The "Missing Rich" in Household Surveys: Causes and Correction Approaches

Nora Lustig

CEQ INSTITUTE
COMMITMENT TO EQUITY
Tulane University

Working Paper 75
November 2019

The CEQ Working Paper Series

The CEQ Institute at Tulane University works to reduce inequality and poverty through rigorous tax and benefit incidence analysis and active engagement with the policy community. The studies published in the CEQ Working Paper series are pre-publication versions of peer-reviewed or scholarly articles, book chapters, and reports produced by the Institute. The papers mainly include empirical studies based on the CEQ methodology and theoretical analysis of the impact of fiscal policy on poverty and inequality. The content of the papers published in this series is entirely the responsibility of the author or authors. Although all the results of empirical studies are reviewed according to the protocol of quality control established by the CEQ Institute, the papers are not subject to a formal arbitration process. The CEQ Working Paper series is possible thanks to the generous support of the Bill & Melinda Gates Foundation. For more information, visit www.commitmentoequity.org.

The CEQ logo is a stylized graphical representation of a Lorenz curve for a fairly unequal distribution of income (the bottom part of the C, below the diagonal) and a concentration curve for a very progressive transfer (the top part of the C).

COMMITMENT
TO EQUITY

CEQ INSTITUTE
COMMITMENT TO EQUITY
Tulane University

THE “MISSING RICH” IN HOUSEHOLD SURVEYS: CAUSES AND CORRECTION APPROACHES

Nora Lustig

CEQ Working Paper 75

NOVEMBER 2019

ABSTRACT

This paper presents a survey of causes and correction approaches to address the “missing rich” problem in household surveys. “Missing rich” here is a catch-all term for the issues that affect the upper tail of the distribution of income: undercoverage, sparseness, unit and item nonresponse, underreporting and top coding. Upper tail issues can result in serious biases and imprecision of survey-based inequality measures. A number of correction approaches have been proposed. A maind distinction is between those that rely on within-survey methods and those that combine survey data with information from external sources such as tax records, National Accounts, rich lists or other external information. Within each category, the methods can correct by replacing top incomes or increasing their weight (reweighting). Correction methods can be nonparametric and parametric. This survey aims to help researchers choose appropriate correction strategies and design robustness tests.

Key words: top incomes, inequality measures, nonresponse, underreporting, replacing and reweighting methods, imputation, poststratification, Pareto distribution, tax records

JEL Classification: C14, C18, C81, C83, D31

THE “MISSING RICH” IN HOUSEHOLD SURVEYS: CAUSES AND CORRECTION APPROACHES

Nora Lustig[1]

1. Introduction

Whether they collect data on income, consumption or wealth, there is reason to believe that household surveys do not capture top incomes well. In this paper, I call this the “missing rich” problem. “Missing rich” here is a catch-all term for the main issues that affect the upper tail of the distribution of income obtained from surveys. Thus, it refers to both the fact that rich individuals may be missing from the sample (due to coverage errors, sparseness or unit nonresponse) or that-- even if they are included-- the information on income is missing (due to item nonresponse), underreported or censored. How do we know that top incomes are not captured well in household surveys? Why is this issue important? What are its causes? What can be done to address the problem? Here I present a synthesis of the factors that give rise to the “missing rich” problem in household surveys, and review the approaches that have been proposed to address it.[2] While there is a vast literature on the topic by economists and statisticians,[3] to the best of my knowledge, there is no comprehensive survey.[4] This is the main contribution of this paper. Its aim is to present and compare the salient correction approaches, discuss their adequacy and limitations, and help researchers choose correction strategies and design robustness tests.

How do we know that the rich are missing in survey data? According to the analysis of Atkinson et al . (2011) and Burkhauser et al . (2012) for the USA, survey-based estimates of the share of total income held by the top 1% are several percentage points less than the estimates from tax return data. Jenkins (2017) shows that the 99.5 centile’s income in the UK household

[1] Nora Lustig is Samuel Z. Stone Professor of Latin American Economics and founding director of the Commitment to Equity Institute at Tulane University (for more information visit www.commitmentoequity.org). She is also a nonresident senior fellow at the Brookings Institution, the Center for Global Development and the Inter-American Dialogue, and non-resident senior research fellow at UNU-WIDER. The author is very grateful for the invaluable comments from Sharon Christ. Very useful comments were also received from Francois Bourguignon, Victor A. Bustos y de la Tijera, Ali Enami, Emmanuel Flachaire, Sean Higgins, Vladimir Hlasny, Christoph Lakner, Marco Mira, Paolo Verme, Andrea Vigorito and Stephen Younger as well as participants of the “Workshop on Harmonization of Household Surveys, Fiscal Data and National Accounts: Comparing Approaches and Establishing Standards,” Paris School of Economics, May 17-18, 2018.

[2] Regardless of its cause, I will call the issue at hand the “missing rich” problem. Other terminology has been used. Jenkins (2017), for example, refers to the problem as “under-coverage” of the rich.

[3] See Figure 2 for a comprehensive analytical summary and a sample of useful references.

[4] A partial survey appears in Lustig (2018).

survey, depending on the year, can be as low as 77% of the equivalent in tax data. With data for 2010, Alvaredo and Londoño-Velez (2013) found that in Colombia the average income of the top 1% is 50% higher with tax data than in surveys. In the case of Brazil, Morgan (2018) finds that the income share of the top 1% in 2015 was 22.5% with fiscal income while only 10.2% with income reported in the survey. By inspection, one can observe that survey top incomes are at most close to the earnings of a well-paid manager. For example, Szekely and Hilgert (1999) 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 of a medium to large size firm, or even less than that. 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 dollars, respectively, a rather low figure in a region with reportedly 4,400 individuals with net worth of 30 million dollars or more.[5]

The fact that top incomes are not well captured in household surveys may explain why there are significant discrepancies in inequality levels and trends, depending on the source of the data. The Gini coefficient for France in 2007, for instance, was equal to 0.39 when measured with incomes reported in the survey but 0.44 when based on tax records (Burricand, 2012). For Colombia in 2010, the analogous figures were 0.544 and 0.587 (Alvaredo and Londoño, 2013). As for diverging trends, Jenkins (2017) showed that, when UK survey-data are combined with tax data, "the Gini coefficient for individual gross income rose by around 7% to 8% between 1996/7 and 2007/8"; in contrast, when only survey data are used, "…the Gini coefficient is estimated to decrease by around 5% over the same period." (Jenkins (2017), p. 285) For Brazil, Morgan (2019) showed a decline of 10 percentage points in the Gini coefficient from 2000 to 2015 when measured with survey income, while with fiscal income the decline shrunk to 3 percentage points. In the case of Colombia, Alvaredo and Londoño-Velez (2013) found that while survey-based estimates showed the share of the top 1% decreasing between 2007 and 2010, tax data showed that it was stable or increasing.[6]

Upper tail issues in may also explain in part the puzzling result that, in many low- and middle-income countries, the survey-based measure of per capita household income (consumption) frequently show levels substantially lower than the per capita household income (consumption) from National Accounts.[7] Analyzing data for Latin America, Bourguignon (2015) found that –between 2000 and 2012-- the ratio of mean income in household surveys to mean household final consumption expenditure per capita in National Accounts could be significantly lower than one.[8] Furthermore, large discrepancies occur not only in levels but also in trends. Deaton states that ".. Taking non-OECD countries as a whole, population-weighted survey

[5] Capgemini and Merrill Lynch (2011).

[6] See also Alvaredo et al. (2015a) and Belfield et al. (2015).

[7] See the pioneer work on this by Altimir (1987). Also, see Fesseau and Mantonetti (2013) and Alvaredo et al. (2018).

[8] Depending on the year, the ratio ranged from 0.78 to 0.84 in Brazil; 0.50 to 0.71 in Colombia; 0.47 to 0.87 in Ecuador; 0.67 to 0.81 in Peru; and, 0.69 to 0.84 in Uruguay. In Mexico, the ratio was the lowest: between 0.42 and 0.49 (!).

consumption in PPP constant dollars grew at only half the rate of population-weighted consumption in the Penn World Tables." (Deaton, 2005, p. 10)

If the rich are missing, the survey-based distributions of income, consumption or wealth, and the concomitant summary inequality indicators should be viewed with caution: actual inequality may be considerably different than survey estimates.[9] The missing rich problem also limits the ability to assess the progressivity of fiscal systems and the impact of reforms.[10] However, it is not necessarily true that correcting the information for upper tail issues will *always* result in higher inequality. If the issue is one of sparsity, first of all, the problem is not bias but precision: inequality measures can experience a high degree of volatility. Second, depending on the type of error and the correction method, corrected inequality measures can be higher or lower than the original uncorrected ones.[11]

The paper is organized as follows. Section 2 discusses the main causes of the "missing rich" problem. Section 3 presents an overview of the correction approaches. Section 4 sums up and concludes.

2. Causes of the "Missing Rich" in Household Surveys

For the purposes of describing the factors that give rise to the missing rich in household surveys, it is useful to define three population groups--the target population (or universe), the frame population, and the respondent population—and the achieved sample survey (household survey).[12] There are, essentially, six main factors embedded in the sampling design, data collection and data preparation process that may give rise to the "missing rich" problem in household surveys. Sampling design issues occur when top incomes are not captured due to frame or noncoverage error or there is sparseness due to sampling error. At the level of data collection, three upper tail issues may occur: unit nonresponse, item (income) nonresponse and underreporting. Top coding (and trimming) occurs at the data preparation level. See Figure 1 for a summary.

