CEQ INSTITUTE
COMMITMENT TO EQUITY
Tulane University

The Missing Rich in Household Surveys: A Survey of Causes and Correction Methods

Nora Lustig

Samuel Z. Stone Professor and Director of CEQ Institute

Tulane University

Nonresident Senior Fellow CGD and IAD

Visiting Researcher, PSE

Workshop on "Harmonization of household surveys, fiscal data and national accounts: comparing approaches and establishing standards"

Paris School of Economics, CEQ Institute, Stone Center on Socio-economic Inequality – Paris, May 17-18, 2018

Cite as:

- Lustig, N. (Forthcoming), “The Missing Rich in Household Surveys: Causes and Correction Methods,” CEQ Working Paper 75 (Commitment to Equity (CEQ) Institute, Tulane University).

Outline

- Motivation
- Causes and types of errors
- Correction approaches
- Impact of approaches on inequality

Motivation

The CEQ Institute: a brief description

Mission: The CEQ Institute works to reduce inequality and poverty through comprehensive and rigorous tax and benefit incidence analysis, and active engagement with the policy community

Objective: To measure the impact of fiscal policy on inequality and poverty across the world using a comparable framework

Workstreams:

- Research-based policy tools (CEQ Handbook) and country studies
- Data Center
- Advisory and training services
- Bridges to policy (IMF, World Bank)

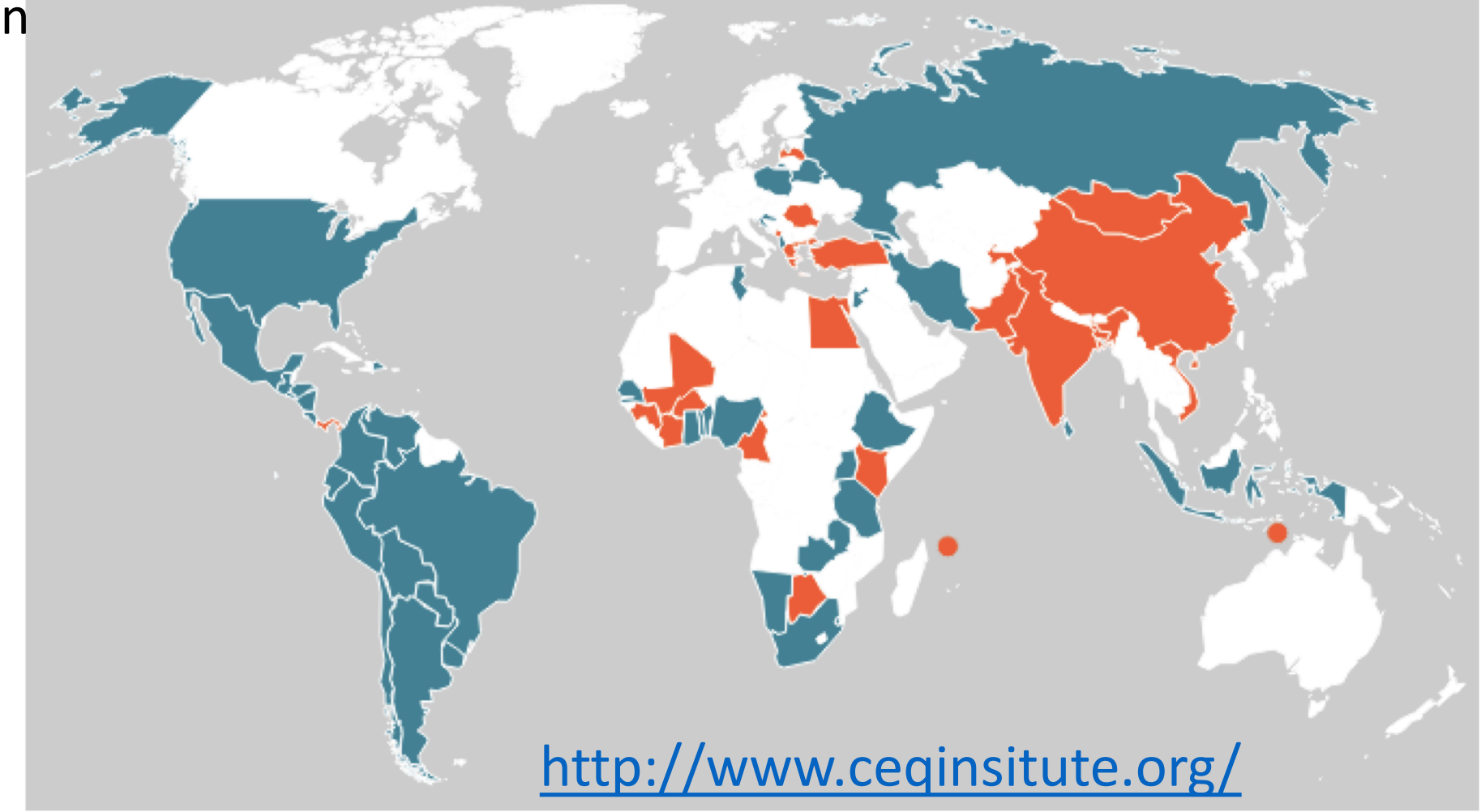
Bill & Melinda Gates Foundation: $4.9 million for 5 years

