FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN
IECON INSTITUTO DE ECONOMÍA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

The "Missing Rich" in Household Surveys: Causes and Correction Approaches. Extended Version with Technical Appendixes

Nora Lustig Andrea Vigorito

INSTITUTO DE ECONOMÍA

Diciembre, 2025

Serie Documentos de Trabajo

DT 03/2025

ISSN: 1510-9305 (en papel)

ISSN: 1688-5090 (en línea)

We thank Ali Enami for sharing a version of the derivation included in Appendix 4. The authors are very grateful to Facundo Alvaredo, François Bourguignon, Mauricio De Rosa, Francisco Ferreira, Emmanuel Flachaire, Stephen Jenkins, Graciela Sanroman, Mathias Silva, Philippe Van Kerm, Paulo Verme and Joan Vilá for providing useful suggestions, insights and clarifications. For their valuable comments and questions, we are also grateful to the participants of the 2nd III-LIS Comparative Economic Inequality conference (Luxembourg, February 27-28, 2025), and of the ECINEQ Conference (Washington DC, July 9-11, 2025). We also thank our research assistants Cliona Poje and Guillermo Sanchez. All errors and omissions rest solely with us.

Suggested citation: Lustig N. and Vigorito A. (2025) “The "Missing Rich" in Household Surveys: Causes and Correction Approaches. Extended Version with Technical Appendixes”, Serie Documentos de Trabajo, DT 03/2025. Instituto de Economía, Facultad de Ciencias Económicas y Administración, Universidad de la República, Uruguay.

The "Missing Rich" in Household Surveys: Causes and Correction Approaches
Extended Version with Technical Appendixes

Nora Lustig Andrea Vigorito[*]

Abstract

Inequality measures based on household surveys may be biased because they fail to capture the upper tail of the income distribution properly. The "missing rich" problem stems from sampling errors, item and unit nonresponse, underreporting of income, and data preprocessing techniques like top coding. This paper reviews salient approaches to address the underrepresentation of the rich in household surveys. Approaches are classified based on information sources and method. In terms of information sources, the distinction is between within-survey data and survey data combined with external sources (e.g., tax records). In terms of methods, we identify three categories: replacing, reweighting, and combined reweighting and replacing. We show that income inequality levels and trends are sensitive to the correction approach. This paper is a companion piece to the chapter of the same name and includes all the appendices that could not be incorporated into the chapter due to space limitations.

Keywords: inequality, missing rich, household surveys, undercoverage, underreporting, replacing, reweighting

JEL Classification: C18, C81, C83, D31

Resumen

Las medidas de desigualdad basadas en encuestas de hogares pueden estar sesgadas porque no logran capturar adecuadamente la cola superior de la distribución del ingreso. El problema de los "ricos faltantes" puede deberse a errores de muestreo, falta de respuesta por ítem y por unidad, subdeclaración de ingresos, y técnicas de preprocesamiento de datos como la codificación superior (*top coding*). En este documento se revisan los enfoques más destacados para abordar la subrepresentación de los ricos en las encuestas de hogares. Los enfoques se clasifican según las fuentes de información utilizadas y el método de ajuste. En términos de fuentes de información, la distinción se establece entre el solo uso de los datos de encuestas y los datos de encuestas combinados con fuentes externas (p. ej., registros fiscales). En términos de métodos, identificamos tres categorías: reemplazo, reponderación, y la combinación de reponderación y reemplazo. Mostramos que los niveles y tendencias de la desigualdad del ingreso son sensibles al enfoque de corrección. Este documento de trabajo complementa al capítulo del mismo nombre e incluye todos los apéndices que no pudieron incorporarse en dicho capítulo debido a limitaciones de espacio.

[*] Nora Lustig is Professor Emerita and founding director of the Commitment to Equity Institute at Tulane University and a Visiting Professor at El Colegio de Mexico. 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 in Uruguay.

1 Introduction

While household surveys have traditionally been used to measure personal income inequality, 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] We outline the factors contributing to the problem, compare the main approaches developed to tackle it, and discuss their strengths, limitations, and practical applications. As we shall see in detail below, the corrected inequality measures are quite sensitive to the approach. This working paper is a companion to the chapter by Lustig and Vigorito (forthcoming).[2] The appendices include a detailed account of how different approaches to incorporating high-income populations into household surveys have been implemented in practice—providing details that could not be accommodated in the main chapter due to space limitations. To the best of our knowledge, the chapter and this companion working paper constitute the first comprehensive analytical survey on the subject.[3]

The main factors that affect the upper tail of the income distribution in household surveys include sampling errors, coverage errors, unit and item nonresponse, underreporting and preprocessing practices by data providers, such as top coding.[4] These factors can lead to biased survey-based inequality indicators which may affect not only their levels but also trends. Beyond measurement, the "missing rich" problem has significant research and policy implications. For example, it could lead to inaccurate estimates of the relationship between inequality and economic growth. The "missing rich" problem in inequality indicators also hinders our ability to accurately assess the progressivity of fiscal systems and evaluate the impact of fiscal reforms.

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

[2] The chapter "The "missing rich" in household surveys: causes and correction approaches" is forthcoming in the Oxford Handbook of Income Distribution and Economic Growth, edited by Gordon Anderson and Sanghamitra Bandyopadhyay, Oxford University Press.

[3] A previous survey by Lustig (2019) focused on the methods that were available until then. This survey was used as a starting point for this chapter.

[4] For a discussion of the factors that lead to the "missing rich" see, for example, 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); and Ravallion(2022).

