The Weight of the Rich

Improving Surveys Using Tax Data

Thomas Blanchet[1]

thomas.blanchet@wid.world

Ignacio Flores[1,2]

ignacio.flores@insead.edu

Marc Morgan[1]

marc.morgan@psemail.eu

*[1] Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris, France
[2] INSEAD, Boulevard de Constance, 77300 Fontainebleau, France

Abstract

Household surveys often fail to capture the top tail of income and wealth distributions, as evidenced by studies based on tax data. Yet to date there is no consensus on how to best reconcile both sources of information. This paper presents a novel method that helps to solve the problem under reasonable assumptions. Our key innovation is to endogenously determine a "merging point" between the datasets, above which we start to incorporate information from tax data into the survey. We use this merging point within a concrete survey correction method that involves both a "reweighting" and a "replacing" step, and which preserves the microdata structure of the original survey. We thus ensure that resulting estimates incorporate all the information at our disposal from both the survey and the tax data. We illustrate our approach with simulations, which show that our method performs better than existing alternatives. We also apply it to five countries, both developed and less developed, and find changes to the levels and trends in income inequality.

Keywords: Calibration, Inequality, Reweighting, Survey Representativeness, Tax Data.

Classification: C83, D31, N30.

Introduction

For a long time, most of what we knew about the distribution of income, wealth and their covariates came from surveys, in which randomly chosen households are asked to fill out a questionnaire. Household surveys have been an invaluable tool for tracking the evolution of society. But in recent years, the research community has grown increasingly concerned with their limitations. In particular, surveys have struggled to keep track of the evolution of the top tail of the distribution, due mainly to heterogeneous response rates, misreporting and small sample bias, which distort all sorts of distributional estimates. These biases end up affecting the way public policy is designed and evaluated.

For this reason, researchers have been increasingly exploring a different source to study inequality: tax data. The idea is not new; we can trace it back to the seminal work of Kuznets (1953), or even Pareto (1896). More recently, Piketty and Saez (2003) and Piketty (2003) applied their method to more recent data for France and the United States. This work was extended to more countries by many researchers whose contributions were collected in two volumes by Atkinson and Piketty (2007, 2010).

But tax data have their own limitations. They usually only cover the top of the distribution and include, at best, a limited set of covariates. They do not capture informal and tax-exempt income. They are often not available as microdata but rather as tabulations summarizing the distribution, which limits their usage. The statistical unit that they use (individuals or households) depends on the local legislation and may not be comparable from one country to the next. This is why many indicators, such as poverty rates or gender gaps, have to be calculated from surveys. The use of different — and sometimes contradictory — sources to compute statistics can make it hard to build consistent and accurate narratives on distributional matters. This explains the ongoing effort to combine the different data sources at our disposal in a way that exploits their strengths and corrects their weaknesses.

There have been a number of suggested approaches to deal with the problem of merging tax and survey data, yet the literature has largely failed to converge towards a standard. Crucially, most of the existing approaches make arbitrary assumptions for key parameters when applying the correction, and overlook the goal of preserving the survey's representativeness in terms of covariates.

In this paper, we develop a methodology to address the underrepresentation of top incomes in surveys that has significant advantages over previous ones. Our method avoids relying, to the extent possible, on *ad hoc* assumptions and parameters. Our key innovation is to present a data-driven way to determine the point in the survey data where the undercoverage of income starts, and use it as the basis for combining data sources. We call this point the "merging point" — the point in the distribution above which survey data and tax data are joined together.

Once this merging point has been chosen, we correct the survey by applying both a "reweighting" and a "replacing" step. The reweighting step enforces the representativeness of the survey in terms of income as well as other covariates. It follows the standard survey calibration theory that is already routinely employed by statistical institutes to enforce survey representativeness in terms of socio-demographic variables such as age, gender, household type, etc. The replacing step aims to overcome the limited sample size of the survey, making sure that the corrected survey data can match the tax data up until the narrowest top income groups, while otherwise preserving the distribution of covariates and their dependency with income. Importantly, we apply these two steps in a way that preserve the microdata structure of the survey, making it possible to use the corrected data with different statistical units, equivalence scales, to calculate complex indicators, and to perform decompositions along any dimension included in the survey.

We illustrate the method with two different types of applications. First, since the true distribution of income and wealth is always unknown, we simulate artificial populations, from which we draw surveys with two types of biases: heterogeneous response rates, and misreporting. We use these experiments to assess the accuracy and precision of our method and to compare its estimates to those derived from both the raw sample and the most common alternative methods that use external data — namely methods that directly replace survey incomes with tax incomes for the same quantiles in the distribution. We demonstrate that our method is superior to available options, not only because it relies on more reasonable assumptions and enables the use of resulting microdata sets (unlike the usual "replacing" alternative), but also because it produces estimates that are consistently closer to true values with lower variance.

In our second application, we apply our method to real data from five countries: France, the United Kingdom, Norway, Brazil and Chile. Our case studies are chosen to showcase the applicability of the method to both developed and less developed countries. The method makes upward revisions to inequality estimates in all cases, with varying degrees of magnitude, depending on the quality of the underlying data and the level of inequality in each country. It can also produce differing inequality trends. Moreover, our empirical results support the findings of our simulations concerning the difference between our method and the replacing alternative.

For practical use, we have developed a Stata command that applies our method. The program works with several input types, income concepts and statistical units, ensuring flexibility for users. Our method may therefore easily be used by researchers interested in analyzing the different dimensions of inequality.[1] The main goal of this paper is to describe the theoretical and practical details behind this readily usable method, as well as its advantages with respect to existing approaches.

The remainder of the paper is structured as follows. In section 1 we relate our paper to the existing literature. In section 2 we lay out the theoretical framework of our method. This is followed by applications to simulated distributions and practical applications to specific countries in section 3, before concluding.

1 Literature on Survey Correction Methods

Numerous studies have adjusted survey data to improve their representativeness and produce a more accurate distribution of income, sometimes using administrative data. We

[1] The package to download is `bfmcorr` for the correction method, which includes two sub-commands: `postbfm` for the post-estimation output and `bfmtoy` for parametric simulations. The command and its sub-commands come with a full set of user instructions.

identify three methodological strands present in this literature. The first strand reweight survey observations. The second strand replaces the income value of observations with a value typically drawn from a parametric distribution or an external data source. Finally, a third strand identifies the need to employ a hybrid procedure by combining reweighting and replacing.

1.1 Reweighting Observations

The studies that focus on reweighting usually formalize the underrepresentation of top incomes as the result of nonresponse. Many papers in this literature estimate a parametric model of nonresponse to adjust survey weights, but do not use direct data on the distribution of income. Korinek, Mistiaen, and Ravallion (2006) make this type of adjustment using nonresponse rates across geographic areas and the characteristics of respondents within regions. This type of approach can be sensitive to the degree of geographic aggregation used calculating response rates. This is an issue explored in more detail by Hlasny and Verme (2017; 2018) respectively for the United States and Europe, using similar probabilistic models. Depending on the nature of the survey data, greater or lesser geographic disaggregation levels are appropriate.

Crucially, these models do not use direct information on the income distribution — often due to lack of availability. Instead, they have to infer relationships between individual nonresponse and individual characteristics based on aggregated data. Because they cannot isolate and observe certain groups directly, it can be difficult with these methods to estimate the behavior of very specific groups that make up a large fraction of inequality, such as the top 1%. Our proposal instead makes use of direct administrative data to determine the relationship between income and nonresponse.

There are a few studies in this literature that combine surveys with external sources to measure inequality. An example of this is the case study of Argentina in Alvaredo (2011), in which the corrected Gini coefficient is estimated by assuming that the top of the survey distribution (top 1% or top 0.1%) completely misses the richest individuals that are represented in tax data. This accounts for underrepresentation of top incomes via an implicit reweighting procedure on the entire survey. However, the assumed nature of the underrepresentation is highly stylized: the survey is perfectly representative until a predetermined quantile, and fails to include anyone beyond it. Furthermore, in both of the empirical applications (the United States and Argentina) the threshold beyond which the tax data is used is chosen arbitrarily. Our method, on the other hand, tries its best to avoid arbitrary choices on the portion of the survey distribution to be corrected or on the form of the bias implied by the correction.

To our knowledge the paper that comes closest to ours, in terms of criteria and methodology, is Medeiros, Castro Galvão, and Azevedo Nazareno (2018) applied to Brazilian data. That is, it is the only study that combines tabulated tax data with survey micro data by explicitly reweighting survey observations. More specifically, the authors apply a Pareto distribution to incomes from the tax tabulation to correct the top of the income distribution calculated from the census. Their method involves reweighting the census population by income intervals above a specified merging point.

However, important differences remain. Contrary to our method, the choice of the merging point is not endogenous, but chosen by the authors. Thus, multiple points can be used, and indeed the authors test two. Our method endogenously determines a single merging point based on a more comprehensive treatment of how the representativeness of the population varies with income. Importantly, our approach preserves the continuity of the density of income — something that only our specific choice of the merging point can ensure. To guide their choice of the merging point, the authors look at the rank at which income in the tax data exceeds that of the survey. However, such a point is not theoretically well defined (see our discussion in section 2.1.2 and in appendix A.3).

Moreover, while they increase the weight of observations above the merging point, they do not reduce the weight of individuals below this point, such that the corrected population ends up being larger than the original official population. The authors do not provide a way to ensure the representativeness of characteristics other than income after the adjustment either — their purpose is solely to remedy the underestimation of top incomes. Moreover, their method does not remedy the lack of precision at the top of the distribution arising from sampling limitations, resulting in downward biased income shares of narrow income groups, especially in small samples. In contrast, our method addresses all of these issues.

1.2 Replacing Incomes

The general feature of the "replacing" approach is that it involves the direct replacing of survey income with income from tax data. Although there is no unified theory or explicit justification behind the applications of this adjustment procedure, most of these methods share some defining characteristics. In practice, they generally adjust distributions by replacing cell means in the survey distribution of income with those from the tax distribution for the same sized cells (i.e., fractiles) of equivalent rank in the population. The size of the cells varies by study (Burkhauser et al., 2016; Piketty, Yang, and Zucman, 2017; Chancel and Piketty, 2017; Czajka, 2017). Furthermore, the overall size of the population group whose income is to be adjusted is sometimes chosen arbitrarily, such as the top 20% in the distribution (Piketty, Yang, and Zucman, 2017), the top 10% (Burkhauser et al., 2016; Chancel and Piketty, 2017), the top 1% (Burkhauser, Hahn, and Wilkins, 2016; Alvaredo, 2011), or the top 0.5% of survey observations (DWP, 2015).

This decision can be made less arbitrarily by comparing thresholds or average incomes by fractile in the two distributions. The size of the group is then chosen as the point in the distribution where the two quantile functions cross (e.g., Czajka (2017)). However, the definition of this point is problematic (see our discussion in section 2.1.2 and in appendix A.3).

Other non-arbitrary choices include the minimum income level that requires mandatory tax filing (Diaz-Bazan, 2015). This keeps the use of survey data to measure the top of the income distribution to a strict minimum, and assumes that the entire tax distribution is reliable. We argue, however, that not all the income in tax data should be considered reliable, given the difference between declarable income thresholds and taxable income thresholds. The quality of tax data generally increases with income in a manner that is often not well defined, and given this uncertainty it makes sense to limit their use to the portion that is absolutely necessary.

In certain cases, the survey distribution stops being reliable before the tax data can be trusted. This happens in particular in countries where only a small part of the population file a tax return. In such cases, from the point at which we stop trusting the survey to the point at which we start trusting the tax data, researchers progressively rescale upwards the income values from the survey distribution, using various profiles of rescaling coefficients (usually linear) (Chancel and Piketty, 2017; Piketty, Yang, and Zucman, 2017; Novokmet, Piketty, and Zucman, 2018). This procedure ensures at least that the quantile function is continuous. These rescaling methods can be seen as an extension of the usual replacing methods (and face similar issues).