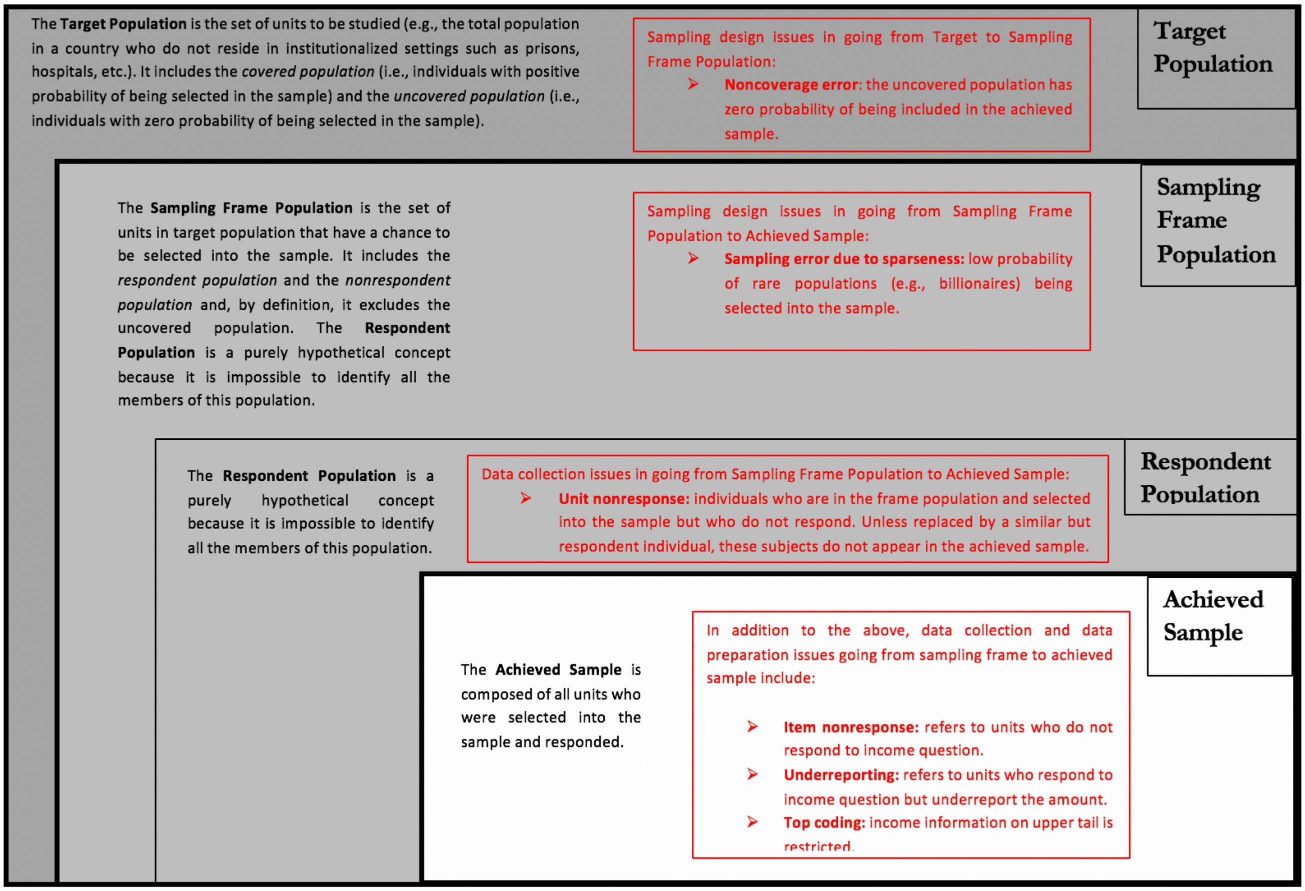
[9] The "Report of the Commission on Global Poverty" (Atkinson, 2016) includes a thorough discussion of these problems at the bottom of the distribution and recommendations on how to deal with them. Here we shall concentrate on the various approaches that have been proposed to address similar problems but at the other end of the distribution, i.e. the high incomes group or the so-called rich.

[10] As discussed in Lustig (2018), the levels and composition of taxes and government spending obtained from fiscal incidence analysis based on household surveys differ significantly from the analogous figures obtained from the governments' budgetary data.

[11] Deaton (2005), for example, shows that correcting for unit nonresponse can result in a decline in measured inequality.

[12] As shown in Figure 1 (adapted from Figure 17.1 in Biemer and Christ, 2008), these three 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). National surveys suffer from a variety of issues related to the representation and measurement of top incomes (Groves and Couper 1998). These range from issues related to sampling (underrepresentation of the very rich) to issues related to data collection (unit nonresponse, item nonresponse, item underreporting and other measurement errors), data preparation (top coding, trimming or censoring, public provision of limited subsamples) or data analysis (trimming of outliers, choices of estimator).

Figure 1: The "Missing Rich" in Household Surveys: Causes



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.). It 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).

Sampling design issues in going from Target to Sampling Frame Population:

- **Noncoverage error**: the uncovered population has zero probability of being included in the achieved sample.

The **Sampling Frame Population** is the set of units in target population that have a chance to be selected into the sample. It includes the *respondent population* and the *nonrespondent population* and, by definition, it excludes the uncovered population. The **Respondent Population** is a purely hypothetical concept because it is impossible to identify all the members of this population.

Sampling design issues in going from Sampling Frame Population to Achieved Sample:

- **Sampling error due to sparseness:** low probability of rare populations (e.g., billionaires) being selected into the sample.

The **Respondent Population** is a purely hypothetical concept because it is impossible to identify all the members of this population.

Data collection issues in going from Sampling Frame Population to Achieved Sample:

- **Unit nonresponse:** individuals who are in the frame population and selected into the sample but who do not respond. Unless replaced by a similar but respondent individual, these subjects do not appear in the achieved sample.

The **Achieved Sample** is composed of all units who were selected into the sample and responded.

In addition to the above, data collection and data preparation issues going from sampling frame to achieved sample include:

- **Item nonresponse:** refers to units who do not respond to income question.
- **Underreporting:** refers to units who respond to income question but underreport the amount.
- **Top coding:** income information on upper tail is restricted.

Note: Adapted by the author from Biemer and Christ (2008), Figure 17.1 and Groves et al. (2009). Definitions in black and type of errors or issues in red.

Noncoverage of Rich Individuals in Household Surveys

The sampling frame error includes errors of exclusion and errors of inclusion in the frame population.[13] In measuring inequality (and poverty), we are primarily concerned with errors of exclusion or 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 to be selected into the sample. These subjects are excluded by design or because they cannot be identified or interviewed: for instance, people living in violent neighborhoods or in areas under conflict, inmates, refugees, and the homeless.[14] If noncoverage error is correlated with income or is more frequent among the richer population, the ensuing inequality measures will be biased.

In general, statistical institutes try not to exclude anybody by design (except for those living in institutions such as prisons and asylums) and try to replace the population who cannot

[13] The frame population can be a mega-sample of the country's population included in the most recent population census or the census population in its entirety.

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

be covered for whatever reason (e.g., people living in violent neighborhoods or in conflict zones) by similar subjects, and over-sample them. To assess the extent to which the frame population in specific countries suffer from noncoverage, the national statistical offices should carry out periodic reviews of the fitness for purpose of the baseline population data (e.g., the Census) for their country.[15]

Sparseness

Even if the achieved sample is flawless –i.e. there are no noncoverage errors and all individuals selected into the sample respond and respond with the truth--very high incomes in surveys tend to be sparse: 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 sub-populations which accrue a disproportionately large part of household income. While sparseness does not cause bias in inequality measures, it produces volatility. Since ultra-high incomes are a low-frequency event, even if there is coverage of the rich and response is positive, they will appear very seldom in a sample. Put differently, the chances to observe Warren Buffett in the US Current Population Survey or Carlos Slim in the Mexican Income Expenditure Survey, even if the ex-ante probability of them being selected into the sample is positive, are almost microscopically small. When high incomes *are* captured, they may appear as outliers even if they are genuine. (Jenkins, 2017, p. 262) In order to avoid volatility in inequality estimates, researchers (and data producers) may drop extreme values on purpose.

Sparseness or low frequency of observations at the top will result in an estimate of inequality and the income share of rich individuals that is not error-free. This error, however, is the typical sampling error which affects any estimate based on a sample and is different in nature from errors caused by the coverage errors and data collection and data preparation issues listed in Figure 1. Sampling errors are expected while these other errors should not happen. Sampling errors create a serious challenge when one wants to estimate with accuracy the upper tail of the income distribution.[16]

One way to address the issue of sparseness is by oversampling rich individuals in the sample frame and the survey sample so that the probability of including someone from the very high-income groups is increased. Oversampling, however, can be costly. An alternative to cope with sparseness and undercoverage has been to replace the upper tail in the achieved sample by a parametric model (e.g., the Pareto distribution), a topic that shall be discussed below.[17]

[15] If the entire population at the top of the income scale beyond a certain threshold is excluded (e.g., people living in gated-communities whose incomes are higher than the highest income of people included in the survey), there is 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 & Flachaire (2015). Cowell and Flachaire discuss methods to address truncation.

[16] See, for example, Flachaire (2018).

[17] See the detailed discussion in Cowell and Flachaire (op. cit.), for example.

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 into the sample because of noncontact (e.g., due to change of address), refusal, or other reasons. As such, and unless the statistical institute is able to replace the nonrespondent individual by a similar subject, the nonrespondent subjects end up not being included in the achieved sample. However, it is possible that none of the theoretical nonrespondent population are selected into the sample which would result 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. Nonresponse can lead to underrepresentation of certain categories (Atkinson, 2016). That is, population groups who are covered but where response rates are lower: for example, slum-dwellers and dwellers of gated communities. In the latter case, the rich will be underrepresented in the survey. Groves and Couper (1998) report that the probability of response is negatively related to almost all measures of socioeconomic status in rich countries and that frequently it is impossible for the survey organizations to penetrate the gated communities in which many rich people live in poor countries.

If one can determine that 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.[18] In other words, one knows that there are individuals above a particular income threshold who will end up being excluded from the survey (achieved sample) and the share of the population these individuals represent. Using Cowell and Flachaire's terminology, we know that, above some threshold, there is an excluded sample; while there are point masses (density) at the boundary that estimate the population share of the excluded part, one does not know the corresponding income.[19]

A potential consequence of unit nonresponse is that one cannot rule out that the population weights supplied by the statistical office for each observation in the achieved sample (i.e., the expansion factors) may be incorrect. In such cases, the achieved sample will not be a representative distribution of the target population. Unit nonresponse bias results if nonresponse is not random but systematically driven by specific factors: e.g., correlated with income or wealth. Given the topic of interest, our concern is if nonresponse is correlated with income. Hlasny and Verme (2015), for example, find that the probability of nonresponse is correlated with income in the US Current Population Survey, the EU-SILC surveys and the household income and expenditure survey for Egypt.[20]

To cope with unit nonresponse in the upper tail and, thus, reduce underrepresentation of rich individuals, national statistical offices can oversample the population groups who are more likely to suffer from unit nonresponse. Statistical offices or researchers can also do expost

[18] Case B in Table 2, Cowell & Flachaire (2015). Cowell and Flachaire discuss methods to address censoring.

[19] See details in Cowell and Flachaire (2015). The case in point is described as case B in their Table 2.

[20] Meyer, Mok and Sullivan (2015) document a rise in unit nonresponse, item nonresponse and measurement error in US surveys.

corrections by changing the weights of the respondent population (known as reweighting or poststratification) or replace the upper tail by a parametric model (e.g., the Pareto distribution). As discussed below, however, within-survey reweighting or replacing can work as long as the achieved sample and the distribution in the target population have the same support: in particular, that the maximum incomes are similar.[21] In the case of reweighting, for example, the maximum incomes in the target population and in the achieved sample must be similar because otherwise reweighting cannot correct for the missing individuals who have incomes beyond those observed in the survey. Replacing the upper tail by a parametric model will have a limited correcting effect because the parameters are estimated on the observations of the achieved sample. If the support is not the same –in particular, if the maximum incomes in achieved sample and target population are not similar--, the use of external sources for the excluded high incomes (e.g., tax records) to correct for the missing rich problem will become essential. This approach will be discussed in more detail below.

