I.A.R.I.W.

IBGE
Instituto Brasileiro de Geografia e Estatística

Measuring Inequality of Opportunity with Latent Variables

Florian Wendelspiess Chávez Juárez (University of Geneva, Switzerland)

Paper Prepared for the IARIW-IBGE Conference
on Income, Wealth and Well-Being in Latin America

Rio de Janeiro, Brazil, September 11-14, 2013

Session 6: Inequality of Opportunity

Time: Friday, September 13, 11:00-12:30

Measuring inequality of opportunity with latent variables

Florian Wendelspiess Chávez Juárez[*]

June 14, 2013

Abstract

In this paper I show that recently proposed methods to quantify the level of inequality of opportunity are likely to be downward biased when the dependent variable is a proxy for an unobserved concept. Using a multidimensional framework of development, such as the capability approach, or a standard utility maximization framework with heterogeneous preferences permits us to show that such measurement errors are the rule rather than the exception. I propose to estimate the latent variable of interest through appropriate multivariate techniques to circumvent the aforementioned bias. Using a simulation and an empirical illustration, I show that the use of multiple indicator variables and appropriate aggregation techniques can reduce the bias substantially. Using data from Mexico, it is found that inequality of opportunity of the broader concept of economic well-being is more than twice as high as inequality of opportunity in log income, which is commonly used as a proxy of the first.

Keywords: inequality of opportunity, multidimensional development, proxy measurement error, capability approach

JEL-Classification: C18, D63, O54

1 Introduction

The concept of inequality of opportunity has become very popular over the last 15 years, especially since Roemer's seminal contribution in 1998. His contribution served as the main theoretical reference for most subsequent work (Roemer, 1998). The fundamental idea of this approach is to look at inequality in the opportunities to achieve a goal rather than at inequality in the final outcome. The underlying idea is that the total outcome is the fruit of opportunities, effort and probably other factors such as luck or inherent abilities. In order to level the playing field, policy should ensure that opportunities are equal for everybody. In this perspective, inequalities due to circumstances beyond the control of the individual are considered to be socially unfair.

[*] Department of Economics, University of Geneva, Uni-Mail, 40 Bd. du Pont d'Arve, 1211 Geneva, Switzerland. Phone +41 22 379 82 16. florian.wendelspiess@unige.ch

In contrast, different outcomes due to different levels of effort or ability might be considered to be socially acceptable.

In this sense, the literature on inequality of opportunity is closely related to the capability approach. To take the words of Sen (2001), inequality of opportunity corresponds to inequality in the freedom to choose. Put differently, inequality of opportunity corresponds broadly to inequality in capabilities while inequality in outcomes would correspond to inequality in functionings. Combining the two approaches is not only interesting from the conceptual point of view, but it allows us also to take advantage of complements in the empirical literature.

The recent empirical literature on inequality of opportunity aimed at improving its measurement. The main difficulty comes from the fact that effort is generally not observed. For this reason, decomposing total inequality in inequality due to effort, inherent abilities and circumstances beyond the control of the individual becomes challenging.

Despite the difficulties, there seems to be a growing consensus that the best way to assess inequality of opportunity is to focus on estimating of the part of inequality due to circumstances. Recent examples can be found in Checchi and Peragine (2010); Ferreira and Gignoux (2011, 2013); Paes de Barros et al. (2009); Yalonetzky (2012). Besides some differences in the methodologies - e.g. due to different response variables - there is the common idea of relating an observed outcome to a set of circumstances to measure the part of inequality that can be attributed to circumstances. Circumstances are defined as characteristics beyond the control of the individual.

Many of these authors acknowledge that using this approach they are likely to underestimate inequality of opportunity due to non-observed circumstances. All inequality that cannot be related to observed circumstances is attributed to effort or luck.

In this article I discuss another source of potential downward bias. I show that the proposed methods to assess *ex-ante* inequality of opportunity are likely to be downward biased when the measure of the outcome is imprecise. Such imprecisions can arise for various reasons. A very simple justification would be a classical measurement error due to approximate responses by households, for instance on their income. However, there are more fundamental reasons why the observed variables might be only an approximation of the concept the researcher wants to analyze. The capability approach offers a comprehensive framework for this discussion. As noted earlier, inequality of opportunity can be seen as inequality in capabilities. Capabilities are generally assumed to be latent and therefore unobserved, while functionings are observable (Krishnakumar, 2007; Krishnakumar and Ballón, 2008; Anand et al., 2011). Realizing one opportunity might reduce the possibilities to realize another one. For instance, taking advantage of the opportunity to send children to school reduces the opportunity for the family to increase income by sending the child to work. These trade-offs between opportunities or capabilities make the observed outcome variable only a proxy variable of the underlying latent concept. The observed functionings can be good proxies of the underlying capabilities, but they are not exact measures. Throughout the paper I follow Gibson and Kim (2013) and use the term *proxy measurement error* to describe the deviation of the observed variable from the concept the researcher wants to analyze. Thus, the proxy measurement error can arise from standard measurement errors, latent concepts or heterogeneous preferences among others.

After discussing the implication of the proxy measurement error on the methods proposed recently to estimate *ex-ante* inequality of opportunity, I propose a small extension of the method introduced by Ferreira and Gignoux (2013) to reduce the bias. The main idea is to use first latent variable models such as factor analysis to estimate the underlying concept we want to analyze. This estimated value is then used as outcome variable in the method proposed by Ferreira and Gignoux (2013).

I illustrate the formal discussion on the bias and the proposed solution using simulations. They confirm that the biases arising from imprecise measures of the outcome variable can be very large and that they can be reduced substantially when using multivariate methods. Finally I present an empirical example with Mexican data and show that the measure of inequality of opportunity more than doubles when including multiple indicators as opposed to a single indicator of income.

