
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:

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
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).
  o For example, in Mexico, the difference is between 50 and 60%

• Assessing progressivity of fiscal system:
  o 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…
  o 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

  ➢ Item nonresponse: hot deck; Unit nonresponse: reweighting

  ➢ Disentangles mixed incomes from the nonwage value added totals in National Accounts

  ➢ Matches survey totals to National Accounts by broad categories

  ➢ 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.

⇒ 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.

⇒ 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.

⇒ 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
<table>
<thead>
<tr>
<th>Method</th>
<th>Survey (support is the same)</th>
<th>Survey &amp; Admin Data (support is not the same)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replacing</td>
<td>Semiparametric</td>
<td>• Semiparametric</td>
</tr>
<tr>
<td>(Weight of top/bottom kept fixed)</td>
<td></td>
<td>• Rescaling to tax data &amp;/or NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Replacing top shares</td>
</tr>
<tr>
<td>Reweighting</td>
<td>Replaces base weights</td>
<td>Replaces base weights of bottom to make room for new observations at the top</td>
</tr>
<tr>
<td>(Weight of top/bottom changes)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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)

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

REWIEIGHTING

Weights of these observations were reduced
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

$\Rightarrow$ Within-survey reweighting can never properly correct.
$\Rightarrow$ Within-survey replacing, limited because parametric distributions are estimated using observations in survey.
Survey & Admin Data
Replacement & reweighting
REPLACING by a Parametric distribution

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

Examples: papers in WID.World

- Replaces incomes (e.g., means by centile) beyond a certain threshold by means from tax records
TOP 1 PERCENT IN SURVEY

TOP INCOMES ARE SCALED-UP TO MATCH NATIONAL ACCOUNTS

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

**REWIGHTING**

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

  o Overview of methods in Flachaire & Cowell, 2015
  o 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
Total Gini = Gini of top + Gini of bottom + Income share of top + Income share of bottom + Population share of top

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