1.3 Combined Reweighting and Replacing

Some voices stress the need to combine the aforementioned correction approaches. Bourguignon (2018), while reviewing the typical adjustment methods employed, highlights that any method must dwell on three important parameters: the amount of income to be assigned to the top, the size of this top group, and the share of the population added to the top in the survey. The definition of these three parameters implies a correction procedure combining reweighting and replacing methods. His analysis goes on to study the ways in which these choices impact the adjustments made to the original distribution. However, this analysis does not shed light on *how* to make these choices. Moreover, in reviewing multiple correction methods and applying them to Mexican survey data (including the combined case, where all three parameters mentioned take non-zero values), he only considers the situation "where nothing is known about the distribution of the missing income, unlike when tax records or tabulations are available" (Bourguignon, 2018). Our approach for correcting survey microdata combines the two previous methods, and seeks to determine how to make these choices by explicitly comparing the survey and the tax data distribution.

2 Theory and Methodology

We describe our method in three steps: first is the choice of the merging point, second is the reweighting step, and third is the replacing step.

2.1 Merging Point

The choice of merging point is the key innovation of our approach. We argue that by looking at how the representativeness of the survey varies with income, we can locate a single merging point that satisfies elementary requirements. That is, our choice of merging point is the only way to ensure that (i) the representativeness of the survey varies continuously with income, that (ii) the representativeness decreases monotonously for top incomes and that (iii) income groups below the merging point are homogeneously represented, and therefore the survey should not be distorted below that point.

We start by formalizing our notion of representativeness by income, by explaining its properties and our assumptions. Then, we explain how we define and estimate our merging point.

2.1.1 General Setting

Survey vs. Tax Data Distribution Let X and Y be two real random variables. We will use Y to represent the "true" (tax data) income distribution, and X to represent the income distribution recorded in the survey. Each random variable has a probability density function (PDF) f_Y and f_X, a cumulative probability function (CDF) F_Y and F_X, and a quantile function Q_Y and Q_X.

Since every individual in the income earning population does not file a tax return, we assume that the tax data only covers a portion of the true income distribution. In reality, however, part of the actual income of this subset of the population may also be missing from the tax data due to tax-exempt income, the subtraction of certain deductions, as well as tax evasion. The extent of these omissions varies by country. As does the presence of top coding in micro-level datasets.[2] Since our methodology does not in itself correct for these issues, we urge users to scrutinize the definitions and coverage of income in both data sources prior to combining the information.

Definition of Representativeness Let $\theta(y) = f_X(y)/f_Y(y)$ be the ratio of the survey density to the true density at the income level y. This represents the number of people within an infinitesimal bracket $[y, y + \mathrm{d}y]$ according to the survey, relative to the actual number of people in the bracket. If $\theta(y) < 1$, then people with income y are underrepresented in the survey. Conversely, if $\theta(y) > 1$, then they are overrepresented. In general terms, this ratio indicates over- or underrepresentation of people at a given income level.

The value of $\theta(y)$ may be interpreted as a relative probability of inclusion in the survey. Indeed, let D be a binary random variable that denotes participation in the survey: if an observation is included in the sample, then $D = 1$, otherwise $D = 0$. Bayes's formula implies:

$$\theta(y) = \frac{f_X(y)}{f_Y(y)} = \frac{1}{f_Y(y)} \times f_Y(y) \frac{\mathbb{P}\{D = 1 | Y = y\}}{\mathbb{P}\{D = 1\}} = \frac{\mathbb{P}\{D = 1 | Y = y\}}{\mathbb{P}\{D = 1\}} \tag{1}$$

If everyone has the same probability of inclusion, then $\mathbb{P}\{D = 1 | Y = y\} = \mathbb{P}\{D = 1\}$, and $\theta(y) = 1$. Hence $f_X(y) = f_Y(y)$ and the survey is unbiased. What matters for the bias is the probability of inclusion at a given income level relative to the average inclusion rate, which is why we have the constraint $\mathbb{E}[\theta(Y)] = 1$. Intuitively, if some people are underrepresented in the survey, then mechanically others have to be overrepresented, since the sum of weights must ultimately sum to the population size. This elementary constraint has important consequences for how we think about the adjustment of distributions. Any

[2] If incomes in the survey data are top-coded, our methodology can at least partially address the issue. The difficulty arises from the fact that with top coding, information on the ranking of observations above a certain threshold is lost, which makes it difficult to know how other variables vary with income at the top. The replacing step of our method will fix the top-coding problem from the perspective of the univariate distribution of incomes, but information about the covariates may be of limited use. Note that the same could be said of the raw survey.

modification of one part of the distribution is bound to have repercussions on the rest. In particular, if the “rich” are underrepresented, then the “non-rich” *as a whole* must be overrepresented — even if the relative sizes of the different “non-rich” subgroups remain valid (and thus the survey is “correct” for the bottom).

[Place figure 1 here]

Figure 1 represents the situation, in the more common case where $\theta(y)$ is lower for top incomes. We show a truncated version of f_Y since tax data often only cover a limited part of the whole distribution. The fact that the dashed red line $f_Y(y)$ is above the solid blue line $f_X(y)$ for top incomes means that they are underrepresented. Therefore, lower incomes must be overrepresented, which is what we see below the point y^*. An appropriate correction procedure here would be to increase the value of the density above it, and decrease its value below it. This can be achieved by reweighting, in which multiply the survey density f_X by a factor $1/\theta(y)$ to make it equal to the true density f_Y. In practice, this means multiplying the weight of any observation Y_i by $1/\theta(Y_i)$.

Assumptions on the Rate of Representativeness When we observe both f_Y and f_X, we can directly estimate θ nonparametrically. But because we do not observe the true density over the entire support, we have to make an assumption on the shape of θ for values not covered by the tax data. We will assume a constant value, an assumption that we justify below. Then, we can write the complete profile of θ as:

$$\theta(y) = \begin{cases} \bar{\theta} & \text{if } y < \bar{y} \\ f_X(y)/f_Y(y) & \text{if } y \geq \bar{y} \end{cases} \tag{2}$$

We call $\bar{y}$ the *merging point.* It is the value above which we merge information from the tax data into the survey. The key aspect of our approach is that this point is endogenously determined by our methodology, not arbitrarily chosen by the researcher: we will return to it in section 2.1.2. For now we will take $\bar{y}$ as given, and only assume that it is below the pivotal point y^* of figure 1. Figure 2 shows how the reweighting using (2) operates.

[Place figure 2 here]

Let $\tilde{f}_X$ be the reweighted survey, i.e. $\tilde{f}_X(y) = f_X(y)/\theta(y)$. By construction, we have $\tilde{f}_X(y) = f_Y(y)$ for $y \geq \bar{y}$. As indicated by upward arrows on the right of figure 2, the density has been increased for $y > y^*$. Since densities must integrate to one, values for $y < y^*$ have to be lowered. The uniform reweighting below $\bar{y}$ creates the dotted blue line. Note that we decrease the weight of observations between $\bar{y}$ and y^*, but by an amount that gets progressively lower, from $\bar{\theta}$ at $\bar{y}$ to 1 at y^*.

Interpretation How should we interpret the assumption that θ is constant below $\bar{y}$? There are two key motivations. First, as we present in section 3, there is evidence that it is empirically verified. Second, this assumption results from our wish to introduce no more distortion than necessary to the survey data. That is, in the absence of any other piece of evidence, we trust that the survey correctly captures the relative size of the different subgroups of the bottom of the distribution, so that the overrepresentation of the non-rich is only the counterpart of the underrepresentation of the rich.

Assuming that there are also large variations of θ at the bottom means assuming that the survey is deficient for both the bottom and the top of the distribution. This situation would be difficult to fix with our data, and in any case would be outside of the scope of the paper. We nonetheless address this issue appendix A.2, and show that our overall method retain good properties even if this assumption is violated. Note that if we assume that the very poor are also underrepresented, so that θ has an inverted U-shape, then our correction will still underestimate the true level of inequality.

2.1.2 Choice of the Merging Point

Trustable Span For many countries, tax data only covers the top of the distribution. We use the term *trustable span* to name the interval over which the tax data may be considered reliable. It takes the form $[y_{\text{trust}}, +\infty[$. This interval is determined by country-specific tax legislation. It relies on the portion of the distribution covered in the data (declarations) or just on the portion of the tax population that pays income tax (taxpayers).

When choosing the merging point $\bar{y}$. It would be tempting to use the tax data over the entire trustable span (i.e., set $\bar{y} = y_{\text{trust}}$), but this will often lead to poor results. First, such choices would generally lead to undesirable features, such as a discontinuous θ. Moreover, the point from which the tax data become reliable is not usually sharp — the reliability of the tax data increases with income in a way that is not well defined, therefore it is more prudent to restrict their use to the minimum that is necessary. Second, once we are past the point where there is clear evidence of a bias, we should prefer to avoid unnecessarily distorting the survey.

Location of the Merging Point We suggest a simple, data-driven way for choosing a merging point with desirable properties. In particular, we seek to preserve the continuity of the underlying density function after reweighting. We start from the typical case where $\bar{y}$ is inside the trustable span $[y_{\text{trust}}, +\infty[$. In Appendix B we consider cases where the trustable span may be too small to observe an overlap between the densities.

Assume that the function $\theta(y)$ follows the form in (2). We introduce a second function, the cumulative rate of representativeness, defined as:

$$\Theta(y) = \frac{F_X(y)}{F_Y(y)} \tag{3}$$

In figure 3, we examine the shape of $\theta(y)$ and $\Theta(y)$ in relation to the density functions presented in figure 2. We have the relationship $\Theta(y)F_Y(y) = \int_{-\infty}^{y} \theta(t) f_Y(t)\,\mathrm{d}t$. Given (2), for $y < \bar{y}$, $\Theta(y) = \bar{\theta}$. As figure 3 shows, we should expect the merging point $\bar{y}$ to be the highest value y such that $\Theta(y) = \theta(y)$.

[Place figure 3 here]

Comparison with Alternative Procedures We can contrast this choice of merging point with the one implicitly chosen in at least some replacing approaches: the point at which the quantile functions of the survey and the tax data cross. In our view, there are two main issues with this approach, for which we provide technical details and formal proofs in appendix A.3.

First, the theoretical justification for why this point is relevant is often unclear. In fact, as we prove in appendix A.3, if we assume that the rate of representativeness is a decreasing function of income, then this point should not exist at all: the survey quantile function is always below the true quantile. In practice, this point tends to exist because tax data fails to cover bottom incomes. But in our view, the fact that this point exists only because of a deficiency of the tax data is precisely a good justification for avoiding to use it.

Second, whenever researchers use this point, they are making implicit assumptions about the shape of the representativeness rate θ that quite unrealistic. Indeed, as we explain in appendix A.3, this alternative choice of merging point leads to a rate of representativeness θ that is both discontinuous and nonmonotonous at the top. We stress that this is an issue regardless of the interpretation we make of θ — that of a response rate or the result of something more complex. In all cases we should expect θ to remain somewhat regular.[3]

Estimation of the Merging Point We can estimate both $\theta(y)$ and $\Theta(y)$ over the trustable span of the tax data. To determine the merging point in practice, we look for the moment when the empirical curves for $\Theta(y)$ and $\theta(y)$ cross.

The estimation of $\Theta(y)$ poses no difficulty as it suffices to replace the CDFs by their empirical counterpart in (3) to get the estimate $\hat{\Theta}_k$. For $\theta(y)$, however, we have to estimate densities. We define m bins using fractiles of the tax data distribution (from 0% to 99%, then 99.1% to 99.9%, then 99.91% to 99.99% and 99.991% to 99.999%). We approximate the densities using histogram functions over these bins. This gives a first estimate for each bin that we call $(\tilde{\theta}_k)_{1\le k\le m}$. The resulting estimate is fairly noisy, so we get a second, more stable one named $(\hat{\theta}_k)_{1\le k\le m}$ using an antitonic (monotonically decreasing) regression

[3] See appendix A.1 for a formalization of the problem with both nonresponse and misreporting.