With renewed interest in economic inequality, there has been correspondingly increased attention to addressing the 'missing rich' problem.[5] This concern has spurred various approaches to generate inequality measures that more accurately capture the upper tail of the income distribution. A first strand of literature attempts to improve the inclusion of the rich in household surveys, either by using within-survey methods (Cowell and Flachaire, 2015) or by combining them with external data such as tax records or National Accounts (Jenkins, 2017; Bourguignon, 2018; and, the long list of references in the online appendix to this paper). A second strand relies primarily on external data sources, such as tax records (Atkinson and Piketty, 2007, 2010). A 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).

We focus here on methods designed to improve the inclusion of the rich in household surveys, as these have been the subject of the most extensive methodological development. Moreover, household surveys, in contrast to tax records or National Accounts, have been designed with the purpose of measuring inequality and poverty in mind. At the time of this writing there were over three hundred published articles and edited volumes. This figure covers studies based on household surveys by themselves or combined with external sources; the literature is considerably larger when including those that rely exclusively on external data and on Distributional National Accounts.

This paper is organized as follows. Section 2 provides evidence on the missing rich in household surveys. Section 3 briefly reviews the causes of the missing rich problem in household surveys. Section 4 presents an analytical description of the approaches developed to address this problem and discusses their application in practice. Section 5 examines the sensitivity of results to the various approaches. Section 6 concludes.

[5] See, for example, Atkinson and Piketty (2007) and Milanovic (2023). 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).

In addition, an online Bibliographical Appendix provides a comprehensive list of papers classified by approach.

2 How do we know the rich are missing in household surveys?

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.[6] 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).

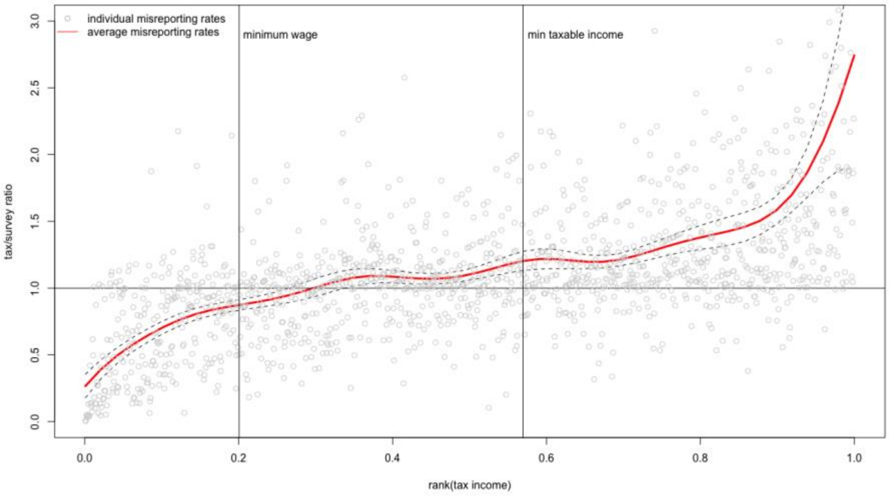
Comparisons with other data sources such as tax records, further highlight the underrepresentation of high incomes in household surveys. Jenkins (2017) showed that the income of the 99.5th percentile of adult individuals in the UK household survey was as low as 77 percent of the equivalent in tax data, depending on the year. 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).

Studies based on household survey data linked at the individual level to external administrative sources provide direct evidence that the upper tail of surveys disproportionately underrepresents income from top earners. In Figure 1 we show the ratio of income from formal employment reported in tax records to income reported in the survey for each observation in linked data for Uruguay in 2012/13. Survey incomes exceed tax return incomes in roughly the bottom half of the tax distribution—likely reflecting payroll tax evasion by small firms—while survey incomes are lower in the top half. On average, the underreporting rate for formal income earners in the top 1 percent of the population is 60 percent.

[6] 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).

Although Figure 1 shows Uruguayan data, similar patterns have been documented for the United States (Bollinger et al. 2019), Sweden (Kapteyn and Ypma 2007), New Zealand (Hyslop and Townsend 2020), the United Kingdom (Jenkins and Rios-Avila 2020), and Italy (Neri and Porreca 2024). Of course, evidence of underreporting using linked data relies on the assumption that administrative incomes are accurate (above a certain threshold) while survey incomes are underreported.[7]

Figure 1 Misreporting Rates, Computed as Ratios of Tax Return to Survey Income: Uruguay Linked Data



Source: Flachaire, Lustig and Vigorito (2023), Figure 2.

There is also evidence that nonparticipation in surveys (unit nonresponse) is not randomly distributed across the population and rises with income (Korinek, Mistiaen, and Ravallion, 2006; 2007; Hundenborn and Verme, 2018a, 2018b, 2021).

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).

[7] Although not the main focus of this paper, Figure 1 also suggests possible “overreporting” in the lower tail, potentially leading to mean reversion (Bollinger et al. 2019). This misreporting pattern is consistent with previous findings in the survey earnings validation literature (Valet, Adriaans and Liebig 2020).

3 The missing rich in household surveys: an overview of the 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[8] 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 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 errors such as income

