IARIW 2021
Monday 23 – Friday 27 August
Statistisk sentralbyrå
Statistics Norway

Underreporting of Top Incomes and Inequality:

An Assessment of Correction Methods using Linked Survey and Tax Data

Emmanuel Flachaire

(Aix-Marseille University)

Nora Lustig

(Tulane University)

Andrea Vigorito

(Universidad de la República de Uruguay)

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
Session 20: Inequality I
Time: Thursday, August 26, 2021 [14:00-16:00 CEST]

Underreporting of Top Incomes and Inequality:
An Assessment of Correction Methods using Linked Survey and Tax Data

Emmanuel Flachaire*, Nora Lustig†, Andrea Vigorito‡

Abstract

Due to underreporting and other factors, household surveys do not capture incomes at the top of the distribution well. This affects inequality measures. Replacing and reweighting methods that combine survey data with tax records have been proposed to address this problem. In this paper we attempt to assess their performance using linked data. We use a novel database in which a subsample of Uruguay's official household survey has been linked to tax records to document the extent and distribution of income underreporting. We find that individuals in the upper half of the income distribution tend to report less income in household surveys than in tax returns, and underreporting is increasing in income. We simulate a true distribution and a distorted distribution that mimics the underreporting pattern found in the Uruguayan data. Inequality estimates based on the distorted distribution are biased. We apply the replacing and reweighting methods to correct the distorted distribution. If the threshold above which true data replaces distorted data is the optimal threshold, we find that the replacing and reweighting methods fully correct the biases of inequality measures. In practice, the optimal threshold, however, is unknown. We assess the sensitivity of the methods to the choice of the threshold. We find that the replacing method is less sensitive to threshold selection.

JEL Codes: D31, C81

Keywords: inequality; income underreporting; tax records; household surveys; linked data; correction methods; reweighting; replacing

February 2021

*Aix-Marseille Université, AMSE
†Tulane University
‡Universidad de la República de Uruguay

1 Introduction

Household surveys suffer from a series of errors.[1] While such errors can affect the entire survey, for inequality measures they are particularly important when they occur in the upper tail.[2] Household surveys do not capture incomes at the top of the distribution well because the rich may be harder to reach, leading to unit nonresponse, more likely to refuse to answer the survey when reached, resulting in item nonresponse, or may report a lower fraction of their income when responding to the survey (Atkinson 2007). In addition, in finite samples the upper tail is not captured well due to sparseness or because data producers truncate or topcode the distributions in the upper tail (Cowell and Flachaire 2007; Biemer and Christ 2008; 2015). These issues can lead to significant bias in inequality measures, and this bias can be either positive or negative (Deaton 2005). Recognizing this, recent papers have made use of data from tax returns to correct survey-based inequality estimates (Burkhauser et al. 2016; Jenkins 2017; Piketty et al. 2019). Correction methods, however, rely on implicit assumptions that are often untestable.[3] In particular, they rely on an appropriate selection of the threshold beyond which survey data tend to underreport income.

Here we are particularly interested in addressing one type of measurement error: underreporting of income at the top.[4] We exploit a novel data set that directly links a subset of individuals from Uruguay's official household survey to the same individuals' tax returns, enabling us to observe income from each of these sources for the same person. However, the linked data is restricted to adults in households with children aged 0 to 3. Although this subsample captures households at the top of the income distribution, this is probably a biased sample if we were interested in measuring the distribution of income in Uruguay. But, the purpose of using this linked data is not to estimate inequality in Uruguay, at least not here. We use the linked data to observe an actual pattern of underreporting, an observation that is usually not possible as there are very few instances for which linked data exist.

In order to assess the implications of underreporting and alternative correction methods on inequality estimates, we resort to simulation. We first consider a true distribution and construct a distorted distribution that mimics the underreporting pattern observed in the linked data for Uruguay's subsample. This approach allows us to focus on underreporting and not consider sampling errors in the upper tail, a common problem featured by finite samples. We calculate the Gini coefficient, the mean log deviation (MLD), the Theil index, and top 10%, 5% and 1% shares for the true and distorted distributions and find that all the inequality indicators estimated with the latter are strongly biased.

With our simulated true and distorted distributions, we then proceed to assess the

[1] The total survey error is composed of the sum of three distinct elements: representation error, error due to item non-response, and measurement error (Groves and Lyberg 2010; Meyer and Mittag 2019).

[2] Poverty measures will be biased primarily because of errors in the lower tail.

[3] For a survey, see Lustig (2019).

[4] However, as we shall see below, when using external data such as tax records, we are also correcting for missing data problems (e.g., top coding or trimming; item or unit nonresponse in the upper tail; etc.).

two main correction methods: replacing (Burkhauser et al. 2016; Hlasny and Verme 2018; Jenkins 2017; Piketty et al. 2019; Chancel and Piketty 2019;) and reweighting (Anand and Segal 2017; Burkhauser et al. 2018; Campos-Vazquez and Lustig 2020; Department for Work and Pensions 2015; Hlasny and Verme 2018).[5] For this purpose, we generate a simulated hybrid distribution that combines data from the distorted distribution and the true distribution.

Replacing consists of replacing the top $k\%$ of the distorted distribution (e.g., the top 50%, 10%, 5% and so on) by the top $k\%$ of the true distribution. A hybrid distribution is then defined with the bottom (100 - k)% of the distorted distribution, and the top $k\%$ of the true distribution. Using different thresholds (including the optimal one), we calculate the inequality measures for the simulated hybrid distributions. We find that inequality measures are biased if the selected threshold is too high (compared to the optimal threshold) and as the selected threshold gets closer to the optimal one, the inequality measures approximate the true ones.

Reweighting consists of using the data from the distorted distribution below an income threshold t, and the data from the true distribution above t. A hybrid distribution is then defined with the bottom (100 - k)% of the survey distribution, and the top $l\%$ of the tax distribution. The reweighting scheme ensures that the upper tail of the hybrid distribution is similar to that of the true distribution. We produce the simulated reweighted hybrid distributions and calculate the inequality measures under different thresholds, including the optimal one. Just as with replacing, we find that inequality measures are biased if the selected threshold is too high (compared to the optimal threshold) and as the threshold gets closer to the optimal threshold, the inequality measures approximate the true ones.

Our analysis shows that threshold selection plays a key role.[6] If the threshold is not correctly chosen, inequality measures may be significantly biased. An important finding is that the replacing method is less sensitive to the choice of the threshold. In other words, the replacing method yields inequality measures that are closer to the true inequality measures for a broader set of thresholds than reweighting.[7] This is because with replacing, the error introduced with corrections is confined to a smaller segment of the distribution. In fact, reweighting affects the entire distribution below the threshold and, thus –unless one was lucky and chose the true threshold–, reweighting may introduce biases in the absolute poverty estimates and in inequality measures that are sensitive to the bottom of the distribution. If one knows the optimal threshold, replacing and reweighting are equivalent and applying either would yield the true distribution. However, the challenge is precisely that the optimal threshold in practice remains unknown.

In addition to the simulations, we explore how to approach the threshold selection challenge in practice using the linked data for Uruguay. We find that underreporting in the survey starts at the median of the tax records income distribution, near the min-

[5]Some of these studies rely on within-survey correction only. For example, Hlasny and Verme (2018).

[6]See Cowell and Flachaire (2015) for a discussion of the challenges around threshold selection.

[7]It is worth pointing out that observing that inequality measures converge to a stable value cannot be interpreted that one has found the threshold which is closer to the optimal. That is, while convergence is a necessary condition to approximating the optimal threshold it is not a sufficient one.

imum taxable income. Thus, it seems natural to choose the median as the threshold above which tax data replaces survey data. However, since administrative information is not exempt from errors, tax records cannot be assumed to be equivalent to the true distribution either.[8] Thus, we examine the sensitivity of inequality measures to correction methods under a range of thresholds. Our results are analogous to our simulation exercise. In other words, with replacing, the inequality measures are less sensitive to the threshold selection.

We make two main contributions. First, we document the extent and distribution of income underreporting using actual linked data for Uruguay. We find that individuals in the upper half of the income distribution (above the minimum taxable income) tend to report less income in household surveys than those same individuals earn according to tax returns, and underreporting is increasing in income. Second, we assess the performance of methods to correct inequality estimates. We find that underreporting leads to biased inequality estimates. We find that as the threshold (above which data in the distorted distribution is replaced by data in the true distribution) approximates the optimal threshold (that is, the one below which there is no underreporting), both replacing and reweighting methods yield inequality measures that get closer to the true ones.[9] The methods will correct well if the threshold is "low enough:" that is, closer to the optimal threshold. However, replacing approaches the true inequality measures more quickly. As indicated above, our analysis shows that threshold selection plays a key role. Both in our simulated distributions and Uruguayan data, we find that the threshold above which data from the survey or distorted distribution should be replaced by data from taxes or true distribution in our simulations, is lower than the threshold usually chosen in empirical studies.[10] In practice, the challenge is that even with linked data we do not know the optimal threshold as the true distribution is unknown. Hence, the recommendation is to test the sensitivity of results to a range of thresholds.

This article is organized as follows. We first describe the databases used in this study and show the misreporting patterns identified in the linked data (Section 2). We then present the correction methods and provide simulations results (Section 3). Based on these findings, Section 4 discusses what to do in practice and presents corrected inequality measures estimated with the linked data for Uruguay to illustrate. Section 5 includes some final remarks. Additional information can be found in the Appendix.

[8]For example, wrong identification numbers, different income concepts and reporting periods, "off the book" payments, exclusion of informal workers, incomes for the same individual may come from formal and informal employment, tax avoidance and evasion, etc. In fact, the survey earnings validation literature concludes the definition of a true distribution largely depends on priors chosen by researchers, which lead to different measurement error estimates (Kapteyn and Ypma 2007; Abowd and Stinson 2013; Jenkins and Rios-Avila 2020). See also Gottschalk and Huynh (2010); Hyslop and Townsend (2020); Adriaans et al. (2020).

[9]As discussed in Section 4.1, this conclusion is still valid with any other problems in survey top incomes, as undercoverage, nonresponse, top coding, etc.

[10]The replacement method is usually applied to the top 10, 5 or 1% of the survey data.

2 Misreporting evidence from linked data

2.1 Data

We use a novel database in which a subsample of Uruguay's official household survey—*Encuesta Continua de Hogares* (ECH)—has been linked to personal income tax records from the *Dirección General Impositiva* (DGI).

ECH collects post-tax information on labor income concepts and social security coverage (formality) for each worker separately considering: i) self-employment earnings; ii) main salaried occupation; iii) remaining salaried occupations. Based on this information we construct post-tax formal labor income defined as close as possible to the income measure from tax records by adding salaries and wages, commissions, incentives, vacation pay and overtime payments. In our analysis, we exclude tips, arrears, transport, food or housing vouchers, other in-kind payments, other fringe benefits and bonuses from formal occupations, but income misreporting patterns remain unchanged if these income concepts are considered.[11]

DGI databases used in this study include the universe of potential personal income tax payers, including all formal workers, pensioners, self-employed (liberal) professionals and capital income receivers for 2012-2013 (Burdín et al. 2014). As a whole, these data cover approximately 75% of the population aged 20 or more. Like all tax records, these data are subject to tax evasion and avoidance (Atkinson, 2007).[12] Personal income taxation (*Impuesto a las Retribuciones de las Personas Físicas*, IRPF) in Uruguay is based on a dual scheme that combines a progressive tax schedule on labor income and pensions, with a flat tax rate on capital income. The tax unit is the individual, but married couples have the chance of filing a joint tax return in the case of labor income. However, only 1.8% of the individuals in the tax records chose this regime. In most cases, the personal income tax is withheld, reported and paid by employers, firms, banks and other agents. Only individuals with more than one occupation or those who receive more than one income source and self-employed file taxes. More information on the tax scheme can be found in Appendix 1.