Item Nonresponse

Another cause for underrepresentation of rich individuals in household surveys can be that within the respondent population there may be people who do not provide a response for the income (expenditures or wealth) variable. Such a situation falls under what in the statistical literature is usually referred to as *item nonresponse* defined as "…failure to obtain data for a particular variable (or item) in an interview or questionnaire when data for other variables in the survey have been obtained." (Groves et al., 2009, p. 354)[22]

Such as with unit nonresponse, a potential consequence of item nonresponse is that the achieved sample may not be representative of the income distribution of the target population. Item non-response biases results if non-response is not random and is related to specific factors such as income.[23] To cope with item nonresponse of the income variable by individuals at the top of the distribution and, thus, reduce underrepresentation of the rich, national statistical offices can oversample the population groups who are more likely to suffer from item nonresponse. Statistical offices and researchers can also use imputation methods or fit a model for the right-hand tail (e.g., the Pareto distribution). As with unit nonresponse, however, the latter works as long as the achieved sample and the distribution of the target populattion have the same support: i.e., there are some respondents in the right tail that can be upweighted or used to impute values for missing others. If the achieved sample suffers from underreporting of incomes by all the rich or, due to sparseness, the rich individuals are not observed in the sample, reweighting or imputing incomes to the nonrespondent cannot correct for item nonresponse.

[21] Formally, the support between a sample and a true distribution is **not** the same when $f_x(x) = 0$ in the sample whereas $f_x(x) > 0$ 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.

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

[23] Campos-Vazquez and Lustig (2018) find evidence that item nonresponse is correlated with income in the Mexican Labor Survey, for example.

Fitting a model with information from the household survey only, will not necessarily work either (more on this below). Using external sources (e.g., tax records) to correct for the rich who are missing in the achieved sample because they did not respond to the income question will, once more, be of the essence.

A possible strategy to cope with item nonresponse is to drop the cases that suffer from it. In statistics, this is called the complete case analysis (versus the achieved case or sample which does not drop the cases with item nonresponse). Complete case analysis results in unit nonresponse because the entire unit is dropped from the analysis. The problem is that the dropped cases are not really being ignored; they are assumed to be missing randomly across income levels. "… Effectively, the complete case analysis 'imputes' or assigns to each of the missing cases the average or result from all of the complete cases. In other words,… the analyst assumes that the result obtained for the respondents applies to the nonrespondents as well." (Grover et al., 2009, p. 356) If item nonresponse is correlated with income, complete case analysis will lead to bias in the inequality estimates.

Underreporting

Underreporting refers to subjects who are selected and respond to the survey but who—when they respond-- report income (or consumption, or wealth) below its actual level. When the rich are included in surveys, severe underreporting may arise because high-income individuals usually have diversified portfolios with income flows that are difficult to value such as capital income invested in pension funds or retained by corporations as undistributed profits; or because they may also be more reluctant to disclose their incomes. Underreporting is a case of measurement error: even when people respond, they may misrepresent their income, whether on purpose or by mistake.

As mentioned in the Introduction, by inspection or through comparison with other sources (such as tax records), it becomes apparent that people at the top of the income distribution tend to underreport their income, especially income from capital. When underreporting is correlated with income, especially income from capital, this can lead to serious biases in the inequality estimates. Using tax-linked survey data for Uruguay, for example, Higgins, Lustig and Vigorito (2018) show that underreporting does increase with income: that is, the *same* individual reports less income in the survey than to tax authorities and that this underreporting is more frequent and higher in magnitude the higher the income of the individual. Using the framework in Figure 1, underreporting occurs at the level of the respondent population and would, thus, contaminate the achieved sample. Reweighting, imputation methods, or fitting a model with the information in the survey will not address the problem of underreporting by the rich when the incomes of the rich in the target population are above the maximum incomes found in the survey. It will be essential to use external sources (e.g., tax records, National Accounts, rich lists, etc.) to complete the information on rich individuals.

Top Coding[24]

Right-censoring in the survey data also occurs when, for instance, survey administrators top-code reported incomes by design in the data that they made available to researchers,[25] or when questionnaires impose an upper limit to the amount that can be reported. When there is top coding, the boundary is the income threshold at which reported incomes are top coded by data administrators. Cowell and Flachaire (2015) review the within-survey methods available to deal with top-coding. However, here again, using external sources (e.g., tax records) to complete the information on rich individuals (i.e., the incomes that occur above the threshold in which top-coding occurs) might be of the essence.[26]

3. Correction Approaches

Coverage errors, unit or item nonresponse, underreporting and top coding will yield biased inequality measures.[27] Even if there are none of these errors in the achieved sample and, therefore, no bias in inequality estimates, sparseness in the upper tail can result in volatile inequality estimates due to sampling errors. Sampling errors create a serious challenge when one wants to estimate with precision the upper tail of the income distribution.[28] While sampling errors can be reduced through a priori sample stratification to ensure selection of observations from the rare population (e.g., billionaires), the data collection costs of oversampling the rich may be quite high.

Existing research has focused on addressing both sampling errors due to sparseness in the upper tail as well as undercoverage, nonresponse (unit and item), top coding (as well as censoring and trimming) and underreporting.

There are a variety of approaches that have been proposed in statistics and the inequality measurement literature to address upper tail issues.[29] It is useful to distinguish between

[24] A similar issue arises with trimming (Cowell and Flachaire, 2015, Table 2). Top coding is the practice adopted by some statistical agencies to modify intentionally the values of some variables to prevent identification of households or individuals. Trimming is the practice of cutting off some observations from the sample.

[25] To protect confidentiality, for example, data providers may top code the information on income (a practice followed with the Current Population Survey in the United States).

[26] Another issue that may be introduced by statistical offices is that they do not share the entire sample with researchers. As put by Hlasny and Verme (2015): "[s]ome statistical agencies cannot provide the entire data sets to researchers for confidentiality or national-security reasons or simply to prevent others from replicating official statistics. In many countries, statistical agencies provide 20% to 50% of their samples to researchers. These subsamples are usually extracted randomly so that statistics produced from these subsamples may be reasonably accurate. As we know from sampling theory, random extraction is the best option for extracting a subsample in the absence of any information on the underlying population. However, only one subsample is typically extracted from the full sample and given to researchers and this implies that a particularly "unlucky" random extraction can potentially provide skewed estimates of the statistics of interest." (p. 5)

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

[28] See, for example, Flachaire (2018).

[29] Cowell and Flachaire (2015), classify the (right-)tail errors into two main types of "data problems:" *measurement error and data contamination;* and *incomplete data.* Their paper discusses a variety of methods to address them.

approaches that rely entirely on information on incomes (or consumption) contained in the survey in question and approaches that use external information from, for example, tax records and other administrative registries, National Accounts, rich lists, other surveys, house prices, etc. to replace, complement, correct or predict information on, in this case, incomes in the survey. Thus, based on the information source that is utilized to address the upper tail issues, the approaches can be classified into three broad categories. *Within-survey corrections:* researchers correct upper tail issues present in the surveys using parametric or nonparametric methods. *Alternate data:* researchers rely entirely on alternative data such as tax records instead of surveys.[30] *Survey-cum-external data:* researchers correct upper tail issues by combining surveys with external data using parametric or nonparametric methods.

Another key distinction among existing methods that correct is whether the method *replaces* the income observations in the upper tail or *reweights* (poststratifies) the population shares of the top and the nontop, increasing the former and reducing the latter.[31] The first approach assumes that the population shares of top incomes (the rich) and the rest (the nonrich) in the achieved sample survey are correct, and that the problem lies in that (some of) the incomes captured in the upper tail are underreported or missing due to undercoverage, sparseness or unit or item nonresponse. The second correction approach assumes that the population weights for the rich and nonrich in the sample are incorrect due to coverage error or unit nonresponse: therefore, one must "add people" in the upper tail and, consequently, reduce weights at the bottom. Figure 2 summarizes the taxonomy just discussed.

Figure 2: Classification of Correction Approaches

Method	Within-survey	Alternate Data	Combining survey with external data
Replacing: assumes population shares (*base* weights) of rich and nonrich in sample are correct.	Replaces the top x% of the distribution by a parametric distribution (e.g., Pareto) or uses imputation methods to estimate missing data.		Replaces the top x% of the distribution by a parametric distribution (e.g., Pareto) but parameters are estimated using external data (e.g., tax records). Replaces incomes (e.g., means by centile) beyond a

[30] In the past, before surveys became pervasive, researchers often relied on census data. See, for example, Fishlow's analysis of inequality in Brazil (Fishlow, 1973).

[31] This classification was also proposed by Hlasny and Verme (2015 and 2017). However, these authors do not make a reference to the main assumption that underlies their distinction.

		Data from tax records are used instead of surveys alone or in combination with wealth surveys and National Accounts.	certain threshold using values obtained from external information (e.g., tax records or National Accounts).[32]
Reweighting: assumes population shares (*base* weights) of rich and nonrich in sample are NOT correct.	Replaces *base* weights by new weights that are the product of the base weights times the nonresponse adjustment factor times the poststratification weights to address noncoverage, unit and item nonresponse.		Reduces *base* weights of bottom of the distribution to make room for new observations at top. These new observations have income levels that were not in the achieved sample or survey. Information on incomes for these new observations is generated from external sources such as tax data.

Within-survey Correction: Replacing

Whenever it can be assumed that underreporting, nonresponse, truncation and/or censorship occur in the upper tail of the distribution, it is possible to view the distribution of income as composed of two segments: the bottom proportion of sampled individuals for which the achieved sample in the observed survey is a reliable representation of the population and a top proportion that suffers from (one or more of) the upper tail issues described in section 2. In other words, the researcher must choose (or know) at which income level or fractile underreporting or top-coding occurs, and generate the income shares of the population above that income level by fitting a statistical function which presumably approximates actual data better than what is observed in the achieved sample.

A large number of the correction methods appear to assume that the "missing rich" is a problem confined to the upper tail of the distribution (broadly defined). In other words, the methods assume that correcting the problem entails adding density by fitting a particular statistical distribution such as the Pareto distribution or –in the nonparametric correction methods-- by adding income to people above a particular threshold, but that the survey population shares above and below that threshold are correct. These approaches correct for the

[32] May or may not use interpolation methods to join the two distributions. May or may not combine with fitting a parametric distribution.

missing rich problem by *adding income* in the right-hand tail of the achieved sample.[33] This method is also referred to as the *replacing* method: the upper tail from surveys is entirely replaced by a simulated parametric distribution (e.g., a Pareto model)[34] or by a variety of imputation methods.

Because it combines the achieved sample with a fitted parametric distribution, this approach is called semiparametric. The semi parametric approach relies entirely on survey data but observations at the top of the income scale are replaced with the density generated by fitting a statistical (theoretical) distribution. [35] Cowell and Flachaire (2015) discuss the various approaches that fall under this category with a primary focus on sparse coverage of top income ranges. In broad terms, inequality among the population excluding the top group is estimated using survey data while inequality among the top is estimated by fitting a Pareto (or other parametric) distribution using the survey information to estimate the parameters. [36] While initially developed to address sparseness, top coding, censoring and trimming, this approach can also be used for unit or item non-responses if these non-responses are concentrated among top incomes.

