New Estimates of Inequality of Opportunity Across European Cohorts

(and Some Insights on the Long-term Impact of Educational Policy)

Francesco Andreoli

(Luxembourg Institute of Socio-Economic Research and University of Verona)

Alessio Fusco

(Luxembourg Institute of Socio-Economic Research)

Iryna Kyzyma

(Luxembourg Institute of Socio-Economic Research)

Philippe Van Kerm

(Luxembourg Institute of Socio-Economic Research and University of Luxembourg)

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
Session 13: Inequalities in Opportunity for Income, Education and Health I
Time: Wednesday, August 25, 2021 [16:30-18:30 CEST]

New estimates of inequality of opportunity across European cohorts (and some insights on the long-term impact of educational policy)*

Francesco Andreoli[1, 2]
Alessio Fusco[1]
Iryna Kyzyma[1]
Philippe Van Kerm[1, 3]

[1] *Luxembourg Institute of Socio-Economic Research (LISER)*
[2] *University of Verona*
[3] *University of Luxembourg*

Conference paper prepared for the 36th IARIW Virtual General Conference
e-Oslo, Norway, August 23-27, 2021

This version: August 6, 2021

*This research is part of the NORFACE project IMCHILD (The Impact of Childhood Circumstances on Individual Outcomes over the Life-Course) supported, for Luxembourg, by the Fonds National de la Recherche Luxembourg (contract INTER/NORFACE/16/11333934). The empirical analysis uses micro-data from Eurostat, the EU Statistics on Income and Living Conditions [EUSILC2004-2018V.1]. The responsibility for all conclusions drawn from the data lies entirely with the authors. We are grateful for comments from Filip Novokmet and participants at the Aix-Marseille School of Economics Big Data and Econometrics seminar, IMCHILD annual workshops in Munich, Stockholm and Luxembourg, the DIAL Mid-Term Conference (University of Turku), the 2020 Welfare Economics workshop at the University of Namur, the Conference of the Economics Section of the German Association for East European Studies (Berlin), and the 2021 ECINEQ conference.

Abstract

This paper provides a set of new estimates of inequality of opportunity (IOp) in Europe, using the European Union Statistics on Income and Living Condition (EU-SILC). Unlike previous research, we estimate inequality of opportunity within birth cohorts, which we argue is the most appropriate population level for inequality of opportunity analysis. Most IOp measures require estimation of the conditional distribution of the outcome of interest given circumstances. With multiple circumstances and the sample sizes available in EU-SILC, we use distribution regression methods combined with local kernel weighting and show how these can be used to estimate a large set of IOp measures. Endowed with cohort-level estimates of IOp, we finally examine the relationship between educational policy variables measured at the time of parental education and offspring generation inequality of opportunity in adulthood. We find a negative relationship between the duration of compulsory education of the parents and IOp among offspring, but the relationship is not very strong.

Keywords: inequality of opportunity; educational policy; distribution regression; semi-parametric regression; bias correction; EU-SILC

JEL classification: C14; D63; D31; I24; I28

1 Introduction

This paper provides new estimates of inequality of opportunity (IOp) across European countries, following up on a series of recent papers that have exploited data from the European Union Statistics on Income and Living Conditions, e.g., Dunnzlaff et al. (2011), Checchi et al. (2016), Andreoli and Fusco (2019*b*), Brzezinski (2020), Ramos and Van de gaer (2021).

The inequality of opportunity theory postulates that the unequal distribution of an outcome of interest (say, earnings, wealth or income) can have legitimate sources—factors that are under the control of the recipients, such as a person's effort or hard work—and illegitimate sources—(dis-)advantages that arise from factors beyond a person's responsibility, such as a person's gender, ethnicity, socioeconomic background, etc. ('circumstances'). Differences in outcome between two people that arise from differences in circumstances are unfair and may justify compensation. Differences in outcome between two people that arise from legitimate sources are fair and should not be compensated—a person should see reward for their efforts. The full normative and theoretical foundations of the inequality of opportunity measurement are reviewed in Roemer and Trannoy (2015).

Operationalization of these concepts typically takes the form of indices of inequality of opportunity. An inequality of opportunity index (or, more formally, a functional) is a mapping of the joint distribution of the outcome of interest Y, of individual circumstances C, and of individual effort E onto a scalar measure reflecting the degree of "unfair inequality" in the distribution of the outcome Y. Effort is however rarely observed. The majority of commonly-used indices are therefore simpler mappings of the joint distribution Y and C onto scalar measures of inequality of opportunity, $\mathrm{IOp}(Y, C)$.

Unlike previous analyses, we derive estimates by *birth cohort* and propose an approach to do so even with moderate sample sizes as available in EU-SILC. IOp studies have been surprisingly mixed in their treatment of age. A person's age is typically either ignored or included as a circumstance. Including age as a circumstance supposes that age is a source of unfair inequality that should be compensated. While it is true that one's age is beyond one's control and has an impact on the outcome of interest—thus satisfies a criteria for being considered a circumstance—it can hardly be considered to represent an *unfair* advantage. In the long-run, from a life-course perspective, age is (mostly) equally shared across the population and therefore does not represent a persistent advantage. Other studies therefore simply ignore age altogether. But this omission may confound the effect of circumstances when circumstances are correlated with age (e.g. parental education). Some studies therefore focus on a limited age range (e.g., 30–50). We trust that IOp measures should be best estimated within birth cohorts of individuals. We would want to estimate how much of pre-adulthood circumstances influence inequality in adulthood outcomes for people born in approximately the same year (so that circumstances are measured in the same period of time). (Ideally, adulthood outcomes should be measured with data over the life-course of individuals. However, this is only rarely available and so not considered here; see Aaberge et al. (2011).) With cross-section data, or repeated cross-section data, this means that we should estimate IOp *conditional* on birth cohort (or age in the presence of a single cross-section). We are aware of only cross-country study–not using EU-SILC however–that adopts this perspective and attempts to estimate cohort-level IOp measures (Bussolo et al., 2019).

Endowed with IOp estimates for a large set of countries and multiple cohorts, we examine if educational policies affecting parental education influence inequality of opportunity in the offspring generation. Most notably, we examine if the duration of compulsory schooling influences inequality of opportunity in the next generation. We observe an overall negative relationship between the duration of compulsory schooling when parents were children and inequality of opportunity among their offsprings: countries in which children were kept longer in schools tend to display lower inequality of opportunity one generation down the road. This is robust to the choice of IOp indicator or the set of circumstances considered in IOp estimation. This is however mostly driven by variations between countries in IOp and educational policies: fixed effects estimation that only exploit variation in policies over time (that is, across cohorts) do not reveal any clear association. These variations over time are however much smaller than variations across countries. We also observe large variations in IOp among countries and cohorts that impose similar minimum school leaving age policies.

Section 2 presents the four IOp indices that we examine. We explicitly express these measures – known in the literature as ex ante and ex post, direct and indirect IOp indices (Ramos and Van de gaer, 2016) – as functionals of the joint distribution of circumstances and income. We describe our strategy to obtain cohort-level indicators and then detail a semi-parametric distribution regression estimation strategy. We also show how we handle the upward bias in estimation of IOp measures. Section 3 describes the EU-SILC data that we examine, including a new module about 'the transmission of disadvantage' collected in 2019 which, to the best of our knowledge, we are first to exploit for IOp estimation. Section 4 presents our new estimates of inequality of opportunity. Section 5 finally examines the relationship between IOp indicators and educational policy parameters.

2 Methods

2.1 The measurement of inequality of opportunity: principles

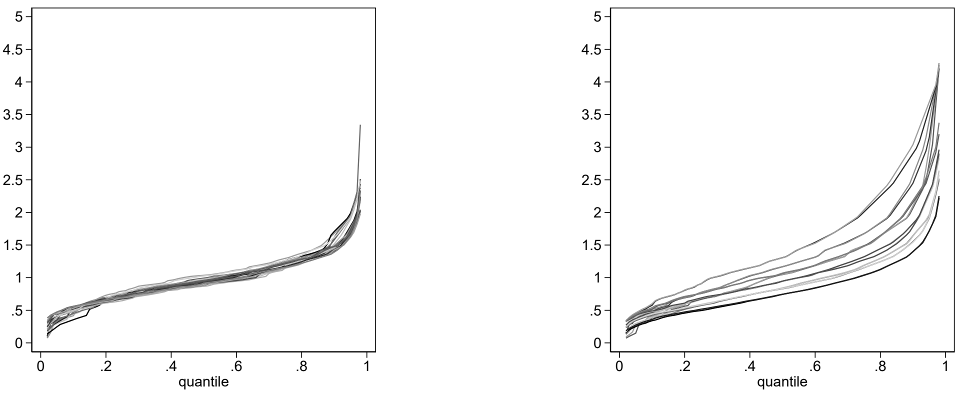
A range of alternative inequality of opportunity indices have been proposed in the literature. Ramos and Van de gaer (2016) distinguish between *direct* and *indirect* measures, between *ex ante* and *ex post* measures and between parametric and non-parametric measures.

The process of deriving IOp measures is typically a two-step process. In a first step, one constructs from the joint distribution (Y, C) a counterfactual outcome distribution that would prevail if we eliminated all differences in outcome due to effort (all inequality is due to circumstances) or, alternatively, a counterfactual distribution that would prevail if we eliminated all inequality due to circumstances. In the first case, we have a counterfactual distribution that only exhibits unfair inequality, so, in a second step, one can measure inequality of opportunity by applying an inequality functional on the counterfactual distribution. This is the *direct* approach. In the second case, we have a counterfactual distribution that only exhibits fair inequality. To measure inequality of opportunity (unfair inequality), we can therefore take, in the second step, the difference between (total) inequality measured in the overall outcome distribution and inequality measured in the counterfactual distribution, according to any usual inequality functional. This is the *indirect* approach.

The distinction between *ex ante* and *ex post* approaches relates to how the counterfactual distribution are constructed. The *ex ante* approaches first makes a welfare evaluation of the conditional outcome distributions given circumstances C which is usually the mean outcome given circumstance C. The counterfactual distributions are then constructed either by assigning everyone the welfare value corresponding to their circumstances (direct approach) or by scaling everyone's income up or down according to the relative deviation between the welfare value of one's circumstances and the average welfare values across all circumstances (indirect approach). The *ex ante* approach does not require observing or imputing a person's effort. The *ex post* approach first classifies individuals by effort level. When effort is not observed, it is proxied by the rank of the person's outcome in the distribution of among individuals sharing the same circumstances. Counterfactual distributions are then constructed by scaling every person's outcome according to the average outcome of all individuals exerting the same effort level. This leads to a counterfactual distribution reflecting variations in outcome according to circumstances only, and therefore to a direct measure of IOp. Alternatively, counterfactual distributions are constructed by assigning to every person the average outcome of all individuals exerting the same effort level. This gives a distribution reflecting legitimate inequality only, the inequality in which can be deducted from total inequality to derive an indirect measure.

Figure 1: Conditional quantile functions of mean-normalised equivalised household income for eighteen combinations of gender, mother and father education levels in Denmark and Poland. Estimates from the EU Statistics on Income and Living Conditions, 2011.

