numerous reforms have been enacted to ensure the financial sustainability and stabilization of pension contributions (Steffen 2020) by raising the retirement age, limiting pension growth, and promoting private pension schemes. With ongoing population aging and especially the imminent retirement of baby boomer cohorts, the pressure for reforms will likely continue. At the same

[2] Previous research tried to answer such questions by statistical matching of survey data with pension records see e.g., Rasner et al. (2013).

time, Germany's labor market has been changing constantly since reunification, with increasing shares of atypical jobs, a growing low-wage sector, rising labor mobility, and more career interruptions, but also more jobs. All of these developments can have significant effects on pension entitlements (Bönke et al. 2015; Westermeier et al. 2017). The interplay between population aging and changing career trajectories challenges the state's efforts at promoting earnings growth and reducing the risk of poverty. It raises questions about the role of the welfare state both during working life and after retirement. In general, the German welfare system works best for individuals who work full-time without interruptions up to retirement: They benefit not only from stable employment but also from good pension prospects. Individuals with career interruptions or precarious work contracts often suffer from both low wages and poor pension prospects. Finally, people's financial situation in retirement depends not only on their statutory pension entitlements but also on any company pensions, private pensions, or other income sources to which they or other household members may be entitled. The comprehensive information contained in the SOEP-RV data should help to advance the research on all these and other important issues.

The remainder of this article is structured as follows: Chapter 2 describes the linked data sources in detail. Chapter 3 deals with potential selectivity of the linked SOEP population. This could be an issue for two reasons: a) by design, as some SOEP subsamples have not been asked for consent (Section 3.1); b) consent to data linkage may vary systematically with respondent characteristics (Section 3.2). The section also proposes a reweighting procedure to correct for selectivity based on observables (Section 3.3). Chapter 4 provides basic comparisons of key variables that are contained in SOEP and the administrative data to illustrate specific characteristics of the data. Furthermore, it provides evidence on the selectivity of the linked population with the base population in the administrative data source. Chapter 5 explains the process of data access. Chapter 6 concludes.

2 Linked Data Sources: SOEP, RTBN and VSKT

SOEP: The SOEP is an ongoing longitudinal survey of private households in Germany that has been running since 1984 (Goebel et al. 2019). Various refresher and supplementary samples have been added over time, including an East German sample in 1990 and a number of special migrant samples. Since 2010, the SOEP has surveyed more than 25,000 individuals annually. Participation in the survey is voluntary; nevertheless, the annual re-survey rates are very high, averaging about 94 percent over many years. Of

the approximately 12,500 individuals in the first two samples from 1984, about 3,500 were interviewed in 2015.

SOEP's survey is interdisciplinary, covering a broad set of individual and household-level variables including socioeconomic status, political attitudes, psychological and health indicators, satisfaction and worries, expectations, family background, and education. Further, SOEP includes information on age, employment and retirement status, income types (including pensions), and various assets and debt components. Overall, these variables provide a very detailed picture of employment and retirement histories at both the individual and household level, with extensive research potential, especially on aging and retirement (Schröder et al. 2020).

RTBN: The RTBN is an administrative dataset containing all monthly pension payments paid out by German Pension Insurance in December of a given year.[3] Every observation represents one pension and distinguishes between old-age pensions and survivor or invalidity pensions.[4] For each pension, in addition to the amount, type, and exact starting point, the data include a range of important information, such as deductions for early retirement or premiums for postponing retirement (Lüthen 2016).[5] The RTBN thus offers detailed information complementing the SOEP, allowing researchers insight into questions such as precisely how and when individuals make the decision to retire, what deductions they were willing to accept, and whether the retirement decision was made due to poor health.[6] Further, the data can proxy time of death and provide new avenues for mortality research. Last, the RTBN includes survivor pensions, which allows researchers to derive the lifetime income of the deceased partner (Haan et al. 2020). However, SOEP-RV cannot directly link data on survivors' pensions, although it collects information on the existence of such pensions if the deceased individual agreed to participate in SOEP-RV prior to his/her death. [7]

[3] DRV Bund: https://statistik-rente.de/drv/extern/rente/documents/RTBN_Renten_nach_SGB_VI_und_sonstige_Renten_Gesamtueberblick.pdf [accessed on January 26, 2021].

[4] In accordance with SGB VI, the RTBN includes all pension types. For SOEP-RV, the most relevant pension types are invalidity pensions, all types of old-age pensions (e.g., disability, old age, unemployment, (very) long-term insured) and survivor's pensions.

[5] For more information, see the code plan of RTBN 2018: http://forschung.deutsche-rentenversicherung.de/FdzPortalWeb/getRessource.do?key=pufrtbn18xvsbb_cdpln.pdf.

[6] For research based on the RTBN, see Haan et al. (2020).

[7] The pension insurance stores survivors' pensions under the deceased person's social security number. Since we are unable to ask for consent here, we cannot retrieve the respective pensions.

VSKT: To calculate pension entitlements, the German Pension Insurance carefully collects information on all contributors' earnings histories. The VSKT is the statistical image of these records. For each month between the ages of 14 and 67, the VSKT provides a monthly history covering employment, unemployment, sick leave, and earnings points, which are used to compute monthly gross earnings. Due to its biographical nature and monthly detail level, the original VSKT sample is frequently used in economic research, for instance, for studies on long-term inequality in lifetime earnings (Bönke et al. 2015) and for research on old age (e.g., Lüthen 2016; Geyer and Welteke 2019). The biographical nature of the VSKT serves as a blueprint for SOEP-RV: If an individual gives consent to SOEP-RV, their biographies are retrieved from pension records in the VSKT format. This is even true for the already retired population. Therefore, SOEP-RV provides a unique possibility for analyzing the entire biographies of the resident population of Germany.

3 Consent and Selectivity

SOEP respondents were asked to consent to data linkage in 2018. Recently integrated new subsamples were exempted from this to reduce the risk of panel attrition. In subsequent waves, the SOEP has made an effort to link these originally exempted individuals as well as SOEP respondents who were too young to give consent in 2018. However, until all of the SOEP samples have been asked for consent, the *SOEP population asked for consent* constitutes a *subsample of the overall SOEP adult population*. The *SOEP population asked for consent* makes up 14,966 respondents (see Appendix A.1). Of those, 8,141 respondents (54.4%) gave consent and thus constitute the *consenting population*. This percentage of respondents consenting is in line with similar record linkage projects in Germany.[8]

The SOEP is equipped with survey weights that allow researchers to draw inferences about the base population: individuals living in non-institutionalized households in Germany. However, since the linked population is a subsample of SOEP's adult population (see Figure 1), the question of selectivity naturally arises. We investigate selection with respect to observable characteristics in two steps: First, we use a multivariate logit model to investigate differences in the characteristics of the *adult population* and the

[8] SHARE-RV has a quota of 55% (http://www.share-project.org/special-data-sets/record-linkage-project/share-rv.html). SHARE-RV also links survey data to administrative pension data in Germany and constitutes the most comparable data research project.

population asked for consent. In the second step, we study differences between the *population asked for consent* and the *consenting population*.

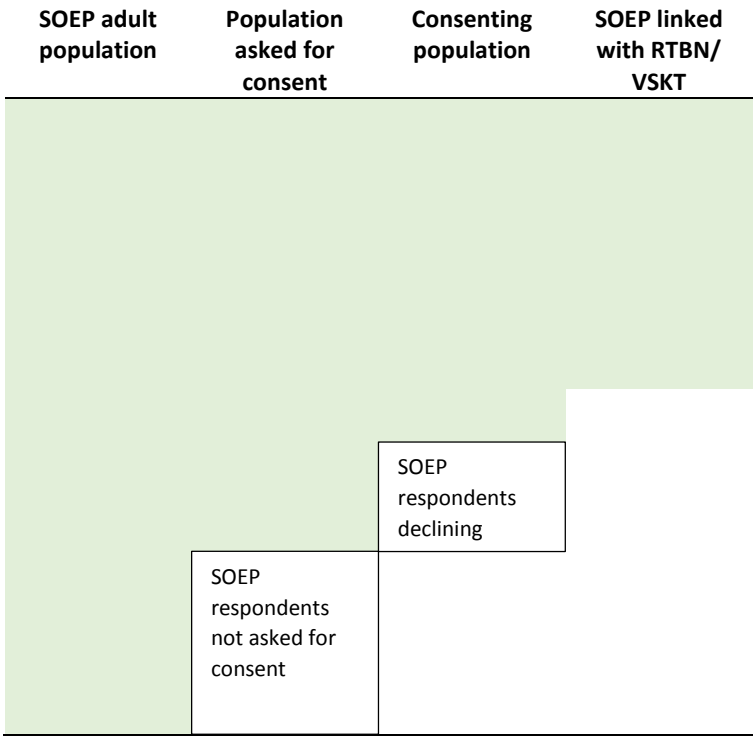


Figure 1: Consent process in SOEP

Our choice of explanatory variables in the multivariate models builds on evidence from comparable record linkage projects (e.g., Jenkins et al. 2006). In the following, we present the variables and briefly review exemplary previous evidence:

1. *Age*. Most studies show that consent decreases with age (Pascale 2011; Sakshaug et al. 2012a; Wahrendorf 2018; Weissman et al. 2016). In our case, closely related to age is *salience*. Here, individual knowledge about the nature of the linked data might influence consent. This implies that individuals who are about to retire may have different consent rates as they are well informed about their pension entitlements (Korbmacher and Schröder 2013).
2. *Health.* Physical limitations might negatively affect people's willingness to share their social security number (Jenkins et al. 2006).
3. *Gender.* Most studies find no gender-driven differences in the willingness to provide consent (Jenkins et al. 2006; Mostafa and Wiggins 2017).