(Brunk, 1955; Ayer et al., 1955; Eeden, 1958). That is, we solve:

$$\min_{\hat{\theta}_1,\dots,\hat{\theta}_m} \sum_{k=1}^{m} w_k(\hat{\theta}_k - \tilde{\theta}_k)^2 \qquad \text{s.t.} \qquad \forall k \in \{2,\dots,m\} \quad \hat{\theta}_{k-1} \geq \hat{\theta}_k$$

where w_k is the size of bin k (i.e., the fraction of the true population covered by that bin). We solve the problem above using the Pool Adjacent Violators Algorithm (Ayer et al., 1955). The main feature of this approach is that we force $(\hat{\theta}_k)_{1\leq k\leq m}$ to be decreasing. This turns out to be enough to smooth the estimate so that we can work with it, without the need introduce additional regularity requirements. We use as the merging point bracket the lowest value of k such that $\hat{\theta}_k < \hat{\Theta}_k$.

2.2 Reweighting

Once we have chosen an appropriate merging point, we still need to provide a concrete approach for correcting the survey. For that we proceed in two steps: the first one is the reweighting step, which we present here.

We stress that it is not enough for the survey to be solely representative in terms of income, we also need to preserve (or possibly enforce) representativeness in terms of other variables such as age, gender, household type, ethnicity, etc. To that end, we adapt a method that is already routinely used by statistical agencies to enforce the representativeness of survey data along several dimensions simultaneously: calibration theory. We start with a general presentation of linear survey calibration, then we explain we apply it to our problem. In appendix B, we discuss potential extensions of the calibration method to deal with more complex settings, which illustrate the flexibility of the framework.

2.2.1 Calibration

Problem Survey calibration considers the following problem. We have a survey sample of size n. Each observation is a k-dimensional vector $\boldsymbol{x}_i = (x_{1i},\dots,x_{ki})'$. The sample can be written $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$, and the corresponding survey weights are $(d_1,\dots,d_n)$. We know from a higher-quality external source the true population totals of the variables $x_{1i},\dots,x_{ki}$

as the vector $\boldsymbol{t}$. We seek a new set of weights, $(w_1, \dots, w_n)$, such that the totals in the survey match their true value, i.e., $\sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t}$.

This problem will in general have an infinity of solutions, therefore survey calibration introduces a regularization criterion to select the preferred solution out of all the different possibilities. The idea is to minimize distortions from the original survey data, so we consider:

$$\min_{w_1,\dots,w_n} \sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i} \qquad \text{s.t.} \qquad \sum_{i=1}^n w_i \boldsymbol{x}_i = \boldsymbol{t} \tag{4}$$

That is, we minimize the χ^2 distance between the original and the calibrated weights, under the constraint on population totals: this is called linear calibration. While alternative distances are sometimes used, linear calibration is advantageous in terms of analytical and computational tractability.

Solution Solving the problem (4) leads to:

$$\frac{w_i}{d_i} = 1 + \boldsymbol{\beta} \boldsymbol{x}_i \tag{5}$$

where $\boldsymbol{\beta}$ is a vector of Lagrange multipliers determined from the constraints as:

$$\boldsymbol{\beta} = \boldsymbol{T}^{-1} \left(\boldsymbol{t} - \sum_{i=1}^n d_i \boldsymbol{x}_i \right) \qquad \text{with} \qquad \boldsymbol{T} = \sum_{i=1}^n d_i \boldsymbol{x}_i \boldsymbol{x}_i'$$

where the matrix $\boldsymbol{T}$ is invertible as long as the constraints are neither redundant nor incompatible.[4] One undesirable feature of linear calibration is that it may lead to weights below one or even negative, which prevents their interpretation as an inverse probability and is incompatible with several statistical procedures. Therefore, in practice, we enforce the constraints $w_i \geq 1$ for all i using a standard iterative method described by Singh and Mohl (1996, method 5). This is known as truncated linear calibration.

Interpretation This procedure can be interpreted in terms of a nonresponse model. In this context, the survey weights are the inverse of the probability of inclusion in the

[4] In practice, we use the Moore–Penrose generalized inverse to circumvent these problems.

survey sample. This probability of inclusion is the product of two components. The first one depends on whether a unit is selected for the survey, regardless of whether that unit accepts to answer or not. We note $D_i = 1$ if unit i is selected, and $D_i = 0$ otherwise. The value $\delta_i = 1/\mathbb{P}\{D_i = 1\}$ is called the design weight. The design weight in constructed by the survey producer and therefore known exactly. The second component depends on whether a unit contacted for the survey accepts to answer or not. We note $R_i = 1$ if unit i accepts to participate in the survey, and $R_i = 0$ otherwise. The value $\rho_i = 1/\mathbb{P}\{R_i = 1\}$ is called the response weight. Since both D_i and R_i must be equal to 1 for a unit to be observed, the final weight is the product of these two components, $\delta_i\rho_i$.

Nonresponse is unknown so it has to be estimated using certain assumptions. The simplest one is that ρ_i is the same for all units, therefore all weights are upscaled by the same factor so that their sum matches the population of interest. More complex models use information usually available to the survey producer, that is, basic socio-demographic variables which we will write $\boldsymbol{U}_i$. The survey producer models nonresponse as a function of these variables: $\rho_i = \phi(\boldsymbol{U}_i)$ and provides weights equal to $\delta_i\phi(\boldsymbol{U}_i)$. If nonresponse is also a function of income, which is not observed by the survey producer, then the estimated nonresponse will fail to accurately reflect true nonresponse, leading to biased estimates of the income distribution. Using the tax data $\boldsymbol{Y}_i$, we can estimate a new model that takes income into account: $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)$. The final weight becomes:

$$\begin{aligned} w_i &= \frac{1}{\mathbb{P}\{D_i = 1\}} \frac{1}{\mathbb{P}\{R_i = 1\}} \\ &= \frac{1}{\mathbb{P}\{D_i = 1\}} \psi(\boldsymbol{U}_i, \boldsymbol{Y}_i) \\ &= \delta_i \phi(\boldsymbol{U}_i) \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \\ &= d_i \times \frac{\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)}{\phi(\boldsymbol{U}_i)} \end{aligned} \tag{6}$$

Comparing equation (5) with (6), we see that the calibration problem suggests both a functional form and an estimation method for $\psi(\boldsymbol{U}_i, \boldsymbol{Y}_i)/\phi(\boldsymbol{U}_i)$. This functional form assumes nonresponse profiles that are as uniform (thus nondistortive) as possible, and only modify the underlying distribution if it is necessary to do so. The preference for nondistortive functional forms is consistent with our decision to impose a constant value for θ below the merging point.

Application to Income Data The calibration problem is presented so as to enforce the aggregate value of variables. In order to use it to enforce the distribution of a variable, we have to discretize this distribution. In the case of income tax data, the income distribution may be presented in various tabulated forms, and we use the generalized Pareto interpolation method of Blanchet, Fournier, and Piketty (2017) to turn it into a continuous distribution.[5] We output the distribution discretized over a narrow grid made up of all percentiles from 0% to 99%, 99.1% to 99.9%, 99.91% to 99.99% and 99.991% to 99.999%. We discard tax brackets below the merging point, whose choice is described in section 2.1.2. We then match the survey data to their corresponding tax bracket. In general, it is necessary to regroup certain tax brackets to make sure that we have at least one (and preferably more) observations in each bracket. Otherwise the calibration will not be possible. We automatically regroup brackets to have a partition of the income distribution at the top such that each bracket has at least 5 survey observations.[6]

Assume that we eventually get m brackets, with the k-th bracket covering a fraction p_k of the population. We create dummy variables $b_1, \dots, b_m$ for each income bracket. If the total population is N and the sample size is n, then the calibrated weights should satisfy:

$$\forall k \in \{1, \dots, m\} \qquad \sum_{i=1}^{n} w_i b_{ik} = N p_k$$

Since these equations are expressed as totals of variables, they can directly enter the calibration problem (4). In practice, we are enforcing the income distribution through a histogram approximation of it.

[5] See `wid.world/gpinter` for an online interface and a R package to apply the method.

[6] That is, whenever a bracket in the tax data has fewer than 5 survey observations, the algorithm groups this bracket with the bracket just above. (Except for the top bracket, for which we group with the bracket just below.) This process is repeated until no bracket has fewer than 5 survey observations. The value of 5 is an argument of our command that can be changed.

The flexibility of the calibration procedure lets us put additional constraints in the calibration problem. In particular, if the survey is already assumed to be representative in terms of age, gender, or any other variable, then their distribution can be kept constant during the procedure. Hence we correct for the income distribution while maintaining the representativeness of the survey along the other dimensions.

For all the observations below the merging point, the dummy variables $b_1, \dots, b_m$ are all equal to zero, so the weight adjustment only depends on a constant and possibly other calibration variables such as age and gender, but not income. This matches the uniform adjustment profile (2) at the bottom of the distribution that we present in section 2.1.1. The calibration, by construction, avoids distorting the bottom of the distribution because it is not necessary to enforce the constraints of the calibration problem.

In practice, we also constrain the number of times the weights are expanded or reduced to avoid disproportionate adjustments to single observations already in the data set.[7] To that end, we first impose the condition that brackets with a $\theta(y)$ outside the boundary defined by $1/\alpha \leq \theta(y) \leq \alpha$ are automatically grouped into larger brackets, until they satisfy the range restriction. Then, we enforce that same condition within the calibration algorithm using the iterative procedure described in (see e.g. Singh and Mohl, 1996, method 5), known as linear calibration, the same procedure we use to enforce that weights remain above 1. The default limit we choose is $\alpha = 5$. Thus, in this case, no observation would have their weight multiplied by more than 5 times or less than 0.2 times.

2.3 Expanding the Support

After applying the methods of the previous sections, the survey should be statistically indistinguishable from the tax data. However, the precision that we get at the top of the income distribution may still be insufficient for some purposes. Indeed, the number of observations in the survey is still significantly lower than what we would get in theory from administrative microdata. The extent to which this represents a problem varies. For

[7] The practice is common in survey calibration: the rationale behind it is that strong adjustments of small subgroups lead to worse small-sample properties for the resulting survey data.

tasks such as regression analysis, the reweighting step is probably sufficient. But problems may arise if we wish to produce indicators of inequality, especially the ones that emphasize the top of the distribution, like top income shares. The combination of a low number of observations with fat-tailed distributions can create small sample biases for the quantiles and top shares (Okolewski and Rychlik, 2001; Taleb and Douady, 2015), and skewed distributions of the sample mean (Fleming, 2007). In most cases, we would underestimate levels of inequality.

Unlike problems caused by, say, heterogeneous response rates, these biases are part of *sampling error*. They do not reflect fundamental issues with the validity of the survey, but arise purely out of its limited sample size. The calibration method (section 2.2) does, to some extent, reduce sampling error. Yet it only does so under asymptotic conditions (Deville and Särndal, 1992) that cannot hold for narrow groups at the top of the income distribution. For this reason, we prefer to consider that the role of survey calibration in our methodology is to deal with *non-sampling error*. We use a different approach to deal with sampling error.

In particular, we aim to solve the case where tax statistics include income declarations beyond the survey's support. That is, we need to account for individuals declaring higher income than the richest persons in the surveys, which cannot be accounted for by reweighting observations. To do so, we start from the original tax tabulations, which were created from the entire population of taxpayers and should therefore be free of sampling error. We use it alongside a generalized Pareto interpolation to estimate a continuous income distribution (Blanchet, Fournier, and Piketty, 2017) that reproduces the features of the tax data with high precision. We then statistically match the information in the calibrated survey data with the tax data by preserving the rank of each observation.

More precisely: we inflate the number of data points in the survey by making k_i duplicates of each observation i. We attribute to each new observation the weight $q_i = w_i/k_i$, where w_i is the calibrated weight from the previous step. We choose $k_i = [\pi \times w_i]$ where $[x]$ is x rounded to the nearest integer. Therefore all new observations have an approximately equal weight close to $1/\pi$. The size of the new data set, made out of the duplicated observations, can be made arbitrarily high by adjusting π, yet any linear weighted statistic will be the same over both datasets.[8]

Let M be the number of observations in the new dataset. The weights are assumed to sum to the population size N. We will associate to each of them a small share $[0, q_{j_1}/N], [q_{j_1}/N, (q_{j_1} + q_{j_2})/N], \dots, [\sum_{k=1}^{M} q_{j_k}/N, 1]$ of the true population. If we attribute to each observation the average income of their population share in the tax data, then by construction the income distribution of the newly created survey will be the same as in the tax data. We rank observations in increasing order by income to preserve the joint distribution between income and the covariates in the survey. This approach is very similar to the one applied in (Burkhauser, Hahn, and Wilkins, 2016).

