Including the Rich in Income Inequality Measures: An Assessment of Correction Approaches

Nora Lustig (nlustig@tulane.edu) and Andrea Vigorito (andrea.vigorito@fcea.edu.uy) *

February 7, 2025

ABSTRACT

Inequality measures based on household surveys may be biased because they typically fail to capture incomes of the wealthy properly. The "missing rich" problem stems from several factors, including sampling errors, item and unit nonresponse, underreporting of income, and data preprocessing techniques like top coding. This paper presents and compares prominent correction approaches to address issues concerning the upper tail of the income distribution in household surveys. Correction approaches are classified based on the data source, distinguishing between those that rely solely on within-survey information and those that combine household survey data with external sources. We categorize the correction methods into three types: replacing, reweighting, and combining reweighting and replacing. We identify twenty-two different approaches that have been applied in practice. We show that both levels and trends can be quite sensitive to the approach and provide broad guidelines on choosing a suitable correction approach.

Key words: income inequality, top incomes, household surveys, correction methods, tax records

JEL Classification: C18, C81, C83, D31

*Nora Lustig is Samuel Z. Stone Professor of Latin American Economics and founding director of the Commitment to Equity Institute at Tulane University. She is also a nonresident senior fellow at the Brookings Institution, the Center for Global Development, the Georgetown Americas Institute, the Inter-American Dialogue, the Paris School of Economics and the CUNY Stone Center on Socio-Economic Inequality. Andrea Vigorito is a researcher at Instituto de Economia, Facultad de Ciencias Economicas, Universidad de la Republica. The authors are very grateful to François Bourguignon, Emmanuel Flachaire, Stephen Jenkins, Graciela Sanroman and Paulo Verme for providing useful clarifications. We thank Ali Enami for sharing the derivation included in Appendix 4.

Table of Contents

1 Introduction

While household surveys have traditionally been used to measure personal income inequality, they have a serious limitation: they often fail to accurately capture income in the upper tail of the distribution, especially income derived from capital. We call this issue the “missing rich” problem.[1] The missing rich problem here refers to the main factors that affect the upper tail of the income distribution in household surveys, including sampling errors, coverage errors, unit and item nonresponse, underreporting and preprocessing practices by data providers, such as top coding (Atkinson and Micklewright, 1983; Cowell and Flachaire, 2007; Korinek, Mistiaen and Ravallion, 2006 and 2007; Atkinson and Piketty, 2007; Biemer and Christ, 2008; Jenkins, 2017; Bourguignon, 2018; Ravallion, 2021).[2] These factors can lead to biased survey-based income distribution and inequality indicators which may affect not only the level of inequality but also its trends, potentially leading to inaccurate estimates, for example, of the relationship between inequality and economic growth.

With the renewed interest in economic inequality, there has been a corresponding interest in addressing the 'missing rich' problem (Atkinson and Piketty, 2007; Milanovic, 2023).[3] This concern has spurred various approaches to generate inequality measures that more accurately capture the upper tail of the income distribution.[4] The approaches proposed in the literature fall into three main strands. The first strand corrects household surveys, either by using within-survey methods (Cowell and Flachaire, 2015) or by combining them with external data (Jenkins, 2017; Bourguignon, 2018). The second relies mainly on external data sources, such as tax records (Atkinson and Piketty, 2009, 2011).

---

[1] Other terminology has been used. The special issue by the Journal of Economic Inequality dedicated to the subject calls is "upper tail" issues (JOEI, 2022). Jenkins (2017), for example, refers to the problem as “under-coverage” of the rich. Here we will use missing rich, upper tail issues, under-coverage, top incomes problems interchangeably.

[2] In the United States, the Census Bureau top codes income to protect the confidentiality of high-income individuals, with 4.6 percent of individuals living in households where some income was top-coded in 2004 (Burkhauser et al., 2012).

[3] In the mainstream academic literature, particularly in the US, the UK, and France, there has been a renewed interest in economic inequality, as evidenced by the extensive list of publications on the subject. Two early and iconic publications are noteworthy: Atkinson's "Bringing Income Distribution in From the Cold" (Atkinson, 1997) and the first handbook on income distribution by Atkinson and Bourguignon (2000).

[4] At the time of this writing there were close to six hundred published articles and edited volumes.

The third strand allocates all income categories in National Accounts to households to generate a distribution of income consistent with macroeconomic aggregates, a methodology known as distributional national accounts (Zwijnenburg, 2019; Blanchet et al., 2024).

Here, we review and assess approaches aimed at correcting household surveys to better represent the rich in inequality measures.[5] To the best of our knowledge, this may be the first comprehensive analytical survey on the subject.[6] Our focus is on methods utilizing household surveys—either by themselves or in conjunction with external data—as these have been the subject of the most extensive methodological development. Moreover, household surveys remain the primary source of data in academic research and policymaking related to income inequality and poverty, providing detailed information at the household level that is not commonly available in administrative data such as social security or tax records.[7]

The absence of high-income individuals in household surveys is often evident by inspection. For example, Szekely and Hilgert (2007) found that the income of the ten richest households in a sample of surveys for Latin America was roughly equal to the average wage of a manager at a medium to large-sized firm--or even less.[8] In Vietnam, the top salaries recorded in the survey were less than half the average executive salaries recorded in corporate salary surveys (World Bank, 2014). In Egypt, the median annual total pay of CEO's was twice as high as the median income of the top .05 percent in the household survey (van der Weide, Lakner and Ianchovichina, 2018).

---

[5] In the online Bibliographical Appendix, we provide a list of all the articles identified.

[6] A previous survey by Lustig (2019) focused on a subset of correction methods to include the rich in household surveys.

[7] As stated by Altimir (1987): "Income and expenditure surveys have long been regarded as the main source for measuring household income and its distribution, since they provide the technical means (i) to investigate income received from all sources, in cash or in kind, by each member of the household, (ii) to impute or check income in kind through the corresponding consumption, (iii) to impute the rent of owner-occupied dwellings and (iv) to differentiate between current income and other financial flows." (pp. 126)

[8] Data from the 2000s showed that the richest two households' monthly incomes in surveys for Argentina, Brazil, Mexico, and Peru were equal to roughly $14,000, $70,000, $43,000, and $17,500, respectively, a rather low figure in a region with reportedly 4,400 individuals with a net worth of 30 million dollars or more (Capgemini and Merrill Lynch, 2011).

Comparisons with other data sources, such as National Accounts, social security registries or tax records, further highlight the underrepresentation of high incomes in household surveys. For example, in the U.S. in 2006, Atkinson, Piketty and Saez (2011) found that the share of total income held by the top 1 percent of adult individuals was 13.7 percent based on survey data, compared to 18 percent from tax returns. Jenkins (2017) showed that the income of the 99.5th percentile of adult individuals in the U.K. household survey was as low as 77 percent of the equivalent in tax data, depending on the year. In 2010, the survey-based average fiscal income of the top 1 percent was 63 percent of its value in tax records in Germany (Yonzan et al., 2022).[9] In China in 2015, the share of the top 10 percent of the population based on tax records and National Accounts was 41 percent compared to 31 percent obtained from survey data (Piketty, Yang, and Zucman, 2016). In Argentina, tax data for the year 1997 showed 698 individuals with incomes above $1 million dollars, whereas there was not a single individual with such an income in the household survey (Alvaredo, 2007). The exclusion of the rich in household surveys may partly explain the significant discrepancy between survey-based per capita household income and National Accounts estimates, especially in low- and middle-income countries (Altimir, 1987; Deaton, 2005; Bourguignon, 2018). [10]

There is also evidence that nonparticipation in surveys (unit nonresponse) is not randomly distributed across the population. For instance, using information on nonresponse rates by Primary Sampling Unit (PSU) for the US Current Population Survey, Korinek, Mistiaen, and Ravallion (2006; 2007) found that unit nonresponse rates were positively correlated with income. Hlasny and Verme (2018a, 2018b, 2021) found the same for Egypt, the European Union and the United States.

[9] Fiscal income is defined as the income observable in tax returns.

[10] For example, with data for Latin America Bourguignon (2015) found that between 2000 and 2012, the ratio of mean income in household surveys to mean household income per capita in National Accounts could be significantly lower than one. Depending on the year, the ratio ranged from 0.78 to 0.84 in Brazil; 0.50 to 0.71 in Colombia; 0.47 to 0.87 in Ecuador; 0.67 to 0.81 in Peru; and 0.69 to 0.84 in Uruguay. In Mexico, the ratio was the lowest: between 0.42 and 0.49 (!). Deaton noted that population-weighted survey consumption in non-OECD countries grew at only half the rate of population-weighted consumption in the Penn World Tables (Deaton, 2005, p. 10). See the pioneer work on this by Altimir (1987). Also, see Fesseau and Mantonetti (2013) and Alvaredo et al. (2018).

The studies based on linked data (where one can observe reported income for the same individual in the survey and the external administrative source) provide direct evidence that the upper tail of household surveys disproportionately misses some of the income from top earners.[11] Linked data studies show that high income earners tend to underreport their earnings in surveys. This pattern has been documented in, for example, the U.S. (Lillard, Smith, and Welch, 1986; Abowd and Stinton, 2013; Bollinger et al., 2019), Sweden (Kapteyn and Ypma, 2007), New Zealand (Hyslop and Townsend, 2020), the U.K. (Jenkins and Rios Avila, 2020), and Uruguay (Flachaire et al., 2023).[12] There is also some evidence that the rate of item (income) nonresponse is higher at the top. In the U.S., Bollinger et al. (2019) found that, on average, 20 percent of linked survey respondents did not report earnings, with nonresponse rates climbing to 24 percent in the bottom 5 percent and 25 percent in the top 5 percent.

Given the widespread use of household surveys in distributional, multivariate, and fiscal incidence analyses, this article describes and assesses the methods developed in the literature to include the rich in survey-based measures of inequality. We begin by outlining the factors contributing to the "missing rich" problem. We then present and compare prominent correction approaches, discussing their strengths, limitations, and practical applications.

Correction approaches typically involve within-survey methods or combining surveys with external data such as tax records, social security registries, or National Accounts. In terms of methodology, correction approaches can be categorized into three main types: replacing, reweighting, and reweighting and replacing combined. Under the replacing method, the original income observations at the top of the income distribution in the uncorrected survey are replaced with a (presumably) more accurate estimate of the upper tail, while the original weights assigned to the top as a whole and to the rest of the distribution remain intact. Reweighting adjusts the weights assigned to different income groups in the survey to (in principle) better reflect the true representation of high-

[11] The external data usually refers to tax records, social security registries, or employer-provided payroll information.

[12] In Uruguay, for example, individuals in the top 1 percent reported 60 percent less income in the household survey than in tax data on average (Flachaire, Lustig, and Vigorito, 2023).

income individuals, without altering the original income observations. The combined approach modifies both the income observations and their associated weights. Within these three broad categories, there are various specific methods, including semiparametric and nonparametric replacement methods, and model based and ad hoc reweighting methods. In total, we have identified twenty-two approaches that have been applied in practice.

Correcting household surveys predictably alters inequality estimates. This should come as no surprise. It is important to note that while including the rich typically increases measured inequality, this isn't always the case, as Deaton (2005) points out. Particularly when correcting for coverage errors or unit nonresponse in household surveys, or when using within-survey semiparametric methods, some inequality indicators (e.g., the Gini coefficient) may be lower than the Gini coefficient estimated from the original survey.

The evidence shows that inequality indicators change, sometimes significantly, after correcting for the missing rich.[13] For example, for the US. in 2006, the Gini was 0.470 for survey data and 0.519 when the survey was corrected using tax data (Atkinson, Piketty, and Saez, 2011). In the UK. in 2009, the survey-based Gini was estimated 0.446 while it was equal to 0.466 after the survey was corrected using tax data (Jenkins, 2017). In South Korea, the Gini coefficient for 2011 rose from to 0.308 to 0.371 (Kim, 2014). De Rosa, Flores, and Morgan (2024) found that correcting household surveys with tax data increased the Gini coefficient by 10 percentage points on average for ten Latin American countries between 2000 and 2020.

Top income shares also differ depending on the data source. For example, in Germany, correcting the survey for the missing rich using tax data raised the top 1 percent's income share from 7 percent to 11.2 percent (Bach, Corneo, and Steiner, 2009). In Latin America, De Rosa, Flores, and Morgan (2024) found that correcting household surveys with tax

[13] There is numerous inequality measures proposed in the literature. The most prevalent in research and in international databases is the Gini coefficient. Another frequently used indicator in the literature, especially in studies focused on "top incomes," is income shares. This approach involves estimating the proportion of total income held by different segments of the population, such as the top 1, 5, or 10 percent. The studies reviewed here present either the Gini coefficient, top income shares, or both.

data led to significantly higher top income shares: for Brazil and Mexico in 2020, for example, the share was 10 and 15 percentage points higher, respectively, when using corrected data. In the Middle East, the top 1 percent income share was 8 percent based on survey data, but 16 percent after correcting survey data with tax record information. (Alvaredo, Assuad and Piketty, 2019).

Trends can also differ depending on the data source. For instance, Jenkins (2017) found that the Gini for individual gross income in the U.K. rose by 7–8 percent between 1996/7 and 2007/8 when survey data is corrected using tax data, compared to a 5 percent decline when only survey data were used. For the United States, the Gini coefficient rose 8.4 percentage points between 1976 and 2006 when correcting the survey with tax records information, whereas official statistics based on the Current Population Survey indicated an increase of 7.2 percentage points (Atkinson, Piketty, and Saez, 2011). In Mexico, while the Gini coefficient from uncorrected surveys declined between 2014 and 2020, the Gini from surveys corrected with tax data remained stable (De Rosa, Flores, and Morgan, 2024).

The evidence also suggests that inequality indicators are highly sensitive to the *specific* correction method. Applying different correction approaches to the *same* data can yield significantly different results for the *same* inequality indicator. This sensitivity is evident in studies employing various correction methods on identical datasets. Moreover, there is no consistent pattern in how different methods influence inequality estimates. The impact of correction methods on inequality indicators varies depending on how each method transforms the data. For instance, De Rosa, Lustig, and Martinez Pabon (forthcoming) found that the uncorrected Gini coefficient for adult individuals for Brazil was 58.2, while corrected Gini coefficients ranged from 58.1 (using a within-survey semiparametric replacing approach) to 69.1 (using a survey and tax data combination method involving replacing and reweighting). However, the method producing the largest or smallest change wasn't consistent across countries. For Chile, the smallest change resulted from a semiparametric replacing method using tax data to estimate the corrected upper tail, while for Colombia, reweighting with tax data yielded the largest change. This variability underscores the need for careful consideration when selecting a

correction method and highlights the importance of multiple robustness checks, as there's no definitive scientific ranking of methods.

Given the above, the question arises as to what criteria should be used to determine which methodological approach brings us closer to the true inequality to be able to analyze its evolution and its relationship with other economic variables. Unfortunately, there are no statistical tests or calibration mechanisms available to make this determination. In practice, researchers often rely on assumptions, often made implicitly, and ad hoc considerations to decide which approach to follow. Our paper includes some broad guidelines that could be used to make decisions on the method and various aspects within each method. Nevertheless, the conclusion remains that it may not be possible to generate a single, definitive inequality indicator. Instead, one will have to accept that the best one can do is obtain a range of possible values of our inequality indicator of choice.

Beyond measurement, the "missing rich" problem has significant research and policy implications. For instance, if inequality trends and patterns differ depending on the data source used, conclusions on whether inequality is conducive to or a hindrance for economic growth might crucially depend on the choice of data. Analyzing the determinants of inequality might be affected as well. The existing literature on determinants relies on household surveys. If the surveys are not an adequate representation of the distribution in the target population, conclusions about the determinants of inequality may be inaccurate. For instance, if earnings in the surveys exclude top earners, the importance of education and experience as determinants of inequality might be exaggerated. The "missing rich" problem in inequality indicators also hinders our ability to accurately assess the progressivity of fiscal systems and evaluate the impact of reforms. Furthermore, excluding the wealthy from inequality measures can skew public perception regarding the fairness of the economic system and the desirability of potential policy changes.[14]

[14] See, for example, Alfani, 2024; Piketty, Saez, and Zucman (2022); and Milanovic (2023).

This article is organized as follows. Section 2 briefly reviews the causes of the missing rich problem in household surveys. Section 3 presents an analytical description of the correction approaches developed to address the missing rich problem in household surveys. Section 4 discusses how these approaches have been applied in practice. Section 5 provides some broad guidelines on how to proceed on selecting a correction approach. Section 6 concludes. In addition, there are four appendixes and an online Bibliographical Appendix that includes the comprehensive list of papers classified by approach.

2 The Missing Rich in Household Surveys: Causes

To understand the factors contributing to excluding the rich in household surveys, it is helpful to define five distinct population categories: the target population, the sampling frame population, the respondent population, the achieved sample (the raw household survey), and the preprocessed sample.[15] These categories help in identifying and understanding the sources of errors, such as coverage errors, unit nonresponse, item nonresponse, income underreporting, and preprocessing practices by data providers, that contribute to the missing rich problem in household surveys.

The target population is the set of units to be studied (e.g., the total population in a country who do not reside in institutionalized settings such as prisons, hospitals, etc.). That is, the entire group of individuals or households that the survey aims to represent. The target population includes the covered population (i.e., individuals with positive probability of being selected in the sample) and the uncovered population (i.e., individuals with zero probability of being selected in the sample). The sampling frame population is the population from which the sample is actually drawn, which may not perfectly match the target population due to limitations in the sampling frame. It includes members of the target population that have a positive probability of being selected into the sample. The sampling frame includes the respondent and nonrespondent subpopulations and, by definition, it excludes the uncovered population. The respondent

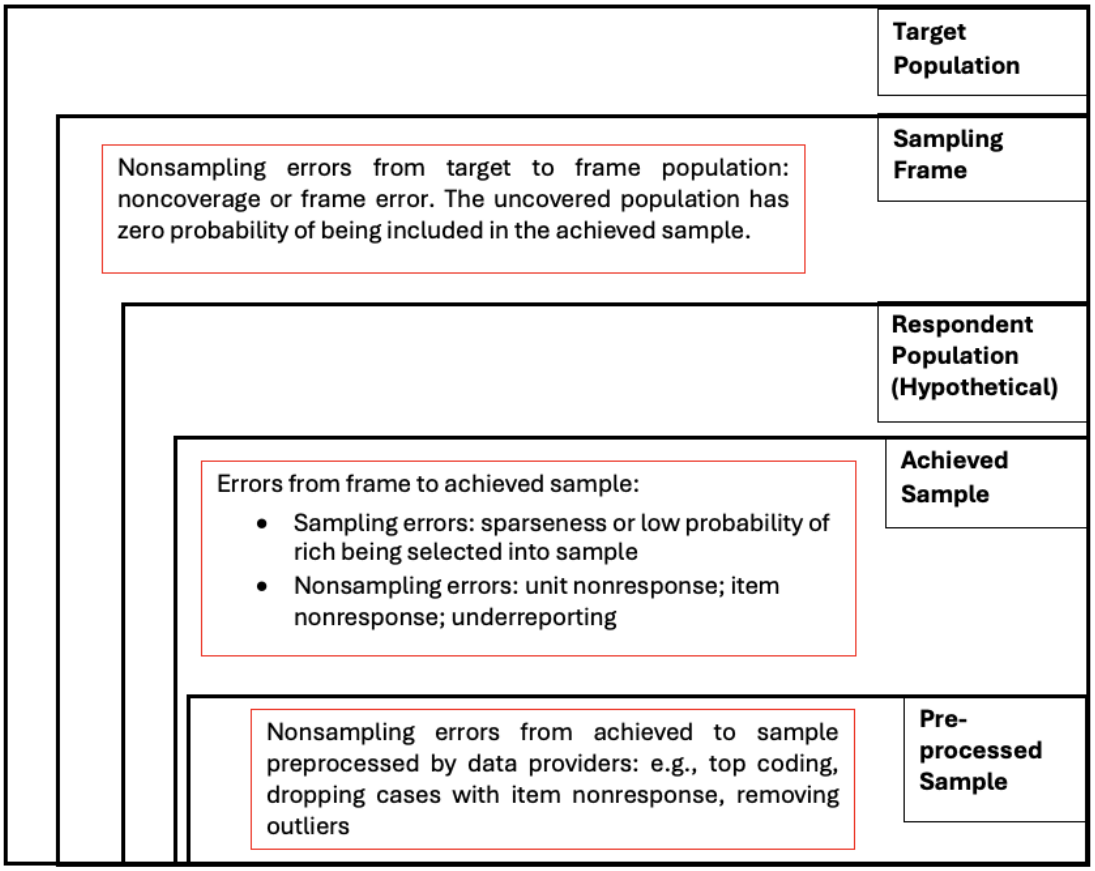
[15] As shown in Figure 1 (adapted from Figure 17.1 in Biemer and Christ, 2008), these four populations "are nested within one another with the target population encompassing the frame population which in turn encompasses the respondent population." (Biemer and Christ, op. cit., p. 318).

population is defined as that subset of the frame population that is represented by units who would respond to the survey if selected. It is a purely hypothetical concept because it is impossible to identify all the members of this population. The achieved sample is comprised by the households who were selected into the sample and who responded to the survey (they may not have responded to all the questions, though; or, responded them with error such as income underreporting). Finally, the preprocessed sample is the survey after the data providers (e.g., government statistical offices) make adjustments.[16]

There are, essentially, two main types of errors that can cause biases in inequality measures due to missing the rich in household surveys: sampling and nonsampling errors (Groves and Couper, 1998; Meyer and Mittag, 2019). For the missing rich problem, sparseness is one of the salient sampling errors. Within the nonsampling errors, there are five main factors embedded in the sampling design, data collection and data preparation process that may lead to the exclusion of the rich in household surveys: frame or noncoverage errors, unit nonresponse, item (income) nonresponse, underreporting, and errors introduced by data preprocessing practices. Nonsampling errors due to sampling design occur when top incomes are not captured due to noncoverage error. At the level of data collection, three upper tail issues may occur: unit nonresponse, item non-response and underreporting. In addition, the achieved sample may be subject to data preprocessing practices that bring their own set of errors and issues. Figure 1 presents the various types of errors in a schematic format. Below, each error is described in detail.