4. *Migration background.* Migrants are usually found to be less likely to provide consent (Carter et al. 2010; Cruise et al 2015; Sakshaug et al. 2016).
5. *Place of residence.* Previous studies suggest differences between East and West Germany, with East Germans exhibiting higher consent rates (Antoni 2011; Coppola and Lamla 2012; Hartmann and Krug 2009; Korbmacher and Schröder 2013).
6. *Education.* Mixed evidence: Whereas Carter et al. (2010) and Knies and Burton (2014) find a positive correlation, Kim et al. (2015) and Sakshaug et al. (2016) find negative effects. Others find different effects for particular levels of education or educational attainment (Beste 2011; Dahlhamer and Cox 2007; Hartmann and Krug 2009; Jenkins et al. 2006)
7. *Income.* Mixed evidence: Some studies suggest a positive correlation (Carter et al 2010; Hartmann and Krug 2009; Huang et al 2007; Mostafa and Wiggins 2017). Others find a higher consent rate for low incomes (Kim et al 2015; Weissman et al. 2016), middle incomes (Coppola and Lamla, 2012), high incomes (Sakshaug et al. 2012b), or no relationship (Antoni 2011; Knies and Burton 2014; Korbmacher and Schröder 2013).
8. *Household composition.* Mixed evidence: Some studies indicate different consent rates across varying household compositions; others document different effects (Carter et al 2010; Coppola and Lamla 2012).
9. *Homeownership.* Homeowners show repeatedly lower consent rates (Cruise et al. 2015; Yang et al. 2019).

3.1 Selectivity of the SOEP Population Asked for Consent

As explained in the previous section, in no wave of SOEP-RV will all respondents in every possible subsample be asked for consent. The most important reason is that asking for consent potentially lowers the willingness of new SOEP respondents to participate. After individuals have taken part in several waves, enough trust has been established for SOEP to ask for consent to record linkage. Hence, it will always be important to analyze who was asked for consent before analyzing the willingness to provide consent. Since SOEP survey weights are constructed for the entire SOEP, controlling for subsample participation by adjusting the survey weights helps in avoiding selectivity bias. This is especially true for the first waves of SOEP-RV: As this is the first time we have implemented the linkage procedure, the aim was to phase-in the linkage of further subsamples consecutively over time and focus on the oldest samples.

To explain statistically who in the adult SOEP population was asked for consent (see column 2 in Figure 1), we use a logistic regression and show the results in terms of marginal effects in Table 1.[9] We use the explanatory variables described above. Of course, depending on the research question, this list may need to be adapted, for example, when it comes to analyses by nationality.

The reference group in the regressions is male respondents of age below 40 with a household post-government income in the bottom quintile, whose health and education is lower than medium; who have no migration background, and who are living in a 1-member household. In the first wave of SOEP-RV, we find that there was a higher probability of being asked for consent among older individuals and individuals with higher incomes. We also find higher probabilities for singles without children, people with medium education, and homeowners. The probability of being asked for consent was lower for respondents with direct (first generation) or indirect (second or third generation) migration backgrounds. Furthermore, the probability of being asked for consent was lower for all types of household combinations in comparison to a single household. The results with respect to migration background, income, and household composition are not surprising, as the first wave of SOEP-RV did not include most of the migration subsamples and only part of the subsamples of low-income families.

Table 1: Marginal effects after logistic regression: Who was asked for consent?

Variables	Margins	SE
Age: 40-49	-0.010	(0.009)
Age: 50-59	-0.021**	(0.009)
Age: 60-69	0.073***	(0.011)
Age: 70-79	0.111***	(0.013)
Age: 80+	0.169***	(0.017)
Medium Health	-0.008	(0.008)
Good Health	-0.015*	(0.008)
Female	0.008	(0.006)
Direct Migration Background	-0.313***	(0.007)
Indirect Migration Background	-0.178***	(0.013)
East German	-0.010	(0.007)
Medium Education	0.034***	(0.009)
High Education	-0.021**	(0.010)
Second Income Quintile	0.082***	(0.010)
Third Income Quintile	0.113***	(0.011)
Fourth Income Quintile	0.156***	(0.012)
Fifth Income Quintile	0.164***	(0.013)
Couple Without Children	-0.053***	(0.010)
Single Parent	-0.173***	(0.013)
Couple With Children	-0.114***	(0.011)

[9] See Appendix A.2 for regression coefficients.

Multiple Generation HH	-0.220***	(0.030)
Other Combination	-0.171***	(0.027)
Homeowner	0.074***	(0.006)
Observations	23,975	
Pseudo R-squared	0.175	
Chi-square test	5,782	
Prob. > Chi^2	0.000	

Note. Own calculations based on SOEP.v36 and SOEP-RV.2018. Standard errors (SE) in parentheses. The base category for age is 18-39. The base category for household-type is 1-member household. *** p<0.01, ** p<0.05, * p<0.1.

3.2 Selectivity of the Consenting SOEP Population

In this section, we implement the variables shown and explained in Section 3.1 in a logistic regression framework to statistically explain the willingness to give consent.

Table 2 presents marginal effects of our logistic regression on consent[10]. The willingness to consent decreases in age, which is in line with evidence from other studies (e.g., Pascale 2011; Sakshaug et al. 2012a; Wahrendorf 2018; Weissman et al. 2016). We find no effects for health, gender, or income, which is in line with the ambiguous or zero effects often reported (e.g., Jenkins et al. 2006; Antoni 2011; Knies and Burton 2014; Korbmacher and Schröder 2013). Migrants and their offsprings are less willing to give consent, which constitutes a typical result (e.g., Carter et al. 2010; Cruise et al 2015; Sakshaug et al. 2016). Highly educated individuals are less likely to consent, which is in line with some studies (Carter et al. 2010; Knies and Burton, 2014). Last, in line with the literature, homeowners are less likely to give consent (Cruise et al. 2015; Yang et al. 2019). In sum, our results are in line with the overwhelming majority of the literature and further confirm the unanimous results of no substantial consent bias.

[10] Appendix A.3 shows the corresponding regression coefficients.

Table 2: Marginal effects after logistic regression: Who gave consent?

Variables	Margins	SE
Age: 40-49	-0.023	(0.015)
Age: 50-59	-0.019	(0.015)
Age: 60-69	-0.026	(0.017)
Age: 70-79	-0.080***	(0.018)
Age: 80+	-0.089***	(0.022)
Medium Health	-0.013	(0.012)
Good Health	0.019	(0.013)
Female	-0.002	(0.009)
Direct Migration Background	-0.080***	(0.015)
Indirect Migration Background	-0.078***	(0.025)
East German	0.036***	(0.011)
Medium Education	-0.029*	(0.015)
High Education	-0.047***	(0.017)
Second Income Quintile	-0.007	(0.015)
Third Income Quintile	-0.014	(0.016)
Fourth Income Quintile	-0.020	(0.018)
Highest Income Quintile	-0.024	(0.019)
Couple Without Children	0.003	(0.015)
Single Parent	-0.024	(0.023)
Couple With Children	0.008	(0.017)
Multiple Generation HH	-0.008	(0.055)
Other Combination	0.117**	(0.051)
Homeowner	-0.046***	(0.010)
Observations	12,869	
Pseudo R-squared	0.009	
Chi-square test	155.7	
Prob. > Chi^2	0.000	

Note. Own calculations based on SOEP.v36 and SOEP-RV.2018. Standard errors (SE) in parentheses. The base category for age is 18-39. The base category for household-type is 1-member household. *** p<0.01, ** p<0.05, * p<0.1.

3.3 A Reweighting Procedure to Adjust for Selectivity

There are two potential sources of selection: Selectivity of the *SOEP population asked for consent* and the selectivity of the consent among those SOEP subjects who were asked. To adjust SOEP frequency weights accordingly, we propose a four-step procedure recommended in a comparable context in Siegers et al. (2020):

Step 1: Estimation of a logistic regression model for the overall *SOEP adult population* where the dependent variable is a dummy variable indicating whether respondents were asked for consent to linkage of their SOEP data with the administrative data (dummy is equal to one) or were not asked (dummy is zero).

Step 2: If at least one explanatory variable is significant (e.g., p-value below 0.05) and at the same time shows a meaningful quantitative effect, the model is re-estimated only including the significant variables, and a correction of the SOEP survey weights is performed by multiplying the survey weights by the inverse estimated probability.

Step 3: Estimation of a logistic regression model for the population asked for consent where the dependent variable is a dummy variable indicating whether respondents consented to data linkage (dummy is equal to one) or not (dummy is zero) using the same explanatory variables as in step 1.

Step 4: If at least one explanatory variable is significant and at the same time shows a meaningful quantitative effect, the model is re-estimated only including the significant variables, and the adjusted weights from step 3 are multiplied by the inverse estimated probability.

These double-adjusted SOEP weights yield the adjusted weight that can be used to calculate population statistics.

4 Comparisons of Key Variables

4.1 Validity of Information for Linked Cases

To validate the linkage, we compare RTBN information to self-reported information for successfully linked SOEP respondents. This section also serves as a warning to read the variable descriptions in both data sources as certain differences lie in the nature of the datasets.

We compare—at the level of each linked individual—the information contained in the two datasets on gender, marital status, age, and monthly retirement payments and display the results in Table 3. Our results suggest a near perfect match for both gender and age, which supports both a successful linkage and a correct collection of age and gender in the survey and administrative data. Further, Table 3 displays that marital status information deviates for about one fifth of the sample.