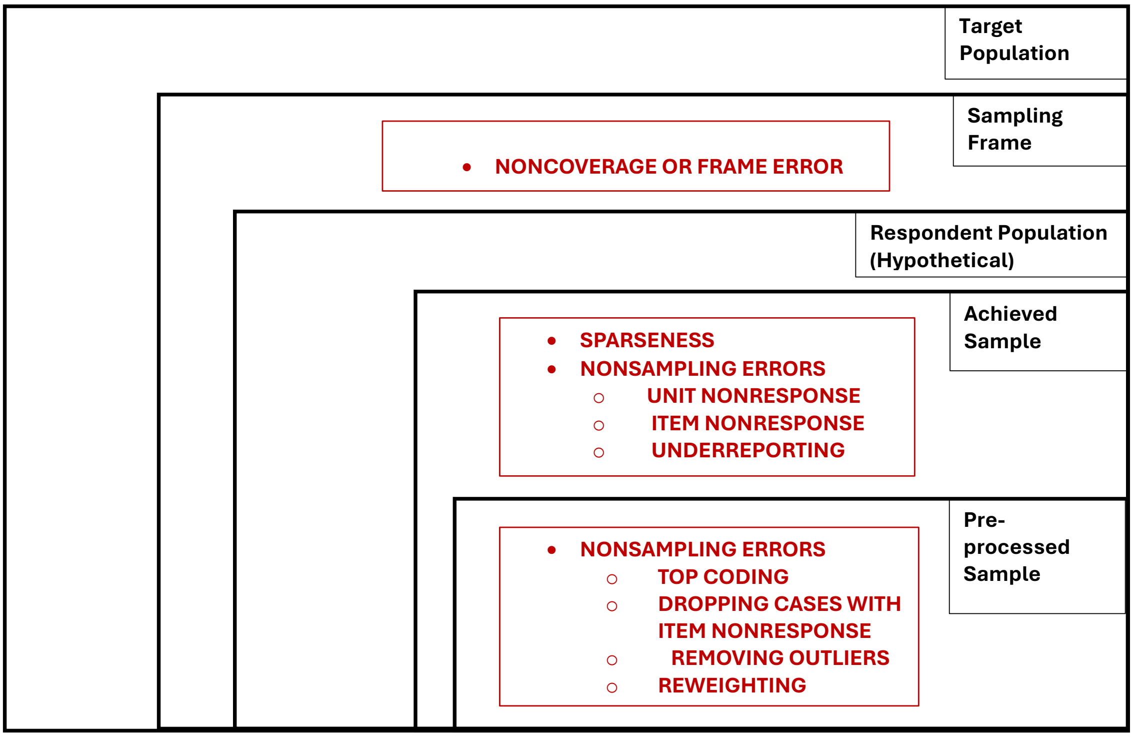
[8] As shown in Figure 2 (adapted from Figure 17.1 in Biemer and Christ, 2008), these four populations "are nested within one another with the target population encompassing the frame population which in turn encompasses the respondent population." (Biemer and Christ, op. cit., p. 318).

underreporting). Finally, the preprocessed sample is the survey after the data providers (e.g., government statistical offices) make adjustments.[9]

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 2 presents the various types of errors in a schematic format. Below, each error is described in detail.

[9] These adjustments can create additional problems in capturing the rich accurately. They include practices such as top-coding and removing outliers. The preprocessed sample may also include addressing 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 2 The Missing Rich in Household Surveys: Causes



Source: authors' elaboration adapted from Biemer and Christ (2008), Figure 17.1. For definitions of the categories see text.

3.1 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. Furthermore, to reduce 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. One way statistical offices can address the issue of sparseness is by oversampling rich individuals when designing the survey.[10] 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. [11]

3.2 Noncoverage error

The *noncoverage or frame error* includes errors of exclusion and errors of inclusion in the frame population.[12] 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 neighbourhoods or in areas under conflict.[13]

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. However, an unintentional noncoverage error on the part of statistical offices may occur if, for example, the

[10] Additionally, wealth and income at the top tail are extremely volatile over short time spans. Thus, even if the upper tail is oversampled to increase the likelihood for the rich to be included, there may be genuine volatile outliers.

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

[12] 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.

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

sampling frame does not adequately include areas where rich individuals live, such as gated communities.[14]

3.3 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 can replace the nonrespondent individual with a similar subject, the nonrespondent subjects end up not being included in the achieved sample.[15] There is evidence that unit nonresponse rates of 20 percent or more are common and that they have been growing over time (see for example, Meyer, Mok and Sullivan, 2015; Hlasny and Verme (2018a; 2018b; 2021; Luiten et al.,2020).

If unit nonresponse is not random, it can lead to the underrepresentation of certain population categories (Atkinson, 2016) as shown by Korinek, Mistiaen and Ravallion (2007) for the US Current Population Survey, and Hlasny and Verme (2018a; 2018b; 2021) for the EU-SILC surveys and the household income and expenditure survey for Egypt.[16] 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. They also report that the probability of response is negatively related to almost all measures of socioeconomic status in rich countries. In such cases, the achieved sample will not be a representative distribution of the target population and the inequality estimates might be biased.

[14] 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).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.

[15] 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.

[16] 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 and Flachaire, 2015). 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.

3.4 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, whereas data for other variables in the survey have been obtained (Groves et al., 2009, p. 354). Such a situation falls under what is usually referred to in the statistical literature on measurement error as item nonresponse. If item nonresponse does not occur at random and is correlated with income, inequality estimates may be biased.[17]

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 only about 75 percent of the top percentile reported their earnings in the survey. Campos-Vazquez and Lustig (2019) found that income nonresponse rises with income in Mexico's employment survey.

3.5 Underreporting

Underreporting refers to subjects who are selected and respond to the survey but report income below its actual level. 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 or 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).[18]

Underreporting, thus, is a case of measurement error. When underreporting is not random and is correlated with income, inequality estimates will be biased. It occurs at the level of the respondent population and would, thus, affect the achieved sample.

One way to unambiguously attribute the upper tail issues to underreporting (and distinguish it from item nonresponse or noncoverage, for instance) is with linked data. Studies for the United States, Norway, New Zealand, the United Kingdom, Uruguay

[17] 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, or if data providers remove observations with item nonresponse.

[18] Other factors, such as the instability of self-employment and business income and the length of the recall period, can contribute to measurement error of the true income.

and Italy 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; Neri and Porreca, 2024).[19] 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).[20]

3.6 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.[21] These are standard data processing techniques, and when used appropriately and transparently, they may not cause insurmountable problems. However, the preprocessed survey may include practices that could exacerbate the "missing rich" problem, such as top-coding or dropping observations when there is nonrandom item nonresponse for the income variable.[22] These practices can lead to further biases and inaccuracies in the data. For more details, see Appendix 1.[23]

As a first step before applying adjustment 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 methods introduced by statistical offices.