A subsample of individuals in the ECH 2012 and 2013 was linked to tax records. The subsample of linked individuals are those included in a follow-up survey: the Nutrition, Child Development, and Health Survey (*Encuesta de Nutrición, Desarrollo Infantil, y Salud*, ENDIS). ENDIS is a longitudinal study that follows 2,649 urban households with children aged 0 to 3 that were interviewed in ECH between February 2012 and August 2013 (Instituto Nacional de Estadística 2013; Instituto Nacional de Estadística 2018). The potentially linked individuals include all adult members (2018 18 years of age or more) of a household.[13]

[11]Results are available upon request to the authors.

[12]In the 2000s, Uruguay experienced rapid economic growth, coupled with a substantial decrease in informal employment: from 40% in 2004 to 23.5% in 2014 (Carrasco, Cichevski, and Perazzo 2018). Thus, formal workers represent the majority of total workers.

[13]By the time of writing this article, there have been three waves of ENDIS. In the first wave (started on November 2013), enumerators collected the unique national identification number (*cedula*) of each

With the linked data, we can compare tax-return incomes with survey incomes for the same individuals. In order to compare incomes reported in ECH with DGI, for each individual, we then create harmonized post-tax total income variables with ECH and DGI data, by adding formal labor income, pensions and taxable capital income (rents, dividends, entrepreneurial profits and interests from bank deposits and other financial assets). As in ECH the reference period for the collection of data on labor earnings and pensions is the previous month, in DGI we consider information corresponding to the month prior to the ECH interview. Capital income is collected in the two databases on an annual basis, and, thus, we include a monthly average.

Of the original ENDIS households, 4,539 adults were income receivers and 2,360 were formal workers (whose incomes are positive or zero in the period of reference) or received income from capital or pensions. Of these, 2,287 had valid ID cards and were linked to tax data. Among linked observations, 1,634 (71%) had positive earnings in tax records in the previous month of the ECH interview. Of the 1,634, a total of 1,471 had positive income in ECH and 163 (10%)[14] did not report their income but have positive income in DGI records.[15] Our exercise uses the 1,471 linked individuals who reported positive income in both sources. Among linked observations, average income is 17% larger in tax records than in survey reports and, perhaps unsurprisingly, maximum income is higher in the tax data (Table 1).[16]

As mentioned in the introduction, our analysis is restricted to couples with children aged 0 to 3 living in urban areas. In order to assess how this restriction might affect the observed pattern of underreporting, we check the distribution of the linked observations in the ECH and DGI. As shown in Figure 1, the linked observations are present along the whole ECH distribution and the full DGI distribution, and they even exhibit a larger share in the upper strata. That is, misreporting does not appear to be pervasive just in one part of the distribution. This means that our subsample can be used as an adequate approximation of the patterns of misreporting that could potential be observed in the full ECH.[17].

---

respondent (principal caregivers of reference children, mainly mothers), whereas in the second wave (2015) this information was also gathered for fathers and other adult household members, allowing INE and DGI to merge all adults (and not just mothers) from 2012-13 ECH that were in ENDIS to DGI tax records. (*Cedulas* are composed of 7 digits plus a verification number. As INE gathered the verification number, it ensured that the numbers provided were correct, minimizing potential linkage errors. We did not have access to the actual card numbers, but to masked identifiers.

[14]In line with the validation literature (Bollinger et al. 2019), this figure can be considered as a measure of item non-response in the survey. For the purposes of our article, this is not an issue that needs to be addressed here. Figure A.2.1 depicts the proportion of item non-response individuals by percentile of DGI tax records, which heavily accumulate in the lower tail of the income distribution, clearly rejecting the hypothesis that non-response is missing at random. This pattern is different to the two tails one identified by Bollinger et al. (2019) for the United States.

[15]Among the 2,749 survey respondents, 1,720 declared labor force participation (1,507 employed and 213 unemployed). The remainder were housekeepers/homemakers (808), full-time students (194), pensioners or rentiers. 1079 were formal workers, rentiers or received pensions and 1027 were merged to the DGI database (95%).

[16]Restricting the comparison to labor income yields similar results in regard to average income, although the DGI maximum is 25% higher than the ECH one.

[17]For a comparison of the characteristics of linked observations and the rest of the individuals in

2.2 Misreporting patterns

To analyze the measurement error in our linked data, we examine the subsample 1,471 individuals who have positive income in both the ECH and DGI during the period of reference. Figure 2 plots the ratio of income reported in tax data to income reported in the survey for each observation in the linked data, and shows how this varies across the tax return distribution (i.e., by tax return income percentile). A local linear regression is estimated with a bandwidth obtained by cross-validation and with bootstrap standard errors. [18] It is clear from this Figure that individuals tend to underreport their income in the survey, above the minimum taxable income. If everyone reported the same income in the two data sources, all points would lie along the y = 1 line. If incomes were reported with noise but income misreporting were orthogonal to income, the points would bounce around, but the average relationship would correspond to y = 1.[19]

It can be noticed that ECH incomes exceed tax return incomes in approximately the bottom half of the tax return distribution, while survey incomes are lower in the top half. Interestingly, this point corresponds to the minimum taxable income threshold for labor earnings, which represent the larger income component both in this subsample and in the full tax records distribution.[20] In fact, the top 1% of the tax return distribution reports only about 60% of the income from their tax returns in the household survey. It is worth noting what happens in the low percentiles, for which the ratio can take on very small values. The confidence bands suggest that the ratio is significantly smaller than one below the minimum wage, for which tax data are known to be unreliable (Atkinson 2007; Burkhauser et al. 2016; Piketty et al. 2019). To some extent, we can question the reliability of tax data under the value of the minimum taxable income, since income from informal employment is more prevalent below the median and, even if income is from the formal sector, misreporting incomes that even if accurately reported are below the minimum taxable income has little consequence.[21]

ENDIS see Table A.2.1.

[18] with the `npreg` function in `R`.

[19] Figure A.2.2 shows the empirical copula (i.e., bivariate density) of percentiles in the survey and tax return income distributions. If the correlation between every individual's rank in the tax return income distribution and her rank in the survey income distribution were the same (which can occur regardless of the extent of misreporting), the copula's density would lie along the gray 45- degree line. At higher incomes, the correlation is stronger: among the top 20% of the income distribution, we see a higher density of observations concentrated near the 45-degree line, although those in the highest survey income percentiles tend to be found at slightly lower tax-return income percentiles due to misreporting.

[20] In this case, the survey reporting pattern we obtain (overreporting in the lower tail and underreporting at the top) is in line with previous findings from the survey earnings validation literature (Adriaans et al. 2020). In addition to social norms (a factor identified in the literature), overreporting at the bottom in the context Uruguay (as well as more generally in low and middle-income countries) is likely the result of the coexistence of income coming from both formal and informal employment. It is worth pointing out that the proportion of income coming from informal occupations is between 0% and 15% among linked cases (Figure A.2.3) and high income individuals also report informal income.

[21] To check whether the misreporting pattern we identify in Figure 2 holds for different population groups, we computed the misreporting ratios for different income variables and population groups. We first restricted merged cases to harmonized labor income only, leaving aside the remaining income sources. Second, we consider full-time workers only, assuming that their income is more stable, and they are less likely to misreport. Third, we consider only survey respondents, assuming that they

3 Correction methods

In this section, we consider a hypothetical *true* distribution, $f_Y(y)$, and misreporting with known shape. We can then derive the corresponding *distorted* distribution, $f_X(x)$, which suffers from average underreporting in high incomes, and study the impact of misreporting on standard correction methods.

3.1 Simulation design

We consider the Singh-Maddala distribution, $SM(2.257, 17393, 1.033)$, as the true distribution.[22] These parameters are obtained by estimating a Singh-Maddala distribution from the Uruguayan linked data, combining survey data below and tax data above the median of the tax distribution. To mimic underreporting obtained from the Uruguayan linked data, we assume that underreporting is designed such that, on average, it increases above the median, as a piecewise linear model:

$$r(p) = \begin{cases} 1 & \text{if } p \leq 0.5 \\ 0.25 + 1.5p & \text{if } 0.5 < p \leq 0.9 \\ -7.85 + 10.5p & \text{if } p > 0.9 \end{cases} \tag{1}$$

where p is the proportion of income smaller than y in the true distribution, $p = F_Y(y)$, and $F_Y(y)$ is the CDF of the true distribution.[23] Under this design, on average, there is no underreporting below the median, underreporting increases slowly above the median, until the $90th$-quantile above which it increases sharply. More generally, underreporting can be defined with a function $r(p)$ such that $r(p) \geq 1$. Thus, we can obtain misreported incomes from the following relationship:

$$y = x\, r(p)\, \varepsilon, \qquad \text{where } \varepsilon \sim N(1, \sigma^2) \tag{2}$$

A misreported income x is then obtained by dividing a true income y by a misreporting factor $r(p)\,\varepsilon$, which is on average equal to $r(p)$. The parameter σ measures the heterogeneity of misreporting rates of individuals with the same tax income. When $\sigma = 0$, individuals with the same tax income misreport exactly the same amount. We use $\sigma = 0.15$ to introduce some heterogeneity. Moreover, we restrict ε to be strictly positive, to ensure positive income $x > 0$. Simulation experiments with different underreporting shapes have been investigated, they provide similar results (not reported).

present a higher probability of providing accurate responses (although in our regression analysis the proxy-respondent variable was not statistically significant). Finally, we consider total income reported by each merged observation in the survey, including informal income from different occupations. As it can be checked in Figure A.2.4, results present slight variations, but are basically similar in the five cases.

[22] The Singh-Maddala density function is equal to $f(y) = aqy^{a-1}/\{b^a[1+(y/b)^a]^{1+q}\}^{-1}$, where a and q are shape parameters, b is a scale parameter, and $y > 0$.

[23] The constant terms 0.5 and -7.6 are defined to have a continuous function at the knots 0.5 and 0.9

Figure 3 shows the average misreporting rates, $r(p)$, and 1000 true misreporting rates generated from the process described above, $y/x = r(p)\varepsilon$. We can see that it mimics the underreporting shape obtained in high incomes from linked data (see Figure 2).[24]

Figure 4 shows the density function of the true distribution (in logs) and a kernel density estimation of the distorted distribution (in logs), obtained from a sample of one million observations generated with the process described above. We can see that the distorted distribution, which suffers from average underreporting in high values, deviates significantly from the true distribution in the upper part of the distribution.

Table 2, rows 1 and 2 (`true`, `distorted`), show several inequality indices computed from the true and distorted distributions. We can see that inequality is always smaller in the distorted distribution. In other words, inequality is downward biased when underreporting occurs in high incomes, due to the fact that the distorted distribution differs from the true distribution, that is, due to *non-sampling* errors.

Overall, we need external information to correct the problem of misreporting. In the following, we study the impact of underreporting in high incomes and the use of several correction methods, when external reliable information is available in the upper part of the true distribution. In practice, the survey distribution is often considered as suffering of underreporting in high incomes, and the tax distribution is often considered as a more reliable external information in top incomes. The opposite is often considered in low incomes, with survey data being more reliable than tax data (when available). Correction methods are then used to combine these two distributions.

