DIW BERLIN

86

Data Documentation

Deutsches Institut für Wirtschaftsforschung 2016

An Integrated Micro Data Base for Tax Analysis in Germany

Stefan Bach, Martin Beznoska and Viktor Steiner

IMPRESSUM
© DIW Berlin, 2016
DIW Berlin
Deutsches Institut für Wirtschaftsforschung
Mohrenstr. 58
10117 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
www.diw.de
ISSN 1861-1532

Data Documentation 86

Stefan Bach*
Martin Beznoska
Viktor Steiner**

An Integrated Micro Data Base for Tax Analysis in Germany

This paper documents methodology underlying the construction of the integrated data base for our study on “Wer trägt die Steuerlast in Deutschland? – Verteilungswirkungen des deutschen Steuer- und Transfersystems” (Who bears the tax burden in Germany? – Distributional Analyses of the German tax and transfer system). Financial support from the Hans Böckler Stiftung for the project is gratefully acknowledged. The paper greatly benefited from comments by the members of the scientific advisory council of the project.

Keywords: Microsimulation models on taxes and transfers, data integration, income distribution.

JEL: H24, C81, D31.

Berlin, December 2016

* DIW Berlin, Abteilung Staat. sbach@diw.de
** FU Berlin, Fachbereich Wirtschaftswissenschaft. Viktor.Steiner@fu-berlin.de

Table of Contents

1 Introduction 1

2 Integrated micro data base 3

2.1 Data 3

2.2 Statistical Matching of the SOEP and the EVS 4

2.3 Integration of Consumption Data 9

2.4 Integration of the Wage and Income Tax Statistic 13

2.5 Data Extrapolation and Uprating 17

3 Microsimulation models 19

3.1 Tax-Benefit Microsimulation Model STSM 19

3.2 Personal Income Tax Microsimulation Model 22

3.3 Modelling VAT and Excises 24

List of Tables and Figures

Table 1 Consumption Expenditures in the EVS Data ... 5
Table 2 Matching Frequencies SOEP-EVS Matching 2008 ... 7
Table 3 Marginal Distributions of Selected Z-Variables ... 8
Table 5 The Distribution of Consumption in the EVS 2008 ... 11
Table 6 The Distribution of Consumption 2015 before and after Adjustments ... 12
Table 7 The Integrated Data 2015 after Adjusting the Weights to LESt Distributional Information ... 14
Table 8 Integrated Data Base (SOEP-EVS-LESt) ... 17
Table 9 Components of Net Household Income ... 19

Figure 1 Estimated Functional Form of the Value Added Tax Share in Gross Household Income ... 16

1 Introduction

This report documents the methodology underlying the construction of the integrated data base for our study on "Who bears the tax burden in Germany? – Distributional Analyses of the German tax and transfer system".[1] An important aim of this research project was to build up a comprehensive micro data set that includes detailed information about income and consumption of German households. The data set should cover and represent the distribution of income for German households adequately in cross section as well as over time. Additionally, the German tax and transfer system should be represented as detailed as possible to allow distributional analyses. Therefore, detailed information on income and consumption are necessary to simulate both direct and indirect taxes. Using micro data combined from various sources, we analyze the representative distribution of pre- and post-government income over the period 1995 to 2015.

The first issue we face is that all available surveys on incomes in Germany have problems of representativeness as they do no cover people at the very top of the income distribution. This is a particularly severe problem for distributional analyses concerning the income tax which is highly concentrated at the top of the distribution. A lack of representativeness in this part of the distribution causes a serious bias in the analysis of the progressivity for the income tax.

The second issue is the range of available household information. The German Socio Economic Panel (SOEP, "Sozio-oekonomisches Panel") provides a rich data set with detailed information on income and surveys the households every year over a large time period but it lacks information on consumption expenditures. The Income and Consumption Survey for Germany (EVS, "Einkommens- und Verbrauchsstichprobe") in contrast covers detailed information on both income and expenditures but is only conducted every 5 years. As the EVS does not include households with very high incomes, it is less representative at the top of the income distribution than the SOEP. We combine both data sets to overcome the respective limitations of the two data sets and integrate them into a new comprehensive data set.

[1] See Bach, Beznoska and Steiner (2016a) for the results and details of the project. A summary of the main results of this study in are contained in Bach, Beznoska and Steiner (2016b). An English version of this summary is available in Bach, Beznoska and Steiner (2016c).

This integrated data set uses the advantages of the SOEP covering a large time period with good representative information on income and adds expenditure information from the EVS by statistical matching.

To further improve the representativeness at the top of the income distribution we impute information from the Wage and Income Tax Statistic (LESt, "Lohn- und Einkommensteuer-statistik"). The data of the LESt, which is an administrative data set and only usable via remote data access, contains all income tax payers in Germany and detailed information on their gross income and income based taxes. We integrate this data at the 0.1-percentile level into our SOEP data base and use it for our distributional analysis. By this imputation procedure way, we can also allocate part of the business taxes to individuals and households.

With this integrated data set, we are able to simulate income and consumption taxes and compare tax burdens at the individual and household level over time. We look at the tax burdens at five points in time, which are the years 1995, 1998, 2005, 2008 and uprated to 2015. We cover the most important taxes paid by households: personal income tax (incl. taxes on business and capital income), solidarity surcharge tax, value added tax, insurance tax, energy taxes, tobacco, alcohol and gambling tax, car tax and real estate taxes. Additionally, we provide an approach to pass the indirect taxes paid by companies on to the households, and we are simulating the social security contributions.

The integration of the various date sources and the structure of the resulting integrated data base for the distributional analyses of the tax burden is described in chapter 2. The microsimulation models used to analyze recent tax reforms or counterfactual tax reform scenarios are described in chapter 3.

2 Integrated micro data base

2.1 Data

The main data base for the analysis is the German Socio Economic Panel (SOEP). It is a representative survey which started in 1984 and currently covers about 30,000 respondents in approximately 11,000 households, which are interviewed yearly.[2] In the starting year of our study 1995, there are about 6,500 households but due to the adding of new subsamples, the sample size has constantly increased over the past years.[3] Importantly, the “high-income subsample”, currently consisting of about 800 households, has been included since 2002 in the SOEP, which improves the representativeness at the top of the income distribution (see Frick et al., 2007). This feature is very important for our purposes, because the top income area is also most significant in research questions on inequality. Furthermore, a big share of the income tax revenue is generated by relatively few tax payers at the top of the distribution. We are interested in representing these households as detailed as possible because an inaccurate treatment of this group could result in serious biases in the analysis. Because of these attributes of the SOEP and its detailed household and income information, we chose this survey as our data base.

