The impact of different data sources on the level and structure of income inequality

Ayala, Luis; Pérez, Ana; Prieto-Alaiz, Mercedes

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Ayala, L., Pérez, A., & Prieto-Alaiz, M. (2022). The impact of different data sources on the level and structure of income inequality. *SERIEs: Journal of the Spanish Economic Association*, *13*(3), 583-611. https://doi.org/10.1007/s13209-021-00258-0

SERIEs (2022) 13:583–611
https://doi.org/10.1007/s13209-021-00258-0

ORIGINAL ARTICLE

The impact of different data sources on the level and structure of income inequality

Luis Ayala[1] · **Ana Pérez**[2] · **Mercedes Prieto-Alaiz**[2]

Received: 12 April 2021 / Accepted: 12 November 2021 / Published online: 30 December 2021

Abstract
This paper aims to analyze the effect on measured inequality and its structure of using administrative data instead of survey data. Different analyses are carried out based on the Spanish Survey on Income and Living Conditions (ECV) that continued to ask households for their income despite assigning their income data as provided by the Tax Agency and the Social Security Administration. Our main finding is that the largest discrepancies between administrative and survey data are in the tails of the distribution. In addition to that, there are clear differences in the level and structure of inequality across data sources. These differences matter, and our results should be a wake-up call to interpret the results based on only one source of income data with caution.

Keywords Inequality · Administrative data · Measurement error · Dependences · Copula

JEL Classification D31 · D63

✉ Luis Ayala
layala@cee.uned.es

Ana Pérez
perezesp@eaee.uva.es

Mercedes Prieto-Alaiz
prietoal@eco.uva.es

[1] Facultad de Derecho, UNED: Universidad Nacional de Educacion a Distancia, C/Obispo Trejo, 2, 28040 Madrid, Spain

[2] Facultad de Ciencias Económicas y Empresariales and IMUVA, Universidad de Valladolid, Avenida Valle Esgueva 6, 47011 Valladolid, Spain

1 Introduction

Studies of the inequality in the distribution of income have gained considerable momentum during the last decade. The increase in inequality in many OECD countries has generated a growing interest both in identifying the possible causes of the increase in income differences among households and in determining the optimal policy design to reduce them. The development of both of these streams of research has led to fundamental contributions, some of them crossing the frontiers of economic analysis and moving to the forefront of social debate (Piketty 2013; Atkinson 2015).

A key reason for this renewed growth in inequality studies has been the increasing availability of datasets that cover extensive periods of time. In many OECD countries, household surveys carried out with homogeneous methodologies over long and continuous periods allow for the reconstruction of changes in the distribution of income over the very long term. These surveys are not free from major problems, which can impose important limits on accurately diagnosing the trends and determinants of inequality as well as correctly determining the design of policies based on them.

As an alternative, administrative records have been increasingly used to analyze changes in the income distribution. A great advantage of these data is their availability over very long periods of time and their better coverage of higher incomes. However, they are also affected by certain limitations, such as those introduced by tax avoidance and evasion, income shifting, theoretical problems when forming households from tax units and, especially, limited coverage of households with incomes below the income tax threshold. Additionally, tax records include only taxable sources of income and do not account for informal sources that may be captured by surveys (Meyer and Mittag 2021). In the same vein, many studies have long recognized that the richest individuals are less likely to participate in surveys. However, some comparisons of survey data with tax records suggest that the two sources only start to diverge appreciably at the very top (Yonzan et al. 2020; Ravallion 2021). In addition, the hypothetical advantage that tax data best capture higher incomes is limited by the fact that several fiscal manipulation strategies are sensitive to changes in marginal tax rates and income reporting rules (Slemrod 1995; Burkhauser et al. 2012; Auten and Splinter 2019; Guyton et al. 2021).

The problems faced by surveys in terms of properly collecting benefit data have also increased the use of administrative records. Some survey respondents may underreport their benefit receipts due to simple forgetfulness, misplacement in time or misclassification, or conscious suppression (Lynn et al. 2012). Because some programs are sharply underreported in survey data, the distributional and poverty-reducing effects of transfer programs may not be accurate. Nevertheless, some recent works show that administrative data—usually considered the "gold standard" for this type of variables—can have missing values, be incorrectly entered, or be outdated (Courtemanche et al. 2019).

In practice, different choices in terms of data source—survey or administrative—may produce different inequality results (Dahl et al. 2011; Burkhauser et al. 2012; Carr and Wiemers 2018). In general, most studies that achieve a higher representativeness of incomes corresponding to the highest percentiles obtain higher inequality indicators with administrative data (Burkhauser et al. 2018; Higgins et al. 2018). However, the incorporation of records on benefits with administrative information that is more

representative of their incidence on low incomes has the opposite effect (Meyer et al. 2015; Meyer and Mittag 2021). The best procedure would be combining the accuracy of administrative information with the rich demographic details and population representativeness of surveys (Oberski et al. 2017; Jenkins and Rios-Avila 2021). Some authors have attempted to implement this strategy by harmonizing the definitions of their variables to improve the representativeness of higher incomes (see, for instance, Burkhauser et al. 2016; Higgins et al. 2018; Lynn et al. 2012; Meyer and Wu 2018).

Starting in 2008, the Spanish Survey on Income and Living Conditions (ECV)—the main dataset for measuring inequality—provides both survey and administrative data on households' income. In this paper, we exploit this information to learn about the accuracy of survey data and to see the implications for measurement of inequality. It is possible to evaluate the impact of the change from one data source to another. How does this change affect the measurement of inequality? What sources of income are modified in their contribution to total inequality? How do the dependences between income sources change? The Spanish ECV data provide an excellent opportunity to answer these questions by allowing a comparison between two different data sources for the same individuals.

This paper addresses this goal by carrying out different types of analysis to identify the change in each income source and its impact on inequality under the new criterion. We pay special attention to possible effects on the structure of inequality and the dependences between sources of income—an issue scarcely studied so far in the literature. Thus, the paper contributes to the literature on the quality of income data by determining whether administrative data yield lower levels of inequality, and, most innovatively, a different structure of inequality across data sources.[1]

Among other results, we find a significant increase in the measured disposable income of households when using administrative data as opposed to survey data, especially in terms of capital income. Moreover, when using administrative data, the measured incomes of both tails of the income distribution increase considerably more than those of the middle strata. An important finding is that the largest discrepancies between administrative and survey data are in the tails of the distribution. We also find lower levels of inequality, in general, when using administrative data, although the magnitude of the change depends on the index considered. Taking advantage of the available information from different years, we find a lower growth of inequality with administrative records than with interview data. Finally, both methods of data collection result in significant differences in the structure of inequality.

The observed reduction in inequality with administrative data, although not in all indices, differentiates the Spanish case from some of the aforementioned studies. Our results are mainly explained by both the higher levels of labor income and cash benefits with administrative data and the remarkable increase in taxes, with a higher growth of the latter in the richest percentiles. These changes would be offsetting the effect of the general increase in capital income. We also show that the reduction in inequality is mainly due to the narrowing of income differences in the lower part of the income

[1] We focus here on accuracy as one of the components of data quality. There are also other components that could be analyzed such as timeliness, relevance, and accessibility (European Statistical System 2019).

distribution. All these findings call into question the conclusions anticipated by the institution producing the data (INE), which predicted only minor changes.

Another contribution of the paper is that it is informative about measurement error and its consequences. Since the pioneering work of Bollinger (1998), many validation studies have confirmed the presence of measurement error in survey data. These studies require some measure of truth or an objective standard by which the accuracy of the data can be judged. In terms of household disposable income, we might think that administrative data, although also affected by errors, have important advantages over interview data due to their larger sample size, high response rates and lower recall bias (Larrimore et al. 2021). Although a priori we cannot unequivocally say that the administrative data in the ECV are optimal, we show how the shift to administrative data better captures some of the incomes usually underreported in household surveys, such as capital income or taxes paid.

The remainder of the paper is organized as follows. Section 2 describes the change in the income collection method. The impact of this change on the income distribution is analyzed in Sect. 3. In Sect. 4, we estimate the effects of the use of administrative data on the measurement of inequality, in terms of both disposable income and each income source. In Sect. 5, we analyze the effect of changing from the use of survey data to that of administrative data on the structure of inequality and dependences between income sources. Section 6 concludes.

2 The income collection method

Since 2004, the European Union countries have utilized the same dataset to collect information on living conditions and household income (EU-SILC; in Spain, the ECV). The ECV provides information on individual and household income, the material and demographic characteristics of households, and a broad set of sociodemographic information. It also provides very detailed information on the material well-being necessary to estimate the incidence of multidimensional deprivation.

The Spanish sample size of the EU-SILC is approximately 16,000 households, divided into 2,000 census sections. Beginning in 2012, the INE made some corrections to this dataset to avoid a lack of representativeness among certain population groups, such as immigrants, and to guarantee the representativeness of the sample. In terms of the measurement of inequality, the main methodological change to this dataset occurred in 2013. Until that year, the income data collected by the survey were declared by households when they were interviewed. Beginning in 2013, the information provided by the Tax Agency and the Social Security Administration was included as the income of households and individuals. Using this new way to collect income data, the INE recalculated the data of the previous waves, which brought the new series back to 2008. Moreover, the INE continued to collect income data through the interview method until 2014. The preliminary analyses carried out by the INE showed that this transition from one method to another did not seem to have an impact on the magnitude of the inequality measures, although its effect on the average levels of different income sources was very significant (Méndez and Vega 2011).