(a) Denmark (b) Poland



In short, direct measures evaluate IOp by the inequality in the “value” of individual circumstances, $I(y^C)$. Indirect measures evaluate IOp by the part of total inequality that is not explained by effort, $I(y) - I(y^{EO})$. Ex ante indicators measure the value of circumstances y^Cor of effort y^{EO} before a person’s effort level is revealed. Ex post indicators measure the value of circumstances y^Cor effort y^{EO} after a person’s effort level is revealed. The variations in the normative content of these alternative approaches are discussed in Ramos and Van de gaer (2016), Ramos and Van de gaer (2021).

To fix ideas, Figure 1 shows estimates of conditional quantile functions of (mean-normalised) equivalised household income for eighteen combinations of gender, mother and father education levels in Denmark and Poland estimated from the EU Statistics on Income and Living Conditions. Each line is the quantile function for a particular configuration of circumstances c. It shows the income at each (fractional) rank within the group c. When we do not observe effort directly, this rank proxies the effort level. Therefore the vertical dispersion *between* lines reflects variation unfair inequality—differences in income for individuals exerting the same level of effort. And the variation in income levels along the quantiles from low to high effort levels—the ‘slope’ of the quantile functions—reflects legitimate inequality. In the two examples shown, Poland appears to exhibit both more unfair inequality and fair inequality. (The overall inequality is driven by the overall dispersion of the points along the quantile functions.)

2.2 Derivation of IOp measures

We will consider four IOp measures. As we show below, we will rely on semi-parametric estimation of conditional quantiles and conditional distribution models and therefore formally express all four measures in terms of these primitives.

To do so, let us first denote by $h : \Omega \times \mathbb{R}^{++} \mapsto \mathbb{R}^{+}$, the joint density distribution of circumstances and income in the population. Ω denotes the (multidimensional) domain of definition of circumstances. We assume incomes to be defined over $\mathbb{R}^{++}$ so as to be able to use the full set of standard inequality functionals, but this can be relaxed. Let $\mathcal{H}$ denote the set of all possible such density functions.

Let us also adopt the following notation. The marginal distribution function of income is denoted $F : \mathbb{R}^{++} \mapsto [0, 1]$ (and let $\mathcal{F}$ denote the set of all possible distribution functions). It can be expressed from h as

$$F(y) = \int_{\Omega} \int_{0}^{y} h(c, x) dx\, dc.$$

The conditional distribution function of income given circumstances $C \in \Omega$ is denoted $F_C : \mathbb{R}^{++} \mapsto [0, 1]$ and is obtained from h, again using elementary probability theory, as

$$F_C(y) = \int_{0}^{y} \frac{h(C, x)}{g(C)}\, dx$$

where $g : \Omega \mapsto \mathbb{R}^{+}$ denotes the marginal density distribution of circumstances:

$$g(C) = \int_{\mathbb{R}^{++}} h(C, x) dx.$$

We will also use the conditional quantile functions defined by the usual generalized inverse function, $Q_C : [0, 1] \mapsto \mathbb{R}$,

$$Q_C(p) = F_C^{-1}(p) = \inf \left\{ y \in \mathbb{R} : F_C(y) \geq p \right\}.$$

Let us then denote the expectation functional by $\mu : \mathcal{F} \mapsto \mathbb{R}$ and, for notational simplicity, the grand mean $m = \mu(F)$ and the conditional means $m_c = \mu(F_C)$. Finally, denote by $\theta : \mathcal{F} \mapsto \mathbb{R}$ a generic inequality functional (such as the Gini coefficient or the mean log deviation).

To derive IOp measures, we first define counterfactual income generation functions (CIGF) $\tilde{y} : \Omega \times \mathbb{R}^{++} \times \mathcal{H} \mapsto \mathbb{R}^{++}$ which associate a value to each income-circumstances pair (C, y) drawn from the joint distribution h:

$$\tilde{F}(y) = \int_{\Omega} \int_{\mathbb{R}^{++}} \mathbf{1} \left[\tilde{y}(c, x, h) \leq y \right] h(c, x) dx\, dc.$$

Measures of IOp are then obtained by applying inequality functionals to the distribution function of these counterfactual values.

We will use four CIGF. First, to obtain the standard direct ex ante (DEA) IOp measure, the CIGF is defined as

$$\tilde{y}_1(C, y, h) := \mu(F_C)$$

with resulting CDF $\tilde{F}_1$ and the DEA index is obtained by the functional

$$\theta^{\text{DEA}} := \theta(\tilde{F}_1).$$

Second, to obtain a direct ex post (DEP) measure, the CIGF is defined as

$$\tilde{y}_2(C, y, h) := y \frac{\mu(F)}{\int_\Omega Q_c\left(F_C(y)\right) g(c)dc}$$

with CDF $\tilde{F}_2$ and the DEP index

$$\theta^{\text{DEP}} := \theta(\tilde{F}_2).$$

Third, to obtain the indirect ex ante (IEA) measure, the CIGF is defined as

$$\tilde{y}_3(C, y, h) := y \frac{\mu(F)}{\mu(F_C)}$$

with CDF $\tilde{F}_3$ and the IEA index

$$\theta^{\text{IEA}} := \theta(F) - \theta(\tilde{F}_3).$$

Fourth, to obtain the indirect ex post (IEP) measure, the CIGF is defined as

$$\tilde{y}_4(C, y, h) = \int_\Omega Q_c\left(F_C(y)\right) g(c)dc$$

with CDF $\tilde{F}_4$ and the IEP index

$$\theta^{\text{IEA}} = \theta(F) - \theta(\tilde{F}_4).$$

2.3 Cohort-level IOp

As elaborated in the Introduction, we trust IOp indices are best measured at the cohort level. To define cohort-level measures, we just need re-define h to denote the joint density distribution of income, circumstances and year of birth in the population of interest, $h : \Omega \times \mathbb{R}^{++} \times B \mapsto \mathbb{R}^+$. We can then introduce a new term for the joint conditional density of circumstance and income conditional on birth year b $h_b : \Omega \times \mathbb{R}^{++} \mapsto \mathbb{R}^+$,

$$h_b(C, y) = \frac{h(C, y, b)}{\int_\Omega \int_{\mathbb{R}^{++}} h(c, x, b) dx dc}.$$

We thereby define indices of inequality for the cohort born in year b as θ_b^{DEA}, θ_b^{DEP}, θ_b^{IEA}, θ_b^{IEP} as above with function h_b used in place of h.

2.4 Estimation and inference

All our IOp measures of interest are expressed above in terms of (functionals of) marginal and conditional (counterfactual) distributions, or, equivalently, of conditional quantile processes conditional on circumstances. This naturally leads to an estimation framework based on semi-parametric distribution regression models for all the underlying components.

Assuming a dataset consisting of N tuples, $S_N := \left\{(y_i, C_i)\right\}_{i=1}^{N}$, where y_i is individual i adulthood income and C_i is a vector of circumstances, estimation proceeds in two steps. The first step consists in obtaining estimates for F_C, the conditional income distributions given circumstances, for all values of C_i observed in the data. With high-dimensional C and sample sizes typically found in surveys on income and living conditions, direct estimation of the empirical CDF for all observed combinations of circumstances is usually not feasible. This has led researchers to limit the dimensionality of the circumstance set – by combining circumstances into broad classes or by limiting focus on a few circumstances, such as parental education only –, or to focus solely on the 'direct ex ante' (DEA) index (which does not require estimation of the whole conditional income distributions F_C but only on conditional averages m_C). Neither strategy is desirable. We propose therefore to specify semi-parametric regression models that allow estimation of conditional distributions even with moderately high-dimensional circumstances. The price to pay is nothing more than the imposition of the additive, linear combination of circumstances inherent to the regression specification. This is however exactly as in the commonly-used regression-based approach to estimation of DEA indices à la (Ferreira and Gignoux, 2011). Furthermore, allowing for interactions and polynomial series in the specification of the model also makes that constraint as soft as one wants to (or as one's sample size allows). Specifying a fully interacted (saturated) model, for example, can make the specification entirely non-parametric and equivalent to calculating empirical CDFs for all possible configurations of circumstances. Endowed with estimates $\left\{\hat{F}_{C_i}\right\}_{i=1}^{N}$ for income distributions conditional on circumstances, the second step consists in generating counterfactual distributions and calculating the four IOP indices.

2.4.1 Estimation of conditional income distributions

We estimate conditional income distributions with the "distribution regression" estimator proposed in Foresi and Peracchi (1995) and for which inference is developed in Chernozhukov et al. (2013).

$$F_C(y) = \Pr[Y \leq y|C] = \Lambda(P(C)\beta_y) \tag{1}$$

where Λ is the logistic function (or any other binary choice model), β_y is a vector of model parameters, and $P(C)$ is a vector of (transformations) of circumstances—which may include interaction terms, polynomial series, etc. For a fixed y, equation (1) defines a standard binary logistic regression model. Parameters can be estimated by standard maximum likelihood.

To obtain an estimate of the entire conditional distribution, the binary choice model is simply estimated over a grid of values for $y \in \{y_1, ..., y_K\}$ spanning the domain of variation of incomes. Model parameter estimates $\{\hat{\beta}_y\}_{y=1}^K$ can then be used to generate predictions for the conditional distributions $\{\hat{F}_C(y)\}_{y=1}^K$ as per (1).

2.4.2 Estimation of counterfactual distributions and IOP measures

Using estimates of conditional distribution functions $\{\hat{F}_C(y)\}_{y=1}^K$, we can now construct the counterfactual distributions needed for estimation of IOP indices.

The first ingredient is the conditional quantile process which can be recovered by numerical inversion of the conditional distribution functions

$$\hat{Q}_C(p) = \hat{F}_C^{-1}(p)$$

over a grid of J with equally-spaced values $p \in \{p_1, ..., p_J\}$ with $p_j = \frac{j}{J+1}$. (In our application, we use standard linear interpolation methods but see Press et al. (2007) for alternative numerical inversion algorithms.)

The J equally-spaced conditional quantiles $\{\hat{Q}_C(p)\}_{p=1}^J$ can be treated as a pseudo-random sample drawn from $\hat{F}_C$. Accordingly, all required distribution functionals from there on – such as the conditional means $\mu(F_C)$ – can be obtained by applying standard estimators for unit-record data on the the J values $\{\hat{Q}_C(p)\}_{p=1}^J$, e.g.

$$\hat{\mu}(\hat{F}_C) = \frac{1}{J}\sum_{j=1}^{J} \hat{Q}_C(p_j).$$

A similar approach is used for unconditional functionals. The $J \times N$ vector of equally-spaced conditional quantile predictions stacked across sample observations, $\left\{\{\hat{Q}_{C_i}(p)\}_{p=1}^J\right\}_{i=1}^N$, is a pseudo-random sample drawn from F. Unit-record estimators can therefore be used on this stacked vector to estimate functionals such as $\mu(F)$ or unconditional quantiles $F^{-1}(p)$.[1]

The final ingredient required is an estimate of the average across the circum-

[1] As in Monte Carlo simulations, random sub-sampling of entries from the stacked vector can be used to ease the computational burden if necessary.

stance distribution of conditional quantiles, $\int_\Omega Q_c\left(F_C(y)\right)g(c)dc$. This is easily obtained as a sample average of predicted conditional quantiles:

$$\bar{Q}(p) = \sum_{i=1}^{N} \hat{Q}_{C_i}(p)$$

where, if p is not a point on the evaluation grid $\{p_1, ..., p_J\}$, $\hat{Q}_{C_i}(p)$ is obtained by linear interpolation across adjacent quantile predictions on the grid

$$\hat{Q}_{C_i}(p) = \left(\frac{p-\underline{p}}{\bar{p}-\underline{p}}\right)\hat{Q}_{C_i}(\underline{p}) + \left(\frac{\bar{p}-p}{\bar{p}-\underline{p}}\right)\hat{Q}_{C_i}(\bar{p})$$

with $\underline{p} = \max\left\{p_j \in \{p_1, ..., p_J\} : p_j > p\right\}$ and $\bar{p} = \min\left\{p_j \in \{p_1, ..., p_J\} : p_j \leq p\right\}$.