Intuitively, this process can be described as replacing the income of observations beyond the merging point with synthetic income observations with equivalent weight and rank in the tax distribution. This step ensures that we reproduce exactly the income distribution from tax data, that we preserve the survey's covariate distribution (including the household structure), and that we preserve the relationship between income and covariates from the survey data.

3 Applications

In section 3.1, we run Monte Carlo experiments in order to assess the accuracy of estimates produced after applying both our adjustment method and the common replacing alternatives found in the literature.[9] In section 3.2, we illustrate how the method operates with actual household surveys and tax statistics, applying it to data from five countries (France, United Kingdom, Norway, Brazil and Chile). These case studies showcase the

[8] Our command uses a default sampling rate of 10%, i.e. a value $\pi = 0.1$. In practice lower values are less demanding computationally for large countries and/or low merging points, with very limited impact in the results.

[9] We choose the replacing alternative as it is the most prevalent one which uses external data to correct surveys.

broad applicability of the method to both developed and less developed countries, covering different data quality.

3.1 Simulations

Our experiments start with the simulation of a “true” distribution with several million individuals. We emulate a typical tax tabulation, which summarizes information on the richest brackets of that distribution. We then draw a number of pseudorandom samples from the original distribution, simulating surveys of a given share of the population each time, which we adjust following both our method and replacing methods that are common in the literature.

We simulate both misreporting and nonresponse biases in all of our samples. The former is defined by a probability of misreporting that is flat for most of the distribution and increases linearly with the income rank for the top. The distribution of misreported income is also defined to ensure a prevalence of underreporting over overreporting. Response rates are flat for most of the distribution and fall linearly with the income rank for the top. In what follows, we present the results of our benchmark experiment. We present in appendix D alternative experiments, conducted under different sets of parameters and assumptions. These include alternative assumptions for each bias, variations in the replacing procedure, the size of the replaced population and the coverage of the simulated tax data. However, despite different — and sometimes extreme — assumptions, these experiments consistently demonstrate that our algorithm is robust and capable of implementing adjustments that push surveys closer to the true distribution when its right tail is biased.[10]

In our benchmark experiment, we study a population of 9 million individuals that are randomly drawn from a standard lognormal distribution. We select a thousand random

[10] All our experiments were conducted using the `bfmtoy` command that comes with the `bfmcorr` Stata package. Not only was it coded to be able to reproduce our experiments, but it also provides a tool for researchers to simulate artificial distributions and change all the parameters involved in testing survey adjustment methods.

subsamples from it, whose size corresponds to 1% of the total population. The response rate, conditional of being sampled, is 50% for most of the population; it then decreases from percentile 90 (P90) onward and tends to 0 for the richest individual, resulting in a general response rate of 47.5%. The probability of misreporting is 20% until the 95th percentile. It then increases, approaching 100% at the very top. The probability of misreporting is close to 22% on average and the distribution of misreported income is also a standard lognormal.

Concretely, in this experiment, all individuals in the simulated distribution have the same probability of being surveyed (1 in 100), yet individuals have their own probability of answering the survey and if they do answer, their response can be either accurate or misreported. Therefore, although the surveyed sample is 1% of the population, only close to 0.5% of the population effectively reports income. Figure 4 graphically depicts the setup of our benchmark experiment, for one of the random samples.

We apply our adjustment method, described in the previous section, and two alternative replacing procedure. The first replacing procedure that we consider simply consists in replacing the 1% of the survey distribution with the tax data: we call it the replacing method with a fixed merging point. We consider a second, more sophisticated procedure which we call replacing with an endogenous merging point. This procedure automatically selects a merging point by searching for the rank at which the quantiles of the survey and the tax data distribution cross.

One difficulty in implementing that last procedure is that, as we have stressed earlier, the existence of this "crossing point" is not theoretically guaranteed. In fact, under basic assumptions for modeling the bias — such as the one described in the experiment above — it does not exist. To be able to apply that method and compare it with in the others, we therefore have to add an additional form of bias to the tax data. We multiply each bracket of the tax data by a coefficient that grows linearly with the income rank from zero at the 50th percentile of the tax data distribution, to one at the 90th percentile. (It is equal to zero for the bottom 50% of the distribution, and one for the top 10%.) This coefficient models the effect of low-income exemption and taxable minimum that make the tax data less reliable below the top, and in fact explain why the “crossing point” tends to exist in practice. (We use that modified tax tabulation to test all methods.)

[Place figure 4 here]

[Place figure 5 here]

Figure 5 compares the accuracy of distributional estimates that result from the raw simulated survey to those resulting from the application of both our method and replacing methods. It displays errors with respect to true values for a series of estimates. Kernel densities present the distribution of measurements across all the 1000 replications. The true values are: an average close to 1.6, a Gini coefficient close to 0.52, a top 1% share close to 9.3% and a top 10% share close to 39%.

Our method’s estimates tend to be more accurate than others, in that they exhibit less bias (they are centered around the true value) and/or less variance (they are concentrated in a narrower range of values). The replacing method with an endogenous merging point performs better that the replacing method with a fixed merging point, especially in terms of variance. But it always shows more bias that our method.

Although all adjustment methods operate differently, in purely distributional terms they always reproduce at least the information on the top 1% that is found in tax data. That is, after applying the adjustment, the average income of the top percentile is identical in all cases. However, the same is not true for the rest of the distribution, and thus not for average income either. Indeed, figure 5a shows that even if the average income gets closer to the true value with replacing methods, on average it still remains underestimated by 15% (fixed merging point) and 5% (endogenous merging point), instead of 15% for the raw survey. The lower total income explains why in figure 5b, the top 1% shares are systematically overestimated with replacing methods: because the numerator of the top share is the same in both, but the denominator is underestimated. In the case of the top 10%, the two replacing methods show errors in opposite directions (figure 5c). This is because the fraction of the distribution corrected by the fixed merging point approach is too small (top 1% only), whereas with the endogenous merging approach that fraction is too large (usually more than the top 10%). When we focus on a synthetic indicator of inequality, such as the Gini coefficient, we find a similar hierarchy of estimates (figure 5d).

Panel 5e shows the distribution of the merging points' locations over the simulations, for both our method and the replacing approach with endogenous merging point (i.e., the point at which quantile functions cross). Our method chooses a merging point around the 90th percentile (although across replications it may in fact be chosen a few percentiles below or above), which is what we would expect given the parameters of pour simulation. The merging point for the other methods is generally lower, slight below the 85th percentile. This is consistent with our discussion of this approach in appendix A.3.

3.2 Real Data

Our method can be applied to all countries with survey micro data covering the entire population and tax data covering at least a fraction of it. We experiment with five real distributions: three European countries and two less-developed Latin American countries. For France, Norway and the United Kingdom, our analysis broadly covers the years 2004–2014. For Brazil, we cover 2007–2015 and for Chile we include the years 2009, 2011, 2013 and 2015. Survey data for the three European countries is from the EU-SILC, for Brazil from PNAD, and for Chile from CASEN. Tax data are sourced from the countries' tax office statistics. Our inequality indicators measure gross (pretax) income of individuals (before income taxes and employee social contributions), with the exception of the French series, which measures income after employee contributions, given that tax data incomes are net of these. Appendix E.1 and E.2 provide more details as to the data sources, data harmonisation, and concepts used.

3.2.1 Empirical Bias and Corrected Population

The Shape of the Bias Our method finds the merging point between surveys and tax data by comparing the population densities at specified income levels, as explained in section 2.1.2. To do so we first interpolate the fiscal incomes in the tabulation using the generalized Pareto interpolation developed by Blanchet, Fournier, and Piketty (2017), which allows for the expansion of the tabulated income values into 127 intervals.[11] Using the thresholds of these intervals, we can construct our key statistics: the rate of representativeness $\theta(y)$ and the cumulative rate of representativeness $\Theta(y)$ for individuals along the income distribution.

[Place figure 6 here]

Figure 6 presents depictions of the shape of the empirical bias within the tax data's "trustable span" for all countries for the latest available year. First of all, the shape of the bias we measure from the data is very similar to the one we expected from our theoretical exposition, as depicted in figure 3.[12] In particular, we always observe a convex shape in the top tail, to the right of the merging point. It thus appears that surveys tend to increasingly underrepresent incomes beyond a certain point in the distribution. For the more developed countries (Norway, France and the United Kingdom), the shape of the empirical bias $\theta(y)$ can be observed for a larger share of the population, due to the greater population coverage in tax data. This enables us to empirically test the validity of our assumptions regarding the rate of representativeness to the left of the merging point. We indeed observe on the left side of Figures 6a, 6b, and 6c, a general stability in the rate of representativeness, with averages trending above one. The extent and quality of tax data below the merging point in less developed countries are such that we cannot observe the same trends.[13]

[11] These comprise 100 percentiles from P0 to P100, where the top percentile (P99–100) is split into 10 deciles (P99.0, P99.1, . . . , P99.9-100), the top decile of the top percentile (P99.9–100) being split into ten deciles itself (P99.90, P99.91, . . . , P99.99-100), and so forth until P99.999. This interpolation technique, contrary to the standard Pareto interpolation, allows us to recover the income distribution without strong parametric approximations.

[12] See also figures E.1, E.2, E.3, E.4 and E.5 in the appendix.

[13] Tax enforcement issues affecting this portion of the distribution could be at play here, as well as the

The merging points found by our algorithm vary by country and by year, again revealing differences in data quality and coverage between them. The Chilean case (Figure 6e) provides an example of our program needing to extrapolate the shape of the bias to find the merging point (see appendix C for more details of this procedure). For this case we rely on parameters observed for Brazil (specifically, values for the elasticity of response to income) above its trustable span as inputs for the Chilean extrapolation.[14] The fit with the existing data seems to work quite well.[15]

Corrected Population Table 1 provides summary statistics on the population corrected by our method, again using the last available year for each country as illustrations.[16] A varying proportion of the total population is adjusted at the top of the survey distribution in each country (column [4] of Table 1), ranging from 5.9% in Chile to 0.05% in France for their most recent years.[17] This corresponds to the share of the population above the merging point in the two datasets. But in both cases, the overwhelming majority that population (over 90%) is located within the survey's support, rather than outside the survey's original support, suggesting that nonsampling issues related to heterogeneous response rates matter more than problems related to sampling for the size of the corrected population.

sharp difference in incomes between the top and the rest in these countries leading to higher inequality levels than developed countries.

[14] I.e., the value of the baseline elasticity of response to income, γ_1^*, extracted from the Brazilian data is -0.99.

[15] The empirical bias that is observed in previous years for all countries is presented in appendix E.3.1.

[16] See Appendix E.2.2 for other years.

[17] Across years there is less variation in this share, with Norway and particularly France being relative exceptions. In the French case, we believe the significant break in the series is due to the use of register data in EU-SILC alongside the household survey from 2008. Despite the EU-SILC survey making use of register data, the goal is not to over-sample the top of the distribution, but rather to improve the precision of responses.

In general, this step of the algorithm is a useful guide to assess the income coverage of surveys across countries. For instance, it appears on the basis of our analysis that the Brazilian surveys do a better job at capturing gross income, given the lower share of the underrepresented population, than the Chilean household surveys. Moreover, comparing France and the United Kingdom, it seems that sampling error is greater in the UK surveys, given the higher share of the population beyond the survey's maximum income that needs to be added. Non-sampling error itself is greatest in Chile, derived from the share of the corrected population found inside the survey's support.

[Place Table 1 here]

3.2.2 Results

We compare the results our approach with the one that simply consists in replacing the top 1%. The latter corresponds to the procedure reproduced in the simulation in section 3.1, whereby the top 1% of the survey distribution is directly replaced by the top 1% of the tax distribution. We present results on top 1% income shares, Gini coefficients and average incomes.[18]

[Place figure 7 here]

Top Income Shares In line with the improved income coverage that our method produces — by more accurately including upper incomes — estimates of the income concentrated at the top of the distribution are revised upwards in all countries. The size of the adjustment, however, varies by country. Figure 7 depicts this for the Top 1% share.[19] Brazil has the most extensive correction, with a top 1% share that increases

[18] Appendix E.3.2 presents results for other income groups in the distribution.

