WPS7458

Policy Research Working Paper 7458

Exploring the Sources of Downward Bias in Measuring Inequality of Opportunity

Gabriel Lara Ibarra
Adan L. Martinez Cruz

WORLD BANK GROUP
Poverty Global Practice Group
October 2015

Policy Research Working Paper 7458

Abstract

This study analyzes the extent of downward bias in the calculation of inequality of opportunity for continuous outcomes such as income. A typically recognized source of bias is the unobserved circumstances as there is a limited set of variables available in household and labor force surveys. Another previously overlooked source is the likely unobservable nature of top incomes. Using Monte Carlo simulations where the underlying inequality of opportunity is predetermined at various levels, the study presents three key findings. First, the omission of a relevant circumstance can bias the inequality of opportunity estimate by as much as 80 percent, depending on how much variation of the outcome such circumstance explains. Second, not observing the top 5 percent of the income distribution can lead to downward biases of anywhere between 12 and 35 percent, and the combination of missing the most favored population and even one relevant circumstance exacerbates the bias of the empirical estimates. The third key result is that the estimated inequality of opportunity is strongly correlated with the amount of variation in the outcome variable explained by the combination of circumstances (measured by the R2). This result suggests that in empirical applications, the inequality of opportunity estimate can be roughly (and quickly) approximated using simple econometric techniques.

This paper is a product of the Poverty Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at glaraibarra@worldbank.org .

Exploring the Sources of Downward Bias in Measuring Inequality of Opportunity

Gabriel Lara Ibarra
The World Bank

Adan L. Martinez Cruz
ETH- Zurich

JEL codes: D63, C15

Keywords: Inequality of opportunity, mean log deviation, Monte Carlo, income distribution, top incomes.

Introduction

Income inequality has become firmly placed at the center stage of economic and policy debate. The reasons for this are the current discussions of how long-term welfare disparities have evolved over time and may continue on such a path (Piketty, 2014), the evidence on large concentrations of wealth in few individuals (OXFAM, 2014), and inequality's role in fueling discontent in several contemporaneous social movements (Los indignados in Spain, the Occupy protests in USA, the events in the middle east region that became known as the Arab Spring).[1]

However, understanding inequality in income, consumption or other such outcomes[2] and the role of policy in addressing it is itself subject to debate because not all inequality can be unambiguously deemed objectionable. On one side, if we take two individuals who exert different levels of effort, whomever exerts higher effort (demonstrated by attaining higher educational levels or working more) should be able to reap higher economic rewards, for instance, in the way of higher incomes. Inequality in this way then, is necessary to produce the right incentives to promote economic development. In contrast, following an egalitarian ethical point of view, inequalities originated in factors beyond individuals' responsibility are inequitable and must be compensated by society (Peragine, 2004). Moreover, this type of inequality may lead to reduction in economic growth as it favors human capital accumulation by individuals with better social origins rather than by individuals with more talent or skills.[3] This reasoning is not only a theoretical artifact. People do make a distinction between circumstances and efforts when judging distribution of outcomes such as income (Ramos and Van de Gaer, 2012).

Arguably, policy makers aiming to reduce inequality should not focus on the inequality caused by choices that individuals can be held responsible for, but instead address the inequality due to circumstances that prevent a "level playing field", circumstances upon which an individual happens to be born into but ultimately affect her available development life paths. Roemer (1993, 1998) has coined the term inequality of opportunity (IOO) to distinguish between inequality due to differences in circumstances beyond an individual's control and inequality of effort (IOE). IOE is inequality caused by choices an individual can be held responsible for. If focus is on IOO

[1] People participating in these social movements share the perceptions of rising inequality and decreasing economic mobility – "when Occupy Wall Street sprang up in parks and under tents, one of the many issues the protesters pressed was economic inequality" (see http://www.nytimes.com/2013/09/14/opinion/blow-occupy-wall-street-legacy.html). Statements about income inequality perceptions of Occupy movement's participants can be found in http://topics.nytimes.com/top/reference/timestopics/organizations/o/occupy_wall_street/index.html. Documentation of Los indignados movements can be found in http://elpais.com/tag/movimiento_indignados/a/.

[2] Other types of inequality such as access to reliable health services are clearly important for the development of the individual. An example of consequences from this type of inequality has been documented in the USA: infant mortality in USA has been linked to inequality at time of birth –infants born to non-white, non-college-educated, non-married US mothers have higher probabilities of post-neonatal mortality (Chen et al., 2014).

[3] From an economic point of view, a multiple state framework with borrowing constraints can be used to show that this type of inequality reduces economic growth as it favors human capital accumulation by individuals with better social origins rather than by individuals with more talent or skills. Following a similar reasoning, income inequality among people exerting different efforts rewards unequal effort and/or unequal talent –incentivizing people to work which in turns stimulates growth (Marrero and Rodriguez, 2013).

instead of inequality in outcomes, then public policies can be reframed. That is, instead of aiming for a policy that equalizes outcomes, public policy makers may put their efforts into designing a policy aiming to nullify, to the greatest extent possible, the effect of circumstances on outcomes, but allowing outcomes to be sensitive to individuals' effort (Roemer and Trannoy, 2013).

By distinguishing between unchosen circumstances and individual choices, equal opportunity theorists have shifted the focus to the individuals' responsibility. Implicit in this approach is the view that an optimal public policy must find an equilibrium between re-allocating initial resources and respect individuals' free will. Thus the goal of a policy must be to equalize what is available to individuals at the beginning of their journey –not at the end. Where and how individuals end up is their responsibility and the society should not interfere in this realm. Notice that, from this point of view, inequality in outcomes is neither desirable nor undesirable –i.e. the inequality of opportunity theory has no aversion to inequality in outcomes (see Roemer and Trannoy, 2013).

Despite the fact that the interest in inequality of opportunity has been fueled by the recent debate on whether large concentrations of wealth in few individuals is due to overcompensation of effort (e.g. Piketty, 2014), Roemer and Trannoy (2013) have highlighted that the theory of equal opportunity is not intended as a theory of distributive justice. They offer two reasons. A first one points the pragmatism behind this theory: this theory does not provide general rules to decide what people are responsible for. Practitioners infer the circumstances, and implicitly the effort, according to what they think a particular society rewards or punishes. A second reason refers to the recognition that this theory does not provide a view on what the proper rewards to effort consist in. That is, the theory of equal opportunity has no stand on the debate on whether individuals are compensated too much or too little.

Thus, after decades of being mostly theoretical, the literature on IOO has become very empirical. IOO has been analyzed in a wide array of realms such as income distribution, income taxation, health conditions, health care, educational achievement, and anti-poverty policy.[4] This rapidly growing empirical literature is based on the work by Bourguignon et al. (2007), Paes de Barros et al. (2009), and Ferreira and Gignoux (2011). In short, for a given outcome, IOO can be obtained as a ratio of the inequality between groups or "types" of individuals to the overall observed inequality. Between-group inequality is calculated as the inequality across types of individuals, where a type results from intersecting categorized variables that capture circumstances beyond an individual's control. The number of types in specific applications depends on the number of categories into which each circumstance is divided. When the types of individuals have been correctly defined–i.e. all relevant circumstances and categories have been taken into account- individuals belonging to a type are treated as homogeneous in their circumstances, and differences in outcomes across types are imputable to differences in circumstances. Thus, the

[4] Extensive recent reviews of the literature are provided by Roemer and Trannoy (2013), Ferreira and Peragine (2015), Pignataro (2012), and Ramos and Van de Gaer (2012).

share of total inequality that is related to the inequality due to differences in circumstances is interpreted as the IOO.

Unfortunately, this empirical strategy suffers from a drawback: it provides at best lower bounds of the true IOO. One well-recognized factor leading to this downward bias is that only a subset of all relevant circumstances are observable in available datasets (Ferreira and Gignoux, 2011). The imprecision in measurement of the outcome variable is another reason recently pointed out by Chavez-Juarez (2015).

An additional factor possibly adding to the downward bias of IOO estimates is the lack of information about a portion of the population under study. Available data are usually gathered through surveys that likely do not reach the most favored population. If this is the case, sample distributions of outcome variables miss the rightmost section of the right tail of the population distribution. Thus sample distributions of outcome variables erroneously look more homogeneous than the true population distribution. Downward bias in IOO estimates is then a consequence of the smaller variation in the sample distribution. The magnitude of the downward bias depends on how much variation is lost. Loss in variation is likely large because most favored individuals usually reach large outcome values.[5]

The possible additional bias from not observing top income populations has largely been overlooked in the IOO literature. [6] For instance, Roemer and Trannoy (2013) provide a lengthy, thorough discussion on the limitations stemming from poor quality data. They emphasize on the need of improving survey, particularly in developing countries. They point out the relevance of gathering physical information such as body mass index and psychological information such as mental health indicators and IQ measures. They also highlight how important is to measure achievements of children around the age of consent. But these recommendations aim to improve the gathering of circumstances, overlooking the possible consequences of not observing subpopulations at all.

In this context, a question with very practical implications naturally arises: is there a way to quantify the downward bias of IOO estimates? This paper answers this question by means of Monte Carlo simulations that experimentally manipulate three factors: i) not observing most favored populations; ii) not observing all relevant circumstances; and iii) interaction between not observing neither most favored population and all relevant circumstances.

[5] The issue on information about the most favored population not being available in household survey datasets has been discussed before in studies of inequality of outcomes (e.g. Korinek et al. (2006) for the case of the U.S., and Hlasny and Verme (2013) for the case of Egypt) but its implications for the estimation of IOO measures has been overlooked so far.

[6] As part of a companion project, we study the impact of an additional empirical issue –living standards across regions. So far, researchers focusing on income or wages have overlooked differences in costs of living across regions of a country. As a consequence, total inequality in income is likely to be overestimated. This is the case because high income individuals usually live in locations with higher costs of living than low income individuals. By not adjusting for differences in purchasing power, high income individuals appear richer than they actually are. Ongoing research is focusing on i) the implications of not adjusting for purchase power on IOO estimates, and ii) the impact of not observing the most favored population on IOO estimates in this context. Implications on the estimation of Growth Incidence Curves from not adjusting for purchase power has been discussed in other studies (e.g. Skoufias and Olivieri, 2013) but it is an overlooked issue in the IOO literature.