The last step is to calculate the sample counterparts of the four counterfactual incomes needed to obtain the counterfactual income distributions based on which the IOp measures are calculated. These sample counterfactual incomes are obtained, for all sample observations as

$$\begin{aligned}
\hat{y}_1(C_i, y_i) &= \mu(\hat{F}_{C_i}) \\
\hat{y}_2(C_i, y_i) &= y_i \, \frac{\mu(\hat{F})}{\bar{Q}\left(\hat{F}_{C_i}(y_i)\right)} \\
\hat{y}_3(C_i, y_i) &= y_i \, \frac{\mu(\hat{F})}{\mu(\hat{F}_{C_i})} \\
\hat{y}_4(C_i, y_i) &= \bar{Q}\left(\hat{F}_{C_i}(y_i)\right)
\end{aligned}$$

The four IOp measures are finally obtained from unit-record estimators of inequality functionals θ applied to the counterfactual sample incomes:

$$\hat{\theta}^{\text{DEA}} = \theta\left(\left\{\hat{y}_1(C_i, y_i)\right\}_{i=1}^{N}\right) \quad (2)$$

$$\hat{\theta}^{\text{DEP}} = \theta\left(\left\{\hat{y}_2(C_i, y_i)\right\}_{i=1}^{N}\right) \quad (3)$$

$$\hat{\theta}^{\text{IEA}} = \theta\left(\{y_i\}_{i=1}^{N}\right) - \theta\left(\left\{\hat{y}_3(C_i, y_i)\right\}_{i=1}^{N}\right) \quad (4)$$

$$\hat{\theta}^{\text{IEP}} = \theta\left(\{y_i\}_{i=1}^{N}\right) - \theta\left(\left\{\hat{y}_4(C_i, y_i)\right\}_{i=1}^{N}\right) \quad (5)$$

2.4.3 Cohort estimates

Estimation of IOp within cohorts using standard cross-section survey data such as EU-SILC is hampered by sample size limitations which may cap the number of circumstances that one can consider. However, semi-parametric empirical specifications are possible.

Derivation of our cohort-level estimates proceeds exactly as above, except for the application of kernel weights at all stages. Birth cohorts are determined by the reported year of birth of the respondent. To avoid handle the potentially small sample sizes, we do not partition the sample in non-overlapping cohort groups but instead use a kernel smoothing approach.

For estimation of IOp of birth cohort b, all sample respondents are assigned a kernel weight

$$w_i(b) = \frac{1}{h} K\left(\frac{b_i - b}{h}\right)$$

where K is the Epanechnikov kernel. and h is a window width around birth year b and b_i is the year of birth of observation i. The Epanechnikov weight means (i) that all sample observations whose birth year fall within a window $(b - h, b + h)$ are used to estimate IOp at b – $\left\{i \in \{1, ..., N\} : |b_i - b| < h\right\}$ – and (ii) that the closer is b_i to b, the higher is the weight given observation i throughout all estimation stages. We set $h = 5$ in our application, but will report sensitivity to variation in the bandwidth. Having determined the weights, all estimation steps are as above.

2.4.4 Upward bias correction

Empirical estimates of IOp based on survey data such as EU-SILC are commonly interpreted as 'lower bound' estimates because only a small number of circumstances are observed. The effect of unobserved circumstances (such as, e.g., social and emotional skills, IQ, physical beauty, or any innate characteristics that may impact later life economic outcomes) are subsumed in 'effort' side of the equation. The lower bound qualification may not however be entirely appropriate because of counter-veiling upward bias in standard measures of IOp.

Brunori et al. (2019) show that standard estimators of IOp indicators can suffer from upward bias. The bias arises in finite samples in case of inclusion of *irrelevant* circumstances in the set C. Even in the absence of any relationship between circumstances and outcomes, empirical estimates of IOp are not zero, by construction. Addition of irrelevant circumstances increases potential bias. Irrelevant circumstances are variables that are treated by the analyst as circumstances but that are, in fact, independent on income, $C \perp\!\!\!\perp y$.

Recall that IOp measures can be viewed as indicators of between-group inequality (considered unfair, given the definition of circumstances) as opposed to within-group inequality. Increasing the number of circumstances increases the dimension of the partition of the population into groups of individuals having the same configuration of circumstances. It is easy to see that, in finite samples, the share of total inequality that is accounted for by between-group inequality can only increase with the inclusion of circumstances. In an extreme case where a partition is so large that there is a one-to-one correspondence between an individual and a configuration of circumstances, all within-group inequality disappears (since all groups are composed of just one person) and inequality in outcome is identical to between-group inequality – all inequality in outcome is considered unfair. This is a biased estimation of IOp. The problem is analogous to the increase in the R^2 statistic achieved by adding irrelevant regressors in linear regression analysis.

This upward bias has motivated the use of machine learning methods, notably penalized regression methods and/or automatic selection (or weighting) of relevant characteristics that penalize the inclusion of too many irrelevant circumstances (Brunori et al., 2019). Those methods are however difficult to implement for indicators beyond direct ex ante IOp measures.

We address the upward bias by implementing a simple bias correction where the IOp indices estimated as above are adjusted by subtracting an estimate of the bias:

$$\hat{\theta}^* := \hat{\theta} - \hat{\theta}^0 \tag{6}$$

where $\hat{\theta}$ denotes any one of the four estimates of IOp described in equations (2)–(5). The correction term, $\hat{\theta}^0$, is an IOp index obtained from exactly the same estimation procedure but applied to a modified input dataset. The modified dataset is constructed in such a way that income is completely independent on circumstances but the marginal distributions of income and circumstances are unchanged. The independence between income and circumstances implies the absence of inequality of opportunity by construction in the modified data. We therefore take the resulting estimate $\hat{\theta}^0$ to reflect *only* the upward bias that is introduced by estimating IOp measures in S_N when circumstances are irrelevant.

In practice, the modified input dataset is obtained by data permutations of S_N: $S_N^0 := \{(y_j, C_i)\}_{i=1}^N$, where y_j are drawn without replacement from the vector of incomes $\{y_i\}_{i=1}^N$. This is similar to the implementation of permutation tests (see, e.g. Pesarin and Salmaso, 2010).[2] Monte Carlo error involved in the random permutation of the data is averaged away by repeating calculations over multiple permutations of the data and using using in (6) the average $\hat{\theta}^0$ obtained across permutations.

3 The European Union Statistics on Income and Living Conditions (EU-SILC)

We exploit data for 26 EU countries or associated states: Austria (AT), Belgium (BE), Bulgaria (BG), Czechia (CZ), Croatia (HR), Denmark (DK), Germany (DE), Greece (EL), Spain (ES), Finland (FI), France (FR), Hungary (HU), Ireland (IE), Italy (IT), Luxembourg (LU), the Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT),

[2] While implementation is trivial for unweighted data, application to weighted data raises computational difficulties. Ensuring that empirical estimates of both marginal distributions G and H are maintained after permutation of the data requires adjustments to the sampling weights. This can be achieved by an iterative proportional fitting algorithm (Stephan, 1942).

Romania (RO), Serbia (RS), Sweden (SE), Switzerland (CH), Slovenia (SI), the Slovak Republic (SK), and the United Kingdom (UK).[3]

EU-SILC is one of the official micro-data source for calculation of EU social indicators and its collection is legally binding in all EU countries. It is available annually since 2003. The data consists of representative samples of the population of all EU Member States. The data contain detailed household and individual-level annual income information either extracted from administrative sources, or elicited directly from respondents. The data are largely, yet not perfectly, comparable across countries.

The analysis uses micro-data from the November 2020 release of EU-SILC containing cross-sectional data up to 2019 (https://doi.org/10.2907/EUSILC2004-2019V.1) We extract three cross-sections of the data, collected in 2005, 2011 and 2019.

To remain in line with official EU statistics on inequality and poverty, we consider as outcome the single-adult equivalent household disposable income of the respondent's household at the time of interview (as in, e.g., Brzezinski, 2020, Ramos and Van de gaer, 2021). The modified OECD scale is used to convert total household income into the single-adult equivalent amount. Households with reported annual income smaller than 100 euros are excluded from the estimation sample. Furthermore, income is top-coded at 150 percent of the 99th percentile of the national income distribution in the corresponding survey year, to mitigate the influence of extreme values.

Previous research exploiting EU-SILC has used a variety of sets of circumstances. Circumstances are obtained from responses to special additional modules collected in 2005, 2011 and 2019, which collected retrospective information about childhood circumstances to EU-SILC respondents aged from 24 to 59. The reference period for the collection of retrospective information when the respondent was aged around 14. The exact information collected in each of these modules varied over time, but on each of the three time points, the survey collected information about parents educational achievement and occupation and household structure when the respondent was a teenager (Andreoli and Fusco, 2019*a*).

We consider two alternative specifications of the set of circumstances:

1. Base model: Respondent's gender and her parental education only, namely father and mother's education each recoded as three different levels: no or primary education, secondary education, tertiary education, with the addition of an 'unknown' category when the respondent does not know her parent (or does not know her level of education);

2. Standard model: Base model augmented with a dummy indicator for the presence of siblings, the household composition at age 14 (whether the respondent

[3] Cyprus, Estonia, Iceland, Latvia, Lithuania and Malta are not included here because of sample size consideration.

lived with both parents, with her father, with her mother or with none of her parents), and father and mother employment status at age 14 in four possible categories (inactive, employed in an elementary occupation, employed in semi-skilled occupation, and employed in high-skilled occupation) occupation with the addition of an 'unknown' category as for education.[4]

These circumstances are all that can be consistently measured in all three periods.[5]

For the analysis by cohort, we estimate cohort-level inequality of opportunity measures for birth cohorts on the grid $b = \{1950, 1960, 1970, 1980\}$. Earlier and later birth cohorts have too small effective sample sizes for estimation of the distribution regression models.

Sampling weights provided by EU-SILC are applied in all calculations to account for unequal sampling probabilities.[6]

Bootstrap inference is performed on the basis of an exchangeably weighted bootstrap with sampling of replication weights from an exponential distribution (Praestgaard and Wellner, 1993). The bootstrap sampling is done at the level of the household, so all household members are drawn together. We are not able to take into account stratification and higher level clustering as the required variables are not systematically provided in EU-SILC users' databases (Goedemé, 2013).

4 New IOp estimates in 26 European countries

4.1 Aggregate IOp indicators and the new 2019 estimates

We first present estimates of inequality of opportunity across countries—ignoring estimation by cohort for the moment. To the best of our knowledge, we present here the first estimates of IOp measures in European countries based on the 2019 EU-SILC module on the transmission of disadvantage.