4 A taxonomy of approaches

[19] 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.

[20] 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.

[21] 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).

[22] See, for example, Burkhauser et al. (2018) in the context of UK data.

[23] 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).

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. A useful way to classify the available approaches to correct household surveys is according to the source of the data and the choice of method (Hlasny and Verme, 2018a; Lustig, 2019). Depending on the source 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 method, one can distinguish between three main categories: replacing, reweighting, and combining reweighting and replacing. As we shall see, there is a clear distinction among these three categories.

At the outset, it is important to note that modifying the survey to include the missing rich will *not* necessarily result in higher inequality. Depending on the type of error and the method, the resulting inequality measures can be higher or lower than the original ones (Deaton, 2005; Ravallion, 2022). Empirically, however, the vast majority of cases show an increase in inequality estimates, as will be shown in a later section.

4.1 Replacing, reweighting, and replacing and reweighting combined

Let β be the upper tail share of households or individuals in the survey (e.g., the top 5 percent) that suffers from one or more of the issues described in section 2, x the incomes observed in the survey and $w(x)$ their corresponding sample weights in the original survey.

The main distinction between replacing and reweighting is whether: 1) the β remains intact or changes; 2) the observations x at the top remain intact or change; and 3) the weights $w(x)$ remain intact or change. In **replacing**, the β and, hence, $(1-\beta)$ remain intact; the x within the upper tail β change; and the sample weights $w(x)$ and incomes x within $(1-\beta)$ remain intact. The $w(x)$ within the upper tail β remain intact or change depending on the method.[24] In **reweighting**, the entire sample weights $w(x)$ are recalibrated while the incomes x remain intact. After reweighting, the β share increases and thus the $(1-\beta)$ share falls. Under replacing, the changes to the

[24] 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. In this case they do: the weights within β change but not β.

original survey occur within β. Under reweighting, in contrast, the entire distribution is affected. In methods that combine reweighting and replacing (in whichever order), both incomes x within the upper tail β and the weights $w(x)$ throughout the distribution change. Figure 3 summarizes the differences between replacing, reweighting and replacing and reweighting combined.

Figure 3 Approaches to Address Upper Tail Issues in Household Surveys: A Simplified Taxonomy

Method	X = Incomes	W= weights	β = Share of population in upper tail
Replacing	change above the threshold but remain intact below	remain intact except in some methods weights change *within* the top	remains the same
Reweighting	remain intact	change throughout	increases
Replacing & Reweighting	change above the threshold but remain intact below	change throughout	increases

Within-survey

or

Combining survey and external data

Source: authors' elaboration.

A crucial step in replacing and in some reweighting approaches is identifying the threshold (e.g., the sample share β or the corresponding income level at the lower bound of β, —the quantile-- that defines the portion of the distribution in the original sample affected by upper tail issues that need to be addressed. There are a variety of ways to select this boundary. See Lustig and Vigorito (2025) for details.

Replacing can work if the presumption is that the weights —except within the top— in the original survey adequately represent the target population. In other words, replacing assumes that undercoverage, unit or item nonresponse or underreporting occur *within* the upper tail but the proportions β and $(1-\beta)$ are not affected. That is, such upper tail issues, if they exist, are not income-linked in a way that the shares β and $(1-\beta)$ are not representative of the target population.

If there is reason to believe that the rich are underrepresented in the survey, then replacing will not suffice to address this problem. To properly include the rich, reweighting will be necessary. Since it affects the weights while the observations remain intact, however, reweighting by itself can work properly under the assumption that there is common support between the sample and the target population. Common support here is understood in the sense that the specific sample drawn includes at least some of the rich (and that they are not dropped because they are "outliers").[25]

As stated by Ravallion (2022), reweighting is attractive because " It will be obvious to any user of micro survey data for empirical analysis that it is desirable to address selective compliance internally to the survey data. Doing so can retain both the statistical integrity of the survey design and the great many applications for micro-data files in distributional analysis, including policy-related applications." (p. 208). With replacing, this is not necessarily possible in general.

The drawback of reweighting is that if there is no common support, then the reweighted distribution or inequality measure will provide limited correction for the missing rich. Reweighting, moreover, affects the entire distribution (in general). Thus, not only inequality but also (absolute and relative) poverty measures will usually change. [26] If there are reasons to believe that the survey's base weights are incorrect, then poverty estimated with the original survey weights may be biased as well and proper reweighting could take care of biases in inequality and poverty indicators in

[25] As stated by Ravallion: "...the specific sample drawn includes at least some rich respondents (even though very few may be found in the final sample). Thus, the sample distribution has the same support as the true distribution." (Ravallion, 2022, p. 212). Formally, if we define 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. 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. The missing rich problem refers to when the latter occurs in the upper tail.

[26] Relative poverty will vary depending on changes to the mean or median income. In this context, both reweighting and replacing could affect poverty estimates.

tandem.[27] Replacing leaves the untreated portion of the distribution unaffected and thus absolute poverty remains the same, except in the case when rescaling (one of the replacing methods) is applied throughout the distribution (as in Altimir, 1987).

Regarding the use of survey data or combining it with external data such as social security or tax records, within-survey approaches will be limited if there is unambiguous evidence that there is undercoverage and/or underreporting by the rich. The advantages and disadvantages and challenges of using external data are discussed in Lustig and Vigorito (2025).

It is important to note that most methods lack a systematic way of computing confidence intervals, and consequently, most studies do not include them.

Within each one of the broad categories, there are several submethods. The next section summarizes the approaches, the specific methods, the type of data used by them, and their respective underlying assumptions.

