Jun, 2025
GRINS DISCUSSION PAPER SERIES DP N° 14/2025 ISSN 3035-5576

GRINS

Biases in inequality of opportunity estimates: measures and solutions

DP N° 14/2025

Authors:
Domenico Moramarco, Paolo Brunori, Pedro Salas-Rojo

Jun, 2025
GRINS DISCUSSION PAPER SERIES DP N° 14/2025 **ISSN 3035-5576**

Biases in inequality of opportunity estimates: measures and solutions

Domenico Moramarco, Paolo Brunori, Pedro Salas-Rojo

KEYWORDS

Equality of opportunity | Machine learning | PhD graduates | Compensation | Reward

JEL CODE

C38, D31, D63

ACKNOWLEDGEMENTS

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CITE THIS WORK

Author(s): Domenico Moramarco, Paolo Brunori, Pedro Salas-Rojo. Title: Biases in inequality of opportunity estimates: measures and solutions. Publication Date: 2025.

In this paper we discuss some limitations of using survey data to measure inequality of opportunity. First, we highlight a link between the two fundamental principles of the theory of equal opportunities – compensation and reward – and the concepts of power and confidence levels in hypothesis testing.
This connection can be used to address, for example, whether a sample has sufficient observations to appropriately measure inequality of opportunity. Second, we propose a set of tools to normatively assess inequality of opportunity estimates in any type partition. We apply our proposal to Conditional Inference Trees, a machine learning technique that has received growing attention in the literature.
Finally, guided by such tools, we suggest that standard tree-based partitions can be manipulated to reduce the risk of compensation and reward principles. Our methodological contribution is complemented with an application using a quasi-administrative sample of Italian PhD graduates. We find a substantial level of labor income inequality among two cohorts of PhD graduates (2012 and 2014), with a significant portion explained by circumstances beyond their control.
Keywords: Equality of opportunity, Machine learning, PhD graduates, Compensation, Reward.
JEL: C38, D31, D63

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
GRINS FOUNDATION

Discussion Paper Series

Biases in inequality of opportunity estimates: measures and solutions.

Discussion paper n. 14/2025

Domenico Moramarco (1), Paolo Brunori (2), Pedro Salas-Rojo (3); (1) University of Bari - Department of Economics and Finance; (2) University of Firenze & London School of Economics - International Inequalities Institute; (3)London School of Economics - International Inequalities Institute.

ISSN 3035-5567

Biases in inequality of opportunity estimates: measures and solutions.

Domenico Moramarco (1), Paolo Brunori (2), Pedro Salas-Rojo (3); (1) University of Bari - Department of Economics and Finance; (2) University of Firenze & London School of Economics - International Inequalities Institute; (3)London School of Economics - International Inequalities Institute.

Keywords: Equality of opportunity, Machine learning, PhD graduates, Compensation, Reward.

Discussion Paper DP 14/2025

In this paper we discuss some limitations of using survey data to measure inequality of opportunity. First, we highlight a link between the two fundamental principles of the theory of equal opportunities – compensation and reward – and the concepts of power and confidence levels in hypothesis testing. This connection can be used to address, for example, whether a sample has sufficient observations to appropriately measure inequality of opportunity. Second, we propose a set of tools to normatively assess inequality of opportunity estimates in any type partition. We apply our proposal to Conditional Inference Trees, a machine learning technique that has received growing attention in the literature. Finally, guided by such tools, we suggest that standard tree-based partitions can be manipulated to reduce the risk of compensation and reward principles. Our methodological contribution is complemented with an application using a quasi-administrative sample of Italian PhD graduates. We find a substantial level of labor income inequality among two cohorts of PhD graduates (2012 and 2014), with a significant portion explained by circumstances beyond their control.

We are grateful to Chico Ferreira for his precious advice and comments. We also thank the participants of the conferences: Ifo Conference Conference on Understanding Socio-Economic Inequalities with Novel Data and Method; Higher Education and Equality of Opportunities, University of Modena and Reggio Emilia; Equality of Opportunity and Intergenerational Mobility: A Global Perspective, University of Bari; Tenth ECINEQ Meeting, Aix-en-Province. Domenico Moramarco gratefully acknowledges ECARES and FNRS for the financial support during the earlier stages of this project. He also acknowledges the GRINS (Growing Resilient, INclusive and Sustainable) foundation; he received funding from the European Union Next-GenerationEU (NRRP - MISSION 4, COMPONENT 2, INVESTMENT 1.3 - D.D. 1558 11/10/2022, PE00000018, Spoke 3).

Biases in inequality of opportunity estimates: measures and solutions.

Domenico Moramarco*Paolo Brunori†, Pedro Salas-Rojo‡

August 2024

Abstract

In this paper we discuss some limitations of using survey data to measure inequality of opportunity. First, we highlight a link between the two fundamental principles of the theory of equal opportunities – compensation and reward – and the concepts of power and confidence levels in hypothesis testing. This connection can be used to address, for example, whether a sample has sufficient observations to appropriately measure inequality of opportunity. Second, we propose a set of tools to normatively assess inequality of opportunity estimates in any type partition. We apply our proposal to Conditional Inference Trees, a machine learning technique that has received growing attention in the literature. Finally, guided by such tools, we suggest that standard tree-based partitions can be manipulated to reduce the risk of compensation and reward principles. Our methodological contribution is complemented with an application using a quasi-administrative sample of Italian PhD graduates. We find a substantial level of labor income inequality among two cohorts of PhD graduates (2012 and 2014), with a significant portion explained by circumstances beyond their control.

JEL: C38, D31, D63

Keywords: Equality of opportunity, Machine learning, PhD graduates, Compensation, Reward.

Acknowledgements: We are grateful to Chico Ferreira for his precious advice and comments. We also thank the participants of the conferences: Ifo Conferece Conference on Understanding Socio-Economic Inequalities with Novel Data and Method; Higher Education and Equality of Opportunities, University of Modena and Reggio Emilia; Equality of Opportunity and Intergenerational Mobility: A Global Perspective, University of Bari; Tenth ECINEQ Meeting, Aix-en-Province. Domenico Moramarco gratefully acknowledges ECARES and FNRS for the financial support during

*University of Bari - Department of Economics and Finance. Email: domenico.moramarco@uniba.it

†University of Firenze & London School of Economics - International Inequalities Institute. Email: P.Brunori@lse.ac.uk

‡London School of Economics - International Inequalities Institute. Email: P.Salas-Rojo@lse.ac.uk

the earlier stages of this project. He also acknowledges the GRINS (Growing Resilient, INclusive and Sustainable) foundation; he received funding from the European Union Next-GenerationEU (NRRP - MISSION 4, COMPONENT 2, INVESTMENT 1.3 - D.D. 1558 11/10/2022, PE00000018, Spoke 3).

1 Introduction

In recent decades, there has been a notable rise in empirical analyses of social justice, particularly concerning equality of opportunity. The Equal Opportunity (EOp) paradigm postulates that inequalities attributed to factors beyond individual's control -circumstances- are unfair and should be eliminated as much as possible. Simultaneously, inequalities due to individual efforts or responsibilities may be considered acceptable because they reflect factors for which the society is willing to hold individuals responsible. These two ideas are respectively known in the literature as compensation and reward principles. Beyond the wide theoretical reasoning developed by prominent political philosophers such as Rawls (1973), Sen (1980) and Dworkin (1981), analyses of EOp lie on compelling empirical evidence that people disapprove inequalities rooted in circumstances more than inequalities arising from choices (Cappelen et al., 2007, 2013).

Most of the early theoretical contributions proposed alternative interpretations of the compensation and reward principles (see Fleurbaey, 2008, for a book-length discussion), exploring their compatibility. In this paper we employ a prominent version of the so-called ex-ante approach to EOp, where reward and compensation principles are compatible. Ex-ante EOp, in its prominent interpretation, is realized when individuals with the same circumstances have the same expected outcome (Ramos and Van de Gaer, 2016). To measure inequality of opportunity (IOp) we follow Roemer (1998) and conceptualize a population as the union of many different types of individuals. Each type is composed of individuals sharing all relevant circumstances such as sex, socioeconomic background or place of origin. If indi-

viduals within a type obtain different outcomes only because they make different choices, then all within-type inequality is ethically acceptable and IOp corresponds to exclusively the between-type inequality.

When IOp is measured as inequality between the average income across types, as assumed in this paper, estimates may suffer from two types of biases. Intuitively, since fewer types lead to lower IOp, partial observation of circumstances induces a downward bias.[1] Given that sample data tend to lack information on all desired circumstances, many researchers (Checchi and Peragine, 2010; Ferreira and Gignoux, 2011; Roemer and Trannoy, 2016; Ramos and Van de Gaer, 2016) considered IOp estimates as lower bounds for the real IOp. More recently, Brunori et al. (2019) showed that measurement errors in estimating types' expected outcomes, including errors due to high sampling variance in small sample sizes, may also cause an upward bias in IOp estimates.[2]

The debate about these opposing biases highlights the importance of type partitioning when studying EOp with survey data. To reduce the downward bias, we should define more types. However, a finer type partition is likely to result in poorly estimated average incomes for each type, biasing IOp upwards. Recently, Brunori et al. (2023) proposed using data-driven approaches, particularly conditional inference regression trees (CITs), to identify the type partition that best balances these biases. Since then, the use of machine learners has become a best practice for robust IOp estimates.

This paper contributes to the literature in three ways. First, we shed new light on the sources of upward and downward biases by highlighting an interesting link between type I and type II errors in hypothesis testing, and the two normative principles of compensation and reward. This allows us to define two indices, the Reward and the Compensation scores, which inform us about the risk of upward and downward biases in IOp estimates,

[1] At the extreme, IOp is zero if only one type is defined.
[2] We come back on this in Section 2.1 and in Appendix A.

respectively.

Violations of the reward principle, reflected in a low Reward score, occur whenever a type partition contains types with statistically indistinguishable mean incomes. In such cases, IOp estimates are upward biased because they capture part of the inequality within types. Conversely, violations of the compensation principle, reflected in a lower Compensation score, can be due to unobservable circumstances (as well-documented in the literature) or to a lack of statistical power, which is necessary to identify differences in types' expected incomes. Since statistical power depends on the sample size, a by-product of our contribution is a criterion to establish whether we have sufficient data to estimate IOp with a given set of circumstances.

The second contribution of this paper is to discuss how CITs, a popular data-driven approach to obtain Roemerian partition, deal with compensation and reward violations. In doing so we discuss how a CIT can partly control for reward violation, by setting the desired confidence level to allow splits, but, can easily violate compensation, producing conservative estimates, especially in cases in which the sample size is small. In discussing this we also highlight two structural limitations of using recursive binary splitting to obtain types' partition: the structural impossibility to test and identify all types, and the possibility that types with same expected income appear in different sub-branches of the tree.

The third contribution of this paper is an algorithm that takes as input any type partition and modifies it to improve the Compensation and Reward scores. This algorithm - the Opportunity tree (O-tree) - merges types that have statistically indistinguishable averages (thus reducing the upward bias) and, whenever possible, it splits a type in new ones that have different expected income (thus reducing the downward bias).