The paper concludes on a short discussion of the more general role of latent variables and the associated proxy measurement errors in empirical research.

2 The measurement of inequality of opportunity and its bias when the outcome is imprecisely measured

2.1 The general idea of the regression approach

Recent contributions to the measurement of inequality share a common idea of relating an outcome variable exclusively to circumstances beyond individuals' control. In a first step the expected outcome conditioned on circumstances is computed. This conditional expectation can be estimated using regression analysis or non-parametric methods. In a second step, these predicted values are used as an argument for a standard inequality measure. For instance, Ferreira and Gignoux (2011) use an OLS regression to relate a continuous variable to a set of circumstances. They then apply the mean logarithmic deviation measure to the predicted outcome variables. Paes de Barros et al. (2009) use a probit model to relate binary outcome variables to circumstances and then use a dissimilarity index on the predicted probability.

The idea behind this is that all the variation in individuals' conditional expectations is exclusively due to circumstances. Let Y be the outcome variable and X the set of circumstances beyond the control of the individual. Using this notation we can define an absolute measure of inequality of opportunity as:

$$\theta_a = I(E[Y|X]) \tag{1}$$

where $I(.)$ is any inequality measure respecting the common axioms. Whenever the predicted values and the original dependent variable have the same scale, a relative inequality of opportunity measure can be computed

$$\theta_r = \frac{I(E[Y|X])}{I(Y)} \tag{2}$$

where Y is the original outcome variable. The intuition is simple: the numerator measures inequality due to X and the denominator measures total inequality. θ_r is therefore a measure of the relative importance of inequality of opportunity. Note that it is not possible to compute θ_r in the case of a binary outcome variable, since the original variable is dichotomous while the predicted probabilities are continuous.

Following the same idea, Ferreira and Gignoux (2013) propose a measure of inequality of opportunity for a continuous variable with no inherent scale, for instance educational achievement. When the outcome variable has no inherent scale, the measurement of inequality of opportunity should be scale and translation invariant, which is not possible with the mean logarithmic deviation measure used in Ferreira and Gignoux (2011). They propose to use the variance instead and to compute only the relative measure of inequality of opportunity. The variance is translation invariant and by using the ratio of two variances, the measure becomes also scale invariant. Hence, their measure is given by:

$$\hat{\theta}_{IOP} = \frac{Var(E[Y|X])}{Var(Y)} \tag{3}$$

A very appealing feature of this index is that it is the R^2 statistic of a simple OLS estimation whenever the conditional expectation of Y given X is estimated through ordinary least squares. Thus, relative inequality of opportunity is defined as being the part of the variation in the outcome explained by circumstances.

The authors acknowledge that $\hat{\theta}_{IOP}$ is a lower bound estimate of inequality of opportunity, since not all circumstances beyond individuals' control can be included. Given that the R^2 increases with each included variable, any excluded circumstance variable decreases this statistic and therefore the estimated inequality of opportunity is underestimated. This is in fact a common problem to all these methods based on the general form described in equation (1) and (2)

2.2 Bias due to proxy measurement error in the dependent variable

In this article I discuss a second source for a downward bias of $\hat{\theta}_{IOP}$ stemming from the outcome variable. To do this, I focus on the methodology proposed by Ferreira and Gignoux (2013), because it is the most intuitive way to discuss this issue[1].

Assume that the true relationship between the outcome Y and the circumstances X is given by

$$Y = X\beta + \epsilon \tag{4}$$

and assume further that we do not directly observe Y but an approximation Z given by:

$$Z = Y + \nu \tag{5}$$

where ν is the deviation of Z from Y. Let us assume for simplicity that $E[\nu] = 0$. Putting (5) into (4) we get

$$Z = X\beta + \epsilon + \nu \tag{6}$$

As long as $E[\nu] = 0$ and $Cov(Y, \nu) = 0$ the estimated $\hat{\beta}$ remains unbiased[2]. In the regression context such a proxy measurement error in the dependent variable does not affect the results. However, this is not true for the inequality of opportunity measure given in (3). Let us first consider the simplest case where $E[\nu] = 0$ and $Cov(Y, \nu) = 0$. This case corresponds to a white noise proxy measurement error.

Case with $E[\nu] = 0$ and $Cov(Y, \nu) = 0$

Under the assumption that ν is white noise, we have the true inequality of opportunity measure given by:

$$\theta_{IOP} = \frac{Var(X\beta)}{Var(Y)} \tag{7}$$

and the estimated value is given by:

$$\hat{\theta}_{IOP} = \frac{Var(X\beta)}{Var(Y + \nu)} = \frac{Var(X\beta)}{Var(Y) + Var(\nu)} \tag{8}$$

[1] The finding can be generalized to all methods using the method proposed in (2).

[2] For the case where $E[\nu] \neq 0$ the coefficients of the circumstances remain unbiased, but the estimation of the constant term becomes biased. This is actually no problem, since the constant is irrelevant for the computation of the R^2.

Dividing (8) by (7) we get a measure of the relative bias:

$$\frac{\hat{\theta}_{IOP}}{\theta_{IOP}} = \frac{Var(Y)}{Var(Y) + Var(\nu)} < 1 \tag{9}$$

which is necessarily less than 1, suggesting that $\hat{\theta}_{IOP} < \theta_{IOP}$. The intuition behind this result is simple: by adding some white noise to the dependent variable we increase its variance. However, the circumstance variables X cannot explain this white noise and therefore the explained part of the variance remains the same. It follows that the estimated R^2 statistic is lower than it would be when observing Y.