[16] These adjustments can create additional problems in capturing the rich accurately These will be discussed further below. They include practices such as top-coding and removing outliers. The preprocessed sample may also include corrections for unit nonresponse through recalibration of weights and for item nonresponse using imputation methods to replace missing incomes. Some statistical offices drop cases with missing incomes (item nonresponse). Others address income underreporting by adjusting survey incomes using external sources such as tax data or National Accounts.

Figure 1 The Missing Rich in Household Surveys: Causes



Note: Adapted by the authors from Biemer and Christ (2008), Figure 17.1. For definitions of the categories see text.

2.1 Sparseness

Sampling error is inherent to the process of sampling. Even if the achieved sample (household survey) is flawless, due to the skewness that characterizes the income distribution, very high incomes in surveys tend to be sparse. That is, there is no density mass at all points of the upper tail of the distribution's support. Random sample selection procedures may leave out very small subpopulations that accrue a disproportionately large part of household income. Put differently, the chances of observing individuals like Jeff Bezos, Warren Buffett, Bill Gates, or Elon Musk in the US Current Population Survey are positive but microscopically small. Sparseness produces volatility and may introduce bias in inequality measures.[17] Furthermore, to reduce

[17] Cowell and Flachaire (2007) show that these effects vary across inequality measures and conclude that the Gini coefficient is les prone to the influence of outliers.

volatility, statistical offices (or researchers) frequently remove observations with very high incomes, even if they are genuine, thereby introducing a bias in the process (more of it below).

In sum, sampling errors can be the underlying reason why the rich are missing in a particular survey: none of the rich individuals made it into the sample in a specific year. One way statistical offices can address the issue of sparseness is by oversampling rich individuals when designing the survey.[18] However, oversampling can be costly. An alternative way to cope with sparseness has been to replace the upper tail in the sample with a parametric model (e.g., the Pareto distribution). For a detailed discussion on how to address sampling errors (among other upper tail issues), see Cowell and Flachaire (2007) and Cowell and Flachaire (2015), for example.

There are other causes for the missing rich in household surveys. These are generically classified as nonsampling errors. Nonsampling errors will usually result in biased estimates of inequality.[19] These errors arise for several reasons. The sampling frame may be incomplete (an error specific to surveys), some respondents may not be reached or refuse to respond, data such as incomes may be missing for some respondents, or some respondents may not accurately report data such as their incomes. In addition, when the sample reaches the users it may have been subject to data preprocessing practices on the part of the country's statistical offices such as top coding.

The nonsampling errors can be classified into the following five categories: noncoverage error; unit nonresponse; item nonresponse; underreporting; and, preprocessing-induced errors.

[18] The word "oversampling" in this context should not be confused with an oversampling (or undersampling) due to, for instance, income-linked unit nonresponse. In the latter case, the concept refers to an error in the base weights that requires correction. In survey-design context, the concept of oversampling refers to stratified sampling practices to ensure that at least some individuals of a sparse population group makes it into the sample. Even if this group is oversampled for the purposes of collecting information, they will be assigned the weights that correspond in the population.

[19] To note is that nonsampling errors not only affect surveys. Non-sampling errors are also present in censuses and administrative data such as tax records.

2.2 Noncoverage Error

The *noncoverage or frame error* includes errors of exclusion and errors of inclusion in the frame population.[20] In measuring inequality (and poverty), we are primarily concerned with errors of exclusion also known as noncoverage error: that is, the exclusion of individuals who should be included in the frame but are not. Noncoverage error refers to individuals with zero probability of being selected into the sample. Noncoverage error refers to both deliberate and unintentional exclusions. Typically, subjects that are excluded from household surveys by design are inmates, subjects in hospitals or institutions, refugees, and the homeless. They may also exclude subjects who live in violent neighborhoods or in areas under conflict.[21]

If the noncoverage error is nonrandom and more frequent among the rich population, the ensuing inequality measures will be biased. In general, statistical offices try not to exclude anybody by design from household surveys samples (except for those mentioned above) and try to replace the population that cannot be covered (e.g., people living in violent neighborhoods or in conflict zones) by similar subjects, and oversample them. There is no ex-ante reason to believe that the rich will be excluded from the sample frame by design or unintentionally. Noncoverage error seems to be more problematic for measuring poverty than inequality because the homeless and the refugees are likely to be poor. However, an unintentional noncoverage error on the part of statistical offices may occur if, for example, the sampling frame does not adequately include areas where wealthy individuals live, such as gated communities.[22]

---

[20] The frame population can be a mega-sample of the country's population included in the most recent population census or the census population as a whole.

[21] For a discussion of issues of noncoverage at the bottom, see Atkinson (2016).

[22] If the entire population at the top of the income scale beyond a certain threshold were excluded (e.g., people living in gated communities whose incomes are higher than the highest income of people included in the survey), there would be truncation of the income variable: one knows that a set of individuals above an upper income threshold are excluded from the frame, but one knows nothing else (Case A in Table 2, Cowell and Flachaire, 2015). In such cases, one would be facing a redefined population and may want to use a parametric method to estimate the truncated part of the distribution, an approach that has been used by some to correct inequality estimates for the missing rich problem, as we will see below.To assess the extent to which the frame population in specific countries suffers from noncoverage, national statistical offices should carry out periodic reviews of the fitness for the purpose of the baseline population data (e.g., the Census) for their country and the sampling frame.

2.3 Unit Nonresponse

The nonrespondent population refers to individuals with a positive ex-ante probability—however small—of being selected into the sample but who do not (or would not) respond if selected because of noncontact, refusal, or other reasons. As such, unless the survey collector is able to replace the nonrespondent individual with a similar subject, the nonrespondent subjects end up not being included in the achieved sample.[23] There is evidence that unit nonresponse rates of 20 percent or more are common and that they have been growing over time. Meyer, Mok and Sullivan (2015) document a rise in unit nonresponse in US surveys. Hlasny and Verme (2018a; 2018b; 2021) and Luiten et al. (2020) show evidence for multiple countries.

In addition, there is evidence that unit nonresponse is not random which can lead to the underrepresentation of certain population categories (Atkinson, 2016). That is, population groups who are covered in the sample frame but where response rates are lower such as slum-dwellers and residents of gated communities.[24] Groves and Couper (1998) report that frequently it is impossible for the survey organizations to penetrate the gated communities in which many rich people live in poor countries: that is, de facto, the rich refuse to respond by making contact impossible. They also report that the probability of response is negatively related to almost all measures of socioeconomic status in rich countries. Korinek, Mistiaen and Ravallion (2007) find that the probability of nonresponse is correlated with income in the US Current Population Survey. Hlasny and Verme (2018a; 2018b; 2022) find a similar result for the US survey, the EU-SILC surveys and the household income and expenditure survey for Egypt.

---

[23] It is possible that none of the theoretical nonrespondent populations are selected for the sample, resulting in no unit nonresponse in the achieved sample. In such cases, one may never know if nonresponse is a problem or how big it is.

[24] If all of the individuals at the top of the income scale and beyond a certain threshold are nonrespondent, the resulting distribution will be right-censored. (Case B,Table 2 in Cowell & Flachaire, 2015). In other words, one knows that there are individuals above a particular income threshold who will end up being excluded from the survey (achieved sample) and what share of the population they represent. Using Cowell and Flachaire's terminology, we know that, above some threshold, there is an excluded sample; while there are point masses (density) at the boundary that estimate the population share of the excluded part, one does not know the corresponding income. Cowell and Flachaire discuss methods to address censoring.

A potential consequence of income-selective (nonrandom) unit nonresponse is that one cannot assume that the population weights supplied by the statistical office for each observation in the achieved sample (i.e., the expansion factors) are correct (Korinek, Mistiaen and Ravallion, 2007). In such cases, the achieved sample will not be a representative distribution of the target population and the inequality estimates might be biased.

2.4 Item Nonresponse

Another cause for the underrepresentation of top incomes in household surveys can be that within the respondent population, there may be people who do not provide a response for the income variable (or the expenditure variable in the case of consumption-based surveys). Such a situation falls under what is usually referred to in the statistical literature on measurement error as item nonresponse. Item nonresponse is defined as "...failure to obtain data for a particular variable (or item) in an interview or questionnaire when data for other variables in the survey have been obtained" (Groves et al., 2009, p. 354).[25]

If item nonresponse does not occur at random and is correlated with income, inequality estimates may be biased. For example, if the observations suffering from income nonresponse are dropped from the sample by the data providers, the survey will not be representative of the income distribution of the target population and, thus, inequality estimates will be biased. Biases can also occur if the missing incomes are imputed using average incomes or other imputation methods that are not able to (roughly) recover the actual incomes of the nonrespondents.

There is evidence that item nonresponse might increase with income. Using linked data that matches individuals in household surveys and tax records, Bollinger et al. (2019) found that in the Current Population Survey (CPS) in the United States, income nonresponse rises with income. Specifically, they found that only about 75 percent of the top percentile reported their earnings in the survey. Campos-Vazquez and Lustig (2019)

[25] This is a case of partial nonresponse where the missing item is income (or consumption or wealth). See Figure 1.1, Little and Rubin, (2014).

found that income nonresponse rises with income in Mexico's employment survey. This pattern indicates that higher-income individuals are more likely to not report their incomes, leading to biased inequality estimates if these observations are excluded from the analysis.

2.5 Underreporting

Underreporting refers to subjects who are selected and respond to the survey but who—when they respond—report income below its actual level. It occurs at the level of the respondent population and would, thus, affect the achieved sample. This is another form in which the rich can be underrepresented or missed. While these observations are included in the frame sample and the individuals agreed to respond and told their income, the latter is below (and could be way below) than the actual income. Thus, even if the rich are in the survey they do not appear to be rich. When the rich are included in surveys, underreporting may arise because high-income individuals usually have diversified portfolios with income flows that are difficult to value, such as capital income invested in pension funds or retained by corporations as undistributed profits; or because they may also be more reluctant to disclose their incomes, for example, to comply with the prevailing social norms of middle class membership (Valet, Adriaans and Liebig, 2018).[26]

Underreporting, thus, is a case of measurement error: even when people respond, they may misrepresent their income, whether on purpose or by mistake. When underreporting is not random and is correlated with income, inequality estimates will be biased.

As mentioned in the Introduction, by inspection or through comparison with other sources (such as tax records), it becomes apparent that people at the top of the income distribution tend to either not make it into the sample due to sparseness or noncoverage errors. Or the rich do not participate in the survey or if they do, the rich do not report their income. Or if they report their income, they underreport it. In general, it is difficult to disentangle the source. One way to unambiguously attribute the upper tail issues to underreporting (and distinguish it from item nonresponse or noncoverage, for instance)

[26] This hypothesis can also explain overreporting among poorer sectors (Valet, Adriaans and Liebig, 2019; Angel et al., 2019). Cognitive problems in understanding survey questions can also explain income misreporting.

is with linked data. Studies for the United States, Norway, New Zealand, the United Kingdom and Uruguay show that misreporting is high and is concentrated in the lower and the upper tail (Lillard, Smith and Welch, 1986; Kapteyn and Ypma, 2007; Abowd and Stinton, 2013; Bollinger et al., 2019; Hyslop and Townsend, 2020; Angel et al., 2019; Jenkins and Rios Avila, 2020; Flachaire, Lustig and Vigorito 2023). In the upper tail, misreporting takes the form of underreporting. [27] In Uruguay, for example, the same individuals in the top 1 percent, on average, report 60 percent less income on the survey than in tax data (Flachaire, Lustig and Vigorito, 2023). [28]

2.6 Preprocessing Practices

Data preprocessing refers to actions undertaken by statistical offices or other data providers that alter the achieved sample. Salient examples of data preprocessing include reweighting to address unit nonresponse, dropping observations with item nonresponse, imputing missing data such as income, rescaling incomes, top-coding, removing outliers, and providing subsamples. These are standard data processing techniques, and when used appropriately and transparently, they may not cause insurmountable problems. However, there are data preprocessing techniques that may introduce new biases.[29] The preprocessed survey may include practices that exacerbate the "missing rich" problem, such as top-coding, or produce inaccurate corrections to the problem, such as dropping observations when there is nonrandom item nonresponse for the income variable. These practices can lead to further biases and inaccuracies in the data. For more details, see Appendix 1.[30]

As a first step before applying correction methods, researchers should check whether the survey was subject to top coding, removal of outliers, rescaling, "hot deck" imputation, removal of cases with item nonresponse, or other data preprocessing

[27] This behavior rules out the hypothesis that the rich underreport their income in surveys to avoid paying taxes or because they are afraid of the tax authority.
[28] Notably, while there is underreporting at the top, in the bottom half of the distribution there is overreporting: for the same individual, income in the household survey exceeds income in tax returns.
[29] See, for example, Burkhauser et al. (2018) in the context of UK data.
[30] In addition, researchers may do their own data corrections to address unit and item nonresponse and measurement errors throughout the survey (and not confined to the top). These data preparation protocols are not focused on the missing rich problem and thus will not be discussed here. Useful resources for the interested reader are (Cowell and Flachaire, 2015 and Little and Rubin, 2014).

methods introduced by statistical offices. Ideally, researchers should have access to the unprocessed original data (the achieved sample) and/or a full or partial description of the methods and information the data producers used to make corrections; and as much information as possible on unit nonresponse, such as nonresponse by geographical units and, preferably, by primary sampling unit (PSU).[31] With this information, a researcher can choose whether to undo or correct the data preprocessing changes introduced by the data providers.[32]

3 Correction Approaches: An Analytical Description

3.1 A Taxonomy of Correction Approaches

Coverage errors, unit nonresponse, item nonresponse, underreporting, and sparseness, as well as some data preprocessing practices that affect the upper tail, could bias inequality measures.[33] There are a variety of approaches that have been proposed in statistics and the inequality measurement literature to address upper tail issues in household surveys. The goal of all these correction approaches is to transform the household survey so that the corrected distribution (or the corrected inequality measure, if it is not possible to recover the entire distribution) is a more accurate estimate of the actual (unobserved true) distribution of income.

A useful way to classify correction approaches is according to the source of the data and the choice of method (Hlasny and Verme, 2018a; Lustig, 2019). Depending on the source

[31] For example, if poststratification was used to address noncoverage and unit nonresponse, one may have access to the original weights. If the preprocessed sample excludes cases with item nonresponse, one may have access to the original sample

[32] However, more often than not there is no access to the original data—the achieved sample—and/or details on the methods used by data producers to generate the preprocessed sample are not available. Researchers then need to make assumptions, and these assumptions will be case-specific. The corrections proposed by Burkhauser et al. (2018) to income distribution data for the UK serve as a fitting example of how to cope with adjustments made by data producers when some information on the adjustments is available and some is not. The UK income distribution statistics are an interesting case because they are one of the few instances (perhaps the only one) where the released survey includes corrections at the top using information from tax data. The study by Hlasny and Verme (2018b) on Egypt is another example of how to cope when the released survey is a subsample of the original survey, which in this case includes 50 percent of the sample. These examples illustrate the challenges and strategies researchers employ when dealing with limited access to original data and detailed information on data preprocessing methods.

[33] This is so because "… the missing-data mechanism is not MCAR (missing completely at random) and the complete cases are not a random sample of all the cases." (Little and Rubin, 2014, location 1195 in ebook).

of the data, the approach can be either within-survey or a combination of the survey with external data, where the latter usually involves tax records, social security registries, or National Accounts. Depending on the correction method, one can distinguish between three main categories: replacing, reweighting, and combining reweighting and replacing. As we shall see below, there is a clear distinction among these three categories. Within each one, there are a number of submethods. Table 1 summarizes the approaches, the specific methods, the type of data used by them, and their respective underlying assumptions.

Table 1 Correction Approaches: A Taxonomy

APPROACH	REPLACING					REWEIGHTING				REWEIGHTING AND REPLACING (or vice versa)		
	Within-survey		Survey and External Data			Within-survey		Survey and External Data		Within-survey	Survey and External Data	
METHOD	Semiparametric	Nonparametric Imputation	Semiparametric	Nonparametric		Weighting Class Adjustment	Model Weight Adjustment	Poststratification	Reweighting Top and Uniform Downweighting of Rest	Model Weight Adjustment and Semiparametric	Reweighting Top and Semiparametric or Rescaling	Nonparametric and Reweighting Top
			External Income Microdata; Regression-based Income Prediction; Income Totals (Nat Accts)	Rescaling; Statistical Matching; Replacing w/External Data	Reweighting Top with External Income Microdata							
Assumes Common Support	No	Yes	No		Yes					No		
Weight of the Upper Tail and the Rest Intact	Yes					No						
Weights within Rest of Distribution Intact	Yes					No						
Observations (incomes) within Upper Tail Intact	No				Yes					No		
Absolute Poverty Indicators Intact	Yes					No						

Within-survey methods have been a long-standing practice in the measurement of inequality to deal with, for example, sparseness to reduce volatility and top coding to mitigate bias (Cowell and Flachaire, 2007). Within-survey corrections have also been used to address unit nonresponse (reweighting) and item nonresponse (imputation methods) (Deville, 2000).

Combining surveys with external data to deal with missing data such as unit and item nonresponse has been used to recalibrate weights, a method known as poststratification (see, for example, Atkinson and Micklewright, 1983; Rubin, 1987; Biemer and Christ, 2008). Rescaling using external data (a form of replacing as shown in Table 1) had been applied in some quarters. For instance, Altimir (1979, 1987) proposed a method to scale up survey-based wage income proportionally and capital income proportionally but for the top 20 percent so that totals matched the respective categories' totals in National

Accounts, a practice that was applied until roughly a decade ago by the UNECLAC in their reported inequality and poverty indicators.[34] In the past couple of decades, there has been a sharp rise in studies that combine surveys with external information (especially, tax records, social security records, and National Accounts) using a variety of correction methods. The presumption in the new generation of studies that combine data sources is that due to noncoverage errors and unit nonresponse, the rich are simply excluded from the surveys and those that are included frequently do not report their income or egregiously underreport it.

At the outset, it is important to note that correcting the survey to include the missing rich will *not* always result in higher inequality. Depending on the type of error and the correction method, corrected inequality measures can be higher or lower than the original uncorrected ones. Deaton (2005), for example, shows that correcting for unit nonresponse can result in a decline in measured inequality. This should not come as a surprise. Correcting for unit nonresponse, for example, usually entails the recalibration of weights and, when weights change, the reweighted and original Lorenz curves may cross each other. As Ravallion (2022) points out, the direction of change will depend on how the weights are adjusted and on the characteristics of the specific inequality measure, particularly with respect to which part of the distribution the indicator is more sensitive to. The corrected inequality could also be lower with the replacing method. This may happen, for example, in cases where the upper tail is replaced by a parametric distribution. Examples of lower corrected Gini coefficients can be found in Flores (2019), Flachaire, Lustig and Vigorito (2023), Hlasny and Verme (2018a; 2018b; 2021) and Morgan (2018). Empirically, however, the vast majority of cases show an increase in the inequality estimates after corrections, as will be shown in a later section. In Appendix 4, we will discuss what can be said in terms of the direction of change in the case of the Gini coefficient.

Let x be income, $f_x(x)$ the original uncorrected distribution (i.e., the household survey), and $F_x(x)$ the cumulative distribution function (CDF). Define $f_z(x)$ as the hypothetical

[34] This practice, discontinued in 2018, was applied by UNECLAC for inequality and poverty indicators and by CASEN the Chilean household survey.

true distribution of the target population and $f_x^c(x)$ as the corrected distribution.[35] The goal of all the correction approaches is to transform $f_x(x)$ so that $f_x^c(x)$ is a better representation of $f_z(x)$.[36]

Correction approaches, as mentioned above, can be classified into replacing, reweighting and reweighting and replacing combined. In replacing, one needs to identify the upper tail of $f_x(x)$ that suffers from one or more of the issues described above and select the method to correct it. In reweighting, one needs to decide the extent to which the original weights w in $f_x(x)$ must be corrected to address biases introduced by unit nonresponse, for example.[37]

Let's define $Q_x(p) = F^{-1}(p)$ as the corresponding quantile function and the p-quantile as x_p.[38] Let β be the upper tail population share that needs correction (e.g., the top 5 percent) and $(1-\beta)$ the population share of the rest. Now let β^c be the top population share for incomes $x > x_{(1-\beta)}$ in the corrected distribution $f_x^c(x)$, $(1-\beta^c)$ the share of the rest and w^c the weights corresponding to $f_x^c(x)$.

The main distinction between the replacing and reweighting approaches is whether β and $(1-\beta)$ remain the same or change in the corrected distribution. In replacing, these shares do not change, that is $\beta^c = \beta$, while in reweighting they do.-In reweighting $\beta^c > \beta$ for $x > x_{(1-\beta)}$ and, hence, the following needs to hold $(1-\beta^c) < (1-\beta)$ for all the incomes $x \leq x_{(1-\beta)}$.

---

[35] Depending on the study and the correction approach, x will be defined in different ways: for example, income per capita at the household level, individual taxable income per earner, etc.

[36] The formulas presented here correspond to the continuous functions case. The formulas for the discrete distributions would be analogous and are not shown here.

[37] Although $w(x)$, for expositional reasons, we do not include the (x). As a reminder, the weight in a survey can be interpreted as the number of individuals in the target population represented by the survey's respondent. Here we are assuming that the w are the probability of being selected into the sample. That is, these weights were not subject to any preprocessing adjustments by statistical offices to take account of unit nonresponse, for example.

[38] Before proceeding a clarification is in order. The word "quantile" is sometimes used to refer to population groupings such as deciles, quintiles, and so on. Here we use it as it has been formally defined (see, for example, Cowell and Victoria-Feser, 2000, p. 1224). The p quantile in the distribution of income is the value (income) that leaves below a p proportion of the population. The p-quantile $x = Q_x(.5)$, for example, is the income level that corresponds to the median.

Under replacing all the correction action occurs within β. Under reweighting, in contrast, the correction affects (in general) the entire sample weights w. In replacing, the $x > x_{(1-\beta)}$ change while the $x \leq x_{(1-\beta)}$ remain intact; the original weights w will also remain intact, except in the methods that reweight *within* the top.[39] In reweighting, the x remain intact while the original weights w are changed to the corrected weights w^c throughout the distribution. In methods that combine reweighting and replacing (in whichever order), both the $x > x_{(1-\beta)}$ and the w change.