While strategies to handle the lack of information about all relevant circumstances already have been proposed (e.g. Niehues and Peichl, 2012), the interactive impact from this downward bias source and the lack of information about the most favored population has not been studied. Results from Monte Carlo simulations are straightforward: i) loss of information about the most favored population produces negligible downward bias when IOO is large (e.g. 0.978); ii) magnitude of downward bias, however, increases when true IOO decreases – e.g. representing 22% of the true value when true IOO is 0.299; iii) loss of information due to not observing a relevant circumstance may or may not be negligible –the magnitude of the downward bias depends on the variation of the outcome the circumstance can explain by itself; and iv) interaction between not observing neither most favored population and a circumstance increases the magnitude of the downward bias –i.e. there are interactive effects between the two sources of downward bias.

By compiling results from the Monte Carlo simulations carried out in this paper, we are able to observe a strong positive correlation between estimated IOO and the amount of variation in the outcome variable explained by the combination of circumstances (measured by the R^2). This association holds both for estimates reported in published studies and under a variety of Monte Carlo scenarios carried out in this paper. This result suggests that in empirical applications, the IOO estimate can be roughly (and quickly) approximated using simple econometric techniques. More importantly, this highlights the steep data requirements of an exercise such as the IOO: only to the extent that variables found in a given survey can explain a larger share of the variation of the outcome of interest, a higher IOO estimate will be estimated.

The rest of this document is organized as follows. Section 2 of this document describes estimation of inequality of opportunity. Section 3 describes the experimental designs behind our Monte Carlo and bootstrapping simulations. Section 4 presents results. Section 5 presents conclusions.

Measurement of Inequality of Opportunity

The estimation of Inequality of Opportunity (IOO) for continuous outcomes followed in this paper relies on the methodology described in Ferreira and Gignoux (2011). Borrowing Roemer's (1998) model of advantages, assume desirable economic outcomes are defined by three types of characteristics: circumstances (C), effort (E), and luck (u). Circumstances are all variables beyond an individual's control. Effort is captured by variables over which individuals have control and may also be correlated with circumstances. Luck refers to the completely randomly variables affecting economic outcomes. Thus, individuals' outcomes we observe can be written as

$$y = f(C, E, u) \qquad (1)$$

As noted by Ferreira and Gignoux (2011), equality of opportunity in Roemer's sense implies that while outcomes can vary by effort (individuals who exert more effort should be rewarded higher incomes) and luck (situations outside the control of the individual or policy), circumstances should not matter in how the outcomes are distributed. That is, equality of opportunity requires that $F(y|C) = F(y)$, where $F(.)$ is the cumulative distribution function of the outcome of interest. Measuring inequality of opportunity then is equivalent to measure the extent to which $F(y|C) \neq F(y)$. To measure this difference, this paper employs the ex-ante non parametric approach of equality of opportunity.[7]

Generally speaking, this approach consists of five steps. First, we define the outcome variable of interest in the survey. In our case, we will focus on individuals' wage earnings. Second, define the set of circumstances that are believed to be relevant to the individuals' observed outcomes. These circumstances include gender, education of the parents, region of birth, etc.[8] Third, allocate individuals into groups or "types" that result from combining circumstances across all their categories. Fourth, calculate the inequality of outcomes from a smoothed distribution (Foster and Shneyerov, 2000). For this distribution, each individual's outcome (y) is replaced by the group-specific mean for her type. Finally, for both the original distribution of incomes and the smoothed distribution calculate the following ratio:

$$\theta_r = \frac{I(\{\mu_i^k\})}{I(\{y_i\})} \tag{2}$$

Where $y_i$refers to the earnings of individual *i*, μ_i^krepresents the average outcome of individuals who belong to type *k*, and I() is an inequality index. While any inequality index could be used, it is preferable that the chosen index satisfies the following axiomatic properties: symmetry (anonymity), transfer principle, scale invariance, population replication, and additive decomposability. In turn, for any inequality index satisfying these properties, θ_r satisfies: i) the principle of population, i.e. the index is invariant to a replication of the population; ii) scale invariance, i.e. the index is invariant to the multiplication of all circumstances by a positive scalar; iii) normalization, i.e. if the smoothed distribution $\{\mu_i^k\}$ is degenerate then the index takes a value zero; and iv) within-type symmetry, i.e. the index is invariant to any permutation of two individuals within a type. Our estimate of IOO (θ_r) is bounded by 0 and 1 and can be roughly interpreted as the share of total inequality that is explained by circumstances.[9] The Inequality index we use here is the mean log deviation (MLD). MLD is defined as $I = (1/N)\sum_i \ln(\mu/y_i)$,

[7] An ex-post approach identifies individuals with the same level of effort; then estimates inequality in outcome; and finally measures how outcomes vary by types. See Ferreira and Gignoux (2009) for details. The authors also propose a parametric approach based on an OLS regression and simple functional assumptions. The non-parametric approach is implemented here.

[8] Continuous variables are broken into categories.

[9] In the numerator, all inequality within types is eliminated, and thus only inequality across circumstances groups is taken into account.

where i stands for individual i, $y_i$is individual i's outcome, and $\mu$is the overall mean of the outcome variable.[10]

The reasoning behind estimation of IOO this way assumes that the i) sample distribution of outcome variable resembles population distribution, and ii) all relevant circumstances beyond an individual's control (and used to create the types that partition the population) impact his/her chances of economic development. If this is the case, the grouping strategy creates groups of individuals facing identical circumstances. Thus, differences in outcome across these homogeneous groups reflect differences attributable to differences in circumstances.

Our strategy to explore potential sources of downward bias in IOO estimates is based on a series of Monte Carlo simulations seeking to quantify the consequences of i) not observing all relevant circumstances (i.e. circumstances that affect one's income); ii) not observing individuals in the top of the income distribution (those most favored, the top incomes); and iii) interaction from not observing neither. We describe this approach in detail next.

Studying Inequality of Opportunity in a Simulated Environment

i. Pseudo-population

The experimental design is implemented on a simulated population of 200,000 pseudo-individuals. In this population, there are five characteristics that define the individuals. We label these characteristics in the same way we would find relevant variables in a household or individual survey: gender, urban/rural setting, region of birth, father's education, and mother's education. Individuals can be either male or female, live in an urban or rural community, and have been born in one of three regions. Each individual's father's and mother's education fall into one of three possible categories: illiterate, literate, or completed primary or above. Taken together, these five circumstances create (2x2x3x3x3=) 108 mutually exclusive groups or *types* to which each pseudo-individual belongs. To keep things simple, we assign approximately the same number of individuals to each group.[11]

To define the characteristics of each individual in the pseudo-population, we apply the following rules. A pseudo-individual is female with probability 0.52. To assign individuals into other circumstances we considered an "ordering" of groups in which group 1 contains the most disadvantaged individuals and the group 108 contains the most favored individuals. Thus for example, an individual is born from an illiterate father with probability $0.0092 * (109 - group)$, where $group = 1,2, \dots ,108$. This pseudo-individual is assigned with probability $[1 - 0.0092 * (109 - group)] * (2/3)$ to a father who reads and writes, and with probability $[1 - 0.0092 * (109 - group)] * (1/3)$ to a father with elementary school or above. A similar

[10] Note that in the smoothed distribution, y_i will be replaced by μ_i^k.

[11] Each group has approximately 0.92 % of the population.

assignment is carried out for mother's education. We note that assigning a weight of 1/3 to category 3, the assignment rule aims to resemble a realistic situation in which a smaller portion of the whole population has completed elementary school in comparison to the portion that can only read and write.

Regarding the region of birth, a pseudo-individual is assigned to region 1 (the most advantageous like the capital region of a country) with probability $[1 - 0.0092 * (109 - group)] * (1/3)$, to region 2 with probability $[1 - 0.0092 * (109 - group)] * (2/3)$ and assigned to region 3 (the least advantageous) with probability $0.0092 * (109 - group)$. In this way, pseudo-individuals with larger probabilities of living in the least advantageous region also face the largest probability of having illiterate parents. Finally, assignment to an urban setting follows a reasoning similar to the one used for assignment to regions. The probability that an individual is born in a urban context is $[1 - 0.0092 * (109 - group)] * (2/3)$. This assignment implies that individuals in the first regions are most likely urban.

To obtain a general idea of the composition of our pseudo-population, Table 1 reports the percentage of pseudo-individuals by circumstances. Percentages in the main diagonal refer to the entire pseudo-population (i.e. 200,000 observations). Percentages outside the diagonal refer to the individuals with the circumstance listed in the corresponding column. For instance, the first element in the diagonal reports that 52% of pseudo-individuals are female, 50% of individuals live in an urban setting, 17% live in region 1, and so on. If we follow column 1, we find that just under 50% of females live in a urban context, or that 33.29% have a father who can read and write. To learn the percentage of pseudo-individuals whose mother and father are both illiterate, we look up such an intersection in Table 1: 66.64%. Similar calculations can be carried out if interested in learning the number of individuals with a given set of circumstances.

Using the pseudo-population distribution, we define the income generating process of individuals as follows:

$$\begin{aligned} Income = {} & 85 - 7 * female + 3 * urban \\ & +9 * region_1 - 3 * region_2 - 4 * region_3 \\ & -0 * father_{illiterate} + 1 * father_{readwrite} + 2 * father_{primary} \\ & -0 mother_{illiterate} + 3 * mother_{readwrite} + 4 * mother_{primary} \end{aligned} \tag{3}$$

According to equation (3), average income in the pseudo-population is 85 units in the omitted category. Being female is associated with lower incomes (8.2% lower). Being born in an urban context is a favoring circumstance increasing average income by 3.5 %. Region 1 is the most advantageous region, increasing income by 10.5%. In contrast, pseudo-individuals born in region 3 receive income 4.7% lower than average income, and pseudo-individuals born in region 2 receive 3.5% lower income.

In terms of parents' educational attainment, equation (3) reflects a monotonic increase in income based on increased parents' education. Resembling typical findings from the empirical literature, impacts from parents' education in equation (3) differ by parent. An empirical regularity with respect to whether father's or mother's education impacts most has not been determined but differences have been documented in case studies (see Dickson et al., 2013). In this simulation, the variable labeled mother's education is associated with higher improvements of income in comparison to the variable labeled father's education. Table 2 presents average income by circumstance across the pseudo-population. Differences across circumstances are evident and consistent with equation (3). For instance, the largest variation is observed across regions (from 80.93 in region 3 to 95.42 in region 1). This distribution of income determines our baseline scenario.