Case with $E[\nu] = 0$ and $Cov(Y, \nu) \neq 0$

Now, let us relax the assumption of $Cov(Y, v) = 0$ and allow it to be nonzero. This change has two consequences for the estimation of θ. First, the variance of Z is no longer the simple sum of two variances, now it includes the covariance between Y and ν and could therefore be smaller than the variance of Y. Second, the estimated $\hat{\beta}$ is no longer unbiased due to the fact that $\sigma_{XY} \neq 0 \ \wedge \ \sigma_{Y\nu} \neq 0 \ \Rightarrow \ \sigma_{X\nu} \neq 0$.

Equation (9) becomes therefore somewhat more complicated due to the fact that we can no longer simplify that many terms:

$$\frac{\hat{\theta}_{IOP}}{\theta_{IOP}} = \underbrace{\frac{Var(X[\beta + \Delta])}{Var(X\beta)}}_{A} \times \underbrace{\frac{Var(Y)}{Var(Y) + Var(\nu) + 2Cov(Y, \nu)}}_{B} \tag{10}$$

where Δ is the bias in the estimation of β taking the form $\Delta = (X'X)^{-1}E[X'\nu]$. The sign of the bias will depend on the relative importance of the bias in the estimation of β and the covariance between Y and ν. Unfortunately, it is impossible to conclude generally on the sign of the bias just by looking at the sign of covariances. Table 1 displays the possible combinations and their effects on the relative bias:

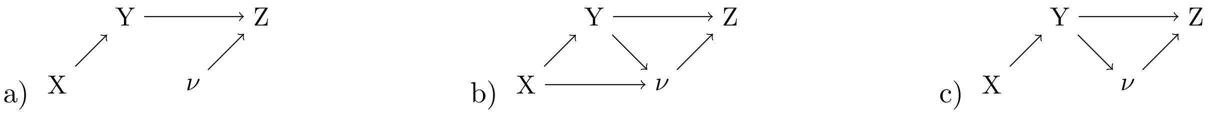
$Cov(Y, \nu)$	$Cov(Y, X)$	$Cov(X, v)$	A	B	Direction of bias
> 0	> 0	> 0	> 1	< 1	unknown
> 0	< 0	< 0	> 1	< 1	unknown
< 0	> 0	< 0	< 1	$\gtreqqless 1$	unknown
< 0	< 0	> 0	< 1	$\gtreqqless 1$	unknown

Table 1: Possible sign of the bias

Whenever $Cov(Y, \nu) > 0$, the second term (B) is smaller than 1, suggesting an underestimation of the true θ. However, the same condition on the covariance implies an overestimation of β and therefore the first term (A) is bigger than 1. In the opposite case where $Cov(Y, \nu) < 0$, the first term is necessarily lower than 1 due to a bias of β towards zero, while the second term can be bigger or smaller than 1. Hence, it is not possible to conclude on the sign in any of the four cases.

Even though it is impossible to make a statement about the bias in general, we can focus on a large subsample of situations where we assume that the association of X and ν comes only through Y. Figure 1 displays three different causal graphs to illustrate how ν can be related to the other variables.

Figure 1: Different possibilities of relationship of ν with Y and X.



In situation a) ν is white noise, in b) ν is related to X and Y and in c) ν is related to X only through Y.

Situation a) corresponds to the easiest case explained in equation (9) where inequality of opportunity is always underestimated. The covariance between X and ν is zero in this case and therefore the estimated parameters of the OLS regression remain unbiased.

Panel b) represents the general situation outlined above (equation 10) in which it is impossible to conclude on the sign of the bias. In this case the sign of the bias depends essentially on the strength of the two relationships between X and ν on the one hand and Y and ν on the other.

Finally, panel c) is a special case of b) where the relation between X and ν is produced by their link through Y exclusively. In this case, which is likely to be realistic for many situations, it can be shown that $\hat{\theta}$ is always smaller than θ. In appendix A I present the proof for this statement for the univariate case and in section 4.1 I include a simulation that shows the bias in function of $Cov(Y, \nu)$.

Overall it is clear that a bias has to be expected in case of a proxy measurement error. For practical reasons, the most relevant cases are almost surely generating a negative bias. Overestimating inequality of opportunity can only arise when the proxy measurement error is more related to the circumstances than to the outcome. Such a situation is rather unlikely to occur in the analysis of inequality of opportunity as long as the outcome variables are chosen in a sound way. Moreover, the solution I propose in the next section is likely to detect such a situation, allowing the researcher to go over his or her model in a more general way.

3 Using a two-step method to reduce the bias

In the previous section I showed that a proxy measurement error generally introduced a downward bias to the estimation of inequality of opportunity. In this section I propose a possible solution to this problem in order to reduce the bias substantially. The general idea is to add a preceding step to the measurement of inequality of opportunity, where we estimate the latent opportunity (or capability) set - say $O(Z)$. In a second step we can then regress this outcome on the set of circumstances and use the R^2 as a relative measure of inequality of opportunity as proposed by Ferreira and Gignoux (2013)

We can write the general from of this index as a weighted sum of all observed indicator (proxy) variables[3]:

$$O(Z) = \sum_{j=1}^{k} w_k z_k \tag{11}$$

where w_k is the weight related to the indicator z_k. The question is how to find the appropriate weights. We must use some aggregation techniques that estimate in the most accurate way the underlying opportunities. The choice of the appropriate model highly depends on the context and is not necessarily limited to statistical argumentation. Imagine the job market where the quality of a candidate is certainly not assessed by considering only one indicator such as the level of education or the experience. Employers rather look at a series of indicators and implicitly give them weights to get a general measure of the candidate's qualification for the job.

