

*Marek Kośny**

UPPER TAIL OF THE INCOME DISTRIBUTION IN TAX RECORDS AND SURVEY DATA. EVIDENCE FROM POLAND

Growing interest in the situation of the rich and in changes in the upper tail of the income distribution entails search for specialized sources of data. The majority of available data come from surveys, which are burdened with the problem of the reluctance of the rich to reveal their income and assets. This is the incentive to look for alternative data sources. One of such sources are tax records. Irrespective of their credibility (limited by tax avoidance and tax evasion), the crucial feature of data collected by tax offices is its completeness for the population of taxpayers: refusal to file a tax return or providing false information is punishable by law. The paper compares the data gathered within the Household Budget Survey, which is commonly used for research and administrative purposes in Poland, with the complete dataset of tax records for Dolnośląskie province (over 2 million taxpayers). Analyses show large discrepancies in the results obtained for both datasets in the upper tail of the income distribution. The observed differences are important for the proper assessment of the actual situation of the rich.

Keywords: affluence, top incomes, survey data, tax records

JEL Classifications: D31, H24, C80

DOI: 10.15611/aoe.2019.1.03

1. INTRODUCTION

Studies on the distributions of income had been primarily focused for many years on the situation of the poor. Vast literature on poverty issues discussed both the changes in the distribution of income, in particular the inequality and polarization, and social consequences of these changes which are relative deprivation, social exclusion and lack of social stability (see e.g. Ferrer-i-Carbonell 2005; Esteban and Ray 2011). The beginning of this century brought an increase in interest in other areas of income distribution. The first is the income situation of the middle class. Particular interest in this area was related to the financial crisis that began in 2008. The increase in unemployment and limited access to credit contributed to the strong deterioration in the situation of households classified as middle class. This

* Department of Econometrics and Operational Research, Wrocław University of Economics.

process became the subject of detailed analysis not only from the point of view of changes in the level of income, but also the widely understood economic security of households in this group; see e.g. Osberg (2009); also Wheary et al. (2007) for studies prior to the start of the crisis. The second area of research, which has become popular over the last decade, concerns analysis of the situation of the most affluent group; see e.g. Piketty (2013). While the situation of the poor is of researchers' interest because of this group itself, to improve the chances of its members for the normal functioning in society, ensuring equal opportunities and the minimum resources necessary for life, or to provide at least basic health care, interest in the situation of the richest people is mainly due to the importance of this group for society as a whole. This group, although relatively small (depending on the definition, but usually consisting of no more than 1 to 10 percent of the wealthiest individuals), generates a big share of income and tax revenue. Because of this, it is able to accumulate significant savings and to invest. Through the process of jobs creation and influence on other aspects of the economy, the condition of the richest part of the population directly affects the situation of the less affluent.

Large discrepancies between the level of income of the poor and the rich are causing problems with the description of the wealth and its diversity in the whole of society. The arithmetic mean and the Gini coefficient, commonly used in assessing the average level and inequality in the distribution of income, depend largely on the structure of the upper part of the income distribution. Sometimes a few very high incomes are able to distort the results for the entire population. In this context, precise estimates of the wealth of the richest parts of the population may have a substantial impact on the assessment of the situation, for example in determining the level of inequality and the direction of its change. Therefore the issue of the assessment of the situation of the wealthiest people is inextricably linked with the search for data sources which provide the most accurate and reliable information on this group.

Analyses presented in the literature are based on data from two main sources. The official surveys conducted by statistical offices and aimed at collecting data on the income and expenditure of households in the given country, are the first type of such sources. For many years, data of this kind were the primary (and virtually only included in the study) source of information on income distributions; see Human Development Reports 1990-2011 and Atkinson and Brandolini (2001) for a review of the use of 'secondary' data sets. But the most serious drawbacks of survey researches

concern the assessment of the situation of the wealthiest. The most important disadvantage of this type of research is the relatively small numbers of individuals (or households – depending on the measurement unit) with very high incomes, and refusal to participate in the study. An attempt to solve this problem is oversampling the highest income individuals (households). There are several such surveys, data from some periodic studies of this type have been collected in the Luxembourg Wealth Study Database; see Sierminska et al. (2006) for a description. Research of this type is also conducted within the Household Finance and Consumption Survey; the results for Poland are presented in the National Bank of Poland (2015) and the National Bank of Poland (2017). However this only allows for reducing, not eliminating, the problems associated with surveying the most affluent part of the population.

An attempt to solve these problems, related to survey researches, is the use of data from fiscal data sets. Although for decades such data were considered rather useless in the analysis of wealth, in the last decade tax records have been one of the primary sources of data in this area; see, among others, Piketty and Saez (2003); Leigh (2009), Atkinson et al. (2011). Despite the undeniable flaw, which is the burden on the data as a result of tax evasion and tax avoidance, their main advantage is the availability for long periods of time. In the long run, in the case of some countries covering more than 100 years, the availability applies only to the basic characteristics of the distribution (e.g. the quantiles which are very useful in the process of approximation of the original distribution of income). Yet some tax data are also available in the form of individual records. Information on the distribution is then complete, assuming that the data cover the entire population, but such data are typically available for much shorter periods which is related to the development of information technology. The main advantage of the fiscal data is the completeness of the data on taxpayers' income, and sometimes also on some other variables such as age and place of residence, for the population of taxpayers; of course, not all individuals are taxpayers. Unfortunately, these sets have same disadvantages as the aggregate tax data.

Both the survey data and those from tax returns are used as an independent basis for assessing the situation of affluent people, an extensive description of the empirical research on the wealth is presented by Leigh (2009). However, both data sets may also serve for each other as a point of reference. The results of the analysis, for the same country and period, may in fact vary largely, depending on the data. A comparison of survey and tax data for the United States conducted by Burkhauser et al. (2009) indicates

that the results for the top one percent of the population are much lower for the survey data, high convergence was observed in the rest of the income distribution. In the case of discrepancies, however, there is a natural question: which results should be considered more reliable? The answer to this question may result from the analysis of the characteristics of both data sets – what types of data are collected, what kind of information is generally ignored, etc. The structure of the data set and the method of gathering data may also be of interest and provide some interesting, additional information.

Detailed analyses of the income of wealthy people in Poland were carried out mostly on the basis of the survey data (see Brzeziński 2010) and journal publications of the so-called ‘rich lists’ (see Brzeziński 2014). Attempts are also made to combine data from surveys with aggregated tax data, published by the Ministry of Finance (see Bukowski and Novokmet 2017a, 2017b). However, no analyses of the upper tail of income distribution have been conducted so far for Poland on the basis of individual data from tax returns.