3.2 Replacing

A first correction method consists to replace the top $k\%$ of the distorted distribution by the top $k\%$ of the true distribution. A hybrid distribution is then defined with the bottom $(100-k)\%$ of the distorted distribution, and the top $k\%$ of the true distribution. Let us consider a cumulative distribution function (CDF), F_X, which is continuous and strictly monotonically increasing, the quantile function is the inverse function of the CDF:

$$Q_X(p) = F^{-1}(p) \tag{3}$$

The *replacing* distribution is then defined as,

$$f_r(x) = \begin{cases} f_X(x) & \text{when } x \leq s \\ 0 & \text{when } s < x \leq t \\ f_Y(x) & \text{when } x > t \end{cases} \tag{4}$$

where s and t are the $(100-k)th$-quantile of the distorted and true distributions, respectively:

$$s = Q_X(1-k/100) \qquad \text{and} \qquad t = Q_Y(1-k/100) \tag{5}$$

[24]Since we do not have the analytical formula for the distorted distribution, we use a huge sample (1 million observations) to approximate it.

This distribution can also be obtained by multiplying misreported incomes x by quantile ratios, as follows:

$$z = \begin{cases} x & \text{when } x \leq s \\ x\, Q_Y(p)/Q_X(p) & \text{when } x > s \end{cases}, \qquad \text{with} \quad p = F_X(x) \tag{6}$$

This rescaling procedure remains to replace misreported incomes above s by their corresponding quantiles in the true distribution, since $Q_X(p) = x$. The density function of z is then $f_r(x)$ defined in (4).

Theoretically, replacing the top $k\%$ of the distorted distribution by the top $k\%$ of the true distribution is equivalent to scaling up misreported data by quantile ratios (6). It is true when we consider population distributions. But in practice, the results can differ significantly (see discussion in section 4.1).

With microdata, we would combine the $(100-k)\%$ lowest misreported data with the $k\%$ highest true data. If the number of observations in the $k\%$ highest misreported and true data are not the same, we have to reweight to guarantee that the selected true data represent $k\%$ of the combined sample.[25] Thus, we would put weights equal to (n_z/n_x) to the selected misreported data, and equal to $(n_z/n_x)(k_x/k_y)$ to the selected true data, where n_x and n_z are the number of misreported and combined data, respectively, and k_x and k_y are the number of observations in the $k\%$ highest misreported and true data, respectively.

From (4), we can see that, when there is no underreporting below the threshold s, we have $f_X(x) = f_Y(x)$ when $x \leq s$ and $s = t$, so the replacing distribution is the true distribution. However, when underreporting occurs below s, the replacing distribution deviates from the true distribution. Specifically, the $(100-k)th$-quantile of the distorted and true distributions may differ, $s \neq t$, and the density is equal to zero between these two values.

Finally, when the top $k\%$ of distorted distribution is replaced by the top $k\%$ of true distribution, additively decomposable inequality measures can be easily estimated from a non-overlapping decomposition, using a breakdown such as this:

$$\begin{aligned} \text{Total inequality} = {} & \text{inequality of the smallest } (100-k)\% \text{ misreported data} \\ & + \text{inequality of the highest } k\% \text{ true data} \\ & + \text{between group inequality} \end{aligned} \tag{7}$$

Decomposition formulas for the Gini and other inequality measures can be founded in Alvaredo (2011) and Cowell (2011). Moreover, top $v\%$ shares above t are defined as follows:

$$\text{TS}_r(v) = \frac{(v/100)\mathbb{E}(y \geq Q_Y(1-v/100))}{\mu_r} = \frac{\mu_Y}{\mu_r}\,\text{TS}_Y(v) \qquad \text{if } v \leq k \tag{8}$$

where μ_r is the mean of the replacing distribution. When $v \leq k$, top shares of the hybrid distribution are then equal to top shares of the true distribution, rescaled by

[25] This reweighting procedure should not be confused with the reweighting method described in 3.3 which is designed to correct for underreporting or missing people.

the mean ratio. Thus, when the mean of the replacing distribution is smaller (larger) than the mean of the true distribution, top $v\%$ shares with $v \leq k$ are biased upwards (downwards).

Figure 5 (a) shows the replacing distribution (in logs), when we replace the top 10% of the distorted distribution by the top 10% of the true distribution. The threshold t is then equal to the $90th$-quantile of the true distribution. A histogram is obtained from a sample of one million observations. We can see that the replacing distribution is similar to the distorted distribution in the bottom, and to the tax distribution in the upper part, but there is no value between the two $90th$-quantiles of distorted and true distributions.

Figures 5 (b), (c), (d), (e) show replacing distributions (in logs), when we replace the top 32.4%, 50%, 60% and 70% of the distribution. The threshold t is then equal to, respectively, the $67.6th$- $50th$- $40th$- and $30th$-quantiles of the true distribution. We can see that the replacing distribution gets closer to the true distribution, as the threshold decreases. Let us define the optimal threshold as the value above which the true and distorted distribution differ, which is around the $30th$-quantile (see Figure 4). An interesting feature of this correction method is that the replacing distribution deviates from the true distribution locally, that is, between the optimal threshold and the selected threshold only.

Table 2, (`replacing`), shows inequality measures obtained from the replacing method with several thresholds. We can see that the inequality measures are much closer to the true values than those obtained from the distorted distribution. Nevertheless, substantial differences remain with the $90th$- and $67.6th$-quantile thresholds, and the top 10%, 5% and 1% shares are overestimated. On the other hand, the results are very good with the $50th$-, $40th$-, $30th$- and $25th$-quantile thresholds.

This correction method is often used in empirical studies. Among others, see Burkhauser et al. (2016); Jenkins (2017); Hlasny and Verme (2018); Piketty et al. (2019); Chancel and Piketty (2019).

3.3 Reweighting

A second correction method consists to use misreported data below a threshold t, and true data above t. A hybrid distribution is then defined with the bottom $(100 - k)\%$ of the distorted distribution, and the top $l\%$ of the true distribution, where:

$$t = Q_X(1 - k/100) = Q_Y(1 - l/100) \tag{9}$$

When $k \neq l$, there is an implicit reweighting: misreported and true data correspond, respectively, to the bottom $100 - m\%$ and to the top $m\%$ of the hybrid distribution, where $m = 100l/(100 - k + l)$. In order to keep a distribution above t similar to the true distribution, it is required to reweight misreported data below t and true data above t

with the following weights:

$$w(z) = \begin{cases} (100 - l)/(100 - m) & \text{if } z \leq t \\ l/m & \text{if } z > t \end{cases} \tag{10}$$

The role of the weights is to increase the density above t, such that the top $l\%$ of the reweighted distribution corresponds to the top $l\%$ of the true distribution, and to decrease the density below t to compensate. This reweighting scheme ensures that the upper tail of the hybrid distribution is similar to that of the true distribution.

The *reweighting* distribution is then given by:

$$f_w(x) = \begin{cases} \lambda f_X(x) & \text{if } x \leq t \\ f_Y(x) & \text{if } x > t \end{cases} \tag{11}$$

where $\lambda = (100 - l)/(100 - m)$. The density of the reweighting distribution is equal to the density of the true distribution above t, and to the density of the distorted distribution rescaled by a factor λ below t. This distribution can also be obtained by using misreported incomes with the following weights:

$$w'(x) = \begin{cases} \lambda & \text{if } x \leq t \\ f_Y(x)/f_X(x) & \text{if } x > t \end{cases} \tag{12}$$

Indeed, the density function is given by $w'(x)f_X(x)$, which is equal to (11)

Theoretically, replacing misreported data above t by true data and using the weights defined in (10) is equivalent to using misreported data with the weights defined in (12). It is true when we consider population distributions. But in finite sample, density ratios are difficult to accurately estimate in high incomes, where densities are close to zero, and weights in (12) may then be unreliable. In practice, the results can differ significantly between the two approaches (see discussion in section 4.1).

Blanchet, Flores, and Morgan (2019), denoted BFM hereafter, proposed a correction method which is used in the *World Inequality Database*, see Alvaredo et al. (2020). This method is based on a first step, where misreported (survey) incomes are used with the weights defined in (12). To correct, among other things, problems induced by the poor estimation of density ratios, a second step is performed where misreported (survey) observations above the threshold are duplicated several times and replaced by observations with equivalent rank and weight in the true (tax) distribution. At the end, numerical results are quite similar to those obtained by replacing misreported (survey) data above t by true (tax) data and using the weights defined in (10).[26]

From (9) and (10), we can see that, when there is no underreporting below t, we have $k = l = m$, $w(x) = 1$, and the reweighting distribution is the true distribution. However, when underreporting occurs below the threshold t, we have $k \neq l$, and the reweighting distribution deviates from the true distribution below t.

[26]There is an issue concerning the threshold selection embedded in the BFM method. See section 3.4 on this.

Finally, when misreported data are used below a threshold t, and true data are used above t, additively decomposable inequality measures can be easily estimated from a non-overlapping weighted decomposition, using a breakdown such as this:

$$\begin{aligned}\text{Total inequality} = &\text{ inequality of the misreported data below } t \\ &+ \text{inequality of the true data above } t \\ &+ \text{between group inequality}\end{aligned} \tag{13}$$

with the weights defined in (10). Since the weights are constant in each group, weighted decomposition formulas for inequality measures with the property of scale independence are similar to unweighted decomposition formulas, where the share of the misreported data below t is equal to $1 - l/100$, the share of the true data above t is equal to $l/100$, and the overall mean μ_w is the weighted mean of the two groups:

$$\mu_w = (1 - l/100)\,\mathbb{E}(x|x \leq t) + (l/100)\,\mathbb{E}(y|y > t) \tag{14}$$

Moreover, top $v\%$ shares above the threshod t in (9) are defined as follows:

$$\mathrm{TS}_w(v) = \frac{(v/100)\mathbb{E}(y \geq Q_Y(1 - s/100))}{\mu_w} = \frac{\mu_Y}{\mu_w}\,\mathrm{TS}_Y(v) \qquad \text{if } v \leq l \tag{15}$$

They are equal to top shares obtained from true data, rescaled by the mean ratio. Thus, when the mean of the reweighting distribution is smaller (higher) than the mean of the true distribution, top $v\%$ shares with $v \leq l$ are biased upwards (downwards).

Figure 6 (a) shows the reweighting distribution (in logs), combining misreported data below t and true data above t, when the threshold is the $90th$-quantile of the true distribution. A histogram is obtained from a sample of one million observations. We can see that the reweighting distribution is similar to the true distribution above the threshold, and it is similar to the distorted distribution pushed downwards below the threshold.

Figures 6 (b), (c), (d) (e) show reweighting distributions (in logs), when we replace misreported data with true data above, respectively, the $67.6th$-, $50th$-, $40th$- and $30th$-quantile of the true distribution. We can see that the reweighting distribution gets closer to the true distribution, as the threshold decreases. A specific feature of this correction method is that the reweighting distribution deviates from the true distribution everywhere below the selected threshold.

Table 2, (`reweighting`), shows inequality measures obtained from the reweighting method with several thresholds. We can see that the inequality measures are much closer to the true values than those obtained from the distorted distribution. Nevertheless, substantial differences remain with the $90th$-, $67.6th$- and $50th$-quantile thresholds. On the other hand, the results are very good with the $40th$-, $30th$- and $25th$-quantile thresholds. Table 2, (`BFM`), shows inequality measures obtained with the BFM method. The results are identical to the reweighting method, which combines survey data below the threshold and tax data above with the weights defined in (10).

Compared to the replacing method, which returns values of inequality measures close to the true values more quickly, as the threshold decreases, the reweighting method

requires lower threshold to obtain very good results. This comes from the fact that the replacing distribution deviates from the true distribution locally (between the optimal threshold and the selected threshold only), while the reweighting distribution deviates from the true distribution more globally (everywhere below the selected threshold). Thus, when the selected threshold is not very far above the optimal threshold, the replacing distribution deviates from the true distribution in a quite narrow interval.