For our purpose, a disadvantage of the SOEP data is the lack of information on consumption expenditures.[4] As we are interested in covering all kinds of taxes and assigning them to the households, information on expenditures is essential for the simulation of the indirect taxes. To overcome this problem, an additional data source is statistically matched to the SOEP: the Income and Consumption Survey for Germany (EVS, “Einkommens- und Verbrauchsstichprobe”). The EVS is also a representative household data set, which is surveyed every 5 years. It has no panel structure and therefore every survey year is an independent cross-section consisting of about 50,000 households. The scientific use-file of the EVS contains an 80%-subsample of the original data. Since the most recent EVS survey conducted in 2013 has

[2] http://www.diw.de/en/soep

[3] See Wagner et al. (2007) for a detailed description of the SOEP.

[4] Consumption expenditures were recorded in the 2010 SOEP special survey in a few categories. These aggregates are too broad for the simulation of indirect taxes, and here are also significant deviations in the distribution of these aggregates compared to the EVS 2008 data. As suggested by Markus et al. (2013) this may be due to problems concerning the definition of the consumption aggregates in the SOEP.

not been available yet when our study was undertaken, we use the EVS surveys for 1998, 2003 and 2008 to analyze household consumption taxation in the period 1995 to 2015. The EVS also reports the household's socio-demographics and incomes very detailed, comparable with the SOEP. Additionally, it has the unique feature of reporting detailed information on household expenditures covering all kinds of consumption expenditures on commodities and services. We integrate this information into the SOEP by statistical matching, which will be presented in the next section. A problem of the EVS is that it excludes households with net household income exceeding 18,000 Euro per month. Therefore, there is no information of the expenditures of the top income households available. We are trying to solve this issue with imputation methods that will also be presented in the next section.

The third micro data set used in our integrated data is the Wage and Income Tax Statistic for Germany (LESt, "Lohn- und Einkommensteuerstatistik"). It is an administrative data set covering all income tax payers in Germany for the particular year. A subsample of the LESt is available as a scientific use file every 3 years. Due to the long tax assessment period and production lag of the scientific use file, the latest year for which data was available when our study was undertaken is 2007. As for top incomes not all relevant tax information is available in the scientific use-file, we had to use remote data access to analyze the whole data set. Thereby we could us the full population of tax payers at the top of the distribution, and therefore avoid any sampling-error, by imputing all relevant tax information from the LESt at the 0.1-percentile level into our SOEP data base, as described below. Thus, we fully cover the part of the income distribution at the top which is not observed in the SOEP. Furthermore, taxes on capital income and business, which are strongly underreported in the SOEP, are much better represented in the integrated data base. In addition, business taxes paid at the company level on distributed profits can be assigned to individual households.

2.2 Statistical Matching of the SOEP and the EVS

The detailed consumption information available in the EVS data allows the simulation of all relevant indirect taxes, e.g. the value added tax, the taxes on the consumption of energy goods and the other excises. For each household, quarterly consumption expenditures are recorded for up to 153 items classified into 12 categories. Table 1 gives a summary of these 12 categories of consumption expenditures available in the EVS for the three survey years.

Table 1
Consumption Expenditures in the EVS Data

Consumption Expenditure Aggregate	Expenditures per Month			Share in Total Consumption Expenditures		
	1998	2003	2008	1998	2003	2008
	in Euro of 2003			in %		
Total Consumption Expenditures	2 198	2 179	2 043	100,0	100,0	100,0
Food, Beverages and Tobacco	308	303	292	14,0	13,9	14,3
Clothes and Shoes	126	112	97	5,7	5,1	4,7
Rents, Imputed Rents and Maintenance	599	575	539	27,3	26,4	26,4
Energy	102	119	127	4,6	5,5	6,2
Furniture and Household Appliance	155	126	103	7,1	5,8	5,0
Health Costs	80	87	85	3,6	4,0	4,2
Mobility Costs	296	307	297	13,5	14,1	14,5
Communication	54	68	59	2,5	3,1	2,9
Leisure, Entertainment and Culture	264	262	233	12,0	12,0	11,4
Education	11	19	19	0,5	0,9	0,9
Accomodation and Service	108	101	103	4,9	4,6	5,0
Other Goods and Services	95	100	89	4,3	4,6	4,4

Source: Own calculations with EVS 1998, 2003 and 2008.

In principle, there are two possibilities to integrate the consumption information into the SOEP data: mean imputation methods (like regression analysis) or statistical matching. Both approaches have advantages and disadvantages (see e.g. O’Hare, 2000, for a review). For our purpose, statistical matching seems the most suitable method as it maintains the variance and covariances of consumption expenditures on the detailed categories in the integrated data set.

The idea of combining the two data sets is to use variables that are observed in both data set and to identify similar households regarding these variables. These matching variables, denoted by “z”-variables below, consist of characteristics that are relevant for household consumption. We include the net household income, the number of persons in the household and the family type, the marital status, the age and the gender of the household head, as well as her educational degree, social status (employee, freelancer, civil servant, unemployed, etc.) and her degree of employment (full-time, part-time). Furthermore, we included the information renter or owner and if the household is located in former East or West Germany. Renters are reporting their rent as part of their expenditures, while there is an imputed rent for owners. Importantly, the expenditures and especially the rent can still be significantly different between East and West Germany for otherwise identical households. The z-variables were similarly defined in both data sets for the matching procedure.

There are several methods to match two data sets. Two of the most commonly used ones are the propensity score matching (see Rosenbaum and Rubin, 1983) and matching based on the Mahalanobis distance (see Mahalanobis, 1936). While the performance of the two methods is similar, we chose the Mahalanobis-metric matching (see Rubin, 1980), as studies find that it performs better with large numbers of covariates (see e.g. Rubin and Thomas, 2000).

The Mahalanobis distance is defined as

$$d(z_i, z_j) = \sqrt{(z_i - z_j)' \hat{S}_Z^{-1} (z_i - z_j)}$$

where *i* refers to an observation from the SOEP and *j* from the EVS, and *z* denotes vectors of matching variables. $\hat{S}$ is the estimated (pooled) covariance matrix of the z-variables. The distance between a household from the SOEP and every EVS household is calculated and the smallest (weighted) distance identifies the match. Multiple matching of the same observation of the EVS is possible and we force every SOEP observation to get a match, which means that we do not apply a caliper.

Table 2 shows the matching frequencies of the EVS households for the 2008 matching (SOEP 2008 with EVS 2008). The first column shows the frequency a single EVS household is used in the matching. The second column counts the individual EVS households and the third column is the product of column 1 and column 2, counting the SOEP households. Single EVS observations are used up to 42 times but over 50% of the SOEP is matched to a unique EVS observation.