Specifically, if one defines 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 distribution (the remaining proportion the $1-\beta$ of lower incomes)." (Cowell and Flachaire, 2015, p. 84) As these authors indicate, there are three important decisions to make if one chooses this path: how should the proportion β be chosen; what parametric model should be used for the tail;[37] and, how should the model be estimated. In the literature, some authors select the β proportion by inspection (heuristic approach) or by an arbitrary assumption. [38] Statistical methods, however, have also been proposed as in Dupuis and Victoria-Feser (2006) and Jenkins (2017). The most commonly used parametric model for the upper tail is the Pareto distribution,[39] but other models have been proposed.[40] This method and its variations have been a long-standing practice to deal with top-coding (censored data), sparse data (e.g. under-representation of rich households), right-truncation, and measurement errors such as underreporting of the incomes of the rich. Their advantages and shortcomings are discussed in detail by Cowell and Flachaire (2015).[41]

[33] One can also think of these corrections as replacing people in the achieved survey's right-hand tail by richer individuals.

[34] This terminology was proposed in Hlasny and Verme (2015 and 2017).

[35] This approach corresponds to Approach A in Jenkins' Figure 1 (Jenkins, 2015, p. 262).

[36] For a discussion of existing parametric models, see Cowell (2009). Also, see survey by Hlasny (forthcoming).

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

[38] An example of the first approach is shown in Figure 2 of Jenkins (2017). An example of the second: in their study for Colombia, Alvaredo and Londoño-Velez (2013) assume that β equals the top 1%.

[39] More precisely, what is called as Pareto I. See Cowell (2009) for description.

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

[41] As surveyed by Cowell and Flaichare (2015), starting with Vilfredo Pareto himself there is a long tradition of using parametric, semi-parametric and non-parametric methods to handle imperfections in data. Cowell and Flachaire state that researchers have adopted a number of work-rounds such as multiplying top-coded values by a

There are also nonparametric approaches that rely on the in-survey available data only. Incomplete data such as item nonresponse in the upper tail can be addressed through single and multiple imputation methods (Little and Rubin, 2014). Little and Rubin classify the single imputation methods into two groups. The explicit modeling methods include *mean imputation (unconditional and conditional), regression imputation* and *stochastic regression imputation.* Pure mean imputation (i.e., replacing the missing income by the mean (or median or mode) of the entire sample population) is roughly equivalent to complete case analysis and, thus, if income nonresponse rises with income, will yield biased results. This bias can be partially reduced if the sample-based means correspond to a relatively homogenous category (e.g., by gender, age, education, etc.). The implicit modeling methods include the *hot deck imputation method* and *composite methods.*[42]

In contrast to the single imputation methods, the *multiple imputation*[43] method refers to replacing each missing value by a vector of two or more imputed values that were generated by 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) discuss in great detail the advantages of multiple imputation and the proper protocols to be followed.[44]

While imputation methods can also be used in conjunction with external sources of information, they were originally designed to deal with the more restrictive case in which the only information that is available is the one contained in the achieved sample (survey). For instance, researchers have often relied on the hot deck imputation method to deal with item (e.g., income) nonresponse.[45] Although this is an advantage of this approach, the limitation is that corrections methods which use the information contained in the surveys such as the hot deck imputation method are not really designed to deal with underreporting or sparse coverage of the top income ranges. If, for example, the respondent population underreports income, the imputation method will propagate the underreporting and the bias in inequality estimates will not be addressed.

Notice that, even though there is no reweighting for the population shares of the rich and nonrich portions of the distribution, the semi-parametric models (fitting a theoretical distribution) and imputation methods described above may implicitly change the weights *within* the pre-selected top share of the population that is subject to being replaced. In other words,

given factor (Lemieux (2006), Autor et al. (2008)) or attempting imputations for missing data (Burkhauser et al. (2010), Jenkins et al. (2011))." As having used this approach, Jenkins (2017) cites Alfons et al. (2013), Burkhauser et al. (2012), Cowell and Flachaire (2007), Ruiz and Woloszko (2016).

[42] In the hot deck method, cases in a survey are sorted by a sociodemographic variable (e.g., gender, education, race). If income is not missing for the first case in the sorted cases, it is stored as the "hot-deck" value. If in the next case sorted list income is missing, "it is replaced by the most recently stored 'hot value.'" (Groves et al., 2009, p. 359) This process is repeated until all missing items are replaced by an income value that corresponds to the most recent reported value (i.e., that of a "neighbor" in the sorted list). The hot deck method thus "uses similarity in sort variables much like predictors in the regression imputation procedure." (Groves et al., 2009, p. 359)

[43] The pioneer of multiple imputation methods is Donald Rubin from Harvard University (Rubin, 1987).

[44] See, for example, chapter 5. Some of the imputation methods have been shown to be less reliable, but this is not the place to discuss their advantages and limitations.

[45] See Campos-Vazquez and Lustig (2017), for example.

the income distribution within the top after fitting a Pareto model, for instance, can (and usually will) be different than the original distribution based on the achieved sample.

Researchers have also tried to address upper tail issues by using several surveys. Fisher et al. (2016), for example, use a combination of surveys to measure the joint distribution of income, consumption and wealth in the United States. By combining surveys that are better at capturing one of the three variables, they are trying to address the missing rich problem as well.

Within-survey Correction: Reweighting

If the achieved sample suffers from unit nonresponse, one cannot rule out that the population shares (weights or expansion factors) of the rich and the nonrich in the achieved sample might be incorrect. This problem has significant implications for the correction method because one needs to go beyond focusing on the right-hand tail and affect the population weights in other segments of the survey: the population weights in the achieved sample must be changed in order to accommodate additional individuals at the top of the distribution. These approaches correct for the missing rich problem by *adding people* in the right-hand tail of the achieved sample. The method is often called *reweighting* and is also known as post-survey weight adjustment or poststratification.[46] Although reweighting is used to correct for unit nonresponse, the method can also be applied to tackle item nonresponse.

As described in Biemer and Christ (2008) and Little and Rubin (2014), reweighting consists in adjusting the expansion factors—also known as base weights--assigned to the complete cases in a sample (that is, the cases with unit or item nonresponse in the available-case sample are discarded) by new weights that take account of, in particular, unit nonresponse (where all the survey items are missing for particular subjects in the sample but not in the frame). Information from respondents and nonrespondents, such as their geographic location, age, gender, and so on available from survey producers (e.g., national statistical offices) can be used to assign new weights. See, for example, Korinek, Mistiaen, and Ravallion (2006) and Hlasny and Verme (2017).[47]

Although the within-survey replacing and reweighting methods are completely different, Bourguignon (2017a) reminds us that, as long as the true distribution and the sample have the same support (that is, there is point-mass at all points in both distributions and their maximum incomes are similar), the results obtained by correcting via reweighting can always find its equivalent using the replacing method. That is, every reweighting exercise, in theory, can be converted into a replacing exercise that will yield the same result, and viceversa. The correction approach will thus be determined by which data is available to the researcher. If information kept by survey producers can be used to correct the survey weights for the presence of unit nonresponse, the reweighting approach should be tried.

[46] See, for instance, Hlasny and Verme (2015 and 2017).

[47] Hlasny and Verme (2017) also apply the replacing method in this paper.

A very important limitation of within-survey correction methods is that, in general, the support of achieved samples is not similar to that of the target population. In particular, the maximum incomes in the achieved sample and target population are not similar. This situation has led some researchers to resort to alternate data sources with more reliable information of incomes at the top such as tax records. This approach is discussed next.

Alternate Data: Tax Records

An approach to capture more accurately the concentration of income and wealth at the top has been to rely on administrative tax data. Inspired by the pioneering work for the United States by Simon Kuznets (1953) and by A. B. Atkinson and Alan Harrison (1978), this approach has been pursued by Piketty (2001) to study the long-run distribution of top incomes in France, by Piketty and Saez (2003) for the United States and in a series of other country studies collected in the two volumes on top incomes edited by Atkinson and Piketty (2007, 2010).[48] To measure inequality, these studies focus on the evolution of income and wealth shares of the population at the top of the distribution, where the "top" can range from the richest 10% to the richest 0.001%.[49]   There are three key methodological challenges when estimating top income shares with tax data: the selection of the total population against which one can define how many tax filers represent a given fractile (such as the top 1%); the selection of the total income used as the denominator in the top income share estimation; and, how to interpolate when the only data available are tabulated by ranges. Atkinson and Piketty (2007, 2010) and Atkinson, Piketty and Saez (2011) describe how to tackle them.

Tax data approximates the upper tail in the target population better for the following reasons. Tax records are less likely to suffer from undercoverage, unit and item nonresponse and underreporting because tax returns are potentially subject to audits and not filing taxes or lying in tax declarations is penalized by the law. Moreover, because it is not a sample, there is no sparseness in the right-hand tail to contend with. However, tax data is no panacea. Although comparisons of top income shares show that tax-based estimates are above survey-based ones, tax records have undercoverage and underreporting problems of their own.

Due to informality, tax avoidance and tax evasion, tax records can also suffer from similar problems to those observed in surveys even if to a lesser degree. In addition, the legal definition of taxable income may leave out (partially or entirely) some very important types of economic resources for the wealthy (e.g. capital gains). While it is true that conventional definitions of income do not include capital gains (or losses, for that matter) because they are changes in wealth, for purposes of calculating income concentration at the very top it may be useful to estimate the shares with and without capital gains. More importantly, it is often the

[48] Also, see Alvaredo et al. (2015a, 2015b) and the surveys by Atkinson, Piketty and Saez (2011) and Alvaredo et al. (2013). Saez (2003)uses tax data to analyze the impact of bracket-creeping in the United States.
[49] An emblematic indicator of this approach is the share of income captured by the top 1%, often reported by the media.

case that a significant portion of income earned by the rich is retained in the corporations as undistributed profits and thus is not captured in personal income tax returns.

To address some of these shortcomings, the DINA (Distributional National Accounts) project led by Thomas Pikettty at the Paris School of Economics and Emmanuel Saez at the University of California, Berkeley, combines tax data with other information sources such as wealth surveys and National Accounts. In particular, Garbinti, Goupille and Piketty (2017) in a study for France and Piketty, Saez and Zucman (2018) in a study for the US combine microfiles from tax returns with information in wealth surveys to impute missing assets and asset income and other income flows that do not appear in income tax returns such as imputed income from owner-occupied housing, life insurance assets or pension funds; and, with National Accounts to impute other missing income flows such as corporate retained earnings. Imputations are carried out so that total income in the corrected microfiles matches total national income and each component matches the corresponding total in National Accounts. The corrected microfiles are generated for pretax (before all taxes and government spending) and posttax (after all taxes and government spending) income. These corrected microfiles are subsequently used to estimate inequality measures for as long a period as data permits.

In principle, inequality measures based on corrected tax data and adjusted to match National Accounts totals should take care of practically all the upper tail issues (an important exception is, for example, incomes kept in tax havens).[50] However, as contended by Deaton (2005), National Accounts may not necessarily be measured with accuracy so some measurement errors could be magnified instead of corrected. In addition, there are significant methodological challenges and a large number of assumptions that must be made in the process of "grossing-up" the information in tax returns to match National Accounts by component and in the aggregate. The significant differences encountered by Piketty, Saez and Zucman (op. cit.) and Auten and Splinter (2019) in their estimates for the US, illustrate how sensitive results can be to particular assumptions.[51]

Combining Survey and External Information

In low and middle-income countries, tax data is likely to cover too narrow a portion of the country's population and, due to weak enforcement mechanisms, declared incomes of the covered population are more likely to be underreported. At the other end of the spectrum, survey data, even if corrected by any of the methods described before will not solve issues of sparseness, undercoverage or underreporting when surveys do not include at least some of the target

[50] Zucman's book on tax havens, for example, reveals the enormous amount of wealth that remains hidden from tax authorities in the world. (Zucman, 2015).