Nevertheless, more powerful methods might be useful when an intuitive aggregation is not clear. In the case of continuous variables, the use of a latent variable model like factor analysis might be a simple but adequate solution. The goal of factor analysis is to estimate an underlying factor of the indicator variables. Formally, the factor analysis can be written as:

$$Z = \Lambda f + \nu \tag{12}$$

where f is the latent opportunity variable and Z a vector of indicator variables, thus $O(Z) = \sum_{j=1}^{k} w(\Lambda_k) Z_k$, where the weight of dimension k is a function of Λ_k. Different estimation methods are possible, including principal factor, iterated principal factor and maximum likelihood. In the case of count data or dichotomous indicators such as having a clean floor or not, the aggregation becomes more difficult. Polychoric factor analysis can be an option in this case, but easier aggregation methods might be useful as well. In section 4.2 I provide different aggregation methods in the empirical application. The best choice might depend on the data and is beyond the scope of this paper[4].

One could argue that instead of first estimating a factor analysis and then an OLS regression, it is better to estimate a MIMIC model. Even though this is appealing, it has a major drawback which completely invalidates the idea in our context. In the case of a MIMIC model the factor f is predicted partially depending on X, hence it is estimated to fit best on the circumstance. This

[3] I assume that the variables z_k are standardized, which allows me to use this simple notation. Otherwise, we would have to add a standardization function.

[4] See for example Hair et al. (1998) for a more general discussion on factor analysis or Muthén (1978) for the estimation of factor analysis when some indicators are binary.

is directly increasing the model fit we use as measure of inequality of opportunity. Thus, the MIMIC framework tries to maximize to some extend the R^2 of the structural equation, which is far from the purpose of the proposed method.

To summarize, a simple but effective way to reduce the bias due to imprecise measures is to use several measures allowing us to get a cleaner proxy of the underlying variable. The method consists of two steps:

1. Estimate the latent variable out of a set of observed indicator variables using an appropriate model (e.g. factor analysis).
2. Use the estimated latent factor as dependent variable in the method proposed by Ferreira and Gignoux (2013)

I insist in the measure proposed by Ferreira and Gignoux (2013) since it is, to the extent of my knowledge, the only one that is translation and scale invariant. This property is needed because the latent factor does not have an inherent scale, so that one can translate or rescale it.

Before turning to the illustrations, I discuss briefly how the factor analysis can help us to identify the case where an upward bias is theoretically possible. Recall situation b) in Figure 1 where ν is directly influenced by Y and X. We have seen that the upward bias could arise whenever ν is mainly explained by X. However, in this case the indicator variables Z depend on two underlying concepts Y and ν and it is therefore very likely to find more than one underlying factor. Hence, finding more than one underlying factor should guide the researcher to a careful reevaluation of the variables used and the concepts they are supposed to measure.

4 Illustrations

The empirical part of this paper is divided into two main sections. The first part is based on simulated data where we know the true value of inequality of opportunity. In the second part I use real world data to see how much the inclusion and exclusion of additional dimensions affect the measurement of inequality of opportunity.

4.1 Simulations

I now present two simulations, starting with the case where $Cov(Y, \nu) = 0$ and followed by a second where I relax this assumption.

Simulation for the case of $Cov(Y, \nu) = 0$

The goal of this first simulation is to see how the number of dimensions included in the factor analysis affects the bias. To do this, I simulate 1000 runs with 1000 individuals each. Individuals maximize their Cobb-Douglas utility function over 10 goods by respecting the opportunity (budget) constraint:

$$max \;\; U(Y) = \prod_{j=1}^{10} y_{ji}^{\alpha_{ji}} \qquad s.t. \sum_{j=1}^{10} y_{ji} = m_i \tag{13}$$

where the $0 \leq \alpha_{ji} < 1 \;\; \forall \; j, i$ and $\sum_{j=1}^{10} \alpha_{ji} = 1 \;\; \forall \; i$. m_i is the total amount of opportunities (you might think of the budget) people have. This budget is simulated using the following model:

$$m_i = 60 + 8x_1 + 8x_2 + 10\epsilon \tag{14}$$

where $x_1, x_2, \epsilon \sim \mathcal{N}(0, 1)$. x_1 and x_2 are the two circumstance variables and ϵ is the part of inequality due to other factors (e.g. effort). This setting gives a true value of the inequality measure θ_{IOP} of slightly above 0.5.

For each sample of 1000 individuals, the method proposed by Ferreira and Gignoux (2013) is applied. First the method is used with only one indicator variable, followed by the estimation using each time one additional indicator in the factor analysis until reaching the maximum number of included variables. I use factor analysis (principal component factor) to aggregate across indicator variables y_j. The estimated $\hat{\theta}_{IOP}$ is then compared to the true value based on the OLS regression of m_i on x_1 and x_2.

