# On Inequality of Opportunity: The Case of India

Tista Kundu

▶ **To cite this version:**

Tista Kundu. On Inequality of Opportunity: The Case of India. Humanities and Social Sciences. Université de Cergy-Pontoise. Ecole doctorale no 405 : Economie, Management, Mathématiques et Physique (EM2P). Théorie _Economique, Modélisation et Applications (THEMA), 2019. English. NNT : . tel-02519121

**HAL Id: tel-02519121**

**https://essec.hal.science/tel-02519121v1**

Submitted on 25 Mar 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# On Inequality of Opportunity: The Case of India

A dissertation submitted in partial fulfillment of the requirements for the degree of

**PHD IN BUSINESS ADMINISTRATION**

**FROM ESSEC BUSINESS SCHOOL**

Was presented and defended publicly on the 2nd of July 2019 by

**Tista KUNDU**

JURY

| | | |
|---|---|---|
| Arnaud LEFRANC | Supervisor | Professor, University of Cergy-Pontoise (France) |
| Cristina TERRA | Supervisor | Professor, ESSEC Business School, Paris (France) |
| Nicolas GRAVEL | Referee | Director, Centre de Sciences Humaines (India) |
| Marta MENENDEZ | Referee | Associate Professor, University of Paris Dauphine (France) |
| Francesco ANDREOLI | Examiner | Researcher, LISER (Luxembourg) |
| Maëlys de la RUPELLE | Examiner | Assistant Professor, University of Cergy-Pontoise (France) |

# Thèse de doctorat

pour l'obtention du titre de DOCTEUR EN SCIENCES ÉCONOMIQUES délivré par

**l'Université de Cergy-Pontoise**

École doctorale no 405 : Économie, Management, Mathématiques et Physique (EM2P)

Théorie Économique, Modélisation et Applications (THEMA)

CNRS UMR 8184

# On Inequality of Opportunity: The Case of India

présentée et soutenue publiquement le 2 Juillet 2019 par


## Tista KUNDU


Directeur de thèse:

M. Arnaud LEFRANC et Mme. Cristina TERRA

**JURY**

| | | |
|---|---|---|
| Arnaud LEFRANC | Directeur | Professeur, Universitè de Cergy-Pontoise |
| Cristina TERRA | Directrice | Professeur, ESSEC Business School, Paris |
| Nicolas GRAVEL | Rapporteur | Directeur, Centre de Sciences Humaines |
| Marta MENENDEZ | rapporteuse | Maître de confèrences HDR, Universitè Paris-Dauphine |
| Francesco ANDREOLI | Examinateur | Chercheur, LISER |
| Maëlys de la RUPELLE | Examinatrice | maître de confèrences, Universitè de Cergy-Pontoise |

**Abstract**

The thesis consists of three empirical chapters on the extent of inequality of opportunity in India using data from the National Sample Survey. Inequality of opportunity is conceptualized as the unequal distribution in an outcome that is generated by factors beyond any individual responsibility. Chapter 1 estimates the index measures of inequality of opportunity in India during the time frame of *2004-12*, for consumption expenditure, wage earning and education. Taking caste, sex, region of residence, parental education and occupation as our fatalistic circumstances, both the non-parametric and parametric index shows that more than one-fourth of the respective inequality in wage and education is due to unequal opportunities among the circumstances. Since the traditional approach of non-parametric and parametric estimation strategy can not point out the meaningful intertwining among the circumstances, we further adopt the recently introduced approach of the regression tree analysis that is able to draw an opportunity tree by hierarchical order of circumstances. The opportunity tree for contemporary India identifies parental education as one of the main circumstance variable behind the unequal opportunity in the country, especially for education. Chapter 2 focuses on a temporal perspective that explores equalization of opportunities among the different caste categories in India over the time span of 1983-2012, adopting the robust methodology of distributional dominance. The relatively upper caste groups are not only found to clearly dominate the historically disadvantageous caste categories, but the premium enjoyed by the former actually increases over time, as far as earning opportunity is concerned. Opportunity in consumption expenditure among the caste groups however have equalized to an impressive extent, especially during *2004-12*. Chapter 3 estimates the degree of inequality of opportunity in access to elementary and post-elementary education among Indian children, using the dissimilarity index and the Human Opportunity Index. Indian children are found to enroll in elementary schools on time, irrespective of their circumstances. However children of lesser educated parents are always less likely to complete the eight years of elementary schooling on time, even in 2012, when elementary education is made free and compulsory by the Government, for all eligible children. The regional analysis reveals that access to basic educational opportunity is relatively worse for East and Central India.

**Résumé**

La thèse se compose de trois chapitres empiriques sur l'ampleur de l'inégalité des chances en Inde à l'aide des données de la National Sample Survey. L'inégalité des chances est conceptualisée comme une distribution inégale dans un résultat qui est généré par des facteurs qui dépassent toute responsabilité individuelle. Le chapitre 1 donne une estimation des indexées de l'inégalité des chances en Inde au cours de la période du 2004-12, pour les dépenses de consommation, les salaires et l'éducation. Si nous considérons la caste, le sexe, la région de résidence, l'éducation parentale et le métier comme circonstances fatalistes, l'indice non paramétrique et l'indice paramétrique montrent que plus du quart de l'inégalité respective en matière de salaire et d'éducation est due à l'inégalité des chances. L'approche traditionnelle de la stratégie d'estimation paramétrique et paramétrique ne permettant pas de mettre en évidence une imbrication significative des circonstances, nous adoptons en outre l'approche la plus récente de l'analyse par arbre de régression, qui permet de tracer un arbre d'opportunité par ordre hiérarchique de circonstances. L'arbre des opportunités de l'Inde contemporaine identifie l'éducation parentale comme l'une des principales variables de l'inégalité des chances, en particulier dans le domaine de l'éducation. Le chapitre 2 se concentre sur une perspective temporelle qui explore l'égalisation des chances entre les différentes catégories de castes en Inde sur la période 1983-2012, en adoptant la méthodologie robuste de la dominance distributionnelle. Les groupes de castes relativement supérieurs non seulement dominent clairement les catégories de castes historiquement désavantageuses, mais la prime dont bénéficient les premiers augmente en fait avec le temps, en ce qui concerne les possibilités de revenus. Les opportunités de consommation des groupes de castes se sont toutefois égalisées de manière impressionnante, en particulier au cours de la 2004-12. Le chapitre 3 évalue le degré d'inégalité des chances d'accès aux études élémentaires et post-élémentaires chez les enfants indiens, à l'aide de l'indice de dissimilarité et de l'indice de chances humaines. On constate que les enfants indiens s'inscrivent à l'école primaire à temps, quelle que soit leur situation. Cependant, les enfants de parents moins scolarisés ont toujours moins de chances de terminer les huit années d'enseignement primaire à temps, même en 2012, année où l'enseignement primaire est rendu gratuit et obligatoire pour tous les enfants éligibles par le gouvernement. L'analyse régionale révèle que l'accès aux possibilités d'éducation de base est relativement moins bon pour l'Inde orientale et centrale.

# Acknowledgments

First and foremost I express my heartfelt gratitude to my supervisors Professor Arnaud Lefranc and Professor Cristina Terra. Dr Lefranc introduced me to the area of inequality of opportunity and guided me throughout since my first year of PhD. I would like to express my greatest appreciation to each of our invaluable discussion that helped me to think more like a researcher, to develop the chapters gradually and to make this PhD a reality. His well articulated constructive feedback on multiple drafts of my works, immensely helped me to improve and organize the thesis. I am extremely grateful to have Dr Terra as my co-supervisor who always motivated me with her positive energy, enthusiastic encouragement and inspirational comments. Her kind words blended with her upright professionalism, helped me to overcome many ups and downs of my graduate life and to rather orient myself more like a future academician. It would be impossible for me to finish the PhD without her support.

I would also like to offer my sincere thanks to Professor Marta Menéndez, Professor Nicolas Gravel, Professor Maëlys de la Rupelle and Dr Francesco Andreoli, for accepting to be part of my thesis committee. I am particularly grateful to Dr Menéndez and Dr Gravel for their kind acceptance to be referees of my thesis and to participate in my defense. I extend my thanks to Dr Rupelle for her helpful feedback on my presentation in THEMA and for agreeing to be the examiner of my thesis. A very special thanks to Dr Andreoli for his indispensable technical help in my 2nd chapter and for accepting to be the examiner of my thesis.

I would very much like to acknowledge the financial support provided by ESSEC Business School (Paris), both in the form of a doctoral fellowship for four years and a teaching contract for the quintessential fifth year. I am deeply grateful to ESSEC for accepting me in the PhD program and providing me the opportunity to learn from the numerous academic professionals in both ESSEC and THEMA (Université de Cergy-Pontoise,

UCP). I express my profound gratitude to Professor Sukanta Bhattacharya (University of Calcutta) and Professor Indraneel Dasgupta (Indian Statistical Institute, Kolkata) for strongly encouraging me and supporting my application to ESSEC.

Throughout my stay in the doctoral program, I am incessantly enriched by the academic excellence of the department of Economics in ESSEC and the laboratoire THEMA in UCP. I take this opportunity to express my sincere thanks to all the faculty members of ESSEC Economics Department and THEMA. I learned a lot from them that eventually assisted me to materialize my PhD dissertation. A special word of thanks to Professor Patricia Langohr and Professor Gorkem Celik for giving me the opportunity to teach under their co-ordination, that made my teaching experience a learning and pleasant memory at ESSEC.

Also thanks to all my colleagues in ESSEC and THEMA. They helped me to learn many necessary skills in life and it has been a real pleasure to have them as my colleagues. Among fellow researchers from other institutions, I am particularly very grateful to Dr Daniel Mahler and Dr Geoffrey Teyssier for their invaluable help in my 1st chapter.

A very special thanks to Ms Lina Prevost, who not only helped me to deal with the various administrative affairs since my first day in the program, but whose friendly assistance often goes beyond the so-called administrative boundary. I express my cordial thanks to you for your generous help and support. My special thanks are extended to Ms Christine Gil and Ms Katy Sylvestre on this account as well. I also like to thank the members of ALEGESSEC for making my stay cozy and comfortable, that enables me to fully concentrate on my work without worrying about many of the household related issues.

Many thanks to my friends Ms Laura Dalmazzini and Dr Akanksha Gaur for providing me the much needed peaceful refuge during the period of my rigorous coursework. I am very lucky to have friends like you in the very beginning of my PhD and I will always cherish the memory of our time spent together. I very much acknowledge the help of my good old colleagues and friends, Dr Debanjana Dey and Dr Meghna Dutta. Without your help I could not even start this PhD project. A most special thanks to Dr Saikat

Mazumdar for bringing me my brand new laptop all the way to Paris. Without this small machine I have no idea how would I finish my PhD at all and I thank you Saikat every time I run a code. Loads of thanks to my fellow battalions Dr Koninika Pal and Dr Arijit Hazra, without whom life in Europe would be non-trivially boring. I will lifelong embrace all of our eventful ventures together. I wish each one of you the very best in your life.