We conclude with an empirical exercise investigating IOp among Italian PhD graduates

in 2012 and 2014 (ISTAT, 2018). Surprisingly, labor income inequality is found to be high among this select group of individuals, reaching 0.261 Gini points. We estimate IOp using sex, area of origin, parental education, and occupation as circumstances. After partitioning the sample into types using CIT, IOp reaches 0.072 Gini points, indicating that the Gini of the variability predicted by circumstances is close to one-third of the Gini in the distribution of income. The Reward and Compensation scores, however, are found to be lower than the desired values, suggesting potential upward and downward biases. We apply the proposed O-tree algorithm to modify the CIT's type partition by merging several types with statistically indistinguishable averages, and identifying two new types. As expected, the resulting type partition performs better in terms of Reward and Compensation scores. Interestingly, despite obtaining fewer types (7 instead on the 12 initial ones), the O-tree leads to a slightly higher IOp of 0.073 Gini points.

The reminder of the paper is structured as follows. Section 2 introduces the definition of inequality of opportunity we focus on, discusses the issues of estimating it using sample data, and elucidates the connection of its basic normative principles – reward and compensation – with the type I and type II errors. Section 3 introduces the CIT algorithm and discusses its limits in dealing with violations of the compensation and the reward principles. Section 4 shows how to measure the risk of these violations, and introduces the O-tree as a potential solution. Section 5 compares CIT and O-tree in an empirical illustration. Section 6 concludes.

2 From statistical to normative principles

2.1 Measuring inequality of opportunity with sample data

Following the philosophical debate on responsibility-sensitive egalitarianism and John Roemer's formal definition of EOp (Roemer, 1998; Fleurbaey, 2008), several economic models of EOp have been proposed in the literature (see Ramos and Van de Gaer, 2016; Roemer and Trannoy, 2016, for recent surveys). Theoretically, EOp roots in two independent normative principles. First, the *compensation principle* formulates that people should be compensated for unequal opportunities associated with circumstances. A prominent formulation of this principle – ex-ante compensation – postulates that opportunity sets should be equalized across people with different circumstances. This compensation principle does not consider the role of effort, as it evaluates opportunities *before* individual's degree of effort is revealed. This interpretation, quite prominent in the literature, characterizes our definition of ex-ante EOp. The second principle at the base of EOp is that individuals should be rewarded for their differential efforts, so that inequalities purely due to effort are fair. This is the *reward principle*, whose most prominent version steams from the principle of utilitarian reward stating that inequalities in outcomes among individuals with the same circumstances are a matter of moral indifference.

Formally, let $Y^N = \left(y_i^N\right)_{i=1}^{|N|}$ be the income distribution of a population $N = \{1, 2, ..., |N|\}$ and $T^N = \{t_1^N, t_2^N, ..., t_{|T^N|}^N\}$ be its partition into non-intersecting types (*type partition*). The type partition is obtained after dividing N into subgroups of individuals with the same circumstances. Formally, let $C = \left\{c_1, ..., c_{|C|}\right\}$ be the set of individual characteristics beyond their responsibility sphere, like sex at birth, father education and ethnicity. The elements of C are assumed to be the result of a social debate about the unfair sources of inequality. Each circumstance $c_j \in C$ can take a number $|c_j| \in \mathbb{N}_{++}$ of possible values;

let us denote with $c_j(i)$ the value taken by characteristic j for individual i. Then, any $t_k^N \in T^N$ is a subset of N such that $\forall i, j \in t_k^N$, $c_h(i) = c_h(j)$ for all $c_h \in C$.

As also shown in Fleurbaey and Peragine (2013), the combination of ex-ante compensation and utilitarian reward leads us to measure IOp as inequality in the *smooth distribution*, where each individual's outcome is replaced by the average outcome of the type she belongs to. Clearly, the smooth distribution is function of $\left(Y^N, T^N\right)$. Letting I denote a generic inequality measure like the Gini coefficient, we have that $I\left(Y^N, T^N\right)$ measures the IOp in the population of interest. In line with this approach, income variability within a type should not affect the IOp measure.

In general, $I\left(Y^N, T^N\right) \neq I\left(Y^N, T'^N\right)$ whenever $T^N \neq T'^N$, so that if one observes only a subset of the relevant circumstances C, then, even if we know the income distribution of the entire population, we may still be unable to measure the *true* IOp. As also noted in Ferreira and Gignoux (2011), we incur in a downward biased IOp measure.

The problem of unobserved circumstances is known in the literature since long, and the solution (more detailed individual information) is easily understood but still a challenge in practice. Recent developments in the literature (Brunori et al., 2019) underlined how sample size and sampling errors may also cause upward biases of the IOp estimate. Intuitively, when the average incomes of types are estimated with errors (for example, due to small sample sizes), these errors can increase the inequality between the types' average incomes, thereby increasing IOp. In Appendix A, using the data described in Section 5, we show how upward biases can prevail, depending on the available sample size.

The problems of missing types (or unobserved circumstances) and limited sample size, with their consequent downward and upward biases, have primarily been considered as statistical issues limiting the reliability of IOp estimates. In the following subsection, we examine these biases from a more theoretical perspective. This approach allows us to

describe them more clearly and, in some aspects, in a new way. We clarify how these biases are tightly connected to the normative principles of compensation and reward, and we reinterpret type I and type II errors in hypothesis testing as risks of violating the reward and compensation principles.

2.2 A normative approach to upward and downward biases

Let Y^S denote the income distribution of a sample $S \subset N$. To clarify the source of potential biases, let us call $T^S = \{t_1^S, t_2^S, ..., t_{|T^S|}^S\}$ the type partition obtained when we observe all circumstances in C, so that $|T^S| = |T^N|$ and, for all $1 \leq j \leq |T^N|$, $t_j^S \subset t_j^N$. Observe that T^S is a hypothetical type partition in which estimation biases – $I\left(Y^S, T^S\right) \neq I\left(Y^N, T^N\right)$ – may come from the sampling process but not from unobserved circumstances.[3]

To fix ideas, suppose that we are interested in measuring ex ante income IOp due to biological sex and race, where both circumstances take two possible values: male or female, and black or white. The population is then divided in four types: $T^N = \left\{t_{MW}^N, t_{MB}^N, t_{FW}^N, t_{FB}^N\right\}$. Let types be of the same population size, and assume that $t_{MW}^N, t_{FW}^N \sim N\left(x, \sigma\right)$ and $t_{MB}^N, t_{FB}^N \sim N\left(x + \Delta, \sigma\right)$, for $x, \Delta > 0$. In words, white (resp. black) men and women have incomes drawn from the same distribution, so they share the same opportunity set. Conversely, there is inequality of opportunity induced by race, since black individuals have higher expected income.

We draw an equal sized sample of observations from each type in T^N. Then, we have $T^S = \left\{t_{MW}^S, t_{MB}^S, t_{FW}^S, t_{FB}^S\right\}$, and $t_j^S \subset t_j^N$ for all j. Let $\mu\left(t\right)$ denote the arithmetic average of the incomes of individuals in t. Despite our assumptions, there exists $Y^S \subset Y^N$ such that $\mu\left(t_{MW}^S\right) > \mu\left(t_{FW}^S\right) = \mu\left(t_{MB}^S\right)$.[4]

[3] The blue solid line in Figure A1 highlights that, since averages are poorly estimated with small samples, we tend to have $I\left(Y^S, T^S\right) > I\left(Y^N, T^N\right)$.

[4] A complete numerical example - with $x = 500$, $\sigma = 50$, $\Delta = 10$ - is available upon request.

In such a sample, the estimate $I\left(Y^S, T^S\right)$ is inaccurate for two reasons. First, since $\mu\left(t^S_{MW}\right) \neq \mu\left(t^S_{FW}\right)$, and given the assumption that white men and women have the same opportunities, part of $I\left(Y^S, T^S\right)$ is inequality due to effort. Including these inequalities among the unfair ones is a violation of the reward principle, and a particular manifestation of the upward bias.[5] Second, $I\left(Y^S, T^S\right)$ fails to include the difference between the expected incomes of white women and black men, violating the compensation principle. We should stress here that this is different that the downward bias highlighted by previous studies because, by assumption, all relevant circumstances are observed. Yet, failing to capture the difference in expected incomes of t^N_{FW} and t^N_{MB} reduces the estimated IOp.

The remainder of this section shows that these violations of the reward and compensation principles are linked to, and can be measured by, the types I and II errors in hypothesis testing.

Our departing assumption is that we are working with sample data, which is typically the case in IOp empirical analyses. Thus, before proceeding with IOp measurement, we should test if there is a statistically significant difference between expected incomes of types. Indeed, if we cannot be confident in claiming that types have different average incomes, then an $I\left(Y^S, T^S\right) \neq 0$ will not be a good estimate of $I\left(Y^N, T^N\right)$, as it could, with an high probability, reflect sampling variance. With this concern in mind, we may rely on simple statistical tools to test whether the average incomes of two types are different.

The most straightforward approach is to test, for each pair $t^S_j, t^S_k \in T^S$, the null hypothesis $H_0 : \mu\left(t^N_j\right) - \mu\left(t^N_k\right) = 0$ against $H_A : \mu\left(t^N_j\right) - \mu\left(t^N_k\right) \neq 0$. This is easily done by computing the statistic

[5] Another instance of upward bias occurs if, for example, we overestimate IOp between white and black men which, by assumption, have different opportunity sets. Formally, this happens if $\left|\left(t^S_{MW}\right) - \left(t^S_{MB}\right)\right| > \left|\left(t^N_{MW}\right) - \left(t^N_{MB}\right)\right| = \Delta$.

$$Z = \frac{\mu\left(t_j^S\right) - \mu\left(t_k^S\right)}{\sqrt{\frac{\sigma\left(t_j^S\right)^2}{\left|t_j^S\right|} + \frac{\sigma\left(t_k^S\right)^2}{\left|t_k^S\right|}}}$$

where $\sigma(t)$ denotes the standard deviation of incomes in group t. By the Central Limit Theorem, Z follows a standardized normal distribution.[6] From standard practices in hypothesis testing we know that the choice of rejecting H_0 is guided by the tolerable risk of Type I error (α): rejecting H_0 when it is true. Formally, if we denote Φ the cumulative standardized normal distribution and Φ^{-1} its inverse, for a given α, the critical values of our hypothesis testing are $\Phi^{-1}(\alpha/2)$ and $\Phi^{-1}(1-\alpha/2)$. Therefore, we reject H_0 if $Z \leq \Phi^{-1}(\alpha/2)$ or $Z \geq \Phi^{-1}(1-\alpha/2)$.

In the context of IOp measurement, if H_0 is true, like for white male and white female in the above example, then α represents the probability of violating the reward principle if t_j^S and t_k^S were treated as different types when estimating IOp.

Hypothesis testing is also sensitive to the Type II error: not rejecting H_0 when it is false. If this error occurs, like for white female and black male in the above example, then we are wrongly concluding that t_j^N and t_k^N have the same average income, implying no IOp between them. The probability of Type II error is denoted as β, and the *power* of a test $(1-\beta)$ is, in our context, the ability to detect a difference in expected income, under the assumption that there actually exist one. Formally, the power of our test is expressed as:

$$\begin{aligned} 1-\beta = \quad & 1 - \Phi\left(\Phi^{-1}(1-\alpha/2) - d\right) \quad + \Phi\left(\Phi^{-1}(\alpha/2) - d\right) \\ \text{with} \qquad & d = \frac{\epsilon}{\sqrt{\frac{\sigma\left(t_j^S\right)^2}{\left|t_j^S\right|} + \frac{\sigma\left(t_k^S\right)^2}{\left|t_k^S\right|}}} \end{aligned} \tag{1}$$

[6] We are assuming, for simplicity, that types are sufficiently big, so we can write the formulas referring to Normal distributions rather than T-students distributions.