This change in methodology was motivated fundamentally by a desire to improve the quality of the information contained in the database and to gain additional knowledge regarding households' income sources. As mentioned above, one of the traditional problems faced by household income surveys is non-response to certain components of income data. Another limitation was the difficulty of determining, through interview data, the gross income of each member of a household, their social contributions and their taxes paid; often, these factors must be simulated. In fact, some EU countries had already begun using administrative information for the collection of income data for the EU-SILC; the Nordic countries, the Netherlands, France, Austria and Slovenia, had begun to implement this method in the case of some income components.

The procedure used to collect individuals' income data from the administrative files consists of using the national identity number of the individuals included in the sample to collect the corresponding income data as recorded in the tax sources and Social Security files.[2] In the case of social benefits, the INE uses the Registry of Public Social Benefits. In the preliminary studies, some divergences were found between the type of benefits declared by some households and the benefits that these households actually received (INE 2010). This phenomenon occurs, for instance, in the case of some old-age pensions classified as non-contributory by the interviewees themselves but that appear in the Social Security files as contributory. Another difficulty is that some benefits that appear in the Registry, such as retirement due to disability, are not classified as such in the survey. The previous evaluations carried out by the INE (2010) revealed that the average level of income as shown by the administrative data was somewhat higher (4.6%) than that shown by the survey data.

Data from the Tax Agency are used for the other income sources, and these data are extracted mainly from the Personal Income Tax (IRPF) files. One typical problem faced by tax records is the large number of people who are not obligated to declare income. To solve this problem, the INE uses tax withholding files, which include income earners without the obligation to declare. One of the main advantages of tax data is that there are many households that do not declare capital income when they are interviewed but do show this income in their tax data. According to the estimates made by the INE (2010), before the adoption of the new procedure, the average capital income according to the tax data was twice as much as it was with the traditional criterion. The opposite phenomenon occurred in the case of self-employment income, although the difference was not as large (6%).

An important issue to understand the type of information taken from the Tax Agency is that INE collects the information on the different incomes subject to the personal income tax before it has been paid. The values of the variables, therefore, are before the application of reductions or exemptions, and there are no differences between the concepts in interview data and administrative records. In other words, the different incomes that appear in the ECV with data from the Tax Agency are the data for each source of income collected by this institution before taxes are paid (e.g., the data that appear on dividends are those collected by the Tax Agency before any reduction is applied). Nevertheless, the income tax rules that determine how income is reported on

[2] The uniqueness of the regional financing system in Spain prevents the same procedure from being used in some other regions, such as the Basque Country and Navarre, where income information is still collected using the interview method.

tax returns may result in some differences between interview and administrative data. Tax filers may have financial incentives to report their income in ways that limit their tax liabilities (Burkhauser et al. 2012).

In this paper, we focus on the survey conducted in 2014 since this was the last year during which data were collected via both methods, namely survey and administrative, and because, as we shall see below, the effects of the change to administrative data are similar in all the years for which both types of information are available. For our analysis to be carried out, a necessary first step is the identification of the main income sources in each survey. We have grouped these income sources into five major categories: labor income, self-employed income, capital income, benefits and taxes.[3] These types of income are recorded in different files. Additionally, the aggregation of the income sources of each member of a household is necessary. While most social benefits are included in an individual's file, some are included in his or her household file—namely family benefits, social exclusion benefits, housing benefits and taxes. Moreover, the INE collects survey and administrative data in different files, and households' identification numbers are not identical in each file. To accomplish the goals of this paper, it was necessary to merge the survey and administrative files. To do this, we used matching methods with the common variables in both administrative and survey data for the same households.[4]

The definitions of the main sources of income are shown in "Appendix A." We take as reference household disposable income. We define it as total gross household income minus tax on income and social insurance contributions, regular taxes on wealth and regular inter-household cash transfer paid. Labor income is total remuneration, in cash or in kind, payable by an employer to an employee in return for work. Self-employment income is income received by individuals as a result of their current or former involvement in self-employed work. Capital income includes interest, dividends, profits from capital investment in an unincorporated business; income from rental of a property or land; and pensions received from individual private plans. The ECV only collects part of realized capital gains—interest generated by assets—and does not include unrealized capital gains. As stressed by some authors, excluding all capital gains may underestimate the flow of resources to tax units near the top of the distribution whose income is heavily dependent on these gains (Larrimore et al. 2021). Cash benefits comprise family-/children-related allowance, housing allowances, social exclusion, unemployment benefits, old-age benefits, survivor' benefits, sickness benefits, disability benefits and education-related allowances. We also include here regular inter-household cash transfers. Taxes include taxes on income and social insurance contributions, regular taxes on wealth and regular inter-household cash transfer paid. Employers' social insurance contributions are part of the employee's gross income but not of the household's disposable income.

[3] For the sake of simplicity, we have grouped taxes and social contributions into a single category. Social insurance contributions refer here to contributions by employees', the self-employed and if applicable, the unemployed.

[4] Less than 5% of the observations remained unmatched and the randomization tests did not detect any patterns in those left out.

3 Effects on the income distribution

A preliminary approach to identifying the possible effects that the change in the income collection method can have on the measurement of inequality is the comparison of the distributions of disposable income provided by the two types of data. That change could also affect the measured level of inequality through the different impact that it has on the distribution of each income source. To analyze these issues, we first compare the densities of disposable income and income sources provided by both the administrative and survey data. Then, we estimate the effect of the data collection change on the mean and the percentiles of disposable income and those of its different income sources.[5]

3.1 Effects on the distribution of disposable income and the distributions by income source

Figure 1 displays the kernel density estimates of disposable income (Panel a) and those of the income sources (Panels b–f) resulting from both the survey and the administrative data.[6] Figure 1a reveals that the change to administrative data has several effects on the income distribution, namely a shift to the right of the distribution and a reduction in the number of households with incomes close to the modal value.

Figure 1b shows the density functions of labor income for each household based on the two income collection methods. Some changes are observed in the two resulting distributions. First, unlike the distribution of disposable income in Fig. 1a, the labor income distribution is shifted to the left in the case of the administrative data, indicating that there are more low-wage earners according to this data source. Second, the bimodal profile of the distribution corresponding to the administrative data stands out, with the first mode probably reflecting earnings received from part-time employment.

The profiles of the two distributions of self-employment income (Fig. 1c) are very similar, except in terms of the decreasing section from the modal value. Regarding capital income, Fig. 1d shows that there is a higher proportion of households with a very low capital income in the distribution corresponding to the survey data than there is in the distribution corresponding to the administrative data. Moreover, the maximum values of this income source are very different in each case, with the maximum value of the administrative data almost doubling that of the data obtained through interviews.

The distributions of taxes do not differ substantially between the two methodologies (Fig. 1e). The tax distribution corresponding to administrative data shifts to the right, and there is also a greater concentration of taxes paid close the modal value of survey data. Finally, the two distributions of cash benefits (Fig. 1f) present somewhat different profiles from the previous ones, which displayed a greater concentration around low values. In the case of cash benefits, the modal values are similar between the distributions, but the concentration around the modal value in the administrative data

[5] In the following analyses, we use non-negative incomes. This may mean that in some of the comparisons the number of individuals is different in each distribution.

[6] To better appreciate the differences between the two data sources, Figs. 1 and 2 use a shorter range in the horizontal axis from 0 to 80,000 euros.