The reweighting method has been use in Anand and Segal (2017); Burkhauser et al. (2018); Hlasny and Verme (2018); Campos-Vazquez and Lustig (2020); Department for Work and Pensions (2015).

3.4 Threshold selection

When there is no average underreporting under the selected threshold, the replacing and reweighting methods are similar. They are based on a distribution which is the true distribution, and thus they correct the problem of misreporting well. However, when underreporting occurs below the threshold, the previous subsections show that the results may be biased. The choice of the threshold is then a key issue.

From Figure 4, we can see that the true and distorted distributions begin to deviate around the $30th$-quantile. It suggests that, in our simulation design, the optimal threshold is around $30th$-quantile. It is below the median, which may be surprising since there is no *average* misreporting below the median, see (1). This is due to the heterogeneity of misreporting behaviours, defined by $\sigma > 0$ in (1). Indeed, above the median, some individuals overreport their income and their income is then replaced by a lower income, which may be smaller than the median. The distribution below the median is then affected by *individual* misreporting. It follows that the optimal threshold may be smaller than the value above which average misreporting rates increase.

In practice, the threshold can be selected a priori, as the $80th$, $90th$, $95th$ or $99th$ quantile of the distribution (Burkhauser et al. 2016; Piketty et al. 2019; Chancel and Piketty 2019). A less arbitrarily method consists to select the threshold based on the quantile ratio function:

$$\text{select} \quad t = \max(Q_X(p)) \quad \text{such that} \quad \frac{Q_Y(p)}{Q_X(p)} = 1 \tag{16}$$

As long as the true and distorted distributions are identical in the bottom (below the optimal threshold), they share similar quantiles. This method is then designed to detect above which value the two distributions differ, when underreporting occurs above a threshold.

To illustrate, Figure 3 shows quantile ratios in our simulation design. We can see that the quantile ratios begin to deviate from 1 below the median. From this Figure, we would select a threshold at around the $40th$-quantile of the tax distribution. The replacing and reweighting methods provide inequality measures very close to the true values with this threshold (see Table 2).

Another threshold selection, proposed by Blanchet et al. (2019), is as follows:

$$\text{select} \quad t = \max(z) \quad \text{such that} \quad \frac{F_X(z)}{F_Y(z)} = \frac{f_X(z)}{f_Y(z)} \tag{17}$$

This method is defined to ensure the continuity of the reweighting distribution in the upper tail. It is not specifically designed to identify when the true and distorted distribution start to differ and, in general, it will select a threshold that is too high. Indeed, the selected threshold is at the highest of possible crossing-points between both densities (or close to it). If the true and distorted distributions deviate above this crossing-point, they will also deviate by the same magnitude/area below this crossing-point, because by definition since they are two density functions the areas under the curve must always equal to one. Thus, by construction, this selected threshold will always miss the deviation below the highest crossing-point between the two densities.[27] In some cases, this may be a price we are willing to pay to have continuity.[28]

Figure 7 shows the threshold selection obtained with (17) on our simulated data. The density ratio function, $\theta(x) = f_X(x)/f_Y(x)$, and the CDF ratio function, $\Theta(x) = F_X(x)/F_Y(x)$, are plotted. We can see that the moving average θ and Θ are close to 1 until the $30th$-quantile, above which they start to deviate, which corresponds to the special case (16) and to the optimal threshold detected in Figure 4. However, the threshold selected with (17) is much higher, it is obtained when θ and Θ cross at the $67.6th$-quantile of the tax distribution. Figure 6 (b) shows the reweighting distribution obtained with this threshold. We can see that the distribution is continuous, but it deviates significantly from the true distribution everywhere below the selected threshold. Moreover, all correction methods applied with this threshold provide inequality measures substantially different from the true values (see Table 2).

4 What to do in practice?

In practice, tax data are often considered more (less) reliable than survey data in high (low) incomes. Correction methods are then used to combine these two distributions. In such case, what would be recommended to do in practice, based on the previous results? And what would the results with the Uruguayan linked data look like if we apply the recommended strategy.

4.1 Lessons from the simulation results

First, we should stress that our simulation results are robust to many different underreporting designs. In particular, we find similar results when underreporting is mainly

[27] In fact, it assumes that this deviation is equally distributed over all the distorted (survey) distribution below the crossing-point.

[28] One can also obtain continuity if one selects the optimal threshold. However, the latter is unknown in general.

concentrated at the top of the distribution, that is, when on average underreporting increases above high thresholds, as the $90th$-, $95th$- or $99th$-quantiles.[29]

The simulation results are focused on misreporting and they are based on population distributions. However, misreporting is not the only reason to believe that surveys do not capture top incomes well. For instance, top-coding or censoring may be imposed by the data provider, because of reasons of confidentiality. Furthermore, some portion of the sampled population may not respond to the survey (item nonresponse), or may be difficult to reach easily (unit nonresponse). Finally, income distributions are often heavy-tailed and very high incomes tend to be sparse in finite samples. These additional reasons are all related to missing data, which is another major issue in surveys.[30]

Overall, if we replace unreliable top income survey data with reliable top income tax data, we can expect to correct both misreported *and* missing data problems in the upper tail. It is not possible if we only correct survey data, by rescaling or reweighting, because missing data cannot be recovered if we only use survey data. Thus, replacing the $k\%$ highest survey incomes by the $k\%$ highest tax incomes (replacing method in section 3.2), or replacing the survey incomes above t by tax incomes (reweighting method in section 3.3), should be preferred in practice, rather than rescaling survey incomes by quantiles ratios or reweighting survey incomes by density ratios.

Our results show that replacing the $k\%$ highest misreported (survey) data with the $k\%$ highest true (tax) data leads to a distribution that deviates from the true (tax) distribution locally, between the optimal and selected thresholds only, where it is discontinuous (section 3.2). Thus, this method is expected to provide good results when the selected threshold is not too far above the optimal threshold, or when summary measures of interest are not very sensitive to this part of the distribution. With the reweighting method, replacing survey data above the threshold t by tax data (section 3.3) leads to a distribution that deviates from the true (tax) distribution everywhere below the selected threshold, due to a reweighting scheme.

The threshold is more difficult to select if tax data are not reliable in low incomes, since quantile ratios may then differ from unity with thresholds lower than the optimal threshold. Thus, we cannot rely so much on the method based on quantile ratios in (16) and, in general, it is not easy to select the threshold in practice.[31] Without an a priori decision on above (below) which value tax (survey) data are reliable, the threshold selection based on quantile ratios in (16) may be used as a starting point in practice, and in addition present results with several thresholds to check robustness/sensitivity to the threshold. On this point, our results show that replacing the $k\%$ highest misreported (survey) data by the $k\%$ highest true (tax) data is less sensitive to the choice of the threshold than the other methods, due to its local deviation property.

[29]Table A.2.2 shows the results with underreporting increasing linearly above the $95th$-quantile, that is, when (1) is replaced by $r(p) = 20p - 18$ if $p > 0.95$, otherwise $r(p) = 1$, and $\sigma = 0.05$ in (2).

[30]Lustig (2019) uses *missing rich* as a catch-all term to refer to the causes of misreporting and missing data affecting the upper tail of survey distributions. These problems are also known as survey *undercoverage* of top incomes (Jenkins 2017, Burkhauser et al. 2018).

[31]This issue is analogous to the challenge of selecting the threshold to fit a Pareto parametric distribution in the upper tail. See Charpentier and Flachaire (2019) for a discussion.

Finally, unlike the previous section, which relies on population distributions, we also have to consider *sampling bias* in finite sample. In particular, it is well-known that the empirical distribution function, or EDF, does not cover well the upper tail of heavy-tailed distributions, due to sparse observations in the top tail. It is still true with tax data, but at a much higher level than with survey data. A Pareto distribution is often fitted in the upper tail of income and wealth distributions to reduce sampling bias errors. An interpolation method with a GPD distribution adjusted to the top can be used with tabulated data (Blanchet et al. 2017). The EDF with a Pareto distribution fitted to the top can be used with microdata (Charpentier and Flachaire 2019).

The price to pay to correct misreporting and missing data problems with these methods is that the covariates are lost, unless we make some strong assumptions. Indeed, when we multiply survey incomes by quantile ratios (replacing), or when we reweight survey incomes by density ratios (reweighting), we can keep covariates only if one of the following conditions holds: (1) misreporting preserves individual rankings in the income distribution ; or (2) individual rankings in the income distribution do not depend on the covariates. Indeed, with replacing, we cannot consider that the survey income and the corresponding corrected income belong to the same individual, because quantile ratios do not measure individual misreporting, except if (1) holds. So, we cannot transfer the covariates of an individual with a given survey income to the individual with the corresponding corrected income, except if (1) or (2) holds. With reweighting, if individuals are reranked at lower/higher levels in the survey, due to misreporting, we cannot use their covariates because their true positions in the income distribution are unknown, except if (1) or (2) holds.[32] Finally, as soon as misreporting implies individual rerankings in the income distribution, the link to individuals and therefore the link to covariates is lost, unless we use linked data or we make unrealistic assumptions such as (1) or (2). With our linked data, we find that there can be a significant extent of reranking when one goes from the survey to the tax distribution, the same individual switches ranks and sometimes by a lot. For inequality measures, since they are anonymous, the reranking does not affect results.

4.2 Application with Uruguayan linked data

In this section, we apply several correction methods with the Uruguayan linked survey and tax data. Tax data are considered more reliable than survey data in high incomes, but they are known to be unreliable below the minimum wage. These linked data have shown evidence of average underreporting in high (low) incomes of the survey (tax) data, see Figure 2. With underreporting of high incomes in survey data, inequality should be biased from the survey. With low income underreporting in tax returns, inequality should also be biased in tax data. Using correction methods, we seek to correct these biases by combining survey and tax data.

[32]To illustrate, let us consider an example which may sound extreme but helps drive the point home. If the richest individual in the true distribution underreports his income in the survey in such a way that he is ranked the poorest individual in the survey, it would make no sense to assign his covariates to the lowest income survey.

Figure 8 shows quantile ratios computed with the Uruguayan survey and tax data. We can see that the quantile ratios are smaller (greater) than 1 below (above) the $50th$-quantile. Thus, we would select a threshold around the median. This is consistent with Figure 2, which shows evidence of average underreporting above the median in the Uruguayan survey data. Moreover, the BFM threshold, which ensures continuity of the reweighting distribution in the upper tail, as defined in (17), is selected at the $72th$-quantile of the tax distribution, which as expected based on the previous discussion is too high.

Table 3 shows inequality measures computed with the Uruguayan data, from the survey and tax data, and from correction methods with several thresholds. The number of observations is equal to $n = 1461$. After merging the two datasets, a Pareto distribution is fitted in the top 5% of each distribution, with the replacing and reweighting methods.[33] The main results can be summarized as follows:

- Correction methods provide values that are larger than those from the survey and smaller than those from tax data. It suggests that the survey underestimate and the tax data overestimate inequality measures.
- Reweighting and BFM provide very similar results when one chooses the same threshold in both, but they are not exactly the same in finite samples. BFM is based on an interpolation method with a Generalized Pareto distribution adjusted in the extreme top, while our implemention of reweighting is based on a Pareto distribution fitted in the top 5%.
- Replacing, reweighting and BFM have the closest results with the $50th$-quantile threshold, which is the threshold selected with quantile ratios, and above which the linked data provide evidence of underreporting (see Figure 2).
- Replacing provides very stable results. For instance, numerical differences do not exceed 0.003 with thresholds at the $60th$- $50th$- and $40th$-quantiles, and they do not exceed 0.007 with thresholds at the $72th$-, $60th$-, $50th$-, $40th$ and $30th$-quantiles.