A naïve interpretation would be to argue that administrative data must be valid and hence the survey data provides false information. However, the devil is in the detail. In the SOEP, respondents are asked to provide their marital status every year. By contrast, the pension insurance asks about marital status only when an individual applies for rehabilitation[11]. Therefore, (a) not everyone is asked this question, and (b) this information corresponds to a certain point in time in a persons' life.

[11] Rehabilitation comprises medical and occupational rehabilitation.

Table 3: Comparison of gender and marital status information in the RTBN and SOEP for successfully linked respondents

Variables	Consistent information	Inconsistent information
Gender	99.95%	00.05%
Age	100.00%	00.00%
Marital status:		
Single, Divorced, Widowed	92.42%	7.58%
Married	87.03%	12.97%
Missings	96.63%	3.37%
Observations	2,108	2,108

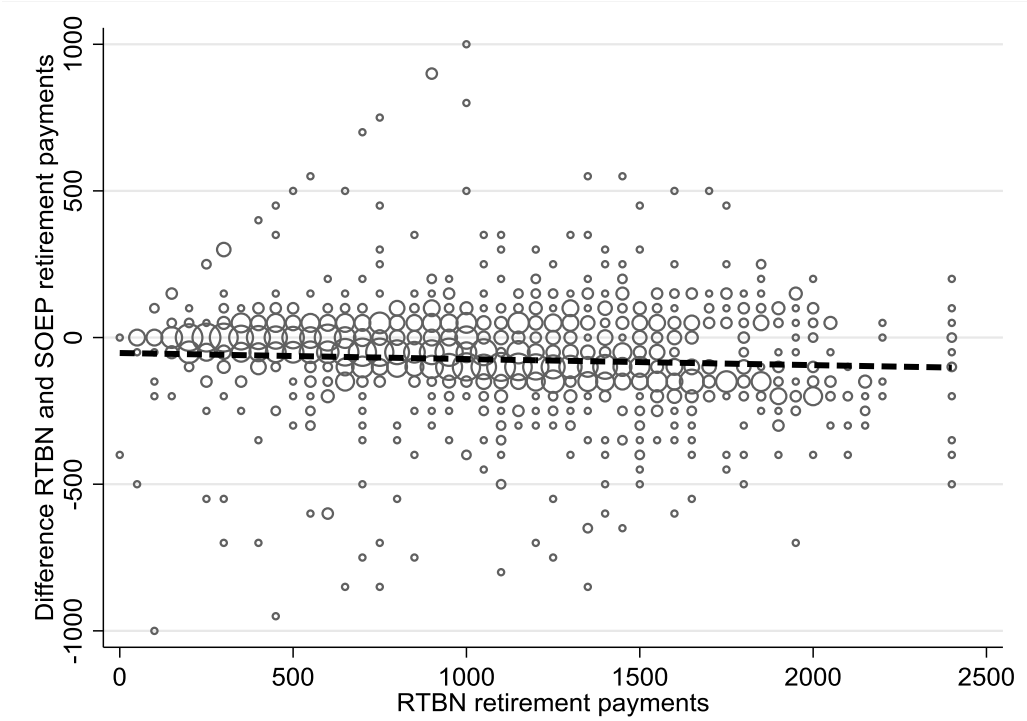
Note. Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018.

Next, we compare individual monthly retirement payments as reported in SOEP and RTBN. The results are illustrated in Figure 2 by means of a scatter plot with the number of cases underlying a particular combination of reported pay being reflected by the size of the bullets. This comparison involves several potential sources of error: First, SOEP values correspond to annual 2017 values because in 2018, the SOEP asked about the annual retirement payments received in the previous year, whereas the RTBN information represents December 2018 values. In most cases, this is the cause of very minor differences. However, some individuals who entered retirement late in 2017 reported 12-month values in the SOEP despite receiving pensions for fewer months, causing outliers: Here, we excluded four observations with extremely large SOEP retirement payments (up to €15,000 per month). All four observations had in common that they entered retirement very late in 2017 and that their self-reported pension values, when used as annual values, correspond to their (much lower) pension values in the RTBN. One approach would be to adjust these values by treating them as 12-month values. A second would be to exclude SOEP respondents who entered retirement after 2017. For the purposes of the present overview, we chose the latter.

Finally, we exclude invalidity pensions in the RTBN. Since these pensions are not awarded on a permanent basis, individuals might leave the insurance between 2017 and 2018. Hence, this temporary pension may distort the results. Still, same-year comparisons in subsequent waves of SOEP-RV improve upon all those results, for instance, when the RTBN 2018 is comparable to the SOEP 2019.[12] However, due to regularly

[12] Such a comparison warrants correction for panel attrition.

occurring adjustments to the pension scheme, some minor deviations are likely to remain even after careful adjustments.



Note. Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The sample refers to 1,787 successfully linked SOEP respondents who entered retirement in 2017 or earlier and do not receive an invalidity pension. The difference refers to RTBN-SOEP monthly retirement payments.

Figure 2: Differences in individual RTBN and SOEP monthly retirement payments

Despite the aforementioned shortcomings, Figure 2 indeed suggests that differences in retirement payments for most individuals are usually small. Many differences are close to zero, especially for lower pensions. Further, SOEP and RTBN retirement payments exhibit a positive correlation of 0.761. Nevertheless, a t-test of equal means suggests a €70 higher SOEP pension, which is significant (Table 4). Still, a t-test with bootstrapped percentiles further supports our results that deviations occur especially among individuals receiving retirement payments in the upper half. Or in other words: Individuals with retirement payments in the upper half report higher retirement payments in the SOEP than what is reported in the administrative data. This result is not surprising and actually underscores the advantages of the linked data: First, the RTBN censors monthly retirement payments of more than €2,199: Here the SOEP complements the RTBN and yields better information. Second, it is conceivable that the slight systematic upward deviation could be a result of older SOEP respondents who partially rounded up their

retirement pay or mistakenly added other pensions such as widow pensions or company pensions.[13] In these cases, the RTBN delivers more precise information.

Table 4: T-tests on equal means of the retirement payment variables

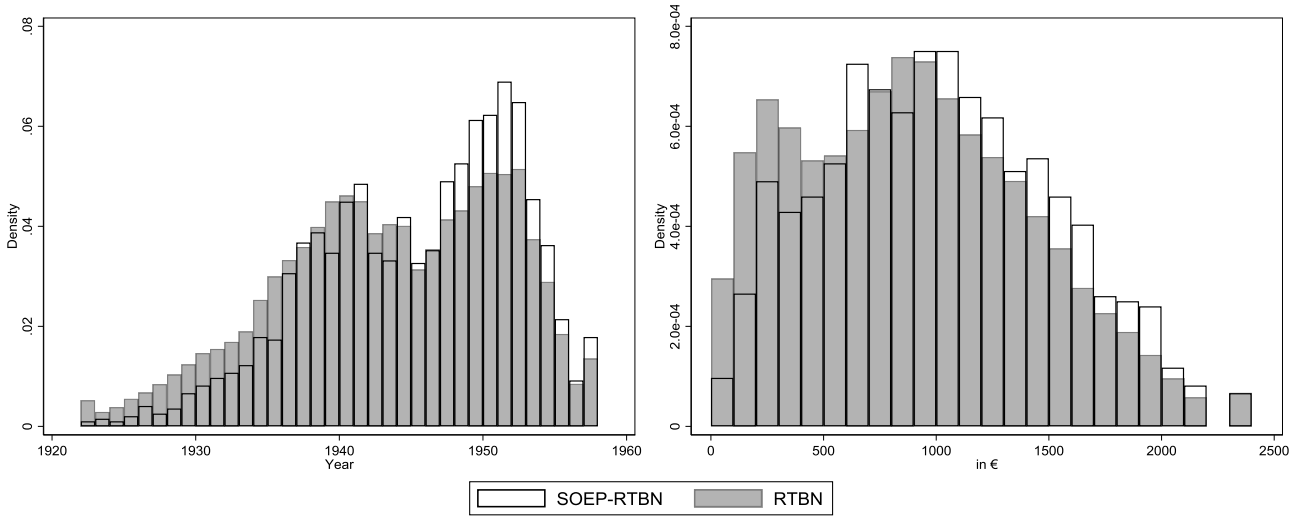
Variables		RTBN	SOEP-RTBN	Difference	SE
Retirement	Mean	1,027	1,097	-70***	10.71
Payments	P10	331	320	11	10.83
	P50	1,006	1,070	-64***	10.71
	P90	1,730	1,850	-120***	20.30
Observations		1,787	1,787		

Note. Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The difference refers to RTBN – SOEP-RTBN. The p-values result from t-tests of equal means and bootstrapped percentiles between the retirement payment variable of the RTBN and the retirement payment variable of the SOEP-RTBN for 1,787 successfully linked SOEP respondents who entered retirement in 2017 or earlier and do not receive an invalidity pension. SE are the standard errors.
*** p<0.01, ** p<0.05, * p<0.1.

4.2 Comparison of SOEP-RTBN and RTBN

To evaluate the overall representativeness of SOEP-RV for statutory pensions in Germany, we compare the RTBN variables gender, age, monthly retirement payments, and pension types for the linked SOEP-RV population to a representative 1% RTBN sample. We restrict both samples to individuals born in 1958 or earlier to ensure a proper comparison. The histograms in Figure 3 show that the SOEP-RV population is younger and receives higher pensions than the RTBN population. Further descriptive results in Table 5 confirm this pattern. Since the SOEP does not include individuals living in care facilities or comparable institutions, differences—especially for the very old—are to be expected.

[13] It cannot be ruled out that retirement income is no longer reported accurately due to the onset of dementia in old age.