A crucial step in replacing and in some reweighting approaches is identifying β: that is, the point p (e.g., .95) in the uncorrected quantile function above which correction must take place. There are a variety of ways to select this boundary or threshold which are discussed below.

Replacing as a correction method can work if the presumption is that the weights--except within the top-- in the original uncorrected survey adequately represent the target population (see Table 1). In other words, replacing assumes that undercoverage or unit or item nonresponse or underreporting are *within* the upper tail; they are not income-linked in a way that the β and $(1-\beta)$ are not representative of the target population.

Since it affects the w and not the x, reweighting as a correction method can work properly under the assumption that there is common support between the sample and the target population (see Table 1). Formally, there is common support when for every x where $f_z(x) > 0$ in the target population it will also be the case that $f_x(x) > 0$ in the sample. As stated by Ravallion (2021), reweighting is attractive because "It will be obvious to any user of micro survey data for empirical analysis that it is desirable to correct for bias in the top-income shares internal to the survey data. Doing so can retain both the statistical integrity of the survey design (with implications for statistical inference) and the great many applications for micro-data files in distributional analysis." (p. 6) With replacing, this is not possible in general except for some imputation approaches.

[39] While it may sound counterintuitive to include a reweighting component under the replacing category recall that the main distinction is whether β and $(1-\beta)$ remain the same after the correction. In this case they do: the weights within β change but not β.

The drawback of reweighting is that if there is no common support, then the correction will be limited. Reweighting, moreover, affects the entire distribution (in general). Thus, not only inequality but also (absolute and relative) poverty measures will usually change after correction. If there are reasons to believe that the survey's base weights are incorrect, then poverty estimated with the uncorrected survey may be biased as well and proper reweighting could take care of biases in inequality and poverty indicators in tandem. However, in some correction approaches the nontop of the distribution is mechanically (proportionally) downweighted to ensure that the density integrates to unity. In those cases, absolute poverty rates after correction may be biased.[40] To avoid the latter, reweighting could be subject to the condition of keeping poverty estimates unaffected as suggested by Bourguignon (2018). However, then other portions of the distribution will need to absorb the downweighting in full. Replacing leaves the untreated portion of the distribution unaffected and thus absolute poverty after the correction remains the same, except in the case when rescaling (one of the replacing methods) is applied throughout the distribution.

Regarding the use of survey data or combining it with external data such as social security or tax records, within-survey corrections will be limited if there is unambiguous evidence that there is undercoverage and/or underreporting by the rich. Below we discuss the advantages and disadvantages and challenges of using external data.

The methods are presented below and in Table 2. For each method, we refer first to within-survey corrections and then to methods using external data to correct the survey.

3.2 Replacing[41]

Let $f_x(x)$ be the original uncorrected distribution as defined above. Let $f_s^T(x)$ be the distribution of the corrected upper tail where the subscript *s* refers to within-survey corrections and let $f_y^T(x)$ be the distribution of the corrected upper tail where the subscript y refers to corrections using external data; the superscript *T* stands for the

[40] Relative poverty will vary depending on changes to the mean or median income. In this context, both reweighting and replacing could affect poverty estimates.
[41] The sampling weights in the replacing formulas are not explicitly mentioned for expositional reasons.

upper tail. Lastly, recall that $x = x_{(1-\beta)}$ is the income threshold above which the original distribution suffers from one or more of the upper tail issues.

In the general case, replacing entails:

$$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_{s,y}^T(x) \text{ for } x > x_{(1-\beta)} \end{cases}$$

As we show below and in Table 2a, the corrected upper tail $f_{s,y}^T(x)$ in the replacing approach can take a variety of forms depending on the method.

3.2.1 Within-survey

In within-survey corrections, $f_s^T(x)$ can be artificially generated with a parametric function or it can be derived from applying nonparametric methods such as within-survey imputation methods--e.g., hot deck, multiple imputation, etc. (as described in section 4).

In the semiparametric case, $f_s^T(x)$ is obtained from estimating a heavy tailed parametric distribution (for example, Pareto I or II, Singh–Maddala, Dagum or Generalised Beta distributions) fitted on the original survey data for $x > x_{(1-\beta)}$ (as in Cowell and Flachaire (2007), Burkhauser et al. (2012), Alfons et al. (2013), Hlasny and Verme, 2018a, 2018b, 2021, for example). In this case, the parametric density function becomes:

$$f_s^T = f_s(x;\, \eta) \text{ for } x > x_{(1-\beta)}$$

Where η is the vector of unknown parameters to estimate (Cowell and Flachaire, 2015).

In this case, the full density function becomes:

$$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_s^T(x) = f_s^T(x;\, \eta) \text{ for } x > x_{(1-\beta)} \end{cases}$$

In the case of a Pareto I distribution[42], the replaced upper tail becomes:

[42] In the case of Pareto distributions there are many tests available to determine t and assess the fitness of the corresponding estimates. However, as Cowell and Flachaire (2015) point out, the upper tail does not behave as a Pareto distribution in all cases.

$$f_s^T(x) = \frac{at^a}{x^{a+1}} \text{ for } x > x_{(1-\beta)}$$

Where a is the Pareto coefficient, a shape parameter that describes the heaviness of the upper tail, i.e. how the number of observations declines as income increases. [43]

Nonparametric methods such as within-survey imputation methods--e.g., so-called hot deck method--can be used, for example, to address item nonresponse (as for example in Hirsch and Schumacher, 2004 or Bollinger and Hirsch, 2006). The formulas that would apply in this case are shown in Table 2.

3.2.2 Combining Survey and External Data

In the replacing approach that combines *survey and external data*, the income $x > x_{(1-\beta)}$ is obtained from an external source such as tax records, social security registries or National Accounts, for example.

As in within-survey corrections, $f_y^T(x)$ may take the form of a parametric distribution but the key difference is that the statistically generated upper tail is estimated with the external data (as for example, in Alvaredo, 2011; Burkhauser et al., 2012; Alvaredo and Londoño, 2013; Lakner and Milanovic, 2016; Burkhauser et al., 2012; Jenkins, 2017; and Van der Weide, Lakner and Ianchovichina, 2018). In this case, the formula is analogous to that shown for the within-survey except that the parameters are obtained with the external data such as tax records, predictions based on housing prices, or National Accounts totals, for example (see Table 2).

Depending on the approach and available data, the parameters are obtained through either statistical estimation of the parametric function (see, for example, Jenkins, 2017) or, in the case of the Pareto I model, using handy formulas derived from the characteristics of the model (see, for example, Atkinson and Bourguignon, 2000; Atkinson, 2007; Alvaredo, 2011). The handy formulas and their use will be discussed in

[43] The inverted Pareto coefficient is $b = a/(a-1)$. Recall that a key property of a Pareto I distribution is that the average income above a given threshold is b (the inverted Pareto coefficient defined above) times that threshold (Atkinson and Bourguignon, 2000; Atkinson, 2007). A higher b (lower a) coefficient implies a heavier right-hand tail (Forbes et al., 2000), which in general means a higher income inequality depending on the indicator. The coefficients of the Pareto function are often defined using Greek letters. Here we adopted the notation in Atkinson (2007) to avoid confusion with other symbols used in this section.

Section 4. Suffice it to say for now that these formulas permit estimations of Gini coefficients and top shares when the external data is limited: for instance, when only tax tabulations or total incomes from National Accounts (or other sources) are available.[44]

A typical nonparametric imputation method consists of rescaling the survey upper tail to match the values observed in the external distribution. The density function takes the following general form:

$$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_y^T = x \ \varphi(x_p) \text{ for } x > x_{(1-\beta)} \end{cases}$$

where φ is the rescaling factor defined as:

$$\varphi(x_p) = x_{py} / x_{ps} \text{ for } x > x_{(1-\beta)}$$

In the case of discrete distributions, the rescaling formula takes the general form of:

$$f_y^T = (x \ \varphi_n) \text{ for } x > x_{(1-\beta)}$$

where φ is the rescaling factor and n is the number of categories.

The rescaling factor φ can be a single scalar or a vector. For example, it can be the ratio of total income in National Accounts to income in the survey (Altimir, 1987; Lakner and Milanovic, 2015; Bourguignon, 2018) or the ratio between the cell mean totals in external data (e.g., taxes) and the cell mean totals in the survey for percentiles in the predefined upper tail for an number of intervals (as in Piketty, Yang and Zucman, 2017). In the first case, n ranges from 1 to the number of National Accounts income categories selected for rescaling.[45] In the second, the number of intervals is usually set at the percentile level.[46] That is, the scaling factor varies across the upper tail depending on how different

[44] The parametric function of course does not have to be a Pareto I. In the literature it has been found that what is called the Pareto II (aka Generalized Pareto) or the Generalized Beta distribution may provide a better fit (see, for example, Cowell and Flachaire, 2007; Jenkins, 2017; and, Hlasny and Verme, 2021). The Pareto II model, however, does not produce the handy formulas.

[45] The method can be applied using other sources that are in totals.

[46] If the number of observations in the percent highest misreported and true data are not the same, we must reweight to guarantee that the selected true data represent *k* percent of the combined sample. In this case, the number of observations included in β in the tax data need to be reweighted to match the corresponding number of observations included in β in the survey. This reweighting it to ensure that population totals match. 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 cell means for each interval are. As the interval decreases, the corrected upper tail approaches the case in which the sample's observations are replaced completely with external microdata (if available). In the continuous case, the quantile adjustment is equivalent to completely replacing the observations in the survey with the observations corresponding to β in an external source. However, this is not the case with discrete distributions (Flachaire et al., 2023).

Other ways to correct the upper tail under the nonparametric replacing approach that combines surveys with external data include statistical matching methods (Bach, Corneo, and Steiner, 2009; Bach, Beznozka and Steiner, 2016) and replacing the upper tail with external microdata (Jenkins, 2017; Bollinger et al., 2019; Hyslop and Townsend, 2020; Jenkins and Rios Avila, 2020; Flachaire et al., 2023).

Finally, reweighting *within* the upper tail is another way to attempt the correction of the missing rich (Medeiros et al., 2014; Bourguignon, 2018; Silva, 2023). Recalibration can be done through poststratification so the new weights match shares in external sources such as tax data (as in Medeiros, de Castro and de Azevedo, 2018). [47] Alternatively, when the method uses National Accounts, for reweighting within the top a scalar can be added to the upper tail to account for missing rich people (Bourguignon, 2018, section 3.3).

The replacing method may entail completely removing the upper tail and replacing by an artificially generated distribution such as in the semiparametric method; in this case, by assumption, the "weights" (mass at different points of the distribution) *within* the top will also change. In rescaling, weights are kept intact while observations above the threshold $x > x_{(1-\beta)}$ are not removed but rescaled. In reweighting *within* the top, original observations of the entire distribution are kept intact (that is, not removed) while weights above the threshold $x > x_{(1-\beta)}$ are recalibrated to match, for instance, population shares in the reference external source (e.g., tax data).

[47] Note that while the formula is analogous to the formula for the reweighting method below, this refers to reweighting *within* the top only. This method is a form of poststratification within the top. In other words, in this case, the weights are only applied in the upper tail of the distribution and integrate to 1 within the top. Recall that poststratification assumes that there is common support between the upper tail in the sample and in the population.

Table 2a Correction approaches in formulas: Replacing

Approach	Within-survey	Survey and external data
Semiparametric	$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_s^T(x) = f_s^T(x;\, \eta) \text{ for } x > x_{(1-\beta)} \end{cases}$	$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_y^T(x) = f_y^T(x;\, \eta) \text{ for } x > x_{(1-\beta)} \end{cases}$
Semiparametric Pareto I	$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_s^T(x) = \dfrac{at^a}{x_s^{a+1}} \text{ for } x > x_{(1-\beta)} \text{ for } x > x_{(1-\beta)} \end{cases}$	$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_y^T(x) = \dfrac{at^a}{x_y^{a+1}} \text{ for } x > x_{(1-\beta)} \text{ for } x > x_{(1-\beta)} \end{cases}$
Nonparametric Rescaling	not applicable	$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_y^T(x) = x\, \varphi(x_p) \quad \text{for } x > x_{(1-\beta)} \end{cases}$
Nonparametric Rescaling by quantile ratios	not applicable	$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_y^T(x) = x\,(x_y/x_s) \text{ for } x > x_{1-\beta} \\ or, equivalently \\ f_y^T(x) = f_y^T(x_y) \text{ for } x > x_{(1-\beta)} \end{cases}$
Nonparametric Imputation and Matching methods[1]	$f_x^c(x) = \begin{cases} f_x(x)\, for\, x \leq x_{(1-\beta)} \\ f_s^T(x) = f_s^T(x_m, \theta/x_{os})\, for\, x > x_{(1-\beta)} \end{cases}$	$f_x^c(x) = \begin{cases} f_x(x)\, for\, x \leq x_{(1-\beta)} \\ f_y^T(x) = f_y^T(x, \theta/x_{oy})\, for\, x > x_{(1-\beta)} \end{cases}$
Nonparametric Reweighting within the top[2]		$f_x^c(x) = \begin{cases} f_x(x) \text{ for } x \leq x_{(1-\beta)} \\ f_y^T(x) = x\,(w\,(f_y(x)/f_s(x)) \text{ for } x > x_{(1-\beta)} \end{cases}$

Note: For definitions see text. Those which are not included in the text are specified in the notes below. Recall that the subscripts s and y refer to within-survey and survey combined with external data.

[1] x_m are the missing income data, x_o are the observed data and θ are the corresponding imputation or matching parameters.

[2] w are the base weights corresponding to $x > x_{(1-\beta)}$ where $\sum_{x_{(1-\beta)}}^{xmax} w\,(f_y(x)/f_s(x) = 1$. The rest of the w are implicit. Recall that for expositional purposes we do not include the base weights since in replacing they remain intact.

3.3 Reweighting

As defined before, let x be income and $f_x(x)$ the original uncorrected distribution and let the corrected distribution be defined as $f_x^c(x) = w^c f_x(x)$. The reweighted distribution can take different forms depending on the method. As with replacing, some methods

rely completely on information from the survey (including the sample frame) and some use external data such as tax records and National Accounts, for instance. As we shall see, some methods select a threshold which results in a new $\beta^c > \beta$ while in others the new β^c is obtained as part of the reweighting procedure (that is, no predetermined threshold is established). The latter can be within-survey or use surveys and external data (Table 2b).

3.3.1 Within-survey

There have been two main strands of within-survey reweighting: weighting class adjustment and model weight adjustment (Biemer and Christ, 2008). In both these cases, the new β^c is determined by the overall reweighting procedure rather than preselected. In both methods, the new weights are generated using information on unit nonresponse. The weighting class adjustment method uses nonresponse rates by geographic location or other group attributes (age, sex, etc.). The method assumes that survey nonparticipation (unit nonresponse) is not linked to observable characteristics--such as income-- within each grouping (Harris, 1977; Atkinson and Micklewright, 1983), an assumption that may be unrealistic. In contrast, the model weight adjustment method proposed by Mistiaen and Ravallion (2003) and Korinek et al. (2006; 2007) assumes that the probability or propensity of nonresponse within a geographic area can increase with income. The method consists of modeling the probability of survey response of participating households as a function of their observable characteristics to generate the new weights.[48] This method requires access to unit nonresponse rates by Primary Sampling Unit or geographic area. In the weighting class adjustment method, let i be the region or group characteristic of interest, h_i be the number of respondent households and h'_i the number of nonrespondent households. The base weights associated to each observation (household) are recalibrated using the nonresponse rate corresponding to each i, h'_i / h_i to obtain the new weights.

[48] A useful overview of this method can be found in Ravallion (2022).

The corrected density function takes the following form:

$$f_x^c(x) \;=\; w^c f_s(x) = (\mathrm{w}\,(1 + h'_i / h_i)) f_s(x)$$

This procedure assumes that the characteristics of respondents and nonrespondents are similar within each (geographic) group. In our context, it assumes that there is no income-linked nonresponse.

To overcome this limitation, Ravallion and Mistiaen (2003) and Korinek et al., (2006; 2007) develop a variant of the model weight adjustment approach that models the determinants of nonresponse. This approach allows the probability of nonresponse to vary with the characteristics of each sampled unit (that is, the household). In this way, the decision to respond is not assumed to be independent of the variable of interest: in this case, income. The corrected density function becomes:

$$f_x^c(x) \;=\; w^c f_s(x) = (w(1 + \hat{\rho}_{ij}) f_s(x)$$

Where $\hat{\rho}_{ij}$ is the estimated probability of a household *i* in a region j to respond to the survey.

The method requires that the variables that affect nonresponse are observable for the respondents. The method assumes that the likelihood of responding based on the variable of interest (i.e., income) is independent of the partitioning by geographic area. In other words, the presumption is that all the households with the same characteristics (including income per capita) will have the same probability of responding to the survey across geographic areas. For details, see Appendix 2.

3.3.2 Combining Survey and External Data

Just as in replacing, reweighting can use survey data combined with external sources such as tax and social security records and National Accounts, for example. In poststratification, weights are recalibrated comparing population shares for similar income intervals (quantiles in the household survey, multiples of the minimum wage, etc.) from survey with external data: e.g., taxes or social security registries (as for

example, Atkinson and Micklewright, 1983; Campos-Vazquez and Lustig, 2019). [49] In this case, the corresponding weights are defined as follows:

$$w^c = \frac{w\, f_y(x)}{f_s(x)}$$

The corrected distribution can be written as:

$$f_x^c(x) \;=\; w^c f_x(x) = \; x\,(w\, f_y(x)/f_s(x))$$

where $\sum w\,(f_y(x)/f_s(x) = 1$

In poststratification and in the within-survey reweighting methods just described, there is no identification of a particular income threshold that predefines where the reweighting should be implemented. Reweighting can happen throughout the distribution or more prominently in parts of it including the zone of higher incomes, but the latter is the outcome of the reweighting method and does not need to be established ex ante.

There is also a strand of reweighting that defines an income threshold $x = x'$ above which the weight of the entire upper tail needs to increase and, to accommodate this, the weights of the observations below that threshold need to be compressed. In some cases, the threshold is determined exogenously (Flachaire et al., 2023; Bourguignon, 2018).[50] In one method, the threshold is determined endogenously (Blanchet, Flores and Morgan, 2022). In the first instance, the threshold is chosen arbitrarily (Bourguignon, 2018) or heuristically. In both cases, the compression of the weights of the observations below the threshold is usually proportional and thus the distribution of income below the threshold remains unchanged but not so absolute and relative poverty indicators.

[49] Some authors define as poststratification the case in which the new weights are obtained from external sources (e.g., Little and Rubin, 2014) while others use the term for any reweighting process (Atkinson and Micklewright,1983). In this paper, we use the first one.

[50] Flachaire et al. (2023) use tax microdata while Bourguignon (2018) uses just one variable: average total income from National Account (or any other external source).

If λ is the uniform factor by which the weights below threshold x' are compressed and f_y^T is the distribution of the reweighted upper tail, then:

$$f_x(x) = \lambda f_x(x) \text{ for all } x \leq x'$$

$$f_y^T = w^c f_y(x) \text{ for all } x > x'$$

where new weights w^c are generated using shares in, for example, tax records. Blanchet et al. (2022) propose a method to identify x' endogenously described below.

The key underlying assumption in reweighting (within survey or combined with external data) is that the sample and the target population have common support. In other words, that (at least) some of the rich individuals made it into the survey. The problem that reweighting wants to address is that due to noncoverage errors or unit or item nonresponse, the weights in the original sample (and not just within the top) are not correct and should be recalibrated. In reweighting, the original observations are always kept intact but the original weights (sometimes called base weights) are removed (all or some) and replaced with new ones. In contrast to the replacing approach, because weights change poverty estimates using the corrected distribution will always differ from those obtained with the original survey (unless by chance or by design, the weights of the portion below the poverty line remained intact after the correction).

Table 3b Correction approaches in formulas: Reweighting

Approach	Within-survey	Survey and external data
Weighting class adjustment[1]	$f_x^c(x) = (w(1 + h'_i / h_i))f_s(x)$	not applicable
Model weight adjustment[2]	$f_x^c(x) = (w(1 + \hat{\rho}_{ij})f_s(x)$	not applicable
Poststratification	not applicable	$f_y^c(x) = x\ (w\ f_y(x)/f_s(x))$
Exogenous threshold[3]	not applicable	$f_x^c(x) = \begin{cases} \lambda f_x(x) \text{ for } x \leq x' \\ (w\ (f_y(x)/f_s(x))f_s(x) \text{ for } x > x' \end{cases}$
Endogenous threshold[3]	not applicable	$f_x^c(x) = \begin{cases} \lambda f_x(x) \text{ for } x \leq x' \\ (w\ (f_y(x)/f_s(x))f_s(x) \text{ for } x > x' \end{cases}$ $x' = \max(x) \text{ subject to } \frac{F(x)_s}{F(x)_y} = \frac{f(x)_s}{f(x)_y}$

Note: For definitions see text. Those which are not included in the text are specified in the notes below. Recall that the subscripts s and y refer to within-survey and survey combined with external data.
[1] w are the base weights.
[2] $\hat{\rho}_{ij}$ are nonresponse rates by geographic unit i and income category j estimated using GMM in model based adjustment (Korinek et al., 2007). See details in text and Appendix 2.
[3] The threshold x' can be selected exogenously or endogenously as in Blanchet et al. (2022). The latter is described in section 4.

3.4 Reweighting and Replacing

Reweighting assumes that the survey and the true distribution have common support and replacing assumes that weights (except within the top in some methods) are correct. If neither assumption holds, the proposed approach is to apply both methods combined. In one approach reweighting goes first and replacing next. Again, this approach can be implemented within the survey (Hlasny and Verme, 2018b; 2021) and using external data such as tax records (Blanchet et al., 2022) or National Accounts (Bourguignon, 2018 or Silva, 2023).

3.4.1 Within-survey

The method consists of two steps. In the first step, there is within-survey reweighting with either the class adjustment or model weight adjustment methods previously described. The original weights are multiplied by the reweighting factor by region (class adjustment) or by region and income interval (model-based weight adjustment). In the case of the latter, the corrected weights take the form:

$$w^c = w\,(1 + \partial_{ij})$$

where ∂_{ij} is the reweighting factor for region *i* for the class adjustment and region and income interval *j* in the case of model adjustment, and $\sum_{ij}(1 + \partial_{ij}) = 1$.