[51] For a summary of the discussion, see, for example, the article by Dylan Matthews "A new study says much of the rise in inequality is an illusion. Should you believe it?"published in Vox on January 10, 2018. https://www.vox.com/policy-and-politics/2018/1/10/16850050/inequality-tax-return-data-saez-piketty

population in the upper tail.[52] When the target population distribution and the sample distribution do not have the same support, reweighting or replacing with survey data will not correct for the missing rich.[53] If support is not the same, reweighting cannot be a solution to underreporting, for example, because there will be incomes whose weights—by definition—cannot be replaced since they don't exist in the sample.[54] Replacing the upper tail by a parametric function will also not yield accurate corrections because if the parameters are estimated with survey data they will fall short of the required correction. To reckon with this problem, researchers have relied on other approaches that use external sources of information to complement, replace, and correct the distribution of income resulting from surveys. Two main external sources have been used: tax records and National Accounts. Authors have also used the so-called rich lists and house prices (or other data) to predict top incomes. The methods are summarized below. As in the case of within survey methods, the methods that combine data sources can also be classified into two main approaches: replacing and reweighting.[55]

Combining Survey and External Information: *Replacing*

One of the commonly used nonparametric method replaces the survey-based mean incomes for percentiles above a certain threshold by tax data cell means.[56] To generate the complete distribution, all incomes within the pre-specified cells are scaled-up by the ratio of the two means (that is, the tax-to the survey-based mean). In general, the replacement by the scaled-up value takes place starting with the fractile in which the tax-based fractile mean is above the survey-based mean for the same fractile; below that threshold, it is assumed that the survey-based means are correct. This approach is similar to what in the statistics literature is called *cold deck imputation*. 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.[57] This method is applied by Bach et al. (2009) to Germany, for example. In the absence of tax

[52] As argued by Jenkins, relying on just in-survey available data to address truncation, censoring, or underreporting, is limited: "… Put differently, fitting a parametric upper tail may obviate the sparsity problem (there is density mass at all points of the distribution's support, by assumption), but the estimate of the 'true' upper tail based on model-based extrapolation from the observed survey observations may not be reliable." (Jenkins, 2017, p. 263) An indication of this issue is, for example, the difference in the magnitude of the inverted Pareto coefficient depending on the source that is utilized to estimate it. In Piketty, Yang and Zucman's analysis for China, for example, the inverted Pareto coefficient estimated with survey data is as low as 1.5 or less while it equals 2.5 or more if estimated with tax data (Piketty, Yang and Zucman, 2019).Recall that the higher the inverted Pareto coefficient, the more unequal the distribution.

[53] As indicated before, formally, the support between a sample and a true distribution is **not** the same when $f_x(x) = 0$ in the sample whereas $f_x(x) > 0$ 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.

[54] Even if the sample had one or two rich cases, in the reweighting method their income values would be used to represent all the other rich people and therefore the variance in the rich income values would be biased downward.

[55] This approach corresponds to Approaches B and C in Jenkins' Figure 1 (Jenkins, 2017, p. 262).

[56] This nonparametric correction for under-reporting has been applied, for instance, by Bach, Corneo, and Steiner (2009), Burkhauser et al. (2016), and the UK Department for Work and Pensions (2015).

[57] Little and Rubin, 2014, location 1682 in ebook.

data, some authors have proposed to use house prices to predict incomes in the upper tail (van der Weide, Lakner and Ianchovichina, 2018).

Another approach corrects survey-based means for wage and capital incomes to match the equivalent in National Accounts. Several decades ago, Altimir (1979) proposed an approach to deal with underreporting in surveys that has been applied by the United Nations Economic Commission for Latin American and the Caribbean (UNECLAC) until 2016: using National Accounts aggregates as control totals for household incomes by source. Roughly, the method consisted in grossing up wage incomes by the ratio of the wage bill in National Accounts to the survey's wage bill. Incomes from capital were similarly grossed up but only for the richest 20% of the population. When compared with the unadjusted estimates, adjusted inequality measures were—by construction-- higher and poverty measures lower. Because the ratios could change by year, trends from adjusted data could also differ from trends with unadjusted data. The limitations of this method are discussed at length by Bourguignon (2015). UNECLAC has now moved away from this method and estimates inequality and poverty indicators directly from survey data.

As mentioned above, some authors are skeptical about using National Accounts to correct surveys because National Accounts may be measured with more significant errors. (Deaton, op. cit) Interestingly, however, there is a "revival" of this approach as exemplified by the already mentioned DINA project, the OECD/Eurostat Expert Group in Integrating Disparities in National Accounts,[58] and the US Census Bureau and others in the United States.[59]

For countries in which tax records cannot be relied upon as the starting point to measure inequality because of their lack of coverage and overall unreliability, the studies produced under the DINA project combine survey data with tax data and National Accounts to generate income and wealth distributions that are corrected for upper tail issues and consistent with totals in National Accounts.[60] The exact method to construct DINA depends on the country and period because data availability varies but the general methodology is described in Alvaredo et al. (2016). In general terms, the method relies on survey data for the bottom $(1 - \beta')100\%$ (for example, the bottom 90 percent) of the population and use tax data for the top $\beta 100\%$ (for example, the top 1 or .5 percent). The fractile β' is the threshold above which the researcher considers survey data is not reliable and β is the fractile above which the researcher considers tax data adequately represents the upper-most tail; if β' and β are not the same threshold, interpolation methods

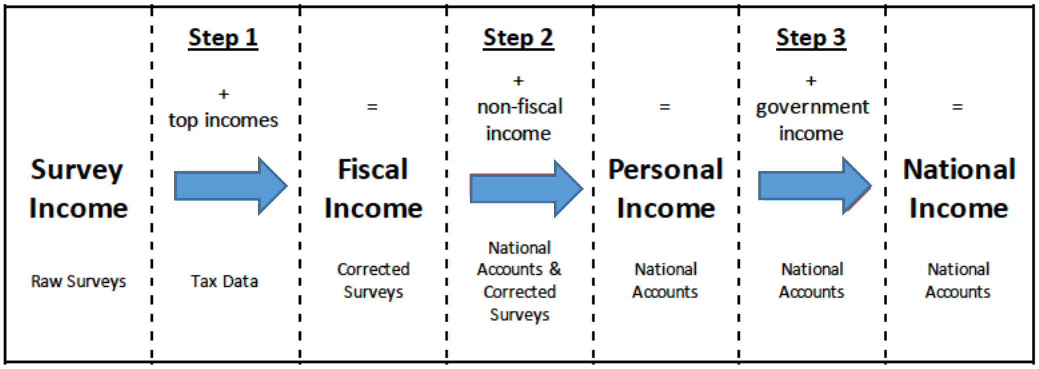
[58] Zwijnenburg, Bournot and Giovanelli (2017).

[59] In contrast to the other methods discussed here, however, the method of adjusting to National Accounts objective is not really (or, not just) addressing issues at the upper tail of the distribution. The main goal is rather to provide indicators of households' economic well-being across countries that go beyond the standard per capita GDP. One such indicator is the level of adjusted households' disposable income by decile or quintile. Consistent measures of GDP growth and income distribution could be used, for instance, to assess whether GDP per capita growth is associated with higher or lower inequality.

[60] See, for example, Alvaredo, Assouad and Piketty (2018); Chancel and Piketty (2017); Novokment, Piketty, and Zucman (2017); Piketty, Yang and Zucman (2019). Morgan (2018) and Flores (2019) also apply this approach but use replacing and reweighting methods.

are used.[61] Combining survey data with tax records to correct for undercoverage and other upper tail issues, authors construct what they call "fiscal income." In the second step, these exercises incorporate all private so-called non-fiscal incomes that can be attributed to the household sector to obtain "personal income." The added income comes from the following items: social insurance contributions (both from employers and employees); imputed rent for owner-occupied housing; investment income attributable to insurance policyholders; investment income payable to pension entitlements; and, pre-tax undistributed corporate profits (retained earnings). The final step involves imputing the remaining categories of income to arrive at a national income distributional series.[62] Figure 3 presents a schematic description of this method.[63]

Figure 3: Combining Surveys, Tax Data and National Accounts: Summary of Steps



Source: Figure 2.3 in Morgan (2018), p. 50.

If one is not interested in producing a corrected version of the entire distribution of income, a simpler alternative to the above consists of estimating inequality measures by directly combining the measures estimated with each information source separately. For this purpose, the method relies on decomposable inequality measures. In the case of the Gini coefficient, the

[61] In the countries for which β does not equal β', these papers assume that the quantile ratio upgrade factor rises linearly in between. They then apply the semi-parametric approach based on the generalized-Pareto-interpolation techniques to complete the distribution (Blanchet, Fournier, and Piketty, 2017). In the case of China, for example, Piketty, Yang, and Zucman (2019) correct the survey assuming that "… the survey data is reliable below percentile p_1=0.9, the fiscal data is reliable above p_1=0.995, and … assume that the quantile ratio upgrade factor f(p) rises linearly from f(p_1)=1 to the observed fiscal/survey ratio f(p_2) between p_1 and p_2." The authors then apply the generalized-Pareto-interpolation techniques to the corrected tabulations to obtain the percentiles for the distribution of income over the period of interest. The authors assess the robustness of their benchmark results through applying different piecewise linear profiles for the rescaling (upgrade) factor between p_1 and p_2.

[62] The remaining categories include government factor (capital) income, net production taxes (e.g., Value Added Tax) received by the government, pension and other social insurance surplus.

[63] Some authors call these methods "consistent income inequality" exercises because incomes are adjusted to be consistent with the same components in National Accounts, tax records and other administrative registries. (Auten and Splinter, 2019)

formula for non-overlapping groups can be written as follows (Dagum, 1997, Atkinson, 2007 and Alvaredo, 2011):

$$G = \frac{b-1}{b+1}\beta S + G^{*}(1-\beta)(1-S) + S - \beta \qquad \textbf{(1)}$$

where β is the top group considered (e.g., β = 0.01 for top 1 percent); S is the tax-based top x percent income share (e.g. the top 1 percent's income share); b is the tax-based inverted-Pareto coefficient;[64] G^{*} 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[65] The Gini coefficient obtained with an estimated parametric function (e.g., Pareto) can be compared to the uncorrected non-parametric estimate for the observed income distribution. A higher semi-parametric Gini would indicate that the observed top incomes are lower than what the modelled (e.g., Pareto) distribution would predict. This could be interpreted as evidence that there is underrepresentation or underreporting of high income units in the achieved sample.

As it happens when applying them to in-survey data only, semi-parametric models face exactly the same set of challenges when data sources are combined: how should the threshold (i.e. the β) be chosen; what parametric model should be used for the tail; and, how should the model be estimated. Using data for the United Kingdom, Jenkins illustrates the sensitivity of results to the choice of the parametric model, and finds that the selection of the threshold and the parametric model for the tail affect—above all—inequality levels but not so the trends, which are quite similar across the board (Jenkins, 2017, Figure 9, p. 282). His paper can be viewed as best practice in terms of robustness checks for this approach.