4.2 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 twelve distinct approaches in the literature and twenty-two different empirical implementations. The approaches, with their respective assumptions are presented in Table 1.

[27] In some approaches the nontop of the distribution $(1-\beta)$ is mechanically (proportionally) downweighted to ensure that the cumulative distribution function of the reweighted distribution integrates to unity. In those cases, absolute poverty rates after reweighting may be biased. 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.

Table 1 Approaches to Address Upper Tail Issues in Household Surveys: Methods and Assumptions

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						

Source: authors' elaboration.

The approaches that combine surveys with external data require aligning the unit of analysis and the income concepts. 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. 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 (see details in Lustig and Vigorito, 2025). Reconciliation exercises therefore involve defining a population of interest to obtain a population control and an income control (Atkinson, 2007; Burkhauser et al., 2012).

Furthermore, household surveys in many countries collect information on consumption (expenditures) instead of income (or they collect income data poorly). 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 to convert consumption into income.

In Appendix 2 we present the twenty-two empirical implementations included in Table A.1, which also lists illustrative bibliographical references. 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.

The approaches that combine surveys with external data require aligning 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 collect income data poorly). We discuss some of the salient data reconciliation challenges and procedures in Appendix 3.

5 Sensitivity of inequality measures

Inequality measures can be quite sensitive to the approach used to address upper-tail issues. This is particularly true for summary inequality indices such as the Gini coefficient. Table 2 shows the sensitivity of the Gini coefficient to the chosen approach in the few available studies where such a comparison is possible. The highest adjusted Gini is shown in red and the lowest in green. Cells in grey show the cases in which the new Gini is lower than the original Gini. This comparison must be done with caution because there are no confidence intervals associated with the Gini coefficients. Therefore, as is common in this literature, the comparisons presented here are based on point estimates.

Table 2 Sensitivity of the Gini Coefficient to Approach

Author	Country	External Data	Original survey	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	Nonparametric							
					Semiparametric	Semiparametric	Reweighting Top with External Data	Rescaling Top Incomes	Replacing Top with External Data in Full	Model Weight Adjustment	Exogenous Threshold	Endogenous Threshold	Model Weight Adjustment and Semiparametric Replacing	Reweighting (endogenous threshold) and Rescaling	Reweighting (exogenous threshold) and Nonparametric Replacing
Hlasny and Verme (2021)	US 2013	none	0.473	1%	0.491	-	-	-	-	0.504	-	-	0.483	-	-
				5%	0.579	-	-	-	-	0.504	-		0.523	-	-
Bourguignon (2018)	Mexico (2008)	Nat. Acc.	0.510	1%	-	-	0.599	0.600	-	-	0.549	-	-	-	0.587
Flachaire, Lustig and Vigorito (2023)	Uruguay (2013)	Tax	0.382	10%&72%	-	-		0.440	0.440	-	0.442	-	-	0.435	-
De Rosa, Lustig and Martínez-Pabon(forth.)	Brazil (2009)	Tax	0.582	1%	0.581	0.692	-	-	-	-	-	0.646	-	0.691	-
	Chile (2013)	Tax	0.529		0.537	0.576	-	-	-	-	-	0.609	-	0.609	-
	Colombia (2014)	Tax	0.538		0.523	0.627	-	-	-	-	-	0.652	-	0.639	-
	Mexico (2014)	Tax	0.567		0.545	0.681	-	-	-	-	-	0.638	-	0.684	-
	Uruguay (2009)	Tax	0.505		0.519	0.617	-	-	-	-	-	0.561	-	0.575	-

Notes: In Flachaire, Lustig and Vigorito the 72 percent threshold corresponds to the second to last column. The highest adjusted Gini is shown in red and the lowest in green. Cells in grey show the cases in which new Gini is lower than the original Gini. In Hlasny and Verme the Gini coefficient under the reweighting method does not vary with the threshold, since this method does not rely on a threshold.
Source: authors' elaboration.

Hlasny and Verme (2021) applied within-survey replacing, reweighting, and reweighting and replacing combined to generate inequality series for the United States. They found that the order of magnitude of the difference between the new Gini and the original one differs with the method. In no case, the corrected Gini is lower than the original one. The interesting feature of their analysis is that it allows to assess the sensitivity of results to the selected threshold. If one does the comparison by method with changing thresholds (by column), one can observe how sensitive the results are just to the threshold. In addition, there is no systematic pattern linking the size of the threshold and the order of magnitude of the change in the Gini by method.[29] When the threshold is set at 1 percent, the maximum change occurred with reweighting and, when it is set at 5 percent, the maximum change occurs with replacing. Although it is often the case that the higher the threshold, the lower the difference, this is neither mathematically true nor empirically pervasive.

That the results by method can vary significantly is also found by Bourguignon (2018), who applies semiparametric replacing (in two variations), reweighting, and replacing and reweighting 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 change between the original and new Gini is obtained 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 and found that pure reweighting yielded the highest change.

De Rosa, Lustig, and Martínez-Pabon (forthcoming) provide an even more illustrative example of how method- and data-dependent the impact on inequality measures can be. The authors applied both within-survey replacing, and replacing, reweighting, and reweighting and replacing combined using surveys and tax data. As observed in Table 2, in some instances, the new Gini coefficient is lower than the original one (cells in grey).[30] Their exercise also shows that there is no systematic association between the method and the size of the change in the inequality indicator. For instance, the maximum change in the Gini coefficient occurs with semiparametric replacing (Uruguay), with reweighting (Colombia), and sometimes with reweighting and replacing (Mexico). Although no systematic pattern emerges from the studies reported in Table 2, within-survey corrections produce smaller changes than when surveys are combined with external data.

[30] 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 approach, as in the case of Chile.