[19] The one exception to this upward correction is Norway in 2006 (see Figure 7b). However, this is likely due to a change in the local tax legislation affecting the distribution of business profits (Alstadsæter et al., 2016), as we explain in the text.

by about 10 percentage points every year (Figure 7d). Conversely, France and Norway experience relatively smaller adjustments, starting from relatively lower levels of inequality. In addition, Brazil offers the clearest illustration of the distinct trends in inequality that can emerge after making a correction to the survey's representation of income. While the raw survey depicts falling top income shares, the corrected survey distribution shows slightly increasing top shares. Distinct trends are also visible, albeit for shorter periods of time, in the other countries.

When we compare the size of the adjustment in Chile and Brazil (Figures 7d and 7e respectively), two highly unequal Latin American countries, the latter has a considerably higher adjustment. One of the reasons that could be behind this phenomenon is the fact that capital income, especially dividends, is better recorded in Brazilian tax statistics. Indeed, the Brazilian tax agency has relatively good means to verify the accuracy of capital income declarations (Morgan, 2018), while Chilean tax authorities are generally constrained by bank secrecy (Fairfield and Jorratt De Luis, 2016). In this case, the limited quality of Chilean tax statistics explains the smaller correction.[20]

Following the same rationale, the inclusion or exclusion of some types of income in a given data set can also affect the size of the correction. In the case of Norway, tax incentives started favoring the retention of corporate profits inside corporations after 2005, with the creation of a permanent dividends tax in 2006. This resulted in less dividend payments, and thus less income to be registered as personal income in tax data. The reform also gave strong incentives for higher-than-normal dividend payouts in 2005, which contributed to the sharp increase in top shares observed for this year (Aaberge and Atkinson, 2010; Alstadsæter et al., 2016). In Figure 7b, it can be clearly perceived that

[20] There is also a considerable difference between these countries' tax systems and their respective incentives. In Chile most dividends received by individuals are taxed, while in Brazil they are not. This, in addition to the fact that Chilean realized capital gains are mostly untaxed, provokes incentives towards the artificial retention of profits that are not as present in Brazil. This is why, in Chile, the imputation of undistributed profits to the distribution of personal income appears to be necessary when making international comparisons ("Top Incomes in Chile: A Historical Perspective on Income Inequality, 1964–2017").

the size of the adjustment appears to drop durably after this year. Additionally, it should be noticed that the Norwegian survey appears to be rather insensitive to this change, implying that dividends were badly represented before 2005. Other potential explanations for the difference in the size of adjustments could have to do with behavioral differences between populations across countries related to response rates and reporting accuracy.

The extent of the adjustment, by definition, depends directly on the shape of the bias that is observed in figure 6. Both the steepness of $\theta(y)$, when it is to the right side of the merging point, and the size of the corrected population (column 4 in table 1) are decisive factors for the size of such an increase. Another way to think about the size of the corrected population is to look at the size of the area between $\theta(y)$ and 1, to the right side of the merging point.

Finally, comparing between correction methods, we can observe — in line with our simulations — that the top 1% share is generally higher in the replacing scenario than in our method due to the fact that while the level of numerator income is equivalent in both settings, average incomes (the denominator) is underestimated in the former scenario, as we show further below.

[Place figure 8 here]

Gini Coefficients Figure 8 shows the time series of the Gini coefficients before and after the correction for all available years. Overall, we find a similar hierarchy of estimates, mirroring our simulations in the previous section — inequality is corrected upwards, more so in countries whose raw survey is not already matched with any administrative source, and to different degrees depending on the year, thus producing distinct trends. This is further evidence that surveys need to be adjusted if they are to better represent the income distribution, in the same manner as they are currently calibrated to better represent the distribution of various demographic variables. Again consistent with our simulations, the replacing procedure seems to undershoot inequality levels compared to our method, which more accurately accounts for higher nonresponse and misreporting at the top. An arbitrary correction of the top 1% is not enough to adjust the under-coverage of income coming from these errors. This is especially the case where the corrected population is larger than the arbitrarily chosen fractile, such as in Brazil and Chile (see Figure 6 and Table 1).

[Place figure 9 here]

Average Incomes As alluded to before, the average income of the top percentile using both correction methods is the same, which is higher than the level observed in the raw surveys. However, the crucial difference between the two methods is that the average income for the other groups in the population are not equal. In our method, the weight of persons with lower incomes are reduced, while the replacing method keeps the same average income for the bottom income groups. This subsequently produces differences in the overall average income of the population in both cases. Figure 9 depicts that our method increases the average income in the surveys in all countries, although with highly varying degrees of magnitude. In the lower-income countries, which have the highest corrections — Brazil and Chile — average incomes increase broadly by 30–50%, with the gap increasing over time. The higher-income countries in Europe experience lower corrections to their average incomes, with the orders of magnitude between them reproducing the rank of countries by size of correction in table 1 — the United Kingdom experiences a larger correction than Norway, which experiences a larger correction than France. Visibly, in figure 9a the gap between the average in the raw data and corrected data is reduced from 2008 onward on account of the reduction in the size of the survey bias coming from the methodological novelties (see table E.4 in appendix for further details).

Conclusion

The main objective of this paper is to provide a rigorous methodological tool that enables researchers to combine income or wealth surveys with administrative data in a simple and consistent manner. The main innovation of our approach resides in the endogenous determination of a “merging point,” i.e., the point at which we should start incorporating tax data information into the survey.

Our choice of merging point naturally emerges as the result of a few basic principles. First the rate of income representativeness (i.e., the ratio of the income survey density to the income tax data density) is a continuous function of income. Second, it decreases monotonously at the top. Third, we should avoid distorting the survey data at the bottom, for which the tax data provides no reliable information: therefore, the rate of representativeness should preferably be kept constant below the merging point. We explain that these three basic requirements are sufficient to unambiguously determine the location of the merging point. Therefore, our methodology provides a straightforward basis for directly estimating this key parameter, rather than choosing it arbitrarily.

Then, we complement our approach for choosing a merging point with a concrete, two-step survey correction method. First, we reweight survey observations using tools from standard survey calibration theory. It allows us to match the survey data to a histogram approximation of the tax data density, while ensuring the representativeness of the survey in terms of other socio-demographic variables such as gender, age, household type, etc. Second, we replace survey observations beyond the merging point by synthetic observations that exactly reproduce the distribution of income observed in the tax data, while keeping the distribution of covariates and their dependency with income according to the survey intact. In doing so, we get rid of the small-sample issues that prevent surveys from accurately describing the top tail of income distributions, retaining as much information as we can from the initial data. Crucially, these two steps preserve the microdata structure of the original survey, making it possible to work with different statistical units, or equivalence scales, as one would do with regular survey data.

We perform simulations that show that our approach improves the survey data, in that it reduces both the bias and the variance of inequality estimates. It also performs better than alternative methods. We apply our approach to several countries, both developed and less developed, and show that both levels and trends in inequality can be significantly revised.

We stress that we still face many obstacles in properly measuring inequality in many countries. First, tax data is not perfect, as it can be insufficiently enforced and subject to tax evasion (Alstadsæter, Johannesen, and Zucman, 2019). Second, surveys may also underrepresent the very poor in addition to the very rich, a problem that tax data cannot correct. Both these issues would suggest that current estimates are a lower bound on inequality. Third, our ability to correct surveys with tax data depends on our ability to match concepts between the two sources (both in terms of taxable income and statistical unit). When surveys do not offer enough details to adjust the income concept, or to identify tax units within households, our ability to combine them with tax data becomes more limited. These issues are complex and require detailed attention to definitions and the legislation of countries. But, to the extent that they can be overcome, our method should help practitioners reconcile surveys with administrative data.

Nonetheless, there are still open questions about the precise mechanisms behind the biases that we observe in surveys. We do not clearly know to date to what extent survey biases reflect different response rates or misreporting of income, and we do not seek to take a stand on this issue in this article. By primarily relying on the standard survey calibration framework, we adhere to a longstanding practice of statistical institutes, which interpret over- and underrepresentation issues in the surveys through the lens of nonresponse. For example, they interpret deviations between the age structure of the survey population and the age structure of the true population as the result of differential nonresponse, even though there is evidence that people who self-report their age can misreport it by wide margins (Preston, Elo, and Stewart, 1999; Newman, 2020).

We probably face similar issues for income: our correction inevitably interprets as "nonresponse" biases that are partly due to misreporting. When looking at the univariate distribution of income, the problem is secondary, as long as we match the distribution of the tax data at the top without introducing clear aberrations (discontinuous density, nonmonotonous rate of representativeness). Indeed, the margin that we use to perform the adjustment (the weight of observations, or their income value) has no consequence on the distribution we observe. However, the dependency between income and covariates is affected differently by nonresponse and misreporting. The issue with the misreporting model is that the real, stochastic misreporting process cannot be identified by anonymously comparing the survey and the tax data density, because it also affects the *ranking* of observations. Therefore, it is only by nominally matching survey observations with administrative records that this issue can actually be solved (an increasingly common practice). We have shown in our simulations that our method is robust to the inclusion of misreporting in the survey when it comes to measuring inequality. However, we do not pretend to be able to fix the measurement of the dependency between income and its covariates that arises from misreporting, and which impacts corrected and uncorrected surveys alike. The question of whether this represents a major flaw in our understanding of, say, the income distribution by age group, is a question that we leave for future research, and better data.

No single source is enough for accurately measuring inequality and its determinants. While administrative data benefits from its large sample sizes and its lack of sampling error, both the breadth and the depth of information that can be collected from surveys is unparalleled. At the same time, the presence of significant gaps between similar concepts in different data sources ought to be a cause for concern. Ultimately, we hope that efforts to reconcile the various sources will improve our understanding and use of both surveys and tax data. We think that the methodology we developed in this paper constitutes a step forward in this direction.

Acknowledgements

We gratefully acknowledge funding from the *Fundación Ramón Areces*, ERC (Grant 340831), Ford Foundation, INET (Grant INO14-00023) and from other partners of the World Inequality Lab. We thank Facundo Alvaredo, Yonatan Berman, François Bourguignon, Lucas Chancel, Mauricio De Rosa, Francisco Ferreira, Emmanuel Flachaire, Pablo Gutiérrez, Amory Gethin, Thanasak Jenmana, Nora Lustig, Brian Nolan, Thomas Piketty, Li Yang and Gabriel Zucman for helpful discussions of earlier versions of this paper, as well as participants at the 2018 INET seminar series at the University of Oxford, the May 2018 Workshop on harmonising surveys and tax data at the Paris School of Economics, the ECINEQ 2019 conference, the inequality seminar at ECLAC Santiago (2019), the STEP Seminar at the Université Paris 1 Panthéon–Sorbonne and the Stone Center of Inequality Research seminar at INSEAD.

References

Aaberge, Rolf and A. B. Atkinson (2010). “Top incomes in Norway”. In: *Top incomes: A Global Perspective*. Ed. by A. B. Atkinson and Thomas Piketty. Vol. 2. Oxford University Press, pp. 448–481.

Alstadsæter, Annette, Niels Johannesen, and Gabriel Zucman (2019). “Tax evasion and inequality”. In: *American Economic Review* 109.6, pp. 2073–2103. ISSN: 19447981. DOI: 10.1257/aer.20172043.

Alstadsæter, Annette et al. (2016). *Accounting for business income in measuring top income shares: Integrated accrual approach using individual and firm data from Norway.* Tech. rep. National Bureau of Economic Research.

Alvaredo, Facundo (2011). “A note on the relationship between top income shares and the Gini coefficient”. In: *Economics Letters* 110.3, pp. 274–277. DOI: 10.1016/j.econlet.2010.10.008. URL: http://dx.doi.org/10.1016/j.econlet.2010.10.008.

Atkinson, A. B. and Thomas Piketty (2007). *Top incomes over the twentieth century: a contrast between continental European and English-speaking countries.* Oxford University Press, p. 585. ISBN: 9780199286881. URL: https://global.oup.com/academic/product/top-incomes-over-the-twentieth-century-9780199286881?lang=en&cc=fr.

— (2010). *Top incomes: a global perspective.* Oxford University Press, p. 776. ISBN: 9780199286898. URL: https://global.oup.com/academic/product/top-incomes-9780199286898?cc=fr&lang=en&#.