Therefore, the subject of this paper is to compare the situation of wealthy individuals in Poland resulting from the data collected in two independent sets. The first data set comes from the household budget survey, conducted annually by the Central Statistical Office. The second contains information on incomes taken from the official register kept for tax purposes by the tax offices of Lower Silesia (one of the 16 Polish regions).

Both data sets include individual data, however, they differ in many respects. The most important differences concern two areas. Firstly, the transfer of data to tax offices by taxpayers is compulsory, and the possible provision of false data is punishable by an administrative penalty. In the case of household budget surveys, participation in the survey is voluntary, and all values are collected based on respondents’ declarations, which allows for both a refusal to participate in the research (non-response) and potentially providing false data (i.e. cheating; the truth of the values given is not subject to verification). The second significant difference is the scope of available information. While in the budget survey, data on various sources of income are collected, tax returns include only income sources that are subject to personal income tax. From the point of view of the purpose of the conducted analysis, i.e. the assessment of the distribution of the highest incomes, the first aspect appears to be of key importance. The lack of information about income not subject to income tax is potentially less important in the analysed area than in lower areas of the income distribution. A more detailed reference to the indicated differences is presented further in the article.

The structure of this paper is as follows. The second section presents a brief characterization of the two data sets and the definition of income. The third section describes the methods used to analyse and evaluate data. The fourth section presents the results of the empirical analyses. The fifth section briefly discusses the results, and the last section concludes.

2. SELECTED SOURCES OF DATA ON INCOME DISTRIBUTION IN POLAND

2.1. Household Budget Survey

The widest survey research on the economic situation of households in Poland is the Household Budget Survey (HBS) conducted yearly by the Central Statistical Office. The outcome of this study is very important for an assessment of living standards in Poland and in determining the number of official social policy indicators. In particular, HBS results are one of the important factors taken into account while deciding on the level of the minimum wage. The HBS data is also used to assess the extent of poverty, to calculate estimates of taxes paid, and to determine the level of social benefits. Data on consumption and prices of goods are also used to set weights for calculating the consumer price index.

To allow, at least a partial, inference about changes in the financial situation of households, about half of the sample is surveyed in the consecutive year. The method of sample selection is a rotating panel with the partial replacement of the sample. This means that households that were surveyed for the first time in a given year are also asked to participate in the survey in the following year. Before the year 2000 a single household could be surveyed for up to four years, but since 2000 the maximum period of participation in the study has been reduced to only two years. This is a very strong limitation from the point of view of the possibility of observing changes over time.

The sampling scheme in the HBS is two-staged, based on the territorial units record TERYT. The units at the first stage are primary sampling units that are based on statistical regions defined in the TERYT system. Before the sampling, primary sampling units are stratified by province and city size. As a result the number of primary sampling units from a given strata is proportional to the estimated number of residential dwellings in this strata. The secondary sampling unit is dwelling, drawn within the primary sampling

units: two dwellings are drawn each month (in total 24 per year, and an additional pool of residential dwellings in cases of refusal to participate in the survey). Since 2005 the first-degree sample includes 1566 primary sampling units, which gives 37,584 households; due to the incompleteness of the information provided, the actual number of households included in the survey is slightly different each year. Owing to such a sampling scheme, the final data set is representative for both the Polish population and the population of each of the 16 provinces. More detailed information on the history and methodology of this study can be found, for example, in Central Statistical Office (2011b) and Kordos et al. (2002).

Because of the relatively large proportion of households refusing to participate in the survey, above 50 percent depending on the round of research, to minimize the impact of differences between the structure of the sample and the structure of the entire population, in the final data sets weights are assigned to individual households. These weights, which denote the inverse of probability of selecting a particular type of household, are estimated on the basis of the 2011 Census. They are not, however, taken into account in the further analysis. As the average weights are systematically lower for the richest (the group of 0.1%, 1% and 10% of the richest) than for the whole sample, their inclusion would result in a decrease in the value of the averages and the respective quantiles. The published weights are, of course, justified in terms of the sampling scheme but do not comply with distribution of income. As is clear from other studies – see e.g. Moore et al. (2000) for a broad discussion of survey data errors – non-response rates are much higher among households with the highest income, which means that higher weights should be applied for this part of the sample.

To maintain comparability with the tax data, further analysis will take into account data for the years 2006-2010, covering a period of rapid economic growth in Poland (2006-2007), and then stagnation, resulting from the financial crisis (2008-2010).

2.2. European Union Statistics on Income and Living Conditions

Besides the household budget survey, income data are also collected within other representative surveys. The most important of these is the European Union Statistics on Income and Living Conditions (EU-SILC). The research is a continuation of the European Community Household Panel (ECHP), conducted in the years 1994-2001. The implementation of the EU-SILC study began in 2003. The study is coordinated by Eurostat, while

its implementation is carried out by statistical offices of individual countries: in Poland the study has been conducted by the Central Statistical Office since 2005. The household budget survey and EU-SILC are similar in terms of the sampling method. As in the case of EU-SILC, the sample is drawn in a two-stage process and is based on the TERYT system, in an analogous way as in the household budget survey (see Central Statistical Office 2008, pp. 35-55), however significantly fewer households take part in this survey (around 13,000 each year). Each randomly selected household is surveyed (direct interview) once a year.

In the context of this analysis, the advantage of the EU-SILC study would be gathering information about the individual incomes of household members, but its important drawback is that the final data set from this study comprises both information obtained directly from respondents as well as imputed data. Therefore, given similar levels of inequalities for both data sets¹, further analysis will be solely based on the data from the household budget survey. An extension of the analysis to other data sets (including EU-SILC or HFCS) are potential directions for future research.

2.3. Tax records

Among the data sets collected by the public institutions, particularly useful in studies on distribution of income and wealth, are data collected by tax offices. Fiscal data sets include not only information on income but also the amounts of taxes paid and contributions for social and health security. Given the diversity of tax forms, it also allows for the distinguishing different social groups – particularly pensioners and the self-employed. In addition, since 2007 some personal income tax forms include a question concerning the number of dependent children – this information is required to be able to benefit from tax exemption. The number of children, even though not always complete (some taxpayers have income too low to benefit from this exemption), allows for the partial analysis of income and public-law burdens in the context of the demographic situation of taxpayers.

¹ The value of the Gini coefficients in both studies was initially very similar, but later systematically lower in the EU-SILC, which suggests the existence of even greater discrepancies between income in the upper tail declared in this survey and in the tax data. Gini coefficients for HBS 2006: 0.340; 2007: 0.340; 2008: 0.339; 2009: 0.336, 2010: 0.342 – see Central Statistical Office 2017, p. 46; for EU-SILC 2006: 0.333; 2007: 0.322; 2008: 0.320; 2009: 0.314; 2010: 0.311 – see Central Statistical Office (2011a, p. 3).