It is easy to observe that power rises along with sample size, and decreases with higher variability in income. Another key determinant of $1-\beta$ is ϵ, which is the *true* difference between the expected incomes of the two types. The bigger the ϵ, the easier will be for the test to detect such a difference, as it will also be less likely not to reject H_0. Finally, necessarily, β is inversely related to α.

High power is essential for measuring IOp correctly. For example, if the sample size is too small, then we may end up not rejecting a false H_0 simply because we do not have enough observations to detect the difference in average incomes. Similarly, if $|\mu\left(t_j^N\right) - \mu\left(t_k^N\right)|$ is positive but close to zero, a high sample variance may not allow us to detect such difference. A test with a high probability of Type II error is more likely to conclude that there is no IOp between types j and k. Indeed, even if $|\mu\left(t_j^N\right) - \mu\left(t_k^N\right)| = \epsilon > 0$, with high β we are likely not to reject H_0. Consequently, β can be interpreted as the probability that, when measuring IOp in our setting, we are violating the compensation principle.[7]

2.3 Normative versus data-driven type partitions

In Section 4.1 we build on the previous remarks to propose some criteria for an overall assessment of a type partition in terms of the compensation and reward principles. From our discussion so far it should be clear that upward and downward biases, as well as violations of the reward and compensation principles in IOp estimates, are related to the specific type partition we choose to measure IOp.

Most theoretical contributions (see Ferreira and Peragine, 2016, for a survey) assume that there exists a normative partition of the population into types. This partition, taken as the same across time and space, results from social and political debates about what should and should not be considered a circumstance. In the previous section, the normative type

[7] The reader may notice that this probability is conditioned on H_0 being false. In other word, it's the probability of not detect a true difference in means.

partition corresponds to T^N. In line with this approach, there exists only one relevant type partition for IOp estimates: the above-defined partition T^S.

As already mentioned, T^S is de facto a theoretical construct, because real-world data typically fail to contain all the individual characteristics that constitute circumstances. All existing IOp estimates are therefore based on type partitions that differ from T^S. This raises issues of comparability across time and countries, either because different databases may contain different (or differently coded) circumstances, or because researchers merge types to obtain more robust estimations of the means (an example, among many, is Checchi and Peragine, 2010).

Following (Brunori et al., 2019, 2023; Escanciano and Terschuur, 2023), the literature has experienced an rising interest in machine learning techniques, like conditional inference regression trees (CITs) and forests, which are implemented to construct data-driven type partitions. These approaches limit arbitrariness in choosing the type partition for IOp estimation, and tend to display better statistical properties. These are both welcomed features for practitioners interested in obtaining statistically robust estimates of IOp.

Skeptics of data-driven approaches defend the idea that IOp is, first of all, a socio-economic issue that can only be tackled when types are precisely defined. Data-driven type partitions may be unstable, as they depend on the sample at hand, and are not always in line with a politically accepted way of grouping individuals. Moreover, even if in the classic approach researchers directly influence the type partition, their choices are more explicit and can be defended on the basis of ethical and practical principles.

On the other hand, one may criticize the idea that the type partition should be the same across time and space. Different societies have different income-generating processes, which are hard to describe with a unique type partition. For example, casts and religion may matter in some countries but not in others. With data-driven approaches, there is no prior

assumption about the structure of society, so that the estimated IOp may be more in line with the country and time specific structure of the socio-economic interactions.

The debate between the normative and data-driven approaches is still open and this paper does not aim at resolving nor taking a stance on it. It however contributes to this debate by showing that, independently of where the type partition comes from - normative approach like in the previous section or data-driven approach like in the following - the simple fact that we estimate IOp using sample data generates biases (violations of the reward and compensation principles) which can be measured, as shown in Section 4.1, and limited, as shown in Section 4.2.

In the following section, we move our focus toward data-driven partitions, specifically the one resulting from CIT. This choice is aimed at (1) underlining structural limitations of the CIT algorithm that have been so far overlooked by the literature, and (2) showing that these limitations result in potential upward and downward biases (i.e. violations of the reward and the compensation principles).

3 Biases in the Conditional Inference Tree

We have seen that IOp estimates may violate the principles of reward and compensation even if they are based on a normative type partition like T^S. Here we argue that data-driven type partitions may suffer from the same biases. We depart from a closer look at conditional inference regression trees, that have received great attention by the recent literature.

3.1 The Conditional Inference Tree

Conditional Inference Trees (CITs), developed by Hothorn et al. (2006), have been recently adopted for IOp estimation (see Brunori et al., 2023, for example). Besides balancing upward and downward biases widely discussed by the literature (Ferreira and Gignoux, 2011; Brunori et al., 2019) this algorithm limits arbitrariness when defining the type partition where IOp is measured. CITs are supervised machine learning algorithms that aim at predicting a dependent variable by partitioning the space of regressors (circumstances). In doing so, they define groups of non-overlapping observations (types) that are homogeneous in terms of the realization of a subsample of the observable predictors. Individual predictions are made by taking the average outcome within groups.

The CIT algorithm takes as input a list of circumstances C and a parameter $\tilde{\alpha}$. At each iteration, the algorithm performs a series of independence tests to identify the circumstance variable that is most associated with income[8]. Formally, for all $\forall c_j \in C$, it tests the null hypothesis that conditional and unconditional income distributions are statistically indistinguishable – $H_0^{c_j} : D(Y^N|c_j) = D(Y^N)$. The resulting p-values are Bonferroni-adjusted. [9] If no adjusted p-value is lower than $\tilde{\alpha}$, then the algorithm stops. Otherwise, the algorithm selects the c_j associated with lowest p-value, and performs a binary split of the population according to the possible values of c_j. The choice of how to split the population is made after performing a series of permutation tests, which in our case corresponds to non-parametric versions of the difference in mean t-test. The performed split is, again, the one with the lowest p-value. Within each resulting sub-populations, the algorithm starts from the independence test phase to attempt further splits. A stopping rule is triggered when the degree of association between the dependent variable and all controls is not sufficiently

[8] As explained by Varian (2014) using a correlation test to select the splitting control eliminates the well-known bias of standard trees in selecting variables with a high number of values or categories.

[9] This adjustment is necessary because of multiple hypothesis are tested in this step.

significant to reject the null hypothesis of independence. In that case, no further split is performed, preventing the model from growing too deep and overfit the data.

The above algorithm gives full control over the parameter ($\tilde{\alpha}$) governing the decision of whether or not to further split the sample. Lowering the $\tilde{\alpha}$ value results in a more conservative tree, with fewer splits and terminal nodes. Standard practices consist in fine-tuning $\tilde{\alpha}$ to improve the out of sample predictions of the model. From a normative perspective, however, there is no direct way of assessing whether the resulting type partition (i.e. the groups defined by the terminal nodes of the tree) respects the compensation and reward principles discussed above. This is because once an independence test passes, a split *must* be performed, independently of the type I and II errors involved in this split. Moreover, the independence test is more demanding than a t-test, in the sense that it is possible that a final node fails to reject the null hypothesis of the independence test while it could still be split according to a difference in means test.

3.2 The flaws of a binary splitting algorithm

As explained, upward and downward biases are statistical manifestations of violations of the normative principles at the base of EOp. We focus now on whether CIT algorithms delivers type partitions that maximize consistency with the compensation and reward principles.

Let us consider a sample S for which we observe only two circumstances: gender (men or women) and skin color (brown or white). Let us call $T_0 = \{t_{MW}, t_{MB}, t_{FW}, t_{FB}\}$ the *full* type partition, that is one in which we divide the population by sex and skin color.[10] In such a setting, the standard approach in the literature (Checchi and Peragine, 2010; Ferreira and Gignoux, 2011) is to estimate ex-ante IOp by $I\left(Y^S, T_0\right)$.

[10] In this section, we do not use the superscript S when referring to types because these may not be subsets of the types in T^N, in particular if $|C| > 2$.

Advocates of data-drive approaches, aware of the potential estimation biases, prefer to let an algorithm like CIT choose the type partition. Indeed, the above sample can be partitioned in several ways other than T_0: fifteen in this case.[11] A hardly mentioned, yet theoretically relevant, limitation of CITs is that, due to its deterministic structure, it chooses type partitions from a subset of the theoretically possible ones. In the previous example, this subset contains only half of the potential partitions. This limitation becomes even more stringent when more circumstances (or categories) are considered (Appendix C provides more details on this issue).

CITs, and any other partitioning algorithm based on binary splitting, are irreversible procedures, in the sense that once a split is performed, the algorithm generates two distinct populations and looks for further possible types within each of them. Consequently, it does not check for the possibility that in the final type partition there exists two or more types with the same average income. Also, since each split is based on the maximum degree of association in the considered sample, the algorithm does not take consider the effect of a split on potential further splits in the two sub-trees it generated. Following our discussion in Section 2.2, this may come at the cost of a more severe violation of the reward principle.

Indeed, it is even possible that in the final type partition one finds two resulting types with statistically indistinguishable densities.[12] Since IOp is computed on the smoothed distribution, this event may have no direct impact on the resulting measure. However, it may indirectly impact estimates by leading to a different type partition.

Let us clarify the previous point with an example. Suppose the CIT algorithm identifies T_0 as the relevant type partition. By construction, this algorithm will never test for the difference in means between t_{MB} and t_{FW}. It may however be the case that, not only the average, but the entire income distribution of these types coincide. Let $T_\star =$

[11] See Appendix C for a deeper exploration on the type partition limitations.

[12] Real-data examples of this are available upon request.

$\{t_{MW}, t_{FB}, \{t_{MB} \cup t_{FW}\}\}$. We may observe that, in this case, $I\left(Y^S, T_0\right) = I\left(Y^S, T_\star\right)$, so that both type partitions lead to the same IOp estimate. Suppose now that there is a third circumstance, say the region of birth, which is found to have no impact on expected income within t_{MB} and t_{FW} separately, but is statistically relevant in $t_{MB} \cup t_{FW}$.[13] In this case, the resulting types cannot appear and our IOp estimate will be based on a type partition that misses a type with different opportunities, violating the compensation principle.

Overall, it seems clear that even if CITs are fine-tuned to reduce estimation biases, they are not immune to violations of the compensation and reward principles biases.

4 Identifying the salient type partition

The discussion so far points out that, whenever we measure IOp using sample data, the implemented type partition and the resulting IOp estimate may violate the reward and compensation principles. In what follows, building on Section 2.2, we show that the risk of these violations can be both measured, - via the Reward and the Compensation scores - and reduced - by the Opportunity tree algorithm.