Table 2

Matching Frequencies SOEP-EVS Matching 2008

Frequency a Single EVS Household is Used in Matching	Number of Single EVS Households	SOEP Households	Share in Total SOEP Households
			in %
1	6 736	6 736	56,6
2	1 336	2 672	22,5
3	350	1 050	8,8
4	114	456	3,8
5	42	210	1,8
6	19	114	1,0
7	16	112	0,9
8	11	88	0,7
9	6	54	0,5
10	5	50	0,4
...			
15	1	15	0,1
...			
42	1	42	0,4
Total	8 654	11 902	100,0

Source: Own calculations with SOEP 2008 and EVS 2008.

Table 3 shows means of the z-variables before and after matching. The second column shows the SOEP values, which of course do not change due to the matching and the third column shows the EVS values. Additionally, we report the relative bias, the bias reduction and the two-sample t-test between SOEP and EVS attributes. While the means of most of the z-variables are getting more similar, we find sporadic worsening, e.g. in the variable “Couple with Children” or in the East Germany dummy. A special variable is the dummy for the bottom 10%. It declares a household that is under a certain income threshold. We included it since the poor are a bit underrepresented in the EVS as seen in the table. Overall, the bias reduction and the two-sample t-test indicate that the matching variables are reasonably well balanced in the matched data set.

Table 3
Marginal Distributions of Selected Z-Variables

Selected Z-Variables		SOEP	EVS	Bias in %	Bias Reduction in %	Two-sample t-Test p > |t|
Single HH	Unmatched	0,282	0,275	1,5		0,139
	Matched	0,282	0,287	- 1,2	21,3	0,358
Single HH with Children	Unmatched	0,063	0,055	3,4		0,001
	Matched	0,063	0,063	0,0	100,0	1,000
Couple without Children	Unmatched	0,346	0,356	- 2,0		0,049
	Matched	0,346	0,349	- 0,7	66,3	0,596
Couple with 1 Child	Unmatched	0,132	0,119	4,0		0,000
	Matched	0,132	0,129	0,9	77,8	0,501
Couple with Children	Unmatched	0,160	0,164	- 1,0		0,353
	Matched	0,160	0,156	1,3	- 30,4	0,328
Other Households	Unmatched	0,016	0,032	- 10,2		0,000
	Matched	0,016	0,016	0,2	98,4	0,877
Married	Unmatched	0,542	0,586	- 8,9		0,000
	Matched	0,542	0,542	0,0	99,6	0,979
East	Unmatched	0,265	0,263	0,5		0,644
	Matched	0,265	0,257	1,7	- 263,2	0,179
Home Owner	Unmatched	0,492	0,541	- 9,8		0,000
	Matched	0,492	0,491	0,2	98,1	0,887
Age of HH Head	Unmatched	53,439	51,950	9,5		0,000
	Matched	53,439	52,737	4,5	52,8	0,001
Male HH Head	Unmatched	0,577	0,646	- 14,4		0,000
	Matched	0,577	0,579	- 0,5	96,3	0,684
Log Household Net Income	Unmatched	10,231	10,303	- 11,0		0,000
	Matched	10,231	10,214	2,7	75,5	0,048
Bottom 10%	Unmatched	0,106	0,067	13,9		0,000
	Matched	0,106	0,102	1,4	89,9	0,318

Note: p > |t| refers to the probability value for the t-test statistic.

Source: Own calculations with SOEP 2008 and EVS 2008.

For the years 1998 and 2008, the SOEP surveys can be directly matched to the respective EVS survey, taking into account that income information in the SOEP is reported retrospectively for the previous year. For the year 1995, we matched the SOEP 1995 with the EVS 1998 after discounting the monetary variables to 1995. The SOEP 2005 is matched with EVS 2003, which is then updated to 2005. For the year 2015, we matched the SOEP survey 2012 (with income information of 2011) with the EVS 2008. The EVS 2008 is updated to 2011 for the matching and then all data is updated to 2015. Additionally, the population weights are adjusted to the marginal distribution of 2015 by the method of static aging. The underlying distributions are derived from the national accounts (VGR) and the micro census (Mikrozensus) published by the German Statistical Office. The expenditures are always uprated or discounted with the official consumer price index and income is nominally extrapolated with factors that also stem from the national accounts.

2.3 Integration of Consumption Data

The integrated data contains all relevant SOEP information and the detailed consumption expenditures from the EVS. Since the EVS excludes households with a net household income above 18,000 Euro per month and is known underrepresent very poor and very rich households, we adjusted household expenditures especially of rich households in the following way.

Firstly, we estimated a consumption function with all available EVS survey years of the form:

$$\frac{C}{Y} = \alpha + \beta_1 \ln(\frac{Y}{P}) + \beta_2[\ln(\frac{Y}{P})]^2 + \beta_3[\ln(\frac{Y}{P})]^3 + \beta_4[\ln(\frac{Y}{P})]^4 + X'\gamma + \varepsilon.$$

C is total consumption expenditures, *Y* is the total household net income, *X* is a vector of control variables, and ε is an error term. Both *C* and *Y* contain imputed rents for owner-occupiers. Control variables include household composition, the social status and the age of the household head, federal state dummies etc. We specify a flexible functional from by including a fourth-degree polynomial of real income and interactions of all income terms with the social status. We also include four polynomials of the household head's age and correct for quarterly effects (since the household report quarterly expenditures) and survey year effects. The R^2 of this regression is 0.89 and the model predicts the distribution of consumption share reasonably well. The estimated relation is not necessarily a "behavioral" consumption function but only used here for the prediction of consumption given income and a vector of other control variables.

We then replaced total consumption of "rich" households if the prediction from our consumption function exceeds their matched EVS total consumption, where a household is defined to be rich if net equivalent household income exceeds the 90th percentile of its distribution (about 70,000 Euro in 2015). For these households, all expenditures on all consumption categories are adjusted proportionally. The adjustments affect less than 10% of all households (e.g. 7.5% in 2015).

A second adjustment is made for poor households for whom the matching procedure yielded too high consumption. In particular, imputed total consumption of the bottom 20% exceeds their net income by up to 60% (see Table 4). Given the assumption that total household consumption cannot exceed disposable household income (incl. transfers) in the long-run, we have censored non-durable consumption for those households at 100% of gross household income.