CEQ INSTITUTE™
COMMITMENT TO EQUITY
Tulane University

-Summary stats
-Harmonized Microdata (2019)

Home | About | Data Center | Handbook | Publications | News

Browse Map for Publications by Country



http://www.ceqinsitute.org/

Fiscal redistribution assessments:

- 41 finished
- 25 in progress

➢Nearly 80% of world's extreme poor

Limitations of fiscal redistribution analysis if top incomes are not well captured:

- Countries covered by CEQI feature among the largest discrepancies between aggregate income/consumption totals from hh surveys and National Accounts (NA).
  - For example, in Mexico, the difference is between 50 and 60%

- Assessing progressivity of fiscal system:
  - Fiscal incidence analysis without top incomes does not provide an accurate description of the extent to which inequality is reduced through taxes and transfers

- Assessing policy options: CEQI supplies the IMF with policy simulations and…
  - Analysis without top incomes seriously limits the assessment on how much can countries expand tax-based redistribution (or mitigate fiscal cuts) through, for example, personal income taxes or wealth taxes

CEQI Research on Top Incomes

- Higgins, S., N. Lustig and A. Vigorito “Top Incomes, Issues with Survey Data and Inequality: Evidence from Simulations and Linked Income and Tax Return Data” (results on F session)

- Scott, J., G. Leyva, N. Lustig, S. Martinez-Aguilar and E. de la Rosa “CEQ-DINA for Mexico” (preliminary results on F session)

- Basu, S., V. Hlasny and N. Lustig “Correcting for Underreporting and Undercoverage of Top Incomes in Household Surveys: An Empirical Assessment of Alternative Approaches for Low and Middle-income Countries,” (just started)

Inequality in the Giants Project
(Alvaredo, Bourguignon, Ferreira, Leibbrandt, Lustig, Tarp)