Note: Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The samples refer to 1.958 successfully linked SOEP respondents in the SOEP-RTBN and a representative sample of 188,910 observations in the RTBN. Samples are restricted to individuals aged 60 or older.

Figure 3: Birth year and retirement payments

Table 5: Descriptive statistics

Variables		RTBN	SOEP-RTBN	Difference	SE
Age	Mean	75	73	2***	9.54
	P10	65	65	0	0.17
	P50	75	72	3***	0.62
	P90	86	83	3***	0.48
Retirement	Mean	904	1016	-112***	511.52
Payments	P10	226	333	-107***	15.09
	P50	878	993	-115***	13.77
	P90	1,620	1,705	-85***	26.54
Observations		188,910	1,958		

Note. Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The difference refers to RTBN – SOEP-RTBN. The p-values result from t-tests of equal means and bootstrapped percentiles between the RTBN and the SOEP-RTBN samples restricted to individuals aged 60 or older. SE are the standard errors. *** p<0.01, ** p<0.05, * p<0.1

For further insight, Table 6 shows a comparison by share of pension types. We find small and significant differences for some pension types, but the only larger deviation is found in regular old-age pensions (about 8 percentage points). Those differences stem from those older than 85, who predominantly receive old-age pensions and are underrepresented in the SOEP (Fig. 2). Table 6 also reveals another strength of the record linkage: SOEP-RV allows investigation of the household situation of certain pension recipients,

such as recipients of invalidity pensions, who experience a higher risk of poverty due to the interruptions in their employment histories.

Table 6: Chi-squared test on equal proportions

Variables	RTBN	SOEP-RTBN	Difference
Male	0.445	0.447	-0.002
Female	0.555	0.553	0.002
Invalidity pension	0.034	0.043	-0.009**
Regular old-age pension	0.404	0.325	0.080***
Unemployment/part time pension	0.108	0.112	-0.005
Old-age pension for women	0.189	0.199	-0.010
Pension for severely disabled	0.098	0.109	-0.011
Pension for long time insured	0.102	0.133	-0.030***
Pension for especially long time insured	0.063	0.080	-0.017***
Other pensions	0.002	0.000	0.001**
Observations	200,791	2,108	

Note. Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The difference refers to RTBN - SOEP-RTBN. The p-values result from chi-squared tests on equal means between the RTBN and the SOEP-RTBN samples restricted to age 60 or older. *** p<0.01, ** p<0.05, * p<0.1

5 Data Access

A central aim of this project is to make the resulting new dataset and its analysis potential available for scientific use as easily as possible and according to the FAIR criteria (Betancort et al. 2020). Both datasets—the SOEP survey data and the administrative data of the German Pension Insurance—are available only for scientific research but free of charge. All datasets are provided for use in the statistical packages Stata and SPSS. Using the data for commercial purposes is forbidden. However, because of the different data sources (survey data versus social data), users must register separately at each Research Data Center according to its access rules.

The SOEP survey data are available through the SOEP Research Data Center (RDC SOEP). After signing a data distribution contract, users can download data from all available years and subsamples with an individual download link. The link is time-limited, encrypted, and can only be used in combination with a personal password, which is sent by text message to the user's cellphone.

The administrative data are stored at and provided by the Research Data Centre of the German Pension Insurance (FDZ-RV). Data use requires registration and submission of an application form. After the registration process is completed, the data are sent to registered users on a hard disc.

The final merging of the two data sources can be done by users themselves using the stable and unique identifiers included in both datasets.

6 Research Potentials and Concluding Remarks

We have provided an overview on the SOEP-RV-project, which connects SOEP survey data to administrative pension data through record linkage, offering many new avenues for research, especially on topics that require detailed pension information or long-term biographical employment and wage information on an individual or household level.

We have also documented that using the data is not as straightforward as it may seem. Because the SOEP has phased in the request for consent to data linkage starting with long-standing samples and asking newer samples only after trust has been established through participation in several waves of the survey, and due to the selectivity in consent, use of the SOEP-RV data requires weighting to be representative. To this end, we have illustrated an exemplary re-weighting procedure. We have also examined SOEP-RV data validity and explained differences in data from the two sources. Finally, we have explained how researchers can obtain the SOEP-RV data.

The project is still ongoing. Future data waves will open up even more avenues for research on topics such as mortality, and will include even greater numbers of individuals, improving representativeness.

Bibliography

Antoni, M. (2011), Linking survey data with administrative employment data: The case of the German ALWA survey. FDZ-Methodenreport.

Beste, J. (2011), Selektivitätsprozesse bei der Verknüpfung von Befragungs- mit Prozessdaten: Record Linkage mit Daten des Panels „Arbeitsmarkt und soziale Sicherung“ und administrativen Daten der Bundesagentur für Arbeit. FDZ Methodenreport 201109_de, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.

Betancort, C. N., Bongartz, E. C., Dörrenbächer, N., Goebel, J., Kaluza, H. Siegers, P. (2020), White Paper on Implementing the FAIR Principles for Data in the Social, Behavioural, and Economic Sciences. RatSWD Working Paper Series. doi: 10.17620/02671.60.

Bönke, T., Giacomo, C., and Lüthen, H. (2015), Lifetime Earnings Inequality in Germany. Journal of Labor Economics 33(1): 171–208.

Bound, J., Brown, C., Mathiowetz, N. (2001), Chapter 59: Measurement Error in Survey Data. Pp. 3705-3843 in: J. Heckman, E. Leamer (eds.), Handbook of Econometrics 5. Chicago/London.

Carter, K., Shaw, C., Hayward, M., Blakely, T. (2010), Understanding the determinants of consent for linkage of administrative health data with a longitudinal survey. New Zealand Journal of Social Sciences Online 5(2): 53-60.

Coppola, M., Lamla, B. (2012), Empirical Research on Households Saving and Retirement Security: First Steps towards an Innovative Triple-Linked-Dataset. MEA Discussion Paper Series 201207.

Cruise, S. M., Patterson, L., Cardwell, C., O'reilly, D. P. (2015), Large panel-survey data demonstrated country-level and ethnic minority variation in consent for health record linkage. Journal of Clinical Epidemiology 68(6): 684-692.

Dahlhamer, J.M., Cox, C.S. (2007), Respondent consent to link survey data with administrative records: Results from a split-ballot field test with the 2007 National Health Interview Survey. Proceedings of the Federal Committee on Statistical Methodology Research Conference. https://nces.ed.gov/FCSM/pdf/2007FCSM_Dahlhamer-IV-B.pdf. Accessed: 07.05.2021

Frick, J., R., Grabka, M. (2005), Item-Non-Response on Income Questions in Panel surveys: Incidence, Imputation and the Impact on the Income Distribution. Allgemeines Statistisches Archiv (ASTA) 89(1): 49-61.

Goebel, J., Grabka, M., Liebig, S., Kroh, M., Richter, D., Schröder, C., Schupp, J. (2019), The German Socio-Economic Panel (SOEP). Jahrbücher für Nationalökonomie und Statistik 239(2): 345-360.

Haan, P., Kemptner, D., Lüthen, H. (2020), The rising longevity gap by lifetime earnings – Distributional implications for the pension system. Journal of the Economics of Aging 17, https://doi.org/10.1016/j.jeoa.2019.100199.

Hartmann, J., Krug, G. (2009), Verknüpfung von personenbezogenen Prozess- und Befragungsdaten – Selektivität durch fehlende Zustimmung der Befragten? Zeitschrift für Arbeitsmarktforschung: 121-139.

Huang, N., Shih, S. F., Chang, H. Y., Chou, Y. J. (2007), Record linkage research and informed consent: who consents? BMC Health Services Research 7(18). https://doi.org/10.1186/1472-6963-7-18.

Jenkins, S. P., Cappellari, L., Lynn, P., Jäckle, A., and Sala, E. (2006), Patterns of consent: evidence from a general household survey. Journal of the Royal Statistical Association 169(4): 701-722.

Kim, J., Shin, H., Rosen, Z., Kang, J., Dykema, J., Muenning, P. (2015), Trends and Correlates of Consenting to Provide Social Security Numbers: Longitudinal Findings from the General Social Survey (1993–2010). Field Methods 27(4): 348-362.

Knies, G., Burton, J. (2014), Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys. BMC medical Research Methodology 14: 14-125.

Korbmacher, J. M., Schröder, M. (2013), Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer. Survey Research Methods 7(2): 115-131.

Kroh, M, Siegers, R., Kühne, S. (2015), Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP), Pp. 409–444 in: J. Schupp, C. Wolf (eds.), Nonresponse Bias: Qualitätssicherung Sozialwissenschaftlicher Umfragen. Springer Fachmedien Wiesbaden. Wiesbaden.

Künn, S. (2015), The challenges of linking survey and administrative data. IZA World of Labor 2015: 214.

Lüthen, H. (2016), Rates of Return and Early Retirement Disincentives: Evidence from a German Pension Reform. German Economic Review 17(2): 206-233.

Lüthen, H., Schröder, C., Grabka, M., Goebel, J., Mika, T., Brüggmann, D., Ellert, S., Penz, H., (2021), SOEP-RV: Linking German Socio-Economic Panel data to pension records. Journal of Economics and Statistics (online first). https://doi.org/10.1515/jbnst-2021-0020.

Mostafa, T., Wiggins, R. D. (2017), What influences respondents to behave consistently when asked to consent to health record linkage on repeat occasions? International Journal of Social Research Methodology 21(1): 119-134.

Pascale, J. (2011), Requesting Consent to Link Survey Data to Administrative Records: Results from a Split-Ballot Experiment in the Survey of Health Insurance and Program Participation (SHIPP). Survey Methodology.