Table 4

The Distribution of Consumption 2008 before and after Adjustments

Deciles, Percentiles of Household Net Equivalent Income	Household Net Equivalent Income		Consumption Expenditures in % of Net Income					
				thereof:			thereof:	
	Class Average	Highest percentile	Total	Food, Beverages	Rent and Imputed Rent	Total	Food, Beverages	Rent and Imputed Rent
	Euro per month		Before Adjustments			After Adjustments		
Bottom 5 %	643	711	164.2	26.8	16.3	113.5	18.4	35.8
1. Decile	767	820	145.0	23.1	14.3	112.2	17.5	36.9
2. Decile	905	993	106.6	18.6	12.9	101.8	16.5	32.2
3. Decile	1 085	1 172	95.7	16.7	11.2	95.8	15.6	27.6
4. Decile	1 245	1 322	93.4	15.0	11.1	92.4	13.9	25.2
5. Decile	1 404	1 489	87.2	13.6	10.7	89.9	13.0	23.3
6. Decile	1 582	1 679	81.4	12.2	10.5	84.7	11.6	21.8
7. Decile	1 795	1 912	78.6	11.0	10.0	82.5	10.5	20.6
8. Decile	2 058	2 217	72.8	9.8	9.2	78.4	9.6	19.0
9. Decile	2 455	2 771	70.9	8.8	9.7	77.9	8.7	18.2
10. Decile	2 996	.	53.6	5.9	8.7	69.1	6.6	17.3
Top 5 %	3 953	.	48.2	5.0	7.9	65.5	5.9	15.9
Top 1 %	8 902	.	32.4	2.8	6.8	55.4	4.0	14.6
Average	1 708	.	78.3	11.2	10.2	82.8	10.7	21.7

Source: Own calculation with the integrated data set of SOEP 2008 and EVS 2008.

The third adjustment refers to the rent. As we have information in the SOEP data for imputed rent and actually paid rent, we replaced the EVS information by the SOEP information. Only if the SOEP information is missing, we used the EVS information.

The comparison of total consumption expenditures[5] before and after these adjustments for the year 2008 clearly shows their effects at the bottom and at the top of the income distribution. The consumption of the bottom 20% slightly decreases after adjustments although expenditures for rent and imputed rent are better recorded. For the top 30% higher consumption expenditures result from these adjustments.

Comparing the distribution with the original one from the EVS 2008 in Table 5 shows similar average shares of consumption expenditures, but differences at the bottom and at the top. Another noticeable difference is the higher household net equivalent income in the EVS. This partly stems from different weighting factors in the SOEP and the EVS.

Table 5
The Distribution of Consumption in the EVS 2008

Deciles, Percentiles of Household Net Equivalent Income	Household Net Equivalent Income		Consumption Expenditures in % of Net Income		
				thereof:	
	Class Average	Highest percentile	Total	Food, Beverages	Rent and Imputed Rent
	Euro per month				
Bottom 5 %	681	757	127.5	22.2	44.7
1. Decile	828	903	112.5	19.8	38.8
2. Decile	1 028	1 151	95.7	15.7	29.6
3. Decile	1 257	1 359	92.0	13.8	26.0
4. Decile	1 458	1 557	88.1	12.6	24.2
5. Decile	1 655	1 753	83.6	11.4	22.5
6. Decile	1 864	1 978	80.3	10.3	21.4
7. Decile	2 108	2 253	77.6	9.6	20.2
8. Decile	2 435	2 649	73.5	8.5	18.7
9. Decile	2 960	3 350	68.8	7.3	16.7
10. Decile	3 677	.	56.7	5.0	12.7
Top 5 %	4 867	.	52.5	4.3	11.4
Top 1 %	8 090	.	44.4	3.1	8.9
Average	1 949	.	75.7	9.5	20.0

Source: Own calculation with the scientific-use file of EVS 2008.

[5] Note that total consumption here does not contain the so called "Other expenditure" in the EVS like private insurances, directly paid taxes like dog tax or inheritance tax, fees for public administration, private transfers etc.

Since there are no EVS surveys for the years 2015 and 1995, expenditures in our integrated data set must be updated, respectively discounted, for the respective years, where we have used the respective consumer price indices. To simulate quantity effects due to real income changes, we estimated Engel curves for the 12 commodity groups shown in Table 1.[6] . As before, all subcategories of the commodity groups were adjusted proportionally.

Table 6 shows the distribution of consumption expenditure for 2015 (SOEP 2011 with EVS 2008, updated). Here, "before adjustments" also means before updating of the expenditure data. We see similar effects of the adjustments as in Table 4 with the exception of an increase in consumption shares from the second decile upwards due to the updating from 2008 to 2015. However, the total consumption share slightly decreases, on average, compared to Table 4.

Table 6
The Distribution of Consumption 2015 before and after Adjustments

Deciles, Percentiles of Household Net Equivalent Income	Household Net Equivalent Income		Consumption Expenditures in % of Net Income					
	Class Average	Highest percentile	Total	thereof:		Total	thereof:	
				Food, Beverages	Rent and Imputed Rent		Food, Beverages	Rent and Imputed Rent
	Euro per month		Before Adjustments			After Adjustments		
Bottom 5 %	702	777	154.8	23.6	16.8	122.8	17.8	39.7
1. Decile	836	901	131.0	21.3	15.2	117.4	17.7	37.9
2. Decile	1 024	1 143	90.8	16.0	12.5	102.7	16.6	30.8
3. Decile	1 263	1 379	82.1	13.9	10.8	96.0	15.0	26.2
4. Decile	1 476	1 580	74.0	12.3	10.7	90.1	13.5	24.1
5. Decile	1 684	1 795	69.0	11.4	9.8	86.1	12.6	22.1
6. Decile	1 911	2 028	67.2	10.1	10.2	84.1	11.2	20.8
7. Decile	2 175	2 332	61.5	9.1	9.5	79.1	10.2	19.5
8. Decile	2 509	2 709	60.7	8.3	9.1	78.2	9.4	18.2
9. Decile	3 016	3 382	56.3	7.1	9.5	75.4	8.1	18.2
10. Decile	3 694	.	44.8	4.9	7.9	68.5	6.4	15.7
Top 5 %	4 907	.	41.2	4.2	7.2	66.4	5.8	15.1
Top 1 %	9 972	.	25.9	2.5	5.8	60.9	4.8	14.6
Average	2 043	.	64.1	9.3	9.7	81.4	10.3	20.6

Source: Own calculation with the integrated data set of SOEP 2011 and EVS 2008, updated to 2015.

[6] The income elasticities are estimated on the basis of the Engel curves using the QUAIDS model (see, Banks, Blundell and Lewbel 1997) and the 2008 EVS

2.4 Integration of the Wage and Income Tax Statistic

The integrated data base created from SOEP and EVS should give a representative picture of the overall income and consumption distribution in Germany. However, households at the very top of the income distribution are still very much underrepresented in the SOEP data despite the inclusion of the "high income" sample described above. Bach, Corneo and Steiner (2009, 2012) merged data from the Wage and Income Tax Statistic (LESt) to analyze the distribution and income tax burden of the richest decile in detail and especially of the top 1% and showed that the median as well as percentiles of the top incomes are distorted in the standard survey data. Therefore, we also integrate the LESt into our data set to get a representative picture of the distribution of the top income households.