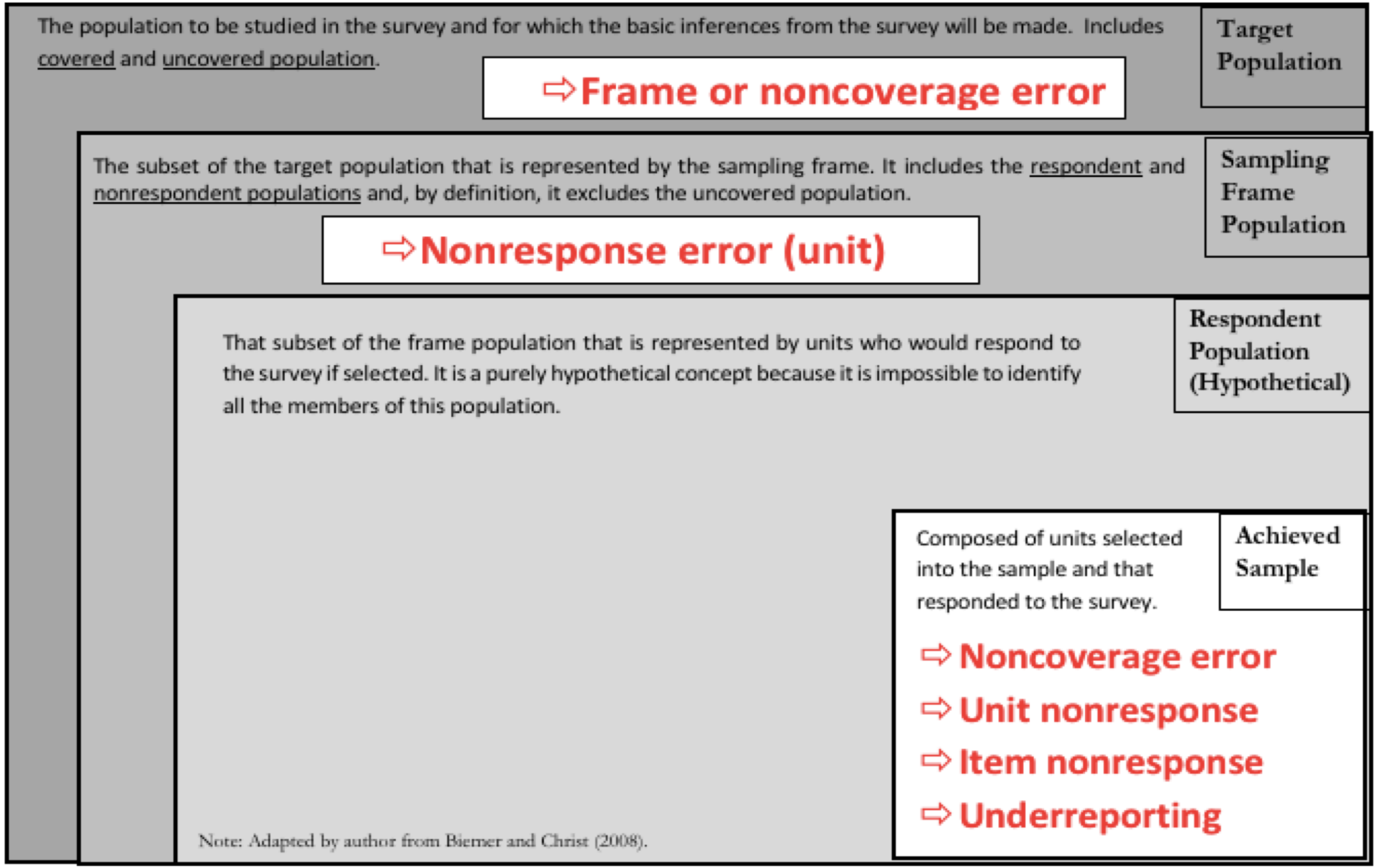
In partnership with UNU-WIDER: Mexico

- Campos, R. and N. Lustig Labour income inequality in Mexico: Puzzles Solved and Unsolved, UNU-WIDER Working Paper 2017/186, November 2017.
  - Item nonresponse: hot deck; Unit nonresponse: reweighting
- Ibarra, C. A, and J. Ros. "The Decline Of The Labour Share In Mexico: 1990–2015," WIDER Working Paper 2017/183 Helsinki: UNU-WIDER, 2017.
  - Disentangles mixed incomes from the nonwage value added totals in National Accounts
- Scott, J., E. de la Rosa, and R. Aranda. "Inequality And Fiscal Redistribution In Mexico: 1992–2015," WIDER Working Paper 2017/194 Helsinki: UNU-WIDER, 2017.
  - Matches survey totals to National Accounts by broad categories
- Alvaredo, F., F. Pinto and S. Garriga. "Household Surveys, Administrative Records, and National Accounts in Mexico 2009-2014. Is a reconciliation possible?" [in progress]
  - Scales-up average income/centile in hh surveys to match average income/centile in tax registries

Objective of correction:

1) Correct for bias generated by ”issues” with top incomes
   - Corrected microdata
   - Corrected distribution
   - Corrected inequality indicators
2) Obtain distributional national accounts

Missing Rich in HH Surveys
A Typology of Causes


The population to be studied in the survey and for which the basic inferences from the survey will be made. Includes covered and uncovered population.
Target Population
⇨Frame or noncoverage error
The subset of the target population that is represented by the sampling frame. It includes the respondent and nonrespondent populations and, by definition, it excludes the uncovered population.
Sampling Frame Population
⇨Nonresponse error (unit)
That subset of the frame population that is represented by units who would respond to the survey if selected. It is a purely hypothetical concept because it is impossible to identify all the members of this population.
Respondent Population (Hypothetical)
Composed of units selected into the sample and that responded to the survey.
Achieved Sample
⇨ Noncoverage error
⇨ Unit nonresponse
⇨ Item nonresponse
⇨ Underreporting
Note: Adapted by author from Biemer and Christ (2008).

Sampling design issues: from target to frame population

- Frame or noncoverage error: erroneous exclusion of units belonging to the target population owing to imperfections in building the frame.
  - Concern: bias in parameter estimates due to frame noncoverage.