Overall, the empirical results are quite similar to our simulation results. The replacing method provides more stable results than the other methods, with different thresholds.

5 Conclusions

Household surveys suffer from sampling and non-sampling errors. Here we are particularly interested in addressing one type of measurement error: underreporting of incomes at the top. We use a novel database in which a subsample of Uruguay's official household survey has been linked to tax records to document the extent and distribution of income underreporting. We find that individuals in the upper half of the income distribution

[33]The number of observations is too small to fit a Pareto distribution in the top 1% or higher.

tend to report less income in household surveys than in tax returns, and underreporting is increasing in income.

We resort to simulation in order to assess the implications of alternative correction methods on inequality estimates. We consider a true distribution and construct a distorted distribution that mimics the underreporting pattern observed in the linked data for Uruguay's subsample. We apply the replacing and reweighting methods to correct the distorted distribution. The corrected distribution, as usual, is a hybrid of the true and the distorted distribution. Both methods entail substituting data from the true distribution for data in the distorted distribution above a certain threshold. If the optimal threshold – that is, the threshold below which there is no underreporting in the distorted distribution – is selected, we find that replacing and reweighting correct the biases of inequality measures in full. In fact, with the optimal threshold, replacing and reweighting methods are equivalent.

In practice, however, the selection of the threshold poses a significant challenge. Even with linked data we do not know the optimal threshold because – given the limitations of tax data – the true distribution continues to remain unknown. Faced with this challenge, the recommended course of action is to use the method that is less sensitive to the choice of threshold. By construction, with replacing, the error in the hybrid distribution is confined to the portion of the distribution between the optimal (unknown) threshold and the one that was selected. In contrast, with reweighting, the error is present in the hybrid distribution everywhere below the selected threshold. The replacing method is then expected to be less sensitive to the threshold. One should still test the sensitivity of inequality measures to alternative thresholds and pick the range for which inequality measures do not change much—i.e., are stable–.

References

Abowd, J. M. and M. H. Stinson (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics 95*(5), 1451–1467.

Adriaans, J., P. Valet, and S. Liebig (2020). Comparing administrative and survey data: Is information on education from administrative records of the german institute for employment research consistent with survey self-reports? *Quality & Quantity 54*(1), 3–25.

Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. *Economics Letters 110*, 274–277.

Alvaredo, F., A. Atkinson, T. Blanchet, L. Chancel, L. Bauluz, M. Fisher-Post, I. Flores, B. Garbinti, J. Goupille-Lebret, C. Martinez-Toledano, M. Morgan, T. Neef, T. Piketty, A.-S. Robilliard, E. Saez, L. Yang, and G. Zucman (2020). Distributional National Accounts Guidelines 2020: Methods and concepts used in the World Inequality Database. Wid.world working paper.

Anand, S. and P. Segal (2017). Who are the global top 1%? *World Development 95*, 111–126.

Atkinson, A. B. (2007). Measuring top incomes: methodological issues. Volume 1, pp. 18–42. Oxford University Press New York.

Biemer, P. P. and S. L. Christ (2008). Weighting survey data. In *International Handbook of Survey Methodology*, pp. 317–341.

Blanchet, T., I. Flores, and M. Morgan (2019). The weight of the rich: Improving surveys using tax data. Wid.world working paper 2018/12.

Blanchet, T., J. Fournier, and T. Piketty (2017). Generalized Pareto curves: Theory and applications. Wid.world working paper 2017/3.

Bollinger, C. R., B. T. Hirsch, C. M. Hokayem, and J. P. Ziliak (2019). Trouble in the tails? what we know about earnings nonresponse 30 years after lillard, smith, and welch. *Journal of Political Economy 127*(5), 2143–2185.

Burdín, G., F. Esponda, and A. Vigorito (2014). Inequality and top incomes in Uruguay: a comparison between household surveys and income tax micro-data. Commitment to Equity Working Paper 21.

Burkhauser, R. V., N. Hérault, S. Jenkins, and R. Wilkins (2016). What has been happening to UK income inequality since the mid-1990s? Answers from reconciled and combined household survey and tax return data. NBER working paper 21991.

Burkhauser, R. V., N. Hérault, S. Jenkins, and R. Wilkins (2018). Survey undercoverage of top incomes and estimation of inequality: What is the role of the UK's SPI adjustment? *Fiscal Studies 39*, 213–240.

Campos-Vazquez, R. and N. Lustig (2020). Labour income inequality in Mexico: Puzzles solved and unsolved. *Journal of Economic and Social Measurement 44*, 203–219.

Carrasco, P., A. Cichevski, and I. Perazzo (2018). Evolución reciente de las principales variables del mercado laboral uruguayo. *Serie Documentos de Trabajo; 9/18*.

Chancel, T. and T. Piketty (2019). Indian income inequality, 1922-2015: From British Raj to Billionaire Raj? *Review of Income and Wealth 65*, S33–S62.

Charpentier, A. and E. Flachaire (2019). Pareto models for top incomes. Working paper hal-02145024.

Cowell, F. A. (2011). *Measuring Inequality* (Third ed.). Oxford: Oxford University Press.

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*(2), 1044–1072.

Cowell, F. A. and E. Flachaire (2015). Statistical methods for distributional analysis. In *Handbook of income distribution*, Volume 2, pp. 359–465. Elsevier.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics 87*, 1–19.

Department for Work and Pensions (2015). Households Below Average Income An Analysis of the Income Distribution 1994/95–2013/14. London: Department for Work and Pensions.

Gottschalk, P. and M. Huynh (2010). Are earnings inequality and mobility overstated? the impact of nonclassical measurement error. *The Review of Economics and Statistics 92*(2), 302–315.

Groves, R. M. and L. Lyberg (2010). Total survey error: Past, present, and future. *Public opinion quarterly 74*(5), 849–879.

Hlasny, V. and P. Verme (2018). Top incomes and inequality measurement: A comparative analysis of correction methods using the EU SILC data. *Econometrics 6(2)*.

Hyslop, D. R. and W. Townsend (2020). Earnings dynamics and measurement error in matched survey and administrative data. *Journal of Business & Economic Statistics 38*(2), 457–469.

Instituto Nacional de Estadística (2013). Encuesta de nutrición, desarrollo infantil y salud, endis.

Instituto Nacional de Estadística (2018). Encuestas continuas de hogares.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica 84*, 261–289.

Jenkins, S. P. and F. Rios-Avila (2020). Modelling errors in survey and administrative data on employment earnings: Sensitivity to the fraction assumed to have error-free earnings. *Economics Letters*, 109253.

Kapteyn, A. and J. Y. Ypma (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics 25*(3), 513–551.

Lustig, N. (2019). The missing rich in household surveys: Causes and correction approaches. CEQ working paper 75.

Meyer, B. D. and N. Mittag (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics 11*(2), 176–204.

Piketty, T., L. Yang, and G. Zucman (2019). Capital accumulation, private property and rising inequality in China 1978-2015. *American Economic Review 109*, 2469–96.

Statistic	ECH income	DGI income	m (ECH-DGI)
Mean	19,363	22,585	-3,221
SD	16,4560	34,260	28,314
Min.	1,062	82	-919,153
Max.	191,185	979,153	143,488
Correlation coefficient			
ECH	1		
DGI	0.569	1	
m (ECH-DGI)	-0.106	-0.878	1

Table 1: Descriptive statistics. Harmonized monthly post-tax income. Linked cases

Note: Descriptive statistics for harmonized post tax income (in Uruguayan currency) were computed with ECH and DGI data for the subset of linked observations that had harmonized income different from zero in the two datasets in the survey reference period. 1 US Dollar=20.45 Uruguayan pesos.

Source: authors´ calculations based on ECH, ENDIS and DGI microdata.

t	Gini	MLD	Theil	Top10%	Top5%	Top1%
true	0.436	0.335	0.385	0.339	0.230	0.093
distorted	0.299	0.174	0.173	0.226	0.140	0.052
			Replacing			
$q90$	0.424	0.324	0.412	0.377	0.257	0.104
$q67.6$	0.441	0.342	0.397	0.346	0.235	0.095
$q50$	0.437	0.338	0.387	0.340	0.231	0.093
$q40$	0.436	0.337	0.385	0.339	0.231	0.093
$q30$	0.436	0.337	0.385	0.339	0.231	0.093
$q25$	0.436	0.337	0.385	0.339	0.231	0.093
			Reweighting			
$q90$	0.422	0.316	0.386	0.355	0.241	0.098
$q67.6$	0.416	0.307	0.358	0.330	0.225	0.091
$q50$	0.427	0.322	0.373	0.335	0.228	0.092
$q40$	0.435	0.335	0.384	0.339	0.230	0.093
$q30$	0.436	0.337	0.385	0.339	0.231	0.093
$q25$	0.436	0.337	0.385	0.339	0.230	0.093
			BFM			
$q90$	0.421	0.316	0.385	0.355	0.241	0.098
$q67.6$	0.416	0.307	0.357	0.330	0.225	0.091
$q50$	0.427	0.322	0.372	0.335	0.228	0.092
$q40$	0.435	0.335	0.383	0.339	0.230	0.093
$q30$	0.436	0.337	0.384	0.339	0.231	0.093
$q25$	0.436	0.337	0.384	0.339	0.230	0.093

Table 2: Simulated data: inequality measures computed from the true and distorted distributions, and from correction methods with several tax-quantile thresholds t.

t	Gini	MLD	Theil	Top10%	Top5%	Top1%
tax	0.472	0.423	0.448	0.359	0.247	0.102
survey	0.382	0.254	0.272	0.300	0.192	0.068
			Replacing			
$q90$	0.440	0.336	0.419	0.373	0.253	0.104
$q72$	0.446	0.343	0.414	0.355	0.244	0.100
$q60$	0.447	0.346	0.412	0.353	0.243	0.100
$q50$	0.447	0.346	0.410	0.350	0.241	0.099
$q40$	0.448	0.347	0.412	0.351	0.242	0.099
$q30$	0.450	0.350	0.414	0.351	0.242	0.099
			Reweighting			
$q90$	0.442	0.338	0.409	0.358	0.246	0.100
$q72$	0.435	0.330	0.391	0.344	0.237	0.097
$q60$	0.441	0.337	0.400	0.348	0.239	0.098
$q50$	0.443	0.340	0.402	0.349	0.239	0.097
$q40$	0.452	0.354	0.414	0.352	0.242	0.098
$q30$	0.458	0.365	0.422	0.354	0.243	0.099
			BFM			
$q90$	0.443	0.341	0.408	0.362	0.250	0.102
$q72$	0.435	0.332	0.387	0.347	0.239	0.097
$q60$	0.441	0.340	0.397	0.350	0.242	0.098
$q50$	0.444	0.344	0.400	0.352	0.240	0.099
$q40$	0.452	0.357	0.412	0.355	0.242	0.102
$q30$	0.458	0.369	0.421	0.358	0.245	0.104

Table 3: Uruguayan linked data: inequality measures computed from the tax and survey samples, and from correction methods with several tax-quantile thresholds t.