While not shown in Table 2, top shares also vary by method. However, the ranking of methods appears to be more stable. A comparison is possible for Bourguignon (2018), Flachaire, Lustig, and Vigorito (2023), and De Rosa, Lustig, and Martínez-Pabon (forthcoming). In all three cases, the authors combined surveys with external data, so it is not possible to compare results by data source (within-survey vs. surveys combined with external sources). In most cases, replacing yields the highest results, followed by reweighting and replacing, with reweighting coming in third place.

We have shown that the orders of magnitude of corrected inequality measures are sensitive—particularly in the case of the Gini coefficient—to the combination of method and data selected. Are changes also sensitive to method and data? Unfortunately, no studies are currently available that would allow a full comparison.

The sensitivity of results to the correction method 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, much less the orders of magnitude.[31] Furthermore, as previously discussed, most methods lack a systematic way of computing confidence intervals, and consequently, most studies do not include them. Therefore, as is common in this literature, the comparisons presented here are based on point estimates, which should be interpreted with caution. In fact, to overcome the problem of identifying inequality levels, some authors recommend creating inequality bands based on the measures obtained from different methods and datasets (Alvaredo et al., 2025). In other words, instead of reporting one data point, one should report a range of possible values.

The above discussion and examples are 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. Selecting one of the available approaches 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

[31] In the case of replacing, the direction of change and order of magnitude for the Gini coefficient will depend on whether inequality for the top portion is higher or lower after implementing the method. If it is higher, then the Gini coefficient will be higher than the original one. If it is lower, however, then the direction of change and magnitude 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.

missing rich problem in the survey, the selection of the approach will be affected by the type of data that is available and the research question. For some types of analysis, such as assessing income inequality levels and trends, producing a revised distribution of income with alternative methods and data combinations should be sufficient. However, to study the determinants of inequality, for instance, the microdata in full (incomes and covariates) is essential. It may well be the case that--aside from pure within-survey reweighting-- relying on covariates from the original survey implies far-stretched assumptions, though. Why should one assume that households with much higher incomes would feature covariates as those observed in the survey, for example.

In the absence of statistical tests or calibration methods to assess which approaches produce more accurate corrections, how should one choose among the possible approaches? Some broad guidelines follow. As shown in Table 1, each 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 will be insufficient because it keeps the observations in the sample intact. If common support does not hold, within-survey replacing 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, the approach will need to combine reweighting and replacing, and utilize both survey and external data.

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. A useful way to classify the available approaches to correct household surveys is according to the source of data --within-survey or survey combined with external data-- and the choice of method: replacing, reweighting, and combining reweighting and replacing. Within each one of the broad categories, there are a number of submethods. In this review, we identified twelve approaches and twenty-two distinct empirical implementations involving unique combinations of replacing and reweighting methods and survey data with external data such as tax records or National Accounts.

We showed that inequality estimates can be highly sensitive to the approach. 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. In fact, an important area of future research is precisely developing calibration mechanisms and methods to compute confidence intervals to assess the statistical significance of the inequality measures after the missing rich issue has been addressed.

Looking into the future, a promising solution to the missing rich problem will likely come from linked data. The ability to obtain more accurate measures of inequality will increase substantially if governments make available this kind of information to validate and correct survey data (Benzeval et al., 2020). 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 (2019), researchers could make use of linked data to address item nonresponse and underreporting and other measurement errors by, whenever appropriate, substituting administrative for survey data.

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 rich 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). Further research should explore methodologies for leveraging linked data to mitigate nonresponse and coverage errors.

Another less developed strand of 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.[32] However, these methods are not exempt of problems and 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 implement several approaches, carry out systematic robustness checks, and report ranges (bands) rather than just one inequality measure.

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.

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.

[32] 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).

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., Bourguignon F., Ferreira F. and Lustig N.(2025), Inequality bands: seventy-five years of measuring income inequality in Latin America, *Oxford Open Economics* 4(Issue Supplement_1), pp. i9–i35, https://doi.org/10.1093/ooec/odae018.

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.

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.

Benzeval, M., Bollinger, C. R., Burton, J., Couper, M. P., Crossley, T. F., and Jäckle, A. (2020). Integrated data: Research potential and data quality. Understanding Society Working Paper Series, 2020–02. https://www.ukri.org/wp-content/uploads/2022/03/ESRC-220311-UnderstandingSociety-IntegratedDataResearchPotentialDataQuality-200113.pdf

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.

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

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

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.

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.

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

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

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.

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.

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.

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., 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.

Journal of Economic Inequality (2022). *Finding the Upper Tail. Special Issue*, 21(1), https://link.springer.com/journal/10888/volumes-and-issues/21-1

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.

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.

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.

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.

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.

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.

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

Neri, A. and Porreca, E. (2024). Total Bias in Income Surveys when Nonresponse and Measurement Errors are Correlated. Journal of Survey Statistics and Methodology, Vol 12 (4), pp 1590–1617.

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.

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 Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC (pp. 2007-2001). IRISS Working Paper Series 2007-01, CEPS/INSTEAD.

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.

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.[33]

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.[34] 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.[35]

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 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. Approaches in practice: a summary

Table A.1 presents the twenty-two empirical implementations that were identified in the literature. The references included in this table 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.

Table A.1

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); Van Kerm (2007); Burkhauser et al. (2012); Alfons et al. (2013); Hlasny and Verme (2018a; 2018b; 2021); Burkhauser et al. (2010)	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)		Bach, Corneo, and Steiner (2009), Bach, Beznozka 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 Gegraphic 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.
Source: authors' elaboration.

A.2.1 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 this 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 process. Whether the resulting inequality measure is a more accurate estimate of the true inequality is directly impacted by the distance between the selected threshold and the true threshold (Flachaire, Lustig and Vigorito, 2023).[36] However, as the true threshold is inherently unknown, 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 in the same proportion. 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).