No word of thanks is sufficient for my entire family, specially my parents and husband. I was first intrigued to research and decided to pursue my doctoral studies someday, upon watching my uncle Dr Rajen Kundu whose academic pursuit I was grown up with. Relentless effort of my mother, Chaitali Kundu, helped me to stay determined in pursuit of my dream and actually made me to finish this entire PhD project successfully. Guru Prasad Kundu, my father, who himself had been a Government executive in his work-life, helped me in developing many perceptible insights on India that are useful to my work at large. Their short visit to Cergy actually made me to realize the beauty of this small serene French town and made my graduate life much more bearable thereafter. This arduous journey very well belongs to my husband, Dr Swarnendu Sil. He is the closest sufferer of all my stress and anxiety, as well as the nearest witness to all my little joy and pleasures of PhD. His scholastic pursuits have always been a great source of motivation for me and his astute appraisal of various aspects of my PhD assisted me to stay focused on my research. Without you, my PhD as well as my life in Europe would have never been such a meaningful quest.

This entire journey was started on the day of my graduate record examination when my uncle, Birendranath Chatterjee (my beloved *thiss*), drove me to the exam center. You were one of the most enthusiastic one about this venture and it feels like yesterday when I was waving at you from the airport lounge. Not even in my worst nightmare, I have ever imagined that you would not be able to see me finish what we kind of began together. Your untimely loss is the biggest loss in my stays abroad. I wish I could be there then! I wish you could be here now!

# Contents

# List of Tables

# List of Figures

# Introduction

The concept of inequality of opportunity is getting popular in the scholarly debates of distributional justice because of its accentuated attention on the rather despondent aspect of social inequality, that results from the ascribed social class of an individual, like race, sex or family background. The precept of equal opportunity talks about an egalitarian society that ensures equal distribution of any social advantage for people across divergent social and family backgrounds.

This PhD thesis draws on the ongoing literature of inequality of opportunity and contributes in this domain by providing internationally comparable estimates of inequality of opportunity in India, using a wide range of empirical approaches. Chapter 1 estimates the different measures of inequality of opportunity in consumption, wage and education, due to differential circumstances like caste, sex, region or parental backgrounds, as well as provides the opportunity tree structure of contemporary India, exhibiting the hierarchical order among the circumstances considered. Chapter 2 gives a temporal perspective by exploring whether the consumption and earning opportunity across the caste groups in India have equalized in the long-term (1983-2012). Chapter 3 evaluates to what extent the access to basic educational opportunity for Indian children is impaired by their ungoverned dissimilar backgrounds in the twenty-first century.

Inequalities can arise from a widely different factors, but not all of them are equally important. Indeed, the classical welfarist tradition was strongly criticized because of the ethical desirability of their motto of assessing social welfare by means of equal distribution for every individuals in the society. Since the motive of equalizing distribution across everyone, fails to hold the individuals responsible for differences in their personal choice or preferences, it is not always morally supportable in the analysis of inequality. A rather conservative egalitarian view of responsibility sensitive analysis of inequality had gradually emerged from the philosophical debate in the late twentieth century, that ques-

tion the moral basis of evaluating social welfare by equalizing outcome across everyone and is rather concerted to focus on a more comparable platform that one can speak of equalizing, by taking into account the differences in individual responsibilities.

The philosophical debate as initiated by Rawls (1971), put forward the important question of 'equality of what' is to be assured in an egalitarian society. The debate is further enriched by the philosophical contributions of Sen (1980), Dworkin (1981*b*,*a*), Arneson (1989), Cohen (1989), that gradually shifts the focus of distributional analysis where *opportunities* rather than economic outcomes become the relevant 'currency of justice'. Given the fair opportunities as provided by a 'level playing field' for everyone, the society is no longer obliged to equalize any further outcome inequality that is generated from differential individual choices or preferences. The metaphor of 'leveling the playing field' is later formalized by Roemer (1993) with the proposition of a concept of *inequality of opportunity*, that prioritizes the analysis of inequality arising only from factors that are *not* subject to individual responsibility.

Since the conceptualization of *inequality of opportunity*, the analysis of inequality got a dichotomous form on the basis of the heterogeneous characteristic attributes of the inequality generating factors. On the one hand there are factors that are beyond the control of individual responsibility and are termed as *circumstances*. The *effort* factors on the other hand are defined as the inequality generating factors that are subject to individual control and are therefore considered as the legitimate sources of inequality. In this dichotomous set-up, inequality of opportunity is the inequality that is generated by the *circumstance* factors only and is therefore strictly unfair from an ethical perspective. Whereas inequality generated by the controllable *effort* factors is no longer morally objectionable from the viewpoint of responsibility sensitive egalitarianism.

The doctrine of inequality of opportunity therefore redefines the notion of an equitable society by emphasizing more on the rather unfair and ethically objectionable part of inequality, for which no one can be held responsible. By definition, individuals can not choose or alter their *circumstances* by themselves. Hence the inequality generated by it should be considered with priority so as to ensure a fair and equal platform for everyone in the society, where an economic advantage is no longer governed by their respective

fortunes or the lack of it. So the portrait of an equitable society from this standpoint is one, where there is neither any premium for individuals with advantageous circumstances nor any penalty for those with unfortunate ones.

The concept of inequality of opportunity thereby changes the focus of social welfare evaluation, from equalizing outcome for everyone, to equalizing opportunities for all individuals irrespective of their circumstances. From this perspective, any economic advantage is therefore usually thought to be generated by two broad classes of factors, *circumstances* and *efforts*, while any inequality due to the former is deemed unfair. In this set-up of circumstance-vs-effort, inequality of opportunity is precisely that part of the outcome inequality that is exclusively generated by circumstances. Therefore the main methodological challenge to estimate the extent of unequal opportunity is to isolate the unfair part of inequality, that is the result of differing circumstances only. Since the proposition of this concept, a large body of literature in this area provides many different measurement strategies to estimate the extent of inequality of opportunity for a number of developed and developing countries. Ramos & Van de Gaer (2012), Roemer & Trannoy (2013) provide some extensive surveys on the existing literature in this field.

Any analysis on this domain thereby starts by identifying the circumstance and/or the effort variables, that divides the population under study in suitable number of groups. When the grouping is done on the basis of circumstance factors, each group is formally called a *type* and individuals belonging to a *type* share the same circumstances but can differ in terms of their effort level. Whereas each group formed by the effort variables are called *tranches* and individuals within a *tranche* differs by their circumstances but are assumed to exert the same level of effort. Provided the type-tranche classification, there are two basic approaches to isolate the unfair inequality, which we describe by the following simple structure.

Consider the simple case where the population can be divided in $n_e$ number of efforts and $n_c$ number of circumstances. Assuming each combination of circumstance and effort variables occur at most once, then borrowing from Ferreira & Peragine (2015) we can

write the following output matrix as -

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1n_e} \\ \vdots & \ddots & \vdots \\ y_{n_c1} & \cdots & y_{n_cn_e} \end{bmatrix} \tag{1}$$

Each row of the above matrix (1) correspond to a *type* and therefore the outcome distribution within each row differs only by the effort levels. On the other hand, the distribution of individual outcome within a column differs by circumstances for exerting the same level of effort. Therefore there are two ways to isolate the unfair inequality, either by inequality *between types* (between the rows) or by inequality *within tranches* (within the column). Notice that in either of the cases the outcome distribution differs only because of the varying circumstances and thereby estimates the extent of inequality of opportunity in the society.

In the literature, the measurement approach of *within-tranche* inequality is called the *ex-post* approach and that of the *between-type* inequalities is known as the *ex-ante* approach (Checchi & Peragine 2010, Ramos & Van de Gaer 2012). Depending on the data availability of circumstance and effort variables, inequality of opportunity can be estimated by either of the *ex-post* or the *ex-ante* approach. Even after a broad spectrum of empirical literature, there is naturally no fixed list of circumstance or effort factors. They are not only limited by the concerned data-set but are also determined from the social and political space of the society under study. Besides many of the effort factors can themselves be shaped by an array of individual circumstances and therefore making it difficult to identify the responsibility cut in the society, for which an individual should undoubtedly be held responsible for. Therefore not unexpectedly, the literature is rather prevalent with the adoption of the *ex-ante* framework, which is also the approach adopted for the present thesis.

The basic principle to equalize opportunities in a society as proposed by Roemer (1998), is to ensure that the distribution of individual outcome should be independent of circumstances. Let $y$ denote the outcome under concern with the corresponding distribution of $F(y)$, that is thought to be generated by two classes of factors, circumstances ($c$) and

efforts ($e$). Efforts however are often unobservable in the data-set and are therefore assumed to be a function of circumstances themselves, so that borrowing from Ferreira & Gignoux (2011), the reduced form outcome generating equation can be expressed as a function of circumstances only. Therefore equal opportunity for everyone requires that the outcome distribution should be identical for any two different circumstances, $c_1$ and $c_2$, such that $F(y|c_1) = F(y|c_2)$. Needless to say that this is never the case in reality and so the difference between the distributions, $F(y|c_1)$ and $F(y|c_2)$, is actually indicative of the extent of inequality of opportunity in the society.

Any analysis of inequality of opportunity in the *ex-ante* framework therefore starts by dividing the population under study in as many as possible *types*, on the basis of available circumstance variables[1]. Some of the most widely used circumstances are parental education and occupation, birth location, race, sex or some other social attributes which are certainly not subject to any individual control. Therefore in the form of the above matrix (1), there will be $K$ rows on aggregate for a total of $K$ number of *types*, where each row corresponds to a different *type*, $c_k$, with the associated distribution of $F(y|c_k)$, where $k \in \{1, \ldots, K\}$[2]. Inequality of opportunity therefore renders to the difference in the outcome distributions between the rows and depending on the purpose of analysis, there are different ways to take account of this difference. The different purposes and approaches of estimating inequality of opportunity that are used in the thesis are discussed below.

Chapter 1 provides the index measures of inequality of opportunity that precisely estimates the unfair part of outcome inequality, generated exclusively due to the differences in individual circumstances. In particular, we quantify unequal opportunity in India during 2004-12, on the basis of five circumstance factors, that of caste, sex, region of residence, parental education and occupation, so that all possible interaction of our circumstance variables yield a total of 324 *types* to compare. In addition, the entire analysis

[1] In the literature, *types* and *circumstances* are the same class of variables which are beyond individual responsibility and are therefore often used interchangeably. But strictly speaking, *types* are all possible combination of the circumstance variables. Consider two circumstances that of sex and race, with two categories in each as (male, female) and (black, white), respectively. Then we have a total of four *types* as (black male, black female, white male, white female).