Data collection issues: from frame population to achieved sample:

- Unit nonresponse: units included in the frame population but do not respond.
- Item nonresponse: units that do not respond to income question.
- Misreporting: units respond to income question, but inaccurately. For top incomes, underreporting is main concern.

    - Concern: bias in parameter estimates due to any of the above.

Data preparation: incomes above/below a certain level or observations above or below a certain quantile are excluded (Cowell & Flachaire, 2015):

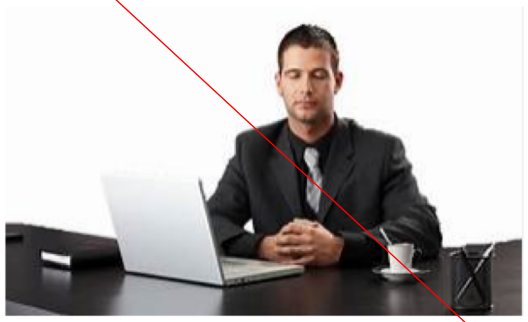
- Truncation: without information on excluded sample above/below a certain threshold.
- Censoring: with information on sample proportion of excluded sample; top coding.
- Trimming: without information on excluded sample above/below a certain quantile.
  - Concern: bias in parameter estimates due to any of the above.

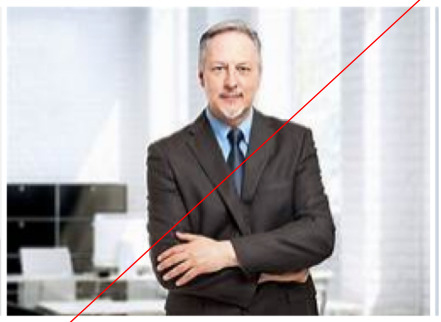
Approaches: A Taxonomy

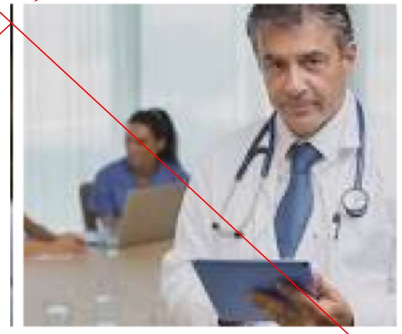
	Survey (support is the same)	**Survey & Admin Data** (support is not the same)
Replacing (Weight of top/bottom kept fixed)	Semiparametric	• Semiparametric • Rescaling to tax data &/or NA • Replacing top shares
Reweighting (Weight of top/bottom changes)	Replaces base weights	Replaces base weights of bottom to make room for new observations at the top

Survey
Replacement & reweighting

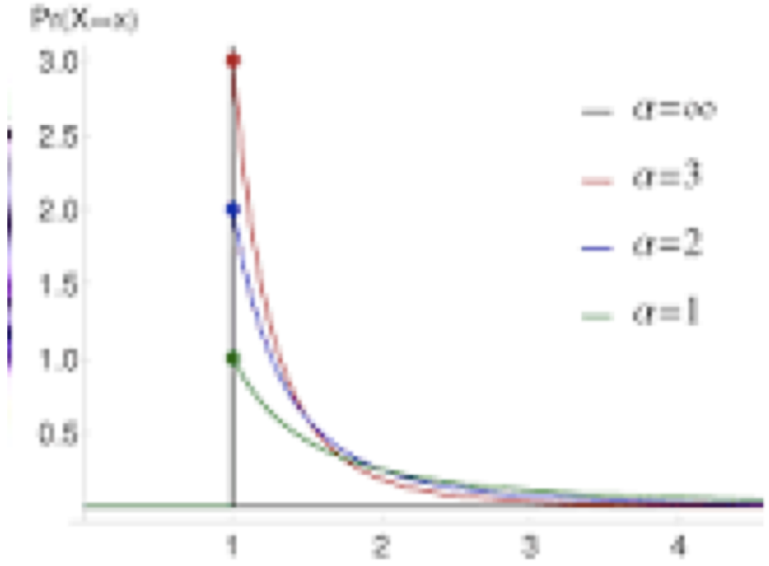
TOP “1 PERCENT” IN SURVEY ARE TOSSED OUT







Method: Replaces the top x% of the distribution by parametric distribution (e.g., Pareto) estimated with observations in survey. (Cowell, 2009; Flachaire & Cowell, 2015)