Rasner, A., Frick, J. R., Grabka, M. (2013), Statistical Matching of Administrative and Survey Data, an application to wealth inequality analyses: Sociological Methods and Research 42: 192-224.

Sakshaug, J. W., Kreuter, F. (2012a), Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data. Survey Research Methods 6(2): 113-122.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., Weir, D. R. (2012b), Linking Survey and Administrative Records: Mechanisms of Consent. Sociological Methods & Research 41(4): 535-569.

Sakshaug, J. W., Huber, M. (2016), An Evaluation of Panel Nonresponse and Linkage Consent Bias in a Survey of Employees in Germany. Journal of Survey Statistics and Methodology 4(1): 71-93.

Schröder, C., König, J., Fedorets, A., Goebel, J., Grabka, M., Lüthen, H., Metzing, M., Schikora, F., Liebig, S. (2020), The Economic Research Potentials of the German Socio-Economic Panel Study, German Economic Review 21 (3): 335-371.

Schnell, R. (2014), Linking Surveys and Administrative Data. Pp. 273-287 in: U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, S. P. (eds.): Improving Surveys Methods: Lessons from Recent Research. Routledge, Taylor & Francis Group, New York.

Siegers, R., Steinhauer, H. W., Zinn, S. (2020), Gewichtung der SOEP-CoV-Studie 2020, SOEP Survey Papers, Series C – Data Documentation, 888.

Steffen, J. (2020), Sozialpolitische Chronik. http://www.portal-sozialpolitik.de/uploads/sopo/pdf/Sozialpolitische-Chronik.pdf.

Wahrendorf, M., Marr, A., Antoni, M., Pesch, B., Jöckel, K.-H., Lunau, T., Moebus, S., Arendt, M., Brüning, T., Behrens, T., Dragano, N. (2018), Agreement of Self-Reported and Administrative Data. European Journal of Population 35(2): 329-346.

Weissman, J., Parker, J. D., Miller, D. M., Miller, E. A., Gindi, R. M. (2016), The relationship between linkage refusal and selected health conditions of survey respondents. Survey Practice 9(5). https://doi.org/10.29115/SP-2016-0028.

Westermeier, C., Grabka, M., Jotzo, B., Rasner, A. (2017), Veränderung der Erwerbs- und Familienbiografien lässt einen Rückgang des Gender-Pension-Gap erwarten. DIW Wochenbericht 12: 235-243.

Yang, D., Fricker, S., Eltinge, J. (2019), Methods for Exploratory Assessment of Consent-to-link in a Household Survey. Journal of Survey Statistics and Methodology 7(1): 118-155. https://doi.org/10.1093/jssam/smx031.

Appendix A

Table A.1: Observations in the SOEP consent process

SOEP population 2018	Observations
Adult population	30,324
Population asked for consent	14,966
Consenting population	8,141
Population linked to the RTBN	2,121

Note. Own calculations based on SOEP.v36 and SOEP-RV.2018.

Table A.2: Regression coefficients after logistic regression: Who was asked for consent?

Variables	Coefficient (SE)
Age: 40-49	-0.053
	(0.044)
Age: 50-59	-0.107**
	(0.047)
Age: 60-69	0.377***
	(0.056)
Age: 70-79	0.577***
	(0.066)
Age: 80+	0.874***
	(0.088)
Medium Health Condition	-0.039
	(0.044)
Good Health Condition	-0.078*
	(0.043)
Female	0.040
	(0.030)
Direct Migration Background	-1.623***
	(0.042)
Indirect Migration Background	-0.922***
	(0.068)
East German	-0.052
	(0.037)
Medium Education	0.174***
	(0.048)
High Education	-0.109**
	(0.053)
Second Income Quintile	0.427***
	(0.054)
Third Income Quintile	0.585***
	(0.058)
Fourth Income Quintile	0.808***
	(0.062)

Highest Income Quintile	0.849***
	(0.067)
Couple Without Children	-0.273***
	(0.053)
Single Parent	-0.900***
	(0.069)
Couple With Children	-0.593***
	(0.056)
Multiple Generation HH	-1.141***
	(0.155)
Other Combination	-0.887***
	(0.141)
Homeowner	0.382***
	(0.034)
Constant	0.176**
	(0.078)
Observations	23,975
Pseudo R-squared	0.175
Chi-squared test	5,782
Prob. > Chi^2	0.000

Note. Own calculations based on SOEP.v36 and SOEP-RV.2018.
Standard errors (SE) in parentheses. The base category for age is 18-39.
The base category for household-type is 1-member household.
*** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Regression coefficients after logistic regression: Who gave consent?

Variables	Coefficient (SE)
Age: 40-49	-0.095
	(0.061)
Age: 50-59	-0.077
	(0.062)
Age: 60-69	-0.105
	(0.069)
Age: 70-79	-0.329***
	(0.075)
Age: 80+	-0.364***
	(0.090)
Medium Health Condition	-0.052
	(0.051)
Good Health Condition	0.079
	(0.051)
Female	-0.007
	(0.037)
Direct Migration Background	-0.329***
	(0.062)
Indirect Migration Background	-0.317***
	(0.102)

East German	0.145***
	(0.044)
Medium Education	-0.119*
	(0.063)
High Education	-0.191***
	(0.070)
Second Income Quintile	-0.029
	(0.061)
Third Income Quintile	-0.058
	(0.067)
Fourth Income Quintile	-0.081
	(0.073)
Highest Income Quintile	-0.098
	(0.078)
Couple Without Children	0.012
	(0.060)
Single Parent	-0.098
	(0.095)
Couple With Children	0.034
	(0.071)
Multiple Generation HH	-0.034
	(0.227)
Other Combination	0.480**
	(0.210)
Homeowner	-0.189***
	(0.040)
Constant	0.605***
	(0.100)
Observations	12,869
Pseudo R-squared	0.009
Chi-squared test	155.7
Prob. > Chi^2	0.000

Note. Own calculations based on SOEP.v36 and SOEP-RV.2018.
Standard errors (SE) in parentheses. The base category for age is 18-39.
The base category for household-type is 1-member household.
*** p<0.01, ** p<0.05, * p<0.1.

Appendix B

B.1 Stata code: Regression asked for consent

```
clear all
set more off
global data ""
global soep ""
global data_rtbn ""
global data_sr ""
global graphs ""
global desk “"
global tables ""

use  $soep\pgen.dta, clear
//merge RTBN mit SOEP = consented population
merge 1:1 pid syear using $soep\bipbrutto.dta
keep if _merge==3

gen befragt = 1 if bireclin_drv_erg == 1 | bireclin_drv_erg ==4
replace befragt = 0 if bireclin_drv_erg == -5 | bireclin_drv_erg ==-2 | bireclin_drv_erg ==2

gen age =.
replace age = 0 if bigeburt >= 1979  //39 und jünger
replace age = 1 if bigeburt <= 1978 &           bigeburt >= 1969 //40-49
replace age = 2 if bigeburt <= 1968 & bigeburt >= 1959           //50-59
replace age = 3 if bigeburt <= 1958 & bigeburt >= 1949           //60-69
replace age = 4 if bigeburt <= 1948 & bigeburt >= 1939           //70-79
replace age = 5 if bigeburt <= 1938  //80+

label define agelabel 0 "bis 39" 1 "40-49" 2 "50-59" 3 "60-69" 4 "70-79" 5 "80+"
label values age agelabel

merge m:1 hid syear using $soep\hgen.dta , keep(match) nogen //Homeownership, Household-type
merge 1:1 pid syear using $soep\pequiv.dta , keep(match) nogen //HH-Net-income
merge 1:1 pid syear using $soep\ppathl.dta , keep(match) nogen //migback
merge 1:1 pid syear using $soep\bip.dta , keep(match) nogen //bip_134, bip_175

*****Auxiliary Variables ******

gen homeowner = 1 if hgowner ==1
replace homeowner = 0 if hgowner >1
//Household-type:
gen hhtype = 1 if hgtyp1hh == 1
replace hhtype = 2 if hgtyp1hh == 2
replace hhtype = 3 if hgtyp1hh == 3
replace hhtype = 4 if hgtyp1hh == 4 | hgtyp1hh == 5 | hgtyp1hh == 6
replace hhtype = 5 if hgtyp1hh == 7
replace hhtype = 6 if hgtyp1hh == 8
label define hhlab 1 "1-Person HH" 2 "Couple Without Children" 3 "Single Parent" 4 "Couple With Children" 5
"Multiple Generation HH" 6 "Other Combination"
label values hhtype hhlab
```

```
//Education:
gen education = 0 if  pgisced11 == 0 |  pgisced11 == 1 |  pgisced11 == 2
replace education = 1 if  pgisced11 == 3 |  pgisced11 == 4
replace education = 2 if  pgisced11 == 5 |  pgisced11 == 6 | pgisced11 == 7 |  pgisced11 == 8 |  pgisced97 == 6

label define bilab 0 "Little" 1 "Medium" 2 "High"
label values education bilab

gen health=0 if bip_134 ==5 | bip_134 ==4
replace health = 1 if bip_134 == 3
replace health = 2 if bip_134 ==1 | bip_134 == 2

label define geslab 0 "Bad" 1 "Alright" 2 "Good"
label values health geslab

gen east_german = 1 if loc1989==1
replace east_german = 0 if loc1989 >1

xtile income = i11102, nq(5)
*xtile ek_alte = i11102 if bigeburt <= 1958, nq(5)

**recode: pgfamstd, pgpsbil, bip_134, bip_175 (Missing-problems)

recode sex (1=0) (2=1)
label define sex 0 "male", add
label define sex 1 "female", modify
label var sex "Female"

label var migback "Migration Background"
label define migback 1 "No Migration Background" 2 "Direct Migration Background" 3 "Indirect Migration
Background", modify

tab age, gen(age_)
label var age_2 "Age: 40-49"
label var age_3 "Age: 50-59"
label var age_4 "Age: 60-69"
label var age_5 "Age: 70-79"
label var age_6 "Age: 80+"

tab health, gen(health_)
label var health_2 "Medium Health Condition"
label var health_3 "Good Health Condition"

tab migback, gen(migback_)
label var migback_2 "Direckt Migration Background"
label var migback_3 "Indirect Migration Background"

tab education, gen(education_)
label var education_2 "Medium Education"
label var education_3 "High Education"

tab income, gen(income_)
label var income_2 "Second Income Quintile"
```

```
label var income_3 "Third Income Quintile"
label var income_4 "Fourth Income Quintile"
label var income_5 "Highest Income Quintile"

tab hhtype, gen(hhtype_)
label var hhtype_1 "1-Person HH"
label var hhtype_2 "Couple Without Children"
label var hhtype_3 "Single Parent"
label var hhtype_4 "Couple With Children"
label var hhtype_5 "Multiple Generation HH"
label var hhtype_6 "Other Combination"

label var east_german "East German"
label var homeowner "Homeowner"

**************************Regressions*****************************************
**Table A.1:
logit befragt age_2 age_3 age_4 age_5 age_6 health_2 health_3 sex migback_2 migback_3 east_german
education_2 education_3 income_2 income_3 income_4 income_5 hhtype_2 hhtype_3 hhtype_4 hhtype_5
hhtype_6 homeowner
vif, uncenteredvif, uncentered
outreg2 using $tables\participation_regression, dec(3) addstat(Pseudo R-squared, `e(r2_p)', chi-square test,
`e(chi2)', Prob > chi2, `e(p)') word label replace
**Table 1:
logit befragt age_2 age_3 age_4 age_5 age_6 health_2 health_3 sex migback_2 migback_3 east_german
education_2 education_3 income_2 income_3 income_4 income_5 hhtype_2 hhtype_3 hhtype_4 hhtype_5
hhtype_6 homeowner
vif, uncentered
margins, dydx(_all) post
outreg2 using $tables\participation_margins, dec(3) ctitle(margins) label  word sideway replace
```