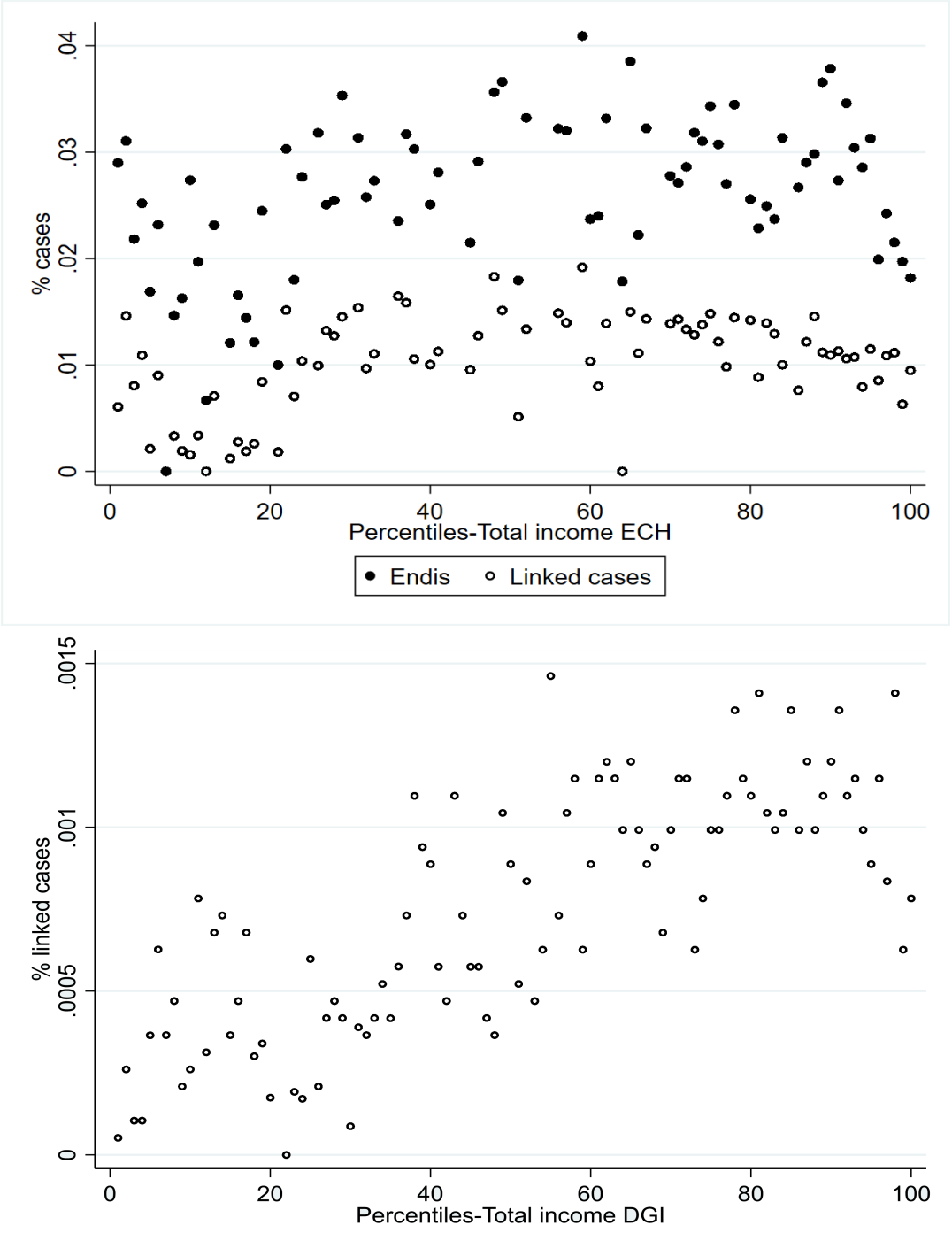

Figure 1: Proportion of ENDIS income receivers and linked observations by income percentile in ECH and DGI

Note: In the first panel, the label *ENDIS* corresponds to adults with positive income in ECH. Percentiles were built with the full set of ECH adults receiving positive income. In the second panel, percentiles were built with the full set of 2012/2013 DGI observations
Source: authors´ calculations based on ECH and ENDIS microdata.

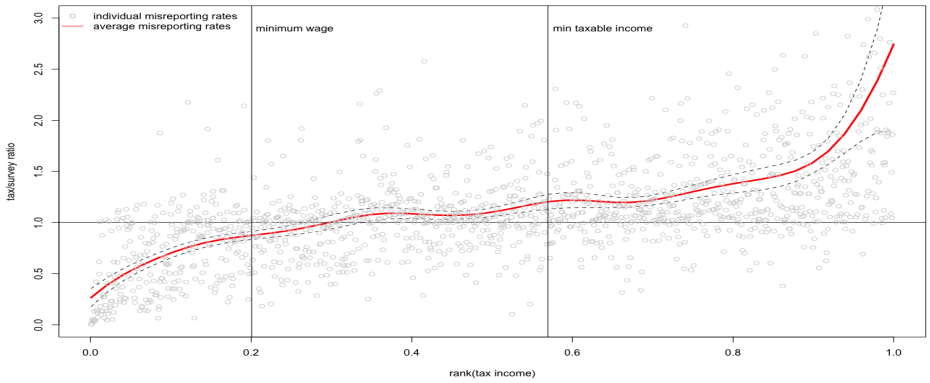

Figure 2: Linked data: misreporting rates, computed as ratios of tax return to survey income (circles), with nonparametric estimation of average misreporting rates (red line)

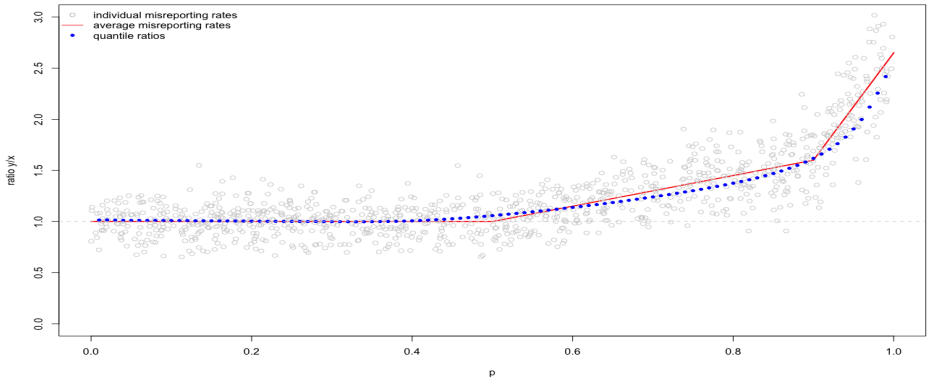

Figure 3: Simulation design: average misreporting rate function $r(p)$, quantile ratios $Q_Y(p)/Q_X(p)$, and 1000 simulated misreporting rates, $y/x = r(p)\varepsilon$.

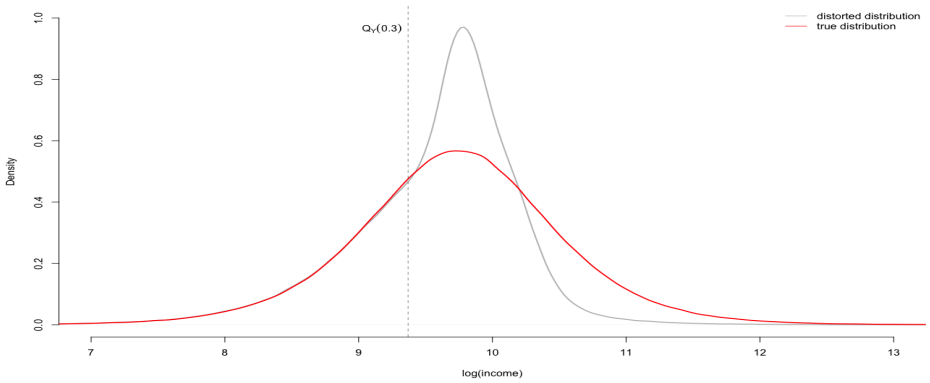

Figure 4: True hypothetical distribution (red line), and distorted distribution (gray line) which suffers from underreporting