Combining Survey and External Information: Reweighting

As discussed above, in the presence of income-correlated unit nonresponse, base weights for the upper tail (and, consequently, also for the rest of the achieved sample) may be incorrect. However, in contrast to the within-survey reweighting method, researchers and statistical offices replace the original expansion factors or base weights by new weights derived from population control totals by age, sex, region, etc., obtained from external administrative registries such as tax and social security records (Burkhauser et al. (2017); Campos-Vazquez and Lustig, 2018; Department of Work and Pensions (2015)). Another approach has been suggested by Bourguignon (2017b). Roughly, it consists in redefining weights in such a way that the distribution in the upper tail resembles the distribution in tax data and the distribution below the upper tail resembles the distribution embedded in the survey. The method proposed by Blanchet,

[64] To avoid confusion, the reader is reminded that Alvaredo (2011) and Jenkins (2017) use the symbol β for the inverted-Pareto coefficient. I decided to use **b** instead to keep the symbol β for the threshold that separates the "rich" from the "nonrich" because this is the symbol used in Cowell and Flachaire (2015).

[65] This semi-parametric approach has been used by Atkinson et al. (2011), Alvaredo (2011), Alvaredo and Londoño-Velez (2013), Diaz-Bazan (2015), Anand and Segal (2015), Jenkins (2017), and Lakner and Milanovic (2016).

Flores and Morgan (2018) and applied by Flores (2019) and Morgan (2019) is an attempt to move in that direction.

A simpler reweighting method was proposed by Atkinson (2007) in the context of the Gini coefficient. It can be shown that the Gini coefficient for the whole population (including the rich) can also be approximated by the decomposition formula:

$$G = G^{**}\beta S + G^{*}(1-\beta)(1-S) + S - \beta \qquad (1)'$$

where $G^{*}$is the Gini coefficient for the entire achieved sample but it is assumed to represent the bottom $(1-\beta)\%$ and G^{**} is the Gini coefficient of the top $\beta\%$ and is calculated from tax data. In essence, the achieved sample is compressed in its totality to make room for the additional $\beta\%$ that represents the share of rich individuals that need to be "added" to complete the distribution. This method could be interpreted as an extreme form of poststratification because the whole achieved sample is assumed not to represent the target population but a subset of the latter. Anand and Segal (2015) apply this method to adjust inequality measures around the globe. In contrast with the methods discussed in the first paragraph of this section, this method generates a corrected inequality measure but not a corrected version of the microdata.

Corrected Inequality Measures and Direction of Change

Will corrected inequality measures be always higher than uncorrected ones? The answer is no. As indicated by Deaton (op. cit.), when correcting for unit nonresponse, the resulting inequality measure can be lower than the uncorrected one: "…with greater nonresponse by the rich, there can be no general supposition that estimated inequality will be biased either up or down by the selective undersampling of richer households. (The intuition that selective removal of the rich should reduce measured inequality, which is sometimes stated as obvious in the literature, is false, perhaps because it takes no account of reduction in the mean from the selection.)" (Deaton, 2005, p. 11). A simple example can illustrate this point. Let's assume that we observe a population of 4 with the first three having $0 income and the fourth $1 (0,0,0,1). The coefficient of variation for this distribution is 2 and the share of income of the richest person is 100 percent. Let's assume that the true distribution is (0,0,0,1,1); the coefficient of variation is 1.37 and the income share of the richest person is 50 percent.[66]

The ambiguity in the direction of change occurs beyond the case of unit nonresponse mentioned by Deaton. Higgins, Lustig and Vigorito (op. cit.), Hlasny and Verme (2018) and Jenkins (op. cit.), for example, replace top observations by a parametric distribution and find that in some cases the corrected Gini is lower than the uncorrected one. In most cases, however, the corrected Gini is higher the original one.

[66] Also see Alvaredo, Atkinson and Morelli (2017), in the case of wealth distribution in the UK.

Let's illustrate how the corrected inequality can be higher or lower than the original inequality with the Gini coefficient using the decomposition formula (1)' above. Let's define the corrected Gini, G^C, as:

$$G^C = G + dG$$

Under which circumstances would G^C be higher or lower than G? To answer this question, let's first take the total derivative of (1)':

$$dG = \alpha\, dG^{**} + \beta\, dG^{*} + \gamma\, dS + \delta\, dP$$

where:

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

$$\beta = [(1\text{-}S)\ (1\text{-}P)] > 0$$

$$\gamma = [G^{**}P - G^{*}\ (1\text{-}P) + 1] > 0$$

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

In replacing methods, $dP = 0$. If bottom distribution is kept the same as in original survey, $dG^{*} = 0$, then the total derivative can be written as:

$$dG = \alpha\, dG^{**} + \gamma\, dS$$

As long as $dG^{**} \geq 0$, any correction method which results in a positive dS (i.e., an increase in the share of income going to the top), will always yield $dG>0$. That is, the corrected Gini G^C will always be higher than the original uncorrected Gini G. However, if inequality within the top declines --if $dG^{**} <0$--, the corrected Gini will be higher than the original $G^C> G$ only if $\gamma\, dS > - \alpha\, dG^{**}$. Otherwise, the corrected Gini will be equal or lower than the original one.

In reweighting methods, whether dG will be positive or negative is hard to predict ex ante because with reweighting dG^{**}, dG^{*}, dS, dP can all change at once. Hence, G^C can be higher or lower than G.

It is worth noting that while in principle (and in practice) corrected inequality can be lower than the original one, empirical studies find that inequality after correcting is more frequently higher. See, for example, Flores (2019), Higgins, Lustig and Vigorito (op. cit.), Hlasny and Verme (2017 and 2018) and Morgan (2018).

4. Summing Up

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: undercoverage, sparseness, unit and item nonresponse, underreporting and 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 unit and item nonresponse and that this problem might have been on the rise. 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 issues that affect the upper tail.

A number of correction approaches have been proposed in the literature. The first important distinction is between those that rely on within-survey methods and those that combine survey data with information from external sources such as tax records, National Accounts, rich lists or other external information. Within each category, the methods can correct by replacing top incomes or increasing their weight (reweighting). Correction methods can be nonparametric and parametric. The previous section discussed the approaches in some detail. Table 1 presents a summary and corresponding references. As in the previous section, the table also makes reference to the approaches that do not rely on household surveys but estimate inequality from administrative registries such as tax records.

Table 1: The "Missing Rich" and Correction Approaches

Approach	Income Survey Data	External (out of survey) Data	References
WITHIN SURVEY CORRECTION METHODS			
REPLACING TOP INCOMES: POPULATION SHARES (WEIGHTS) OF TOP INCOMES (β100%) AND NONTOP INCOMES [(1 - β)100%] UNCHANGED			
Parametric			
Replace upper tail by a Pareto distribution (or other models) estimated from survey and use survey data for incomes below the income threshold that does not suffer from upper tail issues.	Yes	No	Methodology: Cowell and Victoria–Feser (1996); Cowell and Flachaire (2015) Application: Alfons, Temple, & Filzmoser (2013); Burkhauser et al. (2012); Cowell and Flachaire (2007); Higgins, Lustig and Vigorito

			(2018); Hlasny and Verme (2015, 2017, 2018); Ruiz and Woloszko (2016)
Nonparametric Imputation			
Incomplete data such as item nonresponse in the upper tail can be addressed through single and multiple imputation methods.	Yes	No	Methodology: Little and Rubin (2014) Application: Autor et al. (2008); Burkhauser, Feng, and Larrimore (2010); Campos-Vazquez and Lustig (2017); Jenkins et al. (2011); Lemieux (2006)
REWEIGHTING: POPULATION SHARES (WEIGHTS) OF TOP INCOMES (β100%) AND NONTOP INCOMES ((1 - β)100%) CHANGE			
Poststratification: replace the expansion factors in sample (base weights) by new weights generated with information on nonrespondents obtained, for example, from survey producers; it requires information on characteristics (age, gender, education, etc.) on the respondent population.	Yes	Yes (for example, information on nonrespondents from data producers)	Methodology: Atkinson and Micklewright (1983); Biemer and Christ (2008); Korinek, Mistiaen, and Ravallion (2006, 2007); Mistiaen and Ravallion (2003) Application: Hlasny and Verme (2017, 2018); Morelli and Muñoz (2019)
TAX DATA ONLY			
Tax data from individual records or tabulations are used to calculate the income shares of top incomes (e.g., the emblematic 1%).	No	Yes: Tax Data (individual records and tabulations)	Atkinson and Harrison (1978); Atkinson and Piketty (2007, 2010); Kuznetz (1953); Piketty (2001); Piketty and Saez (2003); Saez and Zucman (2016)
COMBINING TAX DATA WITH NATIONAL ACCOUNTS			
WID.World Distributional National Accounts (DINA) are constructed for the adult population (20 yrs or older) starting from tax returns micro-files; household wealth surveys are used to impute missing assets and asset-derived and other income flows; through a series of imputations, national accounts are used to impute other missing income and taxes and transfers (in cash and in-kind) so that labor income, capital income, taxes and transfers in the micro-files are equalized to corresponding totals in the National Accounts. Other "consistent income inequality" exercises; rely on similar data but apply different assumptions and imputation methods.	No	Yes: Tax Data and Other Administrative Registries, Wealth Surveys and National Accounts	Garbinti, Goupille and Piketty (2017); Piketty, Saez and Zucman (2018) Auten and Splinter (2019)
COMBINING SURVEY AND EXTERNAL DATA			
REPLACING TOP INCOMES: POPULATION SHARES (WEIGHTS) OF TOP INCOMES (β100%) AND NONTOP INCOMES [(1 - β)100%] UNCHANGED			
Parametric			