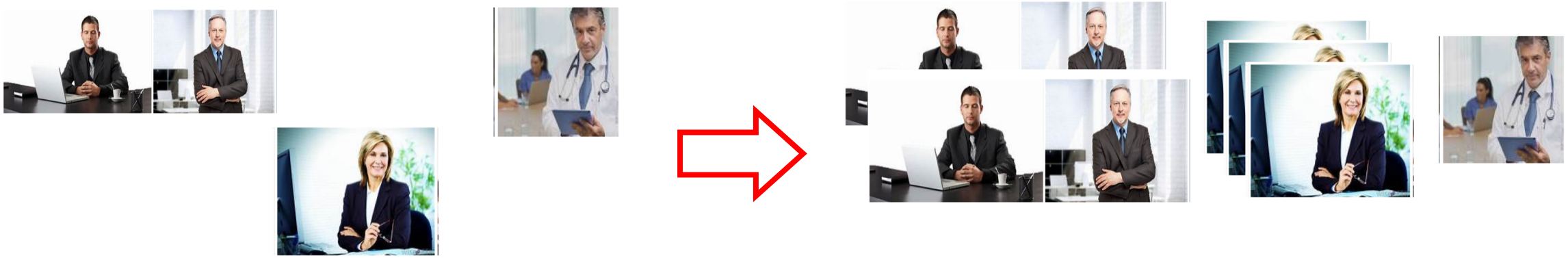


REPLACING by a parametric distribution

Example: Cowell and Flachaire, 2007; Atkinson et al., 2011; Hlasny and Verme, 2018

TOP INCOME OBSERVATIONS IN SURVEY ARE REWEIGHTED TO TAKE INTO ACCOUNT NONCOVERAGE, UNIT AND ITEM NONRESPONSE

New weights product of base weights, nonresponse adjustment factor, and the poststratification weights. (Little & Rubin, 2014; Biemer & Christ, 2008)



REWEIGHTING

Weights of these observations were reduced

Example: Atkinson & Mickelwright, 1983; Mistiaen and Ravallion, 2003; Korinek et al., 2006; Hlasny & Verme, 2018

Survey data may not be enough…

Support between sample and true distribution is **not** the same:

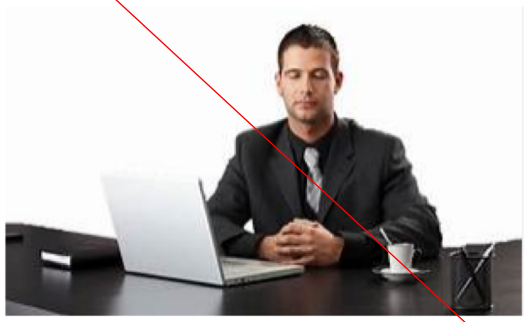
For a discrete distribution, in the sample $P(X = x) = 0$ whereas $P(X = x) > 0$ in the population; or, for a continuous distribution, $f_x(x) = 0$ in the sample whereas $f_x(x) > 0$ in the population

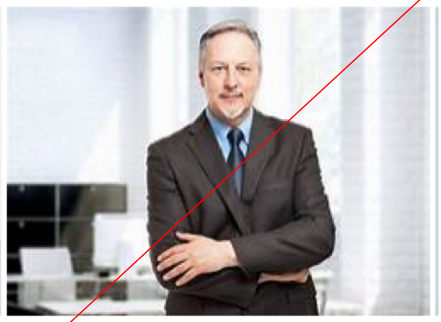
⇨ Within-survey reweighting can never properly correct.

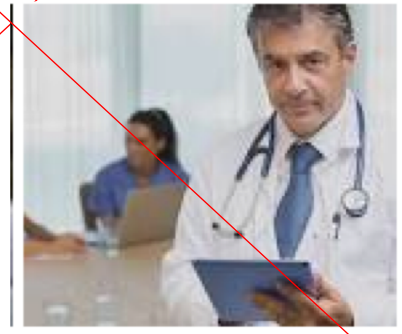
⇨ Within-survey replacing, limited because parametric distributions are estimated using observations in survey.