Since the LESt is an administrative data set containing all income taxpayers of a year, the full information is only accessible by remote processing due to privacy protection. Since the direct integration of the LESt data into our integrated data base is not feasible, we have obtained the required income and tax information for the top percentiles from the LESt by remote processing and, on the basis of this information, have adapted the personal weights in our data set to match the distribution in the LESt. Table 7 shows the new distribution of our 2015 data after adjusting the weights.

While the net equivalent income declines from the first up to the 9th decile, it clearly increases in the top decile. The median equivalent income decreases from 1,795 euros to 1,647 euros. In the right part of the table, we plotted the distributions of net and gross income before and after reweighting. The share of net income in the top decile was 23.4% before reweighting and is now 28.1%, the one of gross income was 25.6% before reweighting and reaches now 31.1%. But at this stage, the editing is still incomplete because the very rich households are missing in the data.

Table 7
The Integrated Data 2015 after Adjusting the Weights to LESt Distributional Information

Deciles, Percentiles of Household Net Equivalent Income	Household Net Equivalent Income		Consumption Expenditures in % of Net Income			Distribution of			
				thereof:					
	Class Average	Highest percentile	Total	Food, Beverages	Rent and Imputed Rent	Household Net Income in %	Household Gross Income in %	Household Net Income in %	Household Gross Income in %
	Euro per month		After Reweighting			After Reweighting		Before Reweighting	
Bottom 5 %	676	748	128.0	17.9	40.7	1.6	1.2	1.7	1.2
1. Decile	803	850	119.6	17.9	38.8	3.7	2.7	3.7	2.8
2. Decile	955	1 055	106.2	17.0	32.7	4.6	3.7	4.9	4.0
3. Decile	1 160	1 259	98.2	15.3	27.6	5.7	4.9	6.1	5.3
4. Decile	1 362	1 452	94.3	14.7	24.9	6.7	6.0	7.2	6.5
5. Decile	1 545	1 647	89.0	13.0	23.4	7.6	7.1	8.1	7.5
6. Decile	1 753	1 859	86.0	12.4	22.0	8.6	8.3	9.4	9.1
7. Decile	1 988	2 119	83.0	10.7	20.9	9.9	9.9	10.5	10.7
8. Decile	2 292	2 483	78.9	9.9	19.7	11.3	11.6	12.0	12.7
9. Decile	2 751	3 127	79.7	8.8	19.5	13.8	14.6	14.6	15.9
10. Decile	3 575	.	66.4	5.8	15.7	28.1	31.1	23.4	25.6
Top 5 %	5 942	.	62.8	5.1	13.8	19.1	21.6	14.4	15.8
Top 1 %	15 232	.	55.6	3.7	11.7	6.9	7.6	4.7	5.0
Average	2 009	.	82.1	10.4	21.1	100.0	100.0	100.0	100.0

Source: Own calculation with the SOEP-EVS data set 2015.

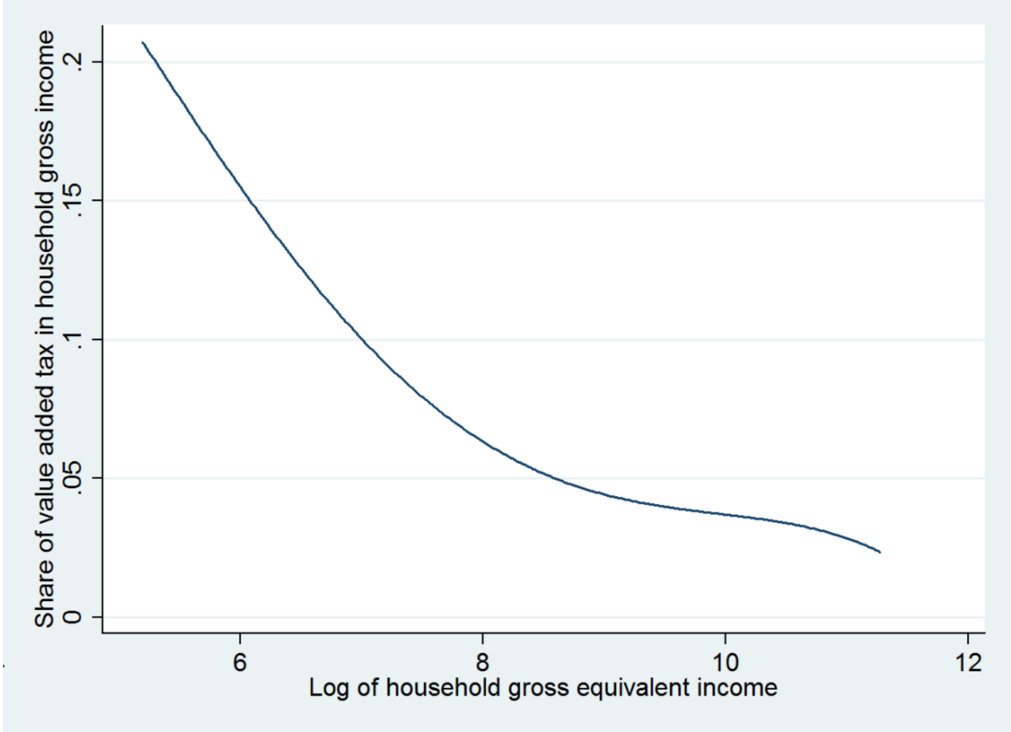
In a second step, we have edited the information provided from tax assessment to fit as closely as possible to the definition of household income in our integrated data base (for details see Bach, Corneo and Steiner, 2009, 2012). Since personal income tax records include most of the income sources and socio-demographic characteristics such as household composition or age, we get a good representation of the income distribution and the income tax burden up to the top income strata. Missing households in the lower income deciles that do not file a tax return are estimated based on demographic statistics and included as dummy observations as we are mainly interested in the higher income deciles. Transfer income is included in the income tax files except for means tested benefits of social security such as social assistance or housing benefits. We neglect these benefits since they do not occur in the high income strata.

The results of the edited LESt data aggregated to the 0.1-percentile level are processed via remote access and provided to us by the Research Data Center of the German Statistical Offices. Since the SOEP data is much less representative for the top income distribution than the LESt, we replaced the SOEP households in the top decile by the LESt observations on the 0.1-percentile level. For the analysis of the income tax burden, we replaced the income tax from the 5th decile up to the 9th decile for the SOEP cases with the information of the equivalent LESt percentiles. The reason is the so-called “middle-class bias” in surveys like the SOEP. Since the richest and poorest persons and households are undersampled, the middle-class is oversampled. Furthermore, the LESt data includes information on the local business and the corporate income tax burden on distributed profits, which could be simulated based on the observed capital income liable to personal income taxation. We allocate these taxes to our integrated data set by deciles.