We report our four IOp indicators: (i) a direct ex ante (DEA) index, (ii) a direct ex post (DEP) index, (iii) an indirect ex ante (IEA) index and (iv) an indirect ex post (IEP) index. All indices are obtained from the same semi-parametric estimation of conditional quantiles estimated separately for all the country and survey year combinations available in our data. The Gini coefficient is used as inequality functional in all estimations. Much empirical research to date has used the mean log deviation because of its connection to the log-linear regression model and, therefore, the calculation of DEA measures. We prefer however to use the Gini coefficient for its generally smaller sampling variance and its limited sensitivity to extreme data. For

[4] Occupations are classified on the basis of the ISCO classification 1–4 for upper-level occupations, 5–7 for semi-skilled occupations and 0, 8 and 9 for elementary occupations.

[5] Information such as country of birth of parents is not collected in all modules.

[6] We use the general personal weight rather than the module-specific weights as the latter are unavailable for a range of countries.

the sake of brevity, we show and comment here on estimates for the extended set of circumstances only.

Our estimates are shown in Figure 2. The four panels pertain to the four different measures and, within each panel, country estimates are displayed on different lines, for 2005 (in light gray), 2011 (in dark gray) and 2019 (in orange). Countries are ordered from top to bottom according to the average IOp index across all years available. The general ordering we obtain is familiar to IOp analysts. When considering the DEA index, clusters of similar countries appear lumped together at similar levels of IOp. At the bottom, we find all four nordic countries (Denmark, Norway, Finland and Sweden) which exhibit the smallest inequality of opportunity. Then come the Netherlands, Austria and Germany. Then come Slovenia and the Slovak Republic. Another cluster follows with Switzerland, France and Belgium. The two anglo-saxon countries (Ireland and the UK) come just after the Czech Republic. All southern European countries appear high up in the ranking—alongside the somewhat peculiar cases of Hungary, Luxembourg and Poland—with, in that order, Croatia, Greece and Spain followed by Hungary and Luxembourg, then Italy, Serbia, Poland and Portugal. At the very top of the list appear the two poorest EU countries, Romania and Bulgaria. With few variations (e.g., in the relative positions of countries such as Luxembourg or Sweden), this general ordering is robust to the choice of IOp measure.

Expressing the IOp measures as a fraction of the overall Gini coefficient, we find that the variation in how much of inequality is accounted for by circumstances is large. As is relatively common in the empirical literature, IOp represents between 5 percent (in nordic countries) and 40 percent (in Portugal, Romania or Bulgaria) of total inequality using the DEA or DEP indices (and excluding the exceptional 48 percent in Bulgaria 2019). The shares are much lower with indirect measures IEA and IEP.

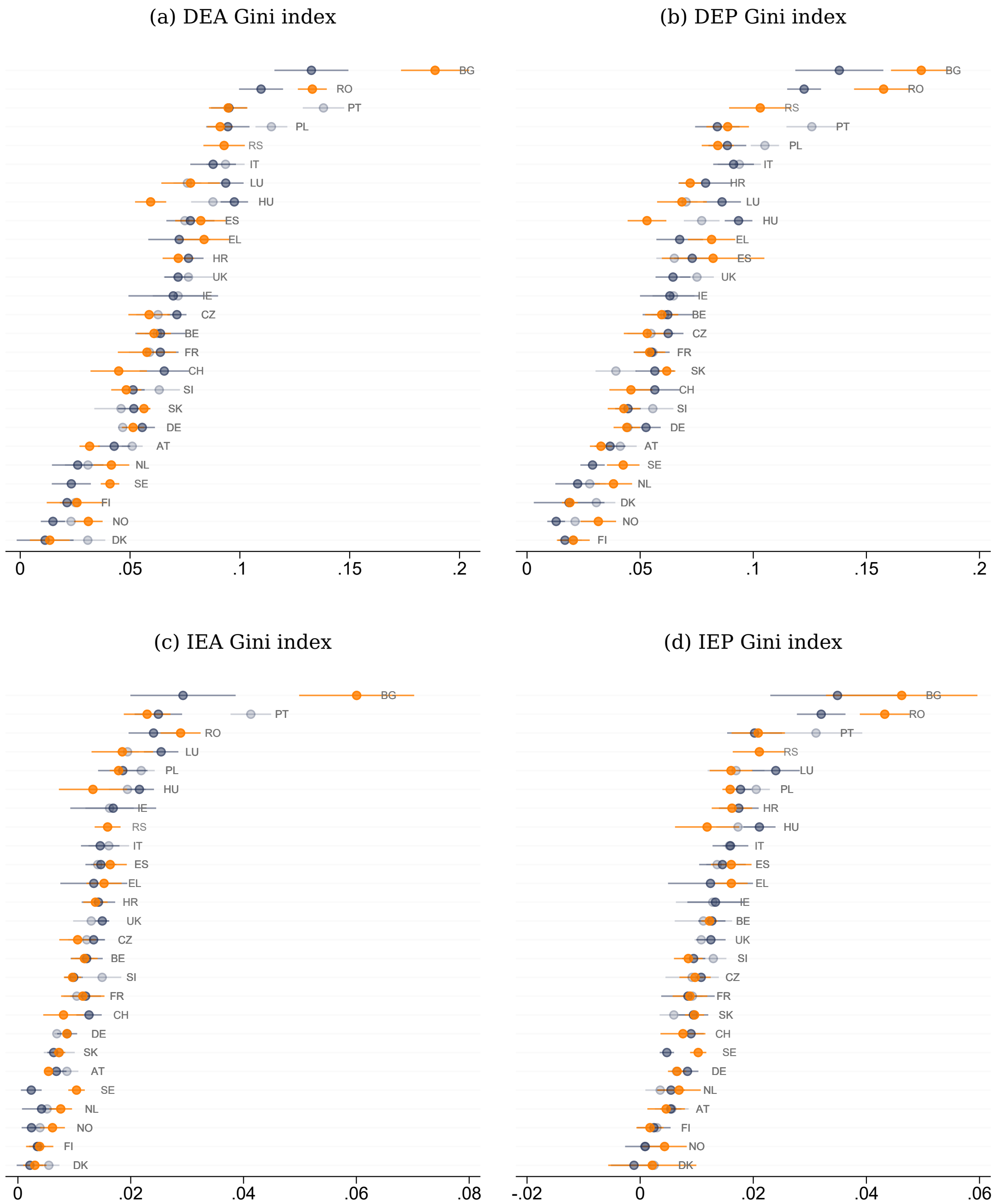
Overall, a similar picture emerges across different measures. Trends and the relative position of different countries is generally consistent across different measures. Indirect indicators however lead to much lower estimates of inequality of opportunity. Unlike Ramos and Van de gaer (2021) we do not find that the choice of the ex post versus ex ante perspective has stronger implications than the choice of direct and indirect measures. The overall rank correlations across the four measures over the 67 estimates pooled across countries and years range between 0.9 (for DEP and IEA) and 0.98 (for DEA and DEP). We use more data than Ramos and Van de gaer (2021) by including the 2019 EU-SILC data, but focus on a smaller set of indices and use the more elaborate estimation strategy presented above.

Given the novelty of the 2019 estimates, we also highlight in Figure 3 the difference between the 2011 and 2019 estimates for all countries for which both are available. New data for 2019 show sharp increases in IOp for a range of countries, notably Bulgaria, Romania and Sweden. The evolution is particularly worrisome for the two former countries which already topped the ranking of countries in 2011 and now exhibit a degree of inequality of opportunity that is well above all other European countries. IOp estimates are also on the increase at the other end of the ranking, with increases observed in all nordic countries and the difference with countries such as the Netherlands, Germany and Austria narrowing down. For the rest, no clear pattern emerges and it seems difficult to draw further lessons from inspection of the change from 2011 to 2019.

On the methodological side, two points are worth noting before moving to cohort-level estimates. First, we show in Appendix A the close similarity of estimates obtained for DEA measures using a classic OLS estimator and our semi-parametric approach. We take this as a re-assuring finding and bolsters our confidence in our empirical strategy which allows us to estimate both ex post and ex anten direct and indirect measures, unlike simple OLS. (Ramos and Van de gaer (2021) also noted in their application the robustness of estimates to alternative estimation strategy.)

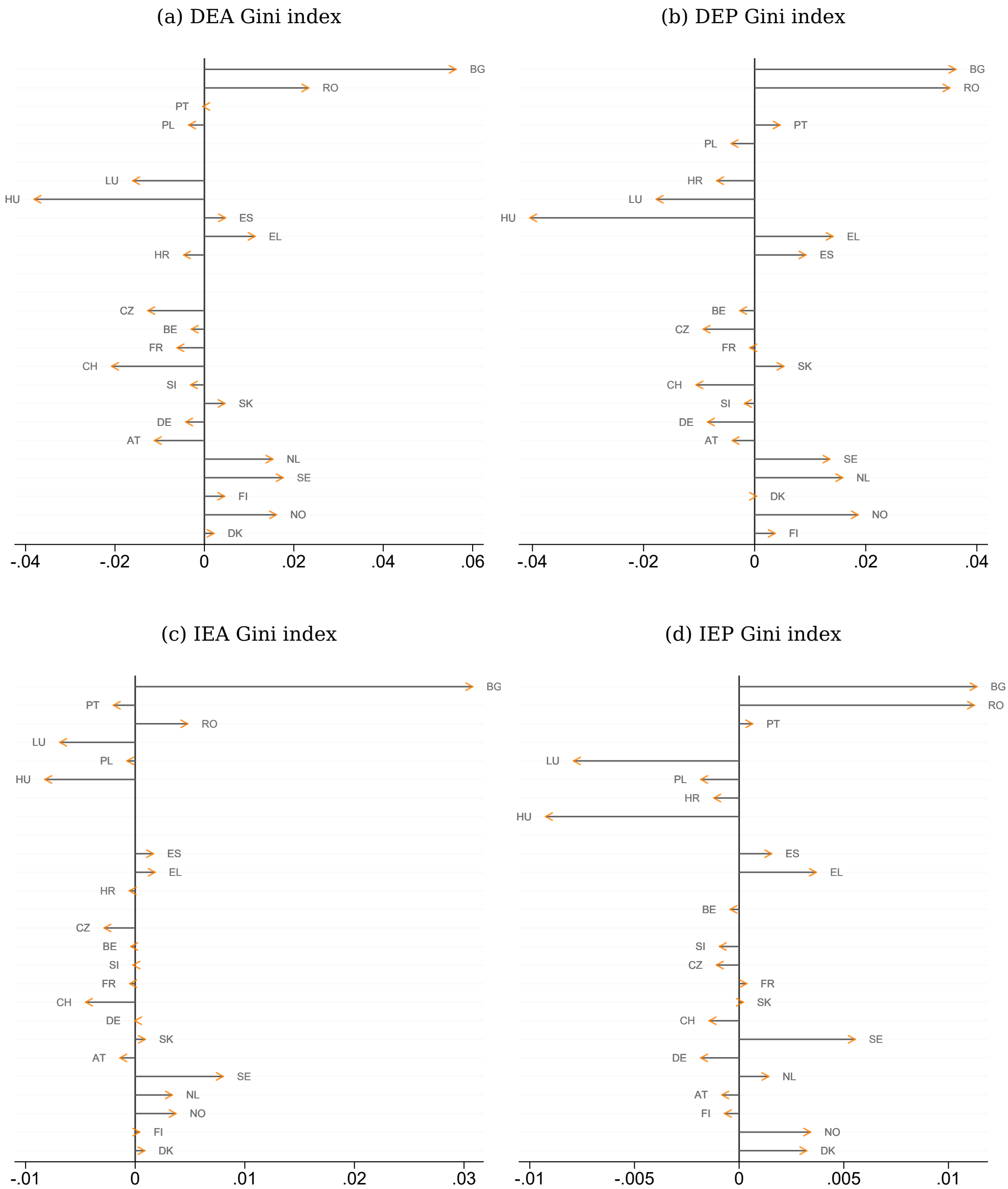
The second point is the impact of the bias correction. Appendix B compares estimates with and without upward bias correction. The upward bias appears to be truly substantial (at least for three of the four measures) – reaching up to half of unadjusted estimates in low IOp countries. The size of the correction is similar to what Brunori et al. (2019) obtain for their DEA measures, although they use a very different estimation strategy and the estimates are not directly comparable.