Replace upper tail by a Pareto distribution (or other models) estimated from tax data and use survey data for incomes below the income threshold that does not suffer from upper tail issues. Calculate total inequality using inequality decomposition formula. (Atkinson, 2007) and Alvaredo, 2011).	Yes	Yes: Tax Data	Alvaredo (2011); Alvaredo and Londoño (2013); Atkinson (2007); Atkinson, Piketty, and Saez (2011); Diaz-Bazan (2015); Jenkins (2017)
Nonparametric Imputation			
Replace the survey-based mean incomes for pre-specified fractiles (e.g. percentiles) by tax data cell-means; cut-off at which replacement takes place varies.	Yes	Yes: Tax Data	Alvaredo et al. (2017a); Bach et al. (2009); Burkhauser et al. (2016); Campos-Vazquez and Lustig (2017); Higgins, Lustig and Vigorito (2018); Dept for Work & Pensions, UK (2015)
Adjust to National Accounts: capital incomes of top β% in survey are grossed-up to match total income from capital in National Accounts. (Method also grosses up labor income).	Yes	Yes: National Accounts	Methodology: Altimir (1987) Application: CEPALStat (UN Economic Commission for LAC) until 2016
Use house prices to predict incomes in the upper tail.	Yes	Yes: House Prices	Methodology and application: van der Weide, Lakner and Ianchovichina (2018)
Combining Parametric and Nonparametric Imputation			
Distributional National Accounts (DINA) ("simplified version") are constructed for the adult population (20 yrs or older) starting from household surveys. Assume survey below percentile β' (e.g., 0.9) is reliable; replace by tax data above percentile β (e.g.,.995 percentile); assume quantile ratio upgrade factor rises linearly in between β' and β (interpolation to "join" both distributions); if data comes in form of tabulations, apply generalized Pareto (Blanchet, Fournier, and Piketty, 2017); add tax-exempt capital income (undistributed profits); gross-up to national accounts totals.	Yes	Yes: Tax Data, National Accounts, Rich Lists[67]	Methodology: Alvaredo et al. (2017b and 2018) Applications: Alvaredo, Assouad and Piketty (2018); Chancel and Piketty (2017); Novokment, Piketty, and Zucman (2017); Piketty, Yang and Zucman (2019) Other applications of parametric and nonparametric imputation methods with combined data: Bricker, Hansen and Henriques Volz (2019); Bustos and Leyva (2017); Lakner and Milanovic (2016)
REWEIGHTING: POPULATION SHARES (WEIGHTS) OF TOP INCOMES (β100%) AND NONTOP INCOMES ((1 - β)100%) CHANGE			
Reweighting Microdata			
Poststratification: replace the expansion factors in sample (base weights) by new weights from external sources (e.g., tax and social security records).	Yes	Yes: Tax Data	Methodology: Biemer and Christ (2008); Bourguignon (2017b) Applications: Blanchet, Flores and Morgan (2018); Burkhauser et al. (2017); Campos-Vazquez and Lustig (2017); Dept. for Work & Pensions (2015); Flores (2019); Higgins, Lustig and Vigorito (2018)

[67] For example, as published by the US-based magazine *Forbes*.

		Yes: Tax Data and National Accounts	Morgan (2018)
“Extreme” poststratification: assume achieved (whole) survey represents only bottom share of population calculate total inequality using inequality decomposition formula. That is, assume survey data is the (1 - β)100% instead of 100%; estimate the Gini for redefined bottom (1 - β)100%; estimate Gini for top (β100%) with tax data; and apply Atkinson (2007) and Alvaredo (2011) formula to estimate total Gini.	Yes	Yes: Tax Data	Methodology: Atkinson and Bourguignon (2000, 2015) Applications: Anand and Segal (2015); Higgins, Lustig and Vigorito (2018)

Note: References are presented in alphabetical order by last name. The mapping of studies to methods under the References column should be viewed as an approximation because studies may apply more than one method and, thus, can also appear more than once.

Can one identify which correction approaches might be better suited for addressing one or more of the issues described in Figure 1? The first matter that a researcher must determine is whether the sample survey and the distribution in the target population have the same support. One way to check this is by comparing the density functions of the sample and, for example, tax-based data, which in general will be closer to the “true” distribution. Fortunately, an increasing number of countries are publishing information from tax records (if only for certain years and often in tabulations rather than unit records) so such comparisons of right-hand tails can be done. Most likely, the comparisons will reveal that the support is not the same; in particular, the maximum incomes will not be similar. This means that within survey corrections will not be able to address the bias (or imprecision) in inequality measures introduced by the missing rich problem in a satisfactory way. Confronted with such a situation, the researcher may decide to rely on tax data only. As discussed above, however, tax data is not problem-free. For the purposes of measuring inequality, one key problem is that in most countries, tax data—if obtained—leaves out significant portions of the population due to informality. Since informality is more likely to occur at lower income levels, tax data is likely to suffer from noncoverage of the bottom portion of the distribution to a greater degree than surveys.

Since neither within-survey correction methods nor using just tax data are satisfactory, combining surveys with external information such as tax records, National Accounts or rich lists appears more promising. However, there is little or no guidance from theory or statistical testing regarding which specific method to pursue next.[68] Bringing out of survey information into the survey distorts the sample frame and there is no way of knowing the counterfactual. Using theoretical distributions at the top does not say anything on whether these distributions mimic real data properly and here too there is no counterfactual. The reweighting methods generally

[68] Using linked tax- and survey-data for Uruguay, Higgins, Lustig, and Vigorito (2018) find that “true” inequality is overestimated in 30% of the simulations.

rely on quite strong assumptions and they require substantial fine tuning based on the data at hand to be viable.

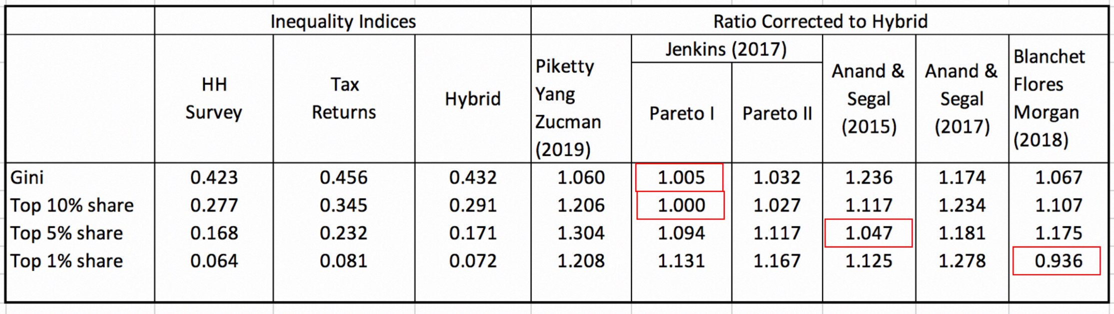
For the methods that adjust data to match National Accounts, there are no statistical tests or calibration methods to assess whether the assumed allocation to specific individuals of gaps between survey totals and National Account totals approximates the true distribution. Inequality measures can be very sensitive to specific assumptions. The current debate between Piketty, Saez and Zucman (op. cit.) and Auten and Splinter (op. cit.) on income inequality trends in the United States is very illustrative. Both set of authors rely on the same information sources. Micro-files from tax returns are combined with National Accounts to generate "consistent income" inequality measures. However, their conclusions about what happened to the top 1% and bottom 50% of the distribution since 1979 differ sharply. For instance, Piketty, Saez and Zucman's estimated increase in the (after tax) income share of the top 1% is almost five times higher than in Auten and Splinter.[69]

As shown by Lustig and Vigorito (forthcoming), there will not necessarily be a single method that outperforms all .the other methods for every inequality measure. These authors tested the methods' accuracy as follows. From a unique linked survey and tax database available for Uruguay, the authors were able to construct what they call a hybrid sample: for every individual observed both in the survey and the tax data, the higher reported income is the one included in the hybrid achieved sample. The hybrid sample is assume to be the closest representation of the true distribution. Table 2 compares the accuracy of alternative methods in reproducing the inequality measures obtained with the income data in the hybrid. As can be observed, for the Gini coefficient and the top 10%, replacing the top 1% by a Pareto I model estimated with tax data (a la Jenkins, op. cit.) performs better than the other methods. However, the income share of the top 5% is more accurately estimated by the reweighting method proposed by Anand and Segal (2015). Finally, the reweighting-cum-replacing method proposed by Blanchet, Flores and Morgan (op. cit.) performs better in estimating the top 1% income share. In other words, there is no dominant method.[70]

Table 2 - Impact of Correction Methods on Inequality Measures for Linked Sample

[69] Between 1979 and 2014, Piketty, Saez and Zucman estimated that pre-tax (post-tax) top 1% shares increased by 9.0 (6.5) percentage points while Auten and Splinter estimated an increase of only 3.2 (1.4) percentage points. For the bottom 50%, the former estimated a decrease of post-tax income share of 6.2 percentage points while Auten and Splinter data found a decline of 2.2 percentage points. In fact, the results are so strikingly different that the latter estimate a real increase in the pre-tax incomes of the bottom 50% of nearly one-third while with the Piketty, Saez and Zucman data, the income of this group remained virtually unchanged.

[70] There is no dominant method either when the authors assume that the true distribution is the one found in the tax data.



	Inequality Indices			Ratio Corrected to Hybrid					
	HH Survey	Tax Returns	Hybrid	Piketty Yang Zucman (2019)	Jenkins (2017)		Anand & Segal (2015)	Anand & Segal (2017)	Blanchet Flores Morgan (2018)
					Pareto I	Pareto II			
Gini	0.423	0.456	0.432	1.060	1.005	1.032	1.236	1.174	1.067
Top 10% share	0.277	0.345	0.291	1.206	1.000	1.027	1.117	1.234	1.107
Top 5% share	0.168	0.232	0.171	1.304	1.094	1.117	1.047	1.181	1.175
Top 1% share	0.064	0.081	0.072	1.208	1.131	1.167	1.125	1.278	0.936

Source: Lustig and Vigorito, forthcoming.

A more promising solution to the missing rich problem will likely come from linked data. Eventually, in counries with reliable administrative registries, the statistical offices themselves could pre-populate the income data for 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 correct for coverage error, unit and item nonresponse, and underreporting and other measurement errors by, whenever appropriate, substituting administrative for survey data. The potential of linked data to address upper tail (and other) issues is high. The ability to obtain more accurate measures of inequality will increase substantially if governments would make available linked survey and tax data. Of prime importance is for governments to make the information from (anonymised) tax records available and allow for the linking through personal identification numbers between surveys and registries.[71] Other administrative registries at the national and cross-national level that trace incomes and wealth to specific individuals will allow for capturing incomes that are not included in tax records due to their characteristic (for example, undistributed profits) or tax evasion. In the meantime, since there is no perfect method and all methods entail some degree of arbitrariness—assumptions whose validity is very hard or impossible to test--, a recommendable strategy is to carry out systematic robustness checks and report ranges rather than single corrected inequality measures.

[71] As indicated above, the government of Uruguay has taken such a step and shared (a partial version of) this type of information with academics.

References

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

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

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

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

Alvaredo, F., L. Assouad and T. Piketty (2018), "Measuring lnequality in the Middle East 1990-2016:The World's Most Unequal Region?," WID.World Working Paper Series No. 2017/15, revised 2018, Paris School of Economics.

Alvaredo, F., A. B. Atkinson and S. Morelli (2016), "The Challenge of Measuring UK Wealth Inequality in the 2000s," Fiscal Studies 37 (1), pp. 13-33.

Alvaredo, F. and J. Londoño-Velez (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, March. http://www.commitmentoequity.org/publications_files/CEQWPNo12%20HighTaxationDevEcon Colombia1993-2010_19March2013.pdf

Alvaredo, F., A. B. Atkinson and S. Morelli (2017), "Top Wealth Shares in the UK Over More than a Century," Research Paper Series 01, Department of Economics, University Ca' Foscari of Venice, January, https://ssrn.com/abstract=2903853

Alvaredo, F., A. B. Atkinson, T. Piketty and E. Saez (2013), "The Top 1% in International and Historical Perspective," Journal of Economic Perspectives 27 (3), pp. 3-20.

Alvaredo, F., R. Campos-Vazquez, S. Garriga, and M. F. Pinto (2017), "Household Surveys, Administrative Records and National Accounts in Mexico 2009-2014. Is a Reconciliation Possible?," Powerpoint Presentation, LACEA Annual Meeting, Buenos Aires, November 11, 2017.