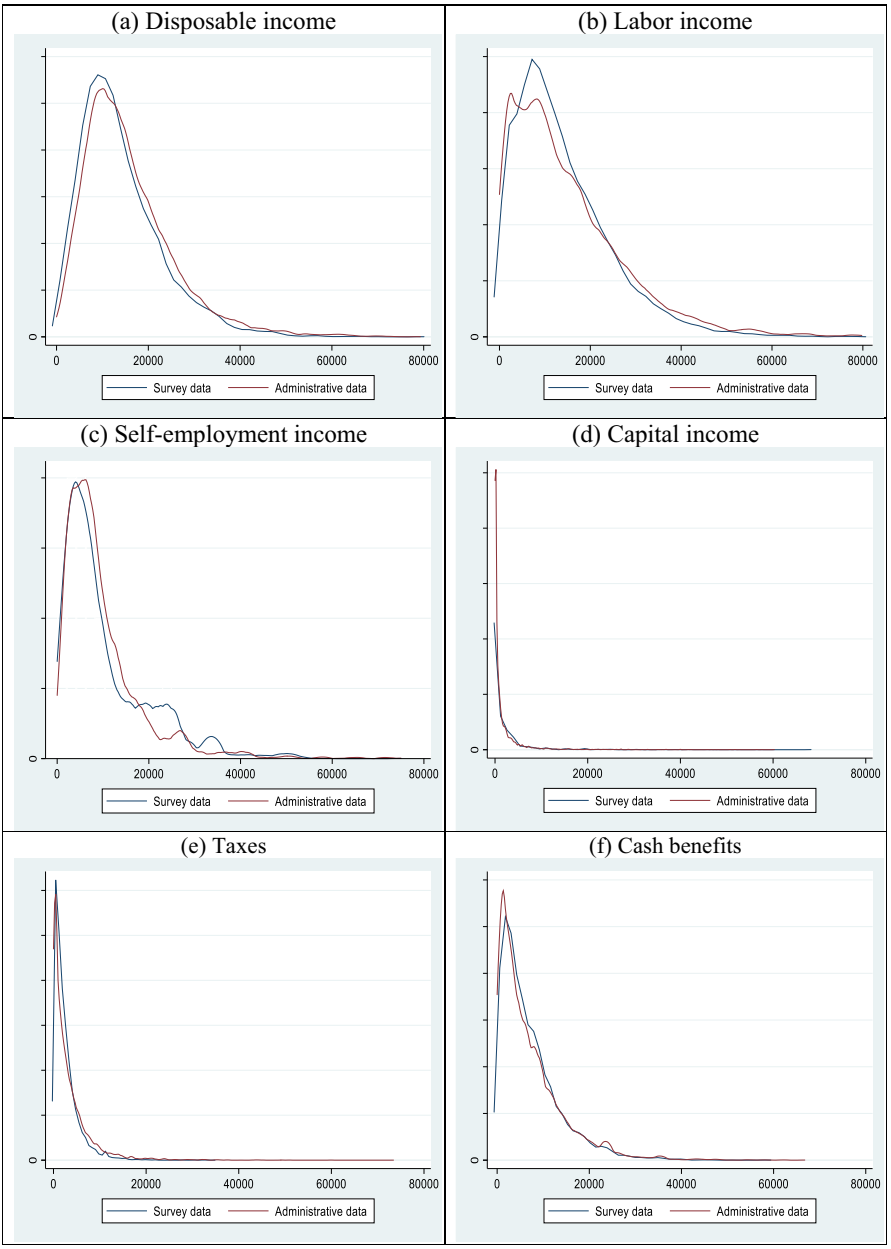


Fig. 1 Distributions of disposable income and income sources

is clearly sparser. Additionally, from this value onward, there are a greater number of households in the distribution with administrative data; however, in the far right of the distribution, both distributions become similar.

In order to get further insights into the differences between the two data sources, Fig. 2 displays the scatter plots of survey against administrative data for disposable income and for each source of income. The red line represents the 45° line. If there were no discrepancies between administrative and survey data, all the points would be aligned along the 45° line. The results in Fig. 2 suggest that this is not the case for the disposable income and for each income source, placing more points above the 45° line. So, in general, the income reported with survey data is lower than with administrative data. Furthermore, the relationship between administrative data and survey data does not seem to be linear, which leads to conjecture that the differences between both types of data do not keep constant throughout the distribution of income. Moreover, this figure confirms some of the features displayed in Fig. 1. For instance, both the distribution of disposable income and taxes shift to the right with administrative data, and the latter displays larger maximum values than with survey data. Also, in the capital income distribution with survey data there is a higher proportion of households with low values, whereas the maximum values with administrative data are larger.

3.2 Effects on the summary measures of the income distributions

The differences in the distributions of income mentioned above will presumably lead to differences in their main summary measures. In this section, we focus on these differences in terms of mean and percentiles, whereas the differences in inequality measures will be discussed in Sect. 4. Figure 3 shows the differences in the average level of disposable income and its various components.[7] Additionally, hypothesis testing for comparing means is done by considering the Wilcoxon rank test.[8]

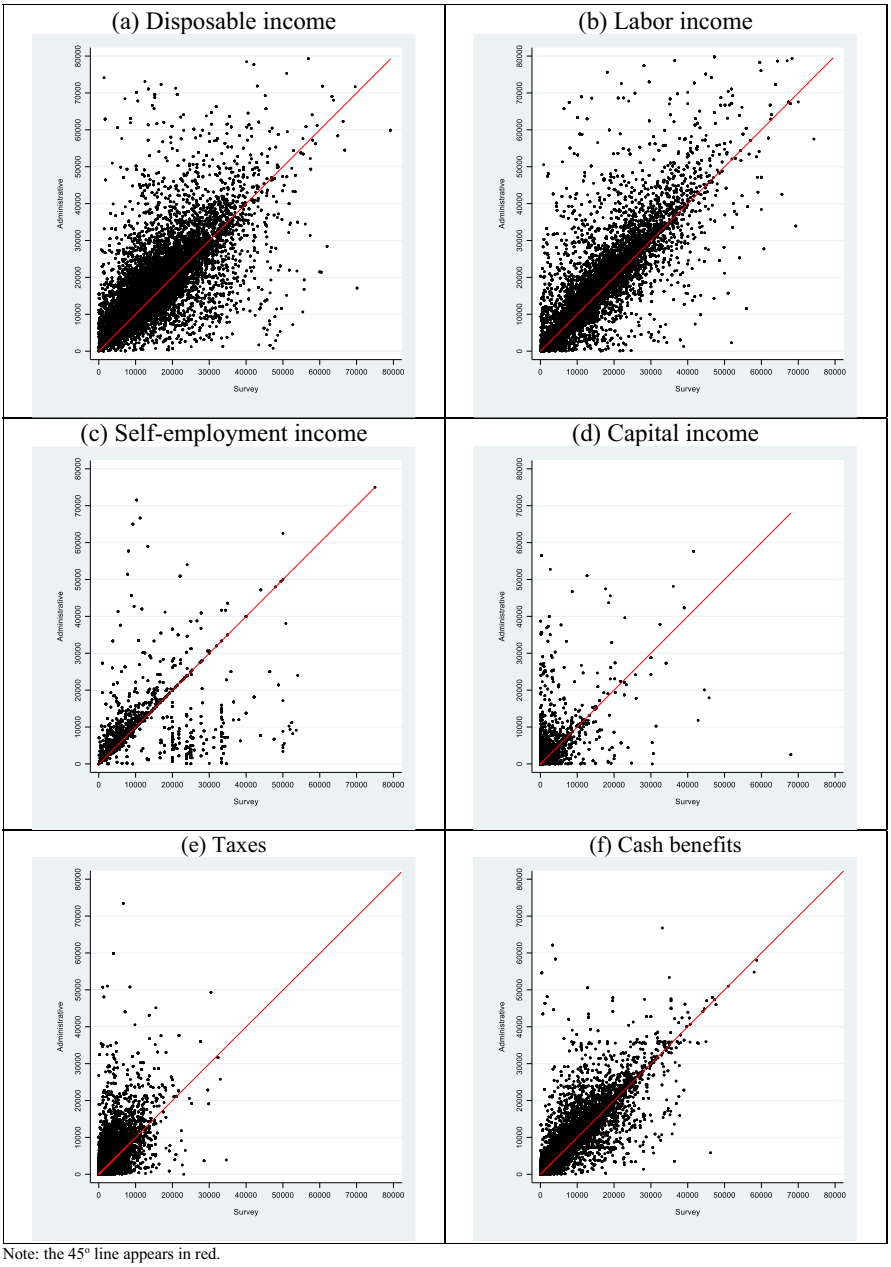
All the variations are significant, and except in the case of self-employment, income increases when changing from interview to administrative data.[9] In line with INE's predictions, we find a relevant increase in disposable income (higher than 14%).[10] This increase is mainly explained by both the higher levels of labor income (given its weight on total income), which increases more than 17%, and those of capital income, which is more than 2.5 times higher in the administrative than it is in the survey data. The better coverage of capital income offered by administrative data is an important improvement for this income source that has traditionally exhibited large amounts of underreporting. The opposite occurs in the case of self-employment income, which is 10% lower in the administrative data than it is in the survey data. The change from the use of survey data to administrative data causes the measured average level of cash

[7] We adjust incomes using the OECD modified equivalence scale.

[8] The same results emerge when a parametric t-test for paired groups is used.

[9] These incomes are usually underreported in household surveys. They receive a specific treatment in the personal income tax in Spain. Thus, in many cases, the related individual's taxable income considerably differs from the income actually received.

[10] This result is very similar to that of Goerlich (2020), who finds that the income reported by administrative data was approximately 16% higher than that reported by survey data.



Note: the 45° line appears in red.

Fig. 2 Scatter plots of the data from administrative versus survey data for disposable income and for each data source. *Note*: The 45° line appears in red

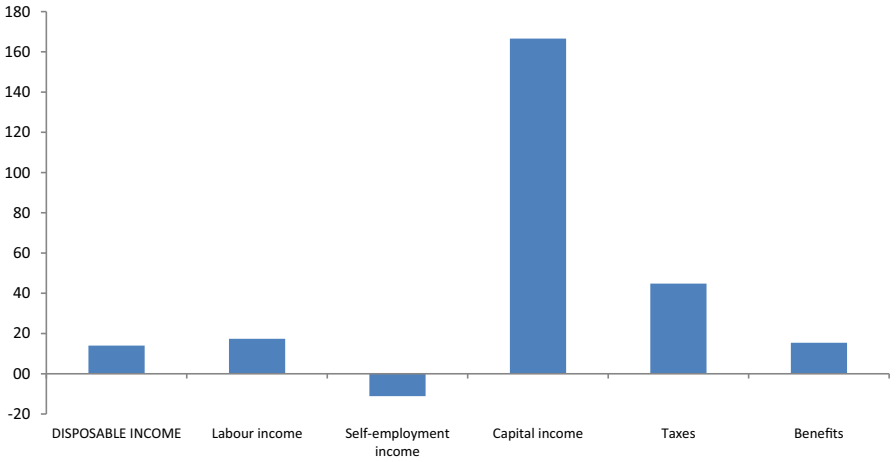


Fig. 3 Change (%) in mean incomes when changing from survey to administrative data

benefits to increase by 15.4%. This change is much larger in the case of average taxes and social contributions (approximately 50%). Again, this difference can decisively affect the measurement of the redistributive effects of taxes.

