Exploring the Sources of Downward and Upward Biases in Measuring Inequality of Opportunity

Gabriel Lara Ibarra
The World Bank

Adan L. Martinez Cruz
ETH- Zurich

Abstract

This study analyzes the extent of biases that affect the calculation of inequality of opportunity (IOO) for monetary outcomes such as income or consumption. Besides the largely recognized bias from unobserved circumstances, we explore two previously overlooked sources of bias: the unobservable nature of top incomes in household surveys, and the differences in costs of living. Using Monte Carlo simulations where the true IOO is known, this study presents four key findings. First, omitting a relevant circumstance can substantially downward bias the IOO estimate, and this bias is correlated with the variation of the outcome explained by the missing circumstance. Second, missing the top of the income distribution exacerbates the downward bias from omission of circumstances. Third, the IOO estimate is strongly correlated with the variation of the outcome variable explained by the set of observed combination of circumstances. This result suggests that the IOO estimate can be approximated using simple econometric techniques –such as OLS. Finally, misrepresenting differential costs of living as circumstances can lead to upward bias in the estimation of IOO, thus raising caution on the current practice of interpreting IOO estimates as lower bounds of the true IOO.

JEL codes: D63, C15

Keywords: Inequality of opportunity, mean log deviation, Monte Carlo, income distribution, top incomes, cost of living, inflation

1. INTRODUCTION

Income inequality has become firmly placed at the center stage of economic and policy debates. In the last few years, discussions have sparked due to the recent evidence of the magnitude of pay gaps between skilled and non-skilled workers (Kiatpongsan and Norton, 2014), on how long-term welfare disparities have evolved over time and may continue to increase (Piketty, 2014), or the apparent large concentrations of world wealth in few individuals (OXFAM, 2014). Inequality has also been argued to play a key role in fueling discontent in several contemporaneous social movements such as Los indignados in Spain, the Occupy protests in the US, and the events in the Middle East region that became known as the Arab Spring.

The implications for policy on how to address inequality have also evolved by first establishing the type of inequality policymakers should tackle. From the earlier discussions on political philosophy in Rawls (1971), Sen (1980), Dworkin (1981a, 1981b), Arneson (1989), and Cohen (1989) to formal treatments in economics by Roemer (1993), Van de Gaer (1993), Fleurbaey (1994) and Bossert (1995), the policy focus has shifted from addressing inequality in income to addressing inequality of opportunities (IOO). A key argument in this shift is that 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 (for example 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 is necessary to produce the right incentives to promote economic development. In contrast, policies should aim to level the playing field for all individuals. Inequalities originated in factors beyond individuals' control are inequitable, morally unjustifiable and must be compensated by society (Peragine, 2004).

A large empirical literature measuring IOO has emerged. Most applications use an indirect ex-ante approach[1] perhaps due to the fact that informational requirements can be somewhat satisfied by household surveys collected in many countries. Versions of this approach were proposed by Bourguignon et al. (2007) in Brazil, Checchi et al. (2010) in European countries, and Ferreira and Gignoux (2011) in Latin America, where measuring IOO amounts to measuring the amount of inequality that comes from individuals' observed "circumstances" (i.e. those characteristics for which an individual exerts no control). However, these estimates suffer several forms of bias that can seriously limit their ability to inform how *uneven* the playing field actually is. In this paper we illustrate the implications of three sources of bias that affect this popular approach to measure IOO in the literature: bias due to unobservable circumstances, bias due to unobservable top incomes, and bias due to misrepresentation of differential costs of living as a circumstance.

[1] See Ferreira and Peragine (2015) for a review of measuring IOO, and Roemer and Trannoy (2013) and Ramos and Van de Gaer (2012) for surveys of research in the field.

The first type of bias comes from the impossible task of observing all relevant circumstances that affect individuals' outcomes. Naturally, this source of bias has been identified in several studies (Barros et al 2010, Ferreira and Gignoux 2011, Luongo 2011, Ferreira and Peragine 2015) and hence the oft-cited label that all IOO estimates constitute a lower bound of the true IOO. The lower bound feature follows from the fact that when more circumstances are observed, the population is partitioned in a larger number of circumstance groups and the estimated IOO will unambiguously be higher (Luongo, 2011).[2] Recognizing the criticisms to the IOO estimation under this approach (Kanbur and Wagstaff 2014), we instead focus on illustrating the potential magnitude of the bias.

The paper contributes to the literature by exploring two other sources of bias that have been overlooked in the literature of IOO. Bias in IOO can also come from a characteristic of household surveys that affects all estimations using this data: unobservable top incomes. Wealthier households have a much lower probability to be captured in household surveys, and those who are captured tend to underreport their income or consumption.[3] This may lead to the underestimation of several income inequality measures such as the Gini (Hlasny and Verme 2013, Korinek et al. 2006). Here we illustrate that IOO is also affected by the missing top incomes.

A third source of bias challenges the lower bound interpretation typically associated with IOO estimates. IOO estimates may be either underestimated or *overestimated* in contexts where the different costs of living are not appropriately taken into account. Individuals' income are explained by several factors including costs of living. Thus, two individuals with the same characteristics, exerting the same effort and holding a similar job could be earning different wages due to the location where such job takes place. Nonetheless, the costs of living component of wages or income should not be deemed part of the inequality of opportunity as it would be hard to argue that policies should try to equalize wages across, say, all urban and rural areas.[4] Instead, we argue that the distribution used as starting point for the estimation of IOO should not be the observed income distribution, but a spatially-deflated distribution where incomes are adjusted to account for different costs of living.

Our results are based on a series of Monte Carlo simulations where we control the distribution of the population, as well as all relevant circumstances and random errors that affect incomes of the individuals in our hypothetical populations. In this controlled environment of income distribution, our results raise further caution on the ability of previous IOO estimates to measure the true IOO. We show that the well-known downward bias from unobserved circumstances can be substantial

[2] This shortcoming has been partially addressed by the use of much richer and long panel datasets to get at upper bounds of IOO (Niehues and Peichl, 2014) or by using latent class analysis as a way to get around the need of panel data (Donni et al. 2015).

[3] Some surveys also limit the amount of information captured of these households by top coding.

[4] Arndt et al. (2015) make a related point and propose an adjustment to the distribution of income to account for differences in costs of the baskets consumed by households and correct the upward bias of several inequality measures.

when the unobserved circumstances explain a large variation of the outcome of interest (i.e. income or consumption). Another source of downward bias is found when incomes from the right tail of the distribution are not observed and can lead to severe underestimation of IOO when the true IOO is low. Finally, major challenges in interpreting IOO estimates arise when incomes are not adjusted for differences in cost of living. Under plausible scenarios we find evidence of potential *upward* bias in IOO estimates, leading to higher uncertainty about the true value of the IOO.

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.