The main advantage of administrative data is – in terms of their use for research purposes – the compulsory character of its collection. Units (individuals, companies, institutions) covered by a register are legally obliged to provide the required information. Refusal to provide information or providing false information is punishable. Administrative coercion does not entail, of course, the completeness and full credibility of the collected data. Some information is intentionally or unconsciously, falsified or concealed which is usually the basis for criticism of this type of data. In the context of the analysis of the situation of high-income taxpayers, it should be noted that the potential incentives to under-report income (in the form of tax avoidance or tax evasion) increase with income and tax duty. The potential threat of being punished, however, means that concealing or misreporting of information cannot simply result from the reluctance to participate in the research, as is in the case of surveys.

Information on aggregated income and taxes paid, estimated on the basis of tax returns, is published annually by the Ministry of Finance. These publications, however, only include income ranges corresponding to the currently applicable tax thresholds (see e.g. Ministry of Finance 2011), so they are not useful in studies on income distribution. Therefore, the analyses presented in the next part of this paper will be based on individual, anonymised data, in which one record corresponds to a single income tax return. The data set includes tax returns filed between 2007 and 2011, reporting the income for 2006-2010. Data are collected in panel form, which allows tracking changes in the situation of individual taxpayers in consecutive years.

The data set includes all the tax returns for personal income and capital gains, filed during that period by residents of the Lower Silesia province (one of 16 provinces of Poland). Therefore the analysed data set, although covering more than 2.3 million taxpayers (the number of taxpayers who file the tax form in any given year is lower – about 1.8 million), cannot be regarded as representative for the population of Poland. Therefore, direct comparison between the results based on survey outcomes and tax records will be carried out for the part of the HBS data set, concerning the Lower Silesia province.

2.4. Definition and sources of income

In HBS incomes from many sources are recorded including, among others, loans, insurance payments and income in-kind. Later in this paper, however, only income from sources subject to personal income tax will be taken into

account due to the possibility of the direct comparison of data from this source with data from tax records. Yet other sources of income recorded in the HBS (such as alimony, social assistance benefits) are not important in the case of wealthy individuals. The following income sources will be considered: income from employment (group 901), income from self-employment (group 902), social security benefits (especially pensions, group 905) and revenue from the sale of products and agricultural services (group 911111). Revenue from the latter group is only partially subject to personal income tax, but there is no possibility of the separation of taxed and untaxed income based on the available data. For this reason all income from this group is included.

Income from these sources, declared by the respondents, is the gross income before personal income tax, but after the deduction of the compulsory social and health security (if applicable). It is recorded in the one month of the year in which the household is surveyed (each month about one twelfth of the whole sample is surveyed). Therefore, when making direct comparisons between the survey and tax data, declared income will be multiplied by 12.

In the case of tax data, income is directly related to the type of tax return filed. The study will take into account the following types of tax returns, characterized by the symbol of the form on which the return is filed.

- Form PIT36, covering income from self-employment and certain types of agricultural production, taxed at the standard, progressive tax scale.
- Form PIT36L, covering income from self-employment and certain types of agricultural production, taxed at a flat rate.
- Form PIT37, filed by taxpayers not running their own businesses, covering income mainly from hired work.
- Form PIT40, including income from hired work, if the tax return is filed by employer on behalf of the taxpayer (upon the request of the taxpayer).
- Form PIT40A, including income from pensions, if the tax return is filed by the Social Insurance Institution on behalf of the taxpayer. If the taxpayer gets an additional income (or files the tax return himself/herself for some other reasons), forms PIT36 or PIT37 are used.
- Form PIT38, including capital gains. Income from interest on deposits and from some investment funds is not declared, since the tax on income from these sources is paid automatically by the bank or fund management company. This means the underestimation of the actual income from this source.

Based on the type of tax form filed, in the next part of the paper three categories of income will be considered. The first will include income from

employment and social security – forms PIT37, PIT40 and PIT40A (equivalent to groups 901 and 905 in HBS). The second category will include income from self-employment and certain types of agricultural production – forms PIT36 and PIT36L (corresponding to groups 902 and 911111 in HBS). The third category – capital gains, reported on form PIT38 – are not recorded in HBS. A taxpayer receiving income from capital has to file form PIT38. Other forms, however, are filed alternatively, depending on the source of income, however form PIT36 has priority over PIT37. This means that in the case of income from both self-employment and some other source, the taxpayer files form PIT36, placing in it the income from various sources. This can lead to some overestimation of income from self-employment.

Forms PIT36 and PIT37 may be filed by the taxpayer individually or jointly with a spouse or dependent child (in the case of single parents). Joint taxation is, however, facultative and the taxpayer may, but does not have to use it. This means that the choice of form of taxation is not uniquely determined by the taxpayer's family situation, as already mentioned the data collected by the tax authorities allow only the partial identification of the family situation. Therefore the data from tax returns will be given for each individual taxpayer, even if joint taxation option has been chosen, most of the information (including income) is declared for each taxpayer separately.

Such a definition of the basic unit is inconsistent with the definition adopted for the HBS data, where the basic unit is the household, but it is not possible to find an unambiguous solution to this problem because in the HBS, income is not recorded separately for each individual in the household. Any division of household income among individuals (based on household composition) would require estimates of the number of people earning income subject to personal income tax. It also needs assumptions about intra-household income distribution, which usually depends on income level and is very asymmetric in the case of households with high incomes. Therefore the original structure will be maintained for both data sets, and the characterized differences will be taken into account in the analysis.

3. METHODS

To assess the situation of the wealthiest part of the population on the basis of data from both presented data sets both traditional methods, usually applied in this type of analysis (see Brzeziński 2010), and some less common ones will be used. To the first of these groups belong estimates of

quantiles of the income distribution, the average income in groups of individuals (households) with the highest income and the income shares of the top p percent, expressed by the formula:

$$IS(\mathbf{x}, p) = \frac{\sum_{i=\lceil n \cdot p \rceil}^n x_i}{\sum_{i=1}^n x_i} \quad (1)$$

where $\lceil n \cdot p \rceil$ denotes rounding up to the integer closest to $n \cdot p$, n – number of individuals (households) and p – quantile of the income distribution ($0 \leq p \leq 1$) for $\mathbf{x} = (x_1, x_2, \dots, x_n)$ being an ordered vector of non-negative values $x_1 \leq x_2 \leq \dots \leq x_n$, representing the distribution of income.