B.2 Stata code: Regression who gave consent

```
clear all
set more off
global data ""
global soep ""
global data_rtbn ""
global data_sr ""
global graphs ""
global desk ""
global tables ""

use $data_sr\SUF_soep-rtbn_mit_pid.dta, clear                    //merge RTBN mit SOEP = consented population
gen syear=2018
merge 1:1 pid syear using $soep\bipbrutto.dta
keep if bireclin_drv_erg == 1 | bireclin_drv_erg ==4

gen consent = 1 if bireclin_drv_erg ==1
recode consent .=0

gen age =.
replace age = 0 if bigeburt >= 1979  //39 und jünger
```

```
replace age = 1 if bigeburt <= 1978 &           bigeburt >= 1969 //40-49
replace age = 2 if bigeburt <= 1968 & bigeburt >= 1959           //50-59
replace age = 3 if bigeburt <= 1958 & bigeburt >= 1949           //60-69
replace age = 4 if bigeburt <= 1948 & bigeburt >= 1939           //70-79
replace age = 5 if bigeburt <= 1938  //80+

label define agelabel 0 "bis 39" 1 "40-49" 2 "50-59" 3 "60-69" 4 "70-79" 5 "80+"
label values age agelabel
merge m:1 hid syear using $soep\hgen.dta , keep(match) nogen //Homeownership, Householdtype

merge 1:1 pid syear using $soep\pgen.dta , keep(match) nogen //Education,
merge 1:1 pid syear using $soep\pequiv.dta , keep(match) nogen //HH-Net-income
merge 1:1 pid syear using $soep\ppathl.dta , keep(match) nogen //migback
merge 1:1 pid syear using $soep\bip.dta , keep(match) nogen //bip_134, bip_175

*****edit Variablen ******

gen homeowner = 1 if hgowner ==1
replace homeowner = 0 if hgowner >1

gen hhtype = 1 if hgtyp1hh == 1
replace hhtype = 2 if hgtyp1hh == 2
replace hhtype = 3 if hgtyp1hh == 3
replace hhtype = 4 if hgtyp1hh == 4 | hgtyp1hh == 5 | hgtyp1hh == 6
replace hhtype = 5 if hgtyp1hh == 7
replace hhtype = 6 if hgtyp1hh == 8
label define hhlab 1 "1-Person HH" 2 "Couple Without Children" 3 "Single Parent" 4 "Couple With Children" 5
"Multiple Generation HH" 6 "Other Combination"
label values hhtype hhlab
//Education
gen education = 0 if  pgisced11 == 0 |  pgisced11 == 1 |  pgisced11 == 2
replace education = 1 if  pgisced11 == 3 |  pgisced11 == 4
replace education = 2 if  pgisced11 == 5 |  pgisced11 == 6 | pgisced11 == 7 |  pgisced11 == 8 |  pgisced97 == 6

label define bilab 0 "Little" 1 "Medium" 2 "High"
label values education bilab

gen health=0 if bip_134 ==5 | bip_134 ==4
replace health = 1 if bip_134 == 3
replace health = 2 if bip_134 ==1 | bip_134 == 2

label define geslab 0 "Bad" 1 "Alright" 2 "Good"
label values health geslab

gen east_german = 1 if loc1989==1
replace east_german = 0 if loc1989 >1

xtile income = i11102, nq(5)

**recode: pgfamstd, pgpsbil, bip_134, bip_175 (Missing-Problems)

recode sex (1=0) (2=1)
label define sex 0 "male", add
```

```
label define sex 1 "female", modify
label var sex "Female"

label var migback "Migration Background"
label define migback 1 "No Migration Background" 2 "Direct Migration Background" 3 "Indirect Migration
Background", modify

tab age, gen(age_)
label var age_2 "Age: 40-49"
label var age_3 "Age: 50-59"
label var age_4 "Age: 60-69"
label var age_5 "Age: 70-79"
label var age_6 "Age: 80+"

tab health, gen(health_)
label var health_2 "Medium Health Condition"
label var health_3 "Good Health Condition"

tab migback, gen(migback_)
label var migback_2 "Direckt Migration Background"
label var migback_3 "Indirect Migration Background"

tab education, gen(education_)
label var education_2 "Medium Education"
label var education_3 "High Education"

tab income, gen(income_)
label var income_2 "Second Income Quintile"
label var income_3 "Third Income Quintile"
label var income_4 "Fourth Income Quintile"
label var income_5 "Highest Income Quintile"

tab hhtype, gen(hhtype_)
label var hhtype_1 "1-Person HH"
label var hhtype_2 "Couple Without Children"
label var hhtype_3 "Single Parent"
label var hhtype_4 "Couple With Children"
label var hhtype_5 "Multiple Generation HH"
label var hhtype_6 "Other Combination"

label var east_german "East German"
label var homeowner "Homeowner"

**********Regressions**********
**Table A.2:
logit consent age_2 age_3 age_4 age_5 age_6 health_2 health_3 sex migback_2 migback_3 east_german
education_2 education_3 income_2 income_3 income_4 income_5 hhtype_2 hhtype_3 hhtype_4 hhtype_5
hhtype_6 homeowner
outreg2 using $tables\consent_regression, dec(3) addstat(Pseudo R-squared, `e(r2_p)', chi-square test, `e(chi2)',
Prob > chi2, `e(p)') word label replace

**Table 2:
```

```
logit consent age_2 age_3 age_4 age_5 age_6 health_2 health_3 sex migback_2 migback_3 east_german
education_2 education_3 income_2 income_3 income_4 income_5 hhtype_2 hhtype_3 hhtype_4 hhtype_5
hhtype_6 homeowner
margins, dydx(_all) post
outreg2 using $tables\consent_margins, dec(3) ctitle(margins) label  word sideway replace
```