2. MEASUREMENT OF INEQUALITY OF OPPORTUNITY

The measure of Inequality of Opportunity (IOO) for continuous outcomes used here follows the indirect ex-ante approach of the compensation principle that is widely used in empirical studies of IOO. We briefly present this measure based on the description 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 can be written as

$$y = f(C, E, u) \tag{1}$$

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 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 becomes equivalent to measuring the extent to which $F(y|C) \neq F(y)$. This paper employs the non parametric approach to measure the magnitude of this difference.[5]

Generally speaking, this approach consists of five steps. First, define the outcome variable of interest in the survey. In our case, we will focus on individuals' income. Second, define the set of circumstances that are believed to be relevant to the individuals' observed outcomes. These

[5] In Ferreira and Gignoux (2011), the authors also propose a parametric approach based on an OLS regression and simple functional assumptions.

circumstances include gender, education of the parents, region of birth, etc.[6] 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). Under 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 certain properties.[7] The Inequality index we use here is the mean log deviation (MLD). MLD is defined as $I = (1/N)\sum_i \ln(\mu/y_i)$, where *i* stands for individual i, $y_i$is individual i's outcome, and $\mu$is the overall mean of the outcome variable.[8] The 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 reasoning behind estimation of IOO this way assumes that i) the sample distribution of the outcome variable resembles the 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, and differences in outcomes across these homogeneous groups reflect differences attributable to differences in circumstances.

We now turn to our strategy to explore the direction and magnitudes of three sources of bias in the empirical application of IOO estimation. IOO estimates are obtained from a series of Monte Carlo simulations that are manipulated to reflect situations frequently faced in empirical applications: i) not observing all relevant circumstances (i.e. circumstances that affect one's income); ii) not

[6] Continuous variables are broken into categories.

[7] These properties include 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 $\{\mu_j^k\}$ 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 (Ferreira and Gignoux, 2011).

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

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

observing individuals in the top of the income distribution (those most favored, the top incomes); and iii) presence of differential costs of living. We describe this approach in detail next.

3. STUDYING INEQUALITY OF OPPORTUNITY IN A SIMULATED ENVIRONMENT

A key advantage of using Monte Carlo simulations is the ability to design scenarios in which researchers know the *true* share of inequality due to Inequality of Opportunity. This knowledge eliminates the uncertainty inherent to real-life settings in regards to the calculation of the bias caused by the factors under study.

To explore how worrisome would the aforementioned sources of bias are, we create a hypothetical population whose individuals have certain characteristics and earn a certain income. Each individual's income is based on one of two possible *income* generating processes or scenarios. In the *baseline* scenario, an *individual*'s income results from an additive process that depends on five *circumstances* and a *random component*.

The five circumstances determining income are labeled as *gender*, *urban*, *region*, *father's education*, and *mother's education*. To resemble empirical regularities, female individuals earn less income than male individuals; urban individuals earn higher incomes than rural individuals; income varies across three regions; and individuals with more educated parents gain more income –with the mother's education causing larger positive impacts than father's education. The random component, mostly introduced to create variation across individuals, is modeled as zero-mean, normally distributed error term that increases or decreases income randomly.[10]

The second income generating scenario aims to resemble an environment where costs of living differ across a spatial dimension. In this *spatially differentiated costs of living* (SDCL) environment, we assume that individuals in urban areas earn higher incomes because they reflect higher costs of living in comparison to costs of living of, say, the rural areas. Under the SDCL scenario, urban impacts income in a multiplicative manner, i.e. in the form of an inflation factor Thus, living in an urban area is not a circumstance *per se*. Individuals' incomes instead are affected by four circumstances only: gender, region, father's and mother's education. In this scenario, we also include a random component used to create random variation across individuals.

Our main interest is to study the direction and the magnitude of the bias in measuring IOO. We use the baseline scenario to explore the impacts on the bias from two factors –namely, the well-recognized inconvenience of not observing all relevant circumstances, and the overlooked issue that the richest population is hardly represented in commonly analyzed surveys. The impact from not observing all relevant circumstances is studied under five treatments –each of which excludes

[10] To keep matters simple, effort is not a determinant of income in this simulated environment.

one of the five circumstances at a time. The impact from not observing the top incomes is studied under two levels of missing top incomes –not observing either the 1% top incomes or the 5% top incomes, respectively. We also explore the combined effect resulting from the interaction of the two sources of bias.

Under the SDCL scenario we show three specific cases: i) correctly adjusting for differential costs of living between urban and rural locations; ii) not including any location control in the IOO estimation; and iii) (incorrectly) including urban as an exogenous circumstance in the IOO estimation. We now describe each of the components of the simulated environment in detail.

3.1. Population

Individuals are randomly assigned to circumstances in both the baseline and the spatially differentiated costs of living scenarios. Assignment into circumstances in the baseline scenario is as follows.[11] An individual may be born i) either female or male; ii) either in an urban or a rural context; iii) in one of three geographical regions; and iv) from a father and a mother who each may be illiterate, or may be literate, or have completed elementary school or above. The intersection of these categorical circumstances generate (2x2x3x3x3=) 108 mutually exclusive groups or *types* of individuals. The size of each of these groups is assumed to be approximately the same –i.e. around 0.92% of the total population.[12]

Assignment of all circumstances is based on an "ordering" of groups in which group 1 is expected to include the most disadvantaged individuals, whereas group 108 contains the most favored individuals. For example, an individual is born from i) an illiterate father with probability 0.0092*(109-*group*), where *group*=1,2,…,108; ii) from a literate father with probability [1-0.0092*(109-*group*)]*(2/3) ; and from a father with elementary school or above with probability [1-0.0092*(109-*group*)]*(1/3). Notice that assigning a weight of 1/3 to the event of completing elementary school or above is meant to resemble the realistic situation in which a smaller portion of a population has completed elementary school in comparison to the portion that can only read and write. Assignment of mother's education is carried out in a similar way as the assignment to the father's education circumstance, but the probabilities are allowed to be independent. By assigning father's and mother's education independently, we allow for more variation of income across types.[13]

Regarding the region of birth, each individual is assigned to the first region - region 1 or the most advantageous region- with probability [1-0.0092*(109-group)]*(1/3); to region 2, with probability

[11] The SDCL scenario follows this approach closely, but urban is not assigned to individuals as a circumstance. The total number of types is 54.

[12] Some variations that allowed different group sizes had little impact in the qualitative conclusions. See Appendix B.

[13] Other assignment rules where mother's education was correlated with father's education were also analyzed. The results were qualitatively similar.

[1-0.0092*(109-group)]*(2/3); and region 3 (the least advantageous), with probability 0.0092*(109-group). This assignment rules imply that individuals with larger probabilities of living in the least advantageous region also face the largest probability of having illiterate parents.

Assignment of the urban circumstance follows a reasoning similar to the one used for assignment of 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 least favored region most likely live in a rural context.