To explore how variations in total inequality impact our simulations, we also simulate an inequality enhanced scenario. The inequality enhanced scenario generates a pseudo population whose income-generating process resembles a context in which, in addition to circumstances, a second inequality inducing process is at play. The relative position of an individual at birth can further affect her income either positively or negatively. If an individual is born at a high income group, the individual's network tends to increase her income. If born in a low income environment (rural setting and the most disadvantageous region), her expected average income further decreases. This stylized description of the negative effect could be consistent with the presence of poverty traps. Particularly, geographical poverty traps. According to these theories, being born in a poverty context reinforce poverty because it permanently limits the decisions available to the poor individual (Kraay and McKenzie, 2014).

The inequality inducing adjustments are introduced to the data-generating process described in equation (3) and adding non-zero mean normally distributed error terms according to the following rule.[12] Individuals who belong to the bottom 1% experience income decreases by the absolute value of a normally distributed random draw, with mean 15 and standard deviation of 10. The income of individuals between the bottom 1% and the 10^{th} percentile is decreased by a random draw of a normal distribution with mean and standard of 5. If pseudo-individuals are between the 10^{th} and 50^{th} percentiles, their income is changed (either increased or reduced) by a normally distributed draw with mean either 5 (10^{th} to 25^{th}) or 15 (25^{th} to 50^{th}). Standard deviations of these draws are both 10. Individuals above the 50^{th} percentile increase their income. Individuals who belong to the range between the 50^{th} and 75^{th} percentiles, experience an income increase equal to the absolute value of a normal draw from a distribution with mean and standard deviation of 5. For individuals between 75^{th} and 90^{th} percentiles, the draw is taken from a normal distribution with mean 15 and standard deviation of 10; for individuals between 90^{th} and 95^{th} percentiles, the mean is 20 and the standard deviation is 10; when between 95^{th} and 99^{th}, the mean is 40 and the standard deviation is 5; and finally, individuals in the top 1%, the draw is obtained from a normal distribution with mean 60 and the standard deviation is 2. Overall, the

[12] All percentiles used are based on the original income distribution.

final distribution of incomes yields a higher inequality than under the first income-generating process.

ii. Experimental design

The experimental section of the Monte Carlo simulations uses as basis the two income-generating processes described above. For each of these, a series of scenarios are developed to study the potential downward bias found in empirical applications of the IOO estimate. Scenarios are created along three possible dimensions: the population effectively observed in the data, the number of observed circumstances, and the true share of inequality of opportunity. Three observed population scenarios are analyzed: a) the entire pseudo-population is observed, b) the top 1% of the income distribution is unobserved, and c) the top 5% of the income distribution is unobserved. Six observed circumstances scenarios are studied: all five circumstances are observed in one scenario, and in each of the remaining five we exclude one circumstance at a time.

Variation in IO share is created by adding a zero-mean, normally distributed error term to the income-generating process (baseline or inequality enhanced). By varying the dispersion of the error added, we obtain different underlying "true" IOO. An error term with standard deviation of 1 produces an IOO of 0.978 in the first data-generating process, and a IOO of 0.326 in the second data-generating process. Standard deviations of 5, 7, 10, 15, and 20 produce IOO estimates of 0.635, 0.468, 0.299, 0.156, and 0.091, respectively, in the first data-generating process. The same standard deviations produce IO shares of 0.272, 0.233, 0176, 0.109, and 0.069 in the second data-generating process.

Table 3 describes true inequality measures under each data-generating/error distribution scenario when the entire distribution is observed. Scenarios have been labeled as baseline (panel A) and inequality enhanced (panel B) to highlight the assumptions behind each data-generating process. In the first panel of Table 3, the true IOO ranges from 0.978 to 0.104, while the corresponding Gini Index ranges from 0.081 to 0.252. An increase in IOO share is associated with a decrease in Gini Index. While this may appear counterintuitive, this results follows from the mechanical increase in relevance of the random component or "luck" when adding zero-mean terms with larger standard errors. Incidentally, column (4) labeled "All" shows the R-squared from an OLS regression of income as the dependent variable and all relevant circumstance categories as regressors. The R-squared obtained when only including one circumstance at a time are also presented in Table 3 (columns [5]-[9]). These R-squared show how much variation can each variable explain by itself. Region can explain up to 0.564 of the variation when the IO ratio is 0.978. As we will see below, this feature becomes relevant when analyzing the magnitude of the downward bias.

As shown in the Table 3 panel B, addition of zero-mean errors results in larger Gini Index measures –ranging from 0.17 to 0.36. In this case, for a given distribution of luck, IOO estimates are smaller in the inequality enhanced scenario. For instance, when the error is distributed with standard deviation of 10, the true IOO is 0.299 (panel A), and only 0.176 in the inequality enhanced scenario (panel B).

Figure 1 illustrates distribution of income under each data-generating/luck scenario. Two features are worth highlighting. First, with exception of one distribution, all of them resemble realistic distributions. The exception corresponds to the baseline scenario in which the error component is distributed with unit standard deviation. In this case, the distribution is multi-modal, reflecting the fact that the random component (or one's "luck") is relatively small in this scenario and groups of individuals are clearly identifiable when considering their circumstances –R-squared under this scenario is 0.978. A second feature of Figure 1 refers to increase in variation of income under the inequality enhanced scenarios. The larger upper tails under such scenarios are evidence of the larger variation induced by our modeled "poverty traps/networking" effects. This increase in variation is what allows for larger Gini Indexes.

Taken together, the experimental design generates 108 study cases for each data-generating process.[13] The combination of observed population scenarios and observed circumstances scenarios generate a range of cases that perhaps could be thought of going from a close to ideal case to arguably more realistic ones. The ideal case corresponds to the scenario in which researchers have access to all relevant information: the entire pseudo-population is observed and all five circumstances are observed. A layer of realism can then be added by excluding one circumstance at a time. In these scenarios, researchers observe the entire pseudo-population but cannot observe all relevant circumstances. Another layer of realism is added by truncating the observed population. Thus a more realistic scenario may correspond to the case in which researchers do not observe individuals at the top of the income distribution and at least one circumstance. Finally, we also explore whether differences in the underlying (or "true") IOO affect the expected bias of IOO estimates.

Results

Monte Carlo simulations

Table 4 reports results for the baseline data-generating scenario. The table shows, in percentage terms, differences between the true IOO (which we defined implicitly based on the error distribution) and the median IOO estimate from 1,000 simulations.[14] Each panel of Table 4

[13] There are three distinct shares of "observed" data (all, truncated top 1%, and truncated top 5%), six sets of observed circumstances and 6 error distributions. These together yield a total of (3x6x6=) 108 cases.

[14] Remember that for each simulation we calculate the IOO. The median from all 1000 IOO estimates us used to calculate the difference shown in the table of results.

reports differences for each IOO level –i.e. 0.978, 0635, 0.468, 0.299, 0.156, and 0.091. The first row of each panel reports differences obtained when the entire pseudo-population is observed. The second row of each panel reports differences obtained under scenarios in which observations above 99th percentile are excluded. The third row report the results for the case when the top 5% is not observed. We also present results when the number of circumstances available to the researcher vary: column (1) shows the case when all variables are observed, while subsequent columns present results when one of the relevant circumstances (notes in the column header) is not observed in the data.

A first feature to underscore from Table 4 is that as true IOO decreases, the associated downward bias when a portion of the population is unobserved increases. This increase is monotonic for the case in which all circumstances are observed but not necessarily monotonic under scenarios in which one circumstance is not observed.

Figure 2 illustrates the pattern just described. The two lines closest to the horizontal axis illustrate how the downward bias increases when all circumstances are observed but a portion of the population is not observed. When observations above 99th percentile are not observed, downward bias is close to 0% when true IO share is 0.978 and closer to 10% when true IOO relatively low at 0.091. Under the scenario in which the top 5% is unobserved, the downward bias appears to be a very small (close to 0%) when true IOO is 0.978. However, the bias is just under 20% when the IOO is 0.47 and reaches 26% when the true IOO is 0.097. That is, missing information from the most favored population appears to be a concern at IOO levels of 0.5 and below. This result seems intuitive: if we observe all relevant circumstances, and these circumstances explain a good deal of the variation in inequality, missing 1% or even 5% of the top incomes does not create large bias. As other factors explain the variation in incomes and circumstances explain a lower share, more information is lost when we miss some portions of the population.

A second feature from Table 4 is that, the downward bias increases when there is interaction between missing information from a portion of the population and not observing a circumstance. An example of this is illustrated in Figure 2, where we plot lines referring to the scenario in which the gender characteristic is not observed. When this information is not observed but the entire population is observed, the downward bias remains practically constant at around 28% across different levels of the true IOO. However, missing 1% of the most favored population is enough to increase downward bias to 36% under the scenario in which true IOO is 0.091. Missing the top 5% incomes increases the bias to around 50% even when the IOO is at mid-level (0.47). Similar patterns can be inferred when studying the downward bias originated in missing other circumstances, as reported by Table 4. In our simulated pseudo-population, the circumstance explaining the most variation is region (see column [7] of Table 3). Accordingly, missing this circumstance produces the largest downward bias –starting at 39 when the entire population is observed and the true IOO is 0.978, and reaching a maximum of 53% when the 5% most favored population is unobserved and the IOO is lowest 0.091.

A third feature to highlight from Table 4 is that the magnitude of the downward bias seems to increase quickly when the top 5% is missing- even at true IOO "medium range" levels of 0.64 and 0.46. For instance, not observing the father's (mother's) education produces a bias of around 15% (19%) when the true IOO is 0.634. Not observing any of these variables leads to a bias of approximately 20% when the IOO is still as high as 0.468. The bias is notable as it comes from variables that explain less than 10% of the variation in income.[15]

Table 5 shows the results for the inequality enhanced scenario.[16] Focusing on the cases where circumstances explain more than 25% of the variation in income (i.e. the top two panels with true IOO higher than 0.25), we see that the biases tend to be larger than in the baseline scenario. Not observing the top 5% of the population leads to a 27% downward bias even if we observe and take into account all the relevant circumstances. Larger biases are present when we fail to observe at least one of the circumstances too. Not observing mother's education produces a 14% bias when the top 1% is missing, and over 33% when the top 5% is unobserved. The largest biases happen when the regional circumstance is not included in the calculation. Even when we observed the full population, not including the region of birth circumstance, the median IOO estimate is 40% lower than the true IOO, when the true IOO is 0.326 and 0.272. Missing the top incomes further exacerbates the problem.