[2] Notice that without any specification of the effort factors, each column simply corresponds to each different individual in matrix (1) and not different efforts.

is done separately for three different outcome variables as well, that of the consumption expenditure, wage earning and education. The purpose of this chapter is to estimate the index measures of inequality of opportunity in India using the widely used econometric methodologies, as well as to provide the opportunity structure of India using the recently introduced method of regression tree analysis, as described below.

As far as the purpose is to estimate inequality of opportunity as the part of unfair inequality by an index measure, each *type* is represented by a counterfactual distribution that by construction, eliminates any inequality within each *type*. Most often the counterfactual is constructed by the mean outcome of the respective *type*, so that each row of matrix (1) is now represented by a singleton, that of the type-mean outcome. The inequality in the counterfactual distributions then isolates the ethically objectionable inequality that is generated only by the circumstance factors and thereby quantifies the degree of unequal opportunity in the society. The construction of the counterfactual distribution differs by the non-parametric and parametric statistical model, that eventually generates the non-parametric and parametric index of inequality of opportunity. The basic set-up of the non-parametric model was provided by Checchi & Peragine (2010) and that of the parametric one by Bourguignon et al. (2007). Chapter 1 nevertheless implements the methodological set-up of Ferreira & Gignoux (2011), who applied either set-up for quantifying inequality of opportunity for some selected countries in Latin America. However in either of the statistical model, the extent of unfair inequality is most often measured by an index from the generalized entropy class of inequality indices, the index of mean log deviation[3].

However depending on the data availability, either of the non-parametric and parametric models are estimated on the basis of a fixed set of circumstance variables that are at the discretion of the researcher. Clearly, higher the number of *types* to compare, more realistic is the index measures of inequality of opportunity. Therefore both the non-parametric and parametric indices are often estimated on the basis of maximum possible

[3]The popular use of the index of mean log deviation is attributed to its additive decomposable property, by virtue of which the total inequality can be written as a sum of fair and unfair inequality. The unfair inequality then quantifies the degree of absolute inequality of opportunity, but it is often expressed in relative terms, as the share of unfair inequality in the total inequality. *See* Lefranc et al. (2009), Björklund et al. (2012) for uses of indices other than that of mean log deviation, as a measure of inequality of opportunity.

interactions of the chosen circumstance variables, while in reality only some of them may be worthy to consider. Neither of the non-parametric or parametric method is capable of identifying the meaningful interactions that are relevant for the underlying unequal opportunity in the society. Besides, given the sample size, indiscriminate interactions of circumstances decrease the observations per *type* and may even generate some vacuous *types* with very few unusual observations. Consider for example the interaction of fathers occupation and education. The few observations corresponding to the interaction of illiterate fathers with an executive managerial profession may well be subject to some very unusual situations and indiscriminate interactions of all such circumstances may eventually have the chance to overestimate the measure of inequality of opportunity. Instead it would be rather interesting to see whether this particular interaction is indeed relevant for the resulting unequal opportunity in the society.

Brunori, Hufe & Mahler (2018) propose a novel alternative to address this issue by introducing the regression tree approach in the analysis of inequality of opportunity. In particular, the regression tree analysis exploits the algorithm of conditional inference tree, that is able to identify the most relevant interactions from the full set of circumstances submitted to the program and is able to generate a visually interpretative opportunity tree for the society under study. The initial node of the tree represents the most important circumstance factor that splits the entire sample into two distinct groups and then for each split, the program goes on dividing the sample based on an algorithm called the recursive binary splitting, potentially based on another circumstance variable and so on. Therefore unlike the indiscriminate interactions of circumstances, each terminal node of the tree here corresponds to a different *type*, that is generated by a non-arbitrary hierarchical order of circumstances. Adopting this method, Chapter 1 draws the opportunity structure for contemporary India, which reveals the educational or occupational background of parents as one of the main circumstance factor for this country as well, similar to the majority of the empirical literature.

However, to quantify the extent of unequal opportunity by an index, all of the statistical set-up as mentioned above counts on the counterfactual distribution that by construction, eliminates any inequality within the circumstances by representing each *type* by their respective mean outcomes and thereby restores the unfair inequality that is present only

between the circumstances. Therefore the index measure of inequality of opportunity relies on the very strict assumption of intra-*type* homogeneity, as representing each *type* by their respective mean outcomes masks their preference towards risk and inherently assumes that the individuals within each *type* are essentially risk-neutral. Lefranc, Pistolesi & Trannoy (2009) overcomes this issue by estimating the existence of inequality of opportunity by comparing the entire distributions for each *types* and hence does not need to assume risk-neutrality within the *types*. Therefore for any pair of different *types*, $c_1$ and $c_2$, the conclusive statistical test of stochastic dominance between the distributions of $F(y|c_1)$ and $F(y|c_2)$, determines the existence of unequal opportunity and points out the former *type* as the advantageous one, if the distribution of it dominates that of the latter, at certain order of stochastic dominance.

Chapter 2 builds on the robust method of equalization of opportunity as proposed in the literature by Andreoli, Havnes & Lefranc (2019), who advances the the distributional dominance approach of Lefranc et al. (2009) in a difference-in-difference set-up and is thereby able to rank different societies in terms of their existing inequality of opportunity, without imposing any restriction on the risk-preference behavior of the individuals within a *type*. While Lefranc et al. (2009) concludes in favor of the presence of inequality of opportunity between a pair of different circumstances, $(c_1, c_2)$, based on the statistical significance of the gap between a pair of distributions, $F(y|c_1)$ and $F(y|c_2)$, it is unable to rank two social states with evidence of unequal opportunity in each. For the same pair of *types*, Andreoli et al. (2019) applies the same concept in comparing the gaps between the *type*-specific distributions, $(F(y|c_1) - F(y|c_2))$, for different social states. For the same pair of circumstances, a reduction in this gap for one society is therefore indicative of a lesser extent of inequality of opportunity there and so the economic opportunity between those pair of circumstances can said to be *equalized* for this society.

India has a century old caste system based on a hierarchical occupational structure, where people with 'purer' occupations like priests, teachers, soldiers or traders form the relatively upper layers. However, *casteism* in its way become hereditary and children inherits the caste of their father that is not changeable for lifetime. That makes *caste* a classic circumstance factor in the context of inequality of opportunity. Even after taking several affirmative policies since 1950, evidence of caste discrimination is still rampant in

the Indian society. Applying the robust method of equalization of opportunity, Chapter 2 therefore estimates how far the consumption and earning opportunity among the different caste groups equalizes in India, over a span of nearly three decades, from 1983 to 2012. While opportunity in terms of consumption expenditure substantially equalizes among the advantageous and the disadvantageous caste groups, we found earning opportunity of the most historically disadvantageous caste category, that of the Scheduled Caste and Scheduled Tribe (SC/ST), actually deteriorates over time, especially since the mid-nineties.

While the focus of the bulk of literature is about measuring the extent of unequal opportunity among adults, Paes de Barros, Ferreira, Molinas-Vega & Saavedra-Chanduvi (2009) introduce a unique way to estimate inequality of opportunity among children, by estimating the unequal probability of access to some 'basic opportunities', like basic education, health immunization, clean drinking water or sanitized residence, for children with varying circumstances. In particular they formulate the 'access to basic opportunity' as a binary outcome variable, that takes the value 1 if children do have access to that facility and 0 otherwise. None of the above mentioned methodologies come handy in their treatment of binary dependent variables as far as measuring unequal opportunity is concerned. Inequality of opportunity in this set-up on the other hand, is measured by the *dissimilarity index* that estimates the unequal probability of access to a basic opportunity for children with different circumstances.

However, since access to a basic opportunity not only differentiates by children's circumstances, but also depends on the overall provision of that facility in the society, Paes de Barros et al. (2009) further adopt the *Human Opportunity Index* that is able to accommodate both the coverage of a basic opportunity in the society as well as its equitable or inequitable distribution among children from differing circumstances. Therefore unlike the conventional indices, the estimation of inequality of opportunity in this approach is not measured typically by an inequality index that isolates the measure of unfair inequality. Instead the Dissimilarity index quantifies the degree of unequal opportunity by the differences in the access probability of children to a basic opportunity, due the differences in their circumstances. Human Opportunity Index, that associates the dissimilarity index along with the overall coverage of a basic opportunity in the soci-

ety, is therefore often interpreted as an opportunity-sensitive development index, that is capable of quantifying development in terms of better provision of opportunities for the children as well as the penalty for development as a result of the the existing unequal opportunity in the society.

Chapter 3 therefore estimates inequality of opportunity in access to elementary and lower-secondary education for all Indian children aged between 6-18 years. Unlike the other two chapters that deals with working adults, Chapter 3 estimates whether access to basic education for children is limited due to the differences in a number of circumstance variables including some parental attributes (parents' education and occupation), social attributes (caste, sex, religion) and some other relevant family backgrounds (household consumption, residential location and sibling composition). Since schooling facilities may widely vary from state to state, we further provide a regional level analysis to estimate the relative performance of different parts of India in terms of providing basic education to all children in the twenty-first century.

For a country with an immensely hierarchical social structure, surprisingly very few works have been done on the ground of inequality of opportunity in India, with two notable exceptions of Singh (2012*b*) and Asadullah & Yalonetzky (2012). One of our contribution is to provide the latest estimates of inequality of opportunity in India until 2012, using the most extensive micro database of the National Sample Survey (NSS). Although unequal opportunity in household consumption expenditure is found to be little more than one-tenth of the total consumption inequality in Chapter 1, the same can not be said for wage earning or education. Even in 2012, nearly 40% of the earning inequality and more than one-fourth of the educational inequality is due to unequal circumstances of caste, sex, region of residence or parental backgrounds.

A particular challenge to estimate unequal opportunity, especially for the developing countries, is the lack of information on one of the most important circumstance factor that of the parental backgrounds (*see* for example Narayan et al. (2018)). There is no direct provision of parental attributes in the NSS database as well and instead they are only available for the selected households where an offspring is enumerated along with his/her parents living in the same households. As far as working adults are concerned

as the respective sample, this may raise the issue of selectivity bias due to the adult inter-generational co-residence. Therefore the estimates of inequality of opportunity for India so far, is either estimated on a different database that incorporates information on parental backgrounds but have a much lesser coverage as compared to the present database of NSS (Singh 2012*b*) or to sacrifice the circumstances of parental backgrounds altogether (Asadullah & Yalonetzky 2012).

Chapter 1 overcomes this issue by imputing the parental attributes for the entire sample from the information provided in the co-resident sample, by the widely used technique of multiple imputation (Rubin 1986). We thereby able to produce the index measure of inequality of opportunity for contemporary India including the most important circumstances of parental education and occupation, using the biggest micro-data archive on India. In fact we found that the measures of unequal opportunity are substantially underestimated when parental backgrounds are omitted from the set of our circumstances. So the information lost for not taking parental attributes, can not be captured well by the other considered circumstances like caste, sex or region of residence. Also the opportunity tree structure finds out parental backgrounds as the main circumstance variables, similar to many other empirical applications in the literature. Speaking of that, estimating the opportunity structure for contemporary India is another novel contribution of Chapter 1, that points out the intertwining among the different circumstances in generating unequal opportunity in the country. The opportunity tree further finds that some social circumstances like sex and caste are actually rather associated with parental education. While better opportunity for the males is rather protruded when parents have lesser education, the forward caste premium is also prominent with higher educated parents as well.