Since the indirect taxes are not observed in the LESt data, we have to impute them for the top decile of the income distribution. Since it is hard to make an out-of-sample prediction for the consumption pattern of the rich households, we do not impute the detailed consumption expenditures directly like for the high income households observed in the SOEP but impute the indirect taxes by extrapolation from the data without LESt. For each indirect tax considered in our distribution analysis, we estimated empirical relationships between the gross household equivalent income and the amount of the respective indirect tax and then obtained the predictions from these estimated relationships to impute the respective tax for the 0.1-percentiles of the LESt cases.

We use two alternative specifications of empirical relationship between taxes and household income: The first is a semi-log linear regression of the share of the respective indirect tax in total household gross income on four polynomials of the log of gross equivalent household income. The second is a log-log linear regression of the log of the respective indirect tax on the same polynomial of gross equivalent household income. These two specifications are estimated separately for each indirect tax considered in our distribution analysis. We use both specifications to predict the respective indirect tax for the high income cases and compare the results. While the semi-log linear regression model with the share of a particular tax as the dependent variable (1) is our preferred model for most of the observations, it seems to overestimate the indirect taxes for the very rich households which seem to be better predicted by the log-log linear specification with the log of the tax as the dependent variable.

Figure 1
Estimated Functional Form of the Value Added Tax Share in Gross Household Income



Source: Own estimation with the integrated SOEP-EVS data 2008.

Figure 1 shows the estimated share function for the most relevant indirect tax, the value added tax. As the consumption share in household income drops with higher income, so does the share of indirect taxes. While households with a monthly gross income under 700 euros (about 6.5 on the scale), which also equals mostly their net income in this income bracket, consume almost their whole income, they have an average share of 13% value added tax in gross income. In contrast, the share of taxes in household income strongly declines with income because of the declining share of consumption.

Table 8 shows the distribution of consumption expenditures for our final integrated data set. Since only the indirect taxes are extrapolated for the LESt cases, the consumption is derived dividing the imputed value added tax by its average relation to total consumption expenditures. In the middle column, we show the final distribution of net income and in the right column the one of gross income. The average net equivalent income clearly increases in the top decile compared to Table 7 from 3,575 Euros to 5,865 Euros. Since the average net equivalent income is now 19,768 Euros per month in the top percentile, one could certainly guess that there is no household from the "real" top 1% in the original EVS survey. The consumption share declines at the top. The top decile obtains 29.2% of total net income and 31.7% of overall gross income, which shows the big concentration of income there.

Table 8
Integrated Data Base (SOEP-EVS-LESt)

Deciles, Percentiles of Household Net Equivalent Income	Household Net Equivalent Income		Total Consumption in % of Household Net Income	Distribution of	
	Class Average	Highest percentile		Household Net Income in %	Household Gross Income in %
	Euro per month		Incl. LESt		
Bottom 5 %	645	748	128.0	1.6	1.2
1. Decile	728	851	119.5	3.6	2.7
2. Decile	956	1 055	106.1	4.5	3.7
3. Decile	1 161	1 261	98.3	5.6	4.9
4. Decile	1 363	1 452	94.3	6.6	6.0
5. Decile	1 546	1 650	89.0	7.5	7.1
6. Decile	1 754	1 860	85.9	8.5	8.3
7. Decile	1 990	2 122	83.1	9.8	9.9
8. Decile	2 295	2 486	77.9	11.1	11.6
9. Decile	2 825	3 441	78.6	13.6	14.0
10. Decile	5 865	.	56.8	29.2	31.7
Top 5 %	8 030	.	51.0	19.9	21.6
Top 1 %	19 768	.	41.3	9.2	9.9
Average	2 072	.	78.8	100.0	100.0

Source: Own calculation with the final SOEP-EVS-LESt data for 2015.

2.5 Data Extrapolation and Uprating

Micro data is typically not up-to-date since the collection and editing procedures take some time. SOEP data has a delay of two to three years. The Income and Consumption Survey for Germany (EVS) is only conducted every 5 years, the survey we could use for the tax incidence project refers to the year 2008. The Wage and Income Tax Statistics available for our analysis refers to the year 2007-08 due to long-lasting assessment and editing procedures of the tax return data. For microsimulation analyses of the distributional impact of the current tax distribution or of counterfactual reforms we update the model data base to current and future periods.

"Static aging"-extrapolation of socio-demographics: To adjust the socio-demographic structures of our model data base we collect actual statistics and consensus projections on population, employment status (self-employed, employees, civil servant, pensioner, and unemployed), age structure, household and family composition, and dependent children. The weighting scheme of the microdata is consistently calibrated to the marginal distributions of these characteristics by using multi-dimensional numerical optimization algorithms.[7]

Income and expenditure growth: In a second step, the different income sources (self-employed income, wage income, capital income, pension income, other public transfers) are uprated using the nominal growth rates of corresponding aggregates derived from macroeconomic statistics and projections. The expenditures are also nominally uprated in a first step. Secondly, the real consumption is adjusted using Engel curves that account for the real income effects on the 12 different expenditure categories.

To uprate our integrated micro data base, we use public available macroeconomic statistics and projections, mainly from to National Accounts[8] which are also the basis for short-term forecasts and medium-term projections. In addition, we take into account information from the yearly micro census and employment statistics on, respectively, household composition and employment. For the medium-term projection we use official population projections[9] and the current macroeconomic projection of the federal government[10] which is used for the official tax revenue forecast.

[7] We use the „rake"-function of STATA module SURVWGT.

[8] https://www.destatis.de/EN/FactsFigures/NationalEconomyEnvironment/NationalAccounts/NationalAccounts.html

[9] https://www.destatis.de/EN/FactsFigures/SocietyState/Population/PopulationProjection/PopulationProjection.html

[10] http://www.bmwi.de/DE/Themen/Wirtschaft/Konjunktur-und-Statistiken/projektionen,did=385026.html

3 Microsimulation models

3.1 Tax-Benefit Microsimulation Model STSM

Our main data source, the SOEP, contains detailed income information at the individual and household level but does not directly record the assessed amount of the personal income tax and social security contributions (SSC) paid by the employee. To simulate these quantities at the individual and household level, we use the tax-benefit microsimulation model STSM (see Steiner et al. 2012). This also allows us to calculate the assessed amount of the personal income tax and SCC under hypothetical regulations and for different structures of taxpayers and employees. The STSM also contains a behavioural part which allows us to estimate the employment effects of tax-benefit reforms.