Survey & Admin Data Replacement & reweighting

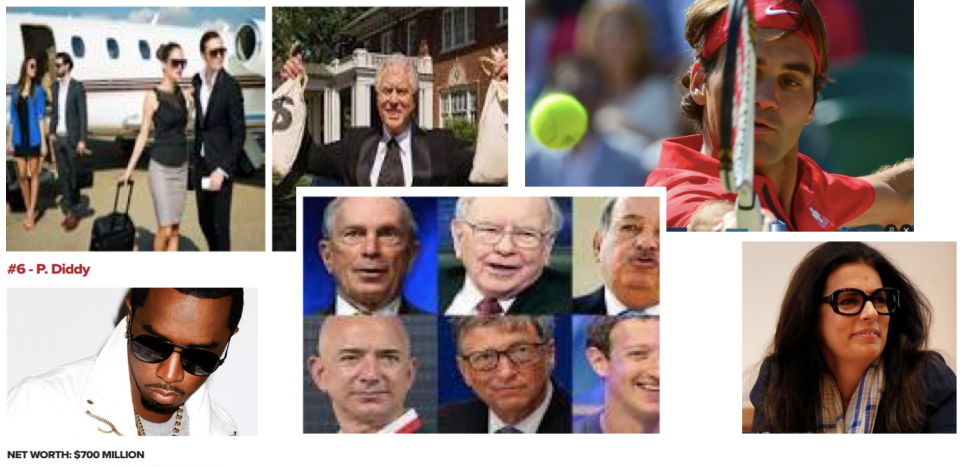
TOP “1 PERCENT” IN SURVEY ARE TOSSED OUT

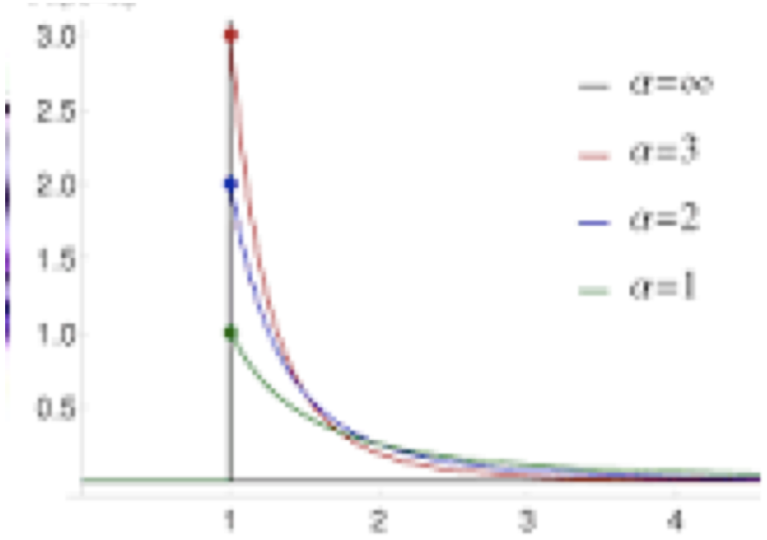






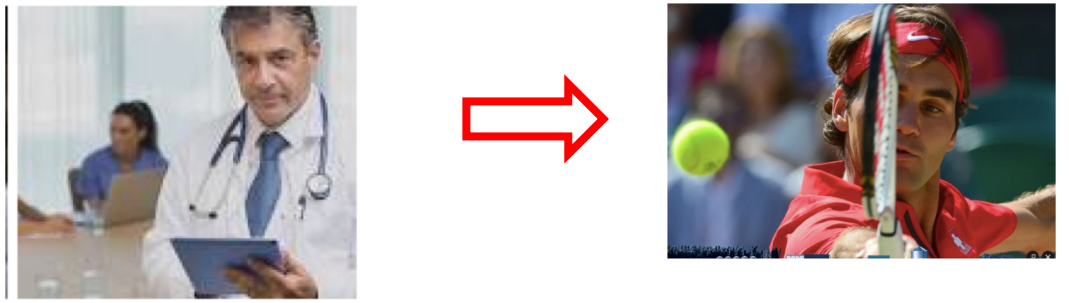
Method: Replaces the top x% of the distribution by parametric distribution (e.g., Pareto) estimated with tax data. (Atkinson, 2007; Atkinson et al., 2011; Alvaredo (2011); Jenkins, 2017)





REPLACING by a Parametric distribution

- Replaces incomes (e.g., means by centile) beyond a certain threshold by means from tax records