To enable a better appreciation of the ways in which this change affects the different components of the distribution of disposable income, Fig. 4 shows the increase (in %) in the average income of each percentile when changing from survey data to administrative data. In the last year with data available for both methods of collecting income, namely 2014, the latter increase the average level of measured disposable income by more than 10% in all percentiles; however, this difference is not uniform throughout the distribution. It is especially prominent in the first percentiles, where income is higher by more than 30%, and in the highest income stratum, where income

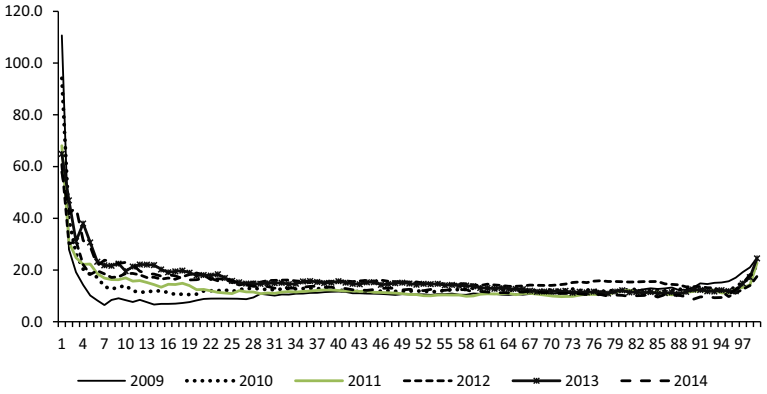


Fig. 4 Growth (%) in the average levels of disposable income by percentile when changing to administrative data

Fig. 5 Growth (%) in the average levels of the different income sources by disposable income deciles when changing to administrative data

is approximately 20% higher and exhibits a greater difference than do the previous percentiles. Moreover, as Fig. 4 shows, this higher income growth in both tails of the distribution is systematically repeated in the different waves for which the two methods of collecting income are available.

Therefore, our results show that changes that occur with the shift to administrative data are much larger in the tails—especially in the lower tail—than in the mid-range incomes. This finding is very important to understand the most general impact of the change in the method of collecting data.

To better characterize the change in the different areas of the distribution when moving to administrative data, Fig. 5 shows the change in the mean values of each income source by disposable income deciles. In the case of labor income—the main source of income—the pattern described above for the distribution of disposable income is repeated: The average value of this source of income increases in all deciles and more so in the extreme deciles than in the intermediate ones. Something similar happens with social benefits, although with a greater increase in the higher deciles. The change in self-employment income and taxes is more "progressive," with a decreasing profile in the former and the opposite in the latter. The most peculiar behavior is that of capital income, with a generally decreasing profile as income increases. It should be kept in mind, however, that there are many households in the first deciles that do not have this type of income or have very low capital incomes with both methods of income collection.

4 Effects on inequality

One of the most important consequences of the examined change in the method of income collection is the possible effect that it can have on the measurement of inequality, since administrative data seem to better report some types of income. In fact, a reduction in measured inequality is expected when changing from interview to administrative data due to the larger increase in income reported by the lower-income

Table 1 Inequality indicators

	Survey data	Administrative data	Change (%)
Gini	0.346	0.339	− 2.0**
GE (c = 0)	0.234	0.219	− 6.4**
GE (c = 1)	0.191	0.190	− 0.3
GE (c = 2)	0.210	0.220	4.9*
Atkinson (e = 0.5)	0.106	0.099	− 6.9**
Atkinson (e = 1)	0.208	0.196	− 5.7**
Atkinson (e = 2)	0.766	0.878	14.6*
P9/P1	5.84	5.16	− 11.6***
P9/P5	2.14	2.07	− 3.4**
P5/P1	2.73	2.49	− 8.5***

***Significant at the 1% level, **significant at the 5% level, *significant at the 10% level

percentiles. They are more likely to underreport their real income in a survey. However, this difference is possibly smoothed by the additional income also recorded by the richest percentiles. In this section, we address this issue by comparing the values of very well-known inequality indices based on both survey and administrative data. First, we carry out an analysis of disposable income, and then we move to a disaggregated analysis of the various income sources.

4.1 Effects on the inequality measures of disposable income

To analyze the effect of the change in the income collection method on inequality, we consider first the following inequality indices: the Gini index, generalized entropy index and Atkinson index. The definition of these indices can be found in "Appendix B." Second, to get a more complete picture of the differences in inequality, we also estimate different percentile ratios (P9/P1; P9/P5; P5/P1).

The top panel of Table 1 shows the estimated values of all these indices computed based on the two income distributions displayed in Fig. 1a, namely the distributions of disposable income using survey and administrative data, as well as their differences in percentage (last column). In order to determine whether these differences are statistically significant, we calculate confidence intervals for the differences. If the (1-α)% confidence interval includes zero, we conclude that the difference is no longer significant at α% level (in a two-sided test); on the other hand, provided that the confidence interval does not include zero, we infer that the difference is significant at α% level. We use the software developed by Duclos and Araar (2006) to compute these confidence intervals, choosing the linearization method to obtain the standard errors.

The main conclusion drawn from Table 1 is that the differences between all the indices are significant, except that of GE(c = 1). For most indicators, inequality is reduced when moving to administrative data, although the magnitude of the change

depends on the index considered—the exceptions are GE (c = 2) and the Atkinson index with parameter 2. This result is related to the different interpretations of inequality that each index summarizes and to the different effect that the change in the income collection method has in both tails of the income distribution.

In the case of the generalized entropy index, the differences between the results corresponding to $c = 0$ (mean logarithmic deviation) and $c = 2$ (half the coefficient of variation squared) are clearly related to the changes shown in Fig. 3. When $c = 0$, the index undergoes its largest decrease, as it is weighted more heavily by the differences between the incomes in the lower tail of the distribution. However, when $c = 2$, the changes in the upper tail of the distribution are weighted more heavily, so the difference in this case is positive and not as considerable. Finally, when the changes are weighted equally throughout the distribution (i.e., when $c = 1$), the change in inequality is not significant.

A similar phenomenon occurs when estimating the Atkinson index for different values of the parameter e. As this parameter increases, the income transfers at the lower tail of the distribution are weighted more heavily than those at the upper tail. Hence, when e is at its highest point ($e = 2$), the use of administrative data induces a very large increase in measured inequality, whereas when this parameter is lower ($e = 0.5$), measured inequality decreases.

The percentile ratios reported in the bottom part of Table 1 provide interesting information, since they show that the reduction in inequality is mainly due to the narrowing of income differences in the lower part of the distribution. This finding helps to understand why according to the ECV sample this decreasing effect on inequality of administrative data—in contrast with a large strand of the literature showing that surveys tend to underestimate the levels of income concentration (Bollinger 1998)—takes place. One of the main effects of the use of administrative data in the ECV would be a further narrowing of the income distribution in its lower part.

The availability of different waves of the survey with administrative data and with the traditional interview method allows us to assess not only how inequality changes at a point in time when using administrative data but also what differences there are in the trend of inequality. It is possible to analyze the two types of distributions and the corresponding change in inequality between 2009, the first year for which microdata were provided with administrative data, and 2014, the last year in which INE continued to collect data using both methods. Figure 6 depicts these changes in percentage.

Figure 6 shows that, regardless of the indicator chosen to measure inequality and the method of income collection, inequality increased during the period considered. This increase is especially remarkable in the Atkinson index with the highest parameter of inequality aversion and is repeated, although more moderate, with the Gini index. A second relevant result is again that the growth of inequality during this period is considerably higher with interview data than with administrative records. This result is important, given that in those years the growth of inequality in Spain was much higher than in most European countries (Ayala and Cantó 2022). The use of interview data would make this difference even greater.

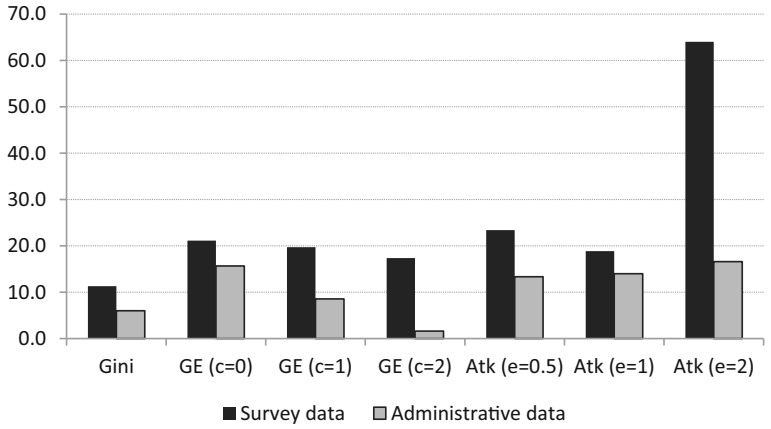


Fig. 6 Change in inequality indicators, 2009–2014 (%)

4.2 Effects on the inequality measures by income source

The different shapes of the distributions of each income source allow us to anticipate that the indicators that summarize inequality in each case will differ when measured by survey data versus administrative data. Specifically, in the case of capital income or taxes and social contributions, the magnitude of the difference observed between the two income collection methods enables the simple prediction of very different levels of inequality as measured by each source. However, the better coverage of capital income offered by administrative data could also affect inequality levels due to the differences in disposable income as measured by survey data and administrative data. Moreover, some recent studies have emphasized the effect that the increase in the weight of capital income can have on inequality in terms of the distribution of disposable income (Milanovic 2017; Bengtsson and Waldenström 2018).