Table 9
Components of Net Household Income

		Income components	**Determined in the STSM**
1	+	Income from dependent employment	
	+	Income from capital	
	+	Income from renting and leasing	
		Income from self-employment, income from agriculture, forestry and business enterprise	
	+	Other income (pensions)	
2	+	Unemployment benefit I	X
	or	Unemployment benefit II	X
	+	Additional child benefit (“Kinderzuschlag”)	X
	+	Child benefit	X
	+	Parental-leave benefit	X
	+	Housing allowance	X
	+	Social assistance	X
	+	Other monetary transfers (education allowance, scholarships, apprentice allowance, special wage replacement payments for short-time work, maternity and widow’s allowance)	
3	–	Employees’ social security contributions	X
	–	Income tax	X
	–	Solidarity surcharge tax (“Solidaritätszuschlag”)	X
	=	Net household income	

Source: Steiner et al. (2012).

Table 9 summarizes the derivation of net household income in the STSM. Gross income by source is listed in the first section of the table, wage replacement benefits and transfers are listed in the second, and the income tax (including the so called "solidarity surcharge tax") and employees social security contributions in the third section. The last column of the table indicates which of the components are simulated in the STSM (denoted by X), and which components are directly recorded in the SOEP under the status quo. Some of the listed monetary transfers for entitled recipients, like unemployment benefits I and II, are recorded in the SOEP and need only be simulated if non-take up is considered or behavioural changes at the household level are taken into account. If policy changes are analysed, some or all of the gross income components also need to be simulated, of course. In the STSM, this can be done with or without accounting for behavioural employment changes.

Gross income is recorded by source in the SOEP which is important because income is not taxed uniformly in Germany. In particular, earnings from interest and dividends are subject to a flat rate withholding tax of 25% since 2009 and only a small part of public pensions have been taxed until 2004. Earnings from dependent employment (salaries, wages, bonuses, renumerations) are the main source of income for the great majority of households in Germany. Income from agriculture and forestry and entrepreneurial income is also included here. Pensions of former civil servants are also included, whereas employee pensions are included as other income. Whereas these income components are recorded at the individual level in the SOEP, income from interest and dividends is only recorded for households and not for individuals. For married or cohabiting couples it is assumed that this income is divided equally between spouses. Capital income and the income tax levied on it have to be partly imputed from the limited information on capital income in the SOEP (for details, see Steiner et al. 2012). The various monetary transfers listed in Table 9 are either directly recorded in the SOEP or can be simulated accurately from the available information (for details, see Steiner et al. 2012). For example, families with children are entitled to either the child benefit or the child tax allowance to guarantee a tax-exempt minimum income for children. These two alternative child benefits are granted by the tax authorities according to a higher-yield test. Both are not directly observed in the data but the relevant child benefit can be simulated given information on the number of children and household income.

Implementation in STSM

We calculate a higher-yield test between child benefit and child tax allowance. We first grant all households who are entitled to either of the two measures the child benefit. In a second step we calculate whether the child tax allowance would yield a higher tax relief than the child benefit. If so, we lower the income tax amount due by this amount.

Social security contributions comprise health and long-term care insurance, old-age insurance (public pensions), and unemployment insurance. For persons who are voluntarily insured in the social health insurance scheme, their social security contributions are deducted up to the maximum amount. After 2010 contributions to health and long-term care insurance are fully deductible for people covered by the social security system. For simulations after 2004 the increased tax exempted share of old-age pension provision expenditures is taken into account.

Except for expenses for commuting, professional expenses, which can be deducted from total wage income as far as they are individually verifiable, are not recorded in the SOEP. For these latter items, the lump-sum allowance for professional expenses is therefore deducted, in addition to the recorded amount of expenses for commuting. For pensions of civil servants the general tax allowance is deducted. Since 2005 an increasing share of pensions of former employees is taxed over a long transition phase.

The income tax amount is calculated by applying the progressive income tax schedule on taxable income and adding the so-called “solidarity surcharge”. We assume that all married partners choose joint filing. Thus, we add the taxable income of married spouses and apply the income tax tariff to half of this sum. Afterwards, the tax amount is doubled in order to get the tax amount due for married couples. If married spouses are living separately, we assume that they choose separate filing.

Unemployment benefits, special wage replacement payments for short-time work (“Kurzarbeitergeld”, “Winterausfallgeld”) and certain family benefits are not taxable themselves but affect the progressivity of the tax on taxable income (“Progressionsvorbehalt”). In the STSM the income tax rate for the other sources of income is thus simulated as if all income, including transfer income, was fully taxable, and the resulting tax rate is then applied to taxable income only.

3.2 Personal Income Tax Microsimulation Model

For the analysis of revenue and distributional effects of income tax reform, we use our Personal Income Tax Microsimulation Model (PIT-MSM) because it contains more information relevant for tax policy than the STSM. The model is based on tax return micro data from the official income tax statistics. We use the most recent data sets available at the Research Data Centers of the German Statistical Offices, the Wage and Income Tax Statistics for Germany (“Lohn- und Einkommensteuerstatistik”) of the tax year 2007 and the yearly Income Tax Statistics of 2008. Due to long-lasting assessment and editing procedures of the tax return data, more recent waves were not available when our study was undertaken. We therefore extrapolate the data to 2015 for our analysis of the current tax law and counterfactual reform scenarios. For reason of data protection restrictions the available scientific use files do not include the full information of all tax returns, in particular of the high income strata which account for a remarkable share of tax revenue. We use remote data access to analyze the whole data set.

The tax return data of the 2008 wave already include the new regulations of the company tax reform 2008 (increased tax credit and elimination of tax allowance for local business tax liabilities, new allowance for retained business income) and the surcharge for high taxable income as of 2007, which noticeably affect the income tax burden of high-income taxpayers. Moreover, the final taxation of capital income at the source introduced in 2009 was not implemented in those tax years. Thus, the tax return data includes capital income as far as taxpayers declared it to the fiscal authorities. Pure wage-tax returns of employees that do not file a tax return are only available for 2007.

We edit the data and recalculate the tax liability using a complex simulation program that captures most of the information available in the data. Using our PIT-MSM, adjusted gross income (“Gesamtbetrag der Einkünfte“) is derived very precisely from the different income items contained in the tax return data. The main itemized deductions such as insurance contributions and other special expenses or extraordinary expenses are simulated using information from the data set, social security contributions are simulated based on wage returns. Moreover, family taxation regulations are appropriately modeled by using the information on dependent children (child benefit, child allowance) and on joint taxation of married couples (full income splitting procedure).