Ayer, Miriam et al. (1955). “An Empirical Distribution Function for Sampling with Incomplete Information”. In: *Ann. Math. Statist.* 26.4, pp. 641–647. DOI: 10.1214/aoms/1177728423. URL: https://doi.org/10.1214/aoms/1177728423.

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty (2017). “Generalized Pareto Curves: Theory and Applications”.

Bourguignon, François (2018). “Simple adjustments of observed distributions for missing income and missing people”. In: *The Journal of Economic Inequality*, pp. 1–18.

Brunk, H D (1955). “Maximum Likelihood Estimates of Monotone Parameters”. In: *Ann. Math. Statist.* 26.4, pp. 607–616. DOI: 10.1214/aoms/1177728420. URL: https://doi.org/10.1214/aoms/1177728420.

Burkhauser, Richard V, Markus H Hahn, and Roger Wilkins (2016). “Top Incomes and Inequality in Australia: Reconciling Recent Estimates from Household Survey and Tax Return Data”.

Burkhauser, Richard V et al. (2016). “What has Been Happening to UK Income Inequality Since the Mid-1990s? Answers from Reconciled and Combined Household Survey and Tax Return Data”. URL: http://www.nber.org/papers/w21991.

Chancel, Lucas and Thomas Piketty (2017). “Indian income inequality, 1922-2014: From British Raj to Billionaire Raj?” URL: http://wid.world/document/chancelpiketty2017widworld/.

Czajka, Léo (2017). “Income Inequality in Côte d’Ivoire: 1985-2014”. In: *WID.world Working Paper* July.

DWP (2015). “Households Below Average Income: An analysis of the income distribution 1994/95 – 2013/14”. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/437246/households-below-average-income-1994-95-to-2013-14.pdf.

Deville, Jean-Claude and Carl-Erik Särndal (1992). “Calibration Estimators in Survey Sampling”. In: *Journal of the American Statistical Association* 87.418, pp. 376–382. DOI: 10.1080/01621459.1992.10475217.

Diaz-Bazan, Tania (2015). “Measuring Inequality from Top to Bottom”. In: *Policy Research Working Paper* 7237.

Eeden, Constance van (1958). “Testing and Estimating Ordered Parameters of Probability Distributions”. PhD thesis. University of Amsterdam.

Fairfield, Tasha and Michel Jorratt De Luis (2016). “Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile”. In: *Review of Income and Wealth* 62, S120–S144.

Fleming, Kirk G (2007). "We're Skewed—The Bias in Small Samples from Skewed Distributions". In: *Casualty Actuarial Society Forum* 2.2, pp. 179–183.

Flores, Ignacio et al. "Top Incomes in Chile: A Historical Perspective on Income Inequality, 1964–2017". In: *Review of Income and Wealth* 0.0 (). DOI: 10.1111/roiw.12441. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12441. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12441. Forthcoming.

Hlasny, Vladimir and Paolo Verme (2017). "The impact of top incomes biases on the measurement of inequality in the United States".

— (2018). "Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data Vladimir". In: *Econometrics* 6.30, pp. 1–38. DOI: 10.3390/econometrics6020030.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion (2006). "Survey nonresponse and the distribution of income". In: *Journal of Economic Inequality* 4.1, pp. 33–55. ISSN: 15691721. DOI: 10.1007/s10888-005-1089-4.

Kuznets, Simon (1953). *Shares of Upper Income Groups in Income and Savings.* NBER. ISBN: 087014054X. DOI: 10.2307/2343040. URL: http://www.jstor.org/stable/10.2307/2343040?origin=crossref.

Medeiros, Marcelo, Juliana de Castro Galvão, and Luìsa de Azevedo Nazareno (2018). "Correcting the Underestimation of Top Incomes: Combining Data from Income Tax Reports and the Brazilian 2010 Census". In: *Social Indicators Research* 135.1, pp. 233–244.

Morgan, Marc (2018). "Essays on Income Distribution: Methodological, Historical and Institutional Perspectives with Applications to the Case of Brazil (1926–2016)". PhD Dissertation in Economics. Paris: Paris School of Economics & EHESS.

Newman, Saul Justin (2020). "Supercentenarian and remarkable age records exhibit patterns indicative of clerical errors and pension fraud". In: *bioRxiv.* DOI: 10.1101/704080. URL: https://www.biorxiv.org/content/early/2020/05/03/704080.

Novokmet, Filip, Thomas Piketty, and Gabriel Zucman (2018). "From Soviets to oligarchs: inequality and property in Russia 1905-2016". In: *The Journal of Economic Inequality* 16.2, pp. 189–223.

Okolewski, Andrzej and Tomasz Rychlik (2001). "Sharp distribution-free bounds on the bias in estimating quantiles via order statistics". In: *Statistics and Probability Letters* 52.2, pp. 207–213. ISSN: 01677152. DOI: 10.1016/S0167-7152(00)00242-X.

Pareto, Vilfredo (1896). *Écrits sur la courbe de la répartition de la richesse.*

Piketty, Thomas (2003). "Income Inequality in France, 1901–1998". In: *Journal of Political Economy* 111.5, pp. 1004–1042. DOI: 10.1086/376955. URL: http://www.journals.uchicago.edu/doi/10.1086/376955.

Piketty, Thomas and Emmanuel Saez (2003). "Income Inequality in the United States, 1913–1998". In: *Quarterly Journal of Economics* CXVIII.1.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2017). "Capital Accumulation, Private Property and Rising Inequality in China, 1978-2015". URL: http://www.nber.org/papers/w23368.pdf.

Preston, Samuel H., Irma T. Elo, and Quincy Stewart (1999). "Effects of age misreporting on mortality estimates at older ages". In: *Population Studies* 53.2, pp. 165–177. ISSN: 14774747. DOI: 10.1080/00324720308075.

Singh, A C and C A Mohl (1996). "Understanding Calibration Estimators in Survey Sampling". In: *Survey Methodology* 22.2, pp. 107–115.

Taleb, Nassim Nicholas and Raphael Douady (2015). "On the super-additivity and estimation biases of quantile contributions". In: *Physica A: Statistical Mechanics and its Applications* 429, pp. 252–260. ISSN: 03784371. DOI: 10.1016/j.physa.2015.02.038. URL: http://dx.doi.org/10.1016/j.physa.2015.02.038.

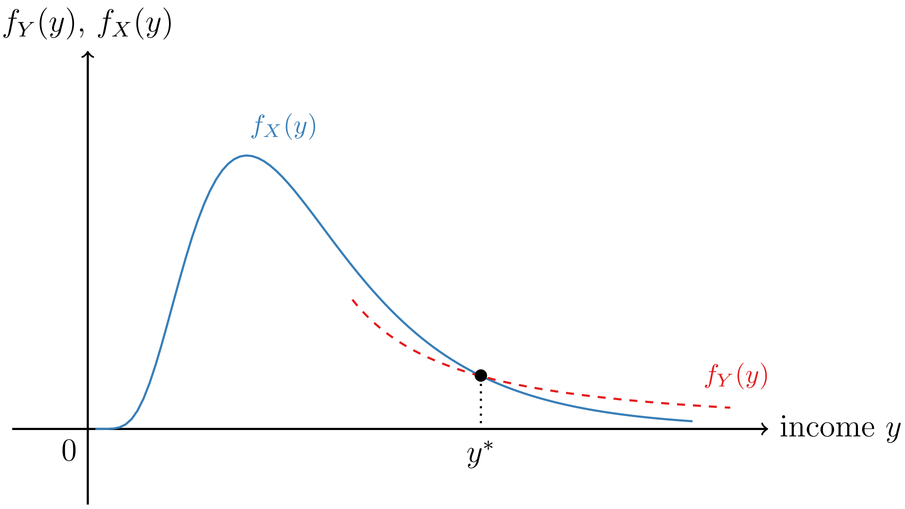
Tables and Figures

Table 1: Structure of Corrected Population: Latest Year

Country	Population over Merging Point (% total population)		Corrected population		
	Tax data	Survey	Total	Share inside survey support	Share outside survey support
	[2]	[3]	[4] = [2] − [3]	[5]	[6]
Chile	17.0%	11.1%	5.9%	99.99%	0.01%
Brazil	8.0%	5.3%	2.7%	99.0%	1.0%
UK	3.0%	2.5%	0.5%	93.6%	6.4%
Norway	5.0%	4.6%	0.4%	96.0%	4.0%
France	0.1%	0.05%	0.05%	99.0%	1.0%

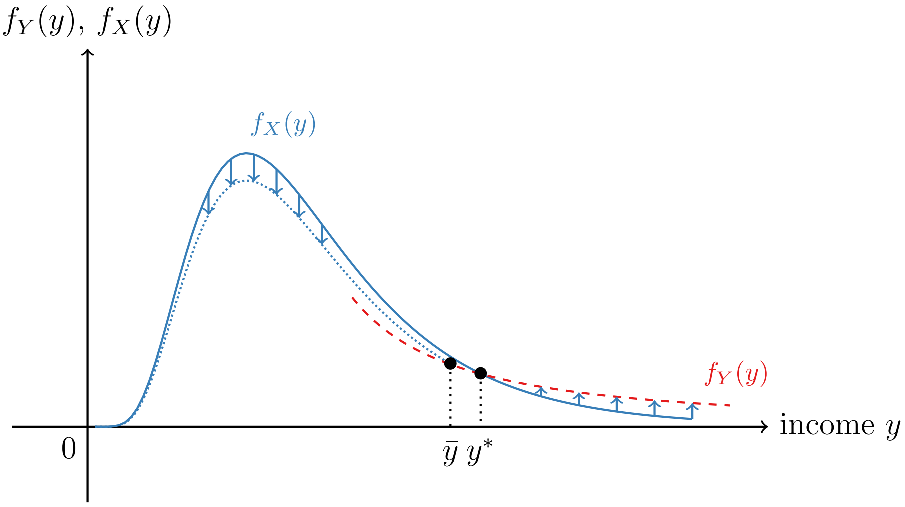
Notes: The table orders countries by the size of the corrected population. Column [2] shows the proportion of the population that is above this merging point in the tax data. Column [3] shows the proportion that is above the merging point in survey data. The difference between the two is the proportion of the survey population that is corrected (Column [4]). As explained in the text, we adjust survey weights below the merging point by the same proportion. The corrected proportion above the merging point can be decomposed into the share of the corrected population that is inside the survey support (up to the survey's maximum income) and the share that is outside the support (observations with income above the survey's maximum). Brazil and Chile refer to 2015, while all the European countries refer to 2014.

Figure 1: A "true" and biased income distribution



The solid blue line represents the survey density f_X. The dashed red line represents the tax data density f_Y, which is only observed at the top. For high incomes, the survey density is lower than the tax data density, which means that high incomes are underrepresented. If some individuals are underrepresented, then others have to be overrepresented: they correspond to people below the point y^*.

Figure 2: The intuition behind reweighting



The solid blue line represents the survey density f_X. The dashed red line represents the tax data density f_Y. Above the merging point $\bar{y}$, the reweighted survey data have the same distribution as the tax data (dashed red line). Below the merging point, the density has been uniformly lowered so that it still integrates to one, creating the dotted blue line.