Results from Monte Carlo simulations can be summarized as follows: i) loss of information about the most favored population produces negligible downward bias when IOO is very high (e.g. 0.978); ii) the magnitude of downward bias, however, increases when true IOO decreases – e.g. representing 22% of the true value when true IOO is 0.299; iii) loss of information due to not observing a relevant circumstance may or may not be negligible –the magnitude of the downward bias depends on the variation the circumstance can explain by itself and on whether the observed circumstances are correlated with the unobserved circumstance; iv) in contrast to the effect from not observing the most favored population, the magnitude of the downward bias originated in not observing a circumstance does not significantly varies across true IOO values; and iv) interaction between not observing neither the most favored population and a circumstance increases the magnitude of the downward bias.

Simulating with "real" parameters

In the previous section we made an effort to provide scenarios that resemble a real-life scenario: income that is related to circumstances of the individual at varying degrees in the context of varying levels of overall inequality. Here we take an additional step and reproduce the simulation exercise above under conditions that follow income differentials obtained from an actual dataset as a way to establish whether the previous findings would translate into under a more 'realistic'

[15] See columns (8) and (9) of Table 3.

[16] In this scenario, lower income individuals had their incomes reduced randomly while higher income individuals experienced a random increase in their income. Under this scenario we have higher inequality (as measured by the Gini index), but individuals' circumstances explain on average a lower share of overall inequality.

context. We carry out Monte Carlo simulations following a similar reasoning as in the previous section, but the baseline data-generating process follows equation (4).

$$\begin{aligned} LN(Income) = {} & 5 - 0.35 * female + 0.15 * urban \\ & +2 * region_1 - 0.10 * region_2 - 0.17 * region_3 \\ & -0 * father_{illiterate} + 0.05 * father_{readwrite} + 0.12 * father_{primary} \\ & -0 mother_{illiterate} + 0.13 * mother_{readwrite} + 0.18 * mother_{primary} \end{aligned} \quad (4)$$

where the coefficients in equation (4) closely resemble those obtained from fitting an OLS on natural log of wages observed in the 2012 Egypt Labor Force Survey. Assuming these are all the relevant circumstances, we can add a normally distributed error term to equation (4) and determine what would be the IOO in this hypothetical country. As before, we construct a baseline scenario following equation (4) and an inequality enhanced scenario where lower income individuals are randomly assigned an even lower income and higher individuals can get even higher incomes. Table 6 reports the parameterization of each error term. When adding an error term with standard deviation of 0.1, the corresponding IOO is 0.984. When adding error terms with standard deviation of 0.3, 0.7, 1.0, 1.2, and 1.5, the corresponding IOO are 0.869, 0.543, 0.358, 0.270, and 0.183, respectively. Table 7 and Table 8 report the magnitude of the downward bias under the baseline and inequality enhanced scenarios.

Using as guidance the most recent estimates of the Gini index for Egypt (around 0.30), we focus on the cases where the true IOO is 0.27 in the baseline scenario (next to last panel in Table 7) and 0.267 in the inequality enhanced scenario (next to last panel Table 8). If we believe that such values are reasonably accurate for the Egyptian case, Table 7 and Table 8 indicate that we may still find ourselves greatly underestimating IOO when we miss the top 1% of the distribution from around 9% in the case we have all the relevant variables, to a substantial 80% when the region of birth is not taken into account.[17] That is, if the true IOO is around 0.30, not controlling for region as a relevant circumstance would lead to and (under-)estimated IOO of 0.06.

Robustness checks

Two additional sets of Monte Carlo simulations are carried out to check the robustness of our results. So far, in both the baseline and inequality-enhanced scenarios IOO decreases as Gini increases (see Table 3 and Table 6). This is a mechanical consequence resulting from increasing total inequality through an increase in the variation explained by the error term. Instead, total inequality can be increased by increasing the variation due to a specific variable in equations (3) and (4). In this way, in contrast to what happens in the scenarios analyzed before, IOO increases at the same time as Gini increases. The first set of additional Monte Carlo simulations studies downward bias of IOO estimates by varying the coefficient of the dichotomous variable urban

[17] If we run a simple OLS regression of income on a series regional dummies, the R^2 is about 80%.

(u) among five values –0, 10, 15, 25, 35, and 50. We label these scenarios as *increase-in-IOO* scenarios. Table 9 describes the data-generating process and corresponding inequality measures. All six scenarios include a normally distributed error term with mean zero and standard deviation of 10. Figure 3 illustrates the distribution of simulated income under the studied scenarios. Table 10 reports the downward bias under scenarios excluding top incomes and circumstances. Results hold similar to those reported in tables 4, 5, 7 and 8.

A second set of simulations modifies the assumption that the same number of pseudo-individuals fall in each type. Instead, individuals are allocated to types according to a normal distributions that assigns a larger number of individuals to middle-income types and relatively small number of individuals to low-income and high-income types. IOO estimates are carried out under baseline, enhanced and increase-in-IOO scenarios. Table 11 reports the data-generating processes with their corresponding inequality measures. Table 12, Table 13, and Table 14 report the downward bias under scenarios excluding top incomes and circumstances. Results hold similar with one nuance: impacts from not observing most favored population is larger in comparison to the case in which individuals are allocated uniformly across types.

The relationship between IOO and R^2

As noted above, one of the main takeaways of the Monte Carlo exercise is that there is a positive correlation between the amount of variation explained by a certain circumstance and the bias that an empirical IOO estimate would suffer with respect to the true value of IOO. A natural question that arises is then, how would the combination of the variation of circumstances observed in the data be correlated with the IOO estimate that we can expect to obtain. To explore this question, we compile information from this study as well as previous studies that have calculated the IOO in other countries.

Ferreira and Gignoux (2011) produced a series of IOO estimates for countries in Latin America while focusing on two related outcomes: labor earnings and household income. In turn, World Bank (2015) produced IOO estimates for labor earnings for a few countries in the MENA region. Finally, we use the data generated in our Monte Carlo Simulations and show how the IOO estimate varies with the R^2 of a simple OLS regression of the outcome of interest on all the circumstances used in the estimation of IOO.

Figure 4 presents the results of this compilation and shows a clear pattern. There is a strong positive correlation between the IOO estimated and the amount of variation of earnings (wages, or household income) explained by the combination of circumstances (measured by the R^2): plotting the IOO estimates and corresponding R^2 yield almost all the data points along the 45 degree line. This result could imply that in empirical applications, the IOO estimate can be roughly (and quickly) approximated using simple econometric techniques. More importantly, this highlights the steep data requirements of an exercise such as the IOO: only to the extent that variables found in a given survey can explain a larger share of the variation of the outcome of interest, a higher IOO estimate will be reached.

Conclusions

Results from Monte Carlo simulations carried out in this paper are straightforward: i) loss of information about the most favored population produces negligible downward bias when IOO is large (e.g. 0.978); ii) the magnitude of downward bias, however, increases when true IOO decreases – e.g. representing 22% of the true value when true IOO is 0.299; iii) loss of information due to not observing a relevant circumstance may or may not be negligible –the magnitude of the downward bias depends on the variation the circumstance can explain by itself; iv) in contrast to the effect from not observing the most favored population, the magnitude of the downward bias originated in not observing a circumstance does not significantly varies across true IOO values; and v) interaction between not observing the most favored population and not observing a circumstance increases the magnitude of the downward bias –i.e. there are interactive effects between the two sources of downward bias. These results imply that strategies proposed to handle the lack of information about all relevant circumstances (e.g. Niehues and Peichl, 2012) may not fully take into account the downward bias originated in the lack of information about top income populations.

Importantly, when pseudo-individuals are not allocated uniformly across types, loss of information about the most favored population has a larger impact than in the scenario in which they are allocated uniformly. This is relevant because a non-uniform distribution is more likely observed in real world applications.

As a by-product of the Monte Carlo simulations carried out in this paper, we are able to observe a strong positive correlation between estimated IOO and the amount of variation in the outcome variable explained by the combination of circumstances (measured by the R^2). This association holds both for estimates reported in published studies and under a variety of Monte Carlo scenarios carried out in this paper. This result suggests that in empirical applications, the IOO estimate can be roughly (and quickly) approximated using simple econometric techniques. More importantly, this highlights the steep data requirements of an exercise such as the IOO: only to the extent that variables found in a given survey can explain a larger share of the variation of the outcome of interest, a higher IOO estimate will be estimated.

References

Bourguignon, François, Francisco HG Ferreira, and Marta Menendez. "Inequality of opportunity in Brazil." Review of Income and Wealth 53, no. 4 (2007): 585-618.

Chen, Alice, Emily Oster, and Heidi Williams. “Why is infant mortality higher in the US than in Europe” (2014). NBER Working paper 20525. Available at http://www.nber.org/papers/w20525.

Dickson, Matthew, Paul Gregg, and Harriet Robinson. "Early, late or never? When does parental education impact child outcomes?." IZA Discussion Paper No. 7123 (2013). Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2203273

Dworkin, Ronald. "What is equality? Part 1: Equality of welfare." Philosophy & Public Affairs (1981): 185-246.

Dworkin, Ronald. "What is equality? Part 2: Equality of resources." Philosophy & Public Affairs (1981): 283-345.

Elbers, Chris and Lanjouw, Peter and Mistiaen, Johan A. and Ozler, Berk. “Re-interpreting Sub-group Inequality Decompositions.” World Bank Policy Research Working Paper No. 3687 (2005). Available at SSRN:http://ssrn.com/abstract=786626

Ferreira, Francisco HG, and Jérémie Gignoux. "The measurement of inequality of opportunity: Theory and an application to Latin America." Review of Income and Wealth 57, no. 4 (2011): 622-657.

Ferreira, Francisco HG, and Vito Peragine. “Equality of Opportunity: Theory and evidence.” World Bank Policy Research Working Paper No. 7217 (2015). Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2584375.

Friedman, Milton, and Leonard J. Savage. "The utility analysis of choices involving risk." The journal of political economy (1948): 279-304.

Friedman, Milton. "Choice, chance, and the personal distribution of income." The Journal of Political Economy (1953): 277-290.

Hlasny, Vladimir, and Paolo Verme. “Top incomes and the measurement of inequality in Egypt.” World Bank Policy Research Working Paper No. 6557. Available at http://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-6557.

Kanbur, Ravi, “How Useful is Inequality of Opportunity as a Policy Construct?” (July 1, 2014). World Bank Policy Research Working Paper No. 6980. Available at SSRN: http://ssrn.com/abstract=2475067

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion. "Survey nonresponse and the distribution of income." The Journal of Economic Inequality 4, no. 1 (2006): 33-55.