Provided the special case of India for the functional casteism, Chapter 2 is concentrated more on estimating equalization of opportunities among the caste groups, over a long span of time. To our knowledge this is the first analysis on India that adopts the robust methodology of equalization of opportunity to provide a rather detailed analysis of unequal opportunity due to casteism in the country. The historically disadvantageous caste categories of *Scheduled Caste* and *Scheduled Tribe* (SC/ST) are not only found to be dominated by the other caste groups till today, the gap with the relatively advantageous caste categories are actually found increasing as far as earning opportunity is concerned.

Further, due to the limited data availability, most of the caste based literature is concentrated on a rather coarse categorization of caste, where SC/ST form the lower layer and all Indian nationals other than SC/ST are amalgamated as the advantageous caste category. However in modern India, the relatively socially and economically backward caste groups among the non-SC/ST are identified as the 'Other Backward Classes' (OBC), who are entitled to several caste based reservation policies since the beginning of nineties. A rather thorough analysis in Chapter 2 reveals that the SC/STs are the worst victim of casteism and the earning opportunity gap with the so-called forward '*general*' caste group (who does not belong to either OBC or SC/ST) actually increases over the time span of 1999-2012. However the gap between the relatively lower caste categories, OBC and SC/ST, is lesser than their corresponding gap with the most advantageous 'general' caste category.

Work on inequality of opportunity among Indian children is rare and to our knowledge, non-existent with the present database. We found in Chapter 3 that children from different social and household backgrounds do have unequal access to educational opportunity even in 2012, although the situation have improved as compared to 2004. This improvement however could be attributed to the actuation of the Government mandate on free and compulsory elementary education in 2010. This chapter therefore explores the unequal opportunity among Indian children, a sample complementary to the other two chapters. A particular advantage of this non-adult sample is that co-residence does not become an issue to incorporate various information on parents into the analysis. While Indian children are found to enroll in primary schools on time irrespective of their varying circumstances, this impressive picture deteriorates as the children ages. For the same age group of children, India therefore depicts an opposite trend than that of the Sub-Saharan African countries, where school attendance as well as timely completion of the basic minimum education improves with children's age (Dabalen et al. 2015). Both the provision of basic schooling as well as its equitable distribution is impressive at the onset of schooling in India. But a sharp deterioration in terms of persuasive continuation of school education reflects on the persistent problem of school drop-outs in this country, especially for children from the lower caste rural households with lesser educated parents. East and Central India seems particularly under-performing as far as ensuring access to

education for all children is concerned.

The thesis consists of the three chapters as mentioned above and uses the data from the National Sample Survey throughout. In particular we use the Employment-Unemployment schedule of the National Sample Survey. The structure and sampling frame of this particular survey is provided in more details in the data appendix A, for each of survey rounds used in the thesis.

# Chapter 1

# Inequality of Opportunity in Indian Society

## *1.1 Introduction*

"*..The service to India means the service of the millions who suffer. It means the ending of poverty and ignorance and decease and inequality of opportunity.*" - Jawaharlal Nehru[1]

Seventy years have passed after this speech is made at the stroke of midnight on the very first day of independence of India. Over this span, India from an impoverished country, made her journey to one of the emerging global economy now. Especially since the late nineties, with a consistent high GDP growth rate of more than 7%, India has now become the sixth largest economy in the world. Much work has been accomplished with significant improvement in overall well-being of the country, but much enough, if not more, remains to be done or even addressed. Numerous studies have showed that the rapid growth of India has been accompanied by increasing inequality as well. However very few studies have yet been done to explore how much of the growing inequality is due to *inequality of opportunity*, that is how much of this high inequality is generated by factors that are purely fatalistic and therefore beyond any human control.

India followed an interventionist central planning for the first forty years after independence followed by 'neo-liberal' economic reforms at the beginning of 1990s. Since then, both the overall growth rate and inequality in India grew almost simultaneously, making

[1]Excerpt from *'Tryst with Destiny'* - a speech delivered on the first day of independence, 15th August 1947, by Jawaharlal Nehru, the first Prime Minister of independent India.

it a very relevant and active area of research concerning India. A sharp increase in consumption inequality along with a slower pace of poverty reduction has almost become a distinct feature of the Indian economy, especially in the twenty-first century[2]. But for a very stratified society like India, while there are wealth of literature on analyzing the problem of inequality, linking it to social mobility, labor market discrimination, urbanization or poverty, only a handful of them analyze how much of this inequality is due to unequal opportunities arising from varying social and family backgrounds, for which no one can be held accounted for.

The present work aspires to quantify the degree of unequal opportunity in India by estimating how much of inequality in consumption, wage and education is due to differences in caste, sex, region, parental education and occupation. Traditionally inequality had been assessed following a welfarist approach, where inequality in the final outcome was the main focus of analysis. Unequal distribution of any desirable outcome (*e.g.* income, education, standard of living, health etc.) are of primary concern for assessing social welfare. However inequality can arise from an array of different factors, some of which are purely fatalistic to the individuals. This heterogeneity in the inequality generating factors had actually triggered a philosophical debate in the late twentieth century, criticizing the fact that the classical welfarist way of inequality analysis is an approach too consequentialist to take into account the multifaceted nature of the inequality generating process (Rawls 1971, Dworkin 1981*b*,*a*). The main point of the debate is that inequality arising from factors on which no individual has any control, like race, sex, ethnicity, religion, birthplace, parental and family background, should be of primary concern from an ethical standpoint and should therefore be considered as rather *unfair*. On the other hand inequality generated from unregulated lifestyle, lack of perseverance, inadequate skill formation or poor managing ability, in other words, factors for which one can arguably be held responsible for, are not unethical and unfair in an egalitarian society.

This new approach of analyzing inequality by splitting it into fair and unfair part, brings about the question of individual responsibility in the domain of distributive justice and started to prioritize the analysis of inequality arising solely from the factors

[2] *See* Deaton & Dreze (2002), Himanshu (2007), Dev & Ravi (2007), for example. For the recent updates on Indian inequality, *see* India inequality report by Himanshu (2018).

that are beyond subjective responsibility (Arneson 1989, Cohen 1989). Inspired by this philosophical debate on the responsibility sensitive egalitarian justice, Roemer (1993) formulates inequality of opportunity as that part of inequality that is generated by factors beyond any individual control. In the jargon of inequality of opportunity (*IOP*), all such factors that are outside the periphery of individual responsibility but are responsible for generating inequality, are called *circumstances*. On the other hand the inequality generating factors that the individual can presumably control, are called *efforts*. In this dichotomous standpoint of *effort* versus *circumstances*, inequality of opportunity is that (unfair) part of inequality that had been generated only by the *circumstance factors* (Roemer 1998).

Methodologically both non-parametric and parametric approaches serve the literature to estimate the measure of *IOP* in a society. The backbone structure of these methods attributes to Checchi & Peragine (2010) (for Italy) and Bourguignon, Ferreira & Menéndez (2007) (for Brazil), respectively for the non-parametric and the parametric estimates. Although the parametric estimates of *IOP* comes at the cost of a specific functional form assumption between the outcome and the circumstance variables, it is often recommended for studies with a broad range of circumstances. Whereas the use of non-parametric approach is more common for multi-country comparison studies that is limited to a comparable set of circumstances across the countries. So far in the literature there is no universal consensus to prioritize one approach over another. But in either set up to quantify the unfair part of inequality as a measure of *IOP*, majority of the literature use an index from the generalized entropy class of inequality indices, that of the index of *mean log deviation*. Using a slightly different non-parametric and parametric set up, Ferreira & Gignoux (2011) nevertheless showed that the estimates of *IOP* are significantly close regardless of the method adopted. This is the methodological set up that we will use for measuring the index of *IOP* in India[3].

[3] *See* Roemer & Trannoy (2013), Ramos & Van de Gaer (2012) for an extensive analysis on the major methodologies used in the literature. For some international estimates of IOP, *see* Brunori, Ferreira & Peragine (2013) (selected developed countries including some Nordic countries, selected Latin American, African, Middle-East and Asian countries), Ferreira & Gignoux (2011) (Latin American countries), Marrero & Rodrìguez (2011) (Unites States of America), Checchi, Peragine & Serlenga (2010) (European countries), Cogneau & Mesplè-Somps (2008) (African countries).

Two of the major shortcomings of the above mentioned approaches are that they are based on a pre-specified number of circumstances and often uses all possible interactions of the chosen circumstances to estimate *IOP*, while in reality only some of the interactions may be meaningful. However there is no way that either of the non-parametric or the parametric set-up can point out the relevant interactions. Besides including all possible interactions also increases the total number of circumstance groups to compare, which may lead to an overestimated *IOP* as the number of observations per cell decreases. The problem is even more aggravated when some of the interactions are almost vacuous, leaving very few unusual observations in some cells. To address this problem, Brunori, Hufe & Mahler (2018) introduced a novel approach of analyzing *IOP* using the *regression tree* analysis that let the algorithm choose the most relevant circumstances in a statistically significant way from the submitted set of circumstances and generates a visually interpretative opportunity tree in the hierarchical order of circumstances. Therefore along with the non-parametric and parametric estimation of *IOP*, we adopt this approach for the present work as well to provide the opportunity structure for contemporary India.

India epitomizes a very hierarchical social structure historically, where the century old caste system is functional even in the twenty-first century. For such a stratified country there is almost no work analyzing unequal opportunity in India, with two notable exceptions. Using the National Sample Survey data, Asadullah & Yalonetzky (2012) analyzed educational opportunity in different states of India due to differences in sex, religion and caste. However for being a state-level study it is naturally focused more on the inter-state differences in terms of unequal opportunity in education rather than the national estimate. Besides due to the structure of the data base, their study can not take into account parental background as one of their circumstances which is repeatedly shown as one of the major driving factor behind unequal opportunities in a number of developed and developing countries. With a different survey Singh (2012*b*) gives a national estimate of *IOP* in India for consumption and income, that includes father's educational and occupational background as two of the major circumstances. But due to the survey structure, the inclusion of parental background limits this study to Indian men only. Besides none of the above studies gives the recent picture of India, as the latest time frame in either work is *2004-05*. The scanty work on *IOP* in India leaves significant scope

of further improvement. The aim of the present paper is to provide the latest estimates of *IOP* in India using both the non-parametric and parametric methodology, as well as to provide the opportunity structure for contemporary India adopting the recently introduced approach of the regression tree analysis.