Figure 2: IOp estimates for 26 EU Member and Associated States, 2005, 2011 and 2019



Note: Estimates of IOp indicators for all countries are vertically positioned from highest IOp (top) to lowest IOp (bottom) according to the average value of IOp in the years for which estimates are available. Light gray markers show estimates for 2005, gray markers show estimates for 2011 and orange markers show the most recent 2019 estimates. Horizontal lines show 95 percent bootstrap confidence intervals.

Figure 3: Change in IOp estimates from 2011 to 2019



Note: Estimates of change in IOp indicators for all countries are vertically positioned from highest IOp (top) to lowest IOp (bottom) according to the average value of IOp in the years for which estimates are available.

4.2 Estimates of inequality of opportunity across cohorts

We now move to estimates by cohort obtained by our kernel weighted estimation procedure. As discussed above, we are primarily interested here in the evolution of IOp across cohorts. We argue that the most meaningful IOp estimates are those estimated across individuals of the same cohort so that estimation is not contaminated by differences in positions in the life-cycle.

Figures 4-7 display estimates of the four IOp measures for all 26 countries and four cohorts born in (a window centered on) 1950, 1960, 1970 and 1980. Estimates obtained from the same survey year are connected by patterned lines. The vertical variations across estimates from different years therefore reflect a combination of an age effect and a potential time effect.

At first sight, IOp is generally flat across cohorts and is otherwise increasing in a few countries. However, we should recognize here that total inequality may also differ substantially within different cohorts and years – notably as per life-cycle variation. We therefore also report estimates of 'relative IOp' by dividing the IOp estimates by an estimate of the Gini coefficient of total inequality *within* the cohort – estimated using the same kernel weighting procedure.

The patterns shown in Figures 8-11 now more clearly reveal an increasing trend over cohorts in the majority of countries: relative IOp tends to be higher among more recent cohorts than among older cohorts. To support visual inspection of the estimates, a linear trend superimposed on the DEA estimates is upward sloping in all countries except Switzerland, Germany and Norway. If we allow for a year effect (assumed common across all cohorts in a given country), the linear trend is positive in all countries, except Germany. Similar results hold for other measures.[7] The gap in IOp between older and younger cohorts is particularly strong in Bulgaria and Romania where it seems the growth in overall IOp is driven by younger cohorts. More generally the upward trends seem more often observed in eastern European countries – see estimates of, e.g., Croatia, Hungary, Poland or Slovenia.

[7] With DEP and a year effect, only Switzerland and Portugal show non-positive trends. With IEA, this is only for Switzerland and with IEP this is only for Serbia and Greece.