B.3 Stata code: Results for Section 4

```
clear all
set more off
if "`c(username)'"=="hpenz"{
global data_rtbn ""
global data_sr ""
global soep ""”
global graphs ""
global desk ""
global temp ""

use $data_sr\SUF_soep-rtbn_mit_pid.dta, clear                    //merge RTBN mit SOEP = consented population
merge 1:1 pid using $soep\bip.dta
keep if _merge ==3
drop _merge
save $temp\descriptives.dta, replace
use $soep\pequiv.dta, clear
keep if syear==2018
keep pid igrv1
merge 1:1 pid using $temp\descriptives.dta
keep if _merge ==3
drop _merge
*Create Dummy before append RTBN to SUF_SOEP_RTBN
gen dummy =1
append using $data_rtbn\SUFRTBN18XVSBB.dta
replace dummy =0 if dummy == .
label variable dummy "RTBN oder SOEP-RTBN"
label define dummylabel 0 "RTBN" 1 "SOEP-RTBN"
label values dummy dummylabel
tab dummy
save $temp\descriptives.dta, replace

********************Cleaning*****************************************************
**Attention: Clean deceased, zero rents and KLG-benefits
*Important: restrict to pension recipients alive
tab rtat //Overview of pension type
                        //Delete spells for rtat>=3 -> those are deceased or surviving dependants 13 Deceased.
keep if rtat <3 & rtzb !=0
tab rtat
************************Auxiliary Variables***********************************
gen age_rtbn =2018-gbjavs
gen age_soep = 2018-bipbirthy
replace age_soep = 35 if bipbirthy >=1983
replace age_soep = 96 if bipbirthy <=1922
label var age_rtbn "age in 2018, RTBN"
```

```
label var age_soep "age in 2018, SOEP_RTBN"

gen gbjahr_kat =.                    //Birthyear categorical
replace gbjahr_kat = 0 if gbjavs <= 1938              //1938 or earlier
replace gbjahr_kat = 1 if gbjavs > 1938 & gbjavs <= 1948              //1939-1948
replace gbjahr_kat = 2 if gbjavs > 1948 & gbjavs <= 1958              //1949-1958
replace gbjahr_kat = 3 if gbjavs > 1958 & gbjavs <= 1968              //1959-1968
replace gbjahr_kat = 4 if gbjavs > 1968 & gbjavs <= 1978              //1969-1978
replace gbjahr_kat = 5 if gbjavs > 1978 & gbjavs <= 1982              //1979-1982
replace gbjahr_kat = 6 if gbjavs >=1983      //1983 or later
tab gbjahr_kat
label define gbjahr_katlabel 0 "80 and older" ///
                                                    1 "70-79" 2 "60-69" 3 "50-59" 4 "40-49" 5 "39-36" ///
                                                    6 "35 and younger"
label values gbjahr_kat gbjahr_katlabel

tab rtzb          //retirement payment
gen rentpay=.
replace rentpay = 0 if rtzb == 0
replace rentpay = 1 if rtzb > 0 & rtzb <=500
replace rentpay = 2 if rtzb >500 & rtzb <= 1000
replace rentpay = 3 if rtzb >1000 & rtzb <=1500
replace rentpay = 4 if rtzb >1500 & rtzb <=2000
replace rentpay = 5 if rtzb >2000 & rtzb <=2200
replace rentpay = 6 if rtzb == 2376
tab rentpay
label define rentpaylabel 0 "zero rent" 1 "retirement payment >0 & <=500" ///
                                                    2 "retirement payment >500 & <=1000" ///
                                                    3 "retirement payment >1000 & <=1500" ///
                                                    4 "retirement paymentg >1500 & <=2000" ///
                                                    5 "retirement payment >2000 & <=2200" ///
                                                    6 "2376 Mean RTZB > 2200"
label values rentpay rentpaylabel
*****generate categorical variables:*****************************************
tab gevs, gen(gevs_)
tab fmsd, gen(fmsd_)
tab leat, gen(leat_)
**************************Graphs*********************************************
** Retirement payments
**** include cases, who entered retirement 2017 and correct these cases
**** Calculate monthly soep retirement payments
gen soep_rtzb = igrv1/12 if rtbej <2017 & dummy == 1              //monthly SOEP-retirement payments for
those who entered before 2017
label var soep_rtzb "monthly retirement pay SOEP"
**** monthly SOEP-retirement payments with retirement payment in 2017:
replace soep_rtzb = igrv1/12 if rtbej == 2017& rtbem == 1
replace soep_rtzb = igrv1/11 if rtbej == 2017& rtbem == 2
replace soep_rtzb = igrv1/10 if rtbej == 2017& rtbem == 3
replace soep_rtzb = igrv1/9 if rtbej == 2017& rtbem == 4
replace soep_rtzb = igrv1/8 if rtbej == 2017& rtbem == 5
replace soep_rtzb = igrv1/7 if rtbej == 2017& rtbem == 6
replace soep_rtzb = igrv1/6 if rtbej == 2017& rtbem == 7
replace soep_rtzb = igrv1/5 if rtbej == 2017& rtbem == 8
```

```
replace soep_rtzb = igrv1/4 if rtbej == 2017& rtbem == 9
replace soep_rtzb = igrv1/3 if rtbej == 2017& rtbem == 10
replace soep_rtzb = igrv1/2 if rtbej == 2017& rtbem == 11
replace soep_rtzb = igrv1/1 if rtbej == 2017& rtbem == 12
tab soep_rtzb if rtbej == 2017

sum rtzb soep_rtzb if dummy == 1 & rtbej<=2017 & soep_rtzb<=4500

tab leat soep_rtzb if dummy ==1 & soep_rtzb >=4500              //4 outliers for long-time and especially long-time
insured. All for entered retirement between October and December 2017

****Figure 2: Scatterplot
preserve
gen count = 1
replace rtzb=rtzb/50
replace rtzb=round(rtzb)
replace rtzb=rtzb*50
replace soep_rtzb=soep_rtzb/50
replace soep_rtzb=round(soep_rtzb)
replace soep_rtzb=soep_rtzb*50
gen rtzb_diff = rtzb - soep_rtzb
corr soep_rtzb rtzb
local corr: di %5.3g r(rho)
collapse (sum) count if dummy ==1 & rtbej<=2017  & leat_1==0 & rtzb_diff<=1000 & rtzb_diff>=-1000, by(rtzb
rtzb_diff)
list
scatter rtzb_diff rtzb [aw=count], msize(vsmall) msymbol(Oh) scheme(s2mono) graphregion(color(white)) ///
                        ytitle("Difference RTBN and SOEP retirement payments") xtitle("RTBN retirement payments")
legend(off) ///
                        || lfit rtzb_diff rtzb, lcolor(black) lwidth(thick)

graph export $graphs\scatter_difference_pension.emf, replace

***Figure 3:
********************Birthyear
tab gbjavs dummy
#delimit ;
twoway hist gbjavs if dummy ==1 & gbjavs<=1958, lcolor(black) fcolor(none) width(1) || hist gbjavs if dummy ==0
& gbjavs<=1958, color(grey%30)
legend(label(1 "SOEP-RTBN") label(2 "RTBN")) width(1) scheme(s1mono) xtitle("Year") saving(birthyear)
;
#delimit cr
******Retirement Pay
#delimit ;
twoway hist rtzb if dummy ==1 & gbjavs<=1958, lcolor(black) fcolor(none) width(100) start(0) || hist rtzb if dummy
==0 & gbjavs<=1958, color(grey%30)
legend(label(1 "SOEP-RTBN") label(2 "RTBN")) width(100) scheme(s1mono) xtitle("in €") saving(retirement_pay)
;
#delimit cr

#delimit ;
grc1leg birthyear.gph retirement_pay.gph, leg(birthyear.gph)
ring(3)
```

```
imargin(0 4 0 0) xsize(10) ysize(4)
graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) icolor(white)) scheme(s1mono)
;
#delimit cr
graph display, xsize(10) ysize(4)
graph export $graphs\by_rp_60plus.emf, replace

**************************Tables*********************************************
**********************Table 3:
**-        Compare administrative data in RTBN with self-reported data in SOEP from
** succesfully linked insured person: RTBN vs. SOEP if dummy == 1
**1) Gender
tab gevs sex if dummy == 1, row matcell(gender)
mat gender=gender/r(N)
matlist gender
mat sum_g = trace(gender)
matlist sum_g
**2) Age
sum alter_rtbn if dummy ==1, detail
sum alter_soep if dummy ==1, detail
tab alter_rtbn alter_soep if dummy == 1, row matcell(age)
mat age = age/r(N)
matlist age
mat sum_a = trace(age)
matlist sum_a

**3) Civil Status
* 0 = not defined - nicht definiert/Altfall/entfällt
* 1 = not married (single/divorced/widowed) - nicht verheriatet (ledig/gesschieden/verwitwet)/nicht in eingetragener lebenspartnerschaft lebend
* 2 = married/remarried/ registered partnership -  verheiratet/wiederverheiratet/in eingetragener lebenspartnerschaft lebend
gen fmsd_soep =.
replace fmsd_soep = . if bip_191 <0
replace fmsd_soep = 1 if bip_191 == 3 | bip_191 == 4 | bip_191 == 5                  //3 = single, never married - ledig, war nie verheiratet, 4= divorced/annulled registered parntership - geschieden/eingetragene gleichgeschlechtliche Partnerschaft aufgehoben, 5 = widowed/deceased partner - verwitwet/gleichgeschleichtliche Lebenspartner verstorben
replace fmsd_soep = 2 if bip_191 == 1 | bip_191 == 2                                  // 1= married and living together - verheiratet, mit ehepartner zuammenlebend, 2= married, not living together - verheiratet, dauernd getrennt lebend