The second step consists in identifying the $x = x_{(1-\beta^c)}$ above which the replacing component of the correction will be implemented. A key difference with the pure replacing method is that the share of the upper tail to be replaced is defined after reweighting. The corrected upper tail can be estimated using parametric or nonparametric methods.

When the upper tail is replaced with a parametric model, the corresponding density function becomes:

$$f_x^c(x) = \begin{cases} (w(1 + \partial_{ij}))f_x(x) \text{ for } x \leq x_{(1-\beta^c)} \\ f_s^T(x) = \big((w(1 + \partial_{ij}))f_s(x);\ \eta\big) \text{ for } x > x_{(1-\beta^c)} \end{cases}$$

3.4.2 Combining Survey and External Data

There are at least four options that have been pursued in this approach. In all four, there is a threshold $x = x'$ above which the weight of the entire upper tail is increased and the weights of the observations below that threshold are uniformly compressed. Reweighting or replacing goes first depending on the method. The formulas are presented in Table 2c.

As discussed in subsection 3.3, some methods determine the threshold exogenously, while others endogenously. Anand and Segal (2017) define the threshold exogenously as the maximum income in the survey. The survey is assumed to represent a portion of the target population (e.g., 99 percent) and the top portion (e.g., the top 1 percent) is generated with a parametric function that can be estimated with external data such as tax records. In other words, the "void" created by assuming the survey doesn't cover the

entire target population is filled by a parametric model. DWP (2015), Burkhauser et al. (2018) and Bourguignon (2018) also determine the threshold exogenously. Next, they apply nonparametric rescaling followed by reweighting. For rescaling, the first two use tax data, while Bourguignon uses National Accounts totals. Combining survey and tax data, Blanchet et al. (2022) endogenously determine the threshold (details below), applying reweighting first and then rescaling incomes in the upper tail to match incomes in tax data.

Table 2c Correction Approaches in Formulas: Reweighting and Replacing

Approach	Within-survey	Survey and external data
Model weight adjustment and semipara-metric replacing[1]	$f_x^c(x) = \begin{cases} (w(1+\hat{\rho}_{ij}))f_x(x) \text{ for } x \leq x' \\ f_s^T(x) = \left((w(1+\hat{\rho}_{ij}))f_s(x);\ \eta\right) \text{ for } x>x' \end{cases}$	not applicable
Exogenous Threshold and semipara-metric replacing[2]	not applicable	$f_x^c(x) = \begin{cases} \lambda f_x(x) \text{ for } x \leq x' \\ (w\ (f_y(x)/f_s(x))f_s(x) \text{ for } x>x' \\ f_y^T(x) = \left(\frac{w\ (f_y(x)}{f_s(x)}\right) f_y^T(x;\ \eta) \text{ for } x> x^{max} \end{cases}$
Endogenous Threshold and nonpara-metric replacing		$f_x^c(x) = \begin{cases} \lambda f_x(x) \text{ for } x \leq x' \\ f_y^T(x) = \left(\frac{w\ (f_y(x)}{f_s(x)} f_s(x)\right)\left(\frac{x_y}{w\ (f_y(x)/f_s(x))f_s(x)}\right) \text{ for } x>x' \end{cases}$ $= \max(x) \text{ subject to } \frac{F(x)_s}{F(x)_y} = \frac{f(x)_s}{f(x)_y}$
Exogenous Threshold and nonpara-metric replacing		$f_x^c(x) = \begin{cases} \lambda f_x(x) \text{ for } x \leq x_{(1-\beta^c)} \\ f_y^T(x) = (w\ \ (f_y(x)/f_x(x)\ )\ f_y^T(x_m, \theta/x_{oy})) \text{ for } x>x_{(1-\beta^c)} \end{cases}$

Note: For definitions see text. Those which are not included in the text are specified in the notes below. Recall that the subscripts s and y refer to within-survey and survey combined with external data. [1] $\hat{\rho}_{ij}$ are nonresponse rates by geographic unit i and income category j estimated using GMM in model based adjustment (Korinek et al., 2007). See details in text and Appendix. [2] x^{max} is the maximum income observed in the survey

"

4 Correction Approaches in Practice

Based on the treatment of original observations and weights, the type of data sources, the application of parametric or nonparametric imputation methods, and how an upper tail threshold is selected (if applicable), we have identified at least twenty-two distinct approaches in the literature. The approaches, with their respective assumptions and illustrative bibliographical references, are presented in Table 3. The references included are meant to be suitable examples of their application, but by no means is an exhaustive list. In the online Bibliographical Appendix, we include a comprehensive list of references classified by approach. [51]

Table 4 Correction Approaches in Practice

APPROACH	REPLACING											REWEIGHTING						REWEIGHTING AND REPLACING (or vice versa)				
	Within-survey		Survey and External Data									Within-survey		Survey and External Data				Within-survey	Survey and External Data			
METHOD	Semiparametric	Nonparametric Imputation	Semiparametric				Nonparametric					Weighting Class Adjustment	Model Weight Adjustment	Poststratification	Reweighting Top w/Exogenous Threshold		Reweighting Top w/Endogenous Threshold	Model Weight Adjustment and Semiparametric	Reweighting Top w/Exogenous Threshold and Semiparametric	Reweighting Top w/ Endogenous Threshold and Rescaling	Nonparametric and Reweighting Top w/ Exogenous Threshold	
			External Income Microdata	Regression-based Prediction of Income	External Income Totals	Reweighting Top with External Income Totals	Rescaling Top Incomes w/External Income Microdata	Rescaling Top Incomes w/Income Totals	Statistical Matching	Replacing Top with External Data in Full	Reweighting Top with External Income Microdata				External Income Microdata	Income Totals					External Income Microdata	External Income Totals
Applications	Cowell and Flachaire (2007); Burkhauser et al. (2012); Alfons et al. (2013); Hlasny and Verme (2018a, 2018b, 2021); Burkhauser et al. (2018)	Hirsch and Schumacher (2004); Bollinger and Hirsch (2006)	Alvaredo (2011), Burkhauser et al. (2012); Alvaredo and Londoño (2013); Jenkins (2017)	Van der Weide, Lakner and Ianchovichina (2018)	Lakner and Milanovic (2016)	Bourguignon (2018)	Piketty, Yang and Zucman (2019)	Altimir (1987); Bourguignon (2018)	Bach, Corneo, and Steiner (2009); Bach, Bennoska and Steiner (2016)	Bollinger et al, (2019); Flachaire et al.(2023)	Medeiros et al. (2018)	Harris (1977), Atkinson and Micklewright (1983)	Mistiaen and Ravallion (2003); Korinek et al. (2006; 2007), Hlasny and Verme (2018a; 2018b; 2021)	Atkinson and Micklewright (1983), Campos-Vazquez and Lustig (2019)	Flachaire et al. (2023)	Bourguignon (2018)	Blanchet et al. (2022), Flachaire et al. (2023)	Hlasny and Verme (2018a; 2022)	Atkinson (2007), Anand and Segal (2015)	Blanchet et al. (2022)	Burkhauser et al. (2018)	Bourguignon (2018)
Type of Data	Survey		Survey, Tax and Social Security	Survey and House Prices (or other variables that predict incomes)	Survey and National Accounts	Survey and National Accounts	Survey and Tax	Survey and National Accounts	Survey and Tax			Survey and Nonresponse Rate by Geographic Area	Survey and Nonresponse Rate by Primary Sampling Unit or Geographic Area	Survey, Census, Tax and Social Security	Survey, Tax and Social Security Survey	Survey and National Accounts	Survey and Tax	Survey and Nonresponse Rate by Primary Sampling Unit or Geographic Area	Survey and Tax		Survey and National Accounts	
Assumes Common Support	No	Yes	No							Yes								No				
Weight of the Upper Tail and the Rest Intact	Yes											No										
Weights within Rest of Distribution Intact	Yes											No		No (Mechanical Uniform Downweighting)				No	No (Mechanical Uniform Downweighting)			
Observations (incomes) within Upper Tail Intact	No									Yes								No				
Absolute Poverty Indicators Intact	Yes											No										
Generates Corrected	Distribution	Microdata	Distribution	Distribution	Distribution	Distribution	Distribution	Microdata**	Microdata	Microdata*	Microdata	Microdata						Distribution	Distribution	Microdata**	Microdata**	Distribution

Note: Some exercises of rescaling to National Account totals involve proportionally upscaling wage income (and other income categories) for the entire income distribution. In such cases, absolute poverty rates will not remain unchanged. Microdata here refers to the feasibility of keeping the observations with their corresponding covariates.
*The microdata can be preserved in the case of linked data. If replacement is with data that is not linked, the microdata cannot be preserved.
**The microdata can be "recovered" under certain assumptions. See text below.

A major distinction between correction approaches is whether they rely solely on the household survey or bring in external data such as tax or social security records, National Accounts, etc. The approaches that combine surveys with external data require aligning of the unit of analysis and the income concepts. In addition, household surveys in many countries collect information on consumption (expenditures) instead of income (or, they

[51] While we cannot claim that the latter is exhaustive given the exponential growth of work in this field, it is the most comprehensive list that exists to the best of our knowledge.

collect income data poorly). We discuss some of the salient data reconciliation challenges and procedures in Appendix 3.

4.1 Replacing

As discussed above, when noncoverage, nonresponse, and underreporting are assumed to affect the upper tail of the income distribution, the distribution can be conceptualized as comprising two segments. The first segment encompasses a top proportion affected by these upper tail issues. The second segment represents the non-top portion of sampled individuals for whom the sample provides a reliable representation of the population. A crucial step in the correction process is to determine the point in the distribution where the upper tail issues begin. In other words, the researcher needs to identify the income level or quantile (e.g., p95, p99) at which the "missing rich" problem occurs.

The selection of the threshold is arguably the most challenging aspect of the correction process. The accuracy of the corrected inequality measures is directly impacted by the distance between the selected threshold and the actual threshold. However, the true threshold is inherently unknown, Consequently, researchers commonly employ sensitivity analyses using multiple thresholds to assess the robustness of their results. Approaches for threshold selection vary widely, ranging from fairly arbitrary assumptions and visual inspection to statistical methods. In some reweighting methods, an upper tail income threshold is also employed, but its purpose differs. Here, the threshold determines the point at which the entire upper tail of the distribution is upweighted, while the remaining portion is downweighted. Notably, this "threshold" is sometimes defined as the highest income observation in the survey (see, for example, Anand and Segal, 2016), implying that the top of the income distribution is entirely missing and that the survey only represents a portion of the population (e.g., the bottom 99 percent).

4.1.1 Semiparametric

The correction method that relies on removing the upper tail and replacing it with a parametric function (e.g., a Pareto model or a generalized Beta distribution) is called semiparametric. In the within-survey semiparametric approach, the upper tail of the income distribution is replaced with the density generated by fitting a statistical

(theoretical) distribution such as the Pareto function using the original observations in the survey.[52] When combined with external data, the upper tail is replaced with the density generated by fitting a statistical distribution to the external data (e.g., tax records) instead. If the external data is limited (for instance, only tax tabulations or National Accounts totals are available), the upper tail is replaced using the parametric model's formula. Whether within-survey or combining survey data with external sources, this approach shares a number of characteristics, so we discuss them together.

Specifically, if we define (as we did above) the affected top incomes population share as β, "it may be reasonable to use a parametric model for the upper tail of the distribution… and to use the empirical distribution function directly for the rest of the…" (Cowell and Flachaire, 2015, p. 84; for the original discussion, see Cowell and Victoria-Feser, 2007). The most commonly used parametric model for the upper tail is the Pareto distribution (Pareto I and Pareto II), but other models have been proposed.[53] This method and its variations have been widely used to address issues such as sparseness, incomplete data (e.g., top-coding), right-truncation, and data contamination. Cowell and Flachaire (2015) provide a detailed discussion of their advantages and shortcomings.[54] While initially developed to address issues like sparseness and truncation in data, the semiparametric replacing approach can also be effectively applied to address unit or item nonresponse and underreporting among high-income individuals (Atkinson, 2007; Bourguignon, 2018).

As Cowell and Flachaire (2015) indicate, if one chooses this path, three important decisions must be made: how should the proportion β be chosen, what parametric model should be used for the upper tail, and how should the model be estimated.[55] These decisions apply to both within-survey and survey combined with external data

---

[52] This approach corresponds to Approach A in Jenkins' Figure 1 (Jenkins, 2017, p. 262).

[53] For example, Singh-Maddala, Dagum and Generalized Beta distributions (Cowell and Flachaire, 2015). For further discussion, see section 6.3 in Cowell and Flachaire (op. cit.).

[54] As surveyed by Cowell and Flaichare (2015), starting with Vilfredo Pareto himself there is a long tradition of using parametric, semi-parametric and non-parametric methods to handle imperfections in data. Cowell and Flachaire state that researchers have adopted a number of work-rounds such as multiplying top-coded values by a given factor (Lemieux, 2006, Autor et al.,2008) or attempting imputations for missing data (Burkhauser, Feng and Larrimore, 2010, Jenkins et al.,2011). See also Alfons, Temple and Filzmoser (2013), Burkhauser et al. (2012), Cowell and Flachaire (2007), Ruiz and Woloszko (2016).

[55] Figure A1 in Cowell (2009, p. 159) presents the various options available and the relationship between them.

approaches. As demonstrated by Jenkins (2017), inequality estimates with corrected data can be sensitive to all three of these decisions. [56]

The selection of β entails choosing the income threshold x' or the quantile q above which upper tail issues are assumed to occur. In the literature, some authors select the β proportion by inspection (heuristic approach) or make arbitrary assumptions that upper tail issues are confined to a certain percent of the distribution (e.g., top 1 percent or top 5 percent).

Statistical methods, primarily focused on Pareto distributions, have also been proposed. Common approaches involve plotting the logarithm of the rank of income observations against the log of income (Zipf plots or minimum excess plots) to identify the quantile at which the income distribution can be described by a Pareto function (Coles, 2001). Dupuis and Victoria-Feser (2010) developed a robust prediction error criterion to estimate the Pareto coefficient and the parameters for other thick-tailed distributions, thus defining the corresponding threshold.

The choice of the threshold can significantly influence the results, regardless of whether one uses just survey data, combines it with external sources, or employs semiparametric or nonparametric methods. Choosing a threshold far from the true value can do more harm than good, as it replaces one biased estimate with another. Since the true threshold is inherently unknown, it is crucial to assess the sensitivity of results to different threshold choices. Using a simulated true distribution, Flachaire, Lustig, and Vigorito (2023) demonstrate that the bias in the "corrected" income distribution increases as the selected threshold diverges from the actual threshold.

Jenkins (2017) illustrates the sensitivity of threshold selection, showing that using the same survey, tax data, and parametric model to replace the upper tail, the Gini coefficient in the UK in 2010 can vary between 0.386 and 0.439 for p99 and p90, respectively. Hlasny and Verme (2021) find similar results for the US. Using the same 2013 survey data and parametric model for the upper tail replacement, they find that the

[56] Even though Jenkins (2017) uses the approach that combines survey data with tax returns, his analysis on this topic equally applies to within-survey corrections.

Gini coefficient for p99 is 0.491 and 0.5792 for p95. Using data for Uruguay, Flachaire, Lustig, and Vigorito (2023) found that choosing the threshold where underreporting in the survey, compared to linked tax data, begins (p30) yields a Gini coefficient of 0.45. Choosing the p90 threshold instead yields a Gini of 0.44.

Regarding the parametric model, while the most common application has been the (one-parameter) Pareto I model, research by Cowell and Flachaire (2015), Charpentier and Flachaire (2022), and others demonstrates that this function may not always provide the best fit. These studies present a thorough review of alternative distributions for estimating the upper tail using a parametric model, along with a set of tests to assess their fit. A comprehensive empirical application focused on the fitness of Pareto models to the upper tail in a study for the UK can be found in Jenkins (2017). After conducting several sensitivity and robustness checks, the author concludes that the Pareto II model generally outperforms the Pareto I model. Hlasny and Verme (2021) find that the Generalized Beta Distribution provides a better fit than Pareto models when applied to a within-survey semiparametric correction for data from the United States.[57]

The semiparametric approach, as mentioned, can estimate the statistically generated upper tail using observations from the survey or external data sources like social security or tax records. Both approaches assume that the uncorrected upper tail and the true upper tail do not share common support. This raises the question of why one should estimate parameters using external data if neither distribution is assumed to have common support with the target population. Jenkins (2017) argues that while fitting a parametric upper tail using observations from the survey may address the sparsity problem by ensuring density mass across the entire distribution's support, the estimated upper tail based on model-based extrapolation from the observed survey data may not be a reliable representation of the "true" upper tail. This is because the parameter

[57] Regarding the estimation of the model, Cowell and Flachaire (2015) recommend using an estimator called Optimal b-robust estimator (OBRE) because the maximum likelihood estimation method for the Pareto model is known to be sensitive to data contamination (see p. 85). For more details, see Jenkins (2017). The interested reader can find more details on estimation procedures and testing in the references provided in this section.

estimates derived from survey data may fall short of the required correction. An indication of this issue can be observed in the difference in the magnitude of the inverted Pareto coefficient depending on the data source used for estimation. For example, in Piketty, Yang, and Zucman's analysis for China, the inverted Pareto coefficient estimated using survey data is as low as 1.5 or less, while it equals 2.5 or more when estimated with tax data (Piketty, Yang, and Zucman, 2019). (Recall that a higher inverted Pareto coefficient indicates a more unequal distribution.) Similar findings are observed in linked data for Uruguay. Flachaire et al. (2023) show that the Pareto coefficient is 1.29 when estimated using survey data for the top 1 percent, while it is 1.8 when estimated using tax records for the exact same individuals.

The semiparametric methods allow for the estimation of corrected density functions, cumulative density functions, Lorenz curves, and inequality measures. The formulas for these calculations can be found in Cowell and Flachaire (2015) and Jenkins (2017). However, a limitation of these methods is the loss of covariate information.

When the external data consists of tax tabulations or National Account totals, the parametric function is not estimated statistically but is directly calculated using readily available formulas specific to the Pareto I model. In the case of tax tabulations, researchers can leverage the group decomposability of the Gini coefficient to estimate the corrected Gini using a formula.[58] The formula for non-overlapping groups is as follows (Dagum, 1997; Atkinson and Bourguignon, 2000; Atkinson, 2007; Alvaredo, 2011):

$$G^c = \frac{1}{2a-1}\beta S + G^{sb}(1-\beta)(1-S) + S - \beta$$

Where G^c is the corrected Gini; a is the external data-based Pareto coefficient (for Pareto I); β is the top population share suffering from upper tail issues considered (e.g., $\beta = 0.01$

[58] Van der Weide, Lakner and Ianchovichina (2018) present decomposition formulas for the Gini, the Theil index and the mean log deviation.

for the top 1 percent); [59] S is the tax-based top percent income share (e.g., the top 1 percent's income share); G^{sb} is the survey-based Gini coefficient for the bottom (1 – β) percent of the population (e.g., the 99 percent); and, $S - \beta$ is the between group inequality.[60]

Due to the properties of the Pareto I function, the α coefficient can be calculated by simply dividing the minimum income of the first centile in the external data by the average income in that data. Since tax data typically represents a fragment of a country's population and income, computing S requires a reference total population and a reference total income (such as total national household income). With the total population control, one can determine the number of individuals in the tax data that belong to the top β percent. This information allows for the calculation of the income corresponding to the β population share in the tax data. Finally, S is obtained by dividing this income by the total income control (Atkinson, 2007).

In some instances, tax records may not be available or may be unreliable for use in semiparametric replacing. In such cases, authors have explored using other sources of external data to predict top incomes. In Table 3, this approach is called Regression-based Top Incomes Prediction. In their study for Egypt, Van der Weide, Lakner and Ianchovichina (2018) explore using house prices as a predictor of top incomes.

As a first step, they utilize a real estate database to verify that house prices follow a Pareto I distribution, a prerequisite for applying the proposed semiparametric method. [61] To replace the upper tail income in the survey, they proceed as follows. First, they run a regression model where total household income is a function of rental payments (self-

[59] To avoid confusion, the reader is reminded that Alvaredo (2011) and Jenkins (2017) use the symbol β for the inverted-Pareto coefficient. I decided to use b instead to keep the symbol β to denote the upper tail share that is affected by one or more of the causes of the missing rich to follow the notation by Cowell and Flachaire (2015). To avoid confusion, the reader is reminded that Alvaredo (2011) and Jenkins (2017) use the symbol β for the inverted Pareto coefficient. To maintain consistency with Cowell and Flachaire (2015), we have chosen to use β to denote the upper tail share instead.

[60] This formula has been frequently applied in the literature for a variety of approaches. See, for example, Alvaredo (2011), Atkinson et al. (2011), Alvaredo and Londoño (2013), Diaz-Bazan (2015), Anand and Segal (2015), Lakner and Milanovic (2016), Jenkins (2017), Bourguignon (2018).

[61] Information on houses for sale is obtained from two real estate firms for the years 2014 and 2015, while the household income and covariates data come from the Household Income, Expenditure and Consumption from 2008/9.

reported rent for owner-occupied housing or paid rent) and other covariates. [62] This regression estimates a coefficient that links rental prices to incomes. Using external information on house prices and rents, they obtain a conversion factor between the two. Assuming a constant relationship between rents and house prices, house prices are converted into equivalent rental values. These imputed rental values are then used to predict a new set of incomes using the regression. These incomes are compared to the reported incomes in the household survey to identify the threshold above which top incomes are likely underrepresented or underreported. The researchers define this cut-off point as the point where the income distribution transitions to a Pareto distribution and apply a correction to the survey data accordingly. They then fit a Pareto I distribution to the predicted incomes for the top 5 percent of the income distribution. Finally, they estimate corrected inequality indices using the corresponding decomposition formulas (Alvaredo, 2011; Shorrocks, 1980). After the correction, the Gini coefficient rises from 0.39 to 0.52, and the income share of the top 1 percent increases from 8.9 percent to 15.1 percent.