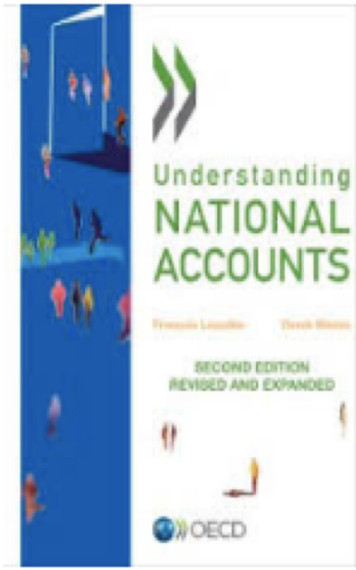


REPLACING incomes for top in survey by incomes from tax data (rescaling)

Examples: papers in WID.World

TOP 1 PERCENT IN SURVEY

TOP INCOMES
ARE SCALED-UP
TO MATCH NATIONAL
ACCOUNTS



NAQ Table Number	Indicators	Quarterly accounts
	GDP, value added and employment	
	Nominal and volume measure of GDP by industry or by expenditure components	*Minimum requirement*
1.1	GDP expenditure at current prices	*Optional minimum requirement*
1.2	GDP expenditure at constant prices	*Optional minimum requirement*
2.1	Value added and GDP in current prices by industry	*Optional minimum requirement*
2.2	Value added and GDP in constant prices by industry	*Optional minimum requirement*
2.3	Value-added components by industry, current prices	*Recommended*
	Employment by industry	*Recommended*
	Integrated accounts and tables, including integrated satellite accounts	
1.3-4.1	Accounts for the total economy (until net lending)	*Minimum requirement*
	Supply and use tables	*Desirable*
	Institutional sector accounts (until net lending)	
4.2	Rest-of-the-world account (until net lending)	*Minimum requirement*
4.3	Non-financial corporations sector accounts	*Recommended*
4.4	Financial corporations accounts (until net lending)	*Recommended*
4.5	General government sector accounts (until net lending)	*Recommended*
4.6	Household sector accounts (until net lending)	*Recommended*
4.7	Non-profit institutions serving households sector accounts (until net lending)	*Recommended*
	Financial accounts	
4.1–4.7	Financial accounts for all sectors	*Desirable*
	Balance sheets and other changes in asset accounts	
	Balance sheets, revaluation and other volume changes in asset accounts for all sectors	*Desirable*

REPLACING top incomes from Survey to match NA (rescaling)

Example: Altimir, 1979, 1987; WID.WORLD DINA; Lakner & Milanovic, 2016

TOP ‘1 PERCENT’ IN SURVEY NO LONGER IS TOP ‘1 PERCENT ‘ (THEY LOST WEIGHT)
AND
NO LONGER AT THE VERY TOP BUT THEY ARE NOT “TOSSED OUT” OR PARAMETRIZED

Changes *base* weight of bottom to make room for new observations at top obtained from, for ex., tax data; simplest version is a proportional reduction of weights

REWEIGHTING

Example: Anand and Segal, 2016

Can we assign methods to “issues”?

Data preparation:
Truncation, Censoring, Trimming

- Replace top x percent in survey by a parametric distribution

  - Overview of methods in Flachaire & Cowell, 2015
  - Application examples: Alfons, Temple, & Filzmoser (2013); Burkhauser et al. (2012); Cowell and Victoria –Feser (1996); Cowell and Flachaire, 2007; Hlasny and Verme, 2018; Ruiz and Woloszko (2015)

Sampling design and data collection issues:
Noncoverage, Unit nonresponse, Item nonresponse, Misreporting

If support is the same:

- Reweighting

  (Recall that if support is the same, in principle, reweighting can be transformed in an equivalent replacing exercise, Bourguignon, 2017a)

  - Overview of methods in Little and Rubin, 2014; Biemer and Christ, 2008
  - Application examples: Atkinson and Micklewright (1983); Autor et al. (2008); Burkhauser, Feng, and Larrimore (2010); Campos-Vazquez and Lustig (2017); Hlasny and Verme, 2018; Jenkins et al. (2011); Korinek, Mistiaen, and Ravallion (2006); Lemieux (2006); Mistiaen and Ravallion (2003)

If support is **NOT** the same:

1. Replacing by a parametric distribution estimated with tax data (with either unitary data or shares)
   - Application examples: Atkinson et al., 2011; Jenkins, 2017

2. Replacing income shares of top in survey by income shares in tax data in inequality measures
   - Application examples: Atkinson (2007); Atkinson, Piketty, and Saez (2011); Alvaredo (2011)