label define fmsd_soeplabel 1 "not married - single/divorced/widowed" 2 "married/remarried/registered partnership"
label values fmsd_soep fmsd_soeplabel

tab fmsd fmsd_soep if dummy ==1 & fmsd ==1, row matcell(cs_nm)
mat cs_nm = cs_nm/r(N)
matlist cs_nm

tab fmsd fmsd_soep if dummy ==1 & fmsd ==2, row matcell(cs_m)
mat cs_m = cs_m/r(N)
```

```
matlist cs_m
mat P = (0, 1 \ 1,0)
matlist P
mat cs_m = cs_m*P
matlist cs_m

****Generate Table 3:
mat mat_gesamt = J(6,2,.)
mat rownames mat_gesamt = "Gender" "Age" "Single/Divorced/Widowed" "Married" "Observations"
mat colnames mat_gesamt ="consistent info" "inconsistent info"
mat mat_gesamt [1,1] = sum_g
mat mat_gesamt [1,2] = 1-sum_g
mat mat_gesamt [2,1] = sum_a
mat mat_gesamt [2,2] = 1-sum_a
mat mat_gesamt [3,1] = cs_nm
mat mat_gesamt [4,1] = cs_m
sum gevs_1 if dummy == 1
mat mat_gesamt [6,1] = r(N)
putexcel set "$desk\results.xlsx", sheet("table_3") modify
putexcel B2 = matrix(mat_gesamt), names

******Table 4: Retirement payments comparison on the individual level
program diff1, rclass
	version 16
	summarize rtzb if dummy == 1 & rtbej<=2017 & leat_1==0, detail
	local rr10 = r(p10)
			local rr50 = r(p50)
			local rr90 = r(p90)
	summarize soep_rtzb if dummy == 1 & rtbej<=2017 & leat_1==0, detail
	local sr10 = r(p10)
			local sr50 = r(p50)
			local sr90 = r(p90)
	return scalar dif10 = `rr10'-`sr10'
			return scalar dif50 = `rr50'-`sr50'
			return scalar dif90 = `rr90'-`sr90'
 end

bootstrap r(dif10) r(dif50) r(dif90), reps(100): diff1
mat mat_r10b = e(b)
mat mat_r10se = e(se)

mat mat_p = J(3,2,.)
mat rownames mat_p = "r_r10" "r_r50" "r_r90"
mat colnames mat_p = "b" "se"
mat mat_p [1,1] = mat_r10b'
mat mat_p [1,2] =mat_r10se'
mat list mat_p

mat mat_gesamt2 = J(5,5,.)
	mat colnames mat_gesamt2 = "RTBN" "SOEP-RTBN" "Difference" "sd" "p_value"
	mat rownames mat_gesamt2 = "Mean" "P10" "P50" "P90" "Observations"

		sum rtzb if dummy ==1 & rtbej<=2017  & leat_1==0, detail
```

```
mat mat_gesamt2 [1, 1]=r(mean)
mat mat_gesamt2 [2, 1]=r(p10)
mat mat_gesamt2 [3, 1]=r(p50)
mat mat_gesamt2 [4, 1]=r(p90)
mat mat_gesamt2 [5, 1]=r(N)

sum soep_rtzb if dummy ==1 & rtbej<=2017  & leat_1==0, detail
mat mat_gesamt2 [1, 2]=r(mean)
mat mat_gesamt2 [2, 2]=r(p10)
mat mat_gesamt2 [3, 2]=r(p50)
mat mat_gesamt2 [4, 2]=r(p90)
mat mat_gesamt2 [5, 2]=r(N)

mat mat_gesamt2 [1, 3]= mat_gesamt2[1,1]-mat_gesamt2[1,2]
mat mat_gesamt2 [2, 3]= mat_p
ttest rtzb = soep_rtzb if dummy ==1 & rtbej<=2017  & leat_1==0
mat mat_gesamt2 [1,4]=r(se)
mat mat_gesamt2 [1, 5]=r(p)

mat list mat_gesamt2

putexcel set "$desk\results.xlsx", sheet("table_4") modify
putexcel B2 = matrix(mat_gesamt2), names
svmat mat_gesamt2

*******************************************************************************
******Table 5: Retirement paments and age comparison RTBN and SOEP_RTBN

sum alter_rtbn if dummy ==0 & gbjavs<=1958, detail
sum alter_rtbn if dummy ==1 & gbjavs<=1958, detail
sum rtzb if dummy ==0 & gbjavs<=1958, detail
sum rtzb if dummy ==1 & gbjavs<=1958, detail

program dif3, rclass
	version 16
	summarize alter_rtbn if dummy == 0 & gbjavs<=1958, detail
	local alter_r10_d0 = r(p10)
	summarize alter_rtbn if dummy == 1 & gbjavs<=1958, detail
	local alter_r10_d1 = r(p10)
	return scalar a_dif10 = `alter_r10_d0'-`alter_r10_d1'
 end

bootstrap r(a_dif10), reps(100): dif3

program diff2, rclass
	version 16
	summarize alter_rtbn if dummy == 0 & gbjavs<=1958, detail
	local alter_r10_d0 = r(p10)
		local alter_r50_d0 = r(p50)
		local alter_r90_d0 = r(p90)
	summarize alter_rtbn if dummy == 1 & gbjavs<=1958, detail
	local alter_r10_d1 = r(p10)
		local alter_r50_d1 = r(p50)
```

```
                local alter_r90_d1 = r(p90)
        return scalar a_dif10 = `alter_r10_d0'-`alter_r10_d1'
                return scalar a_dif50 = `alter_r50_d0'-`alter_r50_d1'
                return scalar a_dif90 = `alter_r90_d0'-`alter_r90_d1'
                summarize rtzb if dummy == 0 & gbjavs<=1958, detail
        local rtzb_r10_d0 = r(p10)
                local rtzb_r50_d0 = r(p50)
                local rtzb_r90_d0 = r(p90)
        summarize rtzb if dummy == 1 & gbjavs<=1958, detail
        local rtzb_r10_d1 = r(p10)
                local rtzb_r50_d1 = r(p50)
                local rtzb_r90_d1 = r(p90)
        return scalar r_dif10 = `rtzb_r10_d0'-`rtzb_r10_d1'
                return scalar r_dif50 = `rtzb_r50_d0'-`rtzb_r50_d1'
                return scalar r_dif90 = `rtzb_r90_d0'-`rtzb_r90_d1'
 end

bootstrap r(a_dif10) r(a_dif50) r(a_dif90) r(r_dif10) r(r_dif50) r(r_dif90), reps(100): diff2
mat mat_b = e(b)
mat mat_s =  e(se)
mat mat_p = J(6,2,.)
mat rownames mat_p = "a_r10" "a_r50" "a_r90" "r_r10" "r_r50" "r_r90"
mat colnames mat_p = "b" "se"
mat mat_p [1,1] = mat_b'
mat mat_p [1,2] =mat_s'
mat list mat_p
local count = 1
mat mat_gesamt1 = J(9,5,.)
        mat colnames mat_gesamt1 = "RTBN_60p" "SOEP_60p" "Difference" "sd" "p_value"
        mat rownames mat_gesamt1 = "a_mean" "r_mean" "a_p10" "a_p50" "a_p90" "r_p10" "r_p50" "r_p90"
"obs"
foreach var of varlist alter_rtbn rtzb {

                sum `var' if dummy == 0 & gbjavs <= 1958, detail                //60plus
                mat mat_gesamt1 [`count',1]=r(mean)

                sum `var' if dummy == 1 & gbjavs <= 1958, meanonly              //60plus
                mat mat_gesamt1 [`count',2]=r(mean)

                mat mat_gesamt1 [`count', 3]= mat_gesamt1[`count',1]-mat_gesamt1[`count',2]
                ttest `var', by(dummy)
                mat mat_gesamt1 [`count',4]=r(sd)
                mat mat_gesamt1 [`count', 5]=r(p)

                local count = `count' + 1
}
mat list mat_gesamt1
sum alter_rtbn if dummy == 0 & gbjavs <= 1958, detail
                mat mat_gesamt1 [3, 1]=r(p10)
                mat mat_gesamt1 [4, 1]=r(p50)
                mat mat_gesamt1 [5, 1]=r(p90)
sum alter_rtbn if dummy == 1 & gbjavs <= 1958, detail
                mat mat_gesamt1 [3, 2]=r(p10)
```

```
                mat mat_gesamt1 [4, 2]=r(p50)
                mat mat_gesamt1 [5, 2]=r(p90)
mat list mat_gesamt1
                mat mat_gesamt1 [3, 3]=mat_p
mat list mat_gesamt1

sum rtzb if dummy == 0 & gbjavs <= 1958, detail
                mat mat_gesamt1 [6, 1]=r(p10)
                mat mat_gesamt1 [7, 1]=r(p50)
                mat mat_gesamt1 [8, 1]=r(p90)
                mat mat_gesamt1 [9, 1]=r(N)

sum rtzb if dummy ==1 & gbjavs <= 1958, detail
                mat mat_gesamt1 [6, 2]=r(p10)
                mat mat_gesamt1 [7, 2]=r(p50)
                mat mat_gesamt1 [8, 2]=r(p90)
                mat mat_gesamt1 [9, 2]=r(N)
mat list mat_gesamt1

putexcel set "$desk\results.xlsx", sheet("table_5") modify
putexcel B2 = matrix(mat_gesamt1), names
svmat mat_gesamt1

*******************************************************************************
******************Table 6: Gender and pension types RTBN and SOEP-RTBN
local count = 1
mat mat_gesamt = J(10,4,.)
        mat colnames mat_gesamt = "RTBN_60p" "SOEP_60p" "Difference" "p_value"
        mat rownames mat_gesamt = "Male" "Female" "invalidity" "reg old-age" "unempl/p-time" "old-age f"
"sev disabled" "long time" "esp long time" "other"
foreach var of varlist gevs_1 gevs_2 leat_1 leat_2 leat_3 leat_4 leat_5 leat_6 leat_7 leat_8 {

                sum `var' if dummy == 0 & gbjavs<=1958, meanonly                    //60 Plus
                mat mat_gesamt [`count',1]=r(mean)

                sum `var' if dummy == 1 & gbjavs<=1958, meanonly                    //60 Plus
                mat mat_gesamt [`count',2]=r(mean)

                mat mat_gesamt [`count', 3]= mat_gesamt[`count',1]-mat_gesamt[`count',2]
                tab `var' dummy, chi
                mat mat_gesamt [`count', 4]=r(p)

                local count = `count' + 1
}
mat list mat_gesamt

mat mat_N_gesamt = J(1,2,.)
mat rownames mat_N_gesamt ="Observations"
sum gevs_1 if dummy == 0
mat mat_N_gesamt [1,1] = r(N)
sum gevs_1 if dummy == 1
mat mat_N_gesamt [1,2] = r(N)
```

```
putexcel set "$desk\results.xlsx", sheet("table_6") modify
putexcel B6 = matrix(mat_gesamt), names
putexcel B22 = matrix(mat_N_gesamt), names
svmat mat_gesamt
```