Assignment of gender is carried out without attention to the "ordering" described above. That is, we assigned to each individual the same probability of being born female (52%).[14]

3.2. Baseline scenario

Using the hypothetical population described above, the circumstances assigned to each individual help define the income she will receive. The baseline income generating scenario is described by equation (3).

$$\begin{aligned} Income = {} & 85 - 7 * female + 3 * urban \\ & +9 * region_1 - 3 * region_2 - 4 * region_3 \\ & -0 * father_{illiterate} + 1 * father_{literate} + 2 * father_{elementary} \\ & -0 mother_{illiterate} + 3 * mother_{literate} + 4 * mother_{elementary} \end{aligned} \quad (3)$$

According to equation (3), average income is 85 units for the omitted category –that comprises rural males. Being female is associated with 8.2% lower incomes, while 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, individuals born in region 3 receive income 4.7% lower than average income, and individuals born in region 2 receive 3.5% lower income.

In terms of parents' educational attainment, equation (3) reflects a positive relationship between income and parental education. Father's and mother's education affect individuals' income differently.[15] In this study, the circumstance labeled mother's education is associated with higher improvements of income in comparison to the variable labeled father's education.[16]

3.3. Spatially differentiated costs of living scenario

[14] Table A.1 describes the composition of the population assumed to be composed by 200,000 individuals.
[15] An empirical regularity on whether father's or mother's education have a differential impact on individual income has not been determined, but differences have been documented in case studies (see Dickson et al., 2013).
[16] Table A.2 reports the average income in this baseline scenario by circumstance across the population.

The focus of the Inequality of Opportunity literature is to gain insights on how much of the inequality in outcomes is deemed unfair. Moreover, all differences in outcomes that are not based on circumstances should not be associated to IOO. In this context, not taking into account spatial differences in costs of living is equivalent to treat these differences as unfair. To clarify, individuals with higher incomes are typically observed in urban areas –as our baseline scenario resembles. At the same time, individuals living in urban areas face higher costs of living. Current practice in the IOO literature either treats urban as a circumstance –i.e. as an unfair condition that translates into higher incomes- or does not account for it in the estimation of IOO. Either way, this practice overlooks the fact that the higher income may result from a market compensation for the higher costs of living, and therefore, in terms of purchasing power no real differences exist across urban and rural areas – nor should they be deemed objectionable. Thus, the main takeaway from this reasoning is that estimates of IOO need to control for differences in costs of living.[17] Nonetheless, recent studies estimate IOO with no adjustments in this regard (e.g. Brzezinkski and Madga 2016; Marrero and Rodriguez 2013; Ferreira and Gignoux; 2011; and Checchi and Peragine, 2010).

To explore the effects from overlooking differences in costs of living, we modify the baseline scenario described in equation (3) to reflect differences in costs of living. The SDCL scenario represents a situation in which income in a urban context is inflated by a factor Π –as described by equations (4a) and (4b). That is, the income of urban individuals is inflated as a reflection of higher costs of living.

$$\begin{aligned} Income = {} & 85 - 7 * female + 9 * region_1 - 3 * region_2 - 4 * region_3 \\ & -0 * father_{illiterate} + 1 * father_{literate} + 2 * father_{elementary} \\ & -0 * mother_{illiterate} + 3 * mother_{literate} + 4 * mother_{elementary} \end{aligned} \quad \text{if urban =0} \quad (4a)$$

$$\begin{aligned} Income = {} & \{85 - 7 * female + 9 * region_1 - 3 * region_2 - 4 * region_3 \\ & -0 * father_{illiterate} + 1 * father_{literate} + 2 * father_{elementary} \\ & -0 * mother_{illiterate} + 3 * mother_{literate} + 4 * mother_{elementary}\} * (1 + \Pi) \end{aligned} \quad \text{if urban= 1} \quad (4b)$$

Results are estimated for two levels of differential costs of living: $\Pi \in \{0.15, 0.30\}$.

3.4. Incorporating variation to the income generating scenarios

We induce variation in the income generating processes in the form of a zero-mean, normally distributed error term. The degree of this random component, and consequently the level of inequality, is determined by the standard deviation of an additive error term included in equations (3), (4a) and (4b). We allow for different levels of observed inequality and "true" IOO by incorporating different variances under each scenario and experiment.

[17] A related argument has been presented in the inequality literature by Arndt et al. (2015), and Skoufias and Olivieri (2013).

Table 1 reports the standard deviations assumed for the error term and the *true* inequality measures. It is evident that the higher the standard deviation of the random error term, the higher the inequality in the outcome variable, and the smaller the share of inequality due to IOO (i.e. lower inequality of opportunity). For instance, in the baseline scenario, the Gini Index is 0.081 when the error term has a unitary standard deviation, and increases to 0.252 when the standard deviation is 20. Similarly, the total Mean Log Deviation increases from 0.003 to 0.035. Based on the predetermined parameters, the IOO estimate takes values between 0.978 and 0.091. The bottom panel of Table 1 reports a similar pattern for the SDCL scenario.

Table 1 also illustrates that, for a given degree of variation, the magnitude of all inequality measures is smaller under the SDCL scenario in comparison to the baseline scenario. This is a consequence of the exclusion of urban as an additive circumstance. That is, the share of IOO is smaller in the SDCL scenario because urban is not treated as a circumstance anymore.

Mechanically, there is a direct relationship between the magnitude of the standard deviation of the error and the measured income inequality, whereas there is an inverse relationship between the standard deviation of the error and the IOO estimate. In the context of a linear regression, increases in the variation of the error lead to lower explanatory power of the circumstances (R^2). Table 2 illustrates this result. For the baseline scenario, the five circumstances explain 97.9% of the variation in the outcome when the standard deviation of the error term is unitary. The explained variation decreases to 10.4% when the standard deviation is 20. For the SDCL scenario, the variation explained by the circumstances represent 61.2% and 14.9% when the standard deviation is, respectively, five and fifteen.

Table 2 also reports the variation that each circumstance explain by itself –measured as the R^2 in a linear regression model where only one circumstance is included. Region explains the most variation by itself under both income generating scenarios. In contrast, father's education explains the smallest share of variation in the outcome. This is a consequence of the relative magnitude of the coefficients assumed for each circumstance in the equations describing the income generating processes. For instance, while region can increase the average income by more than 10%, father's education can only improve it by 2.5% at most. The level of variation explained by each circumstance becomes relevant when studying the magnitude of the bias due to unobserved circumstances –details are provided below.