As the authors note, the main challenges are collecting house price information from diverse sources and ensuring the dataset is nationally representative, as it may be biased towards larger urban centers. They suggest that this approach could be adapted to other databases containing predictors of top incomes, such as mortgage or credit card statements, when tax records are unavailable.

In their attempt to produce global inequality estimates among individuals including as many countries as possible, Lakner and Milanovic (2016) propose an application of the semiparametric approach when the only available external data are consumption totals in National Accounts since for many countries this may be the only external information that could be used, in principle, to correct for the surveys' misrepresentation of the wealthy individuals' incomes. In Table 3, their method is classified as a semiparametric method that for external data uses income totals.

[62] They alternatively fit a linear and a non-parametric kernel regression to the survey data, where total household income is a function of the logarithm of house prices and a set of covariates.

The authors combine 565 national surveys over five reference years between 1988 and 2008, using purchasing parity prices (PPP) to account for differences in the cost of living between countries. Each (country/year) observation is represented by the average income/consumption of ten income/consumption deciles, depending on the concept covered by each survey. The authors used data from 565 national income and consumption surveys conducted between 1988 and 2008. To account for differences in cost of living, they converted income/consumption from local currency units into purchasing power parity prices (PPP) to make the data comparable across countries. Instead of the surveys' microdata, they used average income/consumption by decile because for a number of countries, only the latter is available. Each country-year observation is represented by the average income/consumption of ten deciles.

Assuming that discrepancies between household surveys and National Accounts are not distribution neutral, the authors argue that household surveys provide accurate information for the lower 90 percent of the income distribution. Thus, they correct survey data by allocating the full gap between household final consumption in National Accounts and household surveys to the top 10 percent of the distribution. Then they obtain more disaggregated top quantiles by fitting a Pareto distribution to the upper tail. Assuming that discrepancies between household surveys and National Accounts are not distribution-neutral, the authors argue that household surveys provide accurate information for the bottom 90 percent of the income distribution. Therefore, they correct survey data by allocating the full gap between household final consumption in National Accounts and household surveys to the top 10 percent of the distribution.[63] The top 10 percent averages by decile are replaced with a Pareto distribution to be able to disaggregate the top decile into smaller fractiles.

[63] The survey data are mainly obtained from PovcalNet (75 percent) and complemented with information from the updated World Income Distribution (WID) data (Milanovic 2012), the Luxembourg Income Study, the British Household Panel Survey, the EUSILC and data from country statistical offices. The information from National Accounts corresponding to household final consumption expenditure comes from World Development Indicators for the survey years, complemented with data from the International Monetary Fund and national statistical offices.

The procedure consists of four steps. First, they choose a new mean income or consumption (depending on the variable collected in the household survey) for each country defined as the the survey mean or average consumption in National Accounts, whichever is higher. (In almost all countries, the latter is higher.) Second, they recalculate the decile income/consumption shares for all deciles except the top by taking the ratio of the original deciles' average income/consumption with respect to the new mean. If the new mean is higher than the survey's, then the share of the first nine deciles will shrink. Third, they calculate the new top decile share as the difference between 100 percent and the sum of the revised shares of the bottom nine deciles. Fourth, using the new top 10 and 20 percent shares, they perform a Pareto imputation to split the top 10 percent into smaller fractiles. This is possible because the Pareto coefficient can be arithmetically obtained using the formula described above in subsection 3.2.1. The only information needed is the minimum income and average income of the top decile.

The main results of the paper show that using only household survey information, the global Gini coefficient was 0.705 in 2008. After the correction, it rises to 0.759. Trends are also affected: the survey-based global Gini coefficient shows a decrease of 2 points over the period considered, while the downward trend disappears after the correction.

Bourguignon (2018) introduced another variant of semiparametric replacing when the only available external data are total incomes, such as those found in National Accounts. A key difference between his approach and that of Lakner and Milanovic is that, in addition to income underreporting, Bourguignon's approach assumes that the existing survey weights within the top are inaccurate due to the exclusion of the very wealthiest individuals from the survey. To accommodate this missing portion of the population, the original weights assigned to the top income earners in the survey must be adjusted downward.

Bourguignon demonstrates how certain properties of the Pareto I distribution can be leveraged in this context. Specifically, the formula for the inverted Pareto coefficient can be expressed as a function of the rescaling factor to address underreporting (the ratio of means in National Accounts to survey means), the share of the population in the upper tail (denoted here as β), and a parameter that represents the number of observations that

need to be added to the upper tail to account for the missing wealthiest individuals within the top (due to noncoverage errors or unit nonresponse, for instance).[64] In Table 3, Bourguignon's approach is classified as semiparametric replacing with reweighting within the top.

4.1.2 Nonparametric

4.1.2.1 Within-survey Imputation

Several nonparametric imputation approaches have been developed to address incomplete data throughout surveys, including item nonresponse and issues related to the upper tail. These methods are categorized as single and multiple imputation methods (Little and Rubin, 2014). Within single imputation methods, Little and Rubin further distinguish between explicit and implicit modeling methods. The hot deck imputation and substitution methods fall under the category of implicit modeling methods for within-survey corrections.

In hot deck imputation, missing income observations in a survey are replaced with values from similar units known as "donors." As missing observations are not randomly distributed, leveraging information on covariates can help reduce item nonresponse bias.[65] Substitution methods involve using cases not included in the sample but available in the sample frame. The explicit modeling methods include mean imputation (unconditional and conditional), regression imputation, and stochastic regression imputation. Pure mean imputation involves replacing missing income data with the mean (or median or mode) of the entire sample population. If income nonresponse increases with income, this method will yield biased results. This bias can be partially reduced if the sample-based means correspond to a relatively homogenous category (e.g., by gender, age, education, etc.). However, in the case of the Current Population

[64] To avoid confusion, it is worth clarifying that Bourguignon uses the symbol β for the inverted Pareto coefficient. Recall that here β refers to the upper tail that is subject to correction.

[65] Since missing observations are not randomly distributed, information on covariates can be leveraged to reduce nonresponse bias. One example is the nearest neighbor hot deck imputation method. In this approach, observations are sorted by a relevant sociodemographic variable (e.g., gender, education, race). If income is not missing for the first case in the sorted list, it is stored as the "hot deck" value. For subsequent cases with missing income, the stored "hot deck" value is used as an imputed value. This process is repeated until all missing income values are replaced by the most recent reported value for a "neighbor" in the sorted list (Groves et al., 2009, p. 359).

Survey in the United States, some studies have shown that the probability of a close match declines when characteristics of individuals are less common (Lillard, Smith, and Welch, 1986; Bollinger and Hirsch, 2006). Potentially, this could make it harder to use this type of imputation methods for wealthy individuals.

The regression imputation method predicts missing values by estimating coefficients from a regression model where the variable of interest (in this case, income) is the dependent variable, and covariates are a set of observed variables. This method assumes a linear relationship between missing and nonmissing data and that, conditional on the observed variables, the missing observations are missing at random. Additionally, the range and spread of the imputed data can be limited as it depends primarily on the observed variables. Stochastic regression imputation attempts to address this limitation by introducing randomness into the imputation process. After fitting the regression model, a random component is added to the prediction, drawing errors from the regression residuals. However, the method's effectiveness is sensitive to the chosen model and the availability of relevant variables in the dataset. Examples of this procedure and a general discussion of different imputation methods for the U.S. Current Population Survey can be found in Bollinger and Hirsch (2006).

In contrast to single imputation methods, which replace each missing value with a single imputed value, the multiple imputation method initially proposed by Rubin (1987) involves creating "multiple imputed datasets, each one based on a different realization of an imputation model for each item imputed" (Groves et al., 2009, p. 359). Little and Rubin (2014) extensively discuss the advantages of multiple imputation and the proper protocols to be followed. [66] This approach is used to address item nonresponse in household finance and wealth surveys, as illustrated by Sanroman and Santos (2019) for Chile, Spain, Uruguay, and the US.

Researchers have often relied on the hot deck and other within-survey imputation methods to address income nonresponse. While this approach offers certain advantages, it has limitations when dealing with underreporting of top incomes.

[66] Some of the imputation methods have been shown to be less reliable. For a discussion of their advantages and limitations see, for example, chapter 5 in Little and Rubin (2014).

Correction methods that primarily use information contained in surveys may not effectively address underreporting. For example, if the respondent population systematically underreports income, the imputation method will likely propagate this underreporting, resulting in biased inequality estimates.

To address this limitation, imputation methods have been extended to include the so-called cold deck method, which utilizes external data sources. While nonparametric imputation methods were originally designed for within-survey corrections, they can also be effectively used in conjunction with external sources.

4.1.2.2 Rescaling with External Data

If tax microdata is available, a typical rescaling exercise consists of the following. First, both survey and tax data are divided into 100 cells, with each cell representing 1 percent of the respective dataset. The means for each cell are calculated for both the survey and tax data, and the means are compared. Starting at the point where the ratio of tax-based mean to survey-based mean exceeds unity, survey incomes are scaled up so that they align with the tax data mean. To generate the complete distribution, all incomes within the pre-specified cells are scaled up by the ratio of the two means. Below the point in which the ratio is less than unity, survey incomes are assumed to be accurate. This approach is a type of cold deck imputation, where missing or underreported values in the survey are replaced by values from an external source, in this case, tax data.[67] This method has been applied by, for example, Piketty, Yang and Zucman (2019) and Flachaire, Lustig and Vigorito (2023).

In some cases, tax data start to be reliable above the point where survey data are reliable, as argued by Piketty, Yang and Zucman (2019) in the case of China. In this study, the authors adjust China´s household survey assuming that it is reliable up to the percentile 90 and the fiscal data are reliable above the 99.5 percentile. Using the generalized Pareto

[67] Little and Rubin (2014) describe the latter as the method in which a missing (or underreported) value of an item in the survey is replaced by a value from an external source. Little and Rubin, 2014, location 1682 in ebook.

interpolation on the tax data, they create percentiles and compute quantile ratio upgrade factors that increase linearly between the 90 and 99.5 percentiles.

In the past, when tax data was not readily available, a typical rescaling factor was the ratio of wage and capital income means in National Accounts to the corresponding means derived from survey data, ensuring that the survey totals aligned with the National Accounts total. Several decades ago, Oscar Altimir (Altimir,1987), an economist at the United Nations Economic Commission for Latin America and the Caribbean (UNECLAC), proposed an approach to address underreporting in surveys that was applied by this agency until 2019 (ECLAC, 2018). [68] This method utilized National Accounts aggregates as control totals for household incomes by source. In essence, it involved grossing up wage incomes by the ratio of the wage bill in National Accounts to the survey's wage total wages. Incomes from capital were similarly grossed up, but only for the richest 20 percent of the population. Compared with unadjusted estimates, adjusted inequality measures were—by construction—higher, and poverty measures lower. Because the ratios could change by year, trends derived from adjusted data could also differ from trends observed with unadjusted data. Bourguignon (2015) extensively discusses the limitations of this method, highlighting one crucial drawback: the proportional adjustment that had been applied by UNECLAC assumes that underreporting and other issues are not correlated with income and ignores the heterogeneity within wage-income and capital-income categories.[69]

Some authors are skeptical about using National Accounts to correct surveys because National Accounts themselves may be subject to significant measurement errors (Deaton, 2005). However, as argued by Bourguignon (2018), National Accounts may be the only available external data source in certain cases. In his paper, Bourguignon proposes a method to rescale survey incomes to align with National Accounts. Beyond

[68] Although the method was proposed much earlier, an English version of his approach was published in Altimir (1987). In 2018, UNECLAC discontinued using this method and directly estimates inequality and poverty indicators from survey data (ECLAC, 2018). However, as argued by Bourguignon (2018), National Accounts may be the only other data available so different allocation methods of the gap could be explored.

[69] Bravo and Valderrama (2011) showed that in the case of Chile, this adjustment led to an overestimation of inequality.

the traditional proportional rescaling approach, he explores two alternative allocation rules. These include an "egalitarian adjustment" that adds the same income amount to everyone above a certain threshold and a "progressive adjustment" that increases the allocated income linearly with income. He illustrates these options with an application to Mexico. This paper shows how to generate the corrected Lorenz curve and the corrected Gini coefficient for each case (using a variant of the decomposition formula described above).

Bourguignon (2018) proposes a more general approach to rescaling in which a proportional adjustment is just one possibility. The method consists of allocating the gap (or a portion of this gap) between the mean of the income total in the external source and the income mean obtained from the survey to the observations in the upper tail. There are three allocation options: egalitarian (everybody's income is rescaled by the same amount), proportional (everybody's income is rescaled by the same proportion), and linearly progressive (the rescaling factor increases linearly with income). In the empirical application with Mexican data, the original Gini coefficient is 0.51. After rescaling, it rises to 0.587, 0.593, and 0.600, respectively, using the egalitarian, proportional, and linearly progressive adjustments.

It is important to note that this correction approach is not equivalent to Distributional National Accounts (DINA). Although DINA employs the rescaling method, its primary objective is to allocate all components of the National Accounts, including undistributed profits and government spending on education, health, defense, etc., to households to generate a distribution of income that is consistent with macroeconomic aggregates.

4.1.2.3 Statistical Matching with External Data

Bach, Corneo, and Steiner (2009) merge the German Socio-Economic Panel (SOEP) with a 10 percent sample of individual tax returns provided by the tax authorities. This sample includes all taxpayers within the top percentile which allows the authors to analyze top income shares with granularity. The analysis focuses on gross market income at the individual level for the population aged 20 or older.

To merge these datasets, they employ a constrained matching approach. This approach selects a number of tax records for each potential taxpayer in the survey, ensuring

consistency between the weighting factors of both datasets. The matching procedure relies on a network simplex algorithm to solve the resulting linear programming problems. The matching process is based on income sources, occupational status, and a set of demographic variables. Since the survey represents a larger proportion of the population than the tax records, not all SOEP observations are matched to tax records. Additionally, for individuals receiving interest or dividends below the minimum taxable income derived from savings, their capital income from the SOEP is used as tax records for this group are likely underreported.

The authors compare their merged data with National Accounts totals by income category and find discrepancies in some categories. These discrepancies are attributed to the lack of reliability of National Accounts data regarding business income and how property income and capital gains are recorded in personal income tax records.

In 1992, the Gini coefficient was 0.5973 in the survey and rose to 0.6155 after the correction procedure. In turn, the shares of the top 10 and 1 percent were 35.24 and 6.66 percent in the survey and 39.04 and 11.23 percent after the correction. While the evolution of the Gini coefficient and the share of the top decile are similar in both the survey and the integrated database, the income share of top incomes is considerably higher in the integrated database. For instance, the income share of the top 0.01 percent is 46.6 percent higher in the integrated database.

The authors correctly argue that this procedure has advantages over other imputation methods such as semiparametric models that use external data because it preserves the correlation structure between variables in the survey and ensures that the common matching variables are maintained. Because the structure of constraints can influence the matching process, some survey observations may end up not being matched to an observation in tax records. Like all imputation methods, matching relies on assumptions about unobserved variables and can be susceptible to bias. Finally, the computational resources required can be significant, depending on the data and the number of constraints.

4.1.2.4 Replacing with Linked External Microdata

Another approach to integrating datasets involves what could be termed "full replacement." In such cases, all or a subset of income observations from the survey are replaced with external microdata, such as that from social security or tax records (Bollinger et al., 2019; Hyslop and Townsend, 2020; Jenkins and Rios Avila, 2020; Flachaire et al., 2023).[70] These papers primarily utilize linked, individual-level data from surveys. While they address item nonresponse and measurement error, data limitations preclude them from addressing unit nonresponse.

Bollinger et al. (2019) create a hybrid earnings distribution by substituting missing earnings data from the US Current Population Survey with information from Social Security records for linked cases. For unlinked individuals, they retain the original survey data. To separately identify the impact of item nonresponse and measurement error (income misreporting) on inequality measures, they generated two hybrid distributions. First, they replaced earnings data for all linked individuals, including those with earnings nonresponse. Second, they replaced earnings data only for linked individuals who reported earnings data.[71] The Gini coefficient, calculated using only survey data, was 0.454 in 2005 and 0.455 in 2010. With the hybrid distribution, the Gini coefficient was 0.510 in 2005 and 0.580 in 2010. The authors concluded that earnings inequality levels and the change in inequality are underestimated when based solely on survey data. This

[70] Bollinger et al. (2019) contribute to the earnings validation literature. This field explores how item nonresponse and measurement error affect earnings indicators in surveys, the challenges of merging processes, and the limitations of administrative data. The initial wave of validation studies assumed survey data to be incorrect and administrative data to be correct, generally finding mean reversion processes (Gottschalk and Huynh, 2010). The second wave posits that neither database is entirely correct and recommends that statistical offices develop hybrid measures of "true" earnings based on estimating bivariate mixture models (Kapteyn and Ypma (2007);Meijer, Rohwedder, and Wansbeek, 2012; Abowd and Stinton, 2013; Hyslop and Townsend (2020); Jenkins and Rios Avila, 2020). While this literature is extensive, due to the focus of this review, we concentrate on papers addressing how different combination processes between survey and external data affect inequality measures. For an overview of the validation literature, see Bollinger et al. (2019) and Jenkins and Rios Avila (2020).

[71] To assess the sensitivity of results, the authors also generate a second hybrid earnings distribution by substituting survey earnings only for the top 50 percent of linked observations.

conclusion remained consistent when analyzing item nonresponse and measurement error separately.

Using Uruguayan data, Flachaire et al. (2023) construct an income variable, termed the "true distribution", by starting with the survey income and replacing it with tax record data where the ratio of average survey income to tax income is less than 1. This threshold closely aligns with taxable income. While mathematically equivalent to the rescaling method discussed earlier for continuous functions, this equivalence does not hold for discrete functions (Flachaire et al., 2023). They find that the survey Gini coefficient varies from 0.382 to 0.461 after correction, while the top 1 percent share increases from 0.068 to 0.104.

While integrated datasets are an attractive option to address the missing rich problem, the integration of any two datasets into one requires a careful process of assessing the accuracy of administrative information and its consistency with survey data. The quality of the linking procedure also needs in-depth analysis (Kapteyn and Ypma, 2007; Meijer, Rohwedder, and Wansbeek, 2012; Bollinger et al., 2019; Jenkins and Rios-Avila, 2020). These studies apply methods that permit identifying measurement errors in both surveys and administrative data. They find that income misreporting exists in both. These studies challenge the idea that administrative records are better than household self-reported survey data across the board. Based on this, the authors argue that combining information from both sources will make better use of the available data.[72]

[72] Bollinger et al. (2019) estimate the probability of mismatch in the linked data to be around 3 percent in the case of the United States Current Population Survey and the Internal Revenue Service. This figure rises to 6 percent for the United Kingdom according to the estimates by Jenkins and Rios-Avila (2021) based on data from the Family Resources Survey and the Internal Revenue Service. This study also identifies measurement errors both in survey data (93 percent) and administrative records (33 percent). Besides errors in the matching process, using mixture factor models that generalize Kapteyn and Ypma's (2007) models, they classify individuals into two latent classes according to which type of error their earning measures contain and they also check desirability bias in the survey, differences in reference periods between both data sources, and measurement errors in tax records (we return to this in the section on tax data).

4.1.2.5 Reweighting within the Top with External Data

Another approach involves keeping the observations in the upper tail intact but adjusting their weights to align with the population shares observed in external data (Medeiros, Souza, and Castro, 2015; Medeiros, de Castro, and Azevedo, 2018). For example, in their study for Brazil, Medeiros et al. (2018) reweight the top 2.5 percent of the Census Sample Survey using information from tax records. To identify the starting point for changing the weights in the upper tail (that is, the threshold), they compare the income ratios by quantile from both datasets and find that the ratio is larger than unity at p95. The authors try two different thresholds above which weights within the upper tail are recalibrated to match the population shares observed in the tax data: 95 percent and 97.5 percent. They opt to use 97.5 percent because it preserves a larger section of the survey intact. To reweight within the top 2.5 percent, they create 0.5 percent intervals and multiply all individuals within that interval by the same recalibration factor. These reweighting factors are calculated as the ratio between the population within a certain interval in both data sources. [73]

As mentioned above, it might seem surprising to include a reweighting option within the replacing method. However, it's important to remember that as long as the overall share of the top income group (β) and the remaining portion of the distribution (1 - β) remain unchanged, the method should still be classified as replacing. In addition, it should be remembered that with semiparametric replacing the weights *within* the top can implicitly change: the density within the top after fitting a statistical function such as the Pareto distribution will be different than the one observed in the original sample.

Recall also that if the sample and the external source have common support, reweighting within the top will be equivalent to the rescaling method described above.

[73] As the Brazilian statistical office corrects item non-response using hot-deck (10 percent of the sample), the authors also calculated calibration factors for reweighting only information that was effectively observed. The factors with these restrictions are 19 percent higher that the unconstrained ones.

4.2 Reweighting

If the household survey suffers from non-coverage error or unit nonresponse—that is, lower participation rates—at the top, the weights for the upper tail and the rest of the distribution in the survey might be incorrect. [74] In such a case, one needs to go beyond focusing on the right-hand tail and adjust weights in other segments of the survey. In particular, the weights in the nontop portion of the sample (or, at least, in parts of it) must be adjusted downwards to accommodate additional individuals at the top of the distribution. [75] This method is known as reweighting and corrects for the missing rich problem by adding people to the entire right-hand tail of the sample. In reweighting, the original observations are kept intact while weights are recalibrated. Recall that a crucial assumption underlying the reweighting method is that there is common support between the sample and the target population (Table 1).

4.2.1 Within-survey

As described in Biemer and Christ (2008) and Little and Rubin (2014), reweighting involves adjusting the expansion factors—also known as base weights—assigned to the complete cases in a sample (i.e., cases with unit or item nonresponse in the available-case sample are discarded) using new weights that account for, in particular, unit nonresponse (that is, where all the survey items are missing for particular subjects in the sample but not in the frame). Information on respondents and nonrespondents, such as their geographic location, age, gender, and other relevant characteristics, available from survey producers (e.g., national statistical offices), can be used to assign these new weights.

Although the within-survey replacing and reweighting methods are completely different, Bourguignon (2018) reminds us that, as long as the target population and the sample have the same support (that is, there is point-mass at all points in both distributions), the results obtained by correcting via reweighting can always find its equivalent with

[74] See Biemer and Christ (2008) for an overview of the reasons for reweighting and the corresponding methods.