4.1 Compensation and Reward scores

Let us denote with $T = \left\{t_1, t_2, ..., t_{|T|}\right\}$ a type partition which we plan on using to estimate IOp. It is worth explaining here the change in notation with respect to Section 2. We have defined T^N as a normative partition of the population where IOp should be measured, and T^S the partition of the sample into $|T^N|$ types, which are all subsets of the population's ones. Clearly, for T^S to be well defined, one needs to have information about all circumstances in C. Moreover, even if we do have such information, sample size issues

[13] The example in Figure A2 shows how this can happen because of sample size.

may force us to use a partition, which we denote T, that differs from T^S.[14] When this is the case, it is natural to question the normative relevance of the IOp estimate $I\left(Y^S, T\right)$. While there is no doubt that when $T \neq T^S$ there is room to claim that $I\left(Y^S, T\right)$ is not a perfect estimate of $I\left(Y^N, T^N\right)$, we believe that $I\left(Y^S, T\right)$ is still informative about IOp in the population N if we can claim that T minimizes the risk of violating the reward and compensation principles.

To determine whether T is a sensible type partition we propose to check that *all types in T are salient, and no salient type is missing from T*. Formally, a type t is "salient" if (1) individuals in t have different circumstances than individuals in other types and (2) for all $t' \neq t$, the average income of individuals in t is different than the average income of individuals in t'. If all types in T are salient, then T does not violate the reward principle. Conversely, suppose there are $t, t' \in T$ such that $\mu(t) \neq \mu(t')$ but $|\mu(t) - \mu(t')|$ is not statistically different from zero. In such a case, $I\left(Y^S, T\right)$ will capture part of the inequality due to effort, which violates the reward principle. To assess this risk we propose a *Reward score* defined as :

$$S_R\left(T\right) = \sum_{(t,t') \in T \times T : \mu(t) \neq \mu(t')} \omega(t, t') \times \left(1 - \text{p-value}\left(t, t'\right)\right) \tag{2}$$

where $\omega(t, t') = \left[\, |t| + |t'| \,\right] / \left[\sum_{(t,t') \in T \times T : \mu(t) \neq \mu(t')} |t| + |t'|\right]$ is the relative sample size of the considered pair of types, and $\text{p-value}\left(t, t'\right)$ is the p-value of the difference in means t-test for the considered pair. Observe that, if T contains two types t, t' such that $\mu(t) = \mu(t')$, this will have no impact on the estimated IOp; for this reason Eq. (2) focuses only on pairs of types with different arithmetic mean.

The Reward score is a weighted sum of the statistical significance of the difference between

[14] This is particularly true when the partition is data-driven.

average incomes of types in T. In other words, it measures how confident we can be in claiming that types in T are salient. The higher is $S_R \in [0,1]$, the more the type partition is in line with the reward principle. Following standard practices in the statistical literature, we can easily define values $\underline{S_R} \in [0,1]$ such that, whenever $S_R(T) \geq \underline{S_R}$, we can conclude that T is in line with the reward principle. We believe $\underline{S_R} = 0.95$ to be a sensible threshold, given that 95% confidence intervals are standard in empirical analyses.

Let us now consider the risk that T is missing salient types, downward biasing the estimated IOp and inducing violations of the compensation principle. To assess this, for each $t \in T$, we propose to test if it is possible to binary split this sub-population in two groups, t_j^1 and t_j^2, that are potentially salient types.

Formally, we implement a two step procedure. First, within each $t \in T$, we run a CIT with $\tilde{\alpha} = 1$ and only one split allowed. In this step, a split will be performed because any p-value will be smaller than 1. Clearly, since we are using CIT, this is not going to be a random split, but the one performed by the circumstance most associated with the outcome of individuals in t. Alternatively, one can follow a theoretically driven reasoning to perform the previous binary split. The second step consists in testing the difference in means of the resulting groups, t^1 and t^2 using a standard t-test. We then collect the p-value and estimate the power of this test.[15] The lower the p-value, the more confident we are that t^1 and t^2 should constitute different types, so that we violate the compensation principle by not splitting t. The higher the power, the more confident we are that the final node t has sufficient sample size to identify more types, so that the previously computed p-value provides reliable information.

Let power(t^1, t^2) denote the power of the difference in means t-test conducted on the potential types t^1 and t^2; the p-value of the same test is denoted p-value(t^1, t^2). The power

[15] To compute the power of the test we need to define a set of parameters. We refer to Section 5 for more details.

and p-value computed from a given type are informative per-se. As discussed in Section 2.2, a test with low power suggests, under the assumption of false null hypothesis, potential violations of the compensation principle. Thus, p-value(t^1, t^2) addresses the validity of that assumption, while power(t^1, t^2) measures the risk of violations.

When power(t^1, t^2) is sufficiently high, with high p-value(t^1, t^2) we can confidently claim that t^1 and t^2 should *not* constitute different types. Conversely, when p-value(t^1, t^2) is close to zero, with high power we are confident that t^1 and t^2 should constitute different types. While the former instance is desirable, the latter is a violation of the compensation principle. When power(t^1, t^2) is low, then we do not have sufficient information to assess whether t^1 and t^2 should be different types. While this signals the risk for violations of the compensation principle, this risk decreases with p-value(t^1, t^2). Indeed, one may be less worried about insufficient sample size when p-value(t^1, t^2) is sufficiently high, signalling that t^1 and t^2 have similar means.

In most applications we have partitions composed of many types, so that analysing and comparing powers and p-values within each terminal node may become cumbersome. To account for the interaction between power and p-value in determining whether there is a potential violation of the compensation principle coming from a given $t \in T$, we suggest to look at the product between the two. Moreover, to simplify the normative assessment of a type partition, we propose the following *Compensation score*:

$$S_C(T) = \sum_{t \in T} \frac{|t|}{\sum_{t \in T} |t|} \big(\text{power}(t^1, t^2) \times \text{p-value}(t^1, t^2)\big) \tag{3}$$

where, for all $t \in T$, t^1 and t^2 denote the two types resulting from binary splitting t, $|t|$ is the number of observations in t, and power(t^1, t^2) and p-value(t^1, t^2) are defined as before. The Compensation score is a simple weighted sum of the scores of each type, for

weights depending on the relative sample size. This weighting system emphasizes violations occurring in populated subgroup, as these are likely to have stronger effect on the estimated IOp.

The interpretation of the Compensation score is straightforward: it is an aggregate measure of how confident we are that the given partition is not missing salient types. The higher is $S_C \in [0,1]$, the more a type partition alineates with the compensation principle. Also in this case, we may define a value $\underline{S_C} \in [0,1]$ such that, whenever $S_C(T) \geq \underline{S_C}$ we can conclude that T is in line with the compensation principle. We suggest to define $\underline{S_C} = \underline{pow}\left(1 - \underline{S_R}\right)$, where $\underline{pow}$ is a minimum acceptable power for a test. Since the statistical literature tends to use 0.8 as reference, we take $\underline{S_C} = 0.8\,(1 - 0.95) = 0.04$ to be a reasonable threshold for our Compensation score.

The reader may notice that Eq. (2) and (3), as well as $\underline{S_C}$ and $\underline{S_R}$, are to some extent arbitrary. A deeper analysis and axiomatization of the Reward and Compensation scores is out of the scope of this chapter. Here, we remain in line with the statistical meaning attached to S_R and S_C and consider the product a natural way of aggregating power and p-value across types. More sophisticated scores can be easily constructed; for example, we could replace $(1 - \text{p-value}(t, t'))$ in Eq. (2) with $f_{\underline{S_R}}\,(1 - \text{p-value}(t, t'))$, where $f_\tau : [0,1] \to [0,1]$ is a function such that $f_\tau(x) = 1$ if $x \geq \tau$ and $f_\tau(x) = \tau^{-1}x$ otherwise.[16] While this score does not solve the arbitrariness issue, it has the interesting property of being equal to 1 only if all types in T satisfy the reward principle.

4.2 The Opportunity-tree

We now introduce a procedure aimed at improving the Reward and Compensation scores of a given type partition. We present this procedure within the context of CIT-based

[16] Similar transformations can be performed to the elements of Eq. (3).

partitions, because it also deals with the limitations discussed in Section 3.2.[17]

We suggest a simple refinement of the standard CIT algorithm, which we call "Opportunity-tree" (O-tree). Our approach consists of complementing CIT with a final step in which we check that all resulting types have statistically different average incomes.[18] Types with the same expected incomes are then sequentially merged. After this merging process is completed, we repeat the CIT procedure within each of the new resulting sub-populations to explore possible new types.

The pseudo code of the O-tree algorithm is given in Algorithm 1. The algorithm can be divided in three blocks: an initial one that runs CIT, a second block that merges similar types, and a third block that attempts further splits in the merged types. The first block (Steps 1 and 2) does not require further explanations, since these are given in Section 3. The reader should however notice that $T^\star$ in Step 2 can be any type partition, not necessarily produced by a CIT.

The second part of the algorithm (Step 6 to 16), which is the core of our proposal, takes as input a type partition and merges types that have statistically undistinguishable expected incomes. Specifically, starting from a type partition, it performs t-tests of differences in means for each pair of types. It merges the pair of types with the highest p-value above a 0.05 threshold, and repeats the series of t-tests for the remaining types. If there are pairs for which the p-value is above 0.05, then it merges the pair with the highest p-value and repeats the procedure. Otherwise, it passes to the following block.

The third block of Algorithm 1 (Step 17 to 19) fits a CIT in each resulting type (after merging) to explore whether more types appear. This CIT must have $\tilde{\alpha} = 1$ to ensure that

[17] These limitations are: (1) a high share of the possible type partitions can never be defined as salient; and (2) it is possible that the salient type partition is composed of types with similar expected income.

[18] This part of the procedure can be modified to test that types have statistically different CDF, instead of focusing only on the averages. Minor changes in the O-tree algorithm are required to accommodate this. Here we keep our focus on the average because we measure ex-ante IOp as between-means inequality.

Algorithm 1 The O-tree algorithm

Input: C
run CIT (with a given $\tilde{\alpha}$) and store the resulting type partition in $T^\star$
$T \leftarrow \emptyset$
while $T \neq T^\star$ **do**
    $T \leftarrow T^\star$
    for all $(t_j, t_h) \in T \times T$, $t_j \neq t_h$, **do**
        test $H_0 : \mu(t_j) = \mu(t_h)$ against $H_A : \mu(t_j) \neq \mu(t_h)$
        store the resulting p-value p_{jh} in the set $P = \{p_{12}, ...p_{1T}, p_{23}, ...p_{2T}, ...p_{(T-1)T}\}$.
    end for
    define $P^\star = \{p_{jh} \in P : p_{jh} > \alpha\}$, for $\alpha = 0.05$.
    if $P^\star \neq \emptyset$ **then**
        let $p_{jh} = \max P^\star$.
        merge types to create $t_{jh} = \{t_j \cup t_h\}$
        update type partition: $T \leftarrow T \setminus \{t_j, t_h\} \cup t_{jh}$
        go to Step 6
    end if
    for all $t_j \in T$ **do**
        run CIT with $\tilde{\alpha} = 1$ and only one split allowed
        store the resulting type partition in $T_j^\star$
    end for
    $T^\star \leftarrow \bigcup_{t_j \in T} T_j^\star$
end while
Return: T

a split is performed, and allow only one split in order to have only two new potential types for each of the initial ones. Finally, the algorithm returns to the second block to check that the new type partition does not violate the reward principle. It then keeps running until the process of attempting further splits, after merging types with similar averages, does not modify the resulting type partition.[19]

The proposed O-tree algorithm deals with the second structural problem of CIT by merging types with similar expected incomes. At the same time, it increases the number of possible

[19]It may be possible that this process enters a loop in which a type is split and then merged repeatedly. If this happens, we force the algorithm to stop after the merge, so that the new types are not generated and the final type partition does not violate the reward principle.

type partitions CIT can detect. Finally, by merging types, we generate groups with higher population share, in which it is easier to detect the effects of circumstances that were hidden by the lack of sufficient observations in the original type partition. Consequently, the O-tree algorithm constitutes both, a technical and a normative improvement with respect to the CIT, since it expands the set of possible partitions explored and is likely to improve the compensation and reward scores of the resulting type partition.