Figure 4: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, DEA Gini index

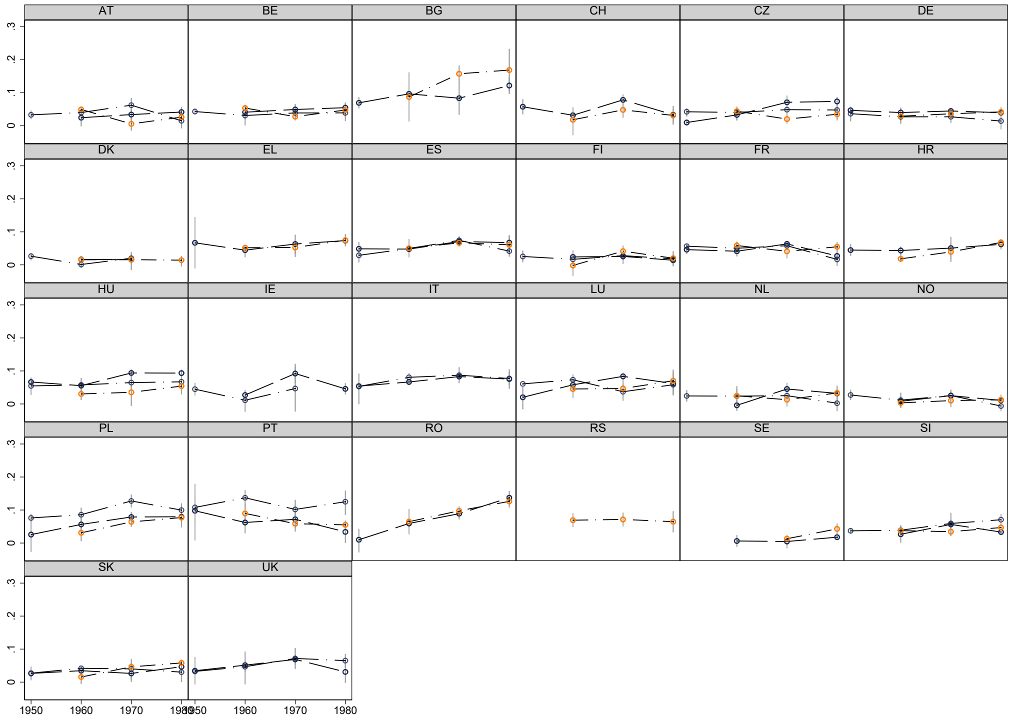


Figure 5: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, DEP Gini index

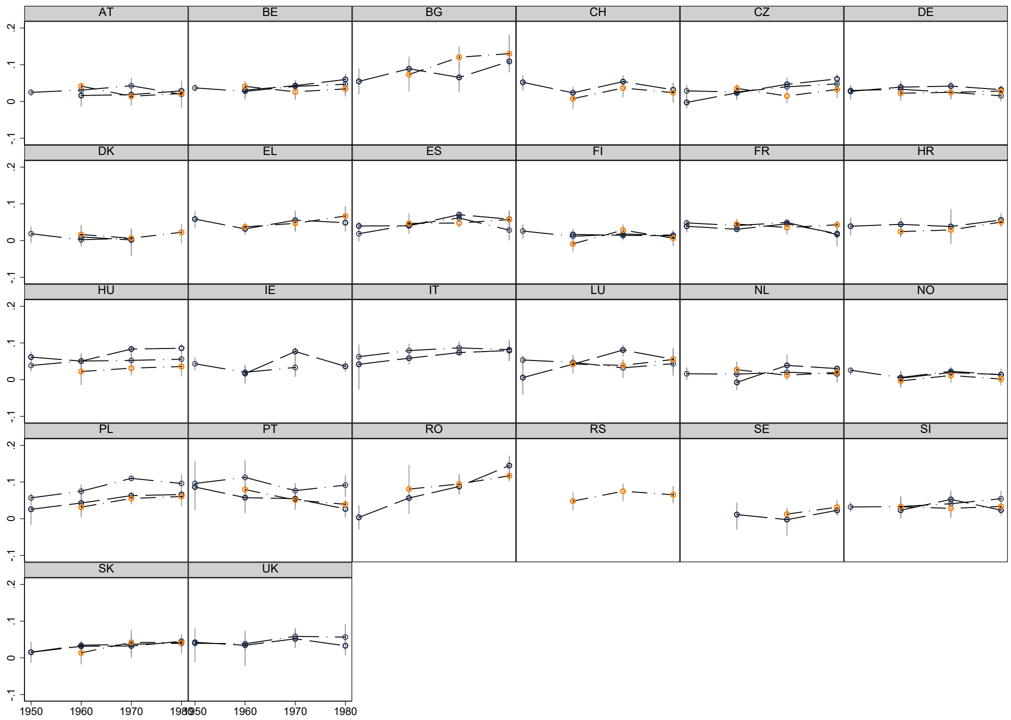


Figure 6: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, IEA Gini index

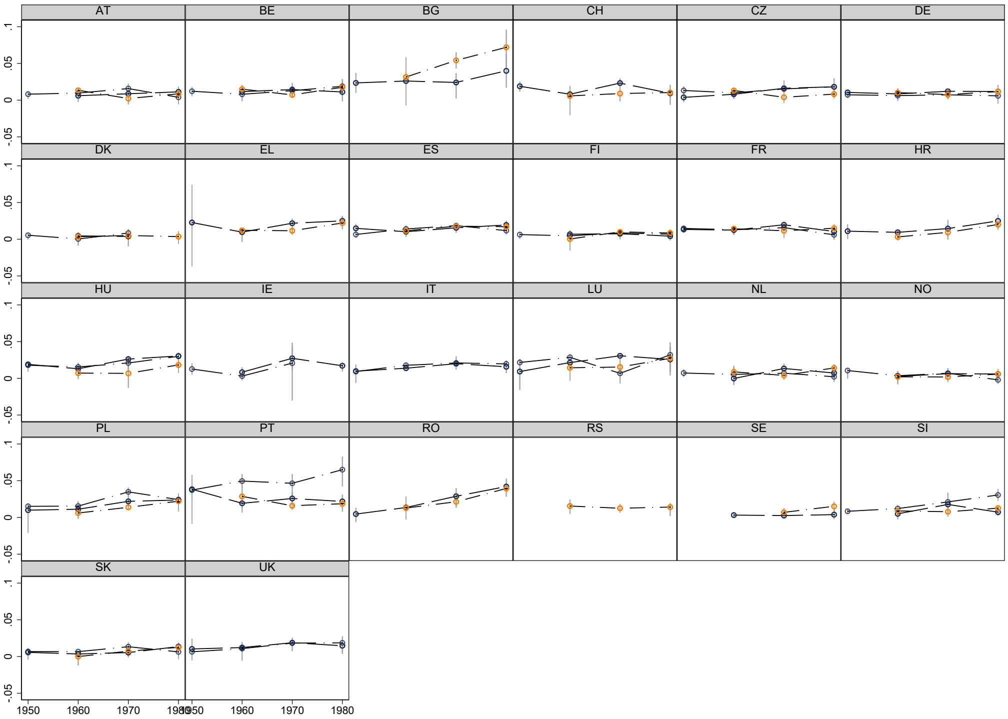


Figure 7: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, IEP Gini index

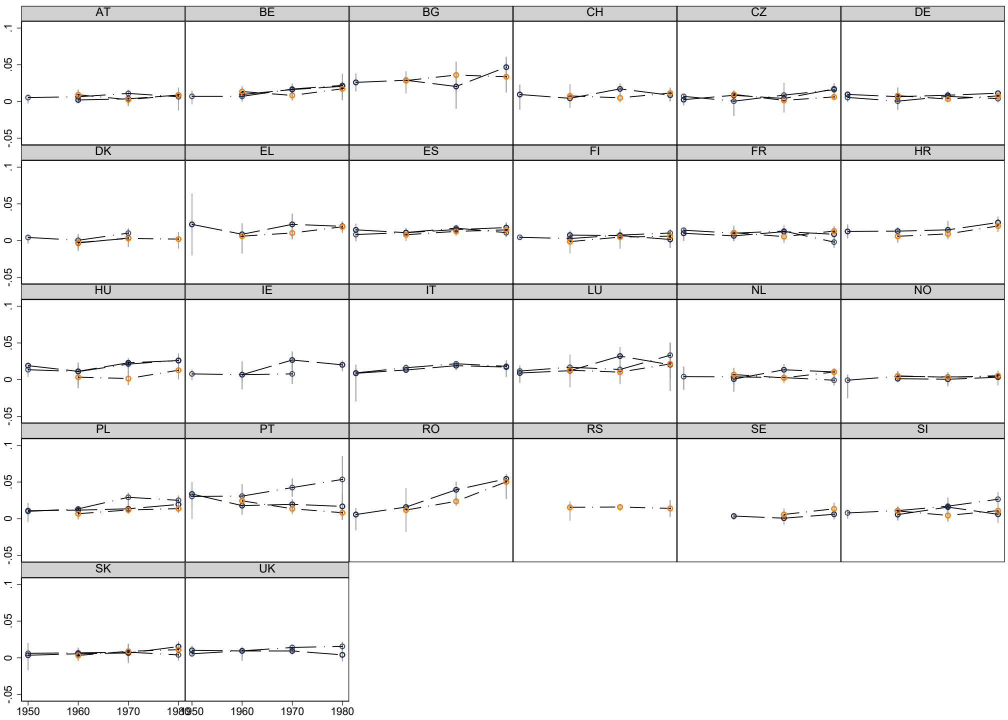


Figure 8: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, DEA Gini index, estimates scaled by overall cohort Gini index

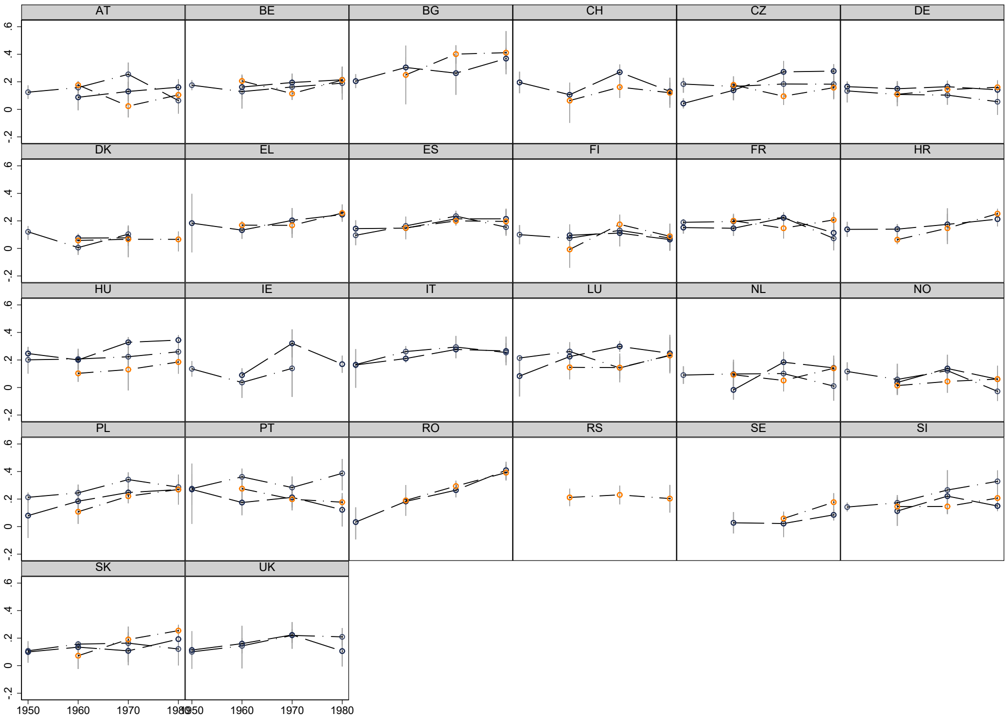


Figure 9: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, DEP Gini index, estimates scaled by overall cohort Gini index

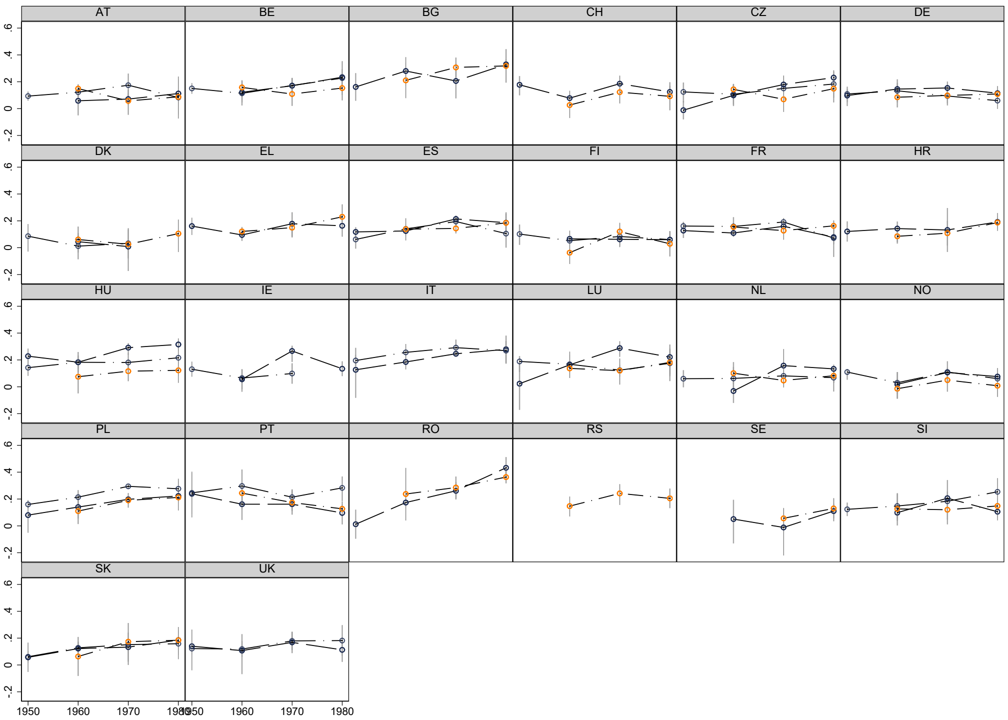


Figure 10: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, IEA Gini index, estimates scaled by overall cohort Gini index

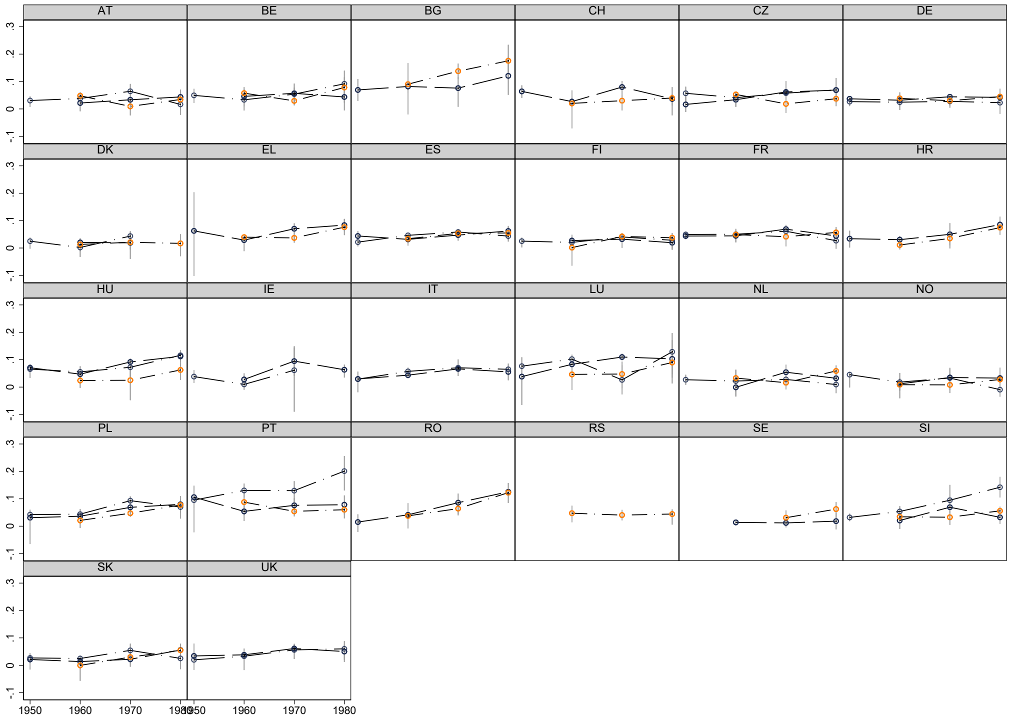
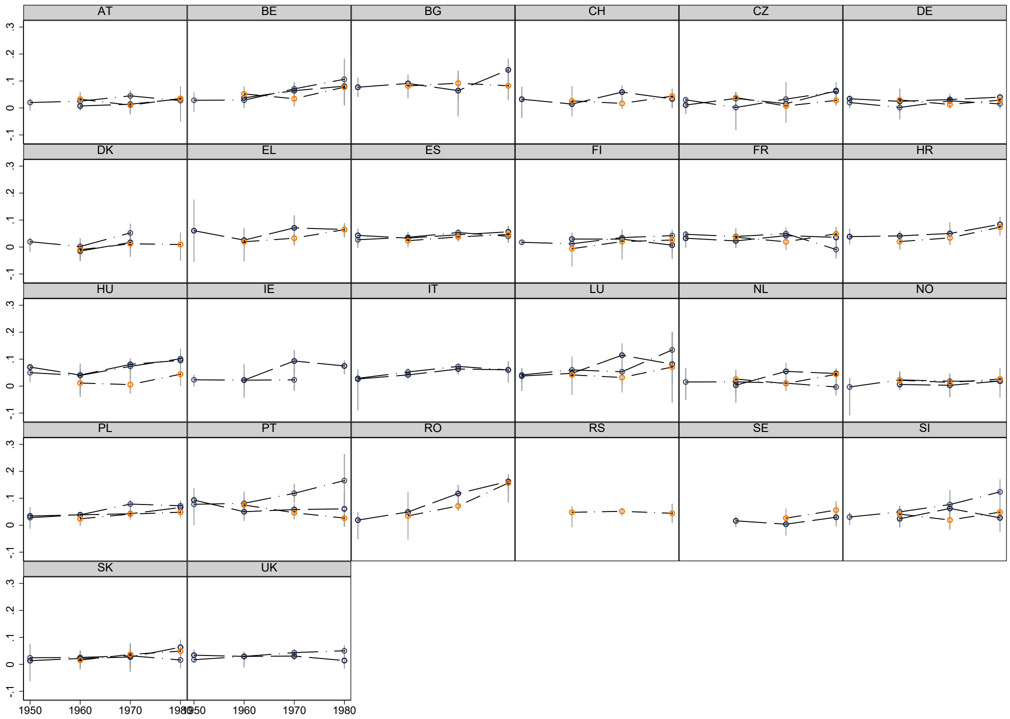


Figure 11: Cohort-level estimates of IOp for cohorts born in 1950, 1960, 1970 and 1980, IEP Gini index, estimates scaled by overall cohort Gini index



5 The relationship between inequality of opportunity and educational policy

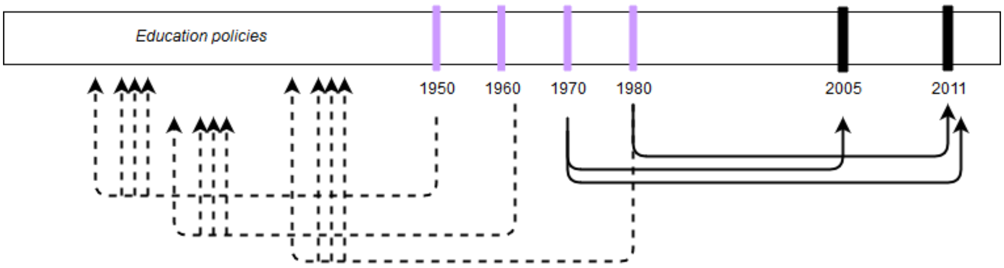
We finally examine the relationship between cohort-level IOp estimates and educational policy that were in place during childhood years of respondents parents. The basic intuition is that education policies, by enhancing educational achievements, have the potential to improve the circumstances of the next generation. While it is well documented that key educational policies (such as the compulsory schooling duration) influence the educational achievement of the generation directly affected, their effect on the *following* generation through improvements of the offspring circumstances is potentially ambiguous.