Kraay, Aart, and David McKenzie. “Do poverty traps exist? Assessing the evidence.” Journal of Economic Perspectives 28 (2014): 127-148.

Marrero, Gustavo A., and Juan G. Rodríguez. "Inequality of opportunity and growth." Journal of Development Economics 104 (2013): 107-122.

Niehues, Judith and Peichl, Andreas, Bounds of Unfair Inequality of Opportunity: Theory and Evidence for Germany and the US (May 15, 2012). CESifo Working Paper Series No. 3815. Available at SSRN:http://ssrn.com/abstract=2060014

OXFAM. “Even it up: Time to end extreme inequality”. (2014). Available at http://policy-practice.oxfam.org.uk/publications/even-it-up-time-to-end-extreme-inequality-333012.

Paes de Barros, Ricardo, Francisco H.G. Ferreira, José R. Molinas Vega, and Jaime Saavedra Chanduvi. “Measuring Inequality of Opportunities in Latin America and the Caribbean”. World Bank (2009).

Peragine, Vito. “Ranking Income Distributions According to Equality of Opportunity,” Journal of Economic Inequality 2 (2004): 11–30.

Pignataro, Giuseppe. "Equality of opportunity: Policy and measurement paradigms." Journal of Economic Surveys 26, no. 5 (2012): 800-834.

Piketty, Thomas. "Capital in the 21st Century." Cambridge: Harvard Uni (2014).

Roemer, John E., and Alain Trannoy. "Equality of opportunity." (2013). Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2345357.

Ramos, Xavier and Van de Gaer, Dirk, “Empirical Approaches to Inequality of Opportunity: Principles, Measures, and Evidence”. IZA Discussion Paper No. 6672 (2012). Available at SSRN: http://ssrn.com/abstract=2096802

Roemer, John E. Equality of opportunity. Harvard University Press (1998).

Sailesh Tiwari, Gabriel Lara Ibarra, and Ambar Narayan (2015). “How unfair is the inequality of wage earnings in Russia? Estimates from panel data”, Policy Research Working Paper; no. WPS 7291. Washington, D.C. : World Bank Group.

Sen, Amartya. Development as freedom. Oxford University Press (1999).

Skoufias, Emmanuel, and Sergio Olivieri. “Inequality and the distribution of gains from growth in Thailand between 2000 and 2009” (2013), mimeo.

Velez, Carlos E. and Al-Shawarby, Sherine and El-laithy, Heba, Equality of Opportunity for Children in Egypt, 2000-2009: Achievements and Challenges (August 1, 2012). World Bank Policy Research Working Paper No. 6159. Available at SSRN: http://ssrn.com/abstract=2127059

Wendelspiess Chávez Juárez, Florian. "Measuring Inequality of Opportunity with Latent Variables." Journal of Human Development and Capabilities (2015): 1-14

Table 1. Percentage of pseudo-individuals by circumstance

	Female	Urban	Region 1	Region 2	Region 3	Father is illiterate	Father reads and writes	Father completed elementary	Mother is illiterate	Mother reads and writes	Mother completed elementary
Female	52										
Urban	49.90	50									
Region 1	16.59	21.89	17								
Region 2	33.39	44.22	0	33							
Region 3	50.09	33.53	0	0	50						
Father is Illiterate	50.15	33.66	33.04	34.15	66.71	50					
Father reads and writes	33.29	43.92	42.95	44.48	22.23	0	33				
Father completed elementary school	16.62	22.07	21.22	22.29	11.39	0	0	17			
Mother is Illiterate	49.96	33.25	32.65	33.69	66.66	66.64	33.41	33.25	50		
Mother reads and writes	33.33	43.99	42.70	44.81	22.40	22.47	44.56	42.97	0	33	
Mother completed elementary school	16.78	22.40	21.85	22.43	11.28	11.38	50.78	21.78	0	0	17

Notes: Percentages in the diagonal refer to the entire pseudo-population (200,000). Non-diagonal percentages refer to the population with the circumstance listed in the column.

Table 2. Average income by circumstance across observed population scenarios

	Male	Female	Rural	Urban	Region1	Region2	Region3	Father is illiterate	Father is literate	Father completed primary	Mother is illiterate	Mother is literate	Mother completed primary
Average	87.79	80.82	81.25	87.09	95.42	83.44	80.93	81.83	86.19	87.15	81.13	86.83	87.9
Std. Dev	5.79	5.79	5.64	6.52	4.42	4.41	4.38	5.91	6.74	6.73	5.57	6.44	6.48
Sample	95,931	104,069	100,356	99,644	33,051	66,614	100,335	100,487	66,193	33,320	99,994	66,488	33,518

Notes: Descriptive statistics refers to the 200 thousand pseudo-individuals generated according to equation (3).

Table 3. Data-generating processes and inequality measures

Error Distribution	Inequality measures			R-squared by observed circumstance					
	Gini	Total MLD	IOO	All	Female	Urban	Region	Father's education	Mother's education
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Baseline scenario									
N(0,1)	0.081	0.003	0.978	0.979	0.260	0.183	0.564	0.120	0.199
N(0,5)	0.100	0.005	0.635	0.647	0.172	0.122	0.372	0.080	0.132
N(0,7)	0.116	0.007	0.468	0.483	0.128	0.091	0.279	0.058	0.099
N(0,10)	0.144	0.011	0.299	0.314	0.086	0.058	0.180	0.038	0.064
N(0,15)	0.196	0.020	0.156	0.170	0.045	0.031	0.099	0.019	0.034
N(0,20)	0.252	0.035	0.091	0.104	0.027	0.019	0.059	0.012	0.020
Panel B. Inequality Enhanced scenario									
N(0,1)	0.165	0.014	0.326	0.354	0.094	0.067	0.206	0.045	0.074
N(0,5)	0.180	0.017	0.272	0.298	0.080	0.056	0.172	0.037	0.062
N(0,7)	0.194	0.020	0.233	0.258	0.068	0.049	0.150	0.032	0.055
N(0,10)	0.219	0.026	0.176	0.199	0.053	0.039	0.116	0.025	0.043
N(0,15)	0.274	0.043	0.109	0.130	0.034	0.024	0.075	0.016	0.027
N(0,20)	0.335	0.069	0.069	0.087	0.021	0.017	0.051	0.012	0.018

Notes: Data generating process is presented in equation (3). Error terms listed in the first column are randomly added to equation (3) to induce variation in incomes across individuals. R-squared is obtained from an OLS regression using income as the dependent variable and the specified circumstance as regressors. MLD is the mean log deviation, IOO is the ratio of the smoothed distribution MLD and total MLD. All estimates are based on observing the full income distribution (i.e. no truncation).

Table 4. Difference between true IOO and median estimated IOO in percentage terms: Baseline scenario

Observed population	Excluded Circumstances					
	None (1)	Gender (2)	Urban (3)	Region (4)	Father's Education (5)	Mother's education (6)
			True IO share = 0.978			
All	0.00	-27.88	-4.14	-39.82	-1.07	-5.35
Top 1% truncated	-0.14	-29.65	-4.56	-40.91	-1.27	-5.82
Top 5% truncated	-0.82	-37.14	-6.41	-42.76	-2.25	-8.02
			True IO share = 0.635			
All	0.00	-27.99	-4.29	-39.83	-1.18	-5.46
Top 1% truncated	-3.25	-32.14	-7.47	-41.66	-4.34	-8.76
Top 5% truncated	-12.81	-45.35	-17.71	-45.07	-14.05	-19.20
			True IO share = 0.468			
All	0.00	-27.95	-4.12	-39.77	-1.10	-5.32
Top 1% truncated	-5.03	-33.58	-9.35	-42.03	-6.24	-10.55
Top 5% truncated	-18.01	-48.01	-22.50	-47.24	-19.14	-23.76
			True IO share = 0.299			
All	0.00	-28.03	-4.36	-40.01	-1.42	-5.53
Top 1% truncated	-7.15	-32.14	-7.47	-41.66	-4.34	-8.76
Top 5% truncated	-22.55	-45.35	-17.71	-45.07	-14.05	-19.20
			True IO share = 0.156			
All	0.00	-28.44	-4.76	-39.99	-1.80	-6.09
Top 1% truncated	-9.28	-36.05	-13.22	-44.20	-10.34	-14.63
Top 5% truncated	-25.45	-49.94	-29.18	-52.60	-26.27	-30.33
			True IO share = 0.091			
All	0.00	-28.79	-5.56	-40.51	-2.26	-6.88
Top 1% truncated	-10.42	-36.77	-14.21	-45.31	-11.36	-15.73
Top 5% truncated	-26.63	-49.83	-30.21	-53.74	-28.26	-31.25

Notes: Results based on 1,000 simulations. Baseline scenario refers to data-generating process described by equation (3)

Table 5. Difference between true IOO and median estimated IOO in percentage terms: Inequality enhanced scenario

Observed population	Excluded Circumstances					
	None (1)	Gender (2)	Urban (3)	Region (4)	Father's Education (5)	Mother's education (6)
			True IO share = 0.326			
All	0.00	-28.28	-4.59	-39.69	-1.28	-5.69
Top 1% truncated	-8.10	-36.19	-12.59	-44.08	-9.18	-13.65
Top 5% truncated	-27.73	-56.20	-32.14	-51.21	-28.73	-33.20
			True IO share = 0.272			
All	0.00	-28.16	-4.54	-39.58	-1.17	-5.64
Top 1% truncated	-8.84	-36.80	-13.13	-43.72	-9.85	-14.20
Top 5% truncated	-28.10	-54.67	-32.22	-52.36	-28.99	-33.31
			True IO share = 0.232			
All	0.00	-28.15	-4.44	-39.50	-1.02	-5.43
Top 1% truncated	-9.24	-36.91	-13.54	-43.67	-10.19	-14.63
Top 5% truncated	-27.90	-53.61	-31.90	-52.76	-29.00	-32.95
			True IO share = 0.176			
All	0.00	-27.91	-4.02	-39.33	-0.84	-5.01
Top 1% truncated	-9.06	-36.55	-13.40	-43.74	-10.21	-14.52
Top 5% truncated	-27.28	-52.02	-31.22	-52.92	-28.42	-32.17
			True IO share = 0.109			
All	0.00	-29.13	-5.47	-40.05	-2.27	-6.77
Top 1% truncated	-10.68	-36.91	-14.78	-45.03	-11.75	-16.17
Top 5% truncated	-27.99	-51.22	-31.81	-54.39	-28.82	-32.75
			True IO share = 0.069			
All	0.00	-28.55	-5.26	-39.86	-1.53	-6.21
Top 1% truncated	-10.13	-36.80	-14.30	-45.03	-11.23	-15.82
Top 5% truncated	-27.22	-49.95	-30.50	-54.19	-28.65	-31.51

Notes: Results based on 1,000 simulations. Inequality enhanced scenario refers to data-generating process described by equation (3), modified as explained in section describing the pseudo-population under analysis.