3.5. Exploring the sources of bias

In the baseline scenario, two factors and their interaction are manipulated to explore the direction and magnitude of the bias in measuring IOO. The first factor is the well-recognized inconvenience of not observing all relevant circumstances. We study the bias in the measurement of the IOO under five cases. For each of these cases, one circumstance is excluded from the IOO estimation and the associated bias with respect to the true IOO is calculated. For instance, in the case when

researchers do not observe the gender of the individuals, the estimate of IOO follows a grouping strategy based on four circumstances and a total of 54 `types' (instead of the 108 types that we know are relevant). This calculation is run over 1,000 simulations. In each simulation, an individual's income depends on the full set of circumstances and a randomly drawn error, but the IOO is estimated as if one circumstance is unobserved. The median IOO estimate resulting from these simulations is compared against the true IOO that has been calculated on the 108 types of individuals. Similar calculations are performed for each circumstance separately, excluding one by one.

The second factor under analysis is the fact that the richest population is hardly ever represented in commonly used household surveys.[18] We study two unobserved top incomes cases. One case assumes that the top 1% incomes are not observed. The second case assumes the top 5% incomes are not observed. Similar to the unobserved circumstances scenarios, we obtain IOO estimates when only the 99% and 95% non-excluded incomes, and then it is compared against the true IOO share (where the entire income distribution is observed). To the best of our knowledge, no previous study in the IOO literature has paid attention to this issue.

Under the SDCL scenario we focus only on the effects of adjusting or not for costs of living before estimating IOO. The true IOO is calculated for different standard deviation of the random component and across the two levels of "inflation" as described above. Next, the IOO estimates for each case are presented where i) living in urban areas is deemed irrelevant for the calculation of IOO (labeled *naïve* approach); and ii) an urban area control is included as an additional circumstance (labeled *current practice* approach).

4. RESULTS

4.1. Bias in baseline scenario

Table 3 reports results for the baseline scenario. The results show the direction and magnitude of the bias in measuring IOO resulting from i) not observing a subset of circumstances; ii) not observing a portion of top incomes is not observed; and iii) not observing neither. Specifically, the numbers reported in table 3 are differences expressed in percentage terms –i.e. *100*(true IOO - median of the estimated IOO)/ true IOO.*

Each panel in table 3 reports results for a different level of IOO –i.e. 0.978, 0.635, 0.468, 0.299, 0.156, and 0.091. The columns in table 3 identify the circumstance that is left " unobserved" in each case. The first column reports results when all circumstances are observed. The subsequent columns report results when one circumstance is excluded one by one –gender, urban, region,

[18] The bias produced by this data limitation has recently been document by Eckerstorfer et al. (2015) in the context of the estimation of the distribution of wealth in Austria.

father's education, and mother's education. The rows in table 3 show results that vary according to the observability of top incomes. The first row reports results when the entire population is observed. The subsequent rows report results when the top 1% incomes are not observed and when the top 5% is not observed. In each panel, the number in the first row and first column is zero because it represents the ideal situation under which all relevant circumstances and the entire population are observed.

We illustrate our findings using the top panel of table 3 where the true IOO is 0.978. Assuming the entire population is observed, failing to take into account any circumstance always results in a downward biased estimate. Importantly, the magnitude of the downward bias varies by circumstance not observed –with a maximum of 39.82 and a minimum of 1.07. The magnitude of the bias is associated to the relative importance of each circumstance in explaining the variation of the outcome variable. For instance, from table 2 we know that region explains the largest portion of the variation in the outcome variable by itself; and father's education, the smallest portion of the variation. Accordingly, when the entire population is observed, and region is not included as a circumstance in calculating IOO, the estimate is biased downward by almost 40%. Not observing (i.e. including) father's education resulted in a downward bias of around 1%. While the latter bias is arguably negligible, the former is substantial. Thus, a first result is pretty much in line with previous discussions on the downward bias effect of not observing a subset of circumstances: the more important the circumstance is in explaining the outcome of interest, the larger the downward bias when such circumstance is unobserved.

If we continue to use the top panel of table 3 as guideline, let's assume that all five circumstances are observed. Not observing the 1% top incomes yields a downward bias of 0.14% and this bias increases to 0.82% when the unobserved population includes the 5% top incomes. This result now points to a source of bias that has been overlooked: not observing top incomes, as it is the case in most real-life applications, leads to a downward bias whose magnitude increases with the unobserved portion of top incomes. By analyzing the patterns across true IOOs, we learn that the smaller the true share of IOO, the larger the downward bias. For instance, assuming a conservative value for true IOO (e.g. 0.299), we find a downward bias of around 7% when the top 1% incomes are not observed, and around 22% when the top 5% incomes are not observed. This second set of results implies that not observing top incomes may be troublesome if we suspect the true IOO share is actually relatively small –say below 0.3.

We focus on the case in which true IOO share is 0.299 to exemplify the interactive effects from not observing circumstances and top incomes. The potential magnitude of the bias is best illustrated when focusing on the case in which father's education is not observed. Remember father's education is the least important circumstance. According to table 3, not observing father's education leads to a negligible downward bias of -1.42%. However, this bias jumps to --4.34% when the 1% top incomes are not observed, and to -14.05% when the 5% top incomes are not observed. That is, the interaction between these two factors may increase the bias substantially.

This result may be of most relevance for the IOO literature as it could be argued that a more realistic scenario is one in which neither all relevant circumstances nor a portion of the top incomes are observed. The interaction of both limitations may produce a downward bias even larger than previous studies had expected.

4.2. Bias in Spatially Differentiated Costs of Living scenario

Table 4 reports results for the Spatially Differentiated Costs of Living scenario. The first column reports the bias resulting from implementing current practices. That is, this approach estimates IOO without controlling for differences in costs of living and, incorrectly, treating urban as a circumstance. The second column reports the bias resulting from using a naïve approach. This approach consists in estimating IOO without controlling for differences in costs of living and, correctly, not including urban as a circumstance.

The direction and magnitude of the bias under the current practice scenario are striking. In contrast to the baseline scenario, the IOO estimate is biased upward. That is, under current practice, the IOO share may be overestimated. In the case where IOO is low (0.136), the overestimation is as large as 115% if there is a 15% inflation of prices, and 261% in the case where differences in average incomes are 30%. This result is relevant to the IOO literature because spatial differentiation of costs of living is a realistic assumption. Consequently, the IOO shares might be overestimated instead of underestimated.

When we focus on the magnitude and direction of the bias under the naïve approach, we learn that foreseeing the direction of the bias in IOO estimates is actually less obvious than previous studies assumed. Under both SDCL scenarios, the bias may be downward or upward. The higher the true IOO, the more likely the estimate will be biased downward.