5 Empirical Aplication

In the above text we have proposed a method to assess the reliability of IOp estimates based on sample data. Given the selected circumstances beyond individual control, Compensation and Reward scores are able to tell us whether we have sufficient information to test the degree of association between circumstances and the outcome of interest. We show now how the two scores can guide our analysis of real data.

5.1 Data

We study IOp among Italian PhD graduates using the "Survey on university doctorate holders' vocational integration" carried out by the Italian National Institute of Statistics (ISTAT, 2018), which contains information about PhDs graduated in 2012 (N=8,172) and in 2014 (N=7,882). In such a highly selected population, one would expect to observe low levels of income and opportunity inequality. Therefore, we ask whether circumstances beyond individual control such as area of origin, sex, and socioeconomic background, are predictive of labour income disparities within our sample.

We take the monthly net labor income of the respondents as the outcome of interest, which is assumed to capture labour market success. Labor income is reported in 134 in-

come brackets ranging from "200 euro/month" and "above 7,000 euros", but we treat it as a continuous variable.[20] We consider five key circumstances out of individual control. A binary variable describing self-reported gender (male and female), mother' s and father' s education coded in five categories (1=elementary education or no education, 2=lower secondary, 3=upper secondary, 4=university diploma, 5=university degree or above), mother and father occupation if employed (1=manager, 2=professionals, 3= technicians and associate professionals, 4= clerical support workers, 5=skilled manual worker, 6=elementary occupations), mother and father occupation if self-employed (1=entrepreneur, 2=professionals, 3=technicians, 4=managing directors (family business), 5=manual worker (family business), 6= administration managers, 7=managing directors). Finally, we consider the region of residence before enrolment in an undergraduate university course as the circumstance describing the area of origin of the respondent. To this circumstance, we add two categories: "other country in European Union" and "outside the European Union".

In order to make the outcome variable comparable across cohorts we remove a fixed cohort effect. Understandably doctors graduated two years earlier earn on average 103.8 euro more per month than their younger pairs. Descriptive statistics of both circumstances beyond individual control and outcome are available in Section D.

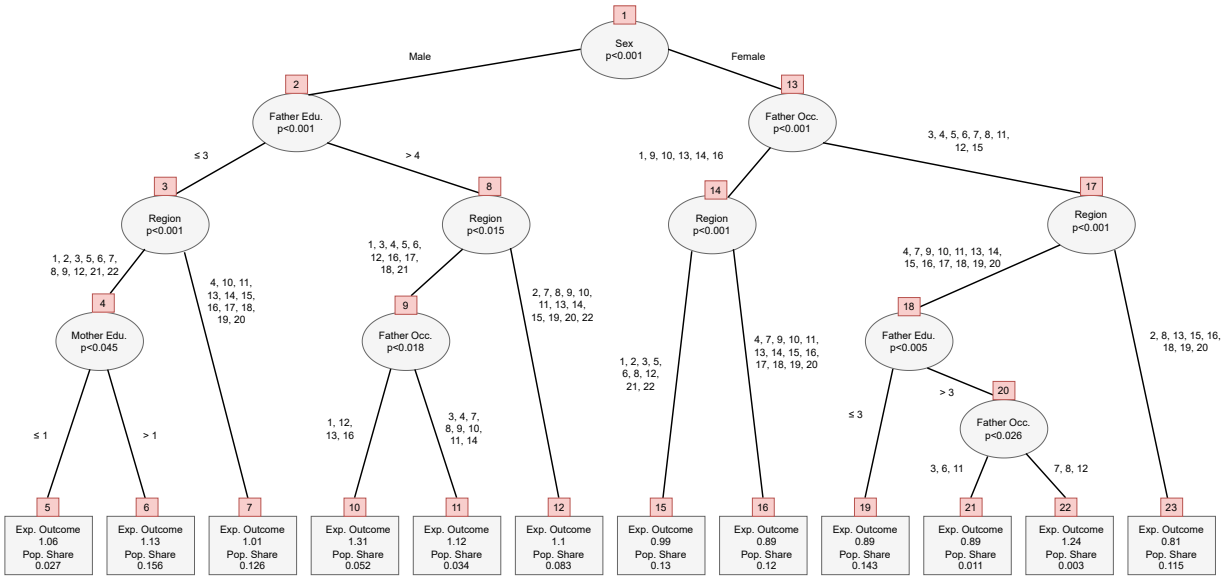
5.2 Main Results

Income inequality in the analysis sample amounts 0.261 Gini points. This value is surprisingly high, given the expected homogeneity of the objective population in terms of education and age. To identify salient types with CITs, we set $\tilde{\alpha} = 0.05$, so the algorithm keeps on splitting until the null hypothesis of independence between the income variability and any observable circumstance can be rejected only with a level of confidence lower than

[20] We assign 7000 euros to observations in the upper bracket.

95%. All remaining parameters defining the tree algorithm are set to the default values (see Hothorn and Zeileis, 2015, for details). We obtain a partition in 12 types reported in Figure 1. The between-type inequality corresponds to our baseline IOp estimate in the resulting type partition, and reaches 0.072 Gini points, meaning about 27.5% relative to total inequality, is inequality that can be accounted for the circumstances employed[21].

Figure 1: Roemerian type partition based on CIT



CIT is estimated setting $\alpha = 0.05$. The outcome is normalized such that 1 = mean sample income, 2072 euros.

Source: Istat, IIPDR 2018

A clear hierarchy of circumstances in terms of their prominence emerges when inspecting the shape of the tree: sex is the first split, dividing the tree in two sub-trees. All terminal nodes made of male have an expected income above the population mean, while none of the types containing women has an expected outcome above the population average, with type 22 being the only exception representing a mere 0.3% of the sample.[22] In the

[21] Note that the Gini index is not a path-independent additive decomposable inequality index, using a perfectly decomposable index such as MLD such share would be smaller.

[22] For clarity, we normalize the type-specific expected outcome with 1 being the sample mean, 2,070 euros.

female sub-tree, the two largest types originate from the splitting node 14, that contains 25% of the sample. Both resulting types are made of PhD graduates whose father has a prestigious occupation (entrepreneur, professional, managing directors, manager), but they differ in their region of origin. Type 15 is endowed with a higher expected outcome (0.99 of the population average), and contains respondents coming from the North of Italy (excluding Trentino-Alto Adige and Liguria), Lazio and outside Italy. Opposed, type 16 mainly contains PhD graduates from the South of Italy, and has an expected income 10% lower.

In the male sub-tree, the richest type (terminal node 10) is made of male PhD graduates with a university graduated father, that also embodies a relatively high-qualification. Respondents in type 10 are generally from Northern Italy (Piemonte, Lombardia, Trentino - Alto Adige, Friuli-Venezia Giulia, Lazio) but also come from some regions in the South (Puglia, Bailicata, Calabria) and the EU.

In addition to the prominent role of sex, circumstances describing the place of origin and the characteristics of the father determine several splits in the tree. Meanwhile, the education of the mother only appears in the sub-tree populated by men. Although the geographical distinction is not always clear, the CIT partition is fairly consistent with the Italian rationale. Coming from the Northern part of the country, Lazio, and from abroad, tends to predict higher incomes than coming frome elsewhere. Father's occupation, although used as an unordered variable, also tends to partition the sample according to a social hierarchy, whereby more prestigious occupations are predictive of higher income, as do the educational levels of both the father and mother.

Remarkably, the types' partition obtained with the CIT produces a number of types with almost indistinguishable expected outcome (terminal nodes 16, 19, and 21). As mentioned above, this is an undesired effect of using recursive binary splitting to obtain a partition

in salient types.

We perform the normative assessment of the type partition in Figure 1 by implementing the procedure discussed in Section 4.1. Table 1 reports the $(1 - \text{p-value})$ is computed for each pair of terminal nodes in Figure 1, together with it's weighted counterpart. There are several pairs of types, such as the combination $Type1 = 5$ and $Type2 = 7$, for which we cannot reject the null hypothesis of equal expected income (1 - pvalue = 0.87).

Table 2, instead, reports the power and p-value computed within each terminal node of Figure 1. To compute power one needs to set the parameters in Eq. 1. For each terminal node we use the respective sample size and empirical standard deviation, and set ϵ to be the difference between the average income in the two resulting types. In doing so, the parameter d corresponds to the standard "Cohen's d". Finally, α is set to 0.05, which corresponds to our favourite threshold for the reward score. In Table 2, there are a few instances where the combined value of power and p-value for the first non-performed split arise suspicion. For example, type 23 has sufficient sample size to test for further splits, and the first non-performed split in this node has a very low p-value. This observation suggests the possibility of refining the CIT algorithm by forcing splits like the one in type 23 to take place.

The Reward Score for the CIT amounts to 0.913 which is just above the lowest $\underline{S_R}$ introduced in Section 4.1. This is in line with the fact that CIT are constructed to somehow avoid violations of the reward principle, so one would expect a good S_R from it. We should however stress that the reward score from CIT falls below 0.95, which is probably the most common confidence level used in practice. The performance becomes even less satisfactory when we consider the Compensation Score. Observe that if we set $\underline{S_R} = 0.95$ and $\underline{pow} = 0.8$, then $\underline{S_C} = 0.04$, while the Compensation Score of the CIT merely reaches 0.0363. Type 21 is particularly worring since the p-value is low and power is also low. The

risk that the split is not performed due to limited sample size (N=157) is high.

Being somewhat insatisfied with the CIT performance, we now estimate the O-tree introduced in Section 4.2. Figure 2 plots the O-tree, where the number of terminal nodes has almost halved: from twelve in the CIT to seven. Specifically, types 5 and 12 have been merged into type 29, type 6 and 11 have been merged into type 26, type 10 and 22 have been merged into type 27, and type 7 and 15 have been merged into type 28. Additionally, type 16, 19, and 21, which initially had the same expected outcome in the CIT, have been merged and subsequently split into types 1200 and 1300. Although the O-tree delivers a more parsimonious type partition, the between-type inequality estimate slightly increases to 0.0732 Gini points.

Furthermore, the O-tree has two types containing a mix of women and men. Type 28, the largest type, includes 25% of individuals with an expected income equal to the population average. Type 27 comprises the most advantaged individuals, characterized by high levels of father's education, high father's occupation, and a mix of regions of origin. This type represents the top 5% top-earners among PhD graduates, earning approximately 30% more than the population average. It is also worth highlighting that the new split is unique on using mother's occupation, revealing a higher expected outcome among women whose fathers were not university graduates but the mother's occupation lied in high-skilled positions.