Table 6. Data-generating processes and inequality measures, using coefficients resembling 2012 Egypt Labor Force Survey

	Inequality measures			R-squared by observed circumstance					
Error Distribution	Gini (1)	Total MLD (2)	IO Ratio (3)	All (4)	Female (5)	Urban (6)	Region (7)	Father's education (8)	Mother's education (9)
Panel A. Baseline Scenario									
N(0,0.1)	0.165	0.012	0.984	0.987	0.041	0.075	0.923	0.060	0.075
N(0,0.3)	0.174	0.013	0.869	0.890	0.037	0.068	0.833	0.054	0.068
N(0,0.7)	0.212	0.022	0.543	0.600	0.024	0.045	0.562	0.036	0.045
N(0,1.0)	0.252	0.033	0.358	0.422	0.018	0.033	0.393	0.026	0.033
N(0,1.2)	0.282	0.043	0.270	0.336	0.015	0.025	0.314	0.020	0.026
N(0,1.5)	0.332	0.064	0.183	0.246	0.011	0.019	0.230	0.015	0.018
Panel B. Inequality Enhanced Scenario									
N(0,0.1)	0.170	0.013	0.973	0.978	0.041	0.074	0.915	0.059	0.075
N(0,0.3)	0.178	0.014	0.860	0.884	0.037	0.067	0.826	0.053	0.067
N(0,0.7)	0.217	0.023	0.535	0.594	0.024	0.045	0.558	0.036	0.046
N(0,1.0)	0.259	0.035	0.353	0.420	0.017	0.032	0.392	0.025	0.032
N(0,1.2)	0.290	0.046	0.267	0.335	0.014	0.026	0.315	0.021	0.025
N(0,1.5)	0.338	0.067	0.179	0.243	0.010	0.018	0.230	0.015	0.019

Notes: Data generating process is presented in equation (4). Error terms listed in the first column are randomly added to equation (4) to induce variation in incomes across individuals. R-squared is obtained from an OLS regression using ln(income) as dependent variable and circumstances included in equation (4) as explanatory variables. MLD is the mean log deviation, IOO is the ratio of the smoothed distribution MLD and total MLD. All estimates are based on observing the full income distribution (i.e. no truncation).

Table 7. Difference between true IOO and median estimated IOO in percentage terms, scenario following parameters from 2012 Egypt Labor Force Survey

Observed population	Excluded circumstances					
	None (1)	Gender (2)	Urban (3)	Region (4)	Father's education (5)	Mother's education (6)
True IO share = 0.984						
All	0.00	-5.10	-0.76	-80.15	-0.27	-0.77
Top 1% truncated	-0.10	-5.47	-0.90	-82.07	-0.38	-0.91
Top 5% truncated	-0.64	-7.49	-1.68	-86.10	-1.00	-1.69
True IO share = 0.869						
All	0.00	-5.12	-0.79	-80.15	-0.29	-0.77
Top 1% truncated	-0.90	-6.22	-1.68	-81.74	-1.17	-1.69
Top 5% truncated	-5.41	-12.16	-6.42	-85.12	-5.75	-6.44
True IO share = 0.543						
All	0.00	-5.42	-1.07	-80.17	-0.59	-1.07
Top 1% truncated	-4.84	-10.16	-5.67	-81.54	-5.21	-5.68
Top 5% truncated	-22.99	-29.34	-23.91	-84.33	-23.37	-23.87
True IO share = 0.358						
All	0.00	-5.03	-0.78	-80.14	-0.36	-0.79
Top 1% truncated	-7.77	-6.22	-1.68	-81.74	-1.17	-1.69
Top 5% truncated	-32.91	-12.16	-6.42	-85.12	-5.75	-6.44
True IO share = 0.270						
All	0.00	-4.80	-0.45	-80.04	0.05	-0.54
Top 1% truncated	-9.29	-14.74	-10.42	-81.74	-9.57	-10.14
Top 5% truncated	-35.88	-41.14	-36.65	-85.46	-36.22	-36.75
True IO share = 0.183						
All	0.00	-4.85	-0.38	-80.09	0.00	-0.46
Top 1% truncated	-11.73	-16.73	-12.06	-81.94	-11.77	-12.40
Top 5% truncated	-37.34	-42.30	-38.15	-85.92	-37.53	-38.23

Notes: Results based on 1,000 simulations. Baseline scenario refers to data-generating process described by equation (4).

Table 8. Difference between true IOO and median estimated IOO in percentage terms, inequality enhanced scenario and following parameters from 2012 Egypt Labor Force Survey

Observed population	Excluded circumstance					
	None (1)	Gender (2)	Urban (3)	Region (4)	Father's education (5)	Mother's education (6)
			True IO share = 0.973			
All	-0.01	-5.12	-0.78	-80.11	-0.28	-0.78
Top 1% truncated	-0.17	-5.55	-0.98	-82.04	-0.45	-0.98
Top 5% truncated	-1.05	-7.90	-2.10	-86.11	-1.41	-2.11
			True IO share = 0.860			
All	0.00	-5.18	-0.84	-80.11	-0.32	-0.82
Top 1% truncated	-1.01	-6.36	-1.80	-81.69	-1.28	-1.81
Top 5% truncated	-5.84	-12.60	-6.87	-85.03	-6.18	-6.88
			True IO share = 0.535			
All	0.00	-4.88	-0.50	-80.01	0.02	-0.48
Top 1% truncated	-4.32	-9.71	-5.19	-81.40	-4.70	-5.18
Top 5% truncated	-22.81	-29.17	-23.74	-84.22	-23.15	-23.68
			True IO share = 0.353			
All	0.00	-4.94	-0.63	-80.05	-0.20	-0.63
Top 1% truncated	-7.66	-6.36	-1.80	-81.69	-1.28	-1.81
Top 5% truncated	-32.88	-12.60	-6.87	-85.03	-6.18	-6.88
			True IO share = 0.267			
All	0.00	-4.81	-0.42	-79.97	0.00	-0.52
Top 1% truncated	-9.49	-14.73	-10.29	-81.76	-9.84	-10.45
Top 5% truncated	-35.83	-41.28	-36.75	-85.54	-36.17	-36.79
			True IO share = 0.179			
All	0.00	-4.22	0.34	-79.81	0.88	0.24
Top 1% truncated	-10.71	-16.04	-11.57	-81.87	-11.02	-11.69
Top 5% truncated	-36.96	-41.88	-37.51	-85.75	-37.54	-37.57

Notes: Results based on 1,000 simulations. Inequality enhanced scenario refers to data-generating process described by equation (4), modified as explained in section describing simulation with "real" parameters.

Table 9. Data-generating processes and inequality measures under increase-in-IOO scenarios

Coefficient in Urban indicator	Inequality measures			R-squared by observed circumstance					
	Gini (1)	MLD (2)	IOO (3)	All (4)	Female (5)	Urban (6)	Region (7)	Father's education (8)	Mother's education (9)
0	0.143	0.010	0.271	0.285	0.089	0.014	0.172	0.025	0.046
10	0.153	0.012	0.422	0.438	0.068	0.230	0.197	0.069	0.098
15	0.163	0.014	0.523	0.540	0.055	0.367	0.201	0.088	0.118
25	0.191	0.019	0.680	0.699	0.037	0.586	0.191	0.108	0.136
35	0.222	0.026	0.779	0.799	0.024	0.724	0.179	0.117	0.139
50	0.268	0.038	0.860	0.880	0.014	0.835	0.164	0.121	0.140

Notes: Data generating process is presented in equation (3). All six scenarios include a normally distributed error term with mean zero and standard deviation of 10. R-squared is obtained from an OLS regression using income as the dependent variable and the specified circumstance as regressors. MLD is the mean log deviation, IOO is the ratio of the smoothed distribution MLD and total MLD. All estimates are based on observing the full income distribution (i.e. no truncation).

Table 10. Difference between true IOO and median estimated IOO in percentage terms: increase-in-IOO scenarios

Observed population	Excluded circumstance					
	None (1)	Female (2)	Urban (3)	Region (4)	Father's education (5)	Mother's education (6)
			True IO share = 0.27			
all	0.00	-33.10	-1.05	-47.31	-2.31	-7.23
Top 1% truncated	-8.85	-40.75	-8.90	-50.23	-10.14	-15.00
Top 5% truncated	-25.18	-55.82	-25.24	-56.52	-26.46	-31.16
			True IO share = 0.42			
all	0.00	-15.92	-26.14	-22.40	-0.30	-2.76
Top 1% truncated	-3.88	-20.23	-30.89	-24.77	-4.56	-7.07
Top 5% truncated	-14.34	-30.75	-41.44	-31.30	-14.99	-17.55
			True IO share = 0.52			
all	0.00	-11.17	-39.82	-15.32	-0.70	-2.35
Top 1% truncated	-2.88	-13.84	-43.36	-16.79	-3.34	-5.02
Top 5% truncated	-10.00	-21.13	-51.31	-21.85	-10.47	-12.18
			True IO share = 0.68			
all	0.00	-5.88	-55.38	-7.83	-0.48	-1.33
Top 1% truncated	-1.23	-20.23	-30.89	-24.77	-4.56	-7.07
Top 5% truncated	-4.39	-30.75	-41.44	-31.30	-14.99	-17.55
			True IO share = 0.77			
all	0.00	-3.56	-63.19	-4.63	-0.28	-0.80
Top 1% truncated	-0.56	-3.95	-64.93	-4.72	-0.69	-1.21
Top 5% truncated	-2.08	-5.51	-68.96	-5.82	-2.21	-2.74
			True IO share = 0.86			
all	0.00	-1.98	-69.19	-2.51	-0.11	-0.40
Top 1% truncated	-0.17	-2.11	-70.47	-2.49	-0.25	-0.55
Top 5% truncated	-0.82	-2.77	-73.52	-2.92	-0.90	-1.19

Notes: Results based on 1,000 simulations. Increase-in-IOO scenarios refer to data-generating process described by equation (3), modified as described in robustness checks section.