Figure 3: Choice of Merging Point when $\bar{y} \geq y_{\text{trust}}$

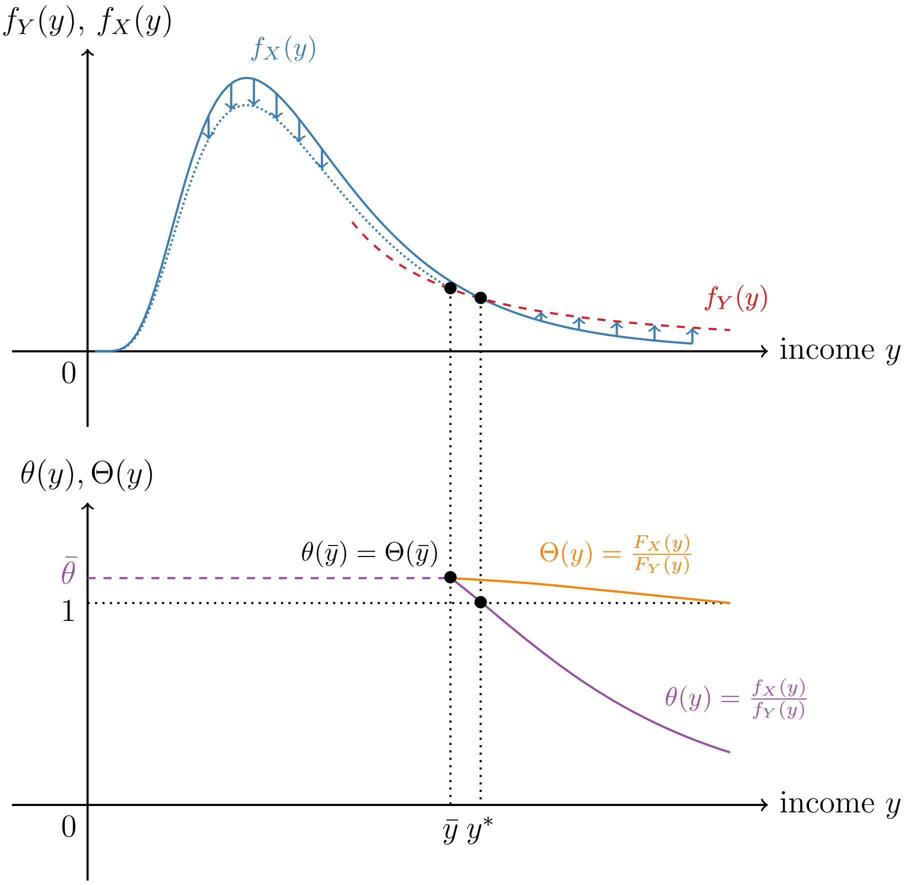


Figure 4: Benchmark Experiment Set Up

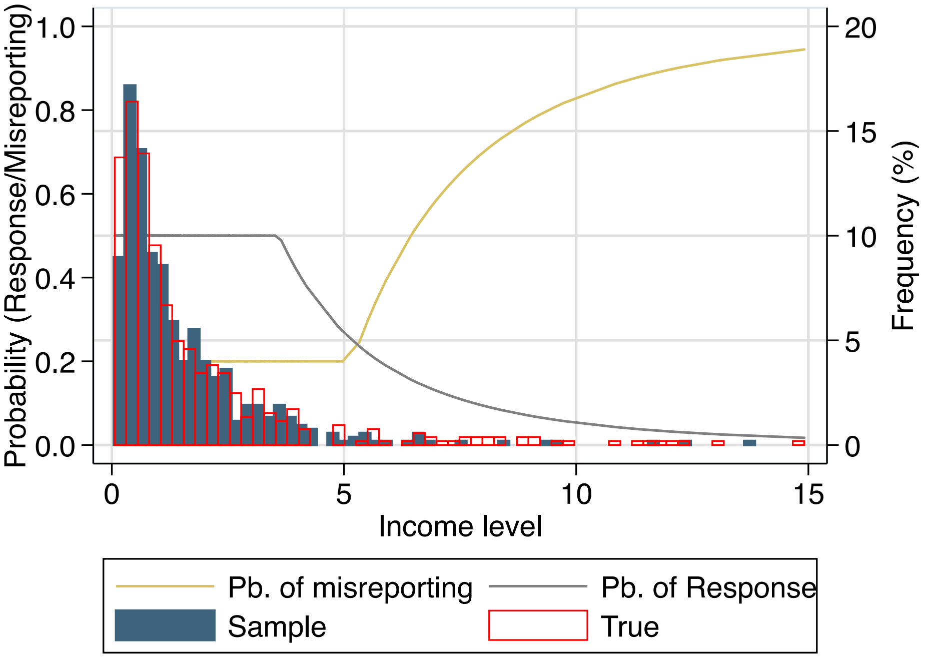
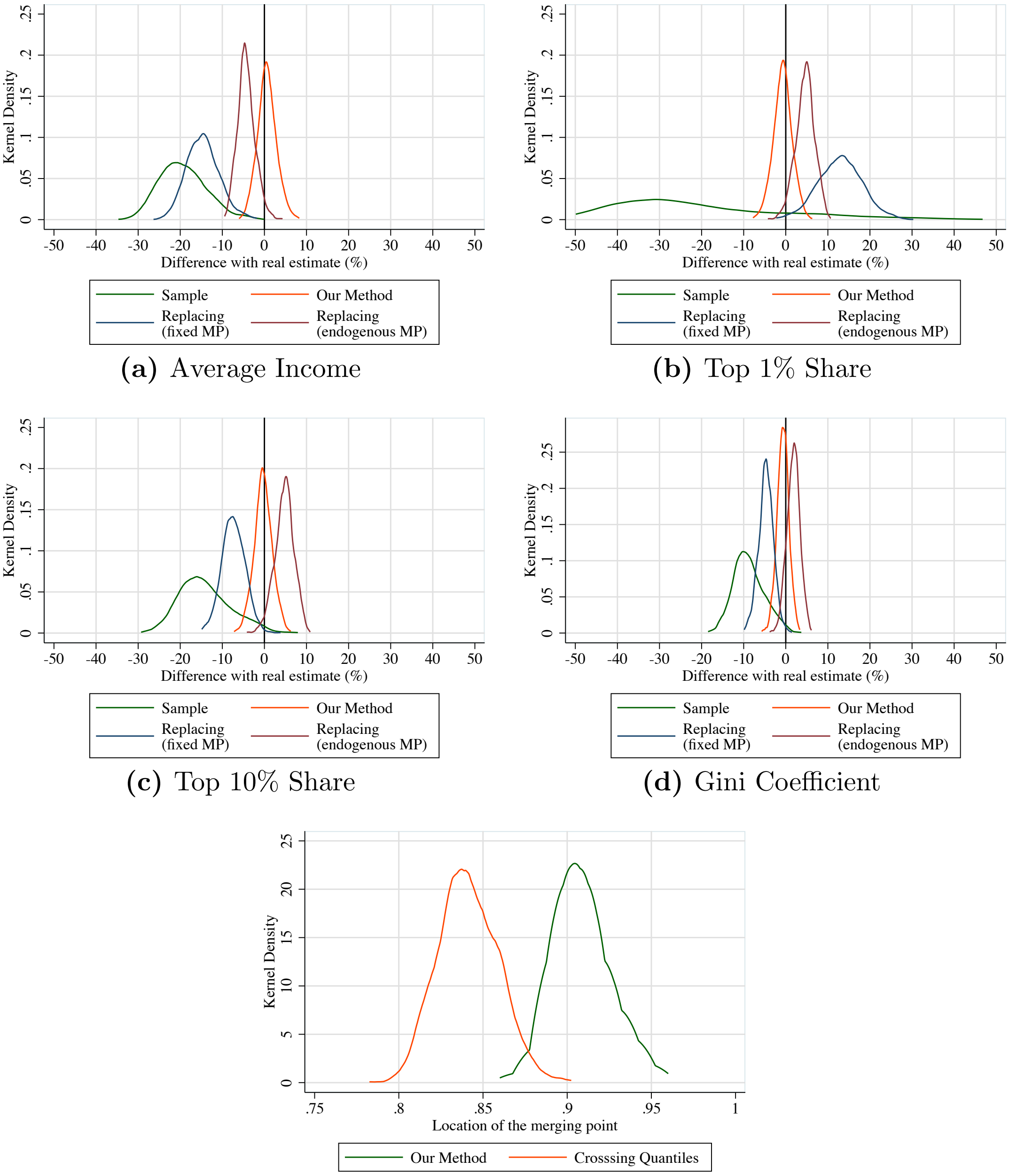