Table 2 shows the values of the first inequality indices described in the previous section for each income source, computed with both survey and administrative data, as well as the differences between them and the statistical significance of these differences.[11] These results are different depending on the index and the income source considered. First, it can be observed that the sign and significance of the changes in the Gini index differ the most from those of the other indices. According to this index, the methodological change causes most income sources to be more equally distributed, although this difference is only significant for capital income and cash benefits. The only exception is taxes that display an increase in measured inequality although it is not significant. A similar pattern is found for the Atkinson index ($e = 0.5$), which represents a lower aversion to inequality; however, in this case, all the changes are highly significant. These particular phenomena could be explained by the fact that the changes that occur with the shift to administrative data are much larger in the tails than in the mid-range incomes.

[11] With each income source, confidence intervals for the differences between indices are constructed to evaluate whether the difference is significant or not.

Table 2 Inequality indicators by income source

	Survey data	Administrative data	Change (%)
Gini			
Labor income	0.588	0.585	− 0.5
Self-employment income	0.959	0.953	− 0.6
Capital income	0.948	0.872	− 8.0***
Taxes	0.688	0.693	0.7
Cash benefits	0.675	0.658	− 2.4***
GE (c = 0)			
Labor income	0.361	0.496	37.4***
Self-employment income	0.452	0.991	119.0***
Capital income	1.765	1.735	− 1.7
Taxes	0.582	0.931	60.1***
Cash benefits	0.483	0.623	29.0***
GE (c = 1)			
Labor income	0.271	0.341	25.9***
Self-employment income	0.365	0.606	66.2***
Capital income	1.167	1.273	9.0***
Taxes	0.445	0.634	42.3***
Cash benefits	0.365	0.438	19.9***
GE(c = 2)			
Labor income	0.640	0.669	4.6*
Self-employment income	6.827	7.028	2.9
Capital income	13.76	6.175	− 55.1***
Taxes	1.058	1.373	29.7***
Cash benefits	1.000	0.989	− 1.1
Atkinson (e = 0.5)			
Labor income	0.407	0.381	− 6.2***
Self-employment income	0.893	0.865	− 3.1***
Capital income	0.896	0.734	− 18.1***
Taxes	0.335	0.348	3.9**
Cash benefits	0.504	0.463	− 8.2***
Atkinson (e = 1)			
Labor income	0.303	0.391	29.0***
Self-employment income	0.364	0.629	72.8***

Table 2 (continued)

	Survey data	Administrative data	Change (%)
Capital income	0.829	0.824	− 0.6
Taxes	0.441	0.606	37.4***
Cash benefits	0.383	0.464	21.1***
Atkinson (e = 2)			
Labor income	0.693	0.959	38.4***
Self-employment income	0.852	0.982	15.3***
Capital income	0.984	0.976	− 0.8
Taxes	1.126	1.049	− 6.8***
Cash benefits	0.757	0.953	26.0***

***Significant at the 1% level, **significant at the 5% level, *significant at the 10% level

Noticeably, different results are found for the other indices. In fact, according to the GE indices with $c = 0$ and $c = 1$ and the Atkinson index with $e = 1$, the change from survey data to administrative data induces a significantly more unequal distribution for all the income sources. Capital income is an exception to this phenomenon, and, in some cases, the change even becomes negative and insignificant for this income source. Moreover, the effect of the methodological change on measured inequality is larger as assessed by the GE index ($c = 0$).

Table 2 also reveals that, depending on the indicator considered, the most important changes in measured inequality arise in terms of the different income sources. According to the Gini, GE ($c = 2$) and Atkinson ($e = 0.5$) indices, the most remarkable changes are those affecting capital income, where a reduction in measured inequality is observed. In the GE ($c = 0$), GE ($c = 1$) and Atkinson ($e = 1$) indices, the change in self-employment income stands out, indicating higher levels of inequality.

5 Effects on the structure of inequality

A very important dimension of inequality that may be affected when changing from the use of survey data to that of administrative data is the structure of inequality in terms of income sources. The inequality observed in the distribution of disposable income, discussed in subsection 4.1, is the result of the inequality observed in the different sources of income, the weight of each source in terms of the total income, and the degrees of association between these sources (Shorrocks 1982).

The methods by which the contribution to income inequality of each income source can be decomposed have been expanded in two ways. One consists of considering a wider variety of inequality indicators, and another consists of combining the changes in individual characteristics and the changes in income sources and their dependences into a single method of analysis. The pioneering contribution to these approaches was that of DiNardo et al. (1996), which was further developed in the field of disposable income inequality by Daly and Valleta (2006). This work consists of a structural method in

which the contribution of each component is identified through a counterfactual. This method compares the distributions with the observable characteristics of the current moment and those that would exist if those characteristics had not changed over time.[12]

In our specific case, this analysis is especially simplified because there are no changes in the characteristics of individuals and households, since they are the same for both datasets. In our approach to the change in the structure of inequality induced by the shift to the use of administrative data, we will combine two types of analysis that attempt to overcome the limits of the cited approaches. First, as proposed by Larrimore (2014), we create a counterfactual that maintains the order of the initial distribution of income—as per the survey data—to determine how the changes in the marginal distributions of each income source induced by the shift to using administrative data affect the differences in the distribution of disposable income. The main limitation of this approach is that it does not explicitly quantify the interrelations between the different income sources. For this reason, in a second analysis (Sect. 5.2), we evaluate the changes in the dependences between the income sources through some extensions of Spearman's coefficient based on the copula functions.

5.1 Effects on the contribution to inequality of income sources

Let X_i denote the total income of individual i in the initial distribution corresponding to the survey data, and let us assume that this variable can be represented as the sum of the incomes obtained from each income source, X_{ki}, with $k = 1,\ldots,d$, i.e.,

$$X_i = X_{1i} + X_{2i} + \cdots + X_{di} \tag{1}$$

To estimate the impact of the change in the distribution of the first income source on the inequality of disposable income, the income of that source based on survey data (X_{1i}) can be replaced with the income of the same source based on administrative data (X'_{1i}) for each individual as follows:

$$X_i^{'1} = X'_{1i} + X_{2i} + \cdots + X_{di} \tag{2}$$

The difference between the inequality of the initial distribution and that of this simulated distribution can be interpreted as the contribution of the first income source to inequality. In a similar way, we could define $X_i^{'j}$ for $j = 2,\ldots,d$ and compute the contribution of the jth income source as the difference between the inequality of the initial distribution and that of the jth simulated distribution. However, the sum of these contributions does not equal 100% of the inequality, since we must add the effects of the dependences between the income sources.

Figure 7 shows the change in the contribution of each income source to income inequality when replacing survey data with administrative data for the different sources considered. For instance, the value of the Gini index and the 1st component is computed

[12] The major criticism that this approach has received is that its results can be very sensitive to the specification of the model (Cowell and Fiorio 2011) and that it can only provide a limited approximation in terms of identifying the interrelations between income sources.

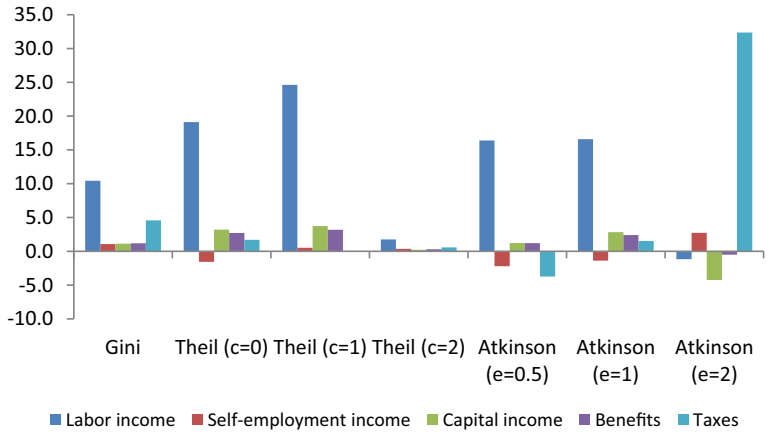


Fig. 7 Change (%) in the measured contribution to inequality of the income sources when changing to administrative data

as follows:

$$100 \times \frac{G\left(X_1'^1, \ldots, X_n'^1\right) - G(X_1, \ldots, X_n)}{G(X_1, \ldots, X_n)}$$

For the other components and indices, the values displayed in Fig. 7 are computed in a similar way.

Several conclusions can be drawn from this figure. First, for all the indices, except the Atkinson index, when the aversion to inequality is at its highest ($e = 2$), the shift from the use of survey data to that of administrative data increases the measured contribution of labor income to inequality, and this increase is the highest exhibited among the income sources. However, in the case of the Atkinson index ($e = 2$), tax income shows the largest change in terms of its contribution to inequality. Second, the changes in the contributions of capital income and benefit income are similar in terms of sign—although they are of different sizes—to that of labor income: The shift from the use of survey data to administrative data increases their measured contribution to all the inequality indices except the Atkinson index ($e = 2$). Finally, the effect of the changes in self-employment income is not relevant.