In particular we choose three outcome variables to analyze, namely, consumption expenditure, wage earning and year of education, and analyze *IOP* for a set of five circumstance variables comprising of caste, sex, region, parental education and occupation. The present work contributes to the literature in several ways. First, using the most recent survey rounds of the National Sample Survey from *2004* to *2012*, our study gives a rather recent picture of unequal opportunity in India. We found that even by *2012*, more than one-fourth of the total wage and educational inequality is due to differences in the taken circumstances. This positions India as one of the high opportunity unequal countries in the global perspective. Second, due to the structure of the National Sample Survey it is difficult to incorporate parental information into the analysis, as the survey questionnaire have no direct provision of reporting this information. Instead parental attributes are only available for the co-resident households where parents are enumerated along with their offspring. This immediately raises the question of selection bias due to co-residence.

The present study overcome this problem by imputing information on parental background for the general sample by the widely used technique of multiple imputation (Rubin 1986). We thereby produce the estimates of *IOP* by taking into account the important circumstances of parental backgrounds but without limiting the study to the co-resident households. In fact we found that ignoring parental backgrounds as circumstances results in considerable underestimation of *IOP*, as the loss in information due to omitting parental attributes can not be captured well by the other social circumstantial backgrounds considered, like caste, sex or region. Besides, in spite of the prevalent evidence of casteism in India, differences only on the basis of caste groups is found to be not enough to capture the differences in economic opportunity arising from other sources like family backgrounds. However the opportunity structure of India shows that while sex become rather relevant when parents have little or no experience of formal schooling, the forward caste premium is not limited to the lesser educated families only. Nevertheless, the historically destitute lower caste categories are most often the most disadvantageous

people, especially if they are from the agricultural or relatively lower educated family backgrounds. This is the third contribution of our paper, that is to show how our circumstances are intertwined in generating unequal opportunity in the society.

The remaining of the paper is organized as follows. Section 1.2 sketch out the methodological framework of the non-parametric, parametric and the regression tree approach. Section 1.3 introduces our data and a clear clarification of all our variables, along with details on our sample selection criteria. Section 1.4 describes our results in different subsections. After discussing the main non-parametric and parametric measures of *IOP* in India, we give a brief account on the relative importance of caste and other social backgrounds, in comparison with the parental backgrounds. The opportunity structure for contemporary India is discussed next, separately for all of the different outcome variables. Section 1.5 concludes.

## *1.2 Theoretical and methodological background*

In the analysis of inequality of opportunity, any social outcome is supposed to be generated by two broad classes of factors. Factors that are beyond individual responsibility or *circumstances (C)* and factors that are within individual control or *efforts* ($e$). Therefore, borrowing from Ferreira & Peragine (2015), the simplified outcome generating process can be written as -

$$y = f(C, e) \tag{1.1}$$

Such that the outcome to be analyzed, $y$, can be determined from a finite set of circumstances, $C$, and efforts, $e$. From the standpoint of responsibility sensitive egalitarianism any outcome inequality generated by $C$ is ethically objectionable, whereas inequality arising from $e$ can be considered legitimate[4].

Any analysis of *IOP* therefore begins with the clear classification of the circumstance and the effort variables. However there are no fixed list of circumstance or effort variables to be taken into account, as they are subject to data availability and are rather determined in the social or political space that varies between different societies (Roemer & Trannoy 2013). Nevertheless as common to any empirical exercise, estimates of *IOP* crucially

[4]Lefranc et al. (2009) introduced a third factor, that of *luck*, in the study of *IOP*, which we did not consider in the present work.

depends on the data structure and partial observabiltiy of circumstance or effort factors severely limit the study. Data availability on effort factors in particular, are even more limited for a large number of surveys. However *IOP* is the amount of inequality generated by circumstances only and *efforts*, the so called legitimate source of inequality, can itself be determined by the existing social circumstances. Hence effort variables themselves are often assumed to be a function of circumstances, so that the outcome generating process in *equation* (1.1) can actually be reformulated as a reduced form equation, $y = g(C)$, where outcome is a function of circumstances only (Ferreira & Gignoux 2011). Of course higher the number of circumstances taken into account, more realistic is the measure of *IOP*. But with addition of new circumstances into the analysis *IOP* will always increase as long as the added circumstances are not orthogonal to the outcome in concern. Since it is impossible for any survey to provide a complete exhaustive list of circumstances, Ferreira & Gignoux (2011) therefore advice to interpret any resulting estimates of *IOP* as the lower bound of the true *IOP* in the society.

Unlike the traditional inequality approach, social welfare in the responsibility sensitive domain is not judged on the basis of total inequality in the outcome variable, $I\{y\}$. Rather *IOP* is the measure of only that part of outcome inequality that is generated by the circumstance factors, $C$, exclusively. So the main methodological challenge for quantifying *IOP* is to quarantine this unfair part of outcome inequality. This is usually done in the literature by constructing suitable counterfactual distributions, $y^{CF}$, such that by construction, $y^{CF}$ is able capture the variability in the outcome arising uniquely from the differences in the circumstance variables, $C$. The measure of *absolute IOP* in the society can then be measured by the inequality in the counterfactual distribution, $I\{y^{CF}\}$. However since *IOP* is estimated as that part of total inequality which is unfair and morally objectionable, it is a common practice in the literature to provide the estimates of *relative IOP* as the share of unfair inequality in the total outcome inequality by $I\{y^{CF}\}/I\{y\}$. The construction of the counterfactual distributions and hence the measurement of *IOP*, varies with the non-parametric or the parametric statistical model of analysis as discussed below.

### 1.2.1 Non-parametric approach

The non-parametric method for the present analysis have been adopted from the work of Ferreira & Gignoux (2011). Consider a finite population set, $i \in \{1, ..., N\}$, characterized by $\{y_i, C_i\}$, standing for outcome and circumstances respectively. Assume that the vector $C_i$ consists of $J$ elements and each of the element can take $x_j$ number of values or categories. Usually groups formed by all possible interactions of the circumstances are called *types*. In this framework, the population under study can thus be partitioned into a maximum of $\bar{K} = \prod_{j=1}^{J} x_j$, exhaustive and mutually exclusive *types*.

From the viewpoint of *IOP* any inequality *between types* is 'unfair'. To isolate this unfair inequality each of the $k$ types are represented by a 'smoothed distribution' of their respective mean outcomes. Thus every individual in a *type*, $i \in \{k, k = 1, \ldots, \bar{K}\}$, are assumed to be characterized by the *type-mean* outcome, $\mu^k$, for each $k = 1, \ldots, \bar{K}$. Therefore the counterfactual distribution to quarantine the inequality generated exclusively from the differences in *types*, is represented by, $y^{CF} = \{\mu_1, \ldots, \mu_{\bar{K}}\}$. The absolute and relative measure of *IOP* can then be estimated as[5] -

$$\theta_a^{NP} = I(\{\mu^k\}) \tag{1.2}$$

$$\theta_r^{NP} = \frac{I(\{\mu^k\})}{I(\{y_i\})} \tag{1.3}$$

Where $I(\{x\})$ denotes inequality in the distribution of $x$. Following the extant literature, $I(\cdot)$ is measured by the index of Mean Log Deviation $(MLD)$[6].

### 1.2.2 Parametric approach

The parametric approach in the present work has been adopted from Ferreira & Gignoux (2011) as well, which also essentially estimates *IOP* by the mean outcome conditional on *types* by the OLS estimates, but differs from the non-parametric set up in its construction of the counterfactual distribution to isolate the ethically unfair part of inequality.

[5] NP stands for Non-parametric, $r$ for relative measure and $a$ for absolute measure

[6] MLD(x)=$\frac{1}{N}\sum_1^N \ln \frac{\bar{x}}{x}$

The parametric set up usually assumes a log-linear relationship between the outcome and the circumstance/effort variables. So the income generating process can be written as -

$$\ln y_i = \alpha C_i + \beta e_i + u_i \tag{1.4}$$

However, as mentioned before, the effort factors can fairly be assumed as a function of circumstances as below -

$$e_i = \gamma C_i + v_i \tag{1.5}$$

with $u_i$ and $v_i$ being the random errors.

Hence, from the structural model (1.4) and (1.5), the *reduced form* income generating process can be summarized as -

$$\begin{aligned} \ln y_i &= \alpha C_i + \beta(\gamma C_i + v_i) + u_i \\ &= (\alpha + \beta\gamma)C_i + (\beta v_i + u_i) \\ &= \Psi C_i + \varepsilon_i \end{aligned} \tag{1.6}$$

From the OLS estimates of *equation* (1.6), $\hat{\Psi}$,*IOP* is then measured in comparison to a hypothesized distribution, $\{\tilde{y}_i\}$, that eliminates any differences in individual circumstances, as -

$$\tilde{y}_i = \exp[\bar{C}_i\hat{\Psi} + \hat{\varepsilon}_i] \tag{1.7}$$

where, $\bar{C}_i$ is the mean of circumstance variables across the population. Thus *equation* (1.7) eliminates the differences in circumstances by replacing them with their mean values and the associated inequality, $I(\{\tilde{y}_i\})$, is therefore segregated as fair, by construction. The measure of absolute *IOP* can then eventually be estimated from the counterfactual distribution, $y^{CF} = (\{y_i\} - \{\tilde{y}_i\})$, that isolates the outcome variations generated from the differences in individual circumstances only. So the relative share of *IOP* in the total inequality is given by[7] -

$$\theta_r^P = \frac{I(\{y_i\}) - I(\{\tilde{y}_i\})}{I(\{y_i\})} \tag{1.8}$$

[7] *Where* $P$ and $r$ in the superscript and subscript stand for parametric and relative measure respectively.

Similar to the non-parametric approach, we use the same index of *MLD* for the parametric estimates of *IOP* as well.

### 1.2.3 Regression tree approach

Circumstances by definition are all possible factors that are beyond individual responsibility and it is physically impossible for any data set to capture all such factors under a single or multiple survey. Research on *IOP* is therefore always restricted to a subset of the total set of circumstances. But as long as the omitted circumstances have non-trivial effect in predicting the outcome variable, addition of each such circumstance will increase the estimate of *IOP* by virtue of finer partitioning of the population. Clearly higher the circumstances taken into account, more realistic is the estimate of *IOP*. However addition of new circumstances also comes at a cost. Not only that this finer sample partitioning leaves fewer observations for each *type*, but some *types* may have very unusual observations due to this unrestricted partitioning. This may contaminate the associated measure of *IOP*. The regression tree analysis coined in the literature by Brunori et al. (2018), makes an attempt to allay this issue in the fashion of machine learning.