A comparison of the taxpayers covered by the income tax statistics in the years 2007-08 with the population statistics shows that approximately 80% of the households in Germany file an income tax return or pay wage tax via the employer. Around 7 million singles and 2 million couples are not represented in the income tax statistics. These are persons or households that mainly live on tax exempted social assistance transfers or wage replacement transfers and have little taxable income from other sources. In order to describe the entire population we include households from the SOEP with a low probability to file a tax return. These non-filers are estimated by a standard probit model based on the socio-demographic variables provided in both datasets. For the analysis of the full tax return data via remote data access we include the missing non-filers as dummy observations.

The detailed data editing and bottom-up simulation of the tax burden allows us to implement recent income tax amendments as well as counterfactual tax reform scenarios. We consider the main reforms of the past years such as the final taxation of capital income at the source as of 2009, the enlargement of deductions for healthcare provisions as of 2010, the repeated increases of the tax-free basic allowance and the children's benefit and allowance, and changes in the tax schedule. We use "static aging" to uprate the socio-demographic structures of the taxpayers to future periods, and uprate monetary variables using nominal growth rates of the main income sources derived from macroeconomic statistics and projections, as described in section 2.5 above.

The PIT-MSM calculates the assessed personal income tax liability including solidarity surcharge. Moreover, the local business income tax and the corporation income tax on distributed profits are simulated using the information on capital income declared by the taxpayers. Since the income tax return data do not contain information on the taxpayers' participation on retained profits of incorporated firms liable to the corporation income tax, we cannot allocate retained profits to individual taxpayers.

The model is used to calculate the first-round revenue and distributional effects of the statutory tax rules and of counterfactual tax reform scenarios. It includes no explicit tools to account for behavioral response to taxation, e.g. with respect to employment, investment or tax avoidance and evasion.

3.3 Modelling VAT and Excises

The detailed information from the EVS on consumption expenditure of around 150 items allows for a proper simulation of the relevant indirect taxes, i.e. VAT and the main excises. We assume full shifting of the tax burden to final consumers and calculate the share of the indirect taxes in the different consumption expenditure items.

For VAT, the main tax exemptions and reduced tax rates could be modeled in sufficient detail. In some cases when expenditure items are not sufficiently differentiated, external information is used to estimate the share of reduced taxation or exemptions, for instance for public transport (only local transport is eligible to the reduced rate) and for sporting and cultural services which are partly taxed at reduced rates and partly exempted. For the main VAT exemptions we assume that the input tax on the intermediate consumption of the suppliers is also shifted to final consumption since the supplier it is not allowed to deduct or to reimburse the input tax. This is in particular the case for housing expenditure which, on average, amounts to one quarter of total consumption expenditure. We estimate the input tax share of the different components of housing expenditure and differentiate between current VAT on running costs as well as historical VAT on the housing investments.[11].

The other excise taxes, namely the taxes on energy commodities, tobacco, alcohol, lottery and betting, and on cars are simulated using the available information on household expenditures. Since the tax base of the excises is related to the quantities consumed rather than to expenditures, we build up tax simulation modules that derive the quantities in each simulation year. For example, we combine price data for the energy goods gasoline, diesel, electricity, fuel oil and petroleum gas with our expenditure data to identify the consumed quantities and then apply the specific tax rates for the relevant year. Expenditures on tobacco and

[11] In a first step, we estimate running costs which are largely liable to VAT from the observed rent or, in case of home owners, the imputed rent applying average rates from rent index statistics used by the real estate industry. In a second step, we calculate the long-run VAT burden on construction investment (for which no input tax could be credited) by applying average multipliers derived from simple investment models. For the details of these calculations, see Bach (2005).

alcohol are not reported explicitly in every EVS survey. Moreover, there is considerable under-reporting of tobacco expenditures compared to macroeconomic statistics. Therefore, we use imputation methods (a tobit model). The motor vehicle tax is directly reported in the data, so we do not have to simulate it but extrapolate it for the following years.

We also estimate the indirect taxes paid in the company sector and assumed to be shifted onto household consumption. This is the case for the energy taxes paid on intermediate energy consumption, the land tax on company buildings, or the taxes on company vehicles. Again, we assume full shifting to the prices paid by households. Using input-output matrices that allocate the tax volume paid by companies to the consumer goods and services they produce, we relate these taxes to the households based on their expenditures.

For the recurrent land tax and the land transfer tax we simply assume that the tax burden is shared equally between landlords and tenants in the case of rented dwellings. For owner-occupied properties the owner is assumed to bear the entire tax burden. We use EVS and SOEP data on housing expenditure and SOEP data on the value of real estate properties surveyed in the waves 2002, 2007 and 2012 to allocate the burden of real estate taxation.

References

Bach, Beznoska and Steiner (2016a): Wer trägt die Steuerlast in Deutschland? – Verteilungswirkungen des deutschen Steuer- und Transfersystems. *Politikberatung kompakt 114. DIW Berlin*.

Bach, Beznoska and Steiner (2016b): Wer trägt die Steuerlast in Deutschland? – Steuerbelastung nur schwach progressiv. *DIW Wochenbericht, 51+52.2016*, 1207-1216

Bach, Beznoska and Steiner (2016c): Who bears the tax burden in Germany? Tax structure slightly progressive. *DIW Economic Bulletin, 51+52.2016*, 601-608.

Bach, Corneo and Steiner (2009): From Bottom to Top: The Entire Income Distribution in Germany, 1992-2003. *Review of Income and Wealth*, 55, 331-359.

Bach, Corneo and Steiner (2012): Effective Taxation of Top Incomes in Germany. *German Economic Review*, 14, 115-137.

Banks, Blundell and Lewbel (1997): Quadratic Engel Curves and Consumer Demand. *The Review of Economics and Statistics*, 79, 527-539.

Frick, Goebel, Grabka, Groh-Samberg and Wagner (2007): Zur Erfassung von Einkommen und Vermögen in Haushaltssurveys: Hocheinkommensstichprobe und Vermögensbilanz im SOEP. *Data Documentation 19. DIW Berlin*.

Mahalanobis (1936): On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science of India*, 12, 49-55.

Markus, Siegers and Grabka (2013): Preparation of Data from the New SOEP Consumption Module: Editing, Imputation, and Smoothing. *Data Documentation 70. DIW Berlin*.

O'Hare (2000): Impute or Match? Strategies for Microsimulation Modeling. In Gupta & Kapur (ed.): *Microsimulation in Government Policy and Forecasting*. North-Holland.

Rosenbaum and Rubin (1983): The Central Role of Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 4155.

Rubin (1980): Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics*, 36, 293-298.

Rubin and Thomas (2000): Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95, 573-585.*

Steiner, Wrohlich, Haan and Geyer (2012): Documentation of the Tax-Benefit Microsimulation Model STSM. Version 2012. *Data Documentation 63. DIW Berlin*.

Wagner, Frick and Schupp (2007): The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. *SOEPpapers on Multidisciplinary Panel Data Research 1, DIW Berlin*.