4.3. Robustness checks

Concerns may arise that implicit or explicit assumptions embedded in the simulated environments may drive the reported results. We carry out four robustness checks on the baseline scenario.[19] First, we explore whether results are driven by the inherent artificiality of the regression parameters associated to the circumstances in the simulated environments. To address this concern, we rerun the full set of simulations under a modified equation (3), where the parameters of the modified equation are obtained from the 2012 Egypt Labor Force Survey (ELMPS). The parameters are the estimates from an OLS regression of the log of income of working age individuals and the five circumstances of equation (3). A second set of robustness exercises explores whether varying the error distribution across circumstance groups would change the results. To do this, we rerun the

[19] A more detailed description of each of these exercises is included in Appendix B and table B.1. Accompanying tables B2 and B3 describe the true IOO that are used in the robustness checks.

set of simulations and incorporate into equation (3) a *self-reinforcing* mechanism that makes individuals more likely to remain in the social groups in which they were born. In this case, individuals from the least advantaged groups are assigned an error component with lower mean and variance than those in the most advantaged groups, leading to higher overall inequality levels. The third check tests whether our assumption of equally sized circumstance groups has any effect on the results. To explore this, we rerun the set of simulations allowing for larger concentration of the population in groups around the "middle" of the income distribution and lower concentration in the extremes – similarly to a normally distribution. The fourth scenario checks whether results are affected by the fact that under our baseline scenario the Gini Index appears to be negatively correlated with the true share of IOO. To do so, the baseline scenario described by equation (3) is replicated under scenarios that vary the total inequality through the increase in the relative importance on one circumstance –namely, urban. The relevance of the circumstance is varied by varying the coefficient associated to urban in equation (3). By doing so, the induced correlation between the Gini Index and the IOO is now positive –and, arguably, more realistic.

Results of all four robustness checks are reported in tables B.4 to B.7. In short, while the point estimates of the bias are different, the qualitative results we have discussed so far remain unchanged.

Under this controlled environment, knowing the true IOO, the estimated IOO and the R^2 of the OLS regressions used to estimate it allowed us to find an empirical regularity worth noting. From our results, we find 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, 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, Krishnan et al. (2016) 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 1 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.

5. CONCLUSIONS

We use a series of Monte Carlo simulations to explore the potential magnitude of downward and upward bias that could be expected in the typical IOO empirical estimations. We provide quantifications of the bias in the presence of the often-cited problem of missing circumstances. We also explore the effects from two features that most frequently characterize the data analyzed in IOO studies: the non-observability of top incomes and the differences in costs of living not captured by household surveys.

Unobserved top incomes are shown to exacerbate the problem of downward bias in the IOO estimates. Furthermore, this issue is still overlooked in Niehues and Peichl (2014) and Donni et al. (2015) recent attempts using panel data or latent class analysis to overcome the lower bound feature of IOO estimates. Incidentally, we show that the variation explained by the circumstances in a simple OLS regression is a very strong predictor of the IOO estimate. This result suggests that in empirical applications, the IOO estimate can be roughly (and quickly) approximated using simple econometric techniques. Finally, we show that differential costs of living can lead to *upward* bias in the IOO estimates. This result provides some evidence against the current practice of interpreting IOO estimates as a lower bound of the true IOO and thus raises further caution on the ability of earlier studies in measuring the inequality of opportunity.

The findings in this study are aimed to provide a glimpse of the potential magnitude of bias that empirical IOO estimates may have. While the magnitudes presented here clearly depend on the parameters used, we regard the analyzed scenarios as plausible income generating processes found in the real world. The main takeaway of the results presented here is that an accurate application of the IOO demands extremely detailed and even more complete data as previously thought. For a quick example, 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.

At the very least, our results suggest that empirical applications of IOO should not escape from the responsibility of data users mentioned in Ferreira et al. (2015). Studies providing IOO estimates need a lengthy discussion on what are the ideal circumstances that should be included in a given context/country, which ones are missing from the available data, whether the most important circumstances are among the missing ones, whether important regional difference in costs of living can be accounted for, and how much of the richest households the household survey may be missing. Fortunately, as recent examples such as Chetty et al. (2015) and Chetty and Hendren (2015) show, there is still a lot of untapped power of moving to richer datasets and exploiting the use of administrative data.

References

Arndt, Channing, Sam Jones and Vincenzo Salvucci "When do Relative Prices Matter for Measuring Income Inequality? The Case of Food Prices in Mozambique" Journal of Economic Inequality, Vol 13 (3), pp. 449-464

Arneson, R., "Equality of Opportunity for Welfare," *Philosophical Studies*, 56, 77—93, 1989.

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

Bossert, Walter "Redistribution Mechanisms Based on Individual Characteristics" Mathematical Social Sciences (1995), 29, 1-17

Brzezinski, Michat and Iga Magda (2016) "Inequality of Opportunity in Central and Eastern Europe: Accounting for Changes over time" IBS Working Paper 5/2016

Checchi, Daniele and Peragine, Vito (2010) "Inequality of opportunity in Italy" Journal of Economic Inequality, 8: 429-450

Checchi, Daniele, Vito Peragine, and Laura Serlenga (2010): "Fair and unfair income inequalities in Europe", ECINEQ working paper 174-2010

Chetty, Raj and Nathaniel Hendren "The Impacts of Neighborhoos on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimates" (2015) mimegraph

Chetty, Raj, Nathaniel Hendren and Lawrence Katz "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment" American Economic Review, *forthcoming*

Cohen, G. A., "On the Currency of Egalitarian Justice," *Ethics*, 99, 906—944, 1989.

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

Donni, Paolo Li, Juan Gabriel Rodriguez and Pedro Rosa Dias "Empirical Definition of Social Types in the Analysis of Inequality of Opportunity: a Latent Class Approach" Social Choice and Welfare, Vol 44 (3), pp. 673-701

Dworkin, R., "What is equality? Part 1: Equality of welfare," *Philosophy & Public Affairs*, 10 (3), 185—246, 1981a.

—, "What is equality? Part 1: Equality of resources," *Philosophy & Public Affairs*, 10 (4), 283—345, 1981b.

Eckerstorfer, Paul, Halak, J., Kapeller, J., Schutz, B., Springholz, F., and Wildauer, R."Correcting for the Missing Rich: An Application to Wealth Survey Data" Review of Income and Wealth (2015)

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.

Ferreira, Francisco H.G., Lustig, Nora and Teles, Daniel "Appraising cross-national income inequality databases : an introduction" Policy Research working paper; no. WPS 7489. Washington, D.C. : World Bank Group. (2015)

Fleurbaey, Marc "On Fair Compensation" Theory and Decision (1994) 36, 277-307

Foster, James and Artyom Shneyerov "Path Independent Inequality Measures", Journal of Economic Theory, (2000) 91: pp. 199-222.

Hlasny, Vladimir, and Paolo Verme. (2013) "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 and Adam Wagstaff, "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

Kiatpongsan, S. and M. I. Norton, "How Much (More) Should CEOs Make? A Universal Desire for More Equal Pay," *Perspectives on Psychological Science*, 9(6), 587—593, 2014.

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.