Figure 5: Benchmark Experiment Results



(a) Average Income

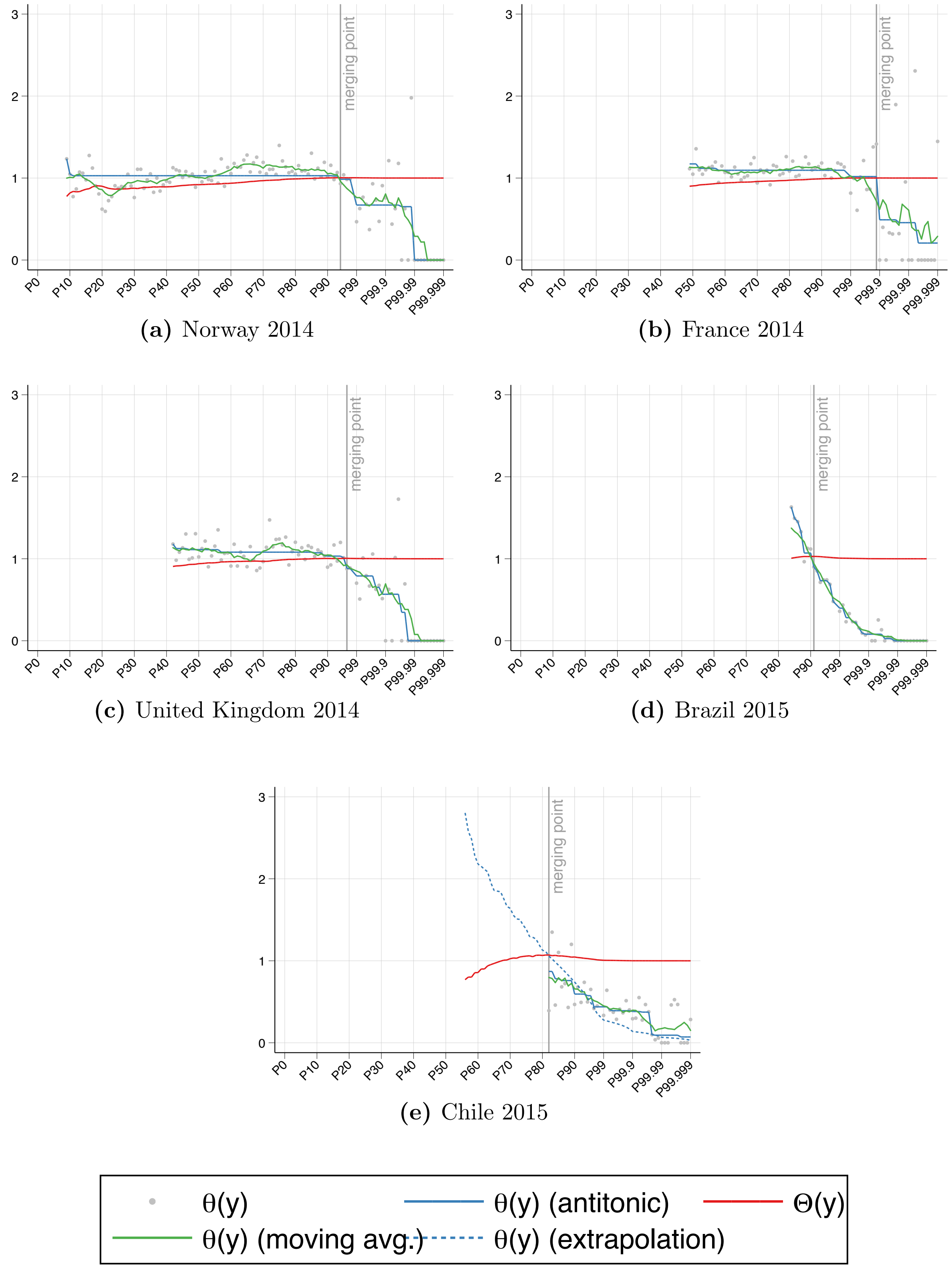
(b) Top 1% Share

(c) Top 10% Share

(d) Gini Coefficient

(e) Location of the Merging Point

Figure 6: Merging Point in 5 Countries, Latest year



(a) Norway 2014

(b) France 2014

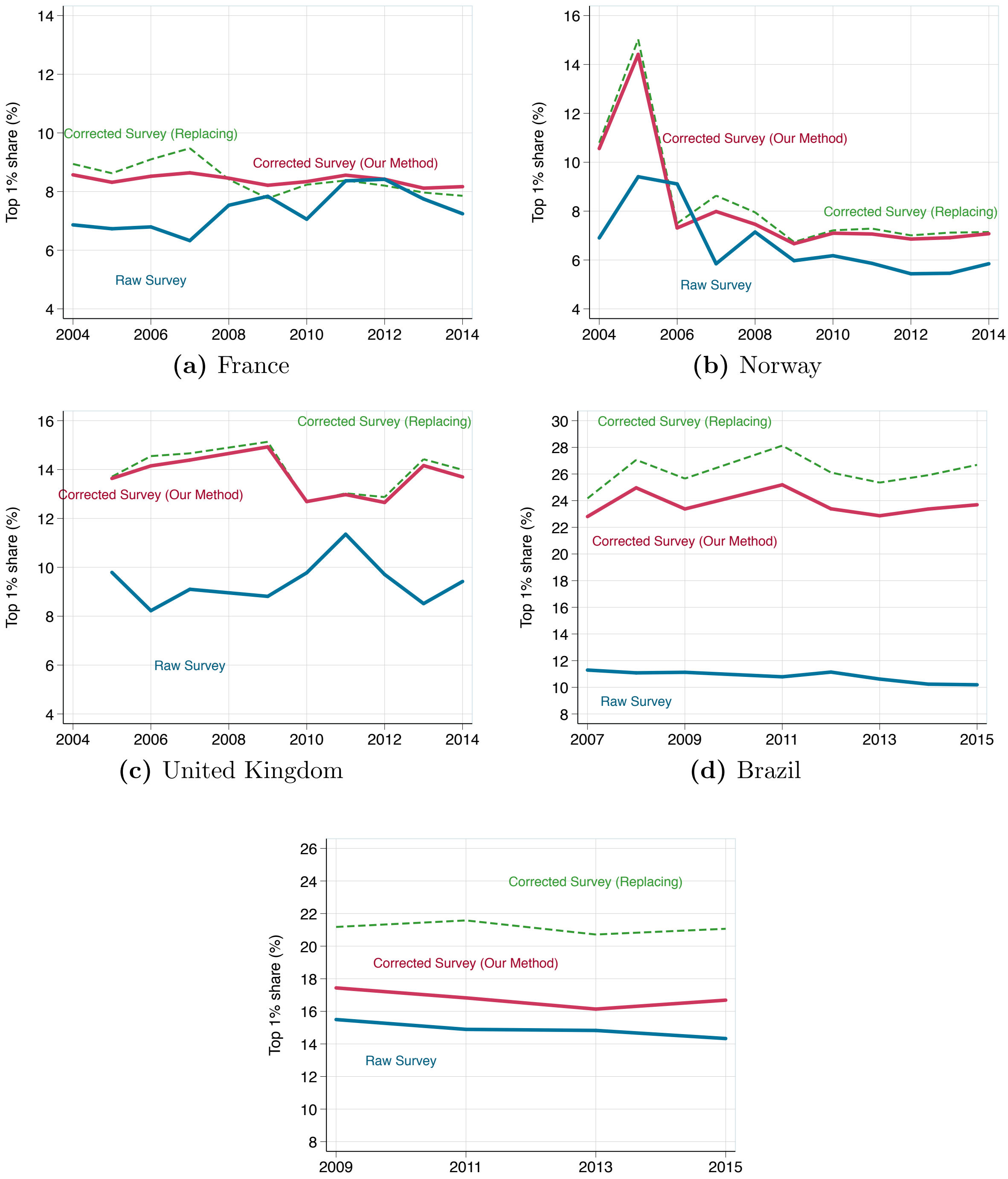
(c) United Kingdom 2014

(d) Brazil 2015

(e) Chile 2015

Notes: the figures depict the estimated bias in the survey relative to the tax data. Gray dots are, for each quantile of the fiscal income distribution, the ratio of income density in the survey over that of tax data. The green line is the centered average of $\theta(y)$ at each quantile and eight neighboring estimates. The blue line is the result of an *antitonic* regression applied to $\theta(y)$. It is constrained to be decreasing as it is used to find a single merging point. The blue dotted line, which only appears in figure 6e,is an extrapolation of the trend described by $\theta(y)$ based on a *ridge* regression (see appendix C). The red line is the ratio of the cumulative densities. For details refer to section 2.1.2.

Figure 7: Top 1% Shares Before and After Correction



(a) France

(b) Norway

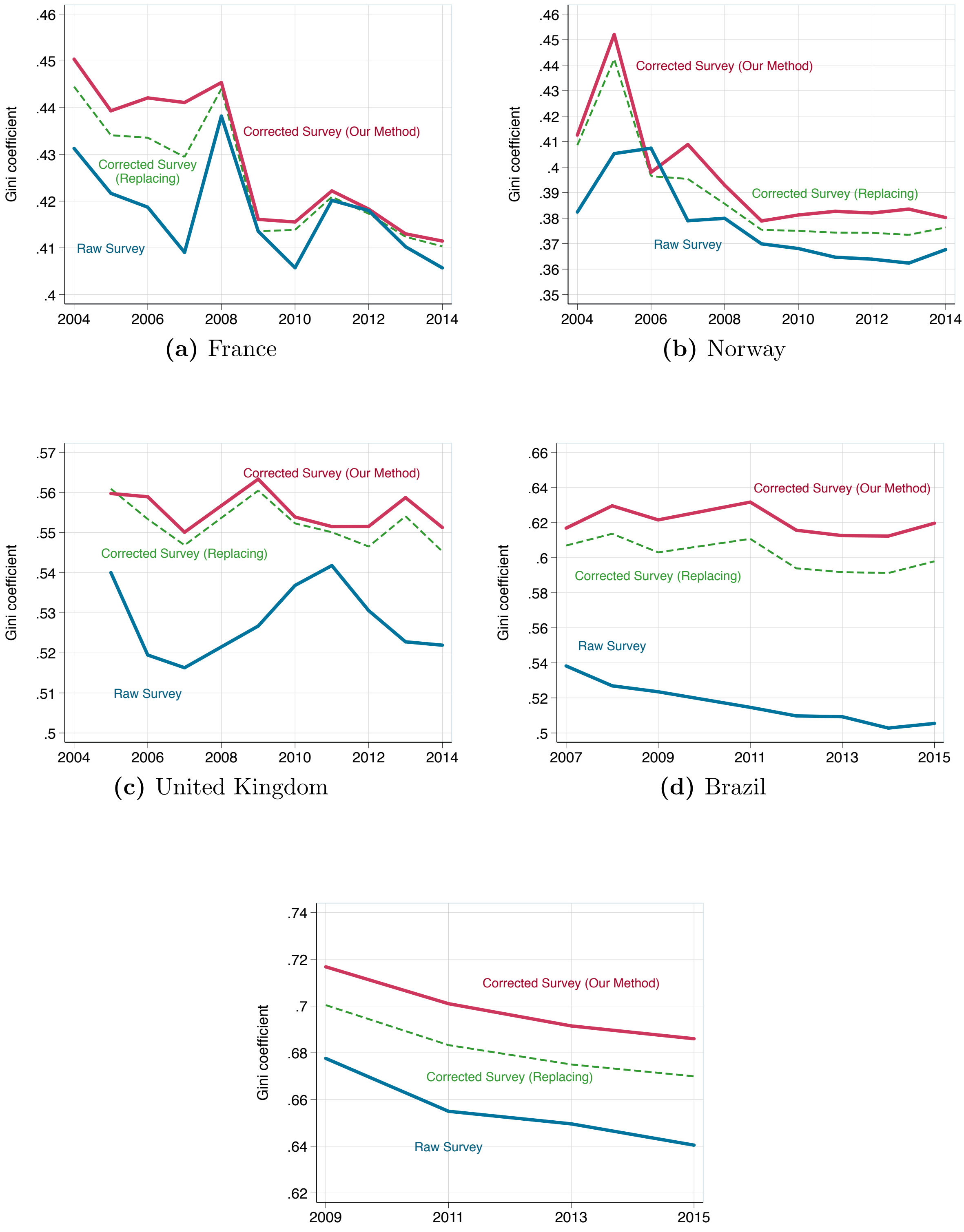
(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution.

Figure 8: Gini Coefficients, Before and After Correction



(a) France

(b) Norway

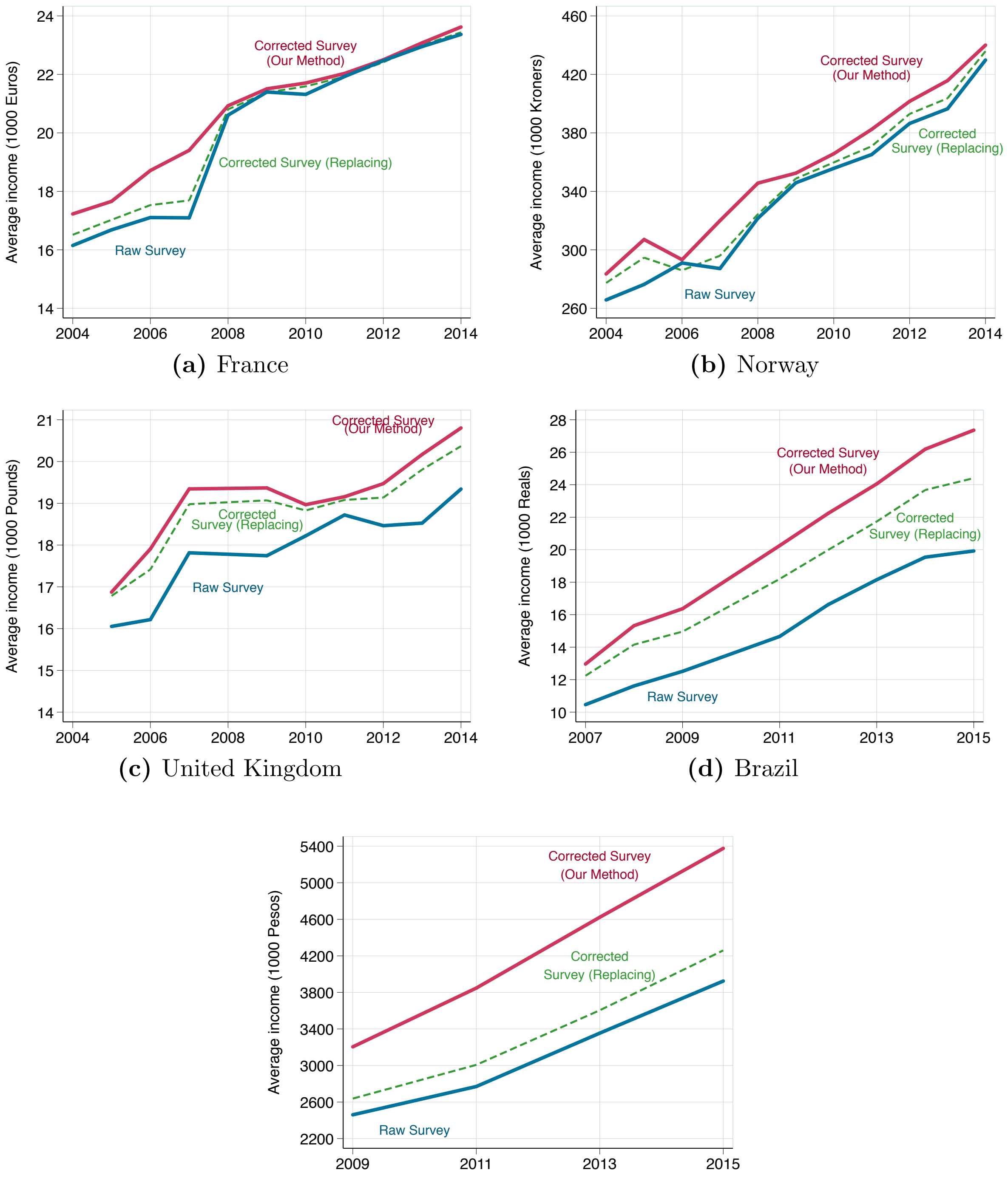
(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution.

Figure 9: Average Incomes Before and After Correction



(a) France

(b) Norway

(c) United Kingdom

(d) Brazil

(e) Chile

Notes: the corrected survey using the replacing method directly replaces the survey distribution above P99 with the distribution above P99 from the tax distribution. Average incomes are rescaled accordingly.