Figure 2: $\frac{\hat{\theta}_{IOP}}{\theta_{IOP}}$ in function of the number of included indicator variables

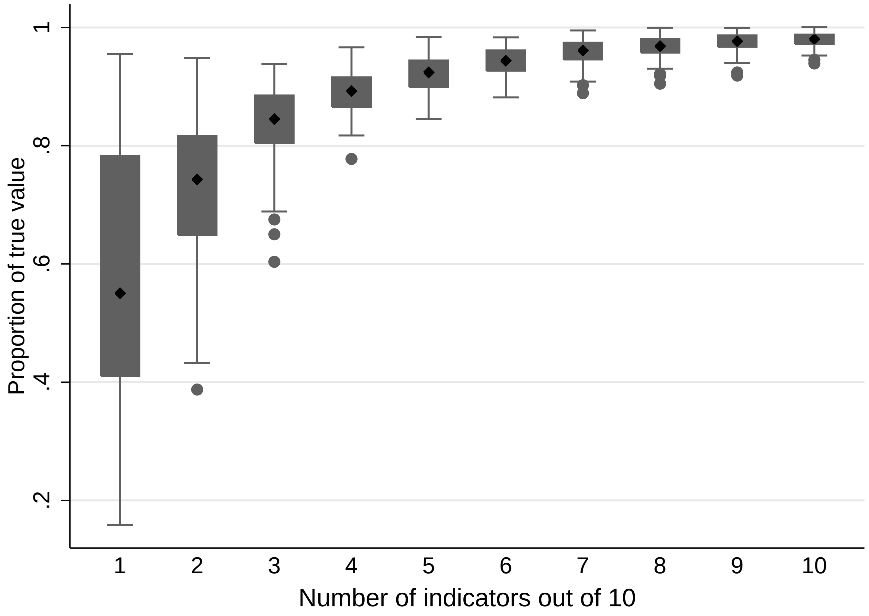


Figure 2 displays the main results of this simulation. On the vertical axis I report the ratio $\frac{\hat{\theta}_{IOP}}{\theta_{IOP}}$ and on the horizontal axis the number of included indicator variables. Two main observations can be drawn from these box plots:

1. The average bias is steadily reduced when including more indicator variables. The additional gain in terms of bias reduction is decreasing with the number of indicator variables. Once about half of the variables included, the bias reduction becomes very small.

2. The variation of the bias is very large for the one dimension case and becomes smaller the more dimensions are included. This stems from the fact that the underlying factor can be estimated more accurately when including more dimensions. Note that even with one indicator it is possible to get almost unbiased estimates. However, due to the large variation, very large biases cannot be excluded. The distribution of the bias becomes much narrower, even when only a few indicators (e.g. 3 or 4) are used.

This simulation shows that in the case of continuous variables resulting from a utility maximization process with heterogeneous preferences, the use of factor analysis can considerably reduce the bias, even when using only 3 or 4 indicators. The marginal decrease of the bias becomes a lot smaller when adding more and more dimensions. This is good news for empirical research where it might be difficult to collect a large number of indicator variables.

Simulation for the case where $Cov(Y, \nu) \neq 0$

In the previous simulation I illustrated the bias introduced by a proxy measurement error that is uncorrelated with the outcome. Even though in many cases we might arguably have this situation, it is important to know how the problem changes if ν is not independent of Y. Using a similar setting as in the previous simulation but focusing this time on the correlation between Y and ν allows me to illustrate the behavior of the bias. The simulation is based on the following relationship between the circumstances x and the outcome Y:

$$Y = x_1 + x_2 + 2\epsilon \tag{15}$$

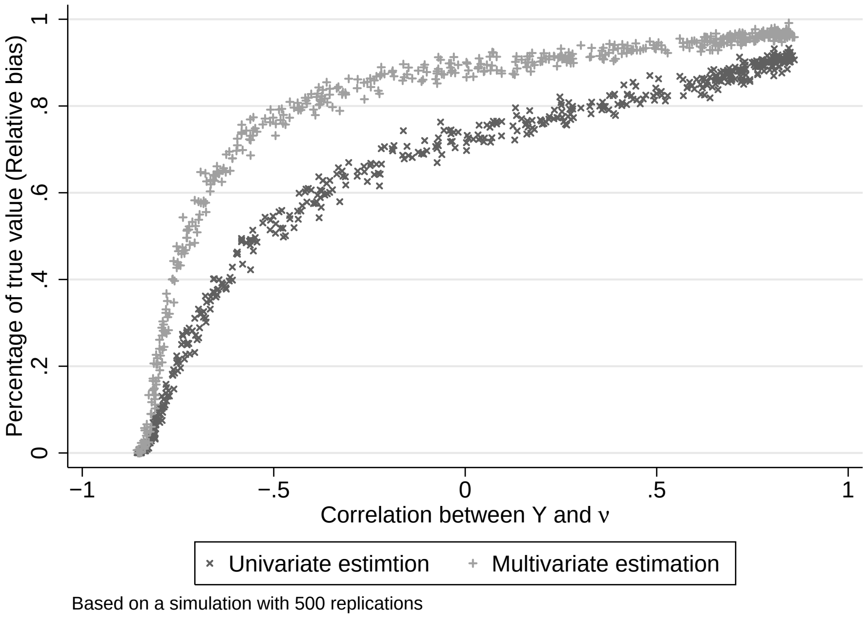
where $x_1, x_2, \epsilon \sim \mathcal{N}(0, 1)$ and the following relationship between the proxy measurement error term and Y

$$\nu = \alpha(Y - \bar{Y}) + \xi \tag{16}$$

with $\xi \sim \mathcal{N}(0, 1.5)$, $\xi \perp Y$ and $\alpha \sim U_{[-1,1]}$. $\bar{Y}$ denotes the mean of Y and is included to ensure the mean of ν to be equal to zero. The simulation includes 500 runs generating several observations for each range of the correlation between Y and ν.

Figure 3 displays the relative bias of $\hat{\theta}_{IOP}$ in function of the correlation between the true outcome variable Y and the proxy measurement error ν. The graphic displays the estimation of the univariate method and the multivariate version based on a factor analysis with only three indicator variables. A first lesson from this illustration is that the problem of underestimation in this method is particularly important when the proxy measurement error is negatively correlated

Figure 3: $\hat{\theta}_{IOP}$ as proportion of θ_{IOP} in function of the correlation between Y and ν



Based on a simulation with 500 replications

with the outcome variable. It becomes relatively less of an issue when the correlation is positive. The intuition of this finding is as follows. Assume we have an outcome where a higher value is better. Adding a negatively correlated proxy measurement error makes the best worse off and the worst better off, reducing therefore the inequality *per se*. The explanatory power of such a model becomes smaller and this corresponds exactly to the definition of our inequality of opportunity index.