Typical measures that are also popular in the analyses of affluence presented in the literature are indexes of mobility, expressed as the probability of leaving (by the person or household) the group of top income earners (depending on the definition of this group) in a given period (see e.g. Saez and Veall 2005).

Besides these popular measures, the analysis presented in this paper will be expanded on the analysis of the curves, based on point Zenga index. This analysis refers to the idea that underlies Zenga's indexes and describes the relationship (and its changes) between groups of people (households), whose income is below and above a specified level. The measure, which is a generalization of the quotient of the quantiles of the income distribution, holds information not only on the situation at a given point (quantile) of income distribution, but on entire groups – the lower and the upper. This is especially important for the analysis of wealth because it takes into account the effects of very few, but very high incomes. A point measure of a relative income change, denoting changes in the distribution of income in a given period of time will be defined as (see Košny and Yalonetzky 2015):

$$RIC(\mathbf{x}^0, \mathbf{x}^1, p^0, p^1) = \frac{\bar{M}(\mathbf{x}^1, p^1)}{\bar{M}(\mathbf{x}^1, p^1)} - \frac{\bar{M}(\mathbf{x}^0, p^0)}{\bar{M}(\mathbf{x}^0, p^0)} \quad (2)$$

where \mathbf{x}^0 and \mathbf{x}^1 denote distributions of income at the beginning and the end of the period, respectively. The line separating two groups – the poorer and

the richer – is set by p^0 and p^1 for these two moments. The lower and the upper mean will be defined as

$$\bar{M}(\mathbf{x}, p) = \frac{\sum_{i=1}^{\lfloor n \cdot p \rfloor} x_i}{\lfloor n \cdot p \rfloor}$$

and

$$M^+(\mathbf{x}, p) = \frac{\sum_{i=\lfloor n \cdot p \rfloor + 1}^n x_i}{n - \lfloor n \cdot p \rfloor}$$

where $\lfloor n \cdot p \rfloor$ denotes rounding down to the integer closest to $n \cdot p$.

The values of the point index given by (2) range between -1 and 1 . They reflect changes in the average income of the lower group with respect to the upper group. For a given p , *RIC* indicates the change (expressed in percentage points) in the share of the average income of the lower group (100 percent $\cdot p$ of the population) in relation to the average income of the upper group (100 percent $\cdot (1-p)$ of this population).

The presented analysis was deliberately limited to the use of descriptive statistics, showing discrepancies between the distributions of income recorded in tax returns and declared in the HBS. Extending the analysis to methods of mathematical statistics, in particular tests of significance of observed differences, would entail serious doubts as to the possibility of their correct application and interpretation of results. In the case of data obtained from tax returns, it is not clear whether we are dealing with a population (if the analysis is limited to the Lower Silesian province, as it was done in the article), or a sample. Assuming that this is a sample, the problem of the independence of samples and their size (and the related sensitivity of significance tests) appears, which significantly affects the credibility of the conclusions.

4. EMPIRICAL EVIDENCE

4.1. Quality of data

Both the analysed data sets are burdened with systematic errors. In HBS, as with other surveys, the basic problem is non-response. This issue is of particular importance in the context of research on wealth. In the case of

households with very high incomes, a very important factor is the desire to protect privacy and the reluctance to share with others the details of both income and consumption. There is also need to fill in quite a complex questionnaire on the structure of consumption, which as a labour-intensive activity is discouraging; the small gifts offered for the participation in the survey is not a sufficient incentive, especially for high income households. Besides the non-response rates, depending on the household income², the deliberate concealment of information has to be classified as non-random distortions of the survey results. In some cases, for example if income is derived from illegal sources, the respondent may misreport the income because of the fear that the data collected will be offered to the tax authorities (tax evasion, resulting from underreporting income from work, is not usually considered as very negative in Poland). In other situations the problem results mostly from social norms, spending on prostitution and drugs is usually zero in HBS data. The third source of systematic errors, important from the point of view of this analysis, is the lack of knowledge of the respondents regarding the definition of certain categories or amounts of their actual spending in these areas. The values of personal income tax declared by the respondents, are an example of this type of error as distribution of personal income tax in the sample is completely different from that of the entire population. Finally, the fourth source of error in assessing the actual situation of households is the lack of some (income) categories in the recorded information. In the context of the most affluent households, capital gains are the most important category.

Besides non-random errors, all surveys also involve random errors. As the group of the most affluent people is relatively small, samples that are said to be representative for the entire population may not adequately reflect this subgroup. Certain drawbacks are also included in the sampling scheme itself (not actual data in the TERYT system) and in the process of collecting the information (reporting the income and expenditure for a single month, filling in the questionnaire by respondents).

Other problems are associated with the data from tax records. In this case, the most important source of discrepancies of the reported data with the actual income is tax evasion (e.g. failure to file a tax return at all) or tax avoidance (formally legal but not approved by the authorities, reduction of tax duty such as transferring the income to tax havens). Although both

² The Central Statistical Office publishes only the overall percentage of households that refused to participate in the study (see Central Statistical Office 2011b).

phenomena relate to taxpayers at all income levels, in the case of lower-income taxpayers the basic problem is the income from unreported employment. In the case of taxpayers with the highest incomes there are many legal (in a formal sense) possibilities to reduce the taxable income by transferring it to sources (or areas) with lower tax rates.

Table 1
Sample size and quantiles of income distribution for tax records and HBS data

Quantile order	Data set	Average income in a given year (in US dollars)				
		2006	2007	2008	2009	2010
0.25	Tax records	2 387	2 453	2 719	2 851	2 895
	HBS – Lower Silesia	4 447	4 552	5 257	5 566	5 736
	HBS – Poland	4 640	4 838	5 400	5 566	5 736
0.50	Tax records	3 953	4 122	4 576	4 777	4 890
	HBS – Lower Silesia	7 079	7 531	8 659	9 044	9 177
	HBS – Poland	7 349	7 761	8 641	8 870	9 177
0.75	Tax records	6 852	7 326	8 186	8 397	8 568
	HBS – Lower Silesia	10 750	11 243	12 961	13 914	13 799
	HBS – Poland	11 178	11 901	13 177	13 566	13 833
0.90	Tax records	11 565	12 297	13 547	13 900	14 153
	HBS – Lower Silesia	15 471	15 880	18 177	19 657	20 215
	HBS – Poland	16 342	17 480	19 041	19 480	20 205
0.99	Tax records	39 161	43 117	46 134	45 031	45 637
	HBS – Lower Silesia	33 964	31 445	37 963	40 594	38 125
	HBS – Poland	38 601	40 166	41 916	42 544	44 654
0.999	Tax records	169 863	206 057	204 199	176 385	175 709
	HBS – Lower Silesia	83 752	79 881	65 240	93 360	144 389
	HBS – Poland	92 488	107 328	101 697	98 058	102 556
Sample size		Number of observations in a given year				
	Tax records	1 772 975	1 807 136	1 833 325	1 824 117	1 823 372
	HBS – Lower Silesia	2789	2797	2824	2829	2831
	HBS – Poland	36 068	36 150	36 332	36 235	36 278

Source: own calculations.