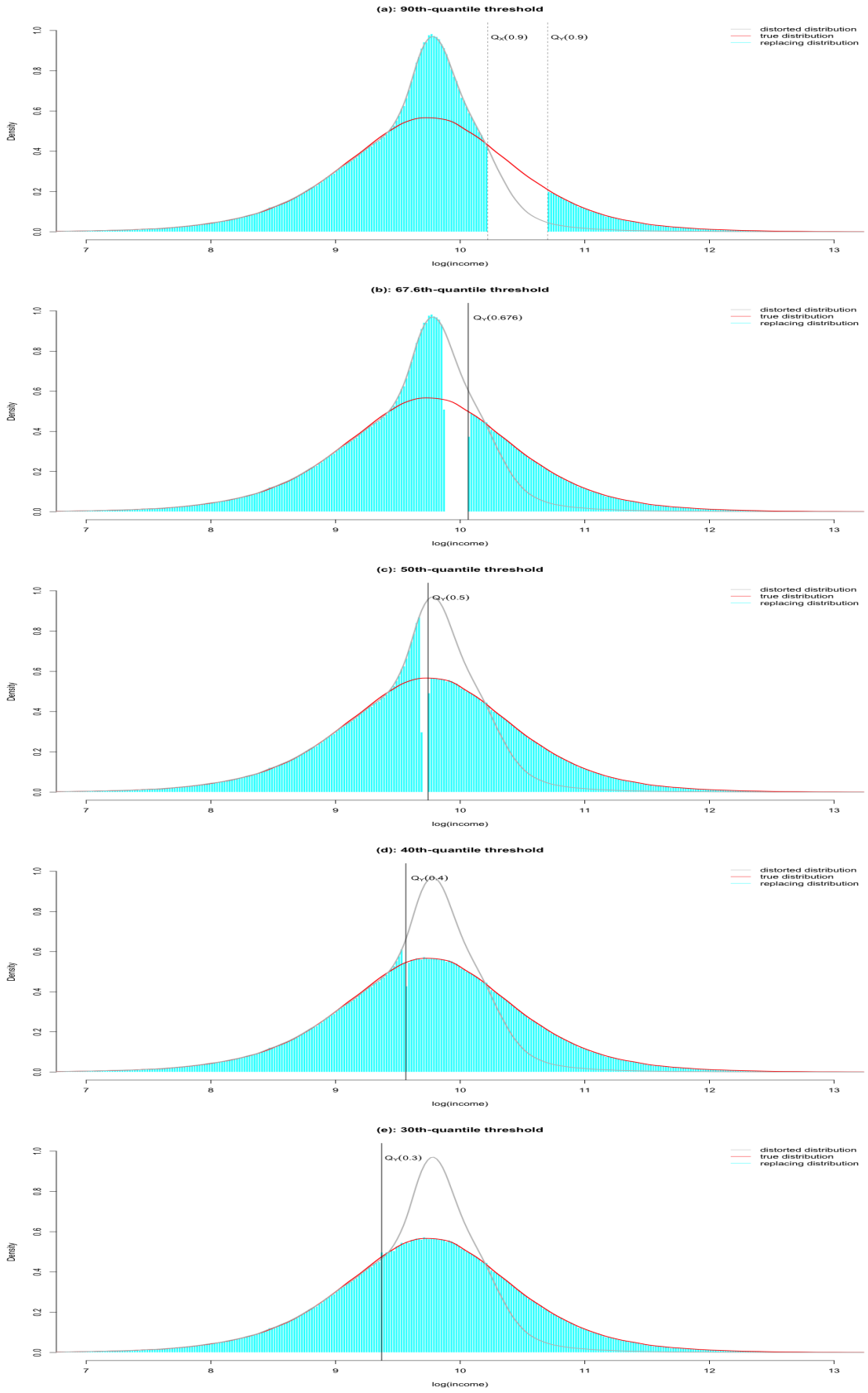

Figure 5: Replacing distributions with several tax-quantile thresholds

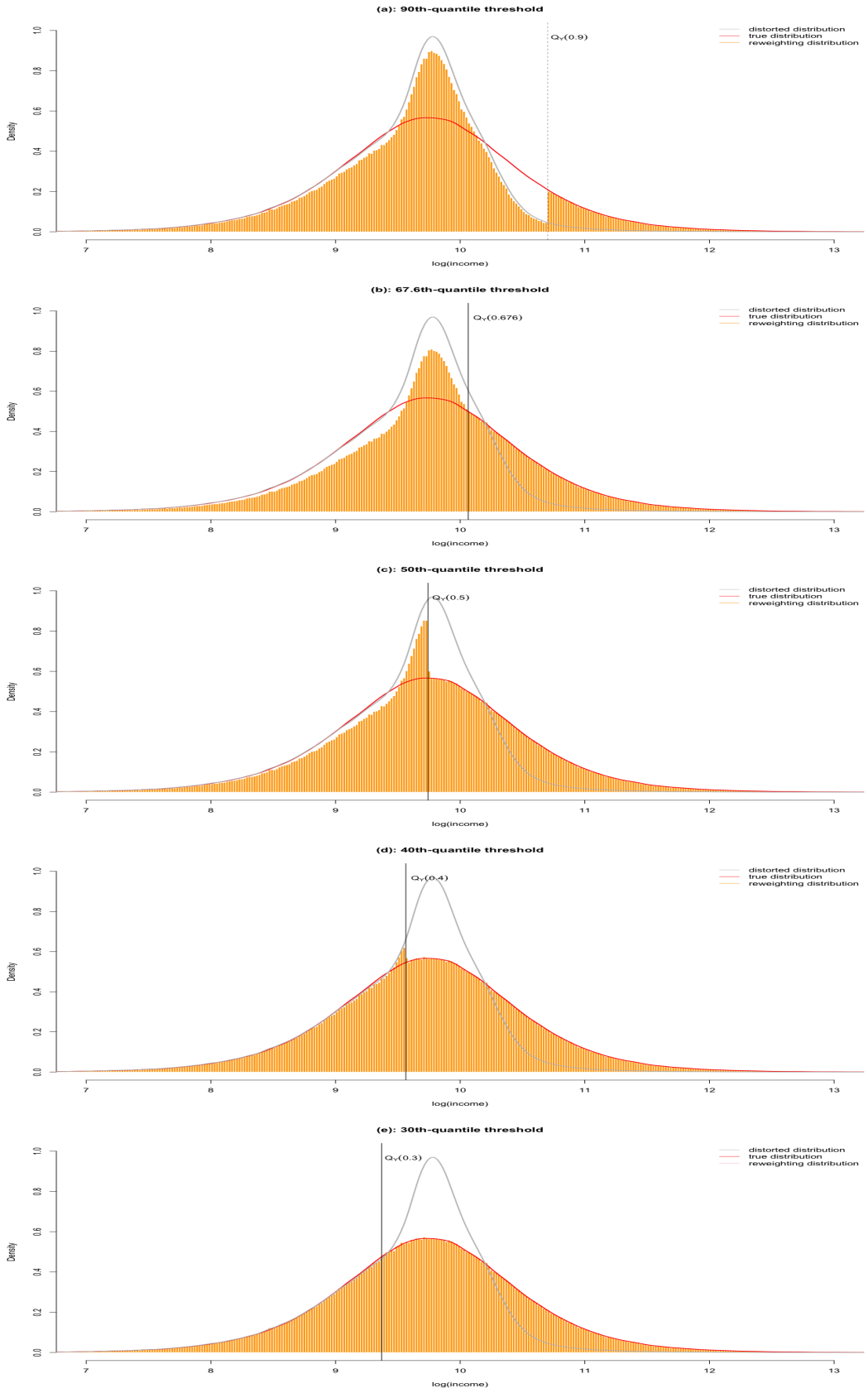

Figure 6: Reweighting distributions with several tax-quantile thresholds

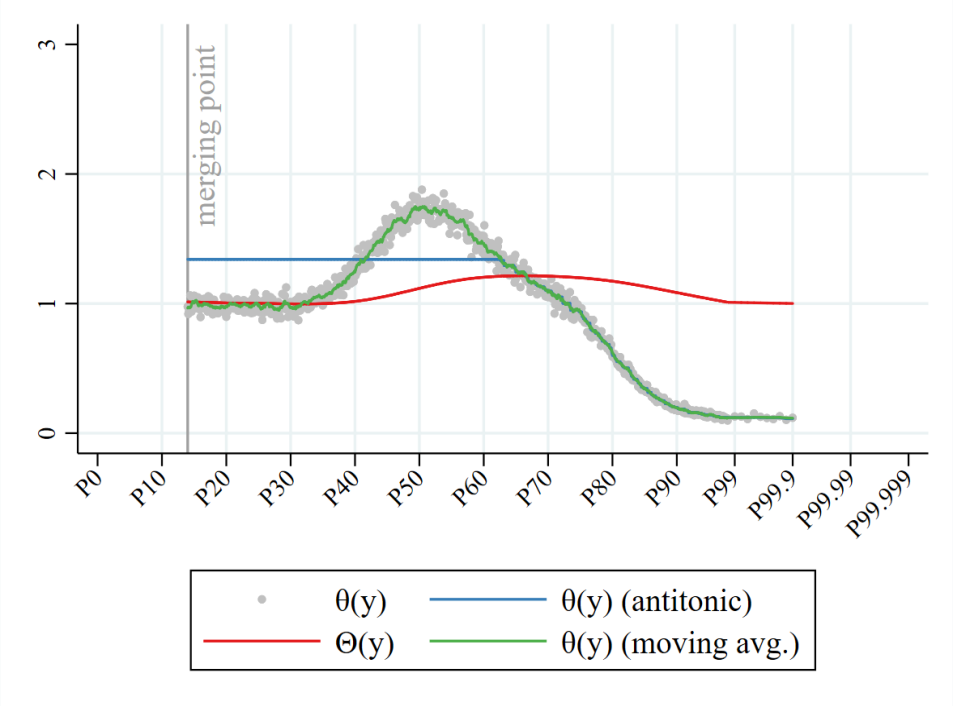

Figure 7: Simulated data: density ratio function, $\theta(y) = f_X(y)/f_Y(y)$, and CDF ratio function, $\Theta(y) = F_X(y)/F_Y(y)$, obtained with BFM method.

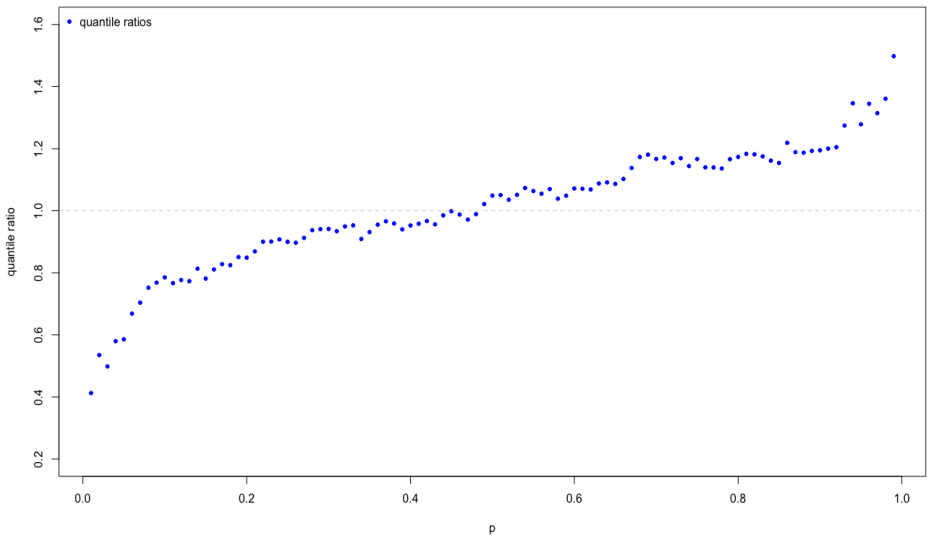

Figure 8: Uruguayan linked data: quantile ratios, $\widehat{Q}_Y(p)/\widehat{Q}_X(p)$

Appendix 1. Personal income taxation in Uruguay

.

As mentioned in section II, the Uruguayan personal income tax scheme is based on *Impuesto a las Retribuciones de las Personas Físicas* (IRPF) and *Impuesto de Ayuda a la Seguridad Social* (IASS). Being a dual scheme, it combines a progressive tax schedule for labor earnings and pensions, and a flat rate on capital income. The information we use in this study is reported in forms 1444, 3100, 1102, 1103, 1104 and 1201.

IRPF

Capital income

Capital income comprises interests from bank deposits and other financial assets, entrepreneurial profits and dividends and rents from real estate capital and lease. The first group includes all cash or in-kind rents coming from bank deposits and other financial assets, business profits and dividends distributed by firms contributing to entrepreneurial income tax (*Impuesto a las Retribuciones de las Empresas*, IRAE) and copyright among others. Public debt interests, gains obtained from private capitalization pension accounts and business profits distributed by firms with total annual revenue below 40,000 US dollars are exempt from IRPF and from filing a tax return. The same holds for liberal professionals if individuals opt to contribute IRAE. Banks, real estate agencies and institutions in charge of payments are set as retention agents. In absence of a retention agent, capital owners can pay in advance and file a tax return at the end of the year. Tax rates are flat but they differ depending on the type of capital rent (Table A1.1).

Individuals receiving housing rents below 40 *Bases de Prestaciones y Contribuciones* (BPC) per year are not subject to IRPF in case they do not have other capital rents higher than 3 BPC a year.

Income concept	Tax rate
Interests- bank deposits in Uruguayan currency	3%
Interests- bank deposits-one year or more, in Uruguayan currency with no indexation clause	5%
Dividends or business profits distributed or credited by IRAE contributors	7%
Copyright	7%
Other capital income sources (real estate rents, lease, etc.)	12%

Table A1.1: Personal income tax rates by capital income concept. Uruguay, 2012/13
Source: DGI (2013).

Labor income bracket (BPC))	Marginal tax rate
0 - 84	0%
84 - 120	10%
120 - 180	15%
180 - 600	20%
600 - 1200	22%
Above 1200	25%

Table A1.2: Personal labor income tax rates by labor income bracket. Uruguay, 2012/13

Source: DGI (2013).

Labor income

This group gathers labor earnings as employee or self-employed as well as unemployement benefits. Wages, salaries, commissions, overtime payments, vacation payments, annual leave, End of the year payments, per diem stipends not subject to return and any other payments received from employers are considered taxable income. Unemployment, illness and maternity subsidies, accident insurance and unemployment benefits and child allowances are excluded. Minimum thresholds and progressional tax rates are depicted in Table A1.2. Throughout the years, the minimum threshold has been increased, although in the period under study it remained unmodified.

Individuals having only one occupation do not need to file a tax return, as IRPF is withheld by their employers. Self-employed workers, contribute for all their labor income generated out of salaried work, and can deduce up to 30% of their income.

IASS

IASS tax rates are depicted in Table A1.3.

Pension income bracket (BPC)	Marginal tax rate
0-96	0
96-180	10
180-600	20
600 and more	25

Table A1.3: Pension income tax rates by bracket. Uruguay, 2012/13

Source: DGI (2013).