Table 11. Data-generating processes and inequality measures under scenarios in which pseudo-individuals are allocated to each type according to a normal distribution

	Inequality measures			R-squared by observed circumstance					
Error Distribution	Gini (1)	MLD (2)	IO Ratio (3)	All (4)	Female (5)	Urban (6)	Region (7)	Father's Education (8)	Mother's education (9)
				Panel A. Baseline Scenario					
N(0,0.1)	0.076	0.003	0.974	0.975	0.302	0.074	0.522	0.026	0.094
N(0,0.3)	0.095	0.005	0.602	0.614	0.189	0.047	0.329	0.016	0.059
N(0,0.7)	0.111	0.006	0.431	0.444	0.137	0.034	0.238	0.013	0.042
N(0,1.0)	0.140	0.010	0.270	0.283	0.088	0.021	0.152	0.008	0.026
N(0,1.2)	0.193	0.020	0.136	0.148	0.044	0.011	0.081	0.004	0.014
N(0,1.5)	0.249	0.035	0.080	0.091	0.028	0.007	0.049	0.003	0.009
				Panel B. Inequality Enhanced Scenario					
N(0,0.1)	0.198	0.020	0.786	0.665	0.155	0.083	0.334	0.031	0.105
N(0,0.3)	0.206	0.022	0.730	0.623	0.145	0.078	0.313	0.029	0.098
N(0,0.7)	0.212	0.023	0.678	0.584	0.136	0.072	0.294	0.027	0.093
N(0,1.0)	0.225	0.027	0.593	0.518	0.121	0.064	0.260	0.024	0.082
N(0,1.2)	0.255	0.036	0.446	0.405	0.094	0.050	0.203	0.019	0.065
N(0,1.5)	0.290	0.048	0.326	0.309	0.071	0.037	0.156	0.015	0.049
Urban Indicator's coefficient				*Panel C. Increase-in-IOO Scenario*					
0	0.141	0.010	0.256	0.269	0.089	0.000	0.155	0.006	0.026
10	0.146	0.011	0.378	0.391	0.075	0.167	0.135	0.008	0.027
15	0.155	0.012	0.479	0.493	0.064	0.305	0.117	0.010	0.026
25	0.181	0.017	0.651	0.668	0.041	0.545	0.081	0.009	0.022
35	0.211	0.023	0.761	0.778	0.027	0.696	0.057	0.009	0.018
50	0.256	0.035	0.851	0.870	0.016	0.823	0.037	0.007	0.014

Notes: Data generating process is presented in equation (3), and modified as described in robustness checks section.

Table 12. Difference between true IOO and median estimated IOO in percentage terms: Baseline scenario

Observed population scenarios	Excluded circumstance					
	None (1)	Female (2)	Urban (3)	Region (4)	Father's education (5)	Mother's education (6)
			True IO share = 0.97			
all	0.00	-31.99	-5.87	-49.55	-1.46	-7.52
Top 1% truncated	-0.18	-34.18	-6.42	-50.97	-1.72	-8.17
Top 5% truncated	-0.92	-42.50	-8.78	-52.50	-2.88	-10.96
			True IO share = 0.60			
all	0.00	-32.01	-5.92	-49.58	-1.52	-7.58
Top 1% truncated	-3.56	-36.71	-9.67	-51.28	-5.08	-11.37
Top 5% truncated	-13.76	-50.54	-20.58	-53.64	-15.45	-22.45
			True IO share = 0.43			
all	0.00	-31.46	-5.19	-49.17	-0.77	-6.88
Top 1% truncated	-4.77	-37.61	-10.83	-51.22	-6.30	-12.53
Top 5% truncated	-18.15	-52.07	-24.44	-54.95	-19.75	-26.21
			True IO share = 0.27			
all	0.00	-31.74	-5.54	-49.38	-1.15	-7.23
Top 1% truncated	-7.09	-36.71	-9.67	-51.28	-5.08	-11.37
Top 5% truncated	-22.50	-50.54	-20.58	-53.64	-15.45	-22.45
			True IO share = 0.14			
all	0.00	-31.03	-4.58	-48.86	-0.16	-6.33
Top 1% truncated	-7.39	-38.65	-13.30	-52.08	-9.08	-14.96
Top 5% truncated	-23.92	-51.92	-29.22	-58.68	-25.37	-30.70
			True IO share = 0.08			
all	0.00	-32.88	-7.19	-50.26	-2.94	-8.86
Top 1% truncated	-10.27	-40.25	-16.04	-53.76	-12.03	-17.59
Top 5% truncated	-26.46	-52.55	-31.47	-60.63	-28.00	-32.96

Notes: Results based on 1,000 simulations. Baseline scenario refers to data-generating process described by equation (3), modified as described in robustness checks section.

Table 13. Difference between true IOO and median estimated IOO in percentage terms: Enhanced inequality scenario

	Excluded circumstance					
Observed population scenarios	**None (1)**	**Female (2)**	**Urban (3)**	**Region (4)**	**Father's education (5)**	**Mother's education (6)**
			True IO share = 0.78			
all	0.00	-39.39	-20.56	-40.60	-6.61	-23.73
Top 1% truncated	-2.51	-43.63	-24.69	-41.52	-9.41	-28.00
Top 5% truncated	-9.66	-55.95	-36.58	-36.95	-17.47	-40.10
			True IO share = 0.73			
all	0.00	-39.47	-20.66	-40.68	-6.73	-23.83
Top 1% truncated	-3.29	-44.12	-25.31	-41.89	-10.14	-28.59
Top 5% truncated	-11.72	-56.96	-37.96	-38.58	-19.34	-41.37
			True IO share = 0.67			
all	0.00	-39.28	-20.43	-40.50	-6.46	-23.60
Top 1% truncated	-3.59	-44.38	-25.57	-41.95	-10.42	-28.84
Top 5% truncated	-13.08	-57.60	-38.81	-39.66	-20.54	-42.16
			True IO share = 0.59			
all	0.00	-39.47	-20.66	-40.68	-6.75	-23.84
Top 1% truncated	-4.96	-44.12	-25.31	-41.89	-10.14	-28.59
Top 5% truncated	-16.00	-56.96	-37.96	-38.58	-19.34	-41.37
			True IO share = 0.44			
all	0.00	-39.38	-20.57	-40.62	-6.64	-23.73
Top 1% truncated	-6.80	-46.46	-28.21	-43.05	-13.40	-31.32
Top 5% truncated	-19.81	-60.07	-42.70	-45.07	-26.42	-45.60
			True IO share = 0.32			
all	0.00	-39.33	-20.46	-40.55	-6.54	-23.66
Top 1% truncated	-8.28	-47.31	-29.38	-43.57	-14.81	-32.41
Top 5% truncated	-22.25	-60.33	-43.71	-47.25	-28.51	-46.41

Notes: Results based on 1,000 simulations. Enhanced inequality scenario refers to data-generating process described by equation (3), modified as described in robustness checks section.

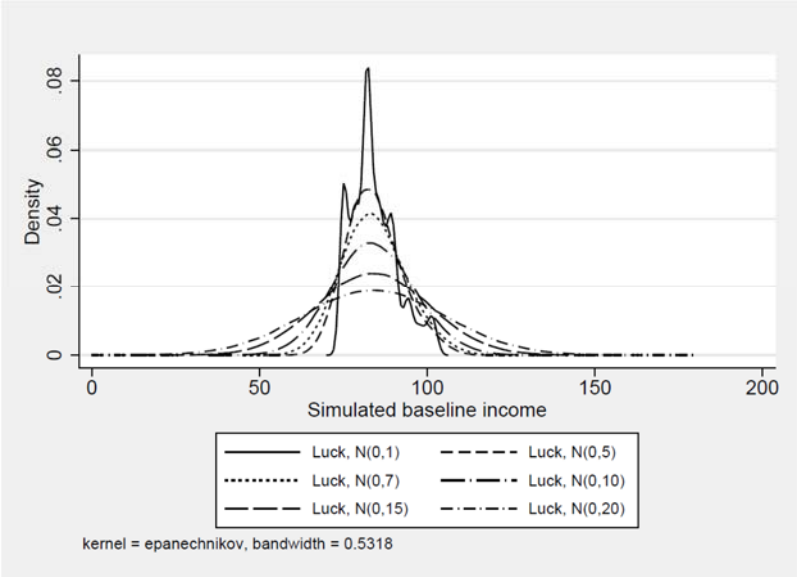
Table 14. Difference between true IOO and median estimated IOO in percentage terms: Increase-in-IOO scenario

	Excluded circumstance					
Observed population scenarios	**None (1)**	**Female (2)**	**Urban (3)**	**Region (4)**	**Father's education (5)**	**Mothers' education (6)**
				True IO share = 0.25		
all	-0.12	-34.75	-0.20	-53.74	-1.78	-8.35
Top 1% truncated	-8.03	-42.27	-8.10	-56.15	-9.65	-16.18
Top 5% truncated	-24.20	-56.73	-24.22	-61.10	-25.75	-31.91
				True IO share = 0.38		
all	0.00	-19.75	-39.84	-30.50	-1.05	-4.76
Top 1% truncated	-4.90	-24.49	-44.89	-32.96	-5.82	-9.55
Top 5% truncated	-15.52	-34.73	-54.80	-38.93	-16.47	-20.11
				True IO share = 0.48		
all	0.00	-13.35	-59.06	-20.46	-0.97	-3.41
Top 1% truncated	-3.31	-16.36	-62.85	-22.13	-3.95	-6.43
Top 5% truncated	-10.69	-23.67	-70.02	-27.36	-11.35	-13.82
				True IO share = 0.65		
all	0.00	-6.78	-78.83	-10.25	-0.65	-1.86
Top 1% truncated	-1.44	-24.49	-44.89	-32.96	-5.82	-9.55
Top 5% truncated	-4.67	-34.73	-54.80	-38.93	-16.47	-20.11
				True IO share = 0.76		
all	0.00	-3.80	-87.11	-5.82	-0.21	-0.92
Top 1% truncated	-0.45	-4.19	-88.83	-5.90	-0.63	-1.33
Top 5% truncated	-1.94	-5.68	-92.16	-7.11	-2.13	-2.83
				True IO share = 0.85		
all	0.00	-2.01	-92.48	-3.11	-0.03	-0.42
Top 1% truncated	-0.06	-2.12	-93.64	-3.08	-0.15	-0.54
Top 5% truncated	-0.64	-2.69	-95.89	-3.51	-0.74	-1.13

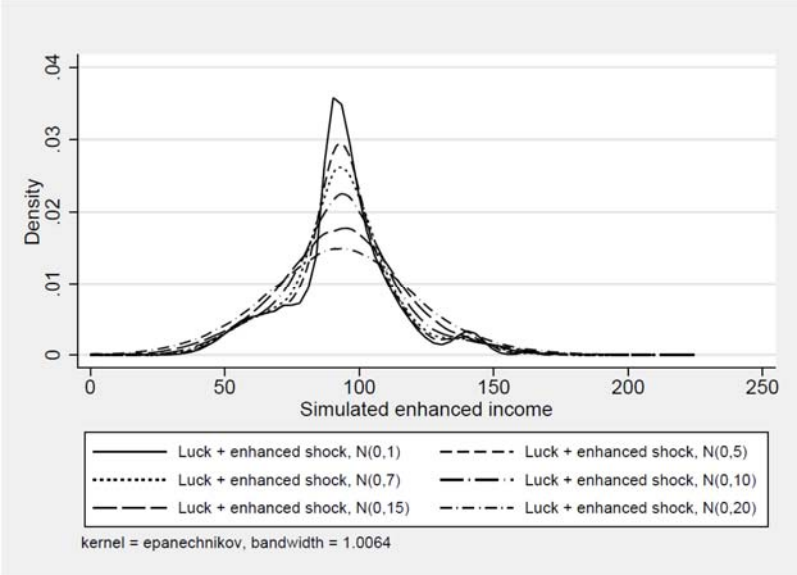
Notes: Results based on 1,000 simulations. Increase-in-IOO scenario refers to data-generating process described by equation (3), modified as described in robustness checks section.