The problem, inevitably related to tax data, is also their selective character. They do not include – despite the efforts of the administration, aiming to maximize the tax base – all sources of income. This means not only the underestimation of the actual income, but also the incompleteness of the records of taxpayers held by tax offices. Untaxed

sources of income, however, have relatively greater importance for the less wealthy taxpayers.

Additionally, in the case of taxes paid by the institution (for example a bank, for gains from investment), the amount of tax paid is not directly attributable to the taxpayer. In this way the income from this source, which could be important in the case of the most affluent, is not reflected in the tax returns.

The above list does not exhaust all the factors that negatively affect the completeness and reliability of the data in both data sets, but the discussed problems are the most important with regard to the group of the most affluent taxpayers and their households (a common feature of both HBS and tax records is that in both cases the actual values of income are understated). Because of the mentioned shortcomings, the very precise comparison of results for both data sets is not justified: it is much more important to assess whether the differences can be considered large or small, and whether the direction of discrepancies is stable over time.

4.2. Situation of the highest income groups

As already mentioned in Section 2, a set of tax data is not representative for the population of Poland. It covers nearly two million taxpayers filing each year a tax return in the tax offices in Lower Silesia. Therefore, besides the results for this data set, Table 1 presents the parameters calculated on the basis of data from HBS, calculated both for the whole Poland and for the province of Lower Silesia. The values of all quantiles in this table take account of inflation and are expressed in US dollars (the average annual exchange rate from 2006 was used).

The differences between Poland and the Lower Silesia province in the HBS data are not big below the quantile of order 0.90, but the values for the Lower Silesia are usually slightly lower. Above this quantile, the differences increase (except the quantile of order 0.999 in 2010, but it can be assumed that this is due to the lack of representativeness resulting from the too small sample size).

Much bigger differences exist between the tax data and the data from HBS. In Table 2 we present the ratios of quantiles calculated for both data sets. The values denote shares of tax record quantiles in HBS quantiles – the minimums and the maximums observed in the period 2006-2010.

Table 2

Relation between tax records quantiles and HBS quantiles for Lower Silesia

Quantile order		Share of quantile values for tax records in quantile values for HBS data (in %)					
		0.25	0.50	0.75	0.90	0.99	0.999
Share of tax record quantile in HBS quantile	Minimum 2006-2010	50	53	60	70	111	122
	Maximum 2006-2010	54	56	65	77	137	313

Source: own calculations.

The differences at the bottom of the distribution are related to, inter alia, the unit adopted in both data sets, which is person (taxpayer) in the case of tax data, and household for HBS. Due to the lack of information on the number of individuals achieving income subject to personal income tax in HBS data, the unification of units is not possible. However, lower values for the tax data in the lower part of the income distribution indicate the potential compatibility of both sets. Assuming the average number of income earners per household³ equals to 2, the share for the first and second quartile, amounting to approximately 50 percent, can be interpreted as the equal distribution of income within the household. The increase in the ratio for the higher quantiles (of order 0.75 and 0.9) suggests, assuming the same interpretation, the increase in asymmetry in the intra-household income distribution. It is reasonable that the probability of achieving a comparable income for all individuals earning income in the household decreases with household income.

In the highest areas of income distribution the relationship between the results for both data sets is, however, completely reversed. Incomes, according to the tax records, are on average more than twice higher than in HBS data (as already mentioned, data from the HBS for Lower Silesia in 2010 seem to be random, as in the case of Poland so big a change with respect to 2009 does not exist). The use of any multiplier that reflects the relationship household-individual person (the taxpayer) would even widen the observed differences between the tax data and the HBS data.⁴ Thus the only explanation for these differences is the very high underestimation of the income from the highest areas of income distribution in the HBS data.

³ In the analyzed period the average number of people aged over 18 years per household was about 2.3, but not every adult person derives income subject to personal income tax.

⁴ Even if this value is slightly below 1, which would mean that in the highest income households asymmetry in intra-household is typically very large (very high incomes are usually achieved by only one person in the household, irrespective of household size).

It is also worth noting that regardless of the economic slowdown that occurred in Poland after 2008, real incomes of people with lower incomes (up to the quantile of order 0.90) grew in real terms over the period under analysis. The financial crisis has affected the situation of the richest, whose incomes have fallen in real terms between 2007 and 2010.

However, quantile measures presented in Table 1 only indicate income thresholds and neglect inequality within each group. Therefore, the results presented in Table 3 supplement the earlier observations – averages, standard deviations (figures in US dollars, adjusted for inflation) and coefficients of variation which facilitate direct comparisons.

Table 3

Averages, standard deviations and coefficients of variation for incomes in the upper part of the income distribution for tax records and HBS data

Quantile	Measure	Data set	Year				
			2006	2007	2008	2009	2010
0.90–0.99	Average	Tax records	17 935	19 296	20 969	21 224	21 689
		HBS – Lower Silesia	20 025	20 701	23 223	25 318	25 401
	Standard dev.	Tax records	6 307	6 994	7 397	7 103	7 251
		HBS – Lower Silesia	4 074	3 835	4 356	5 007	4 355
	Coefficient of variation	Tax records	35%	36%	35%	33%	33%
		HBS – Lower Silesia	20%	19%	19%	20%	17%
0.99–0.999	Average	Tax records	67 110	77 305	80 176	73 878	74 232
		HBS – Lower Silesia	45 688	43 960	48 667	54 422	50 845
	Standard dev.	Tax records	29 542	36 266	35 690	29 571	29 425
		HBS – Lower Silesia	11 616	12 232	7 793	13 852	19 376
	Coefficient of variation	Tax records	44%	47%	45%	40%	40%
		HBS – Lower Silesia	25%	28%	16%	25%	38%
0.999–1.00	Average	Tax records	412 421	524 407	514 425	386 392	396 745
		HBS – Lower Silesia	143 478	115 797	81 079	120 601	247 377
	Standard dev.	Tax records	797 573*	771 934*	814 744*	557 035*	593 261*
		HBS – Lower Silesia	57 188*	45 280*	20 299*	6 543*	79 588*
	Coefficient of variation	Tax records	193%	147%	158%	144%	150%
		HBS – Lower Silesia	40%	39%	25%	5%	32%

* Results not reliable due to the sample size ($n = 2$)

Source: own calculations.