Appendix 2. Additional tables and figures

	Linked to tax records		Not linked to tax records		Total
Variable	INR	ME	ECH h.inc=0	ECH h.inc.> 0	
N	163	1,471	2,179	726	4,539
% Women	78.52	67.48	84.5	26.59	68.98
% Proxy respondents	65.82	64.4	80.95	44.55	69.0
Age (mean)	26.76	34.47	37.3	34.39	39.14
Average years of schooling	8.64	11.68	7.88	10.65	9.75
%Employed	50.31	95.85	30.49	95.33	63.37
% Formal workers	36.10	97.79	14.06	94.65	49.76
% Formal private workers declaring underreporting (a)	9.76	8.36	7.27	7.88	7.92
Worked hours (weekly average)	17.98	36.37	9.59	41.76	23.86
% Poverty (national poverty line)	24.54	6.05	35.97	10.17	21.16

Table A.2.1: Descriptive statistics-ENDIS respondents by linkage status

Note: ENDIS adults were separated in four groups. Among linked individuals, we separate those that reported zero harmonized income in ECH (*ECHh.inc.*) but had positive income in DGI (interpreted as Item non response, INR; n=163) and those who had positive harmonized income in ECH and DGI (used in the Measurement error analysis, ME; n=1,471) in the reference period at the survey. The remaining two groups correspond to individuals not linked to tax data and include those with zero harmonized income that reported informal employment or did not work at all (2,179) or individuals that did not provide an ID AND have positive harmonized income (n=726).

(a) ECH includes a question asking formal workers if their contributions to the social security system correspond to their whole salary.

Source: authors´calculations based on ECH, ENDIS and DGI microdata.

t	Gini	MLD	Theil	Top10%	Top5%	Top1%
true	0.436	0.335	0.385	0.339	0.230	0.093
distorted	0.387	0.268	0.262	0.279	0.163	0.053
			Replacing			
$q99$	0.415	0.309	0.353	0.312	0.201	0.097
$q95.4$	0.434	0.334	0.383	0.336	0.230	0.093
$q93$	0.436	0.336	0.384	0.339	0.230	0.093
$q90$	0.436	0.336	0.384	0.338	0.230	0.093
$q85$	0.436	0.336	0.384	0.339	0.230	0.093
$q80$	0.436	0.336	0.384	0.339	0.230	0.093
			Reweighting			
$q99$	0.422	0.318	0.362	0.319	0.207	0.095
$q95.4$	0.440	0.343	0.381	0.334	0.220	0.089
$q93$	0.436	0.336	0.384	0.338	0.229	0.093
$q90$	0.436	0.336	0.385	0.339	0.230	0.093
$q85$	0.436	0.336	0.385	0.339	0.230	0.093
$q80$	0.436	0.336	0.385	0.339	0.230	0.093
			BFM			
$q99$	0.422	0.318	0.360	0.319	0.207	0.095
$q95.4$	0.440	0.343	0.380	0.334	0.220	0.089
$q93$	0.436	0.336	0.383	0.338	0.229	0.093
$q90$	0.436	0.336	0.383	0.339	0.230	0.093
$q85$	0.436	0.336	0.384	0.339	0.230	0.093
$q80$	0.436	0.336	0.384	0.339	0.230	0.093

Table A.2.2: Simulated data with underreporting concentrated in the top of the distribution: inequality measures computed from the true and distorted distributions, and from correction methods with several tax-quantile thresholds t. The optimal threshold is at the $90th$-quantile. The BFM threshold is selected at the $95.4th$-quantile.

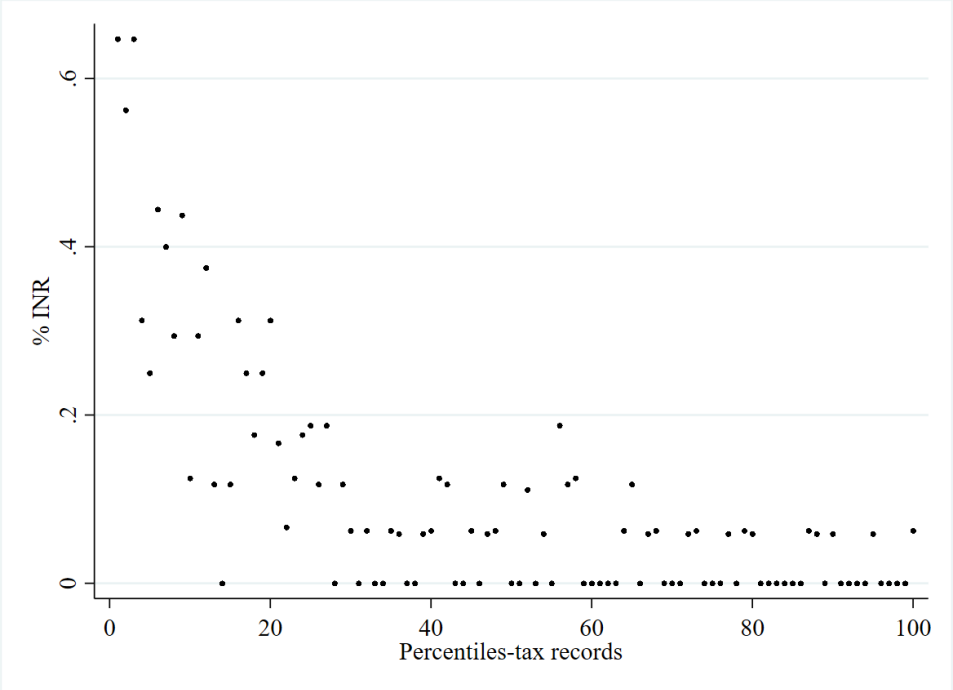

Figure A.2.1: Proportion of linked observations with item non response by percentile of post-tax DGI income

Note: Linked individuals that reported zero harmonized income in ECH but had positive income in DGI in the survey reference period were labeled as item non response observations.
Source: authors´ calculations based on ECH, ENDIS and DGI microdata.

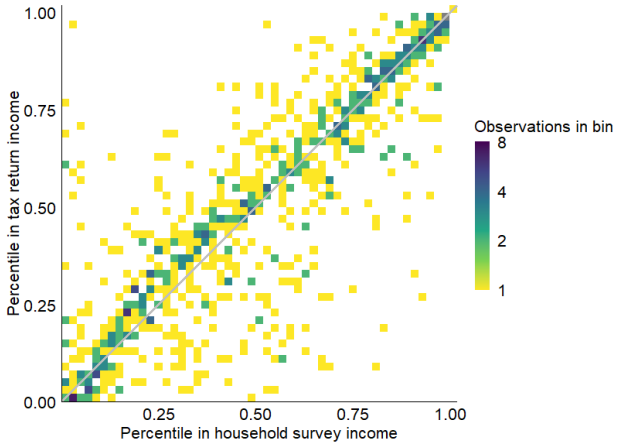

Figure A.2.2: Bivariate distribution of rank in tax return and household survey distributions. Linked observations

Source: authors´ calculations based on ECH, ENDIS and DGI microdata.

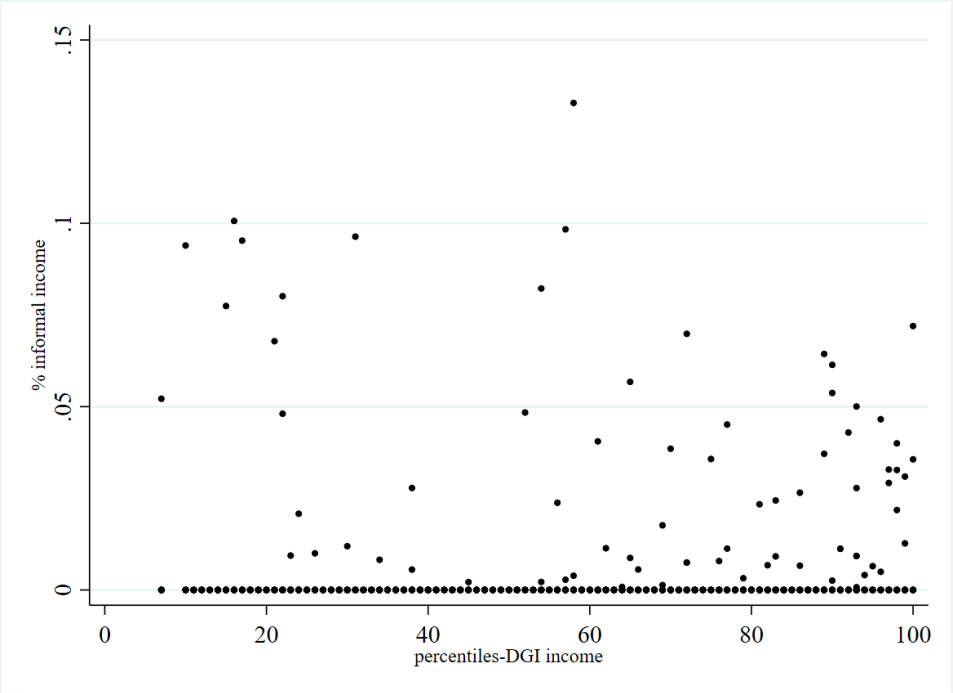

Figure A.2.3: Proportion of informal earnings in total ECH income by DGI income percentile. Linked cases.

Source: authors´ calculations based on ECH, ENDIS and DGI microdata.

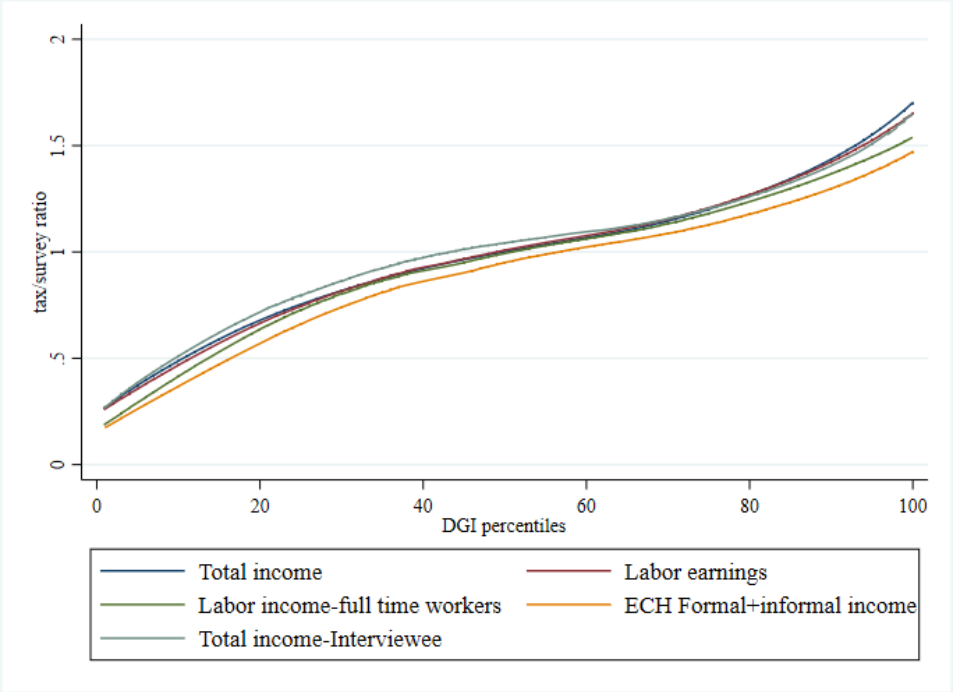

Figure A.2.4: Misreporting rates by population subgroup and DGI income percentile. Linked observations

Note: Misreporting rates were smoothed with a locally weighted regression.
Source: authors´calculations based on ECH, ENDIS and DGI microdata.