Despite the more parsimonious type partition, the O-tree constitutes an improvement with respect to the CIT in terms of reward and, most importantly, compensation score. Specifically, for the O-tree we have $S_C = 0.165$ and $S_R = 0.982$: both above our favourite thresholds $\underline{S}_C = 0.04$ and $\underline{S}_C = 0.95$. This is explained by a refined statistical procedure that parsimoniously employ degrees of freedom by merging non-salient types. The improvement is even more evident after inspecting Tables 3 and 4 below.[23]

[23] We have checked that the O-tree improves the Scores resulting from a non-parametric partition of the

Table 1: Reward score for each pair of terminal nodes of CIT

Type 1	Type 2	1-pvalue	Weighted	Type 1	Type 2	1-pvalue	Weighted
5	6	0,9945	0,0166	15	23	1,0000	0,0222
5	7	0,8706	0,0121	15	11	1,0000	0,0149
5	15	0,9734	0,0139	15	10	1,0000	0,0165
5	16	1,0000	0,0134	15	21	0,9902	0,0127
5	19	1,0000	0,0155	15	22	0,9546	0,0115
5	12	0,8355	0,0084	16	19	0,1078	0,0026
5	23	1,0000	0,0129	16	12	1,0000	0,0185
5	11	0,9358	0,0052	16	23	1,0000	0,0213
5	10	1,0000	0,0072	16	11	1,0000	0,0140
5	21	0,9997	0,0035	16	10	1,0000	0,0156
5	22	0,8638	0,0024	16	21	0,1096	0,0013
6	7	1,0000	0,0256	16	22	0,9947	0,0111
6	15	1,0000	0,0260	19	12	1,0000	0,0206
6	16	1,0000	0,0251	19	23	1,0000	0,0234
6	19	1,0000	0,0272	19	11	1,0000	0,0161
6	12	0,9170	0,0200	19	10	1,0000	0,0177
6	23	1,0000	0,0246	19	21	0,1493	0,0021
6	11	0,2933	0,0051	19	22	0,9949	0,0132
6	10	1,0000	0,0190	12	23	1,0000	0,0180
6	21	1,0000	0,0152	12	11	0,5815	0,0062
6	22	0,6197	0,0090	12	10	1,0000	0,0123
7	15	0,7280	0,0169	12	21	1,0000	0,0086
7	16	1,0000	0,0223	12	22	0,7567	0,0060
7	19	1,0000	0,0244	23	11	1,0000	0,0135
7	12	0,9999	0,0190	23	10	1,0000	0,0152
7	23	1,0000	0,0218	23	21	0,9757	0,0111
7	11	0,9998	0,0145	23	22	0,9992	0,0107
7	10	1,0000	0,0162	11	10	1,0000	0,0078
7	21	0,9975	0,0124	11	21	1,0000	0,0041
7	22	0,9362	0,0110	11	22	0,6578	0,0022
15	16	1,0000	0,0227	10	21	1,0000	0,0058
15	19	1,0000	0,0248	10	22	0,3949	0,0020
15	12	1,0000	0,0194	21	22	0,9923	0,0013

Description: 1-pvalue is the one resulting from a difference in mean t-test between types in the two considered types. The Weighted values multiply 1-pvalue by the weight as defined in Section 4.1.

Source: Istat, IIPDR 2018

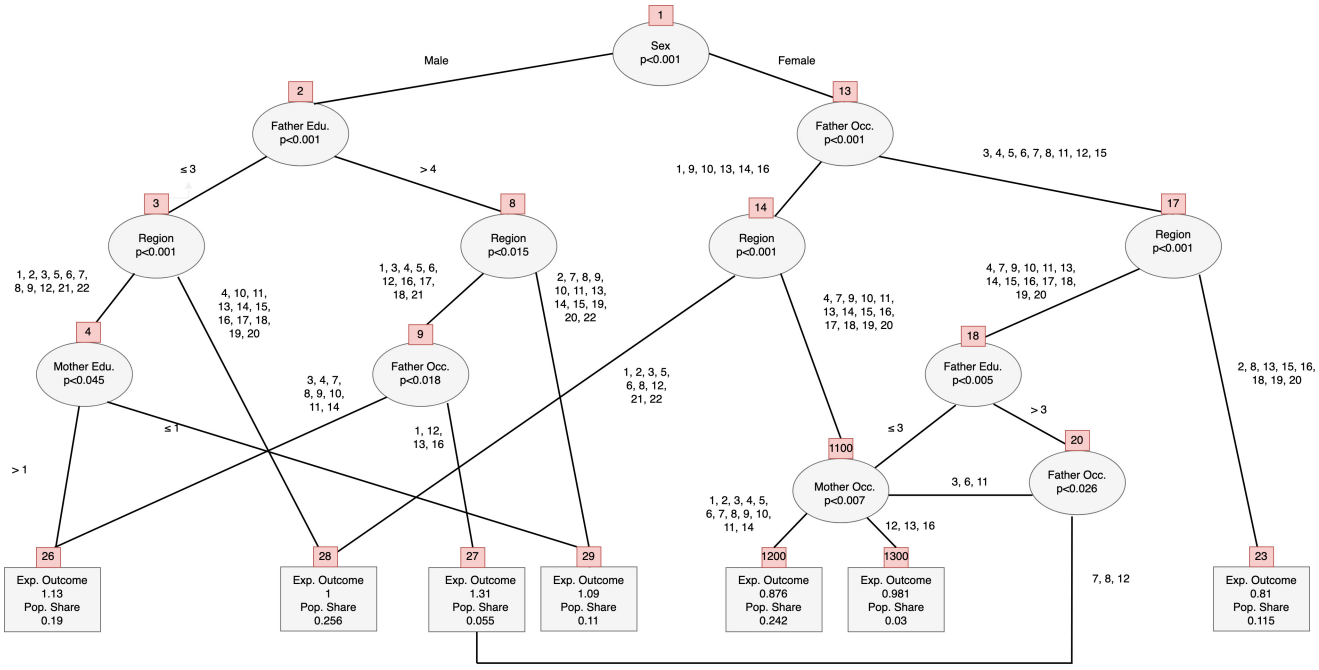
Table 2: Compensation score for each terminal node of CIT

Type	Subsplits	N	Mean	Power	P-value
5	1	320	2121,37	0,2600	0,4301
5	2	69	2427,86	0,2600	0,4301
6	1	2074	2316,1	0,3348	0,1588
6	2	149	2649,44	0,3348	0,1588
7	1	1527	2054,14	0,7059	0,0255
7	2	256	2279,19	0,7059	0,0255
15	1	153	1832,72	0,8257	0,0114
15	2	1689	2067,24	0,8257	0,0114
16	1	1495	1799,17	0,7490	0,0044
16	2	206	2076,98	0,7490	0,0044
19	1	1869	1808,73	0,6428	0,0514
19	2	162	2063	0,6428	0,0514
12	1	400	2167,64	0,8022	0,1933
12	2	785	2312,46	0,8022	0,1933
23	1	1434	1687,77	0,9728	0,0135
23	2	193	1524,56	0,9728	0,0135
11	1	205	2489,69	0,4300	0,0669
11	2	274	2185,82	0,4300	0,0669
10	1	697	2648,71	0,1013	0,1769
10	2	45	3333,32	0,1013	0,1769
21	1	12	1347,74	0,2699	0,0044
21	2	145	1884,55	0,2699	0,0044
22	1	34	2178,72	0,0380	0,2788
22	2	12	3634,93	0,0380	0,2788

Description: Power and P-value are those computed from a difference in means t-test between two types generated from the the initial terminal node.

Source: Istat, IIPDR 2018

Figure 2: Opportunity-tree types partition



O-tree is estimated setting $\alpha = 0.05$. The outcome is normalized such that 1 = mean sample income, 2072 euros.

Source: Istat, IIPDR 2018

Table 3: Reward score for each pair of terminal nodes of O-tree

Type 1	Type 2	1-pval	weighted
29	26	0,9870	0,0495
29	28	1,0000	0,0610
29	1200	1,0000	0,0592
29	1300	0,9999	0,0234
29	23	1,0000	0,0376
29	27	1,0000	0,0277
26	28	1,0000	0,0742
26	1200	1,0000	0,0724
26	1300	1,0000	0,0366
26	23	1,0000	0,0508
26	27	1,0000	0,0409
28	1200	1,0000	0,0832
28	1300	0,6310	0,0299
28	23	1,0000	0,0616
28	27	1,0000	0,0518
1200	1300	1,0000	0,0456
1200	23	1,0000	0,0598
1200	27	1,0000	0,0499
1300	23	1,0000	0,0240
1300	27	1,0000	0,0142
23	27	1,0000	0,0283

Description: 1-pvalue is the one resulting from a difference in mean t-test between types in the two considered types. The Weighted values multiply 1-pvalue by the weight as defined in Section 4.1.

Source: Istat, IIPDR 2018

Table 4: Compensation score for each terminal node of O-tree

Type	Subsplits	N	Mean	Power	P-value
29	1	789	2171,63	0,9263	0,0795
29	2	785	2312,46	0,9263	0,0795
26	1	2534	2315,19	0,4025	0,1224
26	2	168	2624,85	0,4025	0,1224
28	1	1783	2086,45	0,9999	0,47
28	2	1842	2047,76	0,9999	0,47
1200	1	385	1716,76	0,9966	0,0794
1200	2	3084	1819,9	0,9966	0,0794
1300	1	184	1942,09	0,53	0,4076
1300	2	236	2080,04	0,53	0,4076
23	1	1434	1687,77	0,9728	0,0135
23	2	193	1524,56	0,9728	0,0135
27	1	730	2627,63	0,1189	0,0542
27	2	58	3373,66	0,1189	0,0542

Description: Power and P-value are those computed from a difference in means t-test between two types generated from the the initial terminal node.

Source: Istat, IIPDR 2018

The results confirm that our proposal, the O-tree, improves the respect for the compensation and reward principles in the delivered type partition. Moreover, the comparison between Figures 1 and 2 highlights how the interaction between individual circumstances may be more complex than what results from the standard CIT. Interestingly, and maybe counterintuitively, the type partition that better respect the compensation principle is composed of fewer types.

Finally, aimed at providing more evidence about the suitability of the O-tree algorithm to improve the reward and the compensation scores, we have analyzed all CIT tree results for Europe produced at the Global Estimates of Opportunity and Mobility (GEOM) database.[24] For each CIT type partition, we have estimated ex-ante IOp as in GEOM, and

data. Using a simplified version of the father's and mother's education, the non-parametric delivers an IOp of 0.065 Gini points, a reward score of 0.876 and a compensation score of 0.52. The O-tree delivers an almost identical IOp of 0.064 Gini points, improving the reward score to 0.999 and the compensation score to 0.718. Results for other type partitions are available upon request.

[24]The GEOM is a comprehensive database aimed at estimating IOp for as many countries as possible.

the associated reward and compensation scores. Then, we have produced the O-tree, and estimated IOp Gini and the scores in the resulting type partition.