Taking into account all the observations, the average incomes in the highest group are on average almost four times higher for tax records than for the HBS data (this ratio varies between 1.60 in 2010 and 6.34 in 2008). The differences between the data sets have also increased for the two other groups. For all the analysed groups and periods, tax records are also characterized by higher values of standard deviation and coefficient of variation. This means that the survey data only partially reflect the variance among the highest incomes, and the variance in this area of income distribution largely affects the measures of inequality and polarization estimated for the entire distribution of income.

Changes in the relative situation of different groups, being a consequence of the observed changes in wealth, are presented in Figure 1. The presented values of the *RIC* indexes denote the change (in percentage points) in the share of average income for the group of people with incomes lower than a given quantile in the average income of a group with higher incomes. In each case the reference point was the preceding year. The results for the quantiles of order 0.999-1.00 for the HBS data have not been presented due to the sample size.

As one can see from the presented graphs, the differences between both data sets do not include the absolute value of the income of wealthy individuals. While the tax data represent an improvement in the relative situation of the more affluent with respect to the poorer in 2007 (negative values indicating a reduction of income share of the poor in relation to the income of more wealthy individuals) and a decline (or no change) in subsequent years, whereas the HBS data indicate the relative improvement in the situation of wealthy households in 2009 and a deterioration in the remaining periods. Due to the fact that the financial crisis mostly affected the income of wealthy individuals, the relative improvement in the situation of the poor in the early years of the crisis seems to be a more likely scenario.

The analyses presented so far were aimed at describing the income situation (relative and absolute) of particular groups, but from the standpoint of society as a whole, particularly important is the share of income derived by individuals (households) in the group in the total income of the population. Analyses of this type, often presented in the literature, are designed to answer the question of the concentration of income. The data presented in Table 4 show the share of income of the group in the aggregate income of the entire population.

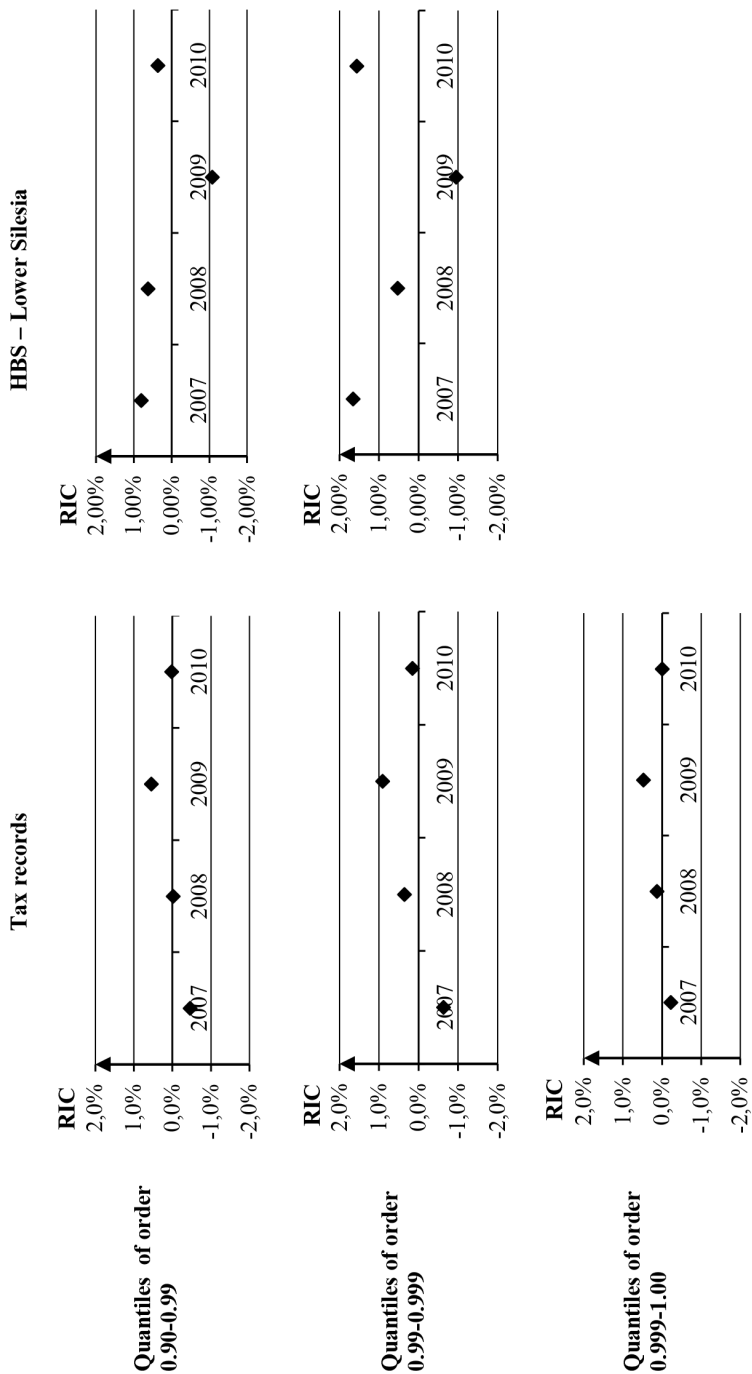


Fig. 1. Relative income changes in the upper part of the income distribution for tax records and HBS data
Source: own calculations.

Table 4

Income shares in the upper part of the income distribution for tax records and HBS data

Quantile range	Data set	Sum of incomes in a quantile group, given as a share of the sum of all incomes in respective years (in %)				
		2006	2007	2008	2009	2010
0.90–0.99	Tax records	25	25	25	26	26
	HBS – Lower Silesia	21	21	21	21	21
0.99–0.999	Tax records	9	10	10	9	9
	HBS – Lower Silesia	5	4	4	5	4
0.999–1.00	Tax records	6	8	7	5	5
	HBS – Lower Silesia	1*	1*	1*	1*	2*

* Results not reliable due to the sample size ($n = 2$)

Source: own calculations.

Particularly large gaps between the analysed data sets can be seen in the case of the highest income taxpayers whose share in total income of the population according to the HBS is underestimated. In the case of a group with incomes between the quantile of order 0.99 and quantile of order 0.999, the share is two times higher for tax data. Such an underestimation of income in the highest income groups denotes the misstatement of the actual ability of these groups to accumulate income and to influence economic growth.

The understatement of income in groups of individuals (households) with the highest income is the result of several factors, characterized in general at the beginning of this section. One of the possible reasons is the lack of a ‘capital gains’ category in HBS, while the importance of this source of income increases with income. An analysis of the percentage of income derived from particular sources is presented in Table 5.