[75] As indicated, this is different from reweighting within the top where, the weights for the nontop and within the latter are not changed. All the recalibration of weights takes place within the top.

rescaling, one of the nonparametric replacing methods. That is, every reweighting exercise, in theory, can be converted into a replacing exercise that will yield the same result, and vice versa.

As with replacing, reweighting can be conducted within a survey or by combining survey data with external information. There are at least two main within-survey reweighting methods: weighting class adjustment and model weight adjustment.

4.2.1.1 Weighting Class Adjustment

In the weighting class adjustment method, the base weights associated to each observation are recalibrated using nonresponse rates by category to obtain the new weights. Atkinson and Micklewright (1983) apply this method to the UK total income and its components adjusting nonresponse rates by region of the Family Expenditure Survey using the response rates provided by Kemsley (1975), Kemsley et al. (1980) and the Northern Ireland FES. They also carry out a second adjustment considering differential nonresponse by age. The main caveat of this method is that it assumes that respondents and nonrespondents are similar within each geographic/age group. Harris (1977) uses a similar procedure using educational groups.

4.2.1.2 Model Weight Adjustment

Mistiaen and Ravallion (2003) and Korinek, Mistiaen, and Ravallion (2006, 2007) develop a variant of the model weight adjustment that allows the probability of nonresponse rates to vary with income within geographic areas. In this way, the decision to respond is not assumed to be independent of the variable of interest. They model the determinants of income nonresponse and apply it to the US Current Population Survey. In this case, information from respondents and nonrespondents at the Primary Sampling Unit available from survey producers (e.g., national statistical offices) can be used to assign new weights. The method assumes that all the households with the same characteristics (including income per capita) will have the same probability of responding to the survey across geographic areas. Using a GMM estimation the number of nonresponding households for each income interval of interest and region is estimated and new weights are obtained. Details about the method can be found above and in Appendix 2.

Hlasny and Verme (2018a, 2018b, 2021) apply this method using household surveys for Egypt, the European Union and the US. They use nonresponse rates by geographic area instead of Primary Sampling Unit. In fact, the method can be applied using nonresponse rates by other sociodemographic characteristics such as age, gender, and education.

4.2.2 Combining Survey and External Data

4.2.2.1 Poststratification

Researchers and statistical offices have also used weights obtained from external information in lieu of the base weights of the achieved sample. In this method, the original expansion factors or base weights are substituted with new weights derived from population control totals by age, sex, region, etc., obtained from external administrative registries such as tax and social security records.

This method is applied by Campos-Vazquez and Lustig (2019) to correct for item nonresponse in the Mexican Labor Force Survey (ENOE) from 2006-2017. They found that item nonresponse in income data was around 33 percent both for formal and informal workers and increased over time. To address this issue, they recalibrate the survey weights using social security information to align the distribution of formal workers by category in the survey with that observed in social security records.[76] The categories considered are defined by multiples of the minimum wage, gender, and age. This is because social security records provide information on the number of formal workers based on these characteristics. Using only survey data, the labor income Gini coefficient declined from 0.424 in 2006 to 0.382 in 2017. After the correction, the decline was much smaller.

In the above methods, there is no predefined income threshold above which the weight of the upper tail needs to be upweighted. Other proposed reweighting methods select a threshold beforehand.

Reweighting with external data instead of within-survey model weight adjustment may be an option when, for instance, one cannot obtain unit nonresponse rates by PSU or

[76] The income of informal workers is recovered using hot deck imputations.

geographic area. It may also make sense if one suspects that unit nonresponse within a PSU is correlated with income in such a way that within-survey reweighting methods may not be able to detect or correct for income-related nonparticipation in surveys.

4.2.2.2 Reweighting with Exogenous Threshold

Based on Uruguayan linked data, Flachaire et al. (2023) assess changes in inequality measures using the reweighting method to recover the upper tail above different predefined income thresholds. In the case of the Gini coefficient, they find that the lower the threshold, the higher the corrected inequality estimate. The Gini coefficient is 0.382 in the survey and rises to 0.458 when the threshold is p30, to 0.443 when p50 is chosen, and to 0.442 when p90 is chosen. The income share of the top 1 percent increases after correction but the increase is not systematically inversely related with the threshold and the differences between shares associated with alternative thresholds is significantly smaller. The income share is 6.8 percent for the uncorrected survey and 9.9, 9.7 and 10 percent thresholds p30, p50 and p90.

Bourguignon (2018) proposes a method to reweight the upper tail when the only external information available are totals from National Account. He predefines an income threshold above which the right-hand tail of the distribution needs to be upweighted. In his empirical application for Mexico, the Gini coefficient rises from 0.510 to 0.549.

4.2.2.3 Reweighting with Endogenous Threshold

Although the reweighting component can be applied by itself, the full-scale method proposed by Blanchet, Flores and Morgan (2022) combines reweighting and replacing so the details are described in subsection 4.3. De Rosa, Flores and Morgan (2024) is an example of applying the reweighting component only.

4.3 Reweighting and Replacing

4.3.1 Within-survey

4.3.1.1 Model Weight Adjustment Reweighting and Semiparametric Replacing

After correcting sampling weights for item nonresponse using the model weight adjustment method, Hlasny and Verme (2018b; 2021) estimate Pareto type I and II, and a generalized Beta type 2 distributions to correct measurement error in the reweighted

data, using different thresholds. Then they compute the semiparametric Gini coefficient using the reweighted survey data for the nontop incomes and the Pareto-replaced top tail in a reweighted income distribution.

The authors conclude from their empirical exercises for Egypt and the US that reweighting is more stable than replacing because the latter is sensitive to the shape of the income distribution found in between the lower and upper thresholds as well as to the arbitrarily set threshold (as discussed above under semiparametric replacing).

They also find that the joint correction method mitigates implausible corrections that can occur with the replacing method. Additionally, they observe that in most years, corrections in the Gini index are larger when both methods are used than with reweighting alone, but smaller that under replacing alone.

4.3.2 Combining Survey and External Data

4.3.2.1 Reweighting with Exogenous Threshold and Semiparametric Replacing

Some authors consider that the sample completely excludes the upper tail so the "threshold" is in effect the last observation in the survey. In this case, reweighting the survey throughout is a necessary step. The weights of the whole survey are compressed to make room for the additional tail that is presumed to include the observations of rich individuals that need to be “added” to complete the distribution. This method could be interpreted as an extreme form of poststratification because the whole achieved sample is assumed not to represent the target population but a subset of the latter. After uniformly downweighting the whole survey, the new nonexistent upper tail is replaced by a new upper tail obtained by fitting a model on external data such as tax records. Anand and Segal (2015) apply this method to adjust inequality measures around the globe. They cross-checked their top income estimates against data from rich lists, assessing their country and income source composition against Forbes list.[77]

[77] Several organizations, like Forbes and Credit Suisse, create lists of billionaires. Individuals on the Forbes list have a net worth of more than one billion current dollars in any given year. Wealth estimations are based on an examination of potential billionaires' holdings and transactions, which include investments in businesses, property and real estate, art, and cash, among others. These lists are used to estimate Pareto distributions to adjust wealth inequality estimates based on wealth surveys.

4.3.2.2 Reweighting with Endogenous Threshold and Rescaling

Blanchet, Flores and Morgan (2022) propose a method that uses survey and tax records and combines reweighting and replacing. This method is used by the DINA (Distributional National Accounts) project at the Paris School of Economics and is the first step in producing the survey-based inequality indicators housed in WID.World. For details, see the February 27, 2024 version of the Distributional National Accounts Guidelines (Blanchet et al., 2024).

The authors argue that the method distinguishes itself because it proposes a data-driven threshold selection process where the continuity of the corrected income density is ensured by the method itself and, under certain assumptions, covariates can be recovered.[78] After identifying the point above which tax data appears reliable, the method has three main steps: first, select the threshold; second, reweight the sample weights; and third, replace incomes at the top with a parametric function.[79]

The threshold in this approach is the income level in the survey above which it is presumed that the survey underrepresents the upper tail in the target population, and it is selected as follows. First, calculate the ratio of the cumulative density functions of the survey to tax records and the ratio of the density functions of the survey to tax records and identify the points where these ratios are equal (that is, where these two curves cross). There can be one point or more where these two ratios are equal. The authors call them merging points. If there is more than one merging point, the threshold is set at the maximum merging point. Selecting the threshold as one of the merging points helps ensure continuity between the two functions. While other merging points may exist, the method chooses the maximum merging point at which these ratios coincide to preserve the original survey data as much as possible.

Once the threshold has been selected, the reweighting step follows. The new weight of the upper tail --that is, the $\beta^c > \beta$ --is defined as the population share with incomes above

[78] In the other methods, the continuity may not happen and hence in those cases one relies on the formula presented by Alvaredo (2011).

[79] Identifying the point above which tax data is reliable is especially recommended for countries with large pockets of informality, which introduces noise at the bottom end of tax data (note that the starting point of the "bottom" depends on the country). The tax data below the trusted section is removed.

that threshold in the tax data. The weights within the upper tail are adjusted to reproduce the distribution of income observed in the tax records. The survey weights below the threshold are uniformly adjusted downwards so that the corrected weights sum to unity. In other words, in the reweighting step, the weight of the whole upper tail is increased and the weights within the upper tail are changed while the weights for the rest of the distribution are uniformly compressed.

The replacing step involves replacing survey observations above the threshold with observations with equivalent weight and rank in the tax distribution, that reproduce the distribution of income observed in the tax data and match survey covariates. For the replacement component they use a similar approach to the rescaling one implemented by Piketty, Yan and Zucman (2019), where cell means in the survey by fractile are substituted by the corresponding information in tax data. To improve the precision at the top of the distribution (sampling error), they use a generalized Pareto interpolation in the tax data (Blanchet, Fournier and Piketty, 2022).[80]

After implementing the correction in their empirical application, the authors find that the income share of the top 1 percent increases by 10 percentage points in the case of Brazil, 8 percentage points in the case of Chile and 4 percentage points in the case of the UK. However, France and Norway experience small adjustments.

The authors rightfully emphasize that their method allows for the continuity of the combination of the distribution of income in the two sources. However, this is not the only method that allows this. Determining the threshold as the maximum merging point is not without problems. Using a synthetic true distribution (equivalent to the tax records in Blanchet, Flores and Morgan, 2022) and a simulated distribution that suffers from income-linked underreporting and item nonresponse, Flachaire et al. (2023) demonstrate that selecting the maximum merging point may be incorrect. In their experiment, the correct threshold—the one that allows for recovery of the true distribution after correction—coincides with the minimum merging point. This is crucial

[80] While this approach uses a parametric model to replace the uppermost section of the upper tail, we classified as nonparametric replacing. Same was done in the case of Flachaire et al. (2023). The rationale being that the parametric function is used to address sparsity at the very top of the top.

because the selection of the merging point significantly impacts the magnitude of the corrected inequality estimates. This should serve as a warning against a mechanical application of this method, or any other method that selects a single threshold without any sensitivity analysis.

The authors argue that, in contrast to other methods, their approach allows for the recovery of the entire microdata and the preservation of covariates. First, this is not the only method that allows this. Any reweighting approach allows for the covariates to be preserved. Same is true for imputation methods such as matching, rescaling and the typical within-survey imputation methods (see Table 3). Although an advantage of rescaling methods is that, in principle, one can recover the corrected microdata and not just the distribution, there is an implicit assumption for this to be the case, as Blanchet, Flores and Morgan (2022) acknowledge. The assumption is that the ranking of observations in the corrected distribution is identical to the ranking of observations in the original sample. In other words, rescaling, for instance, assumes that misreporting is uniform within the cells that are upscaled. If misreporting within the cells in the true distribution is not uniform, then the ranking after correction could change. This could happen if misreporting within cells depends on covariates for example.

4.3.2.3 Nonparametric Replacing and Reweighting with Exogenous Threshold

In the above, the sequence of the combined approach is to recalibrate weights first and proceed to replace the upper tail subsequently. The inverse sequence has been applied too: rescale incomes first and then recalibrate weights. For instance, this is the practice followed by the UK's Department of Work and Pensions to correct the household survey (the HBAI) used to produce the official inequality and poverty figures (DWP, 2015, Burkhauser et al., 2018). In this approach, the first step is to identify top income individuals in the survey, using a threshold that varies by income source and region. This definition is based on observing incomes in the survey that are considered to be very volatile, and hence, the threshold varies by year. After that, the income of this very rich individuals in the survey is replaced by the mean of their group in tax records. Finally, the number of top income individuals is estimated from tax data and survey weights are recalibrated. In 2014/15, the unadjusted Gini coefficient was 0.324 and rose to 0.338 after the correction. Burkhauser et al., (2018) raise important concerns on this procedure

highlighting the method used to identify top incomes, the effects of the lags in the availability of tax data, and the fact that external users cannot reproduce it among others.

Bourguignon (2018) implements reweighting and replacing using Mexican data, assuming that no information is available about the income distribution beyond the survey data, except for the National Accounts totals. Like Hlasny and Verme (2018b; 2021), he concludes that combining both methods results in inequality estimates that are higher than those obtained using only reweighting, but lower than those obtained implementing only rescaling.

5 Selecting the Correction Approach

5.1 Sensitivity of Inequality Measures to Correction Approach

Corrected inequality measures can be highly sensitive to the correction approach. This is particularly true for summary inequality indexes such as the Gini coefficient, the sensitivity is the highest. Top income shares, in contrast, tend to be less sensitive to the correction approach. For top fractiles, the impact will depend on the proportion of added observations and added income allocated to the top.

This should not come as a surprise, since, as argued by Deaton (2005) and shown in Appendix 4 for the Gini coefficient, it is not possible to predict ex ante the direction of change after correction, much less the orders of magnitude. The total derivative formula reveals how both depend on the data. For the replacing method, the direction of change and order of magnitude will depend on whether inequality for the top portion is higher or lower after correction. If it is higher, then the corrected Gini coefficient will be higher than the uncorrected one. If it is lower, then it depends on the order of magnitude compared with other parameters. For reweighting (and reweighting and replacing combined), it depends on the extent to which the Gini coefficient for the top and nontop portions of the distribution, and the population and income shares of the top, change after corrections.

Table 4 shows the sensitivity of the Gini coefficient to the correction approach even when the same data is utilized for the few available cases in which such a comparison was implemented by the authors. For example, Hlasny and Verme (2021) applied within-survey replacing, reweighting, and reweighting and replacing combined to generate corrected inequality series for the United States. As expected, they found that the order

of magnitude of the difference between the corrected Gini and the uncorrected one differs with the method. They also found that the order of magnitude of the difference depends on the threshold. However, there is no pattern between the threshold's location (higher or lower in the distribution) and the order of magnitude of the correction. They found that the method producing the maximum (minimum) correction for the Gini coefficient changed depending on whether the threshold for upper tail issues was set at 1 percent or 5 percent of the distribution. Although it is often the case that the higher the threshold, the lower the difference between the corrected and uncorrected Gini, this is neither mathematically true (as the formula in Appendix 4 reveals) nor empirically pervasive.

Table 5 Sensitivity of Inequality Measures to Survey-based Correction Approach

Author	Country	External Data	Uncorrected	Threshold	REPLACING					REWEIGHTING			REWEIGHTING AND REPLACING (or vice versa)		
					Within-survey	Survey and External Data				Within-survey	Survey and External Data		Within-survey	Survey and External Data	
					Semiparametric	Semiparametric	Semiparametric	Nonparametric		Model Weight Adjustment	Exogenous Threshold	Endogenous Threshold	Model Weight Adjustment and Semiparametric	Endogenous Threshold and Rescaling	Nonparametric and Exogenous Threshold
							Reweighting Top with External Data	Rescaling Top Incomes	Replacing Top with External Data in Full						
Hlasny&Verme (2021)	US	none	0.4725	1%	0.491	-	-	-	-	**0.5038**	-	-	0.483	-	-
				5%	**0.5792**	-	-	-	-	0.5038	-		0.5226	-	-
Bourguignon (2018)	Mexico	Nat Acc	0.51	1%	-	-	0.599	**0.6**	-	-	0.549	-	-	-	0.587
Flachaire et al. (2023)	Uruguay	Tax	0.382	10%&72%	-	-		0.44	0.44	-	**0.442**	-	-	0.435	-
De Rosa et al (forth.)	Brazil	Tax	0.582	99%	0.581	0.605	-	-	-	-	-	0.646	-	**0.691**	-
	Chile	Tax	0.529		0.537	*0.536*	-	-	-	-	-	0.609	-	0.609	-
	Colombia	Tax	0.538		0.523	0.547	-	-	-	-	-	**0.652**	-	0.639	-
	Mexico	Tax	0.567		0.545	0.594	-	-	-	-	-	0.586	-	**0.633**	-
	Uruguay	Tax	0.505		0.522	0.554	-	-	-	-	-	0.561	-	**0.575**	-

Notes: In Flachaire et al. the 72 percent threshold corresponds to the second to last column, endogenous threshold and rescaling approach. The highest corrected Gini is shown in bold.

That the results by method can vary significantly is also found by Bourguignon (2018), who applies semiparametric replacing (in two variations), reweighting and replacing combined to Mexico's household survey with the objective of allocating the gap between per capita household income in the survey and National Accounts. In this case, the maximum correction is with one of the variants of the replacing method. However, this is not the case, for instance, in Flachaire, Lustig, and Vigorito (2023), who apply replacing, reweighting, and both combined to Uruguay's household survey using tax data to correct

the upper tail and found that pure reweighting yielded the highest correction (in the case of the Gini coefficient but not with top shares).

De Rosa, Lustig, and Martínez-Pabon (forthcoming) provide an even more illustrative example of how case-specific the correction methods' impact on inequality measures can be. They applied within-survey replacing, and replacing, reweighting and reweighting plus replacing to several Latin American countries, combining surveys with tax data. First, in some instances the corrected Gini coefficient is lower than the original one. This is the case with within-survey semiparametric replacing for Brazil, Colombia and Mexico. Their exercise shows how even with the same threshold, it is not always the case that semiparametric replacing with survey and tax data combined is higher than the equivalent within-survey correction, as in the case of Chile where the correction that uses tax data is lower. And, as with the other examples, the maximum correction is sometimes with reweighting and sometimes with reweighting and replacing. Moreover, although not shown on the table, at a threshold of 90 percent, in the case of Uruguay, the maximum correction occurs with semiparametric replacing.

5.2 Criteria for Selection of the Approach: Some Broad Guidelines

The above section and the examples presented in the Introduction are resounding evidence that the choice of approach is a decision that requires careful assessment of the factors that may be the cause of the missing rich. As mentioned, at least twenty-two approaches have been put into practice (Table 3). Selecting one of them mechanically (because the software or a certain dataset are easily available, for example) can exacerbate the problem of biased inequality measures. In addition to the underlying factors for the missing rich problem in the survey, the selection of the approach will be affected by the type of data that is available and the purpose for which the corrected inequality measures will be used. For some types of analysis, such as such as assessing inequality levels and trends or the redistributive impact of fiscal policy, prefiscal and postfiscal incomes are sufficient. However, to study the determinants of inequality, for instance, the microdata in full (incomes and covariates) is essential.

As shown in Table 1, each correction approach implies some key underlying assumptions. Two key assumptions are i) whether the weight of the upper tail is well-represented in the sample and ii) whether there is common support between the sample

and the target population. If the weight of the upper tail is presumed correct, then replacing is the method of choice. If there are reasons to believe that the weight of the upper tail in the survey is not representative of the target population, then reweighting takes precedence. If there is no common support, however, reweighting as a correction method will be insufficient because it keeps the observations in the sample intact. In the absence of common support, within-survey replacing as a correction method will be limited as well. Thus, if the weight of the upper tail in the survey is not representative of the weight of the upper tail in the target population and there is no common support, correcting the survey will need to combine reweighting and replacing, and utilize both survey and external data.

5.2.1 Assessing Underrepresentation of the Rich

If there is evidence that survey nonparticipation rises with income (nonrandom unit nonresponse), the weights in the survey (including the population share of the upper tail) may be incorrect so reweighting is a necessary step.

Ideally, to assess whether the upper tail in the survey adequately represents the upper tail in the target population one would like to check whether survey nonparticipation is higher for the rich. In order to do this, one could follow the procedure proposed by Korinek, Mistiaen and Ravallion (2006; 2007). Obtain the rate of unit nonresponse by Primary Sampling Unit (PSU) or the more disaggregated geographic unit possible within the sampling frame. Calculate the average income for each geographic unit based on the observations in the survey and then tabulate, plot or regress the rate of unit nonresponse against the average income. If the plot is not a horizontal line, for example, then unit nonresponse is not missing at random, and reweighting is a necessary correction step to reduce bias. Evidence of underrepresentation of the rich would be indicated if the plot is upward sloping or U-shaped.

In the absence of access to nonresponse rates by geographic unit, which the data providers may be reluctant to share, one common tool for analyzing nonresponse bias is to compare the respondent-based distribution of the variable of interest in the survey with the distribution from another more accurate source (Groves, 2006, p. 655). In their analysis for the UK, for example, Burkhauser et al. (2018) conclude that unit nonresponse does not seem to be a problem because the proportion of individuals above the income

threshold corresponding to the top 1 percent in the household survey is very similar to the proportion of individuals above that same income level in the tax data.

Within-survey reweighting to correct for the missing rich has a great advantage: it allows one to retain both the statistical integrity of the survey design (with implications for statistical inference) and the great many applications for micro-data files in distributional analysis (Ravallion, 2022). However, if support is not the same, the correction obtained through reweighting will be limited.

5.2.2 Assessing Common Support

As Table 1 indicates, a second key factor influencing the decision of which method and data to use is the assumption regarding common support. Formally, recall that we defined income as x and the mass at x in the density functions for the probabilistic distributions as $f_x(x)$ for the sample and $f_z(x)$ for the population. There is no common support between a sample and the target population when for some x, $f_x(x) = 0$ in the sample, whereas $f_z(x) > 0$ in the population.[81] The missing rich problem refers to when the latter occurs in the upper tail.