The choice of the threshold can significantly influence the results, regardless the correction approach. Choosing a threshold far from the true value can do more harm

[36] Using data for Uruguay, Flachaire, Lustig, and Vigorito (2023) found that choosing the threshold where underreporting begins (p30) yields a Gini coefficient of 0.45. Choosing the p90 threshold instead yields a Gini of 0.44.

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 modified income distribution increases as the selected threshold diverges from the actual threshold.

A.2.1.1 Semiparametric

The method that relies on removing the upper tail and replacing it with a parametric function 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 using the original observations in the survey.[37] 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. Whether within-survey or combining survey data with external sources, this approach shares a number of characteristics.

The semiparametric replacing approach can be applied to address unit or item nonresponse and underreporting among high-income individuals (Atkinson, 2007; Bourguignon, 2018).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 Cowell and Flachaire (2015).[38]

As Cowell and Flachaire (2015) indicate, if one chooses this path, three important decisions must be made: how the proportion β should be chosen, what parametric model should be used for the upper tail, and how should the model be estimated.[39] These decisions apply to both within-survey and survey combined with external data approaches.

[37] This approach corresponds to Approach A in Jenkins' Figure 1 (Jenkins, 2017, p. 262).
[38] For example, Singh-Maddala, Dagum and Generalized Beta distributions.
[39] Figure A1 in Cowell (2009, p. 159) presents the various options available and the relationship between them.

Jenkins (2017) illustrates the sensitivity of threshold selection, showing that using the same survey, tax data, and a 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.[40] In the case of the US, Hlasny and Verme (2021) find that in 2013, the Gini coefficient for p99 is 0.491 and 0.5792 for p95.

Regarding the parametric model, the most common application has been the (one-parameter) Pareto I model. Research by Cowell and Flachaire (2015), Charpentier and Flachaire (2022), and others demonstrate 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 estimation for data from the United States.[41]

The semiparametric approach can estimate the statistically generated upper tail using observations from the survey (van Kerm, 2007) or external data sources like social security or tax records (Jenkins, 2017). Both approaches assume that the original survey's upper tail and the true upper tail do not share common support.[42] Jenkins (2017) argues that while fitting a parametric upper tail using observations from the

[40] 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. Silva and Lubrano (2024) developed a procedure that departs from a parametric model with a Pareto tail and treats the threshold as an unknown parameter, using a Bayesian strategy for its selection, resulting in a wide range of estimates rather than a single threshold. Their empirical exercise tests the sensitivity of results to the obtained thresholds.

[41] Regarding the estimation of the model, Cowell and Flachaire (2015) recommend using the Optimal b-robust estimator (OBRE) because the maximum likelihood estimation method for the Pareto model is known to be sensitive to data contamination. Jenkins (2017) and Hlasny and Verme (2021) 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.

[42] 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.

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 estimates derived from survey data may fall short. 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, Lustig and Vigorito (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.[43]

The semiparametric methods allow for the estimation of combined 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). A limitation of these methods is the loss of covariate information. To overcome this problem, Van Kerm (2007) imputes extreme incomes by replacing the observed values in the survey with random draws from robustly estimated Pareto distributions (that is, applying a multiple-imputation method).

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 new Gini coefficient using a formula.[44] The formula for non-overlapping groups is as follows (Dagum, 1997; Atkinson and Bourguignon, 2000; Atkinson, 2007; Alvaredo, 2011):

[43] 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.

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

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

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., β = 0.01 for the top 1 percent);[45] 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.[46]

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. The calculation of S using tax data is as presented in section 3.2.

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 A.1, 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. After implementing this procedure, 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.

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. In Table A.1, their method is classified as a semiparametric method that for external data uses income totals.

The authors assume that household surveys provide accurate information for the bottom 90 percent of the income distribution.-Consequently, they assign the full difference between household final consumption in National Accounts and household

[45] 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.

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

surveys to the top decile of the survey's income distribution. Using only household survey information, the global Gini coefficient was 0.705 in 2008. After implementing their method, 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 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 survey weights within the top are inaccurate due to the exclusion of the very wealthiest individuals from the survey. Thus, 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). In Table A.1, Bourguignon's approach is classified as semiparametric replacing with reweighting within the top.

A.2.1.2 Nonparametric

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).

Among the first group, hot deck imputation, mean imputation (unconditional and conditional), regression imputation and stochastic regression imputation have been widely used.[47] However, 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).[48]

In contrast to single imputation methods, 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. This approach is used to address item nonresponse in household finance and wealth surveys (Sanroman and Santos, 2019).

Researchers have often relied on the hot deck and other within-survey imputation methods to address income nonresponse. However, if the respondent population systematically underreports income, the imputation method will result 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.

Rescaling with External Data

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. [49] This method has been applied by, for example, Piketty, Yang and Zucman (2019) and Flachaire, Lustig and Vigorito (2023).[50]

[47] 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).

[48] A general discussion of different imputation methods for the US Current Population Survey can be found in Bollinger and Hirsch (2006).

[49] 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.

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

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.

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. Some authors are skeptical about the use of National Accounts to address upper tail problems in surveys because National Accounts themselves may be subject to significant measurement errors (Deaton, 2005).

Oscar Altimir, 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 (Altimir,1987; ECLAC, 2018).[51] 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.[52]

[51] 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).

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

Bourguignon (2018) proposes a method to rescale survey incomes to align with National Accounts for those cases in which they are the only available external data. He also presents an application to Mexico. His paper shows how to generate the corresponding Lorenz curve and the 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 approach is not equivalent to Distributional National Accounts (DINA) discussed in subsection 3.3. Although DINA employs the rescaling method, its primary objective is to allocate all components of the National Accounts to generate a distribution of income that is consistent with macroeconomic aggregates.