The second lesson we can draw from this graph is that the multivariate estimation performs always better, independently of the level of correlation. In practice the region around zero correlation is the most relevant. The reduction of the bias is substantial, despite the fact that in this simulation I used only three indicator variables to estimate the latent concept.

4.2 Empirical application: Inequality of opportunity in economic well-being

The goal in this illustration is to estimate the level of inequality of opportunity in economic well-being, which is a broader concept than only income or wages.

4.2.1 Data

I make use of the Mexican Social Mobility Survey 2006 (MSMS) which was especially designed to study phenomena of intergenerational transmission. The survey is nationally representative and covers 25 to 64 year old adults living in Mexico. The choice of the respondent within a selected household gave priority to men, as women were only interviewed if no male was living in the household. This makes the female respondents likely not to come from representative households. I present the analysis for both the full sample and a sample limited to men in order to account for this data issue.

Table 2: Summary statistics

	Only males (N=4690)		Males & females (N=5277)	
	Mean	StdDev	Mean	StdDev
Circumstances				
Father's schooling in years	3.760	(3.904)	3.780	(3.932)
Mother's schooling in years	3.299	(3.513)	3.301	(3.544)
Parents owned house	0.776		0.772	
Estimated socioeconomic situation at age of 14	4.249	(2.219)	4.252	(2.236)
Belongs to an indigenous group	0.030		0.028	
Gender dummy	0.000		0.111	
Indicators				
Log income	8.083	(0.760)	8.052	(0.775)
Number of bedrooms per capita	0.708	(0.496)	0.738	(0.540)
Number of goods at home [0,14]	7.400	(2.565)	7.392	(2.578)
Schooling in years	8.198	(4.437)	8.170	(4.465)
Can read/write	0.957	(0.202)	0.955	(0.207)
Current HH owns a car	0.582	(0.766)	0.552	(0.756)
Quality of house				
high	3.80 %		3.83 %	
medium	45.07 %		45.25 %	
low	51.13 %		50.92 %	

Table 2 presents the summary statistics of circumstances and indicators for economic well-being for both samples.

The circumstance variables I include in this study refer all to characteristics beyond the control of the individual. Parental education is measured in years of education of both mother and father. A dummy capturing if the parents owned their house and a self-reported socioeconomic level of the household (compared to other households) are included to describe the general socioeconomic situation during childhood. To account for racial discrimination I include a dummy for indigenous people. Finally, in the full sample the gender dummy is included as an additional circumstance.

Regarding the indicators of economic well-being in the large sense I use the current log income, the number of 14 specific consumption goods present in the household, the number of bedrooms per capita, an indicator of the presence of a car and a qualitative indicator for the quality of the house to capture the purely economic well-being. Additionally I include education and literacy since they give individuals more freedom and opportunities to influence the economic well-being and are themselves part of a larger definition of well-being.

4.2.2 Results

Table 3 displays the estimates of θ_{IOP} for both samples for a series of univariate and composite dependent variables. For each dependent variable four estimates are presented according to the inclusion of women in the sample or the prior correction of the dependent variable for cohort effects[5]. This correction for the age effect might be reasonable because a potentially large part of the variation in economic well-being is due to age effects, especially when taking into account elements of wealth, rather than just income.

Table 3: Estimation of inequality of opportunity

	Only males		Males & females	
Standardized by age	no	yes	no	yes
Dependent variable				
Log income	0.167	0.152	0.183	0.165
Number of goods at home [0,14]	0.287	0.300	0.287	0.299
Number of bedrooms per capita	0.057	0.073	0.089	0.104
Factor analysis (ml)	0.353	0.293	0.360	0.301
Factor analysis (pf)	0.383	0.366	0.390	0.371
Factor analysis (pcf)	0.370	0.353	0.377	0.359
Factor analysis (ipf)	0.395	0.374	0.401	0.379
Polychoric factor analysis	0.396	0.358	0.403	0.364

Notes: The table shows the value of θ_{IOP} using each time the same set of circumstances and the same sample within a column. For the factor analysis, the following abbreviations are used: pf=principal factor, pcf=principal component factor, ipf=iterated principal factor, ml=maximum-likelihood factor

While the first three rows represent the standard univariate case where a proxy is regressed on circumstances, in the last 5 rows I use different factor analysis techniques to aggregate all the indicator variables mentioned in Table 2. This composite variable is then regressed on circumstances. The reported statistic of these regressions is the R^2, which corresponds to the definition of θ_{IOP} in equation (3). Under the assumption that we really want to measure inequality of opportunity in the three indicated variables, these figures should be unbiased with respect to the proxy measurement error bias discussed earlier[6].

In contrast, if these variables are supposed to proxy inequality of opportunity in economic well-being, the problem of proxy measurement errors arises. The large differences among the three univariate estimates - ranging from 0.05 to 0.3 approximately - are a first way to look at it. By using a larger number of indicators and aggregating them into a measure of the latent economic well-being, the estimated inequality of opportunity index goes up to almost 0.4. With respect to the log income - which is normally seen as a good indicator - the estimated inequality of opportunity more than doubles in almost all samples. Depending on the aggregation method,

[5] The standardization is done following O'Donnell et al. (2008, p. 61), where the standardized variable is given by: $Y^{std} = Y_i - E[Y_i|age_i] + E[Y]$, where $E[Y_i|age_i]$ is the conditional expected outcome of individual i given his/her age and $E[Y]$ is the unconditional expected outcome for the population.