Once again assume that for individual $i$, the circumstance vector, $C_i$ consists of $J$ elements, $C_i \in \{C_i^1, \ldots, C_i^J\}$, each of which can take $x_j$ number of values, where $j \in \{1, \ldots, J\}$. Unrestricted partitioning will then divide the population into, $\bar{K} = \prod_{j=1}^{J} x_j$, number of *types*, considering all possible interactions among the circumstances. However for a large number of $C_i$ and/or $x_j$ variables, observations in all or some of the cells in $\bar{K}$ may get too crunched to allow the researcher to use all available *types*, especially when sample size is relatively less. Besides in case of unrealistic vacuous interactions, some cells may suffer from no observations at all. Since there is no way to point out the relevant interactions in either of the non-parametric or the parametric modeling, the conventional resort is either to regroup the circumstances in broader categories (less $x_j$) or sacrificing some of the circumstances (less $C_i$) or both. In the regression tree approach instead, the researcher submits the full set of available circumstances, $C_i$, to the program and let the algorithm choose the relevant partitioning of the sample under study in a non-arbitrary way, by *recursive binary splitting* to be precise.

The recursive binary splitting is a type of permutation test, because it rearranges the labels on the observed data set multiple times and computes test statistic (and $p$-value) for each of this rearrangement. It starts by dividing the full sample into two distinct groups based on one circumstance factor and then continue the same for each split, potentially based on another circumstance, into more subgroups and so on. The criteria for the selection of splitting circumstances depends on the type of regression tree used. Brunori et al. (2018) uses the conditional inference tree algorithm to determine the splitting criteria as follows.

The algorithm runs in two stages as -

- **Stage - I**: Selecting the initial splitting circumstance
  - It starts with the simultaneous testing of the $J$ partial hypothesis, $H_0^{C^j}$ : $D(Y|C^j) = D(Y)$ for $j \in \{1, \ldots, J\}$. Notice, this precisely is the testing of the existence of $IOP$, to see if any circumstances have any effect on the outcome.
  - Adjusted $p$-values, $p_{adj}^{C^j}$, are then computed with the standard adjustment for multiple hypothesis testing[8] and identifies the circumstance, $C^*$, with the highest degree of association, that is, the circumstance with the minimum $p$-value, $C^* = \{C^j : \text{argmin} \quad p_{adj}^{C^j}\}$[9].
  - The algorithm stops if the $p$-value associated to $C^*$ is greater than some pre-specified significance level, $\alpha$[10]. Hence, if $p_{adj}^{C^*} > \alpha$, the null of equality of opportunity for the society, can not be rejected at $\alpha\%$ level of significance. Otherwise, the circumstance, $C^*$, is selected as the initial splitting variable.
- **Stage - II**: Growing the opportunity tree
  - Once $C^*$ is selected, it is split by the binary split criterion to grow the tree. For each possible binary partition, $s$, involving $C^*$, the entire sample can be split into two distinct parts as, $Y_s = \{Y_i : C_i^* < x_j\}$ and $Y_{-s} = \{Y_i : C_i^* \geq x_j\}$.

[8] The adjustment is the Bonferroni correction, $p_{adj}^{C^j} = 1 - (1 - p^{C^j})$ (Brunori et al. 2018, p. 8).

[9] To test the association between the outcome variable and the covariates, the linear statistics form, along with its mean and variance, is provided in Hothorn, Hornik & Zeileis (2006), where from the relevant test statistic and $p$-value can be formulated.

[10] Like Brunori et al. (2018) we also choose $\alpha = 0.01$

– For each binary split, $s$, the goodness of split is tested by testing the discrepancy between $Y_s$ and $Y_{-s}$[11]. The split, $s^*$, with the maximum discrepancy, that is with the minimum $p$-value, is then selected as the optimum binary split point, based on which the sample is now partitioned into two sub-samples, constructing the initial two branch of the opportunity tree.

– The entire algorithm is then repeated for each branch separately, to construct the full opportunity tree.

## *1.3 Data, variables and sample selection*

### 1.3.1 Data

For the present analysis of inequality of opportunity in India we have taken data from the National Sample Survey (*NSS*). This is the biggest nationally representative micro level database for India, collected by the National Sample Survey Organization (*NSSO*), India. Among the many national level surveys conducted by *NSSO* we have taken the *Employment Unemployment Survey* in particular. This survey is conducted for a year in every five years, covering the whole country except some remote inaccessible area[12]. For focusing on the recent scenario in India, we have taken the latest two employment-unemployment survey rounds of *NSS*, covering years *2004-05* and *2011-12*[13].

These rounds on average, survey 110000 households enumerating about 0.4 to 0.6 million individuals. India is predominantly rural even to date with a rural-urban ratio around 70 : 30 on average. Initially we have to drop about 1000 observations per round to clean for valid age, sex, sector, caste specification, marital status and some other criterion, depending on different rounds. *NSS* provides details on several household and individual characteristics. Some of the major household provisions include household size, religion, caste and consumption expenditure, whereas age, sex, education, occupation and many other demographic characteristics are recorded for each member of the household. How-

[11]This is tested by the two sample test statistics, provided in Hothorn et al. (2006). The entire algorithm can be executed by an R package, developed by the same authors.

[12]So conflict areas of Ladakh & Kargil districts of Jammu & Kashmir, some remote interior villages of Nagaland, few unreachable areas of Andaman & Nicobar Islands and those villages recorded as uninhabited by respective population census, are kept out of these surveys.

[13]This means we have taken *Schedule 10* survey of *NSS*, for rounds 61 (2004-05) and 68 (2011-12). Details of these database are in the NSSO data appendix A.

ever not everybody reported as 'employed' do have information on their income, rather wage earning is selectively reported in the *NSS* data only for the regular and the casual wage earners who are not self-employed. Another possible drawback in the structure of *NSS* data base is that it does not report information on parental background directly for every individual. Rather this crucial information is only available for households where the offspring is enumerated along with his/her parents.

### 1.3.2 Definition of variables

#### Circumstance variables

For the present analysis, we have chosen a set of five circumstance factors, that of caste, sex, region of residence, parental education and father's occupation. We can label the first three of them as social backgrounds, while parental education and occupation constitute parental background. With all possible interactions of these five circumstance variables, we have a total of 324 *types*.

Caste system in India is a century old hierarchical social structure based on occupation. However the historical occupational perspective in its way became hereditary over time and children always inherit the caste of their father that is unchangeable for lifetime. There are thousands of castes in the country, which are regrouped in fewer caste categories by the constitution of India for the purpose of caste based affirmative policy or reservations. We consider three caste categories in our analysis. The lower caste category consists of the *Scheduled Castes* and the *Scheduled Tribes* caste categories together ($SC/ST$). They are the most historically disadvantageous caste groups in India and are designated the reservation status since 1950. Around mid-eighties, the socially and economically backward castes among the non-$SC/ST$s are further categorized as the *Other Backward Classes* ($OBC$) who are entitled to certain reservation quotas in higher education and Government jobs since the beginning of nineties. Indian nationals do not belong to any of the above mentioned caste categories are formally called as the *General* category individuals and are excluded from any caste based affirmative policies by rule. $OBC$s can be thought of as the middle level caste category who are usually little more advantageous than the historically disadvantageous caste categories of $SC/ST$, but have lesser economic advantage as compared to the forward *General* caste category.

Considering the bulk of literature on gender discrimination in India, we take two categories of sex, male and female, as our next social circumstance. To consider region as one of our circumstances, we have to take region of residence, although the ideal circumstance factor would be the birth region. Due to unavailability of information on birth place, we have to consider the present residing region as a proxy for birth region, which is not a far fetched assumption given the low rate of inter-state and inter-district migration in India as per the recent migration survey report of NSSO (2008). To further minimize migration related contamination, we take six broad regional categories for our analysis as - North, East, Central, North-East, South and West[14].

Our next batch of circumstances consists of parental background that includes two kind of parental attributes, that of parental education and occupation. By combining father's and mother's education, we take three categories of parental education as - (i) both parents have no formal schooling (ii) at least one parent has primary or below primary schooling (that means the other parent, either have the same level of schooling or less) and (iii) at least one parent has above primary schooling. It is worth a mention, that 'no formal schooling' is not equivalent to illiterate parents, as they may have exposed to other informal adult literacy programs, but have never experienced formal schooling. Due to considerably low information on mother's occupation, we took three categories of father's occupation as a proxy of parental occupation. The occupational categories are taken as father's employment in - (i) white collar job (ii) blue collar job and (iii) agricultural occupation. White collar job category includes all sorts of professional, executive and managerial jobs. Whereas, sales and service workers falls in the domain of blue collar workers. Agricultural job includes horticulture, fishing and hunter-gatherers as well.

### Outcome variables

The analysis of *IOP* on India is executed for three different outcome variables - consumption, wage and education. All the three outcome variables are considered as continuous variables. While first two of them is measured in monetary units (Indian Rupee, INR),

[14]Statewise composition: Jammu & Kashmir, Himachal Pradesh, Punjab, Haryana and Uttarakhand - *North*; Bihar, Jharkhand, Orissa, West Bengal - *East*; Uttar Pradesh, Rajasthan, Madhya Pradesh, Chattisgarh - *Central*; Sikkim, Arunachal Pradesh, Assam, Nagaland, Meghalaya, Manipur, Mizoram, Tripura - *North-East*; Karnataka, Andhra Pradesh, Tamilnadu, Pondichery, Kerala, Lakshadeep - *South* and Gujrat, Daman & Diu, Dadra & Nagar Haveli, Maharashtra, Goa - *West*.

education is measured as the years of school/college education.

Consumption is considered as the monthly per capita consumption expenditure (*MPCE*). This is the total monthly expenditure on certain durable and non-durable goods incurred by the household over the last thirty days prior to the date of the survey. This data is therefore reported at the household level, which we divide by the respective household size to get the individual level values. The list of goods, expenditure on which is to be reported is a selection of goods that has been considered as the most important ones by the respective survey. Borrowing from Hnatkovska, Lahiri & Paul (2012), we use the real *MPCE* as our outcome variable, upon dividing *MPCE* by the state level absolute poverty lines[15].

Our next outcome variable is the wage earning, which is reported only for the regular and casual wage earners. Therefore the wage data is not available for a large chunk of self-employed individuals who constitute nearly 40% of the working adults in India. Unlike *MPCE*, wage is reported as the weekly wage received or receivable for multiple activities, by each regular/casual earning members of the household over the last week prior to survey. The main reason for reporting wage as an weekly input is that many of the Government or non-Government public work programs in India are transitory in nature that employ a huge number of rural casual laborers. However we consider wage corresponding to the major activity that had been pursued for the maximum number of days over the reference week. In case of equal number of days spent on more than one activity, we prioritize those having valid wage entry and occupation information. In particular we consider the daily real wage earning as our outcome variable by dividing total weekly wage by the number of days engaged in that major activity. Similar to *MPCE*, the corresponding real wages are generated upon division by the state level absolute poverty lines.