It is however difficult—theoretically—to anticipate the direction of the effect since many potential factors are at play (Bussolo et al., 2019). First, there is the question as to how much educational policy has an impact on the educational achievement of the generation immediately affected. Assuming educational policies meant to expand achievement do reach their goal, the impact on inequality of opportunity depends on whose parents are affected – is educational expansion across the board or primarily increasing education of those who would otherwise obtain lower levels of education? Then comes the question of what the impact on the distribution of educational achievement among the parents implies for the overall socio-economic circumstances of the future children. We expect education expansion to ameliorate the general socio-economic background in which offsprings are born, but whether it really does so and how much it does so cannot be taken for granted (simple supply and demand arguments suggest that educational expansion implies higher overall returns only if demand for education grows too). Moving to the offspring generation—among whom we are interested in evaluating inequality of opportunity—we expect that improved circumstances individually improves later life outcomes. But again whether an overall increase in the supply of individuals with better background leads to an overall increase in outcomes depends on supply and demand arguments and composition effects. Adding consideration of 'selection' of what types of offspring benefit from the enhanced parental education, it is ultimately difficult to predict whether educational policies have long-run impacts across generations.

We modestly set out to provide some (highly) 'reduced form' estimates of the empirical relationship between IOp in cohorts of offsprings and some key educational policy parameters that affected the parents of these cohorts. Our objective is essentially to examine whether there is any relationship at all – and whether there is therefore interest in pushing future research into examining the mechanisms more causally.

We examine perhaps the two most important educational policy parameters:

Figure 12: Assigning educational policy based on parental birth year



the minimum legal school leaving age and the number of compulsory school years (which depends on the age at entry). These policy parameters are matched to every EU-SILC respondent on the basis of the year of birth of their parents, as Figure 12 illustrates. The match between year of birth and policy parameter is taken from the database built by Braga et al. (2013). The database allows us to know what is, say, the school leaving age in place in a given country when a person born in year x enters secondary education, or the school entry age when the person is 3 years old, etc. (That is, people are matched to the policy in place when they are affected by them.)

Practically the policy parameter estimates matched to the cohort-level IOp are obtained as follows. We know from EU-SILC the year of birth of respondents' parents. We therefore can associate the policy parameters to one or both parents and take the average thereof for each respondent. These are then averaged across all respondents from a given cohort b, as illustrated in Figure 12. Note that not all individuals in a cohort have parents born in exactly the same year, so there can be variation in the value of policy parameter for parents of different individuals from the same cohort. This explains why the average policy parameter is used and regressed against the IOp indices.

The analysis here exploits only a subset of the data used in the previous subsection. First, policy parameters in the database compiled by Braga et al. (2013) do not cover the whole set of 26 countries. Second, EU-SILC stopped collecting information about the year of birth of parents in 2019. We therefore base our analysis on cohort-level estimates obtained from the 2005 and 2011 datasets and for 18 countries.

Figure 13 plots the cohort-level IOp estimates by the (average) duration of compulsory education when the parents of the individuals in these cohorts were at school. The duration of compulsory education varied from about three years in Portugal up to around 10 years in the UK. Most of the variation in policy parameters is *across* countries but there is also some variation over time within countries that reflect the gradual expansion of the duration of compulsory education in most countries.

The scatter plot and the overlaid regression lines suggest the existence of a negative relationship – whatever the index used. Simple Regression analysis show in Table 1 confirms the negative relationship when we use school leaving age instead of duration of compulsory education. In the first two columns, the table shows coefficients on univariate regression of the data as shown in Figure 13. When the two policies are combined (column 3), the negative coefficient remains for the duration of compulsory education but it turns positive on the school leaving age. This pattern suggests that the negative relationship is mostly driven by school entry age rather than school leaving age. These coefficients remain robust to the addition of average parental birth year in the regressions in order to capture potential secular trends in educational achievement beyond policy variables (column 4). The inclusion of country fixed effects however makes coefficients close to zero: unsurprisingly given the pattern shown in Figure 13, the coefficients on policy parameters are driven by between country variations. Note that we use cluster robust standard error estimators with clustering at the 'dataset' level (formed by combinations of country and year) and this leads to very few of the coefficients being statistically different from zero.

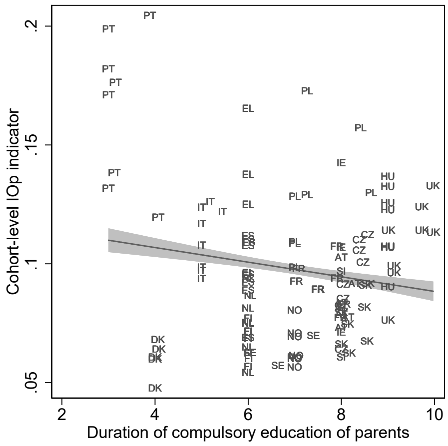
6 Conclusion

This paper set out to derive new estimates of inequality of opportunity across European countries. The 'novelty' comes from four elements. First, we implement a semi-parametric estimator for both ex post and ex ante, direct and indirect measures of IOp. The estimator exploits simple distribution regression models and overcomes the somewhat artificial distinction often made between 'parametric' and 'non-parametric' estimation of IOp (see, e.g., Ramos and Van de gaer, 2021). The key advantage is that we can estimate different IOp measures without being constrained by the number of circumstance variables. Second, our estimates account for the upward bias that finite sample estimation of IOp face (Brunori et al., 2019) – a bias that we show can be substantial. Third, we focus on cohort-level inequality of opportunity. We argue that IOp indicators are best measured within cohorts of individuals that are in the same position in their life-cycle. Fourth, we exploit newly available data from the newly released 2019 European Union Statistics on Income and Living Conditions. The latest estimates presented in similar studies now dated ten years with the 2011 release.

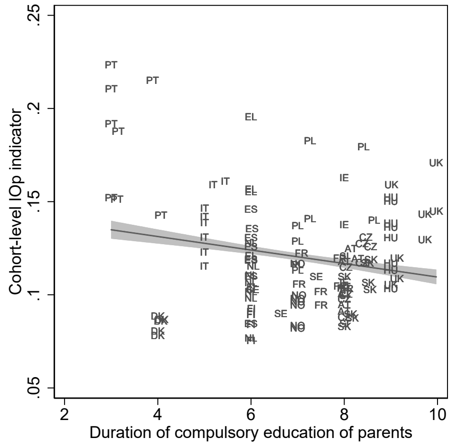
On the methodological side, our results indicate that the semi-parametric estimator is both feasible and reliable and that our bias correction approach is both important and easy to implement. On the substantive side, the 2019 estimates reveal no strong changes from 2005 or 2011 except for what appears to be a dramatic worsening of IOp numbers in the two poorest EU countries, Romania and Bulgaria. Examination of the cohort-level estimates show a general upward trend, with younger cohorts experiencing higher IOp than older cohorts in most countries

Figure 13: Inequality of opportunity at cohort level and duration of compulsory education of parents

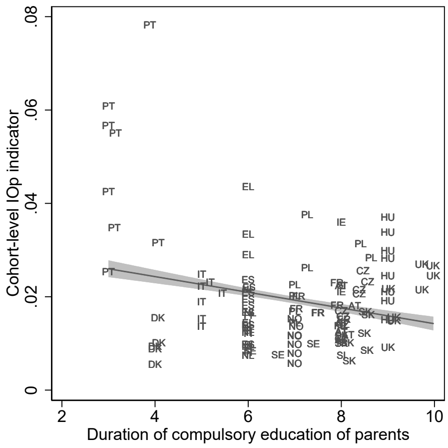
(a) DEA Gini index



(b) DEP Gini index



(c) IEA Gini index



(d) IEP Gini index

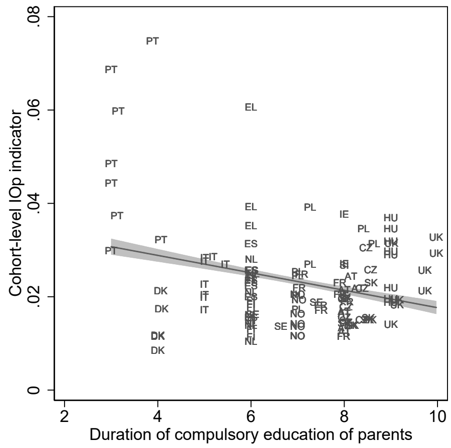


Table 1: Regressions results: inequality of opportunity at cohort level (divided by Gini) and compulsory education policies of parents

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) FE
DEA Gini index					
Compulsory schooling	-0.00242		-0.0224	-0.0182	0.00834
Minimum school leaving age		0.00147	0.0273	0.0175	–
Average parental birth year				0.00331*	0.00326*
DEP Gini index					
Compulsory schooling	-0.00412		-0.0218^{+}	-0.0159	-0.00571
Minimum school leaving age		-0.000991	0.0242^{+}	0.0105	–
Average parental birth year				0.00462*	0.00461*
IEA Gini index					
Compulsory schooling	-0.00334		-0.0131	-0.0116	0.00467
Minimum school leaving age		-0.00177	0.0133	0.00997	–
Average parental birth year				0.00114*	0.00109*
IEP Gini index					
Compulsory schooling	-0.00407		-0.0124^{+}	-0.0109	-0.00224
Minimum school leaving age		-0.00290	0.0115^{+}	0.00778	–
Average parental birth year				0.00125*	0.00127*

$^{+}$ $p < 0.10$, * $p < 0.05$. Cluster-robust standard errors with clustering at the dataset level (country-year pairs).

– and in particular in many eastern European countries. While we do not research the causes and the consequences of this pattern, it is tempting to link this with some of the anti-EU sentiment and the support for populist parties observed in some of these countries.