[6] Other biases might prevail.

slightly different values are obtained. However, they are all much closer to each other than the univariate case. The most adequate latent variable model in the list is the polychoric factor analysis, as it takes into account the non-continuity of some indicator variables.

Assuming that we correctly estimate the underlying factor, circumstances beyond the control of the individual account for about 40% of total inequality, which is a massive amount. Moreover, as Ferreira and Gignoux (2013) mention and due to missing circumstances in the regressions, these figures remain lower bound estimates of the true inequality of opportunity.

Table 4: Estimation of inequality of opportunity

	Only males		Males & females	
Standardized by age	no	yes	no	yes
Indicator variables in polychoric FA				
Only log income	0.167	0.152	0.183	0.165
+ Number of bedrooms per capita	0.159	0.165	0.161	0.168
+ Number of goods at home	0.296	0.296	0.299	0.299
+ Quality of house	0.317	0.326	0.319	0.328
+ Ownership of car	0.296	0.308	0.302	0.313
+ years of education and literacy	0.396	0.358	0.403	0.364

Notes: The table shows the value of θ_{IOP} using each time the same set of circumstances and the same sample within a column. Polychoric factor analysis is used on sequentially included indicator variables

Now let us see which indicator variables are responsible for the changes. Table 4 displays the same statistics as Table 3, but for different sets of indicator variables. The aggregation method is always the polychoric factor analysis. The idea is as follows: by adding more variables, the underlying concept should be measured more appropriately. The general pattern is that by including more variables, the estimated amount of inequality of opportunity rises. Nevertheless there are exceptions. For instance, the inclusion of the car ownership indicator slightly reduces the measured level of inequality. It could be argued that this variable might overestimate the level of economic well-being for the lower middle class and underestimates it for the very rich people. Indeed, plotting the average number of cars against the log income shows that for a large range of incomes the numbers are quite stable. Only for the upper class an increase and for the very poor a decrease in the average number of cars is observed. This example shows that adding variables does not mechanically improve the estimation of an underlying concept. Hence, the choice of the correct variables must be made with caution. Moreover, when using factor analysis, it is important to make sure that only one underlying factor is present. In case of finding more than one factor, the variables in the model should be reconsidered more generally. As I discussed at the end of section 3, more than one factors might suggest that the problem does not stem from imprecision but from the fact that the indicator variables are proxies for different underlying concepts.

A final point I would like to highlight with respect to Table 4 is the effect of the inclusion of education and literacy as indicator variables. The factor analysis clearly shows the presence of a unique underlying factor[7] and the level of inequality of opportunity increases substantially from about 0.3 to about 0.35 for the age standardized estimations. This result suggests that the inclusion of education in the measurement of the broader concept of economic well-being is useful.

To conclude the discussion of this illustration, I would like to stress the fact that this illustration does not invalidate the use of a single indicator to measure inequality of opportunity in general, but it shows that one has to be very precise about the measured concept. The use of a single indicator as a proxy for a broader concept might be very risky. Researchers interested in such a broader concept might want to consider an analysis with multiple indicators instead of multiple analyses with univariate indicators.

5 Conclusion

In this paper I argue that recent methods to estimate inequality of opportunity are likely to be downward biased when the used outcome variable is a proxy for an underlying concept. I define the proxy measurement error as a deviation of the observed variable with respect to the underlying concept we actually want to measure.

Proxy measurement errors can arise from simple misreporting of data in a survey. However, there are also more conceptual reasons justifying the presence of measurement errors in empirical research. The capability approach offers such an example. Capabilities are by definition latent and only functionings can be observed. Using a single variable of functionings to approximate the underlying capabilities introduces precisely a proxy measurement error.

An underestimation of the part of inequalities due to opportunities and therefore an overestimation of the part due to effort and luck can have severe policy implications. While inequality of opportunity can and should be tackled by policies, much less consensus exists about the need of policy interventions when inequality is due to different effort levels. When we systematically underestimate the part due to circumstances, we probably underestimate the problem at stake and in consequence policy responses might be too weak.

I propose a simple solution to reduce the downward bias substantially. Instead of using only one proxy variable, I suggest to use several at a time and to estimate first the underlying latent variable. The exact nature of aggregation largely depends on the context. In this paper I make use of factor analysis to estimate the latent concept. Once the latent concept is estimated, I suggest using it as dependent variable in the method proposed by Ferreira and Gignoux (2013). In a simulation exercise I show that the bias can be reduced substantially even with some few indicator variables. An illustration using Mexican data shows that the measure of inequality more than doubles in most settings when using the multivariate estimation rather than the log income. The analysis of inequality of opportunity through various univariate proxy variables will

[7] The eigenvalue of the first factor is 2.83 and of the second factor 0.72, hence only 1 factor should be retained.

not allow us to estimate the true inequality of opportunity, because each proxy has a component of proxy measurement error causing an underestimation of univariate analyses.

The role of proxy measurement errors is rarely discussed in the empirical literature. However, this paper shows that in some cases the consequences might be substantial. It is therefore advisable to discuss the possible consequences of proxy measurement errors in empirical studies. This is particularly true when applying multidimensional concepts like the capability approach. In such cases it is important to consider multivariate techniques and not exclusively rely on univariate proxy variables, since this might introduce systematic biases to the estimation.