Krishnan, Nandini, Lara Ibarra, G., Narayan, A., Tiwari, S., and Vishwanath, T. 2016 "Uneven Odds, Unequal Outcomes : Inequality of Opportunity in the Middle East and North Africa" Directions in Development--Poverty;. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/24596 License: CC BY 3.0 IGO

Luongo, Patricia (2011) "The Implication of Partial Observability of Circumstances on the Measurement of IOp" Research on Economic Inequality 19 (2): 23-49

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” Social Choice and Welfare, June, Vol 43 (1), pp. 73-99

OXFAM, “Even it up: Time to end extreme inequality,” available at https://www.oxfamamerica.org/static/media/files/even-it-up-inequality-oxfam.pdf, 2014.

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

Peragine, Vito (2004): "Measuring and implementing equality of opportunity for income," Social Choice and Welfare, 22(1), 187-210.

Piketty, T., *Capital in the 21st Century*, Harvard University Press, Cambridge, Mass., 2014.

Ramos, Xavi, and Dirk van de Gaer (2012): “Empirical Approaches to Inequality of Opportunity: Principles, Measures and Evidence”, Discussion Paper No. 6672, Institute for the Study of labor (IZA), Bonn.

Rawls, J., *A Theory of Justice*, Harvard University Press, Cambridge, Mass., and Clarendon Press, Oxford, 1971.

Roemer, John (1993): “A Pragmatic Theory of Responsibility for the Egalitarian Planner”, Philosophy & Public Affairs, 10: 146-166.

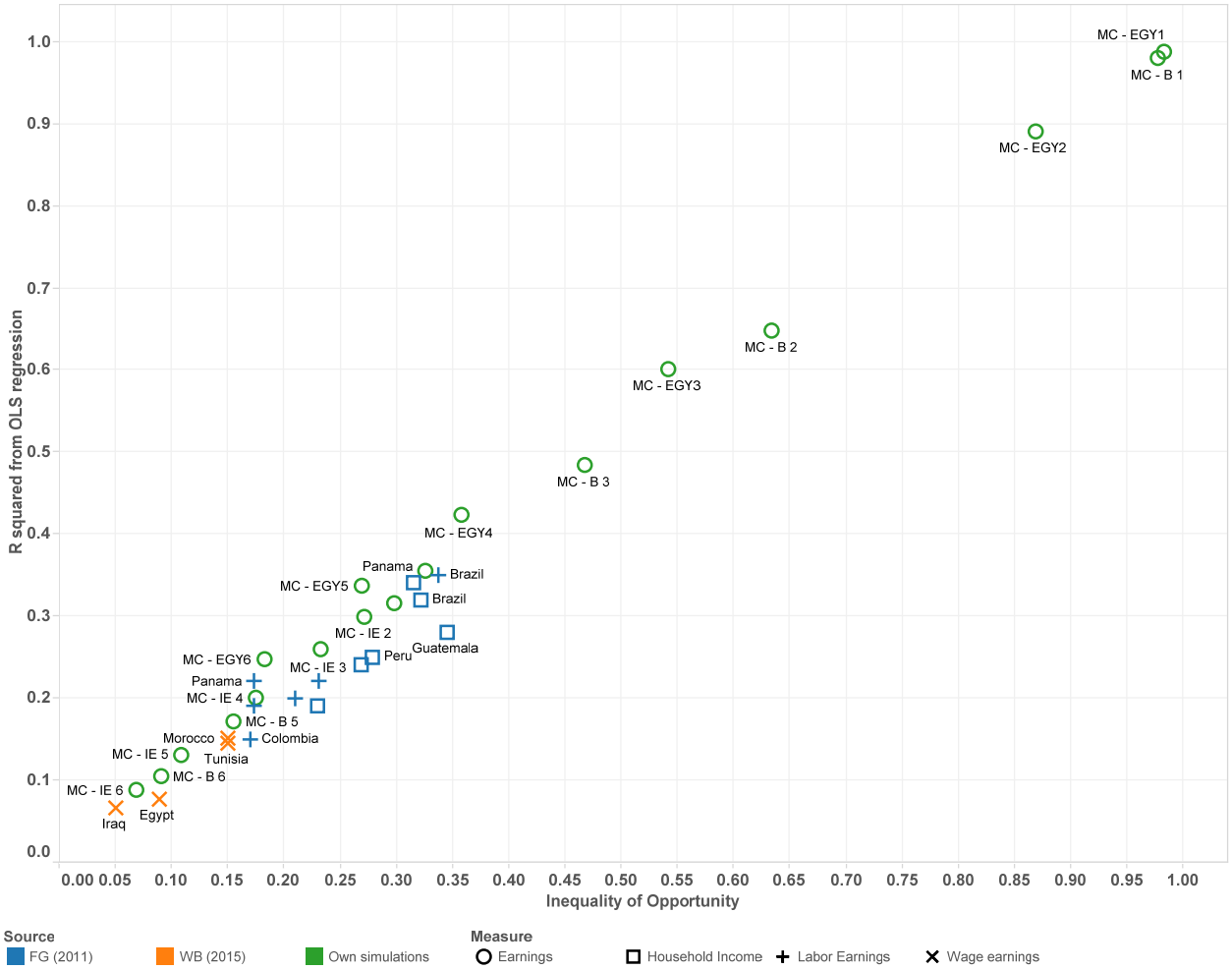
Roemer, John (1998): “Equality of Opportunity”, Cambridge, MA: Harvard University Press.

Roemer, John, and Alain Trannoy (2013): “Equality of Opportunity”, Cowles Foundation Discussion Paper n.1921, (October 25, 2013), Handbook of Income Distribution, A. Atkinson and F. Bourguignon (eds.).

Sen, Amartya (1980): “Equality of what?” in S. McMurrin (ed.) The Tanner Lectures on Human Values, Salt Lake City:University of Utah Press.

Van de Gaer, Dirk (1993): “Equality of opportunity and investment in human capital”, Ph.D. Dissertation, Katholieke Universiteit Leuven

Figure 1. Correlation between overall variation explained and the inequality of opportunity estimate



Source: Authors' compilation using data from this study,, 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 the baseline scenarios from Tables 1 and 2 (labeled MC – B), Table B5 for the inequality enhanced scenarios (MC – IE) and Table B4 results (MC - EGY).

Table 1. True inequality measures, by standard deviation of random error

Normally distributed error term	Gini Index	Total Mean Log Deviation	Share of IOO
		Baseline Scenario	
N(0,1)	0.081	0.003	0.978
N(0,5)	0.100	0.005	0.635
N(0,7)	0.116	0.007	0.468
N(0,10)	0.144	0.011	0.299
N(0,15)	0.196	0.020	0.156
N(0,20)	0.252	0.035	0.091
		SDCL Scenario	
N(0,5)	0.097	0.005	0.598
N(0,7)	0.113	0.006	0.429
N(0,10)	0.142	0.010	0.266
N(0,15)	0.196	0.021	0.136

Notes: Baseline scenario is described by equation (3). Spatially Differentiated Costs of Living (SDCL) scenario is described by equations (4a) and (4b). The error terms listed in the first column are added to equations (3), and (4a) and (4b) to represent different levels of statistical noise. Share of IOO is the ratio of the smoothed distribution MLD and total MLD. All estimates are carried out on the entire income distribution (i.e. no truncation is in place).