As follows from the presented data, the share of income from capital increases with the increase in total income. Yet the amount of capital gains – not included in the HBS data – does not fully explain the discrepancy between the values observed for the highest income group in both analysed data sets. It is also worth noting that the importance of capital gains was greatly reduced during the financial crisis – the share of capital income decreased approximately three times between 2006 and 2009. This is because of the greater risk aversion and the transfer of assets to more secure investments, and due to the losses caused by decline in the stock market.

Table 5

Sources of income in the upper part of the income distribution for tax records and HBS data

Data set	Quantile range	Source of income	Sum of incomes from a given source in a quantile range, given as a share of the sum of all incomes in this quantile range in respective years (in %)				
			2006	2007	2008	2009	2010
Tax records	0.90-0.99	Own business	28.1	27.6	27.4	27.9	27.9
		Hired work, social security	70.3	70.8	71.9	71.6	70.9
		Capital	1.6	1.6	0.7	0.5	1.2
	0.99-0.999	Own business	64.9	67.9	70.2	66.2	64.3
		Hired work, social security	31.9	28.4	28.6	32.7	33.9
		Capital	3.2	3.7	1.2	1.1	1.9
	0.999-1.00	Own business	75.1	78.3	82.9	86.3	84.1
		Hired work, social security	6.9	4.3	4.5	6.8	7.9
		Capital	17.9	17.4	12.6	6.9	7.9
HBS – Lower Silesia	0.90-0.99	Own business	23.3	24.0	20.6	21.1	21.2
		Hired work, social security	76.7	76.0	79.4	78.9	78.8
	0.99-0.999	Own business	38.5	39.9	36.7	34.8	30.5
		Hired work, social security	61.5	60.1	63.3	65.2	69.5
	0.999-1.00	Own business	70.5 *	75.4 *	51.6 *	77.0 *	25.0 *
		Hired work, social security	29.5 *	24.6 *	48.4 *	23.0 *	75.0 *

* Results not reliable due to the sample size ($n = 2$)

Source: own calculations.

In addition to capital income, with the increase in total income, also income share from self-employment is growing. In the case of the highest income group this is the main source of income. The contribution of income from this source to the total income, however, is systematically lower in the case of the HBS data. To some extent this may be due to reporting some part of income from labour or social security benefits using form PIT36 – in the case of revenue from several sources. However, the almost two times lower share of income from self-employment resulting from the HBS data for the group between quantiles of order 0.99 and 0.999, suggests serious underestimation.

A very important element in assessing the situation in the upper part of the income distribution is the stability of the composition of each group. Estimates of mobility of the high income earners, defined as the probability of leaving the group in a given period of time, are presented in Table 6.

Table 6
Mobility of the top income earners for tax records and HBS data

Data set	Group definition – quantile range	Share of taxpayers leaving given group in respective periods (in %)				
		2006-2007	2007-2008	2008-2009	2009-2010	2006-2010
Tax records	0.90-1.00	22	22	22	20	39
	0.99-1.00	31	31	33	30	53
	0.999-1.00	42	42	44	42	63
HBS – Lower Silesia	0.90-1.00	35	36	37	42	-
	0.99-1.00	56	55	48	58	-

Source: own calculations.

Tax data are gathered in panel form and allow analysis of the taxpayers' situation throughout the period 2006-2010. The panels in the HBS data are only two-year long and cover about half of the sample. Hence this gives the opportunity to analyse the situation in the next year, but does not allow estimating the probability of leaving the group for the entire period.

The figures presented in Table 6 increase with income, but are relatively stable over time (for annual periods) especially for the tax data. Income mobility for both data sets differs significantly as much higher values were obtained with the HBS data. Taking into account the differences in research units – taxpayers for tax records and households, in which usually more than one person gains income, for HBS – greater stability should be observed for the HBS data, even taking into account the already mentioned impact of intra-household income distribution. This may suggest that among households with a high income which are to be surveyed in the second year, non-response rates are lower for households which have experienced a decline in their income (and left the upper group), so the payment associated with participation in the study may be relatively more important for them. For the tax data, some impact on estimates of mobility can come from the difference between the size of the entire analysed population of taxpayers (approximately 2.3 million) and the number of taxpayers who file a tax return each year (about 1.8 million). This difference is to some extent due to the migration between provinces. If a high-income taxpayer migrates from the area of Lower Silesia region because of the worsening of his/her economic situation, the case would not be included in the analysis, yet such situations are not frequent.

5. DISCUSSION

The assessment of the situation of wealthy individuals (households) with respect to the absolute changes in the income situation is similar for both data sets. In the analysed period, real income grew in subsequent years for all groups below the quantile of order 0.99. Differences between data sets appeared only in the case of the highest percentile for which the real value of income declined in different periods in both data sets.

Different results were obtained in the assessment of the relative situation of wealthy people. Analysis of the relative income change – the relationship between income of wealthy and non-wealthy individuals – made for the tax data, indicated the improvement in the relative situation of wealthy individuals only in 2007 (compared to 2006). A similar analysis carried out for the HBS data showed the improvement of the situation of wealthy individuals in 2009 (compared to 2008). This difference is so big that the actual situation in these two periods was very different because of the financial crisis.

The most important differences between the data sets concerned not the direction of change, but the estimates of the absolute level of wealth. Given the differences in the definition of units for which data were collected in both sets, it can be assumed that the estimates are similar for less wealthy people (households). In the case of the highest incomes – above the quantile of order 0.99 – the differences are big and they cannot be justified other than by systematic errors in the HBS data set. Big differences were also observed regarding the estimates of mobility. The results for the tax data show an approximately two times lower probability of leaving the group of wealthy individuals than analogous estimates for the HBS data.

These findings lead to asking the question about the reliability of the results obtained on the basis of data from both sources. Given the disadvantages of both data sets characterized briefly in Section 4.1, the results indicate the higher reliability of tax data in the assessment of the situation of top-income earners. Systematic errors associated with the collection of the data on income, in the case of both sets indicate the underestimation of the real values of income. Thus if the estimates of averages and quantiles calculated for the tax data are higher despite the smaller (or at least no bigger) unit (a household may not consist of less than one person), the HBS data concerning the upper tail of the income distribution is certainly not reliable. Differences in the values of the parameters are very large for the group of people with the highest incomes,

so using data from the HBS to describe the situation of the richest is burdened with a very high risk to draw completely wrong conclusions, both in terms of assessing the level of wealth of this group as well as changes in this level.

It is worth noting that despite the large discrepancy between the results obtained in analysing results for both data sets, estimates of various parameters are relatively stable over time⁵ for each data set separately. This means that the underestimation of the highest income has a systematic character and is not just a consequence of random errors in a particular edition of the survey.