References

Anand, Paul, Jaya Krishnakumar, and Tran Ngoc Bich, "Measuring welfare: Latent variable models for happiness and capabilities in the presence of unobservable heterogeneity," *Journal of Public Economics*, 2011, *95* (3), pp. 205–215.

Checchi, Daniele and Vito Peragine, "Inequality of Opportunity in Italy," *Journal of Economic Inequality*, 2010, *8*, 429–450.

Ferreira, Francisco H.G. and Jérémie Gignoux, "The Measurement of Inequality of Opportunity: Theory and an Application to Latin America," *The Review of Income and Wealth*, 2011, *57* (4), pp. 622–657.

_ **and** _ , "The Measurement of Educational Inequality: Achievement and Opportunity," Forthcoming in The World Bank Economic Review. Advance Access published February 20, 2013. doi: 10.1093/wber/lht004 2013.

Gibson, John and Bonggeun Kim, "How reliable are household expenditures as a proxy for permanent income? Implications for the income-nutrition relationship," *Economics Letters*, 2013, *118* (1), pp. 23–25.

Hair, Joseph F., Rolph E. Anderson, Ronals L. Tatham, and William C. Black, *Multivariate Data Analysis*, Prentice Hall, Upper Saddle River, NJ, 1998.

Krishnakumar, Jaya, "Going Beyond Functionings to Capabilities: An Econometric Model to Explain and Estimate Capabilities," *Journal of Human Development and Capabilities*, 2007, *8* (1), 39–63.

_ **and Paola Ballón**, "Estimating Basic Capabilities: A Structural Equation Model Applied to Bolivia," *World Development*, 2008, *36* (4), 992–1010.

Muthén, Bengt, "Contributions to Factor Analysis of Dichotomous Variables," *Psychometrika*, 1978, *43* (4), pp. 551–560.

O'Donnell, Owen, Eddy van Doorslaer, Adam Wagstaff, and Magnus Lindelow, *Analyzing Health Equity Using Household Survey Data - A Guide to Techniques and Their Implementation*, The World Bank, Washington D.C., 2008.

Paes de Barros, Ricardo, Francisco Ferreira, José Molinas Vega, and Jaime Saavedra Chanduvi, *Measuring Inequality of Opportunity in Latin America and the Caribbean*, The World Bank, Washington DC., 2009.

Roemer, John E., *Equality of Opportunity*, Harvard University Press, Cambridge, 1998.

Sen, Amartya, *Development as Freedom*, new ed., Oxford University Press, 2001.

Yalonetzky, Gaston, "A dissimilarity index of multidimensional inequality of opportunity," *Journal of Economic Inequality*, 2012, *10* (3), pp.343–373.

A Proof

Let us assume for simplicity the case of only one circumstance variable x. The R^2 equals in this case the squared correlation between the dependent and the independent variable, which is for the case of y simply:

$$\theta = \rho_{xy}^2 = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}} \tag{17}$$

where σ_{xy} denotes the covariance between x and y and σ_{xx} is the variance of x. We can rewrite the ratio between the estimated and the true inequality of opportunity measure as follows:

$$\frac{\hat{\theta}}{\theta} = \frac{\frac{\sigma_{xz}^2}{\sigma_{xx}\sigma_{zz}}}{\frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}} = \frac{\sigma_{xz}^2}{\sigma_{xy}^2}\frac{\sigma_{yy}}{\sigma_{zz}} \tag{18}$$

Now, let us take the definition of ν used in the simulation to generate positive and negative correlations between ν and y.

$$\nu = \alpha(y - \bar{y}) + \xi \tag{19}$$

where ξ is white noise. Therefore, z is defined as:

$$z = (1 + \alpha)y - \alpha\bar{y} + \xi \tag{20}$$

It follows that the variance of z can be written as:

$$\sigma_{zz} = (1 + \alpha)^2\sigma_{yy} + \sigma_{\xi\xi} \tag{21}$$

Using this definition of z, we can rewrite equation (18) as follows:

$$\frac{\hat{\theta}}{\theta} = \frac{(E[xz] - E[x]E[z])^2}{(E[xy] - E[x]E[y])^2} \times \frac{\sigma_{yy}}{(1+\alpha)^2\sigma_{yy} + \sigma_{\xi\xi}} \tag{22}$$

We can now substitute z by its definition given in (20)

$$\frac{\hat{\theta}}{\theta} = \frac{(E[x\,\{(1+\alpha)y - \alpha\bar{y} + \xi\}] - E[x]E[z])^2}{(E[xy] - E[x]E[y])^2} \times \frac{\sigma_{yy}}{(1+\alpha)^2\sigma_{yy} + \sigma_{\xi\xi}} \tag{23}$$

and use the fact that $E[y] = E[z]$

$$\frac{\hat{\theta}}{\theta} = \frac{(E[x\,\{(1+\alpha)y - \alpha\bar{y} + \xi\}] - E[x]E[y])^2}{(E[xy] - E[x]E[y])^2} \times \frac{\sigma_{yy}}{(1+\alpha)^2\sigma_{yy} + \sigma_{\xi\xi}} \tag{24}$$

$$= \frac{((1+\alpha)(E[xy] - E[x]E[y]))^2}{(E[xy] - E[x]E[y])^2} \times \frac{\sigma_{yy}}{(1+\alpha)^2\sigma_{yy} + \sigma_{\xi\xi}} \tag{25}$$

$$= \frac{(1+\alpha)^2\sigma_{yy}}{(1+\alpha)^2\sigma_{yy} + \sigma_{\xi\xi}} \leq 1 \tag{26}$$

Since the variance of the white noise term is necessarily positive, the denominator is at least as big as the numerator. Hence, we should expect a downward bias for as long as y and z are not perfectly correlated.