After averaging results, the average Gini IOp rises from 0.095 (GEOM) to 0.102 (O-tree). While the reward score sligtly decline (0.954 and 0.950 for the GEOM and O-tree, respectively), the compensation scores rises from 0.019 in GEOM to 0.075 for the O-tree. We find the O-tree to make a more appropriate use of the available data. The merging - splitting process algorithm delivers more types (on average, 9.5 in GEOM and 10.6 in the O-tree), which is also associated with higher IOp levels. This is reflected in a much higher compensation score, which is substantially improved without worsening the reward score. All these results and further analysis on this robustness check are available upon request.

6 Conclusion

Empirical approaches to ex-ante IOp are often limited by data availability. Most studies are based on survey data, including just a few circumstance variables from the many that a society could blame for affecting individual outcomes. Due to this partial observability of circumstances IOp estimates were traditionally considered as lower-bound estimates of the "true" values ((Checchi and Peragine, 2010; Ferreira and Gignoux, 2011). Recent contributions have challenged this reasoning, because scarcely populated types may bear errors when estimating the average outcomes, thus provoking upward-biased IOp estimates (Brunori et al., 2019). This paper expands on these ideas by first identifying a connection between Type I and Type II errors in hypothesis testing and the normative principles of reward and compensation that underlie the equality of opportunity principle. Moreover, we show how these errors are closely related to upward and downward biases in the es-

There are 80 estimates for Europe, steaming from EU-SILC data 2005, 2011 and 2019. The outcome is disposable household income, and the avialable circumstances are sex, parental education and occupation and place of origin. Details about the data and all results are available at https://geom.ecineq.org/.

timation of IOp. When analysts use sample-based survey data, they face a trade-off: to reliably assess differences between groups, they need a sufficient sample size.The richer the information about circumstances beyond individual control included in the analysis, the more Roemerian types can be identified. However, this richness exacerbates the sample size constraint: with the typical survey data available today, the number of possible types often exceeds the sample size. This makes it impossible to fully account for all interactions of circumstances or risks introducing a substantial upward bias. Consequently, analysts, constrained by sample size, must limit the number of types considered in the analysis. This can be done theoretically, by excluding some circumstances or merging categories, or empirically, by following data-driven methods to identify the most statistically salient groups. Whatever methodology is followed, once a partition in types is defined, to make the problem of IOp estimation statistically meaningful, it is key for researchers to evaluate whether they have a sufficient degree of freedom to reliably estimate differences in types' means. A sample size that is too small will result in an upward bias in the IOp (see Brunori et al. (2023) for a discussion).

Unfortunately, the typical empirical study of IOp does not include an assessment of the sufficiency of the data used. This is where this paper makes its most important contribution: we propose two scores to assess whether the type partition adopted is appropriate given the available data. The Compensation and Reward scores do this by signaling the magnitude of the risk of violating the two principles, which would translate into downward and upward biases, respectively.

Practitioners might find that, after computing the compensation and reward scores for a preferred type partition, they do not meet what they consider to be the appropriate normative thresholds. One potential solution could be to expand the sample size, for instance, by imputing missing observations or merging information obtained by different sources.

Moreover, the researcher may evaluate whether the initial partition can be improved by merging redundant types in order to free degree of freedom. If the type definition is found to be too narrow, the researcher can group categories within circumstances (e.g., broadening the definition of parental education) or even eliminate certain circumstances from the analysis. The O-tree introduced in Section 4.2 is an example of how a data-driven partition of Roemerian types can be adjusted to obtain improved scores.

However, if the researcher prefers to measure IOp using that specific type partition, we recommend always addressing and reporting the associated reward and compensation scores, highlighting the limitations and biases that might affect the estimates. While we are all constrained by data availability, one should never demand more from the data than it can deliver.

To demonstrate the usefulness of our proposals, we analyzed the distribution of income in a sample of Italian PhD graduates. Using conditional inference trees, we show a substantial degree of inequality among Italian doctors, with a surprising share of variability predicted by circumstances beyond individual control, such as gender and region of birth. Moreover, we evaluated the compensation and reward scores for the partition obtained using standard conditional inference trees and then for the partition obtained with the opportunity tree, showing improvement in both scores. Our evaluation demonstrates that the sample used is sufficiently rich to estimate inequality of opportunity with a limited risk of biases.

References

Brunori, P., Hufe, P., and Mahler, D. (2023). The roots of inequality: Estimating inequality of opportunity from regression trees and forests. *Scandinavian Journal of Economics*.

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.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Cappelen, A. W., Konow, J., Sørensen, E. Ø., and Tungodden, B. (2013). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, 103(4):1398–1413.

Checchi, D. and Peragine, V. (2010). Inequality of opportunity in italy. *The Journal of Economic Inequality*, 8(4):429–450.

Dworkin, R. (1981). What is equality? part 2: Equality of resources. *Philosophy & public affairs*, pages 283–345.

Escanciano, J. C. and Terschuur, J. R. (2023). Machine learning inference on inequality of opportunity.

Ferreira, F. H. 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.

Ferreira, F. H. and Peragine, V. (2016). Individual responsibility and equality of opportunity. In *The Oxford Handbook of Well-Being and Public Policy*.

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. OUP Oxford.

Fleurbaey, M. and Peragine, V. (2013). Ex ante versus ex post equality of opportunity. *Economica*, 80(317):118–130.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research*, 16(1):3905–3909.

ISTAT (2018). Survey on doctorate holders vocational integration: microdata for research purposes. *ISTAT*, 1.

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.

Rawls, J. (1973). A theory of justice. *Harvard University Press*.

Roemer, J. E. (1998). Equality of opportunity. In *Equality of Opportunity*. Harvard University Press.

Roemer, J. E. and Trannoy, A. (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, 54(4):1288–1332.

Sen, A. (1980). Equality of what? *The Tanner lecture on human values*, 1:197–220.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

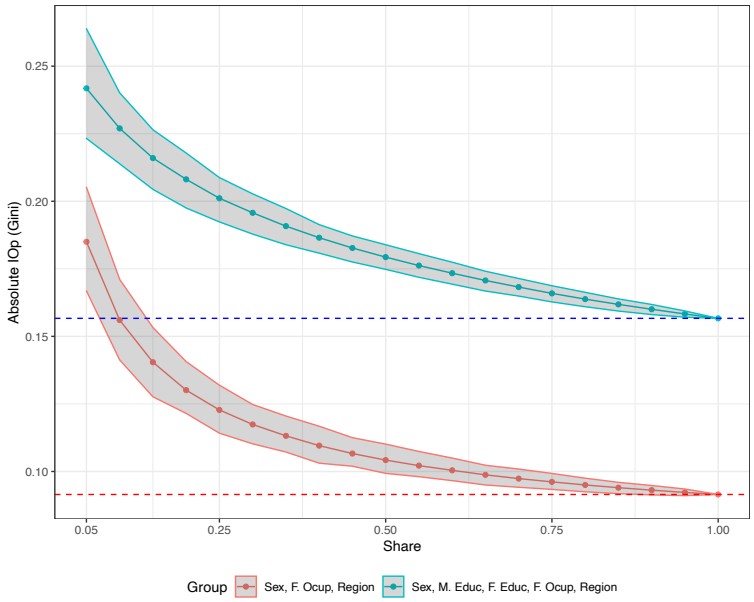
Appendix A Upward and downward biases

To clarify the dynamics between up and downward biases in IOp estimation, let us consider Figure A1 which reports the estimated IOp for different type partitions and sample sizes based on the data described in Section 5.

The dashed horizontal lines in Figure A1 are the IOp levels computed on the whole sample which, for the scope of this exercise, is assumed to correspond to the population. Suppose that our $I\left(Y^N, T^N\right)$ is the IOp computed on the whole population when sex, father occupation, father education, mother education and region of birth are taken as circumstances (the dashed blue line). The solid lines report the IOp estimates for the same partitions on random subsamples of the population (subsamples are obtained without replacement from the 5%, around 800 observation, to 100%). Analogously, the red dashed line exemplifies the downward bias, representing IOp estimated on the same sample but excluding the parents' education from the set of circumstances. In this case, for sample sizes above 10%, the red solid line is always below the dashed blue one. The upward bias is, instead, quite evident by observing how small sample sizes lead to IOp estimates higher than the respective dashed blue line. Interestingly, this upward bias may reach a point where, even with fewer circumstances and types, a small sample we lead to IOp estimates above the true value $I\left(Y^N, T^N\right)$. In Figure A1 this happens in correspondence of the 0.05 sample size (red solid line).

The upward and downward biases illustrated in the previous example are well known in the literature (Ferreira and Gignoux, 2011; Brunori et al., 2019). For a long time, the downward bias has led many researchers (Checchi and Peragine, 2010; Ferreira and Gignoux, 2011; Roemer and Trannoy, 2016; Ramos and Van de Gaer, 2016) to interpret IOp estimates as mere lower bounds for the real IOp. Upward biases have been recently shown by Brunori et al. (2019), and the literature (see Brunori et al., 2023) has looked at

Figure A1: Upward and downward biases.



Source: Own elaboration from Istat IIPDR 2018. Confidence intervales estimated after 100 bootstrapped repetitions.

data-driven approaches for a solution.

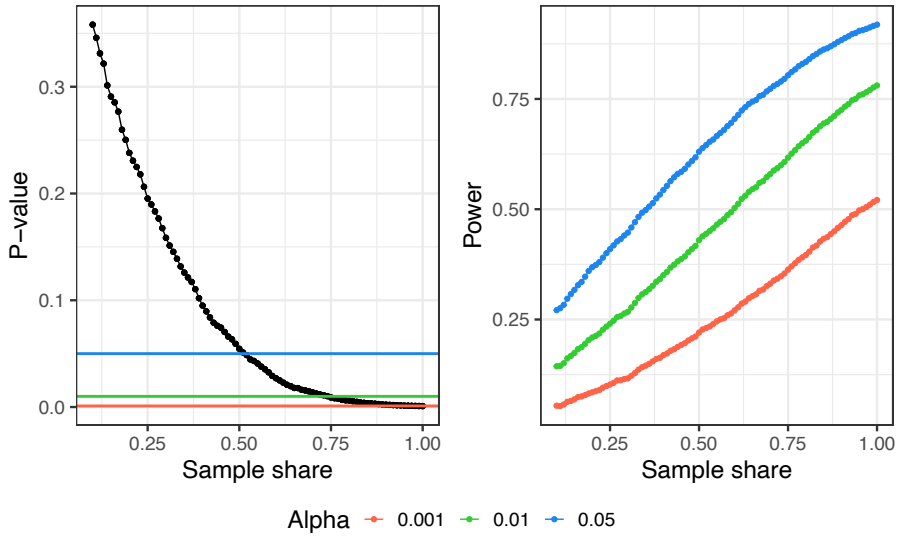
Appendix B Appendix: Sample size simulation

Here we clarify that much of our ability to satisfy the compensation and reward principles in hypothesis testing depends on the sample size. Therefore, researchers should be particularly careful when trying to measure IOp with small samples.

Consider two groups of 250 individuals each. An outcome of interest, such as income, is distributed as a $N(160, 25)$ in the first group, while income in the second is distributed as a $N(170, 40)$. In the left-hand plot in Figure A2 we show the p-values of a difference-in-means t-test for different sample shares, and the α thresholds that would make us reject the null hypothesis of equal means. In the right-hand plot in Figure A2, we show the power

of each test, for the same sample sizes and α values. We bootstrap each test 500 times. By construction, the two groups have different expected income. Nevertheless, with small sample sizes (less than 75 individuals per group) for none of the standard values of α we will be rejecting the null hypothesis of equal means. This is reflected, in the right-hand panel, by the low power in correspondence of small sample shares.