If there is no common support, there will not be any respondents in the uppermost right tail that can be used to impute values or upweighted to correct for others that are missing or have underreported their income. In other words, if none of the rich make it into the sample (or, even if they make it, they underreport their income) within-survey imputation methods (including the semiparametric ones) or reweighting (whether within survey or poststratifying with external data) will not be able to correct the bias in inequality measures in full. [82]

---

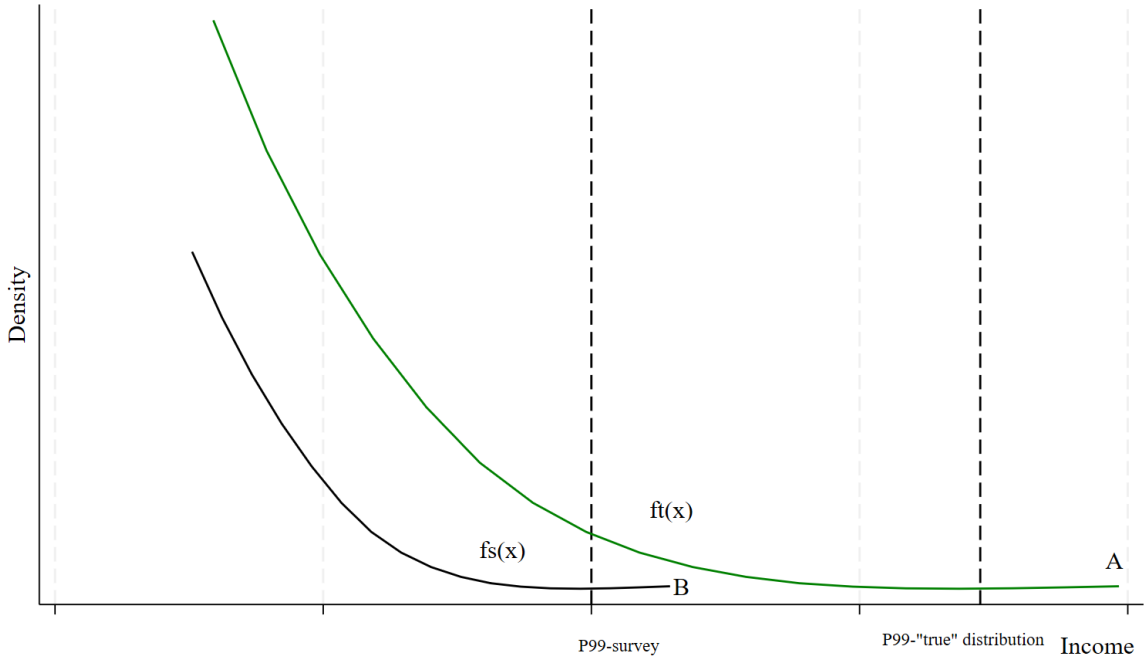
[81] For a discrete distribution, support is not the same when in the sample $p\,(X = x) = 0$, whereas $P(X = x) > 0$ in the population.

[82] If there is no common support, then the within-survey semiparametric approach will also yield a limited correction. Using the notation introduced above, in the parametrically replaced upper tail it is the case that $f_x^T\,(x) > 0$, for $x > x_{(1-\beta)}$ However, even if there is always positive mass at all points of the distribution by assumption, the magnitude of the mass at income x in the function replaced with a within-survey parametric model will be different from the one replaced with a parametric function estimated with, for example, tax data. That is, $f_s^T\,(x) \neq f_y^T\,(x)$. (Recall that the subscript s and y refer to within-survey and

In practice, to assess whether the common support assumption holds the comparison is not between the sample and the unknown true population but between the sample and an external source that is supposed to be closer to the true distribution in the upper tail such as tax data. To make comparisons between the sample and the external data and to apply correction methods that combine sources, the income concept must be compatibilized (see Appendix 3). For example, income obtained from tax data is taxable income for adults and for those adults who pay taxes; income in surveys includes taxable, nontaxable, and income from informal sources from individuals belonging to a household. Figure 2 illustrates the absence of common support in linked data for Uruguay (Flachaire et al., 2023). The figure is a depiction of the upper tail of the survey's (grey) and tax records' (green) density functions. The lack of common support is evident: for incomes between A and B there is no mass in the survey's density function.

survey and external data replaced upper tail, respectively.) In particular, the mass is likely to be higher for super high incomes in the model estimated with tax data because the latter is estimated from data that empirically include observations with incomes that are usually higher than the top incomes in a survey. Using Uruguayan linked data, Higgins, Lustig and Vigorito (2019) show, for example, that when implementing a within survey semiparametric correction, the share of the top 10 percent is 29 percent, which rises to 31.4 percent when using tax data. The results are very similar when fitting either the Pareto I or Pareto II distributions.

Figure 2 Absence of Common Support



Source: Author's elaboration based on linked data for Uruguay in Flachaire et al. (2023). This figure was inspired by the hypothetical example in Bourguignon (2017b).
Note: the grey line is based on the survey and the green one is based on tax records.

Determining whether there is common support can also be done heuristically by, for example, comparing the maximum incomes in the surveys with the typical incomes of high-earning individuals, such as CEOs of large companies. As mentioned, Szekely and Hilgard (2007) found that the income of the ten richest households in a sample of surveys for Latin America was roughly equal or even lower to the average wage of a manager of a medium to large size firm. [83]

Lack of common support can stem from any of the causes discussed: noncoverage error, unit nonresponse, item nonresponse, underreporting, and data preprocessing practices.

[83] Some correction approaches ignore the differences in the definitions of income. For example, those that combine surveys with rich lists or those that compress the surveys to represent the nonrich and resort to secondary data cannot carry out a careful harmonization process. For example, Anand and Segal (2017) acknowledge that one limitation of their world inequality estimates based on 129 countries might rely on non comparability, since they combine household survey data at the household level (per capita household income) with tax records information for individuals or tax units.

Any of these factors can result in the sample and the population lacking common support. Interestingly, as previously indicated, the assumption of the presence or absence of common support is not typically made explicitly. Furthermore, as Ravallion (2022) notes, when authors assume a lack of common support, they do not always provide empirical evidence.

If one is convinced that there is no common support, then all *within-survey* methods and reweighting may result in a limited correction of the missing rich problem. As stated, however, when recurring to external data such as tax records there is a cost: the corrected distribution will not be for per capita (or equivalized) household income but for the adult population and taxable (or taxed) income. Reconstructing income at the household level will not always be feasible.[84]

6 Conclusions

This paper presented a survey of the causes and correction approaches to address the "missing rich" problem in household surveys. "Missing rich" here has been used as a catch-all term for the main issues that affect the upper tail of the distribution of income: sparseness, undercoverage, unit and item nonresponse, underreporting and preprocessing practices by data providers such as top coding. Comparing top incomes in surveys with data from taxes or other sources reveals that the rich are not well captured in surveys. There is also evidence that surveys suffer from income-linked unit and item nonresponse. Upper tail issues can result in serious biases and imprecision of survey-based inequality measures. Hence the overriding importance of properly correcting the surveys for the sampling and nonsampling errors that affect the upper tail.

Several correction approaches have been proposed in the literature. In this review, we identified twenty-two distinct approaches involving unique combinations of replacing and reweighting methods and survey data with external data such as tax records or

[84] This problem does not arise when the external data is an income total from National Accounts (or other sources) and the correction consists of adding income or adding individuals or both so that the average income in the corrected survey matches the average in the external data.

National Accounts. Above we showed that inequality estimates can be highly sensitive to the approach.

Given the sensitivity of inequality estimates to the correction approach, as mentioned in the introduction, the question arises as to what criteria should be used to determine which methodological approach brings us closer to the true level of inequality and, therefore, to analyze its evolution and its relationship with other economic variables. Unfortunately, there are no readily available statistical tests or calibration mechanisms to make this determination.

Here we provided some broad guidelines to diagnose whether the wealthy are missing from the survey because the sample underrepresents them or the reported incomes are below the incomes of the target population, or both. This diagnosis is the first step to determine whether one needs to apply reweighting and/or replacing techniques, and whether it is advisable to combine survey with external data such as tax records. As for the specific method within the broad categories of replacing and reweighting, it is crucial to assess the advantages and limitations of each of the methods reviewed.

Looking into the future, a promising solution to the missing rich problem will likely come from linked data. Eventually, in countries with reliable administrative registries, the statistical offices themselves could pre-populate the income data for consenting individuals selected into the sample from registers (as it is done to some extent for France in the EU-SILC survey). Simultaneously, as suggested by Meyer and Mittag (2021), researchers could make use of linked data to correct for coverage errors, nonresponse, underreporting and other measurement errors by, whenever appropriate, substituting administrative for survey data. The potential of linked data to address upper tail (and other) issues is high.

The ability to obtain more accurate measures of inequality will increase substantially if governments would make available linked survey and tax data. Of prime importance is for governments to make the information from (anonymized) tax records available and allow for the linking through personal identification numbers between surveys and

registries.[85] Other administrative registries at the national and cross-national levels that trace incomes and wealth to specific individuals will allow for capturing incomes that are not included in tax records due to their characteristics (for example, undistributed profits) or tax evasion. It is important to remember, however, that linked data will not improve the accuracy of reported incomes for the wealthy if those incomes are not reported to begin with due to tax havens, illegal sources, or other factors (Zucman, 2015; Londoño-Velez and Avila-Mahecha, 2021).

Another less developed strand of the literature but which will probably grow in next years is machine learning methods. Machine learning tools can be used to combine different data sources to improve data availability in terms time frequency, spatial coverage and missing observations. New non-traditional data sources, such as satellite images, digital fingerprints, purchase data, and social media can be combined with existing data-sources (surveys, census and administrative data) to improve income and wealth predictions.[86] However, these methods are not exempt of problems. More importantly for our purposes, thus far these approaches have proven more useful for identifying and measuring poverty than for capturing information on the incomes of the wealthiest individuals.

In the meantime, since there is no perfect method and all methods entail some degree of arbitrariness—assumptions whose validity is often very hard or impossible to test--, a recommendable strategy is to carry out systematic robustness checks, and report ranges rather than single corrected inequality measures.

[85] As indicated above, the statistical offices of New Zealand, Norway, United Kingdom, Uruguay and the United States have taken such a step and shared (a partial version of) this type of information with academics.

[86] Sosa-Escudero et al. (2022) provide an overview of the use of machine learnings methods for poverty, inequality and development studies. Other examples in this direction are Henderson et al. (2012), Blumenstock, 2018, Chi et al. (2022), Chetty, Friedman and Stepner (2024), Abbate et al (2024).

References

Abbate, N., Gasparini, L., Quiroga, F. M., and Ronchetti, F. (2024). Deep Learning with Satellite Images Enables High-Resolution Income Estimation: A Case Study of Buenos Aires. Available at SSRN 5026760.

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

Alfani, G. (2024). Inequality in history: A long-run view. Journal of Economic Surveys. https://doi.org/10.1111/joes.12616

Alfons, A., Templ, M. and Filzmoser, P. (2013). Robust estimation of economic indicators from survey samples based on Pareto tail modelling. Journal of the Royal Statistical Society 62 (C), pp. 271–86.

Altimir, O. (1979). La dimensión de la pobreza en América Latina. Cuadernos de la CEPAL N27, Santiago de Chile.

Altimir, O. (1987). Income Distribution Statistics in Latin America and their Reliability. Review of Income and Wealth, Vo. 33, Issue 2, June, pp. 111-155.

Alvaredo, F. (2007). The rich in Argentina over the twentieth century: From the Conservative Republic to the Peronist experience and beyond 1932-2004.

Alvaredo, F. (2011). A Note on the Relationship Between Top Income Shares and the Gini Coefficient," Economics Letters 110 (3), pp. 274-277.

Alvaredo, F. and Londoño-Velez, J. (2013). High Incomes and Personal Taxation in a Developing Economy: Colombia 1993-2010. CEQ Working Paper 12, Center for Inter-American Policy and Research and Department of Economics, Tulane University and Inter-American Dialogue.

Alvaredo, F., Assouad, L., and Piketty, T. (2019). Measuring lnequality in the Middle East 1990–2016: The World's Most Unequal Region?. Review of Income and Wealth, 65(4), 685-711.

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). Distributional National Accounts in the Context of the WID.World Project, chapter in For Good

Measure: Advancing Research on Well-Being Metrics Beyond GDP, edited by Martine Durand, Jean-Paul Fitoussi, and Joseph E. Stiglitz, OECD report by the High Level Expert Group on Measuring Economic Performance and Social Progress.

Anand, S., and Segal, P. (2015). The global distribution of income. In A. B. Atkinson, and F. Bourguignon (Eds.). Handbook of income distribution (2A, pp. 937–979). Amsterdam: North-Holland.

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year?: explaining measurement error in survey income data. Journal of the Royal Statistical Society Series A: Statistics in Society, 182(4), 1411-1437.

Atkinson, A. B. (1997). Bringing income distribution in from the cold. The Economic Journal, 107(441), 297-321.

Atkinson, A. B. (2007). Measuring Top Incomes: Methodological Issues, in Atkinson, A. B. and T. Piketty (eds.), Top Incomes over the Twentieth Century - A Contrast between Continental European and English-Speaking Countries, Oxford and New York: Oxford University Press.

Atkinson, A. B. (2016). Monitoring Global Poverty, Report of the Commission on Global Poverty, World Bank, Washington, DC: World Bank.

Atkinson, A. B. and Bourguignon, F. (eds.). (2000). Handbook of Income Distribution, Vol. 1, North-Holland, Amsterdam: Elsevier.

Atkinson, A. B. and Micklewright, J. (1983). On the Reliability of Income Data in the Family Expenditure Survey 1970- 1977. Journal of the Royal Statistical Society. Series A (General) 146, no. 1 (1983): 33-61. doi:10.2307/2981487.

Atkinson, A. B. and Piketty, T. (2007). Top Incomes in the Twentieth Century, Oxford: Oxford University Press.

Atkinson, A. B. and Piketty, T. (2010). Top Incomes. A Global Perspective, Oxford: Oxford University Press.

Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top Incomes in the Long Run of History. Journal of Economic Literature 49 (1), pp. 3-71.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in U.S. Wage Inequality: Revising the Revisionists. Review of Economics and Statistics 90(2), pp. 300–323.

Bach, S., Beznoska, M., and Steiner, V. (2016). Who bears the tax burden in Germany? Tax structure slightly progressive. DIW Economic Bulletin, 6(51/52), 601-608.

Bach, S., Corneo, G., and Steiner, V. (2009). From Bottom To Top: The Entire Income Distribution In Germany, 1992-2003. Review of Income and Wealth 55(2), pp. 303-330, 06.

Bartels, C., and Metzing, M. (2019). An integrated approach for a top-corrected income distribution. The Journal of Economic Inequality, 17(2), 125-143.

Biemer, P. and Christ, S. (2008). Weighting Survey Data, Chapter 17, in De Leeuw, E., J. Hox, and D. Dillman International Handbook of Survey Methodology. Great Britain: Psychology Press.

Blanchet T., Chancel L., Flores I. and Morgan M. (2024) Distributional National Accounts Guidelines. World Inequality Lab

Blanchet, T., Flores, I., and Morgan, M. (2022). The Weight of the Rich: improving surveys using tax data. The Journal of Economic Inequality, 20(1), 119-150.

Blanchet, T., Fournier, J., and Piketty, T. (2022). Generalized Pareto curves: theory and applications. Review of Income and Wealth, 68(1), 263-288.

Blumenstock, J. E. (2018). Estimating economic characteristics with phone data. In *AEA papers and proceedings* (Vol. 108, pp. 72-76), American Economic Association.

Bollinger, C. R., and Hirsch, B. T. (2006). Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching. Journal of Labor Economics, 24(3), 483-519.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., and Ziliak, J. P. (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.

Bourguignon, F. (2015). Appraising Income Inequality Databases in Latin America. Journal of Economic Inequality 13 (4), pp. 557-578.

Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. The Journal of Economic Inequality, 16, 171-188.

Bravo, D., and Valderrama Torres, J. A. (2011). The impact of income adjustments in the Casen Survey on the measurement of inequality in Chile. Estudios de Economía, 38(1), 43-65.

Bricker, J., Hansen, P., and Henriques Volz, A. (2019). Augmenting the upper tail of the wealth distribution in the Survey of Consumer Finances. Paper presented at ECINEQ Conference, Paris School of Economics, July 3-5, 2019.

Burdín, G., De Rosa, M., Vigorito, A., and Vilá, J. (2022). Falling inequality and the growing capital income share: Reconciling divergent trends in survey and tax data. world Development, 152, 105783.

Burkhauser, R. V., Feng, S., and Larrimore, J. (2010). Improving Imputations of Top Incomes in the Public-Use Current Population Survey by Using Both Cell-Means and Variances. Economic Letters 108 (1), pp. 69-72.

Burkhauser, R. V., Feng, S., Jenkins, S. P., and Larrimore, J. (2012). Recent trends in top income shares in the USA: reconciling estimates from March CPS and IRS tax return data. Review of Economics and Statistics 94, pp. 371–88.

Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2018). Survey Under-Coverage of Top Incomes and Estimation of Inequality: What is the Role of the UK's SPI Adjustment?. Fiscal Studies, 39(2), 213-240.

Campos-Vazquez, R. M., and Lustig, N. (2019). Labour income inequality in Mexico: Puzzles solved and unsolved. Journal of Economic and Social Measurement, 44(4), 203-219.

Capgemini and Merrill Lynch. (2011). World Wealth Report, New York.

Chancel, L., and Piketty, T. (2019). Indian income inequality, 1922-2015: from british raj to billionaire raj?. Review of Income and Wealth, 65, S33-S62.

Chancel, L., Cogneau, D., Gethin, A., Myczkowski, A., & Robilliard, A. S. (2023). Income inequality in Africa, 1990–2019: Measurement, patterns, determinants. World Development, 163, 106162.

Charpentier, A., and Flachaire, E. (2022). Pareto models for top incomes and wealth. The Journal of Economic Inequality, 20(1), 1-25.

Chetty, R., Friedman, J. N., and Stepner, M. (2024). The economic impacts of COVID-19: Evidence from a new public database built using private sector data. *The Quarterly Journal of Economics*, *139*(2), 829-889.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, *119*(3), e2113658119.

Coles, S. (2001). Threshold models. An Introduction to Statistical Modeling of Extreme Values, Springer Series in Statistics.

Cowell, F. A. (2009). Measuring Inequality, Series London School of Economics Perspectives in Economic Analysis, Oxford, UK: Oxford University Press.

Cowell, F. A. and Flachaire, E. (2007). Income Distribution and Inequality measurement: The Problem of Extreme Values. Journal of Econometrics 141(2), pp. 1044–1072.

Cowell, F. A. and Flachaire, E. (2015). Statistical Methods for Distributional Analysis, Chapter 6 in Atkinson, A. B. and F. Bourguignon (eds.), Handbook of Income Distribution, Vol. 2, North-Holland, Amsterdam: Elsevier.

Cowell, F. A. and Victoria-Feser, M. P. (1996). Robustness properties of inequality measures. Econometrica, 64, 77-101.

Cowell, F. A., and Victoria-Feser, M. P. (2000). Distributional analysis: A robust approach. Putting Economics to Work, Volume in Honour of Michio Morishima. Edited by Anthony Atkinson, Howard Glennerster and Nicholas Stern. London: STICERD.

Cowell, F. A., and Victoria-Feser, M. P. (2007). Robust stochastic dominance: A semi-parametric approach. The Journal of Economic Inequality, 5, 21-37.

Dagum, C. (1997). A new approach to the decomposition of the Gini income inequality ratio. Empirical Economics 22, 515–531.

De Rosa, M., Lustig, N., and Martinez Pabon, V. (forthcoming). Fiscal redistribution when accounting for the rich in Latin America. CEQ Working Paper Series.

De Rosa, M., Flores, I., and Morgan, M. (2024). More unequal or not as rich? Revisiting the Latin American exception. World Development, 184, 106737.

Deaton, A. (2005). Measuring Poverty in a Growing World (or Measuring Growth in a Poor World). The Review of Economics and Statistics 87 (1), pp. 1-19.

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

Deville, J.-C. (2000). "Generalized calibration and application to weighting for non-response". In: COMPSTAT: Proceedings in Computational Statistics 14th Symposium held in Utrecht, The Netherlands, 2000. Ed. by J. G. Bethlehem and P. G. M. van der Heijden. Heidelberg:Physica-Verlag HD, pp. 65–76 (cit. on p. 111).

Diaz-Bazan, T. (2015). Measuring Inequality from Top to Bottom. World Bank Policy Research Paper 7237, World Bank.

Dupuis Lozeron, E. and Victoria-Feser, M. P. (2010). Robust Estimation of Constrained Covariance Matrices for Confirmatory Factor Analysis. Computational Statistics and Data Analysis (54) pp. 3020–3032.

ECLAC (2018). Social Panorama of Latin America 2018. ECLAC. Santiago.

Fesseau, M. and Mantonetti, M. L. (2013). Distributional Measures Across Household Groups in a National Accounts Framework. Working Paper 53, EUROSTAT and OECD.

Fisher, J. D., Johnson, D. S., Smeeding, T. M., and Thompson, J. P. (2022). Inequality in 3-D: Income, Consumption, and Wealth. Review of Income and Wealth, 68(1), 16-42.

Flachaire, E., Lustig, N., and Vigorito, A. (2023). Underreporting of top incomes and inequality: a comparison of correction methods using simulations and linked survey and tax data. Review of Income and Wealth, 69(4), 1033-1059.

Flores, I. (2019). On the Empirical Measurement of Inequality. Ph.D. dissertation, Universite Paris I - Pantheon-Sorbonne, Sciences Economiques, Paris, France.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2000). Statistical distributions. John Wiley & Sons.

Groves, R. M. and Couper, M. P. (1998). Nonresponse in Household Interview Surveys, New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2009). Survey Methodology, New Jersey: John Wiley & Sons.

Harris R.P. (1977). Differential response in the Family Expenditure Survey: the effect of estimates of the redistribution of income, Statistical News 39.

Henderson, J. V., Storeygard, A. and Weil, D. N. (2012), Measuring Economic Growth from Outer Space, American Economic Review 102(2), 994–1028.