Table 2. Variation in outcome explained by circumstances (R^2 of linear regression), by standard deviation of random error

Normally distributed error term	All five circumstances are observed	Observed circumstance				
		Female	Urban	Region	Father's education	Mother's education
			Baseline scenario			
N(0,1)	0.979	0.261	0.183	0.564	0.121	0.199
N(0,5)	0.647	0.172	0.122	0.372	0.081	0.132
N(0,7)	0.483	0.128	0.091	0.279	0.058	0.099
N(0,10)	0.314	0.086	0.058	0.18	0.038	0.064
N(0,15)	0.171	0.045	0.031	0.099	0.019	0.034
N(0,20)	0.104	0.027	0.019	0.059	0.012	0.021
			SDCL scenario			
N(0,5)	0.612	0.198	0.031	0.364	0.055	0.101
N(0,7)	0.445	0.137	0.023	0.267	0.039	0.075
N(0,10)	0.281	0.086	0.013	0.169	0.025	0.045
N(0,15)	0.149	0.047	0.007	0.089	0.013	0.025

Notes: Baseline scenario is described by equation (3). Spatially Differentiated Costs of Living (SDCL) scenario is described by equations (4a) and (4b). The error terms listed in the first column are added to equations (3), and (4a) and (4b) to represent different levels of statistical noise. R-squared is obtained from an OLS regression using income as the dependent variable and the specified circumstance as regressor. All estimates are carried out on the entire income distribution (i.e. no truncation is in place).

Table 3. Difference (in percentage terms) between true IOO and median estimated IOO: Baseline scenario

Observed population	Unobserved Circumstances					
	None (1)	Gender (2)	Urban (3)	Region (4)	Father's Education (5)	Mother's education (6)
	True IOO share = 0.978					
Entire population	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 IOO share = 0.635					
Entire population	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 IOO share = 0.468					
Entire population	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 IOO share = 0.299					
Entire population	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 IOO share = 0.156					
Entire population	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 IOO share = 0.091					
Entire population	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 is described by equation (3).

Table 4. Difference (in percentage terms) between true IOO and median estimated IOO: Spatially Differentiated Costs of Living (SDCL) scenario

Normally distributed error term	True IOO	Bias in Empirical Approach (%)	
		Current practice	Naïve
SDCL15 scenario			
N(0,5)	0.598	32.73	-7.82
N(0,7)	0.429	53.70	6.74
N(0,10)	0.266	82.57	26.77
N(0,15)	0.136	115.30	49.44
SDCL30 scenario			
N(0,5)	0.598	50.25	-26.03
N(0,7)	0.429	90.15	-6.38
N(0,10)	0.266	157.69	26.87
N(0,15)	0.136	261.25	77.80

Notes: Results based on 1,000 simulations. Spatially Differentiated Costs of Living scenario is described by equations (4a) and (4b). SDCL15 assumes outcome variable is 15% inflated in urban contexts. SDCL30 assumes outcome variable is 30% inflated in urban contexts.

Appendix A. Description of population in baseline environment

Table A.1. Percentage (%) of 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 population (200,000 individuals). Non-diagonal percentages refer to the population with the circumstance listed in the column.

Table A.2 Average income in the baseline scenario by circumstances

	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
Individuals	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 200,000 individuals.

Appendix B. Robustness check scenarios

As in every Monte Carlo simulation study, concerns arise that implicit or explicit assumptions embedded in the simulated environments may drive the reported results. We carry out four robustness checks on the baseline scenario. The motivation and description of these robustness checks are summarized in table B.1.

The first robustness check is motivated by the concern that results may be driven by the inherent artificiality of the regression parameters associated to the circumstances in the simulated environments. In response to this concern, we modify equation (3) by replacing the parameter values with values that resemble regression parameters obtained from the 2012 Egypt Labor Force Survey. Equation (B.1) parameterizes this *Real Parameters* (RP) scenario.

$$\begin{aligned} LN(Income) = {} & 5 - 0.35 * female + 0.15 * urban \\ & + 2 * region_1 - 0.10 * region_2 - 0.17 * region_3 \\ & - 0 * father_{illiterate} + 0.05 * father_{literate} + 0.12 * father_{elementary} \\ & - 0 mother_{illiterate} + 0.13 * mother_{literate} + 0.18 * mother_{elementary} \end{aligned} \qquad \text{(B.1)}$$

Assuming equation (B.1) lists all the relevant circumstances, we can add a zero-mean, normally distributed error term, and determine the true IOO in this RP scenario. 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.

The second robustness check is motivated by the concern that results may be driven by the unrestricted social mobility embedded in the simulated scenarios. That is, in a more realistic context, individuals are more likely to remain in the social group in which they were born. Documentation of this *self-reinforcing* conditions have been characterized in the Media as marriages of power and their reinforcing effect on income equality (e.g. Cowen, 2015). In academic literature, for instance, Jenderny (2016) has documented that persistence in German top incomes fractiles is comparatively high and fairly stable in the 2001-2006 period. In addition, German top income recipients are less prone to downward mobility and see less variation in annual ranks than less rich tax payers. This self-reinforcing reasoning, when applied to the bottom incomes, is consistent with the presence of poverty traps. Particularly, the theory of geographical poverty traps argue that being born in a poverty context reinforce poverty because it permanently limits the decisions available to the individual (Kraay and McKenzie, 2014).

Thus, we modify equation (3), incorporating two self-reinforcing mechanisms simultaneously. These modifications introduce a dynamic under which an individual born in the least (most) favored group is more likely going to remain in such a group or transition to an even less (more) favored group. Specifically, we add a non-zero mean, normally distributed error terms according

to the following rule. 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 10th percentile is decreased by a random draw of a normal distribution with mean and standard of 5. If individuals are between the 10th and 50th percentiles, their income is changed (either increased or reduced) by a normally distributed draw with mean either 5 (10th to 25th) or 15 (25th to 50th). Standard deviations of these draws are both 10. Individuals above the 50th percentile increase their income. Individuals who belong to the range between the 50th and 75th 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 75th and 90th percentiles, the draw is taken from a normal distribution with mean 15 and standard deviation of 10; for individuals between 90th and 95th percentiles, the mean is 20 and the standard deviation is 10; when between 95th and 99th, 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 final distribution of incomes yields a higher inequality than under the baseline scenario.

The third check is motivated by the concern that results may be driven by the size of the groups of individuals. We have assumed equal size for simplicity but variation in the size of social groups can be induced by allocating individuals 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. In this way, sizes of each group is different and both the least favored group and the most favored groups are relatively small.