Figure A2: The importance of sample size



Source: Own elaboration.

Appendix C Appendix: Missing partitions in conditional inference trees

Here we exemplify why Conditional Inference Trees (CITs) cannot explore all the possible type partitions.

Let us assume to have a sample (S) for which we observe only two circumstances: gender (men or women) and skin color (brown or white). Let us call $T_0 = \{t_{MW}, t_{MB}, t_{FW}, t_{FB}\}$

the *full* type partition, that is one in which we divide the population by sex and skin color. In such a setting, the standard approach in the literature (Checchi and Peragine, 2010; Ferreira and Gignoux, 2011) assumes T_0 to be the salient type partition so that the ex-ante IOp estimation is $I\left(Y^S, T_0\right)$. The literature, however, often neglects that the same population can be partitioned in several other ways: fifteen in this case. To these partitions, listed below, one should add $T_\emptyset = S$ which corresponds to the case in which no type is generated, so that IOp is zero by definition.

- $T_0 = \{t_{MW}, t_{MB}, t_{FW}, t_{FB}\}$
- $T_1 = \{t_M, t_F\}$
- $T_2 = \{t_{MW}, t_{MB}, t_F\}$
- $T_3 = \{t_M, t_{FW}, t_{FB}\}$
- $T_4 = \{t_W, t_B\}$
- $T_5 = \{t_{WM}, t_{WF}, t_B\}$
- $T_6 = \{t_W, t_{BM}, t_{BF}\}$
- $T_7 = \{t_{MW}, t_{-MW}\}$
- $T_8 = \{t_{MB}, t_{-MB}\}$
- $T_9 = \{t_{FW}, t_{-FW}\}$
- $T_{10} = \{t_{FB}, t_{-FB}\}$
- $T_{11} = \{t_{MW}, t_{FB}, \{t_{MB} \cup t_{FW}\}\}$
- $T_{12} = \{t_{MB}, t_{FW}, \{t_{MW} \cup t_{FB}\}\}$
- $T_{13} = \{t_{FW}, t_{MB}, \{t_{FB} \cup t_{MW}\}\}$
- $T_{14} = \{t_{FB}, t_{MW}, \{t_{FW} \cup t_{MB}\}\}$
- $T_{15} = \{\{t_{MW} \cup t_{FB}\}, \{t_{FW} \cup t_{MB}\}\}$

Clearly, $I\left(Y^S, T_0\right)$ is only one of the possible measures of IOp: the one based on the assumption that T_0 is the correct way of partitioning the population.

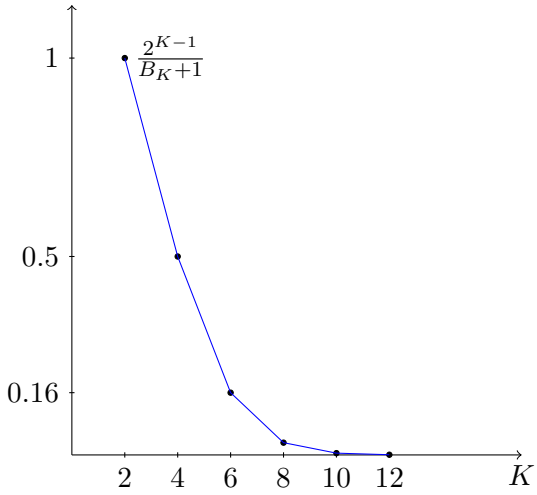
The CIT algorithm starts with the identification of the circumstance (sex or skin color, in this case) most associated with income. Suppose it was sex at birth, so the algorithm tests whether there exist a difference in expected income between men and women. If we reject the null hypothesis of both groups having same incomes, the type partition T_1 is generated. We can now treat each type in T_1 as a different population and test, within each sex group, whether skin color is correlated with income. If skin color is associated with men's income, but not in women's income, we get T_2. Vice versa, the algorithm may also deliver T_3.

A hardly mentioned, yet theoretically relevant, limitation of CITs is that, due to its deterministic structure, the salient partitions one can hope to identify are $T_\emptyset$, T_0 and T_1 up to T_6, which constitute half of the possible partitions of T_0. This limitation becomes even more stringent when more circumstances (or categories) are considered.

Formally, let $c_1, ..., c_m$ be the list of observed circumstances and $|c_j| \geq 2$, $j = 1, ..., m$, the number of values c_j can take. Without loss of generality, assume that all circumstances are nominal variables, not ordered.[25] Then, for each c_j, we have $2^{|c_j|-1}$ possible binary splits, including the cases of no split and total split. Let $K = \prod_{j=1}^{m} |c_j|$, then the number of possible salient type partitions CITs can explore is 2^{K-1}. On the other hand, the number of all the possible partitions of a K-dimensional set into disjoint subsets whose union forms the initial set is measured by the K-th *Bell number*. The series of Bell numbers is defined by $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_n$, with $B_0 = 1$. As shown in Figure A3, the share of type partitions that CITs can possibly identify quickly converges towards zero.

[25] Since the number of possible binary splits of a nominal variable is higher than the ones for an ordinal one, the result below holds true even if some or all circumstances are ordinal.

Figure A3: Share of observable type partition using recursive binary split



Source: Own elaboration.

Appendix D Appendix: Descriptive Statistics

Table A1: Income summary statistics

N	Mean	Sd	Gini
14,205	2,070.42	1,098.70	0.26

Source: Istat, IIPDR 2018

Table A2: Mother occupation

Label	Original Variable	Our variable	2014 (%)	2012 (%)
Entrepreneur	Aut_m_1	Mother Occ.=9	1.4	1.3
Professionals	Aut_m_2	Mother Occ.=13	2.6	2.35
Technicians	Aut_m_3	Mother Occ.=12	3.77	3.94
Managing Directors (Family Business)	Aut_m_4	Mother Occ.=14	0.32	0.45
Manual Worker (Family Business)	Aut_m_5	Mother Occ.=5	0.32	0.38
Managing Director	Aut_m_6	Mother Occ.=15	0.03	0.04
Administration Managers	Aut_m_7	Mother Occ.=6	0.06	0.04
Manager	Dip_m_1	Mother Occ.=16	5.01	4.25
Professional (Supervisor, official)	Dip_m_2	Mother Occ.=10	26.58	26.73
Technicians and Associate Professionals	Dip_m_3	Mother Occ.=11	14.54	12.92
Clerical Support Workers	Dip_m_4	Mother Occ.=8	7.02	7.48
Skilled manual worker	Dip_m_5	Mother Occ.=7	1.69	1.6
Elementary Occupations	Dip_m_6	Mother Occ.=4	4.2	3.87
Other	Cond_m_4	Mother Occ.=3	1.52	1.33
Dead	Cond_m_6	Mother Occ.=1	1.13	1.28
Housewife	Cond_m_3	Mother Occ.=2	29.83	32.05

Source: Istat, IIPDR 2018

Table A3: Father occupation

Label	Original Variable	Our variable	2014 (%)	2012 (%)
Entrepreneur	Aut_p=1	Father Occ.=9	4.78	4.59
Professionals	Aut_p=2	Father Occ.=13	9.55	9.19
Technicians	Aut_p=3	Father Occ.=12	10.45	10.79
Managing Directors (Family Business)	Aut_p=4	Father Occ.=14	0.2	0.2
Manual Worker (Family Business)	Aut_p=5	Father Occ.=5	0.14	0.16
Managing Director	Aut_p=6	Father Occ.=15	0.05	0.11
Administration Managers	Aut_p=7	Father Occ.=6	0.08	0.13
Manager	Dip_p=1	Father Occ.=16	14.5	14.1
Professional (Supervisor, official)	Dip_p=2	Father Occ.=10	15.97	16.56
Technicians and Associate Professionals	Dip_p=3	Father Occ.=11	18.26	17.66
Clerical Support Workers	Dip_p=4	Father Occ.=8	6.51	6.83
Skilled manual worker	Dip_p=5	Father Occ.=7	9.54	9.84
Elementary Occupations	Dip_p=6	Father Occ.=4	4.34	3.93
Other	Cond_p=4	Father Occ.=3	1.78	20.03
Dead	Cond_p=6	Father Occ.=1	3.84	3.89

Source: Istat, IIPDR 2018

Table A4: Mother education

Label	Original Variable	Our variable	2014 (%)	2012 (%)
Elementary or no education	Titstu_m=1	Mother Edu.=1	10.52	12.52
Lower secondary	Titstu_m=2	Mother Edu.=2	19.64	20.09
Upper secondary	Titstu_m=3	Mother Edu.=3	38.18	37.76
University diploma	Titstu_m=4	Mother Edu.=4	5.16	4.59
University degree or above	Titstu_m=5	Mother Edu.=5	26.5	25.04

Source: Istat, IIPDR 2018

Table A5: Father education

Label	Original Variable	Our variable	2014 (%)	2012 (%)
Elementary or no education	Titstu_p=1	Father Edu.=1	8.74	10.13
Lower secondary	Titstu_p=2	Father Edu.=2	20.24	20.26
Upper secondary	Titstu_p=3	Father Edu.=3	36.37	36.12
University diploma	Titstu_p=4	Father Edu.=4	3.18	2.5
University degree or above	Titstu_p=5	Father Edu.=5	31.46	30.98

Source: Istat, IIPDR 2018

Table A6: Region of origin and sex

Label	Original Variable	Our variable	2014 (%)	2012 (%)
Piemonte	Unireg=1	Region=1	5.52	6.13
Valle d'Aosta	Unireg=2	Region=2	0.15	0.13
Lombardia	Unireg=3	Region=3	12.81	11.34
Trentino - Alto Adige	Unireg=4	Region=4	1.81	1.14
Veneto	Unireg=5	Region=5	6.31	7.01
Friuli-Venezia Giulia	Unireg=6	Region=6	2.11	2.48
Liguria	Unireg=7	Region=7	2.25	2.21
Emilia-Romagna	Unireg=8	Region=8	5.46	5.56
Toscana	Unireg=9	Region=9	5.9	6.73
Umbria	Unireg=10	Region=10	1.4	2.18
Marche	Unireg=11	Region=11	2.77	2.83
Lazio	Unireg=12	Region=12	12.79	11.75
Abruzzo	Unireg=13	Region=13	2.54	2.69
Molise	Unireg=14	Region=14	0.56	0.6
Campania	Unireg=15	Region=15	10.02	9.4
Puglia	Unireg=16	Region=16	6.71	7.92
Basilicata	Unireg=17	Region=17	1.1	0.95
Calabria	Unireg=18	Region=18	3.89	3.98
Sicilia	Unireg=19	Region=19	8.15	9.08
Sardegna	Unireg=20	Region=20	3.32	2.37
Europe-EU	Unipaese=101	Region=21	1.6	1.19
Rest of the World	Unipaese=301	Region=22	3.48	2.33
Sex	Sesso_2	Female	52.83	52.9

Source: Istat, IIPDR 2018