Hirsch, B. T., and Schumacher, E. J. (2004). Match bias in wage gap estimates due to earnings imputation. Journal of Labor Economics, 22(3), 689-722.

Hlasny, V. (2021a). Redistributive effects of fiscal policies in Mexico: Corrections for top income measurement problems. Latin American Policy, 12(1), 148-180.

Hlasny, V. (2021b). Parametric representation of the top of income distributions: Options, historical evidence, and model selection. Journal of Economic Surveys, 35(4), 1217-1256.

Hlasny, V. and P. Verme (2018a). Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data. Econometrics 6(2):30.

Hlasny, V., and Verme, P. (2018b). Top incomes and the measurement of inequality in Egypt. The World Bank Economic Review, 32(2), 428-455.

Hlasny, V., and Verme, P. (2021). The Impact of Top Incomes Biases on the Measurement of Inequality in the United States. Oxford Bulletin of Economics and Statistics, 84(4), 749-788.

Hundenborn, J., Woolard, I., and Jellema, J. (2019). The effect of top incomes on inequality in South Africa. International Tax and Public Finance, 26, 1018-1047.

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

Jenkins, S. P. (2017). Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. Economica 84 (334), pp. 261-289.

Jenkins, S. P. (2022). Top-income adjustments and official statistics on income distribution: the case of the UK. The Journal of Economic Inequality, 20(1), 151-168.

Jenkins, S. P., and Rios-Avila, F. (2020). Modelling errors in survey and administrative data on employment earnings: Sensitivity to the fraction assumed to have error-free earnings. Economics Letters, 192, 109253.

Jenkins, S. P., Burkhauser, R. V., Feng, S., and Larrimore, J. (2011). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174 (1). pp. 63-81.

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

Kemsley, W. F. F. (1975). Family Expenditure Survey: a study of differential response based on a comparison of the 1971 sample with the census. Statistical News, November.

Kemsley, W. F. F., Redpath, R. U and Holmes, M. (1980). Family Expenditure Survey Handbook, Social Studies Division, London.

Kim (2014). Income inequality in Korea. Piketty Fever and Income Distribution in Korea: Reality and Prescription. EAF Policy Debates 13, East Asia Foundation.

Korinek, A., Mistiaen, J. A., and Ravallion, M. (2007). An Econometric Method of Correcting for UnitNonresponse Bias in Surveys. Journal of Econometrics,136(1), 213–235.

Korinek, A., Mistiaen, J.A. and Ravallion, M. (2006). Survey nonresponse and the distribution of income, Journal of Economic Inequality, 4, 33-55.

Kuznets, S. (1953). Economic Change, New York: Norton.

Lakner, C. and Milanovic, B. (2016). Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession. The World Bank Economic Review, Volume 30, Issue 2, Pages 203–232.

Lemieux, T. (2006). Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?, American Economic Review 96 (3), pp. 462-498.

Lillard, L., Smith, J. P., and Welch, F. (1986). What do we really know about wages? The importance of nonreporting and census imputation. Journal of Political Economy, 94(3, Part 1), 489-506.

Little, R. J. A. and Rubin, D. B. (2014). Statistical Analysis with Missing Data, Second Edition, Wiley Series in Probability and Statistics, New Jersey: John Wiley and Sons, Inc.

Londoño-Vélez, J., and Ávila-Mahecha, J. (2021). Enforcing wealth taxes in the developing world: Quasi-experimental evidence from Colombia. *American Economic Review: Insights*, *3*(2), 131-148.

Luiten, A., Hox, J., and de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. Journal of Official Statistics, 36(3), 469-487.

Lustig N. (2019). The "Missing Rich" in Household Surveys: Causes and Correction Approaches, Working Paper 75, Commitment to Equity, Tulane University.

Medeiros, M., de Castro Galvão, J., and de Azevedo Nazareno, L. (2018). Correcting the underestimation of top incomes: combining data from income tax reports and the Brazilian 2010 census. Social Indicators Research, 135, 233-244.

Meijer, E., Rohwedder, S., and Wansbeek, T. (2012). Measurement error in earnings data: using a mixture model approach to combine survey and register data. Journal of Business & Economic Statistics, 30(2), 191-201.

Meyer, B. D., and Mittag, N. (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.

Meyer, B. D., Mok, W. K. C., and Sullivan, J. X. (2015). Household Surveys in Crisis. Journal of Economic Perspectives 29 (4), pp. 199-226.

Meyer, B. D., Mok, W. K., and Sullivan, J. X. (2015). Household surveys in crisis. Journal of Economic Perspectives, 29(4), 199-226.

Milanovic, B. (2012). Global inequality recalculated and updated: the effect of new PPP estimates on global inequality and 2005 estimates. The Journal of Economic Inequality, 10, 1-18.

Milanovic, B. (2023). Visions of inequality: from the French Revolution to the end of the Cold War. Harvard University Press.

Mistiaen, J. A., and Ravallion, M. (2003). Survey compliance and the distribution of income (Vol. 2956). World Bank Publications.

Morgan, M. (2018). Essays on Income Distribution. Methodological, Historical and Institutional Perspectives. Ph.D. dissertation, Ecole Doctorale n°465, Ecole des Hautes Études en Sciences Sociales, Paris, France.

Piketty, T (2001). Les Hauts revenus en France au 20e siècle: inégalités et redistribution, 1901-1998, Paris: Ed. Grasset.

Piketty, T. (2003). Income inequality in France, 1901–1998. Journal of Political Economy, 111(5), 1004-1042.

Piketty, T., Saez, E., and Zucman, G. (2022). Twenty years and counting: Thoughts about measuring the upper tail. The Journal of Economic Inequality, 20(1), 255-264.

Piketty, T., Yang, L., and Zucman, G. (2019). Capital Accumulation, Private Property, and Rising Inequality in China, 1978–2015. American Economic Review, 109 (7): 2469-96.

Ravallion, M. (2022). Missing top income recipients. The Journal of Economic Inequality, 20(1), 205-222.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys, New York: Wiley.

Ruiz, N. and Woloszko, N. (2016). What Do Household Surveys Suggest About the Top 1% Incomes and Inequality in OECD Countries?. OECD Economics Department Working Paper 1265, January.

Sanroman, G., and Santos, G. (2021). The joint distribution of income and wealth in Uruguay. *Cuadernos de Economía*, *40*(83), 609-642.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. Econometrica, 613-625.

Silva M. (2023). Parametric Models of Income Distributions Integrating Misreporting and Non-response Mechanisms. AMSE Working Paper 2023/11.

Sosa-Escudero, W., Anauati, M. V., and Brau, W. (2022). Poverty, inequality and development studies with machine learning, en Chan, F. and Mátyás, L. (eds.), Econometrics with Machine Learning, Springer International Publishing. pp. 291-335.

Souza, P. H. G. F., and Medeiros, M. (2015). Top income shares and inequality in Brazil, 1928-2012. Sociologies in Dialogue, 1(1), 119-132.

Szekely, M., and Hilgert, M. (2007). What's behind the inequality we measure? An investigation using Latin American data. Oxford Development Studies, 35(2), 197-217.

Valet, P., Adriaans, J., and Liebig, S. (2019). Comparing survey data and administrative records on gross earnings: nonreporting, misreporting, interviewer presence and earnings inequality. Quality & Quantity, 53, 471-491.

Van der Weide, R., Lakner, C., and Ianchovichina, E. (2018). Is Inequality Underestimated in Egypt? Evidence from House Prices. The Review of Income and Wealth, Vol. 64, Issue s1, pp. S55-S79.

World Bank. (2014). Taking Stock. An update on Vietnam´s recent economic developments. Available at: https://www.worldbank.org/content/dam/Worldbank/document/EAP/Vietnam/Taking_Stock_July2014_EN_final.pdf

Yonzan, N., Milanovic, B., Morelli, S., and Gornick, J. (2022). Drawing a line: Comparing the estimation of top incomes between tax data and household survey data. The Journal of Economic Inequality, 20(1), 67-95.

Zizzamia, R., David, A., and Leibbrandt, M. (2021). Inequality in sub-Saharan Africa: A review paper. AFD Research Papers, (207), 1-33.

Zucman, G. (2015). *The hidden wealth of nations: The scourge of tax havens*. University of Chicago Press.

Zwijnenburg, J. (2019). Unequal distributions: EG DNA versus DINA approach. In AEA Papers and Proceedings (Vol. 109, pp. 296-301). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Appendix 1 Data Preprocessing Practices by Providers

Statistical offices may address unit nonresponse using reweighting or post-stratification methods. For item nonresponse, data producers often complete the information with imputation methods such as "hot deck" (Little and Rubin, 1987). When implemented correctly, these methods generally do not pose significant issues. Transparency is key, allowing survey users to assess the techniques used. However, some practices may be more problematic.

One such approach is complete case analysis, where cases with item nonresponse are simply discarded from the sample. This differs from achieved case analysis, which includes all cases regardless of missing data. Although seemingly straightforward, complete case analysis implicitly assumes that dropped cases are missing at random across all income levels. If item nonresponse is correlated with income, as research suggests, complete case analysis will introduce bias into the estimates of inequality.[88]

Top coding is another preprocessing practice that exacerbates the "missing rich" problem. It involves replacing values above a certain threshold in the achieved sample with that threshold value. In some countries, statistical offices use top coding procedures to protect the identity of high-income respondents.[89] When top coding is applied, the boundary is the income threshold at which reported incomes are replaced by data administrators. This practice introduces bias into inequality measures. Cowell and Flachaire (2015) review the within-survey methods available to address top coding.[90]

Data providers may remove outliers (observations with incomes that are many times higher than the closest observation in the income scale) to reduce volatility in inequality estimates or because they are considered errors (data contamination). If the outliers are

[88] This is so because "… the missing-data mechanism is not MCAR (missing completely at random) and the complete cases are not a random sample of all the cases." (Little and Rubin, 2014, location 1195 in ebook).
[89] For instance, a practice followed by the Current Population Survey in the United States.
[90] Right-censoring in the survey data also occurs when, for instance, questionnaires impose an upper limit to the amount that can be reported..

genuine observations, however, removing them can introduce bias into the inequality measures. A very high income could simply belong to a genuine billionaire.

It should be noted that, while not common practice, some statistical offices have attempted to address underreporting by rescaling survey incomes to align with totals from administrative sources like tax records and National Accounts.

For instance, since 1992, the incomes of very wealthy respondents in the UK Family Resources Survey have been adjusted using information from income tax records to reduce volatility and improve data quality (Burhauser et al., 2018). In this approach, the mean income of very wealthy respondents in the survey is replaced with the mean income of the same group in tax data. Subsequently, the survey weights are recalibrated to match the number of wealthy individuals reported in the tax data. Although documentation on this procedure is limited and replication is challenging, Burkhauser et al. (2018) reconstruct its main features and highlight potential caveats. These include the stratification criteria used to identify which incomes to adjust and the limitations of replacing individual incomes with the group mean, as well as the challenges associated with projection methods due to lags in the availability of tax data.

Another example of preprocessing practices is the Chilean household survey, CASEN. Although this practice has been discontinued, for many years the government agency rescaled incomes to align with National Accounts, and the unscaled survey data was not publicly released.

The United Nations Economic Commission for Latin America and the Caribbean (UNECLAC) used to rescale survey incomes to align with National Accounts. Earnings were rescaled proportionally to match earnings in National Accounts, and the differences in capital income were allocated proportionally but only to incomes from capital for the top 20 percent. One key limitation of this approach is the assumption that underreporting rates were proportional across all income levels. While this practice aimed to address the issue of missing high-income individuals in household surveys and misreporting in general, the lack of public disclosure of the scaling factors prevented any assessment of the method's quality and accuracy. See Bourguignon (2015) for a critical assessment of this practice. UNECLAC discontinued this practice in 2018.

In addition, to mitigate the impact of unit and item nonresponse and underreporting, some statistical offices have adopted a strategy of directly linking survey data with information from administrative registries, rather than solely relying on interview-based data collection. This practice is implemented in France, for example (INSEE, 2016). This approach holds significant promise for improving the representation of high-income households in surveys.

Appendix 2 Within-survey Model Weight Adjustment

Let *H* be the total number of households, *I* the number of nonoverlapping groups in which the population can be partitioned so that households belonging to each group are observationally identical according to their observable characteristics *X,* and J the number of regions as in the previous method. Then, the set of households H_{ij} defines the intersection of region and observable characteristics of weight W_{ij}. Let S_{ij} be the corresponding households in the sample with weight w_{ij}. For each sampled household *h*, the probability of responding the survey is $D_{ijh} = 1$ if the household responds the survey and $D_{ijh} = 0$ otherwise.

This probability can be modeled based on observable variables:

$$P(D_{ijh} = X_i, \theta) = P_i$$

Where θ is an unknown vector of parameters. The authors assume that this probability can assume a parametric form and use a logistic function:

$$P(D_{ijh} = X_i, \theta) = \frac{e^{g(X_i\theta)}}{1 + e^{g(X_i\theta)}}$$

Where g is a stable function of the observable characteristics of interest of the responding households. The set of parameters θ can be estimated on the sample in *j* using the generalized method of moments:

$$\hat{\theta} = \arg min_{\theta} \sum_{j} \left[(\widehat{h_j} - h_j) w_j^{-1} (\widehat{h_j} - h_j) \right]$$

Where $\widehat{h_j}$ is the estimated number of households in the region. As the total number of sampled households in area h_j is observed, the probability of response by area is known but as h_{ij} is unknown, the probability of response for each subgroup m_{ij} is unknown. $\widehat{h_j}$ can be imputed as the sum of the inverted response probabilities in the region. The estimation is carried out at the geographical area level and in the left hand side requires to define the X_{ij} groups that are assumed to be identical within the interval *i* in terms of non response. The set of that are different from zero give an indication of a systematic relationship between X_i and non-response bias.

In the empirical applications, X_i can be income as in Korinek, Mistaen and Ravallion (2006; 2007) and Hslany and Verme (2018a; 2018b; 2021) or other demographic characteristics as in Korinek, Mistaen and Ravallion (2006). However, Korinek, Mistaen and Ravallion (2006) conclude that the better fit is obtained with income. This method is sensitive to the number of groups and most empirical applications conclude that it is better to use medium size geographical areas.

Appendix 3 Reconciling Survey and External Data

a. Income-based Surveys

In the approaches that combine survey data with external sources, the income variable should be reconciled. Reconciling the income variable in household surveys and the external source entails using the same (or most similar) income concept. In addition, when the external source involves data such as tax records (whether tabulations or microdata), the income-sharing unit and the unit of analysis should also be reconciled.

Some surveys collect detailed information for different income sources (capital, labor income, pensions, direct transfers, imputed rent for owner-occupied housing, consumption of own production) and the characteristics of the occupations of each household member (number of occupations, contributions to the social security system and access to social benefits), whereas others collect data at a more aggregated level. In some surveys the reported income is after tax but before transfers, in others it is before both. In some surveys households report income by category (or total) net of taxes and some report gross income: that is, taxable income from the market plus taxable government transfers and before income tax and social security contributions deductions. What is worse, sometimes it is not clear whether the reported income is before or after taxes. In some surveys the reported incomes include imputed rent for owner-occupied housing and consumption of own production while in others it does not. In low- and middle-income countries reconciling incomes gets more complicated because the share of nontaxpaying income is significant due to informality. Additionally, in some cases, not all income sources are reported at the individual level.

Usually, survey data are more flexible in terms of the type of information needed to carry out a comparison with external data, so the main procedures will be to redefine income and the population of interest according to the information available in the external data. Reconciliation exercises therefore involve defining a population of interest to obtain a population control and an income control (Atkinson, 2007; Burkhauser et al., 2012).

The definition of the population control will largely depend on the characteristics of the external data. Reconciliation entails comparing the income-sharing unit: households (with its corresponding definition), individuals, or married couples (in the countries

where there is joint tax filing). Data reconciliation also entails comparing the unit of analysis (for instance, households, individuals, adults, tax units). For example, many empirical exercises use the population aged 15 or 20 and over because the external data covers the adult population only.

To define the income control, it is necessary to assess the characteristics of the external data and then to create the corresponding variable in the survey. For example, to match most tax records, pre- or post-tax income at the individual level for individuals aged 15 and over who contribute to the social security system and/or receive pensions and/or receive taxable capital income sources. It also implies reconciling the time span that corresponds to the income information in both data sources. Since the process depends largely on the specific characteristics of the survey and the external data to be used, it is crucial to have access to the methodological details of the survey and, in particular, to the specific characteristics of the external data. In the case of administrative records, these are often less standardized than data produced by statistical offices and typically require familiarity with tax or social security regulations, as well as common practices in filing tax returns. For example, Burkhauser et al. (2012), Alvaredo and Londoño (2013) and Burdin et al. (2020) describe how this reconciliation process was performed in the UK, Colombia and Uruguay respectively.

b. Consumption-based Surveys

While household surveys in advanced countries, Latin America, and a few other places in the world report the income variable, in most of the rest of the world household surveys report consumption (or expenditures) only. In these cases, researchers use various methods to transform expenditures into incomes (Zizzamia, David and Leibbrandt, 2021; Chancel et al., 2023). For instance, in the case of West Africa, Chancel et al. (2023) convert consumption percentiles into income percentiles by modelling income-consumption profiles using survey data that contain both types of information and then use this information for the countries that only collect consumption data. To construct these profiles, they divide the corresponding average for each percentile and use these ratios as multipliers. As expected, they find an S-shaped relationship, with the ratio of average income to consumption being less than one for the poorer households and

growing exponentially at the top. They also estimate these ratios parametrically using a scaled logit model.

c. National Accounts

The adjustment to National Accounts requires three main definitions: a) the amount to be allocated to the top of the distribution, b) the fractile of the population to be corrected, and c) the share of the population to be added to the top (Bourguignon, 2018). In section 4 we have given several examples of the choices that researchers have made on these three issues.

Although comparisons with National Accounts allow to infer that household surveys underestimate personal income and consumption, per capita GDP is not a suitable measure of household income (Anand and Segal, 2015). National accounts include imputations that are not usually made in household surveys, such as depreciation, retained earnings of corporations and taxes that are not distributed back to households, financial intermediation services, savings done by corporations, the government or foreigners, and consumption of non-profit institutions serving households (Deaton, 2005; Anand and Segal, 2015. Analyzing almost 300 household surveys from 157 countries, Deaton concludes that household income represents around 57 percent of GDP. This figure rises to 70 percent in the case of the United States.

Taking these problems into account, Deaton (2005), Anand and Segal (2015) and Lakner and Milanovic (2016) argue that it is more suitable to compare total household income from surveys with the concept known as income from the household sector in the National Accounts, acknowledging that the latter aggregate will also include financial intermediation services and consumption of non-profit institutions serving households.

As total household income from National Accounts is not always available, these authors argue that for international comparisons it is better to use household final consumption expenditure (HFCE) estimates from national accounts. In this case, the aforementioned Deaton (2005) study estimates that, on average, household survey consumption and income are 86 and 90.4 per cent of HFCE, respectively. However, comparisons of survey data with this aggregate are not free of problems, as its estimation is based on a residual and may involve the use of outdated ratios and correction factors

that may fail to capture intermediate consumption and overstate the levels and growth rates of HFCE (Deaton, 2005). In addition, HFCE includes indirectly imputed financial services, consumption of risk-bearing services by non-profit institutions serving households, and errors and omissions. There are also differences in definitions, such as the imputation of rents for owner-occupiers, the coverage of nonexchange goods (consumption of own production, gifts, etc.), which are not included in the HFCE.

Appendix 4 Estimating Inequality from Corrected Household Surveys: Direction of Change with the Gini Coefficient[91]

Let's examine how the corrected inequality changes for the Gini coefficient using the group decomposition formula for nonoverlapping categories (Dagum, 1997; Atkinson and Bourguignon, 2000; Atkinson, 2007; Alvaredo, 2011):

$$G = G^T \beta S + G^{NT}(1-\beta)(1-S) + S - \beta$$

where G is the Gini coefficient. G can be expressed as the weighted sum of the Gini coefficient for the top G^T and the Gini coefficient for the nontop G^{NT} plus the between inequality component $(S-\beta)$. As in the main text, β is the top population share (e.g., β = 0.01 for the top 1 percent) and S is the income share that corresponds to β.

Let's define the corrected Gini, G^c, as:

$$G^c = G + dG$$

where dG is the total derivative of G:

$$dG = \alpha\, dG^T + \zeta\, dG^{NT} + \gamma\, dS + \delta\, d\beta$$

where:

$$\alpha = [\beta S] > 0$$

$$\zeta = [(1-\beta)(1-S)] > 0$$

$$\gamma = [G^T\beta - G^{NT}(1-\beta) + 1] > 0$$

$$\delta = [G^T S - G^{NT}(1-S) - 1] < 0$$

Is it possible to predict the direction of change? In particular, in which cases will $G^c > G$?

In the case of the replacing method, by assumption $d\beta = 0$. If the bottom part of the distribution is kept the same as in original survey, $dG^{NT} = 0$. In this case, the total derivative becomes:

[91] We thank Ali Enami for sharing this derivation.

$dG = \alpha\, dG^T + \gamma\, d\mathrm{S}$

Thus, any correction method which results in a higher dS, will yield a higher corrected Gini G^c if the Gini for the top increases or remains the same after the correction $dG^T \geq 0$.

If $dG^T < 0$, then $G^c > G$ if $\gamma\, d\mathrm{S} > -\alpha\, dG^T$ (recall that the right-hand side is a positive value because $dG^T < 0$). If this condition does not hold, then $G^c < G$.

Examples of lower corrected Ginis with replacing are not rare. For instance, in the within-survey semiparametric replacing for the EU where Hlasny and Verme (2018a) found that the Gini corrected Gini coefficient, after correction, is 0.2–3.3 percentage points lower than the uncorrected one. In Jenkins (2017) study for the UK, the top observations are removed and replaced by a parametric distribution using tax data. Here too, there are cases in which, after correction, the Gini coefficient is lower than the survey-based Gini.

With reweighting methods, whether dG will be positive or negative is hard to predict ex ante because with reweighting dG^T, dG^{NT}, dS, $d\beta$ can all change at once. Even when the nontop is uniformly downweighted as in some of the reweighting methods described in Table 3 (which means that $dG^{NT} = 0$), there are still three other elements that can change.