CONCLUSIONS

Studies of affluence are limited by the data availability, even more than studies on inequality and other aspects of income distribution. In some situations the relevant data are not attainable at all, but sometimes similar data can be derived from several independent sources collected within the unrelated databases. In the latter case, assessment of the situation requires a decision as to which set should be considered more reliable. Indications in the process of making such a decision may be, in addition to the results of basic analyses (based on the compatibility of the results of observations with theoretical considerations), some additional information regarding the structure of the data, the methods of data collection, etc.

In the case of an analysis of the situation of wealthy people, the tax data (although burdened with certain defects) are considered more reliable than survey data (see e.g. Moore et al., 2000). This assessment was essentially confirmed in this analysis. The basic conclusion that flows from the analysis is the limited usefulness of the data from household budget surveys (usually applied to the analysis of income distribution in Poland) to assess the situation of the top-income earners. At the same time it can be expected that for the part of the population with lower incomes, survey data covering a much wider range of income sources will allow a more reliable assessment of the situation.

Summing up, it should be emphasized that both data sets covered a relatively short period of time (only five years). Although the overall relationship between them appears to be stable over time, the period is far too short to formulate completely unambiguous conclusions.

⁵ For the full sample (covering households from all provinces in Poland) in cases of the HBS data set.

REFERENCES

- Atkinson, A. B., Brandolini, A., *Promise and Pitfalls in the Use of 'Secondary' Data-Sets: Income Inequality in OECD Countries as a Case Study*, "Journal of Economic Literature", 39(3), pp. 771-799, 2001.
- Atkinson, A. B., Piketty, T., Saez, E., *Top Incomes in the Long Run of History*, "Journal of Economic Literature", 49(1), pp. 3-71, 2011.
- Brzeziński, M., *Income Affluence in Poland*, "Social Indicators Research", 99, pp. 285-299, 2010.
- Brzeziński, M., *Wealth Distributions and Power Laws: Evidence from 'Rich Lists'*, "Physica A", 406, pp. 155-162, 2014.
- Bukowski, P., Novokmet, F., *Inequality in Poland: Estimating the Whole Distribution by G-percentile, 1983-2015*, WID.world Working Paper Series No 2017/21, 2017a.
- Bukowski, P., Novokmet, F., *Top Incomes during Wars, Communism and Capitalism: Poland 1892-2015*, International Inequalities Institute Working Paper 17, London School of Economics and Political Science, London 2017b.
- Burkhauser, R. V., Shuaizhang, F., Jenkins, P. S., Larrimore, J., *Recent Trends in Top Income Shares in the USA: Reconciling Estimates from March CPS and IRS Tax Return Data*, Center for Economic Studies Working Paper, CES 09-26, 2009.
- Central Statistical Office, *Incomes and Living Conditions of the Population in Poland* (report from the EU-SILC survey of 2006), Warsaw 2008.
- Central Statistical Office, *Europejskie badanie dochodów i warunków życia (EU-SILC) w 2010 r.* Notatka informacyjna [*European Study of Income and Living Conditions (EU-SILC) in 2010*. Information note], Warsaw 2011a.
- Central Statistical Office, *Metodologia badania budżetów gospodarstw domowych [Methodology of Household Budget Survey]*, Warsaw 2011b.
- Central Statistical Office, *Household Budget Survey in 2016*, Warsaw 2017.
- Esteban, J., Ray, D., *Linking Conflict to Inequality and Polarization*, "American Economic Review", 101, pp. 1345-74, 2011.
- Ferrer-i-Carbonell, A., *Income and well-being: an empirical analysis of the comparison income effect*, "Journal of Public Economics", 89, pp. 997-1019, 2005.
- Human Development Reports, United Nations Development Programme, New York 1990-2011.
- Kordos, J., Lednicki, B., Żyra, M., *The household sample surveys in Poland*, "Statistics in Transition", 5, pp. 555-589, 2002.
- Kośny, M., Yalonetzky, G., *Relative income change and pro-poor growth*, "Economia Politica. Journal of Analytical and Institutional Economics", 32, pp. 311-327, 2015.
- Leigh, A., *Top incomes* [in:] Salverda, W., Nolan, B., Smeeding, T. (eds.), *The Oxford Handbook of Economic Inequality*. Oxford University Press, Oxford 2009.
- Moore, J. C., Stinson, L. L., Welniak, J. E., *Income Measurement Error in Surveys: A Review*, "Journal of Official Statistics", 16, pp. 331-361, 2000.
- Ministry of Finance, *Information Concerning Personal Income Tax in 2010*, Warsaw 2011.

- National Bank of Poland, *Zasobność gospodarstw domowych w Polsce. Raport z badania pilotażowego 2014 r.* [The households' wealth in Poland. Report from the 2014 pilot study], edited by: Bańbuła, P., Żółkiewski, Z. Warsaw 2015.
- National Bank of Poland, *Zasobność gospodarstw domowych w Polsce. Raport z badania 2016 r.* [The households' wealth in Poland. Report from the 2016 study], edited by: Bańbuła, P., Żółkiewski, Z. Warsaw 2017.
- Osberg, L., *Measuring Economic Security in Insecure Times: New Perspectives, New Events, and the Index of Economic Well-Being*, Centre for the Study of Living Standards Research Report 2009-12, 2009.
- Piketty, T., *Capital in the Twenty-First Century*. Harvard University Press, 2014.
- Piketty, T., Saez, E., *Income Inequality in the United States, 1913-1998*, "Quarterly Journal of Economics", 118(1), pp. 1-39, 2003.
- Saez, E., Veall, R. M., *The Evolution of High Incomes in Northern America: Lessons from Canadian Evidence*, "The American Economic Review", 95(3), pp. 831-849, 2005.
- Sierminska, E., Brandolini, A., Smeeding, M. T., *The Luxembourg Wealth Study – A cross-country comparable database for household wealth research*, "Journal of Economic Inequality", 4(3), pp. 375-383, 2006.
- Wheary, J., Shapiro, M. T., Draut, T., *By a Thread: The New Experience of America's Middle Class*, Demos: A Network for Ideas & Action, 2007.

Received: March 2016, revised: June 2018

Acknowledgment: *This paper is part of the research programme The importance of the shape of the income distribution to the economic growth in Poland (No. 3921/B/H03/2011/40) of the National Science Centre of Poland whose financial support is gratefully acknowledged. I would also like to thank the Directorate of the Tax Chamber in Wrocław for providing tax data for research purposes, and because of their openness and kindness, conducting this research was possible.*