Figure 1. Distribution of simulated incomes

Panel A. Distribution in the Baseline scenario

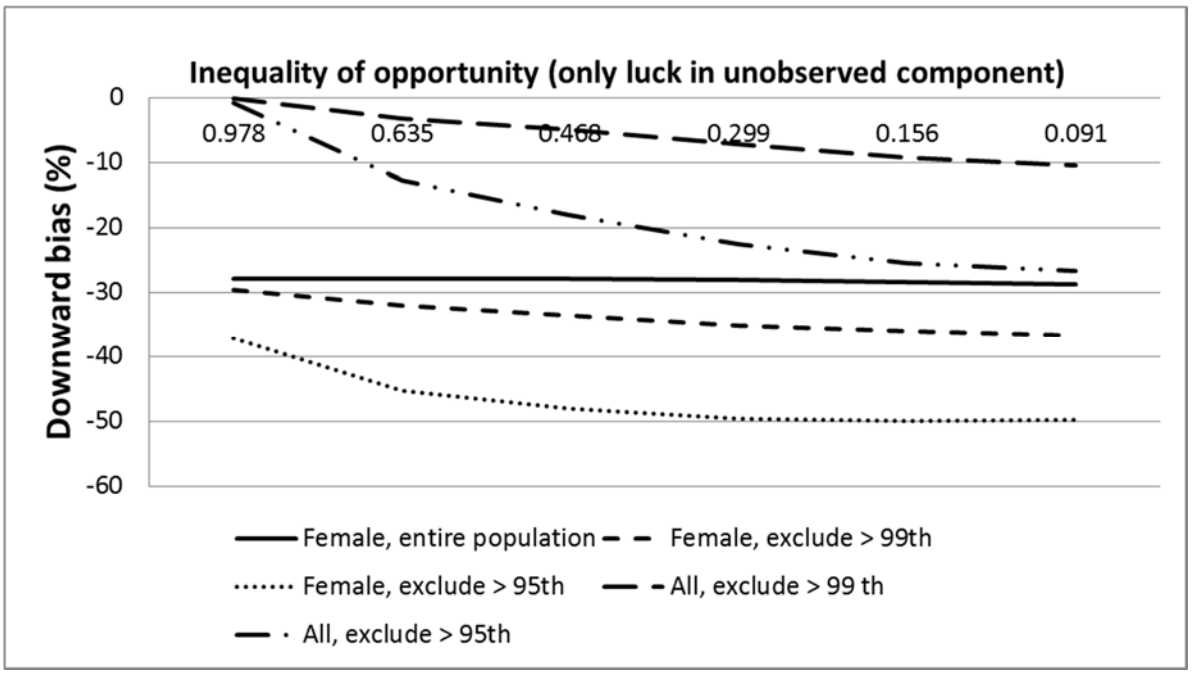


Panel B. Distributions in the Inequality enhanced scenario



Notes: Distributions in income generated according to equation (3) –panel A—plus enhanced shocks as described in experimental design section –panel B.

Figure 2. Downward bias (in percentage terms) across observed circumstances scenarios



Notes: Values are taken from tables 4 and 5.

Figure 3. Distribution of simulated incomes under increase-in-IOO scenarios

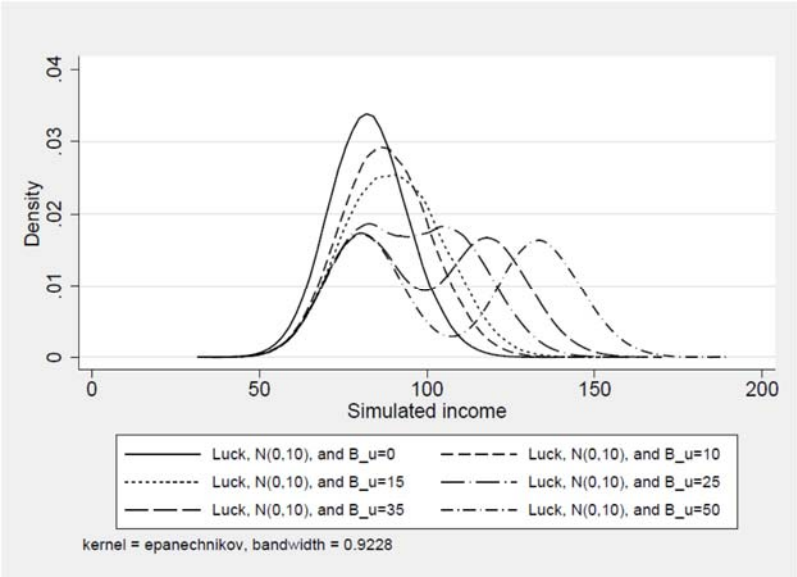
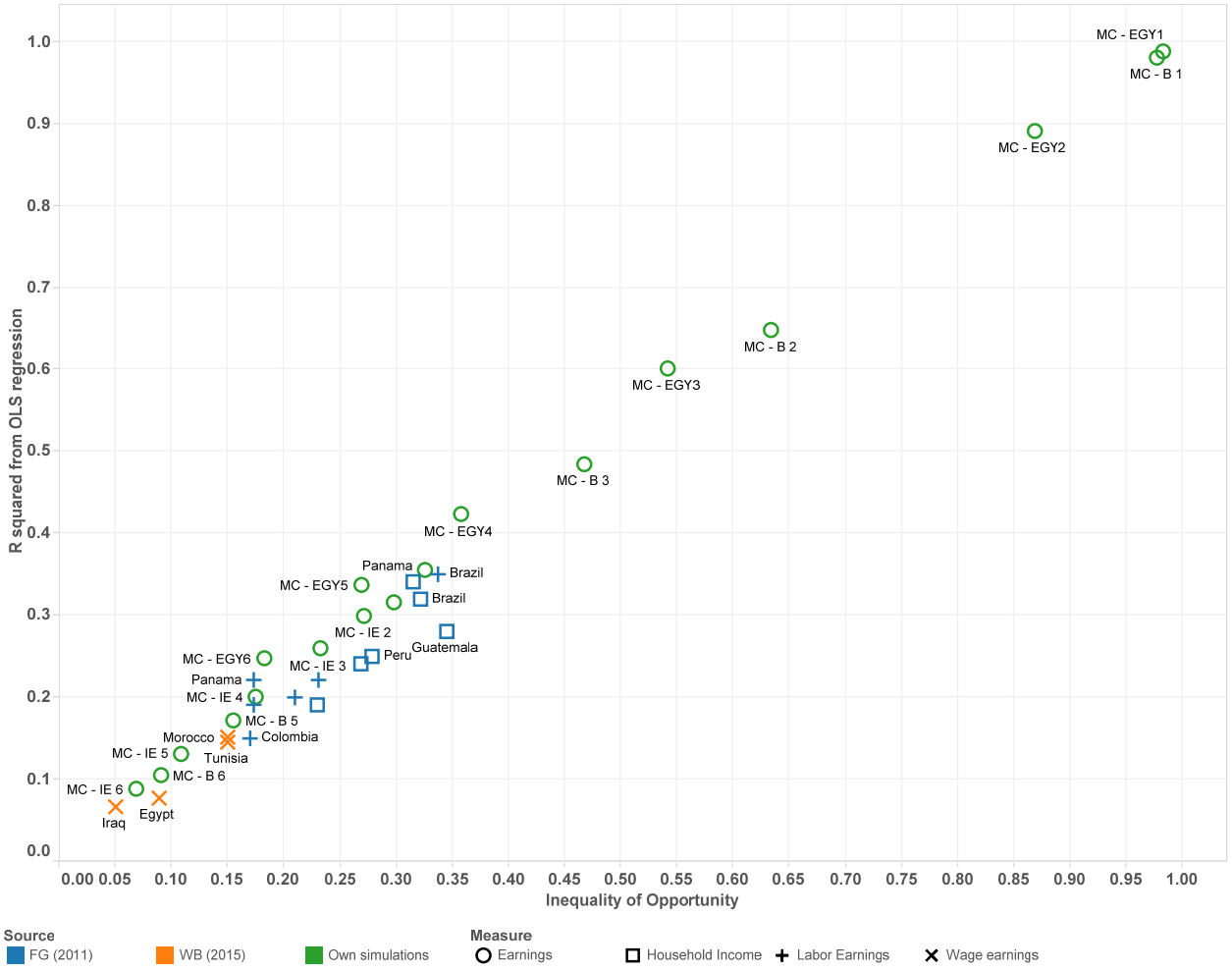


Figure 4. Correlation between overall variation explained and the inequality of opportunity estimate



Source: Authors' compilation using data from this study (LIMC, 2015), Ferreira and Gignoux (2011) and World Bank (2015) study in inequality in the MENA region. Notes: The results from LIMC (2015) are obtained from the Monte Carlo simulations in Table 3's Baseline scenarios (labeled MC – B) and Inequality enhanced scenarios (MC – IE) and Table 6's baseline scenario (MC - EGY).