Statistical Matching with External Data

Statistical matching methods have also been used to replace survey data with external information. 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. 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.[53]

The authors 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.

In 1992, the Gini coefficient was 0.5973 in the survey and rose to 0.6155 after implementing the matching procedure. In turn, the shares of the top 10 and 1 percent rise from 35.24 and 6.66 percent in the survey to 39.04 and 11.23 percent. 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.

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, Lustig and Vigorito, 2023).[54] These papers primarily utilize linked,

[53] 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.

[54] 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).

individual-level data from surveys. While they address item nonresponse and measurement error, data limitations preclude them from addressing unit nonresponse and noncoverage.

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.[55] 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.

Using Uruguayan data, Flachaire, Lustig and Vigorito (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.[56] They find that the survey Gini coefficient varies from 0.382 to 0.461 after combining the two data sources, 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. Validation studies challenge the idea that administrative records are better than household self-reported survey data across the board and apply methods that permit identifying measurement errors in both surveys and administrative data (Kapteyn and Ypma, 2007; Meijer, Rohwedder, and Wansbeek, 2012; Bollinger et al., 2019; Jenkins

[55] 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.
[56] While mathematically equivalent to the rescaling method discussed earlier for continuous functions, this equivalence does not hold for discrete functions (Flachaire, Lustig and Vigorito, 2023).

and Rios-Avila, 2020). Based on this, the authors argue that combining information from both sources will make better use of the available data.[57]

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. For example, in their study for Brazil, Medeiros, De Castro Galvao and Nazareno (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. To reweight within the top, 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.

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. 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.

A.2.2 Reweighting

If the household survey suffers from noncoverage error or unit nonresponse—that is, lower participation rates—at the top, the weights for the upper tail and the rest of the

[57] 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).

distribution in the survey might be incorrect.[58] 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.[59] This method is known as reweighting and addresses 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).

A.2.2.1 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) points out 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 reweighting can always find its equivalent with rescaling, one of the nonparametric replacing methods.

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. Additional variants within these two approaches can be found in Table A.1

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

[59] 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.

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 alternatively adjusting nonresponse rates by region or age of the Family Expenditure Survey. The main caveat of this method is that it assumes that respondents and nonrespondents are similar within each geographic/age group.

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, using information from respondents and nonrespondents at the Primary Sampling Unit available from survey producers (e.g., national statistical offices) to assign new weights. This 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.

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.

A.2.2.2 Combining Survey and External Data

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 address 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.[60] Using only survey data, the labor income Gini coefficient declined from 0.424 in 2006 to 0.382 in 2017. After reweighting, 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 Exogenous Threshold

Based on Uruguayan linked data, Flachaire, Lustig and Vigorito (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 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 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 original 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.

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

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

so the details are described in subsection 0. De Rosa, Flores and Morgan (2024) is an example of applying the reweighting component only.

A.2.3 Reweighting and Replacing

A.2.3.1 Within-survey

Model Weight Adjustment Reweighting and Semiparametric Replacing

After modifying sampling weights to address 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 address 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 implementation of reweighting and replacing mitigates implausible modifications that can occur with the replacing method. Additionally, they observe that in most years, the Gini index is larger when both methods are used than with reweighting alone, but smaller that under replacing alone.

A.2.3.2 Combining Survey and External Data

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, 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.[61] After 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

[61] 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.

records. Anand and Segal (2015) apply this method to adjust inequality measures around the globe.[62]

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 combined income density is ensured by the method itself and, under certain assumptions, covariates can be recovered.[63] 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.[64]

The threshold in this approach is the income level in the survey above which 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

[62] They cross-checked their top income estimates against data from rich lists, assessing their country and income source composition against Forbes list. 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. Alvaredo, Assouad and Piketty (2019) also use rich lists combined with national accounts information to estimate top income shares in Middle East countries.

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

[64] 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.

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 is defined as the population share with incomes above 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 new sample 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).[65]

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

[65] 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, Lustig and Vigorito (2023). The rationale being that the parametric function is used to address sparsity at the very top of the top.

suffers from income-linked underreporting and item nonresponse, Flachaire, Lustig and Vigorito (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—coincides with the minimum merging point. This is crucial because the selection of the merging point significantly impacts the magnitude of the new 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 Tables 1 and A.1). Although an advantage of rescaling methods is that, in principle, one can recover the modified 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 combined distribution is identical to the ranking of observations in the original sample.[66]

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 also been applied. For instance, this is the practice followed by the UK's Department of Work and Pensions to adjust 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 these very rich individuals in the survey is replaced by

[66] 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 could change. This could happen if misreporting within cells depends on covariates for example.

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 adjustment. 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 repla.

cing 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.

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, househods, 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 affected, and c) the share of the population to be added to the top (Bourguignon, 2018). In section 4.2 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 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 after addressing the missing rich: direction of change with the Gini coefficient[67]

Let's examine how the Gini coefficient might change after implementing reweighting and replacing 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:

[67] We thank Ali Enami for sharing the derivations included in Appendix 4.

dG = α dG^T + γ dS

Thus, any method which results in a higher dS, will yield a higher Gini G^c if the Gini for the top increases or remains the same (dG^T ≥ 0). If dG^T < 0, then $G^c > G$ if γ dS > - α 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 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 coefficient, after adjustment, is 0.2–3.3 percentage points lower than the original one. In Jenkins (2017) study for the UK, the top observations are removed and replaced by a parametric distribution using tax data.

With reweighting methods, whether dG will be positive or negative is hard to predict ex ante because with reweighting dG^T, dG^{NT}, dS, dβ can all change at once. Even when the nontop is uniformly downweighted as in some of the reweighting methods described in Table 2 (which means that dG^{NT} = 0), there are still three other elements that can change.