[15]We use poverty lines, that can account for the differences in standard of living across the states of India. Besides, the measure of absolute poverty line is provided by the Planning Commission of India using data collected by the same survey, that of the National Sample Survey, the one we use for the present analysis. Another commonly used deflator is the consumer price index, which we did not use, as it was measured on the basis of a different survey and prior to 2011, the combined rural and urban price indices are not provided (instead, consumer price index used to comprise of multiple series like, urban non-manual labor, agricultural labor, rural labor and industrial workers).

Our third outcome variable is education. However our data base provides information on education in different categorical codes, that is recorded for each individual as their highest educational achievement at the time of the survey. We converted the education codes in suitable number of school/college years, based on the standardized norm in the country and use years of education as our outcome variable in concern. We assign 1 year of education to the lowest category of 'without formal schooling'. As 'without formal schooling' incorporates literacy through other informal medium, we reserve one year of education for this category as a cognitive margin of low education. Further, 2 to 4 years of education is assigned to educational categories corresponding to primary or below-primary level of schooling. For education above the formal primary schooling 8 to 16 years of education are assigned, that covers a broader range of reported educational codes from below secondary level to graduate level college education. Since our data has the further provision of some additional technical education (like certain under-graduate or graduate levels diploma/certificate course), we update the years of education accordingly, for those who have reported to have some technical education[16].

### 1.3.3 Sample selection

As mentioned before, *NSS* does not provide information on parental attributes for every individual, making this data limited to the co-resident households that consists of both offspring and parents as the respondents. Provided the instrumental role of parental backgrounds in the analysis of unequal opportunities for a number of countries, the study on India will remain incomplete had we not consider that. Therefore given the data structure, the biggest challenge in the sample selection procedure is how to best incorporate the valuable information of parental backgrounds in our analysis of *IOP* in India.

Studies for which parental information may be important, like the analysis of inter-generational mobility or inequality of opportunity, when using the *NSS* data base, usually deal with this issue either by restricting their analysis to the co-resident households (*e.g.* Hnatkovska, Lahiri & Paul (2013) for inter-generational mobility analysis) or by sacrific-

[16]We draw upon the work of Hnatkovska et al. (2012) while updating the year of education for technical education. The mapping of education codes to years of education is provided in the NSSO data appendix A.

ing the parental background data (*e.g.* Asadullah & Yalonetzky (2012) for educational opportunity analysis). As mentioned before we already ruled out the second option considering the importance of parental attributes in *IOP*. However to analyze *IOP* we want our sample to be restricted to working adults who have reportedly finished their education. Provided that, the other option to include parental attributes is to limit our analysis to households with adult inter-generational co-residence. Although adult parent-child co-residence is not an uncommon social pattern in India, it may raise the issue of selectivity bias. So to provide estimates of *IOP* in India with a nationally representative sample, we impute the parental attributes for our sample using the technique of *multiple imputation*. Our sample therefore consists of working adults who are aged between 18 to 45 years, are not currently enrolled in any educational institution, are from male-headed households (who also are the only head of the household) and have valid information on education and occupation, both for themselves and for their parents[17]. However for estimating *IOP* in wage, we further restrict our sample to those who additionally provide valid data on wage.

The theory of multiple imputation was introduced by Rubin (1976, 1986) for dealing with the problem of missing data due to non-response in large survey data sets. Although mostly popular in the statistical and medical research, the use of multiple imputation to handle missing values is expanding in economics as well, especially in the survey data based econometric analysis[18]. In particular, Teyssier (2017) showed the efficacy of multiple imputation for imputing parental information for a data set on Brazil, for which this information is also available without the co-residence issue. We want to impute two parental attributes in particular, that of parental education and father's occupation, both of which are considered as categorical variables in our estimation of *IOP*.

We first form our sample as per the sample selection criteria mentioned above, except the criteria related to parental attributes. We can now think of this sample as the union of two exhaustive and mutually exclusive parts - the 'response' and the 'non-response' part. While the 'response' part have valid information on parental background, this crucial

[17]We exclude multi-headed and female headed households in India, as they are rare and subject to special constraints. Over 90% of heads are male and 99% households are single-headed-household.

[18]For application of multiple imputation technique in poverty and inequality analysis, *see* Alon (2009), Jong-Sung & Khagram (2005), for example. Whereas, Salehi-Isfahani, Hassine & Assaad (2014), Teyssier (2017), provide estimates of IOP using multiply imputed circumstances.

information is missing for the other part. The exercise of multiple imputation is to use information from the 'response' part to *impute* values for the 'non-response' part, using all possible auxiliary information provided by the data set that are non-missing for both of the 'response' and the 'non-response' part. In our case the 'response' part consists of the co-resident data points for which parental background is observed[19]. Table 1.B.1 in Appendix 1.A reports the summary statistics of the 'response' and the 'non-response' sub-samples. It shows that co-residence does not seem to make a marked difference in terms of caste, occupation and rural-urban composition. But notice that the samples of the 'response' part, as expected, are relatively younger. Hence is the justification of taking relatively younger adults (18-45 years) for our analysis, so as to keep parity between the 'response' and the 'non-response' part.

The two parental variables in concern, that of the parental education and father's occupation are then estimated for the 'response' part by a suitable imputation model (an ordered logistic regression, in our case), using a broad range of predictors including households, individuals and some survey related characteristic variables that are strictly non-missing for both the 'response' and the 'non-response' part[20]. Parental attributes for the 'non-response' part is then imputed from simulated draws of the posterior distribution of these estimates. However as the name suggests, the imputation of the missing values is done for a multiple number of times generating multiple number of 'completed' data-sets, where none of the attributes are missing any longer. We adopt the sequential regression multiple imputation algorithm of Raghunathan, Lepkowski, Van Hoewyk & Solenberger (2001) and use 20 imputations in particular. Both the non-parametric and parametric measures of IOP are then analyzed separately over each of the 'completed' data-set and combined by Rubin's rule (Rubin 1986) to give the final measures of *IOP*.

[19]In particular we consider our 'response' part to constitute of samples who are living with their parents, with father as the household head. However a co-resident household may consist of other members with information on parents as well. Two cases in particular are excluded. First we did not take grandchildren of the household head for simplicity. Secondly, households where the adult working child share the headship and is living with one of his/her parents should also be taken into account, but could not be, because in this case *NSSO* reports father/mother/father-in-law/mother-in-law by a single code, making it impossible to extract information on biological parents. However these two cases together do not exclude more than 10% of the sample, as far as adults are considered.

[20]This includes some household characteristics like household size, caste, sector (rural/urban), religion, consumption expenditure and offspring's' characteristics like their age, relation to head, marital status, region of residence, sex, occupation, education, along with some other survey-specific attributes. Further details of our imputation model are provided in Appendix 1.A.

However the exercise of multiple imputation does *not* mean to 'create' the missing values in a deterministic fashion, but rather to capture the additional features of the 'response' part to use it in the final analysis. Therefore two of the important criteria for a successful imputation are, that the imputation model should provide good estimates of the missing parental attributes from a bunch of non-missing variables and that the relation between them remain the same for the 'non-response' part as well. While the former can be tested by the imputation model diagnostics, given the data-set the latter can at best be reasonably assumed (Marchenko & Eddings 2011). In particular the second criteria of a good imputation requires that the probability of the missing data does not depend on any unobservable factor and hence can be imputed successfully from the imputation model (Allison 2000). Our imputation exercise and eventually the measures of *IOP* also bank on this assumption, which is the so called assumption of 'missing at random' (*MAR*)[21]. Summary statistics of our *sample*, as well as our sub-sample for the wage analysis (wage sample), is given in Table 1.1.

| | age | hhsize | %rural | %married | %noschool | %agri | %wage | N |
|---|---|---|---|---|---|---|---|---|
| ***Working sample*** | | | | | | | | |
| *2004-05* | 32.11 | 5.5 | 0.76 | .82 | .36 | .53 | .41 | 127002 |
| [61] | (0.03) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *2011-12* | 32.74 | 5.0 | 0.72 | .82 | .24 | .45 | .48 | 90574 |
| [68] | (0.05) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| ***Wage sample*** | | | | | | | | |
| *2004-05* | 31.80 | 5.0 | .69 | .80 | .37 | .42 | 1.0 | 48127 |
| [61] | (0.05) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *2011-12* | 32.24 | 4.7 | .65 | .79 | .24 | .33 | 1.0 | 41619 |
| [68] | (0.07) | (0.01) | (0.00) | (0.00) | (0.00) | (0.01) | | |

Table 1.1: Work sample summary statistics [a]

[a] standard errors are in parentheses and round numbers are in squared brackets. 'age' and 'hhsize' reports the mean age and household size of our sample. %rural, %married, %noschool, %agri and %wage reports the share of rural sample, married individuals, samples without any formal schooling, samples engaged in agricultural jobs and samples who further have the information on wage data, respectively. The last column (N) reports the respective sample size.

[21] Since we can never actually test whether the missing-ness depend on some unobservable factor not provided by the data-set, we have to assume MAR. However, since adult inter-generational co-residence is the rather prevalent social pattern for most part of India, it is quite reasonable to assume that parental attributes does not depend on some hidden unobservable factors beyond the provision of the survey. Another assumption that of 'missing completely at random' (MCAR) is also mentioned in the literature, which assumes that the probability of missing-ness is random. This is rarely the case for any survey data and so for NSS, because co-residence is clearly more probable for younger males and less for females (for female migration due to marriage). However a number of literature suggests that the assumption of MAR is good enough for a reasonable imputation (Rubin 1976, Little 1988, Allison 2000, Raghunathan et al. 2001). Appendix 1.A provides further details of our imputation algorithm and diagnostics.

Table 1.1 reports the mean age, household size (hhsize), share of rural sector, share of married samples, share of individuals without any formal schooling (noschool) and share of population engaged in agriculture (agri) in our working sample along with the respective sample size. First of all, similar to the general picture of the whole country, our sample is predominantly rural with a substantial population in agricultural occupations. However even in *2012*, nearly one-fourth of our sample have no experience of formal schooling ever. The last but one column reports the percentage share of our working sample to have information on wage data. It shows that more than half of our working sample do not have information on wage data, which explains the massive reduction of sample size for our wage sample. The lower panel of Table 1.1 shows that the regular and casual wage earners are usually less rural and less agricultural.

| circumstances → | Caste | Sex | Region | Parental education | Father's occupation |
|---|---|---|---|---|---|
| share of → | SC/ST | male | north | no schooling | agriculture |
| ***Working sample*** | | | | | |
| *2004-05* | 29.1% | 77.2% | 6.6% | 55.0% | 62.8% |
| *2011-12* | 29.7% | 82.1% | 6.9% | 46.9% | 54.3% |
| ***Wage sample*** | | | | | |
| *2004-05* | 32.0% | 80.0% | 7.9% | 56.5% | 56.7% |
| *2011-12* | 34.1% | 83.9% | 7.8% | 46.4% | 48.1% |

Table 1.2: Circumstance specific summary statistics[a]

[a]Each column shows the percentage share of our samples who are - SC/ST, males, residents of Northern region, have both parents without any formal schooling and have agricultural fathers, respectively.