5.2 Effects on the dependences between income sources

One key aspect in the decomposition of total income by income sources is the interaction among these sources because of its possible effect on income inequality; see Shorrocks (1982), Lerman and Yitzhaki (1985) and Aaberge et al. (2018). For instance, a situation where one individual scores the lowest in all income sources, another individual scores the second lowest in all income sources, and so on up to another individual who scores the highest in all sources, conveys a different degree of interdependence than a situation where some individuals score high in some sources and low in others.

In these situations, even if the distribution by different income sources was equal, the level of inequality of total income would be different due solely to the different dependence between the income sources. Hence, incorporating such dependence will provide new insights into the measuring of inequality.

However, measuring the dependence between income sources is a multidimensional problem that involves more than two variables and special care is required, since some bivariate dependence properties are not preserved in higher dimensions; see Durante et al. (2014). For example, taking the average of pairwise bivariate measures, over all distinct bivariate margins, ignores the multivariate structure and could conceal important aspects of multivariate dependence. Actually, there are examples of variables that are pairwise independent but display some type of multivariate dependency; see Nelsen (1996). Moreover, since the variables we are dealing with are not Gaussian (see Fig. 1), we need to capture other types of dependences beyond linear correlation (Embrechts et al. 2002).

In order to do that, in this paper, we use three copula-based multivariate extensions of the well-known Spearman's rank coefficient.[13] These coefficients focus on the positions of the individuals across variables, rather than on the specific values that the variables attain for such individuals, and capture to what extent the positions in the different variables are aligned. (The more aligned they are, the stronger the dependence.)[14] In a bivariate setting, Spearman's rank coefficient can be defined either as a functional of the joint cumulative distribution function or as the joint survival function, with exactly the same result; see Joe (1990). However, in a multivariate setting, this equivalence no longer holds and this leads to (at least) two different coefficients that capture distinct aspects of multivariate dependence; see Nelsen (1996, 2002).

The first coefficient we consider, denoted as ρ_{d}^{-}, measures, through a rescaled average, "how far" from independence the joint cumulative distribution function of the three income sources is. In a similar fashion, the second coefficient we use, denoted as ρ_{d}^{+}, measures departure from independence by using the survival function rather than the cumulative distribution function. In doing so, ρ_{d}^{-} measures whether the probability is concentrated around the lower part of the distribution, whereas ρ_{d}^{+} stresses the dependency in the upper part of the distribution. The third coefficient we use is the average of the two generalizations above, namely $\rho_d = \left(\rho_d^{-} + \rho_d^{+}\right)/2$. Noticeably, when $d = 3$, this coefficient is the average of the three possible pairwise Spearman's coefficients.

If income sources are independent, these three coefficients will be equal to zero, whereas in the case of maximal dependence, i.e., when the positions of individuals in all income sources are the same, all the coefficients will reach their maximum value 1. Moreover, if the joint distribution of the three income sources was symmetric, the three coefficients would coincide ($\rho_{\mathrm{d}}^{-} = \rho_{\mathrm{d}}^{+} = \rho_d$). Furthermore, a positive value of

[13] For a further description of these coefficients, see Appendix C and the references therein. An application of these coefficients and other related measures in Welfare Economics can be found García-Gómez et al. (2021). For other copula-based measures of multivariate dependence, see Schmid et al. (2010).

[14] To ease the interpretation of the results, we aggregate the income sources into three components (d = 3) having the same polarity with respect to inequality, namely, labor income (labor income plus self-employment income), capital income, and n taxes (taxes minus cash transfers).

Table 3 Copula-based measures of orthant dependence between the income sources

	Survey data	Administrative data	Change (%)
ρ_d^+	0.311^{***}	0.328^{***}	$5.3\%^{***}$
ρ_d^-	0.274^{***}	0.258^{***}	$-5.8\%^{***}$
ρ_d	0.293^{***}	0.293^{***}	0.1%

***Significant at the 1% level, **significant at the 5% level, *significant at the 10% level

ρ_{d}^- indicates that income sources are more likely to simultaneously take small values than in the case of independence, whereas if ρ_{d}^+ is positive, the income sources are more likely to simultaneously take large values than in the case of independence. For our purposes, and given that the variables we are dealing with are not symmetric (see Fig. 1), the coefficients ρ_d^- and ρ_{d}^+ are preferable, as they reveal some forms of dependences that ρ_d fails to detect. For example, the joint distribution of income sources might become more concentrated around its upper tail or its lower tail and such differences, captured by ρ_d^- and ρ_{d}^+, respectively, could be missed by ρ_d.

To check whether the degree of dependence between income sources changes depending on the type of data used, Table 3 presents the estimates of the three coefficients described above computed with both data sources. Since our variables are not strictly continuous, we use the estimators proposed by Genest et al. (2013). To compute their standard errors, we approximate their bootstrap distributions by resampling with replacement repeatedly from the original data (1,000 subsamples) and we test whether the difference between the survey and administrative data is significant based on the bootstrap distributions obtained.

Several conclusions emerge from this table. First, regardless of the coefficient and the data source used, there is a positive and significant multivariate dependence between the income sources. Second, regardless of the data collection method, the largest coefficient of multivariate dependence is ρ_d^+. This means that the simultaneous occurrence of high values in the three income components—i.e., households with simultaneously high labor incomes, high capital incomes and high net taxes—is more likely than the simultaneous occurrence of low values of the three income components. Moreover, the change of data collection method affects not only the marginal distributions of the income sources and inequality measures, but also the dependency of the income sources. In particular, the coefficient ρ_d^+ is significantly lower with survey data than with administrative data, that is, the simultaneous occurrence of "good" rankings in the three income components is more likely to arise with administrative data than with survey data. By contrast, the coefficient ρ_d^- is significantly greater in the case of survey data than in administrative data. This means that low values in the three income components are more likely to simultaneously occur with survey data than with administrative data. These features could be related to the changes in the tails of the income distribution when moving from one data source to another (see Sect. 3.2). Finally, it is worth pointing out that there is no significant difference between the two types of data in terms of the coefficient ρ_d. These results underline the importance of using ρ_d^- and ρ_d^+, as they capture some types of dependences that could be important,

but left undetected if only computing the coefficient ρ_d. Hence, simple methods, like averaging pairwise Spearman's coefficients, conceal important patterns that could be important for a better understanding of the effects of income collection methods on the structure of inequality.

6 Conclusions

The past decade has witnessed an intense debate regarding the trends and consequences of inequality. To understand its changes, determinants and implications, robust datasets are required. In most countries, these data come from household surveys that provide detailed information on the different income sources that each individual receives. However, survey income data are affected by various problems related to measurement error that may limit their ability to provide accurate diagnoses for decision-making. Due to these limitations, some countries are replacing the income data collected through these surveys with administrative data. Such a process can have important effects on the measurement of inequality, and it can also affect the optimal design of redistributive policies. Hence, it is necessary to evaluate how this change affects not only the general indicators of inequality but also the structure of inequality by income source as well as that of the dependences between these sources.

In this paper, we have addressed this problem using data from the main household survey in Spain (ECV), which provides both survey and administrative data for the same individuals and households. Specifically, we have shown that the shift to administrative data has effects on the averages of the different income variables included in the survey as well as on the magnitude of measured inequality based on the distribution of income and its structure by income source.

First, our analysis confirms a significant increase in measured household disposable income when administrative data are used. This increase is especially prominent in the case of capital income, as the rate of underreporting in survey data is usually very high for this type of income. The opposite occurs in the case of self-employment income, which reflects the special treatment that these incomes receive in the personal income tax. Second, our results show that the examined change in the method of income collection also affects the shape of the income distribution. Our main finding is that changes that occur with the shift to administrative data are much larger in the tails—especially in the lower tail—than in the mid-range incomes.

Administrative data also yield lower levels of inequality in disposable income according to most inequality measures used, although the magnitude of the change depends on the index considered. Another relevant result is that the growth of inequality for the years in which both types of data are available is considerably higher with interview data than with administrative records.

In general terms, this potential reduction in inequality with administrative data differentiates the Spanish case from what is obtained in other studies for other countries. Nevertheless, those indicators that assign considerably more weight to the upper tail of the distribution suggest higher levels of inequality. Our results are mainly explained by both the higher levels of labor income and cash benefits with administrative data and the remarkable increase in taxes, with a higher growth of the latter in the richest

percentiles. It must also be noted that what we do in the paper is to compare the distribution of income of two different data sources with the ECV sample. This is different to compare the distribution of income of a survey with the distribution of income of the universe of administrative data. It may be the case that the lower inequality that we find with administrative data is due to the sampling method used in the ECV which fails to capture the very top of the income distribution.

A third contribution of this paper is the identification of the income sources whose inequality levels are most sensitive to the use of administrative data. Employing administrative data in the ECV rather than survey data leads to increased equality in the distributions of all the income sources. However, it induces a substantial modification to the structure of inequality that increases, above all, the contributions of capital income, cash benefits and taxes to inequality.

Finally, we have shown that there are also significant differences in the dependences between income sources depending on which income data collection method is used. Specifically, changing from the use of survey data to administrative data induces a higher simultaneous occurrence of high values in the different income sources and a lower one of low values. This effect could be important for an adequate design of redistributive policies, as it affects the type of relationships among the different income sources that could determine the effectiveness of public intervention. If the relationships between income sources differ across data sources, the design of policies based on these relationships may be affected.