The fourth robustness check is motivated by the concern that results may be driven by the unrealistic assumption that the Gini Index is negatively correlated with the true share of IOO. Strictly speaking, this situation is not an explicit assumption but a mechanical consequence from the decision of modelling increase in total inequality as due to increase in statistical noise. Consequently, the baseline scenario described by equation (3) is replicated under scenarios that vary the total inequality through the increase in the relative importance on one circumstance – namely, urban. The relevance of the circumstance is varied by varying the coefficient associated to urban in equation (3). By doing so, the induced correlation between the Gini Index and the IOO is now positive –and, arguably, more realistic.

For the first three robustness checks, table B.2 reports the true inequality measures and the variation explained by circumstances under each statistical noise scenario. Table B.3 reports the corresponding figures for the fourth robustness check under each scenario modifying the coefficient of urban in equation (3).

Table B.1 Motivation and description of robustness check scenarios

Label	Concern	Reasoning behind concern	Approach
	Are results driven by the assumption that …	***Making sure results do not depend on this assumption is relevant because …***	***Baseline scenario is replicated …***
Real Parameters	regression parameters in the simulated environment can be chosen with no reference to a specific real-world contexts?	parameters chosen for statistical convenience may not deliver results applicable to real-world contexts.	using parameters that resemble regression parameters estimated on the 2012 Egypt Labor Force Survey.
Self-reinforcing	social mobility is unrestricted?	inequality in real-world contexts may also be caused by a self-reinforcing layer making baseline scenario conclusions to be impractical.	adding two self-reinforcing mechanisms simultaneously. These mechanisms imply that an individual born in the least (most) favored group is more likely to remain in such a group or transition to an even less (more) favored group.
Different group sizes	all groups are composed by the same number of individuals?	realistically, each group should have a different size. In general, there are relatively few individuals in the most favored group. The size of the groups with the least favored individuals depends on the specific context of interest and it will depend on the size of the group in the middle.	by assigning individuals to each group according to a probabilistic rule that is normally distributed. In this way, sizes of each group is different and both least favored and most favored are relatively small groups.
Gini-IOO positively correlated	increase in inequality is due to increase in the statistical noise?	this modelling decision translates into a negative correlation between the Gini index and the IOO share which is a counterintuitive correlation. More realistically, the increase in inequality summarized by the Gini index is expected to reflect a higher IOO share.	under scenarios that increase total inequality by increasing the relative importance of one circumstance. In this way, Gini and IOO are positively correlated.

Source: Own compilation.

Table B.2 True inequality measures and variation in outcome explained by circumstances (R^2 of linear regression) under different levels of statistical noise for robustness check scenarios

Noise (normally distributed error term)	True inequality measures			R-squared					
	Total			All five	Observed circumstance				
	Gini Index	Mean Log Deviation	Share of IOO	circumstances are observed	Female	Urban	Region	Father's education	Mother's education
Real Parameters scenario									
N(0,0.1)	0.165	0.012	0.984	0.987	0.041	0.075	0.923	0.061	0.075
N(0,0.3)	0.174	0.013	0.869	0.891	0.037	0.068	0.833	0.054	0.068
N(0,0.7)	0.212	0.022	0.543	0.601	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	0.273	0.336	0.015	0.025	0.314	0.02	0.026
N(0,1.5)	0.332	0.064	0.183	0.246	0.011	0.019	0.231	0.015	0.018
Self-reinforcing 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)	0.18	0.017	0.272	0.298	0.08	0.056	0.172	0.037	0.062
N(0,7)	0.194	0.02	0.233	0.258	0.068	0.049	0.15	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	0.13	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
Different group sizes scenario									
N(0,1)	0.076	0.003	0.974	0.975	0.302	0.074	0.522	0.026	0.094
N(0,5)	0.095	0.005	0.602	0.614	0.189	0.047	0.329	0.016	0.059
N(0,7)	0.111	0.006	0.431	0.444	0.137	0.034	0.238	0.013	0.042
N(0,10)	0.14	0.011	0.271	0.283	0.088	0.021	0.152	0.008	0.026
N(0,15)	0.193	0.02	0.136	0.148	0.044	0.011	0.081	0.004	0.014
N(0,20)	0.249	0.035	0.08	0.091	0.028	0.007	0.049	0.003	0.009

Notes: The three income generating scenarios are described in table B.1. The error terms listed in the first column are added to represent different levels of statistical noise. Share of IOO is the ratio of the smoothed distribution MLD and total MLD. All estimates are carried out on the entire income distribution (i.e. no truncation is in place). R-squared is obtained from an OLS regression using income as the dependent variable and the specified circumstance as regressor.

Table B.3 True inequality measures and variation in outcome explained by circumstances (R^2 of linear regression) under different levels of relevance of urban as a circumstance

Coefficient of urban [equation (3)]	True inequality measures			R-squared					
		Total		All five	Observed circumstance				
	Gini Index	Mean Log Deviation	Share of IOO	circumstances are observed	Female	Urban	Region	Father's Education	Mother's education
0	0.143	0.011	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	0.231	0.197	0.069	0.098
15	0.163	0.014	0.523	0.541	0.055	0.367	0.201	0.088	0.118
25	0.191	0.019	0.681	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	0.861	0.881	0.014	0.835	0.164	0.121	0.141

Notes: The Gini-IOO positively correlated scenario is described in table B.1. The values listed in the first column are the parameter values associated to the urban circumstance in the equation (3). Share of IOO is the ratio of the smoothed distribution MLD and total MLD. All estimates are carried out on the entire income distribution (i.e. no truncation is in place). R-squared is obtained from an OLS regression using income as the dependent variable and the specified circumstance as regressor.

Table B.4. Difference (in percentage terms) between true IOO and median estimated IOO: Real Parameters scenario

Observed population	Excluded circumstances					
	None (1)	Gender (2)	Urban (3)	Region (4)	Father's education (5)	Mother's education (6)
			True IO share = 0.984			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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. Real Parameters scenario described by equation (B.1).

Table B.5. Difference (in percentage terms) between true IOO and median estimated IOO: Self-reinforcing 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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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. Self-reinforcing scenario incorporates two mechanisms to increase the probability that individuals remain in the groups they are born in.

Table B.6. Difference (in percentage terms) between true IOO and median estimated IOO: Different group sizes 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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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			
Entire population	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. Different group sizes scenario allocates individuals according to a normal distribution that allows middle class to be the largest one.

Table B.7. Difference (in percentage terms) between true IOO and median estimated IOO: Gini-IOO positively correlated scenario

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					
Entire population	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					
Entire population	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					
Entire population	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					
Entire population	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					
Entire population	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					
Entire population	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. Positive correlation between Gini Index and IOO share is induced by increasing the relative importance of urban in equation (3).