Alvaredo, F., A. B. Atkinson, T. Piketty, E. Saez and G. Zucman (2015a), The World Top Incomes Database, http://wid.world/wid-world/

Alvaredo, F., A. B. Atkinson, T. Piketty, E. Saez and G. Zucman (2015b), The World Wealth and Income Database-WID, http://www.parisschoolofeconomics.eu/en/research/data-production-and-diffusion/the-world-wealth-income-database/

Alvaredo, F., L. Chancel, T. Piketty, E. Saez and G. Zucman (2018), "Distributional National Accounts in the Context of the WID.World Project," chapter in *For Good Measure: Advancing Research* *on Well-Being Metrics Beyond GDP,* edited by Martine Durand, Jean-Paul Fitoussi, and Joseph E. Stiglitz, OECD report by the *High Level Expert Group on Measuring Economic Performance and Social Progress.*

Alvaredo, F., A. B. Atkinson, L. Chancel, T. Piketty, E. Saez and G. Zucman (2017b), "Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world," WID.World Working Paper, June http://wid.world/document/dinaguidelines-v1/

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

Auten, G. and D. Splinter (2019), "Income Inequality in the United States: Using Tax Data to Measure Long-term Trends.", unpublished working paper.

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

Atkinson, A. B. and F. Bourguignon (eds.) (2015), Handbook of Income Distribution, Vol. 2, North-Holland, Amsterdam: Elsevier.

Atkinson, A. B. and A. J. Harrison (1978), Distribution of Personal Wealth in Britain, Cambridge, UK: Cambridge University Press.

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

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

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

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

Bach, S., G. Corneo and V. Steiner (2009), "From Bottom To Top: The Entire Income Distribution In Germany, 1992-2003," Review of Income and Wealth 55(2), pp. 303-330, 06.
Belfield, C., J. Cribb, A. Hood and R. Joyce (2015), "Living Standards, Poverty and Inequality in the UK: 2015," Report 107, London: Institute for Fiscal Studies, http://www.ifs.org.uk/publications/7878

Biemer, P. and S. Christ (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., I. Flores and M. Morgan (2018), "The Weight of the Rich: Improving Surveys Using Tax Data," WID.world Working Paper Series N° 2018/12, October.

Blanchet, T., J. Fournier and T. Piketty (2017) "Generalized Pareto Curves: Theory and Applications," WID.world Working Paper 2017/03.

Bourguignon, F. (2015), "Appraising Income Inequality Databases in Latin America," in Ferreira, F. H. G. and N. Lustig, Appraising Cross-National Income Inequality Databases, special issue, Journal of Economic Inequality 13 (4), pp. 557-578.

Bourguignon, F. (2017a), "Equivalence Between Adjusting Income Levels and Sample Weights in Correcting Income Distributions," unpublished document, Paris School of Economics, June.

Bourguignon, F. (2017b), "Correcting Survey Data for Top Incomes," Powerpoint Presentation, Methodological Advances in Fiscal Incidence Analysis, Commitment to Equity Institute (Tulane University), Universidad de San Andrés, Buenos Aires, November 7 -8, 2017.

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

Burkhauser, R. V., S. Feng and J. Larrimore (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., J. Larrimore and S. Lyons (2016), "Measuring Health Insurance Benefits: The Case of People with Disabilities," Contemporary Economic Policy 35 (3), pp. 439-456, doi:10.1111/coep.12213

Burkhauser, R. V., S. Feng, S. P. Jenkins and J. Larrimore (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., N. Herault, S. P. Jenkins and R. Wilkins (2017), "Survey Under-coverage of Top Incomes and Estimation of Inequality: What is the ole of the UK's SPI Adjustment?," NBER Working Paper 23539, June.

Bustos, A. and G. Leyva (2017), "Towards a more realistic estimate of the income distribution in Mexico," United Nations Economic Commission for Europe, Conference of European Statisticians, Expert meeting on measuring poverty and equality 26-27 September 2017, Budva, Montenegro.

Campos-Vazquez, R. and N. Lustig (2018), "Labour income inequality in Mexico: Puzzles Solved and Unsolved," UNU-WIDER Working Paper 2017/186, November.

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

CEPALSTAT (UN Economic Commission for Latin America and the Caribbean), http://estadisticas.cepal.org/cepalstat/WEB_CEPALSTAT/Portada.asp

Chancel, L. and T. Piketty (2017), "Indian income inequality, 1922-2015:
From British Raj to Billionaire Raj?," WID.World Working Paper Series No. 2017/11, revised 2018, Paris School of Economics.

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

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

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

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

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

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

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

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

Fesseau, M. and M. L. Mantonetti (2013), "Distributional Measures Across Household Groups in a National Accounts Framework," Working Paper 53, EUROSTAT and OECD, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=STD/CSTAT/WPNA(2013)10/RD&docLanguage=En

Fisher, J., D. Johnson, T. Smeeding and J. Thompson (2016), "Inequality in 3D: Income, Consumption and Wealth," Working Paper Series, Washington Center for Equitable Growth, September.

Fishlow, A. (1973), "Some Reflections on Post 1964 Brazilian Economic Policy," in Stepan, A. (ed.), Authoritarian Brazil, New Heaven: Yale University Press.

Flachaire, E. (2018), "On the Estimation of Inequality Measures," powerpoint presentation at the "Workshop on Harmonization of Household Surveys, Fiscal Data and National Accounts: Comparing Approaches and Establishing Standards," Paris School of Economics, May 17-18, 2018.

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

Garbinti, B., J. Goupille-Lebret and T. Piketty (2017), "Income Inequality in France, 1990-2014: Evidence from Distributional National Accounts (DINA)," WID.world Working Paper, December.

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

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

Higgins, S., N. Lustig, and A. Vigorito (2018), "The Rich Underreport Their Income: Assessing Biases In Inequality Estimates And Correction Methods Using Linked Survey And Tax Data." CEQ Working Paper 70, CEQ Institute, Tulane University, September. Also published in ECINEQ, Working Paper 475, September 2018.

Hlasny, Vladimir (forthcoming), "Parametric Representation of the Upper Tail of Income Distributions: Options, Historical Evidence and Model Selection" CEQ Working Paper 90, CEQ Institute, Tulane University.

Hlasny, V and Verme, P. (2015), Top Incomes and the Measurement of Inequality: A Comparative Analysis of Correction Methods using Egyptian, EU and US Survey Data. Paper presented at the 6th meeting of the Society for the Study of Economic Inequality (ECINEQ), Luxembourg, July 13-15, 2015.
http://www.ecineq.org/ecineq_lux15/FILESx2015/CR2/p145.pdf

Hlasny, V. and P. Verme (2017), "The Impact of Top Incomes Biases on the Measurement of Inequality in the United States," ECINEQ Working Paper 452, November.

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

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 P. Van Kerm (2009), "The Measurement of Economic Inequality," in Salverda, W., B. Nolan and T. M. Smeeding (eds.), The Oxford Handbook of Economic Inequality, Oxford, UK: Oxford University Press.

Jenkins, S. P., R. V. Burkhauser, S. Feng and J. Larrimore (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.

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

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

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

Lakner, C. and B. Milanovic (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, http://dx.doi.org/10.1093/wber/lhv039.

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

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

Lustig, N. (2018), "Measuring the Distribution of Household Income, Consumption and Wealth: State of Play and Measurement Challenges," chapter 3 in *For Good Measure: Advancing Research on Well-Being Metrics Beyond GDP,* edited by Martine Durand, Jean-Paul Fitoussi, and Joseph E. Stiglitz, OECD report by the *High Level Expert Group on Measuring Economic Performance and Social Progress.*

Lustig, N. (ed.) (2018), Commitment to Equity Handbook. Estimating the Impact of Fiscal Policy on Inequality and Poverty, Washington, DC: Brookings Institution Press and CEQ Institute, Tulane University. http://www.commitmentoequity.org/publications/handbook.php

Lustig, N. and A. Vigorito. (Forthcoming), "Correction Methods and the Upper Tail: An Assessment using Linked Survey and Tax Data for Uruguay." CEQ Working Paper 89, CEQ Institute, Tulane University.

Matthews, Dylan (2018), "A new study says much of the rise in inequality is an illusion. Should you believe it?"published in Vox on January 10, 2018. https://www.vox.com/policy-and-politics/2018/1/10/16850050/inequality-tax-return-data-saez-piketty

Mistiaen, J. and M. Ravallion (2003), "Survey Compliance and the Distribution of Income," Policy Research Working Paper 2956, World Bank, Development Research Group, Washington, DC.

Meyer, B. D. and N. Mittag (2019), "Combining Administrative and Survey Data to Improve Income Measurement," NBER Working Paper No. 25738, April.

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

Morelli, S. and E. Muñoz (2019), "Unit Nonresponse Bias in the Current Population Survey," unpublished paper, Stone Center on Socio-Economic Inequality and CUNY Graduate Center, New York, USA.

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

Novokmet F., T. Piketty, and G. Zucman (2017), "From Soviets to Oligarchs: Inequality and Property in Russia 1905-2016," WID.world Working Paper Seires 2017/09, Paris School of Economics.

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

Piketty, T. and E. Saez (2003), "Income Inequality in the United States 1913-1998," Quarterly Journal of Economics 118 (1), pp. 1-39.

Piketty, T., E. Saez, and G. Zucman (2018), "Distributional National Accounts: Methods and Estimates for the United States", *The Quarterly Journal of Economics*, Volume 133, Issue 2, May 2018, Pages 553–609.

Piketty, Thomas, Li Yang, and Gabriel Zucman (2019), "Capital Accumulation, Private Property, and Rising Inequality in China, 1978–2015." *American Economic Review*, 109 (7): 2469-96.

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

Ruiz, N. and N. Woloszko (2016), "What Do Household Surveys Suggest About the Top 1% Incomes and Inequality in OECD Countries?" OECD Economics Department Working Paper 1265, January. http://www.oecd-ilibrary.org/economics/what-do-household-surveys-suggest-about-the-top-1-incomes-and-inequality-in-oecd-countries_5jrs556f36zt-en

Saez, E. (2003), "The effect of marginal tax rates on income: a panel study of 'bracket creep'," Journal of Public Economics 87 (5), pp. 329-347.

Saez, E. and G. Zucman (2016), "Wealth Inequality in the United States Since 1913: Evidence from Capitalized Income Tax Data," The Quarterly Journal of Economics 131 (2), pp. 519-578.

Szekely, M. and M. Hilgert (1999), "What's Behind the Inequality We Measure: An Investigation Using Latin American Data", Research Department Working Paper Inter-American Development Bank.

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

Zwijnenburg, J., S. Bournot and F. Giovanelli (2017), "Expert Group on Disparities in a National Accounts Framework,: Results from the 2015 Exercise," OECD Statistics Working Papers, 2016/10, OECD Publishing, Paris, http://www.oecd-ilibrary.org/economics/expert-group-on-disparities-in-a-national-accounts-framework_2daa921e-en

Zucman, G. (2015), *The Hidden Wealth of Nations. The Scourge of Tax Havens,* London: The University of Chicago Press.