3. Replacing (rescaling) average income by quantile in survey with the average income by quantile in tax data
   - Application examples: papers in WID.World: Garbinti, Goupille and Piketty (2017); Novokment, Piketty, and Zucman (2017); Piketty, Saez and Zucman (2016); Piketty, Yang and Zucman (2016); Alvaredo, Garriga and Pinto (2017)

4. Replacing (rescaling) incomes for top earners in survey to match totals of in NA
   - Application examples: Altimir (1979, 1987); Lakner and Milanovic, 2016

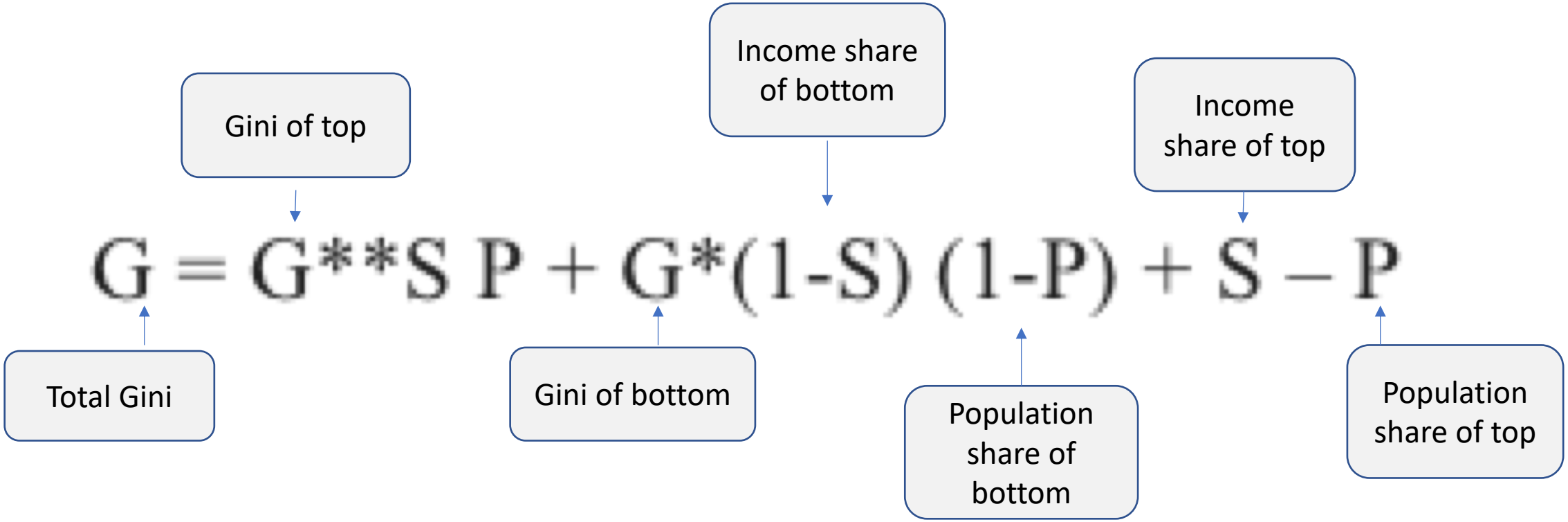
5. Reweighting survey observations to represent “bottom” and adding shares for top obtained from tax data
   - Application examples: Anand and Segal, 2016; Bourguignon (2017b)

Sampling design and data collection issues:
Noncoverage, Unit nonresponse, Item nonresponse, Misreporting

If support is NOT the same and tax data is fairly comprehensive, start from tax records and use survey to complement/complete the information

- Application examples: Piketty, Saez and Zucman (2016)

Correction Methods & Impact on Corrected Inequality



Source: Dagum (1997), Atkinson and Bourguignon (2000), Alvaredo (2011)

Total derivative

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

where:

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

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

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

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

- Impact depends on the sign of changes of 4 components that can change
- If we can't determine the sign ex ante, there can be 16 possible outcomes (2 x 2 x 2 x 2)

However, it is possible to show that…

- Replacing methods that result in an increase in the share of income accruing to the top, will always yield a higher inequality indicator if dG** is positive or equal to zero; and, result depends when correction yields a negative dG**
  - Replacing by a parametric distribution can go either way
  - Replacing by rescaling to tax data and/or NA will (almost) always result in an increase in inequality
- Reweighting a la Anand and Segal, will always increase inequality because it adds a positive term to the uncorrected inequality indicator which, by assumption, is inequality based on the survey

References

- Bibliography will be available on the workshop’s shared folder

Thank you