Finally, we finish with examination of what had initially triggered our efforts to obtain new cohort-level estimates of IOp. We wanted to explore how policy decisions influencing a generation can have implications for inequality of opportunity in the following generation. We present some exploratory analysis of the empirical relationship between inequality of opportunity in European cohorts and education policy parameters in place when the parents of these cohorts where in education. We observe some negative association – based mainly on variations across countries – between IOp and how long parents are kept in school: expansion of compulsory education is associated with lower inequality of opportunity among children. The promotion of education by educational policies may therefore not only be beneficial for children kept in the school but also reduces the long-term association between circumstances and income across generations. Our results are however exploratory and, of course, cannot be interpreted as any robust causal parameters. Further research will need to clarify potential mechanisms (as, e.g., Bussolo et al., 2019, do) and develop a refined strategy for causal estimation of the policy parameters.

References

Aaberge, R., Mogstad, M. and Peragine, V. (2011), 'Measuring long-term inequality of opportunity', *Journal of Public Economics* **95**(3–4), 193–204.

Andreoli, F. and Fusco, A. (2019*a*), 'Application of the EU-SILC 2011 data module "intergenerational transmission of disadvantage" to robust analysis of inequality of opportunity', *Data in Brief* **25**(1043016), 1–13.

Andreoli, F. and Fusco, A. (2019*b*), 'Robust cross-country analysis of inequality of opportunity', *Economics Letters* **182**, 86–89.

Braga, M., Checchi, D. and Meschi, E. (2013), 'Educational policies in a long-run perspective', *Economic Policy* **28**(73), 45–100.

Brunori, P., Peragine, V. and Serlenga, L. (2019), 'Upward and downward bias when measuring inequality of opportunity', *Social Choice and Welfare* **52**(4), 635–661.

Brzezinski, M. (2020), 'The evolution of inequality of opportunity in Europe', *Applied Economics Letters* **27**(4), 262–266.

Bussolo, M., Checchi, D. and Peragine, V. (2019), Long-term evolution of inequality of opportunity, Policy Research Working Paper 8700, The World Bank Group, Washington D.C., USA.

Checchi, D., Peragine, V. and Serlenga, L. (2016), Inequality of opportunity in Europe: Is there a role for institutions?, *in* L. Cappellari, S. W. Polachek and K. Tatsiramos, eds, 'Inequality: Causes and Consequences', Vol. 43 of *Research in Labor Economics*, Emerald Publishing Ltd.

Chernozhukov, V., Fernandez-Val, I. and Melly, B. (2013), 'Inference on counterfactual distributions', *Econometrica* **81**(6), 2205–2268.

Dunnzlaff, L., Neumann, D., Niehues, J. and Peichl, A. (2011), Equality of opportunity and redistribution in Europe, *in* J. G. Rodríguez, ed., 'Inequality of Opportunity: Theory and Measurement', Vol. 19 of *Research on Economic Inequality*, Emerald Publishing Ltd, chapter 5, pp. 99–129.

Ferreira, F. H. G. and Gignoux, J. (2011), 'The Measurement Of Inequality Of Opportunity: Theory And An Application To Latin America', *Review of Income and Wealth* **57**(4), 622–657.

Foresi, S. and Peracchi, F. (1995), 'The conditional distribution of excess returns: An empirical analysis', *Journal of the American Statistical Association* **90**(430), 451–466.

Goedemé, T. (2013), 'How much confidence can we have in EU-SILC? Complex sample designs and the standard error of the Europe 2020 poverty indicators', *Social Indicators Research* **110**(1), 89–110.

Pesarin, F. and Salmaso, L. (2010), 'The permutation testing approach: a review', *Statistica* **70**(4), 481–509.

Praestgaard, J. and Wellner, J. A. (1993), 'Exchangeably weighted bootstraps of the general empirical process', *The Annals of Probability* **21**(4), 2053–2086.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007), *Numerical Recipes: the Art of Scientific Computing*, 3rd edn, Cambridge University Press, New York.

Ramos, X. and Van de gaer, D. (2016), 'Approaches to inequality of opportunity: Principles, measures and evidence', *Journal of Economic Surveys* **30**(5), 855–883.

Ramos, X. and Van de gaer, D. (2021), 'Is inequality of opportunity robust to the measurement approach?', *Review of Income and Wealth* **67**(1), 18–36.

Roemer, J. E. and Trannoy, A. (2015), Equality of opportunity, *in* A. B. Atkinson and F. Bourguignon, eds, 'Handbook of Income Distribution', Vol. 2, North-Holland, Amsterdam, chapter 4, pp. 217–300.

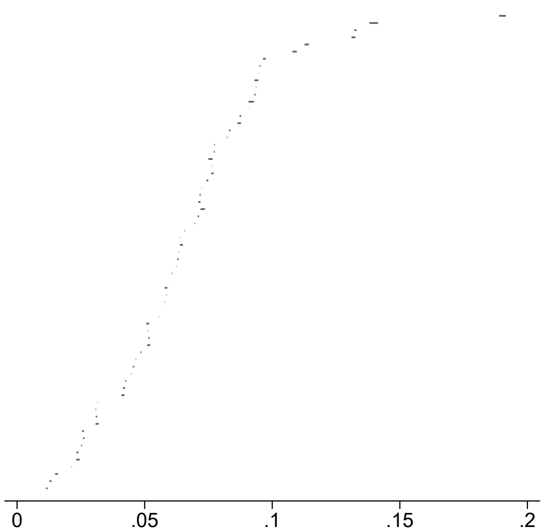
Stephan, F. F. (1942), ‘An iterative method of adjusting sample frequency tables when expected marginal totals are known’, *The Annals of Mathematical Statistics* **13**(2), 166–178.

Appendices

A Semi-parametric versus regression-based estimates

Figure 1 shows the difference between estimates of direct ex ante indices obtained from the classic OLS regression and estimates obtained from our semi-parametric distribution regression (DR) model. The difference in estimates is represented by the length of horizontal segments. Estimates obtained from both estimation strategies are very similar.

Figure 1: Direct ex ante indices of IOp: regression estimates vs. semi-parametric regression estimates



Note: Estimates of IOp indicators for all countries and the three years in our data are vertically positioned from highest IOp (top) to lowest IOp (bottom), with estimates reported on the horizontal axis. OLS and DR estimates are connected by horizontal segments.

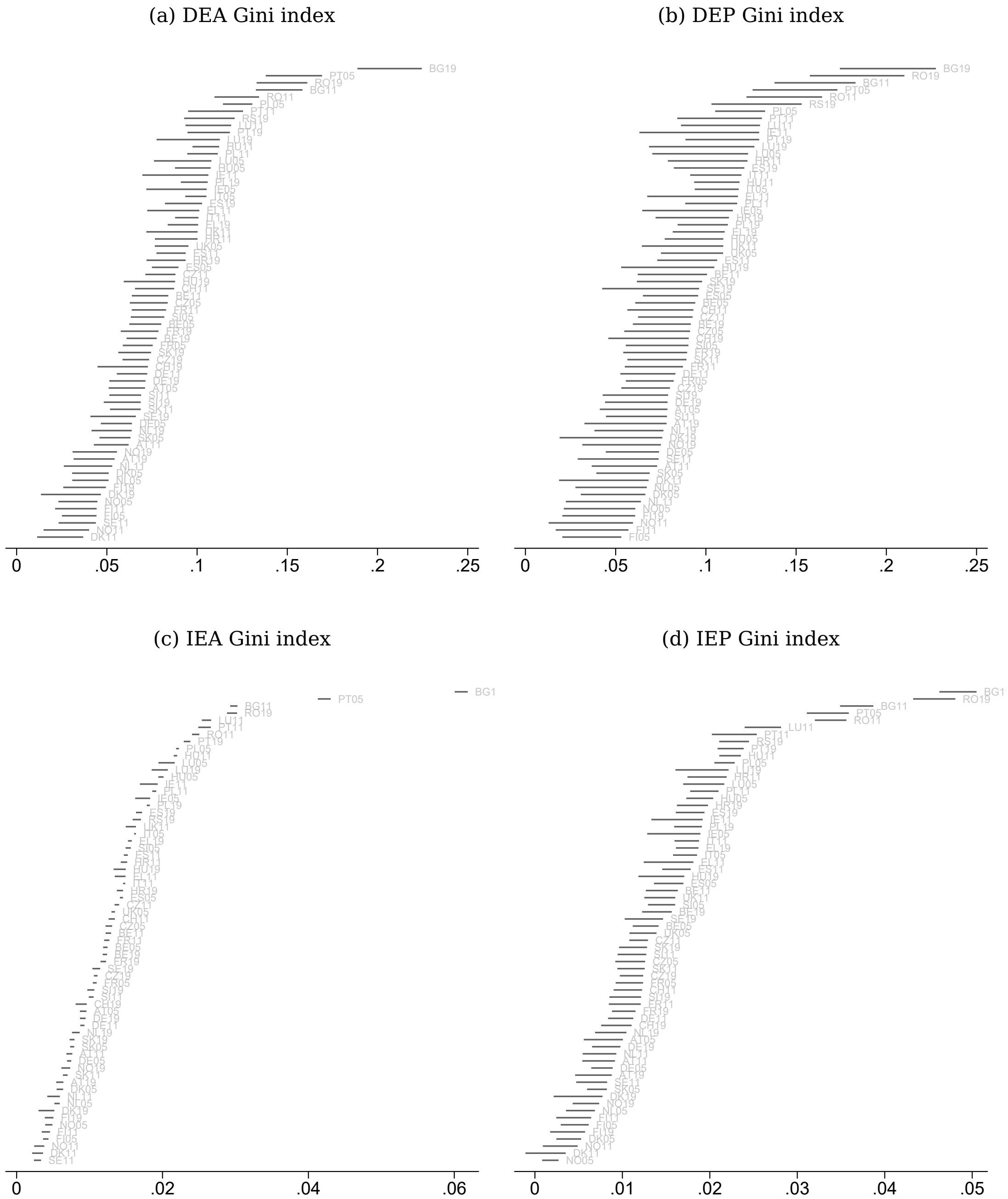
B The impact of bias-correction on IOp indices

Figure 2 shows both unadjusted and bias-corrected estimates of IOp measures for the extended set of circumstances. The upward bias appears to be substantial for three of the four families of indices – only the indirect ex ante measure shows little bias. The bias can be large – reaching up to half of the IOp estimates for low IOp countries. In relative terms, the correction is largest for low IOp countries, such as, e.g., Denmark. This is unsurprising since in a country with *no* inequality of opportunity, *any* unadjusted IOp would be the result of bias.

It is the direct ex post index which appears to exhibit most upward bias.

Overall, the size of the correction is similar to what Brunori et al. (2019) obtain for their direct ex ante measures using an alternative estimation strategy based on data-driven selection of circumstances.

Figure 2: IOp estimates with and without bias-correction



(a) DEA Gini index

(b) DEP Gini index

(c) IEA Gini index

(d) IEP Gini index

Note: Estimates of IOp indicators for all countries and the three years in our data are vertically positioned from highest IOp (top) to lowest IOp (bottom), with estimates reported on the horizontal axis. Unadjusted and bias-corrected estimates are connected by horizontal segments alongside their country-year identifiers. The right-hand side of the segment is at the unadjusted value and the left-hand side is at the bias-corrected estimate. The length of the segments show the size of the bias correction.