The important differences that we document, therefore, call into question the conclusions anticipated by the institution producing the data (INE), which predicted only minor changes. This paper therefore contributes to the literature on the quality of income data by determining how the shift to administrative data may affect the observed levels and characteristics of inequality, and, most innovatively, how they show a different structure in terms of the dependences between income sources. Our results should be a wake-up call to interpret the results based on only one source of income data with caution.

7 Appendix A: Income definitions

The concept of income we use is household disposable income as defined by Eurostat in the European Union Living Conditions Survey (EU-SILC). Total gross household income is computed as the sum for all household members of gross personal income components. Gross means that neither taxes nor social contributions have been deducted at source. Total disposable income is the total gross household income minus tax on income and social insurance contributions, regular taxes on wealth and regular inter-household cash transfer paid.

The different variables used to calculate household disposable income are as follows:

– *Employee income*: total remuneration, in cash or in kind, payable by an employer to an employee in return for work done by the latter during the income reference period. The employee income is broken down into gross employee cash or near

cash income; gross non-cash employee income; and employers' social insurance contributions. The latter are part of the employee's gross income but not of the household's disposable income.

- *Self-employment income*: income received, during the income reference period, by individuals, for themselves or in respect of their family members, as a result of their current or former involvement in self-employed work. Self-employment income is broken down into gross cash profits or losses from self-employment (including royalties) and value of goods produced for own consumption.
- *Capital income*: income received, less expenses, occurring during the income reference period by the owner of a financial asset or a tangible non-produced asset (land) in return for providing funds to, or putting the tangible non-produced asset at the disposal of another institutional unit. It is broken down into interest, dividends, profits from capital investment in an unincorporated business; income from rental of a property or land; and pensions received from individual private plans.
- *Transfers received* (benefits): current transfers received during the income reference period by households intended to relieve them from the financial burden of a number of risks or needs, made through collectively organized schemes, or outside such schemes by government units and nonprofit institutions serving households. The social benefits collected at the household level are the following: family-/children-related allowance; housing allowances; and social exclusion not elsewhere classified. The benefits collected at the individual level are the following: unemployment benefits; old-age benefits; survivor' benefits; sickness benefits; disability benefits; and education-related allowances. Regular inter-household cash transfers received are also classified under current transfer received. They refer to regular monetary amounts received, during the income reference period, from other households or persons.
- *Taxes*: sum of tax on income and social insurance contributions, regular taxes on wealth and regular inter-household cash transfer paid. Tax on income refers to taxes on income, profits and capital gains. They are assessed on the actual or presumed income of individuals, households or the tax unit. They include taxes assessed on holdings of property, land or real estate when these holdings are used as a basis for estimating the income of their owners. Social insurance contributions refer to contributions by employees', the self-employed and, if applicable, the unemployed. Regular taxes on wealth refer to taxes that are payable periodically on the ownership or use of land or buildings by owners, and current taxes on net wealth and on other assets (jewelry, other external signs of wealth). Repayments/receipts for tax adjustments refer to the money paid to/received from the Tax Agency related to the income received. This applies only in cases where taxes at source are deducted from income received and the Tax Agency compares the amount of taxes of income paid at source with the taxes that correspond to those paid over the total income received for the "tax unit." We also include under taxes the regular inter-household transfers paid.

8 Appendix B: Inequality indices

Let us denote the number of individuals considered as n, let y_i denote the income of person i with $i = 1,..,n$, and let $\overline{y}$ denote the income mean. The indices are defined as follows:

- Gini index

$$G = \frac{1}{2n^2\overline{y}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| y_i - y_j \right|$$

- Generalized entropy index

$$GE(c) = \begin{cases} \frac{1}{c(c-1)} \frac{\sum_{i=1}^{n} \left[\left(\frac{y_i}{\overline{y}} \right)^c - 1 \right]}{n}, c \neq 0, 1 \\ \frac{1}{n} \sum_{i=1}^{n} log\left(\frac{\overline{y}}{y_i} \right), c = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\overline{y}} \right) log\left(\frac{y_i}{\overline{y}} \right), c = 1 \end{cases}$$

- Atkinson index

$$Atk(e) = \begin{cases} 1 - \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\overline{y}} \right)^{1-e} \right]^{1/(1-e)}, e > 0 \,\&\, e \neq 1 \\ 1 - \frac{\prod_{i=1}^{n} (y_i)^{\frac{1}{n}}}{\overline{y}}, e = 1 \end{cases}$$

9 Appendix C: Copula-based generalizations of Spearman's correlation coefficient

The copula approach focuses on the positions of the individuals across variables, rather than on the specific values that the corresponding variables attain for such individuals. In our setting, this entails transforming the outcomes of one individual in all income sources into the positions that this individual attains in these sources as compared to other individuals. The transformation is achieved as follows. Let the random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$ represent the d income sources, and let F_j denote the marginal distribution function of source j, with $j = 1,\ldots,d$. We transform each original variable X_j by applying the probability integral transform, so that we obtain the random variable $U_j = F_j(X_j)$, which describes the relative positions of the individuals in the jth income source and follows a uniform U(0,1) distribution. Moreover, the random vector $\boldsymbol{U} = (U_1,\ldots,U_\mathrm{d})$ captures the distribution of these positions in all income sources and its joint cumulative distribution function turns out to be a copula function denoted as C. In particular, for every position vector $\boldsymbol{u} = (u_1,\ldots,u_\mathrm{d}) \in [0,1]^d$, the value $C(\boldsymbol{u})$ represents the probability that the position of an individual is outranked by $\boldsymbol{u}$, i.e.,

$C(\boldsymbol{u}) = p(U_1 \leq u_1,\ldots,U_d \leq u_d)$. By contrast, the probability that the position of an individual outranks the position vector $\boldsymbol{u}$ is given by the survival function $\overline{C}$, which is defined as $\overline{C}(\boldsymbol{u}) = p(U_1 > u_1,\ldots,U_d > u_d)$. Hence, while the copula function looks "downwards" at the proportion of individuals who are outranked, the survival function looks "upwards" at the proportion of individuals who are outranking (Decancq 2020).

From the statistical point of view, the importance of copulas comes up in the Sklar's theorem. This theorem establishes that, if $\boldsymbol{X} = (X_1,\ldots,X_d)$ is a d-dimensional random variable with joint distribution function F and univariate marginals $F_1,\ldots,F_d$, then, for all $\boldsymbol{x} = (x_1,\ldots,x_d) \in R^d$, there exists a copula C such that $\boldsymbol{F}$ can be written as:

$$\boldsymbol{F}(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{A1}$$

If $F_1,\ldots,F_d$ are all continuous, then C is unique; otherwise, C is uniquely determined on $\mathrm{Ran}F_1 \times \cdots \times \mathrm{Ran}F_d$. Conversely, if C is a d-copula and $F_1,\ldots,F_d$ are univariate distribution functions, then the function $\boldsymbol{F}$ in (A1) is a distribution function with margins $F_1,\ldots,F_d$. Thus, copulas link joint distributions functions to their univariate marginals.

Each copula function C is bounded by its so-called Fréchet–Hoeffding bounds, $W(\boldsymbol{u}) \leq C(\boldsymbol{u}) \leq M(\boldsymbol{u})$, where $W(\boldsymbol{u}) = \max(u_1 + \ldots + u_d - d + 1,0)$ and $M(\boldsymbol{u}) = \min(u_1,\ldots,u_d)$. M is always a copula but W is a copula only if $d = 2$. M represents the "maximal concordance," i.e., the state where each component of $\boldsymbol{X}$ is an almost surely increasing function of every other component. Unlike, if $\boldsymbol{X}$ is a vector of independent random variables, then its copula is the independent copula Π defined as $\Pi(\boldsymbol{u}) = u_1 \cdots u_d$.

Copulas are closely related to some measures of multivariate dependence beyond linear correlation. In this paper, we focus on multivariate dependence measures which are copula-based multivariate generalizations of the well-known bivariate Spearman's rho.

The first coefficient we consider, denoted as ρ_{d}^{-}, is a measure of multivariate dependence derived from average lower orthant dependence, that is based on the work of Wolff (1980) and Nelsen (1996), and it is defined as follows:

$$\rho_{\mathrm{d}}^{-} = \frac{(d+1)}{2^d - (d+1)} \int_{\boldsymbol{I}^d} [C(\boldsymbol{u}) - \Pi(\boldsymbol{u})]\mathrm{d}\boldsymbol{u}. \tag{A2}$$

To some extent, this coefficient measures the rescaled "average distance" between our multivariate data (C) and independence (Π) from a lower perspective, i.e., by measuring whether the probability that all variables are simultaneously low is at least as large as in the case of independence. In a similar fashion, Nelsen (1996) defined a measure of multivariate association ρ_{d}^{+} derived from average upper orthant dependence, given by:

$$\rho_{\mathrm{d}}^{+} = \frac{(d+1)}{2^d - (d+1)} \int_{\boldsymbol{I}^d} \left[\overline{C}(\boldsymbol{u}) - \overline{\Pi}(\boldsymbol{u})\right]\mathrm{d}\boldsymbol{u} \tag{A3}$$

This measure can be regarded as a rescaled "average distance" between our multivariate data and independence from an upper perspective, i.e., by comparing the probability that all current variables take simultaneously high values ($\overline{C}$) with the value of this probability if the variables were independent ($\overline{\Pi}$). The third multivariate version of Spearman's rho, developed by Nelsen (2002) as a measure of multivariate concordance, is the average of the two generalizations described above, namely:

$$\rho_d = \frac{1}{2}\left(\rho_d^- + \rho_d^+\right) \tag{A4}$$

when $C = M$, the three measures defined above, ρ_{d}^-, ρ_{d}^+ and ρ_d, attain their maximum value, 1, and they all become zero when $C = \Pi$. The lower bound for these three coefficients is $[2^{\mathrm{d}}—(\mathrm{d}+1)!]/\{\mathrm{d}![2^{\mathrm{d}}—(\mathrm{d}+1)]\}$. When $\mathrm{d} = 2$, the three coefficients above reduce to the well-known bivariate Spearman's rho. When $d = 3$, ρ_3 is the average of the three possible pairwise Spearman's coefficients.

In practice, copula C is unknown, and the coefficients in (A2)–(A4) must be estimated from the data with the empirical copula. Pérez and Prieto-Alaiz (2016) propose feasible nonparametric estimators of ρ_d^- and ρ_d^+ for continuous variables that are easy to compute and have good asymptotic properties. The average of these estimators could be used to estimate ρ_d. If the variables are not continuous, the underlying copula C in (A1) is no longer unique; thus, the coefficients in (A2)–(A4) and their corresponding estimators should be modified; see Genest et al. (2013) and the references therein.

Funding The authors acknowledge financial support received from Comunidad de Madrid (H2019/HUM-5793).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Aaberge R, Atkinson AB, Königs S (2018) From classes to copulas: wages, capital, and top incomes. J Econ Inequal 16(2):295–320

Atkinson AB (2015) Inequality. Harvard University Press, Cambridge, Ma

Auten G, Splinter D (2019) Top 1 percent income shares: comparing estimates using tax data. AEA Pap Proc 109:307–311
Ayala L, Cantó O (2022) *Radiografía de medio siglo de desigualdad en España*. Palma: Observatorio Social de la Fundación la Caixa
Bengtsson E, Waldenström D (2018) Capital shares and income inequality: evidence from the long run. J Econ Hist 78:712–743
Bollinger CR (1998) Measurement error in the current population survey: a nonparametric look. J Law Econ 16:576–594
Burkhauser RV, Feng S, Jenkins SP, Larrimore J (2012) Recent trends in top income shares in the United States: reconciling estimates from march CPS and IRS tax return data. Rev Econ Stat 94:371–388
Burkhauser RV, Hérault N, Jenkins SP, Wilkins R (2016) What has been happening to UK income inequality since the mid-1990s? Answers from reconciled and combined household survey and tax return data. IZA Discussion Paper 9718
Burkhauser RV, Hérault N, Jenkins SP, Wilkins R (2018) Survey under-coverage of top incomes and estimation of inequality: what is the role of the UK's SPI adjustment. Fisc Stud 39:213–240
Carr MD, Wiemers EE (2018) New evidence on earnings volatility in survey and administrative data. Am Econ Rev Pap Proc 108:287–291
Courtemanche C, Denteh A, Tchernis R (2019) Estimating the associations between SNAP and food insecurity, obesity, and food purchases with imperfect administrative measures of participation. South Econ J 86:202–228
Cowell F, Fiorio CV (2011) Inequality decompositions—a reconciliation. J Econ Inequal 9:509–528
Dahl M, DeLeire T, Schwabish JA (2011) Estimates of year-to-year volatility in earnings and in household incomes from administrative, survey, and matched data. J Hum Resour 46:750–774
Daly MC, Valletta RG (2006) Inequality and poverty in the United States: the effects of rising dispersion of men's earnings and changing family behavior. Economica 73:75–98
Decancq K (2020) Measuring cumulative deprivation and affluence based on the diagonal dependence diagram. Metron 78(2):103–117
DiNardo JE, Fortin N, Lemieux T (1996) Labour market institutions and the distribution of wages, 1973–1992: a semi-parametric approach. Econometrica 64:1001–1044
Duclos J-I, Araar A (2006) Poverty and equity measurement, Policy and estimation with DAD. Kluwer Academic Publishers, London
Durante F, Nelsen RB, Quesada-Molina JJ, Úbeda-Flores M (2014) Pairwise and global dependence in trivariate copula models. In: Laurent A, Strauss O, Bouchon-Meunier B, Yager RR (eds) Information processing and management of uncertainty in knowledge-based systems. Springer, pp 243–251
Embrechts P, McNeil A, Straumann D (2002) Correlation and dependence in risk management: properties and pitfalls. In: Dempster M (ed) Risk management: value at risk and beyond. Cambridge University Press, Cambridge, pp 176–223
European Statistical System (2019) Quality assurance framework of the european statistical system. Eurostat
García-Gómez C, Pérez A, Prieto-Alaiz M (2021) Copula-based analysis of multivariate dependence patterns between dimensions of poverty in Europe. Rev Income Wealth 67:165–195
Genest C, Nelšehová J, Rémillard B (2013) On the estimation of Spearman's rho and related tests of independence for possibly discontinuous multivariate data. J Multivar Anal 117:214–228
Goerlich FJ (2020) La encuesta de condiciones de vida: evaluación de los cambios metodológicos en relación a la obtención de los ingresos. Hacienda Pública Española 233:85–116
Guyton J, Langetieg P, Reck D, Risch M, Zucman G (2021) Tax evasion at the top of the income distribution: theory and evidence. NBER Working Paper 28542
Higgins S, Lustig N, Vigorito A (2018) The rich underreport their income: assessing bias in inequality estimates and correction methods using linked survey and tax data. ECINEQ WP 2018—475
INE (2010) Oportunidades de aprovechamiento de registros administrativos en la ECV. Análisis realizado con la encuesta 2007. INE (mimeo)
Jenkins SP, Rios-Avila F (2021) Measurement error in earnings data: Replication of Meijer, Rohwedder, and Wansbeek's mixture model approach to combining survey and register data. J Appl Economet 36:474–483
Joe H (1990) Multivariate concordance. J Multivar Anal 35(1):12–30
Larrimore J (2014) Accounting for United States household income inequality trends: the changing importance of household structure and male and female labour earnings inequality. Rev Income Wealth 60:683–701

Larrimore J, Burkhauser RV, Auten G, Armour P (2021) Recent trends in US income distributions in tax record data using more comprehensive measures of income including real accrued capital gains. J Polit Econ 129:1319–1360

Lerman RI, Yitzhaki S (1985) Income inequality effects by income source: a new approach and applications to the United States. Rev Econ Stat 67(1):151–156

Lynn P, Jäckle A, Jenkins SP, Sala E (2012) The impact of questioning method on measurement error in panel survey measures of benefit receipt: evidence from a validation study. J R Stat Soc Ser A 175:289–308

Méndez JM, Vega P (2011) Linking data from administrative records and the Living Conditions Survey. INE Working Papers 01/2011

Meyer BD, Mittag N (2021) Combining administrative and survey data to improve income measurement. In: Chun AY, Larson M, Reiter J, Durrant G (eds) Administrative records for survey methodology. Wiley, New York

Meyer BD, Wu D (2018) The poverty reduction of social security and means-tested transfers. NBER Working Paper No. 24567

Meyer BD, Mok W, Sullivan J (2015) Household surveys in crisis. J Econ Perspect 29:199–226

Milanovic B (2017) Increasing capital income share and its effect on personal income inequality. In: Boushey H, DeLong JB, Steinbaum M (eds) After Piketty. The agenda for economics and inequality. Harvard University Press, Cambridge, MA

Nelsen RB (1996) Nonparametric measures of multivariate association. In: Rüschendorf L, Schweizer B, Taylor MD (eds) Distributions with given marginals and related topics, 28: 223–232

Nelsen RB (2002) Concordance and copulas: a survey. In: Cuadras CM, Fortiana J, Rodriguez-Lallena JA (eds) Distributions with given marginals and statistical modelling, Springer, Dordrecht, pp 169–177

Oberski DL, Kirchner A, Eckman S, Kreuter F (2017) Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. J Am Stat Assoc 112:1477–1489

Pérez A, Prieto-Alaiz M (2016) A note on nonparametric estimation of copula-based multivariate extensions of Spearman's rho. Statist Probab Lett 112:41–50

Piketty T (2013): Le capital au XXIe siècle. Seuil, París

Ravallion M (2021) Missing top income recipients. NBER Working Paper 28890

Schmid F, Schmidt R, Blumentritt T, Gaißer S, Ruppert M (2010) Copula-based measures of multivariate association. In: Jaworski P, Durante F, Härdle WK, Rychlik T (eds) Copula theory and its applications, pp 209–236, Berlin. Springer

Shorrocks AB (1982) Inequality decomposition by factor components. Econometrica 50:193–212

Slemrod J (1995) Income creation or income shifting? behavioral responses to the tax reform act of 1986'. Am Econ Rev Pap Proc 85(1995):175–180

Wolff EF (1980) N-dimensional measures of dependence. Stochastica 4(3):175–188

Yonzan N, Milanovic B, Morelli S, Gornick J (2020) Drawing a line: comparing the estimation of top incomes between tax data and household survey data. Stone Center Working Paper 27, City University of New York