Table 1.2 gives the circumstance specific composition for each of our five circumstance variables (caste, sex, region, parental education, father's occupation). The samples across the rounds looks comparable with identical proportion of circumstances, especially in terms of social background circumstances (caste, sex, region). Due to low female labor force participation, both of our working and wage sample are rather male dominated with even higher proportion of males in the wage sample[22]. Besides as per with the national population, Northern India has relatively less number of samples[23]. Although our wage sample has relatively more lower caste individuals, caste composition for either

[22]In the chosen age group (18-45 yrs.), about 30% females are currently employed, while, on average, 65% are reported as not in labor force for attending domestic duties.

[23]Notice that many parts of the Northern and North-Eastern India are more likely to be out of the *NSS* sample coverage for having relatively more conflict zones and remote areas.

of our sample is close to the national proportion. For both of the survey rounds, nearly 30% of our sample are from the destitute caste groups of *SC/ST* which is similar to the caste proportions in the country as a whole. About 46-56% of both of our working and wage sample have neither parents with any formal schooling experience. Besides most of the samples are from agricultural households where fathers are engaged in agro-based occupations. However similar to their own education and occupation as provided in Table 1.1, parental education seems better for the latest round as well, along with a lesser share of agricultural fathers.

## *1.4 Results and discussion*

### 1.4.1 Measures of IOP in India

To quantify the degree of unequal opportunity in Indian society for consumption, wage and education, we adopt both the non-parametric and the parametric approaches for at least two good reasons. First, it will serve as a robustness check to our measures of *IOP*. With the same set of circumstances, the amount of unfair inequality should not have much variation under the non-parametric and the parametric set up. Second, most of the international measures of *IOP* have used either or both of these methods. Estimating *IOP* for India under both the approaches will therefore be helpful for international comparisons. Following the extant literature, both inequality and *IOP* are always measured by the index of mean log deviation. Besides both the non-parametric and the parametric measures of *IOP* are based on all possible interaction of the full set of circumstances, *viz.* caste, sex, region, parental education and occupation, leaving us a total of 324 *types* to compare[24].

Table 1.3 reports the *relative IOP* as well as the measure of total inequality, for MPCE (consumption), wage and education. The first row reports the amount of total inequality in each of the outcome variables separately. Inequality is highest for education that lies between 0.39 to 0.46 over the time frame of *2004-12*. Whereas over the same time span inequality in consumption and wage is close by and hovers around 0.24 on average. Notice that other than education, inequality in all other outcome variables are actually

[24]The 324 *types* correspond to the interaction of - caste(3)×sex(2)×region(6)×parental education(3)×father's occupation(3), where number of categories for each circumstances are in parentheses.

increasing over the eight years time span considered here. Particularly consumption inequality for our sample shows a rather sharp increase for the latest survey year which is at par with the recent trend in Indian economy. A number of literature documents the increasing consumption and earning inequality in India since the country switched from a centrally interventionist policy to a rather neo-liberal open-market policy regime around early nineties[25].

| | MPCE | | Wage | | Education | |
|---|---|---|---|---|---|---|
| | 2004-05 | 2011-12 | 2004-05 | 2011-12 | 2004-05 | 2011-12 |
| Inequality | 0.19681 | 0.28527 | 0.22519 | 0.25101 | 0.46248 | 0.39063 |
| Measures of relative IOP | | | | | | |
| Non-parametric | 0.11051 | 0.11172 | 0.32896 | 0.39310 | 0.31968 | 0.26707 |
| Parametric | 0.10378 | 0.10661 | 0.31461 | 0.37747 | 0.30591 | 0.24803 |

Table 1.3: Measures of Inequality of opportunity in India[a]

[a]All IOP measures are the relative measures of IOP and therefore reports the percentage share of IOP in the total inequality upon multiplied by 100. So the non-parametric estimation of IOP in education for 2011-12 tells that 26.7% of educational inequality is due to unequal circumstances in that survey year.

The last two rows of Table 1.3 reports the non-parametric and the parametric measures of relative *IOP* respectively, using all possible interaction of the chosen circumstances. So the non-parametric *IOP* for education says that 26.7% of the high educational inequality is due to differences in the chosen set of circumstances during the survey year of 2011-12 and therefore strictly unfair from an ethical perspective. Similar to Ferreira & Gignoux (2011), we found the non-parametric measures for each outcome to be always little higher than the corresponding parametric measures of *IOP*. However for all the respective outcome variables, the measures of *IOP* are close-by under both of the statistical set-ups (non-parametric and parametric), indicating that our results are actually robust to the method adopted.

Among the three outcome variables considered, Table 1.3 shows that the share of ethically unfair inequality is relatively low for *MPCE*. About 11% of consumption inequality is due to unequal opportunities arising from the differences in the chosen circumstances. The degree of consumption *IOP* in India is still a bit higher than most of the developed countries and in fact positions India closer to the Sub-Saharan African countries

[25] *See* for example, Deaton & Dreze (2002), Himanshu (2018).

(Cogneau & Mesplè-Somps 2008). The same can not be said for wage and education though. Over the time span of *2004-12*, about 33-39% of wage inequality in India is conditioned by unequal social and parental backgrounds. At least in terms of wage *IOP* with a comparable set of circumstances, India seems worse than Brazil that has found to be as one of the most opportunity unequal country in Latin America (Ferreira & Gignoux 2011). Education on the other hand, in spite of having a much higher level of inequality than wage, shows a comparable degree of *IOP* that on average accounts for about 29% of total inequality. However even by *2012*, more than one-fourth of educational inequality and more than one-third of earning inequality in India is due to unequal opportunities, arising from differences in circumstances that are beyond any subjective control.

Although Consumption and wage are often analyzed side by side in many of the development studies as two comparable source of standard of living, this is not the case for the present analysis. This is because *NSS* data does not report these two variables in a comparable format and we can point out at least three major sources of variation in the reporting of the consumption and the wage data in our data base. First of all, *MPCE* is a household level data reported as the total expenditure of the household and is therefore unable to capture any intra-household differences. Wage on the other hand is likely to be rather varying in nature, as it is reported not only for every regular/casual earning members of the household but also for multiple number of activities. Second, *MPCE* is recorded for a larger recall period of a month. Whereas due to the transitory nature of many casual wage earning jobs, wage is reported for the reference week prior to the date of the survey. Together a shorter recall period along with a finer reporting unit makes the wage data to be more variant and responsive to changes in the individual circumstance factors. Finally, wage and consumption are estimated for different samples and the same set of circumstances may have a differentiated effect for different samples. In particular a large body of self-employed individuals are excluded exclusively from the wage analysis.

### 1.4.2 Effect of caste in comparison with parental background

India is one of the very few countries where the century old caste system is well embedded even to date. The origin of the caste system was found in the ancient Hindu text, where the society was divided in hierarchical occupational structure. Upper castes are

supposed to be engaged in occupations that are more pure in nature like worshiping deities or serving the country as soldiers. Whereas the major occupation of the lower caste categories is to serve the upper caste 'masters'. Caste in its way became hereditary and is identified at birth that is not convertible for lifetime. Although that makes caste a classic circumstance factor in the context of *IOP*, it is certainly not the only source of hierarchy in the Indian society and may have its effect through many channels. The purpose of the present section is not to explore these different channels, rather to show the relative importance of caste as a circumstance factor as compared to parental background and other social backgrounds, in the context of estimating *IOP* for India.

Table 1.4 reports the non-parametric relative measures of *IOP* with different set of circumstances. The first row gives the non-parametric *IOP* with the full set of circumstances and is therefore the same as the non-parametric measures in Table 1.3. From the second row onward we provide the associated estimates of *IOP* after omitting one or more of the circumstances from our analysis. Measures corresponding to the second row reports the index of non-parametric relative *IOP* after caste is omitted from our set of circumstances. Similarly the third row estimates *IOP* without taking any parental attributes (parental education and father's occupation) as our circumstances and the last row reports the same when all circumstances other than caste are omitted from the analysis. However unless the omitted circumstances are completely orthogonal to the outcome in concern, *IOP* will always increase with addition of new circumstances. It is the reason why Ferreira & Gignoux (2011) suggested to interpret the resulting *IOP* estimates as a lower bound of the true *IOP* in the society because no study can ever take into account the complete exhaustive set of circumstances. Therefore as expected, *IOP* mostly decreases as we move down in Table 1.4 from more to lesser number of circumstances.

| | Measures of relative *IOP* (non-parametric) | | | | | |
|---|---|---|---|---|---|---|
| | MPCE | | Wage | | Education | |
| Taken circumstances | 2004-05 | 2011-12 | 2004-05 | 2011-12 | 2004-05 | 2011-12 |
| *caste+sex+region+parental backgrounds* | 0.111 | 0.112 | 0.329 | 0.393 | 0.320 | 0.267 |
| *sex+region+parental backgrounds* | 0.069 | 0.099 | 0.313 | 0.363 | 0.310 | 0.248 |
| *caste+sex+region* | 0.086 | 0.047 | 0.131 | 0.161 | 0.123 | 0.102 |
| *caste* | 0.049 | 0.014 | 0.030 | 0.079 | 0.029 | 0.045 |

Table 1.4: Effect of omitted circumstances in the measure of IOP[a]

[a]'Parental background' is abbreviated to indicate circumstances related to parents and therefore includes parental education and father's occupation.

Notice that as compared to the first row with full set of circumstances, *IOP* decreases both for the second and the third row of Table 1.4, but it is the latter for which the fall in the value of *IOP* is larger for most cases. Even after omitting caste, earning and educational *IOP* in India are still mostly over 30%. Consumption *IOP* for the latest round (2011-12) also decreases marginally when only caste is omitted. On the other hand after omitting parental backgrounds from the analysis, only about 10-16% of the total inequality is deemed unfair for the presence of *IOP* in wage and education. So *IOP* more than doubled for most of the outcomes when parental background is considered as additional circumstances along with the social backgrounds (caste, sex, region), whereas it decreases marginally when only caste is omitted from the analysis. This implies that the omitted effect of caste can be captured to a large extent by the other social and parental attributes considered. The only exception is the outcome of *MPCE* for 2004-05, where the omitted caste effect is higher than that of parental backgrounds. But even after controlling for caste, sex and region, differences in parental background have non-trivial additional effect in generating unequal opportunities for all the outcome variables. Hence is the necessity of multiple imputation of information on parental backgrounds, as the social attributes alone are not sufficient to take into account the discriminatory effect of parental backgrounds.

With caste as the only circumstance variable, *IOP* in India is no more than 8% for any outcome which is even less than some of the developed countries. However a comparison in this regard is not really appropriate as most of the international studies on quantifying *IOP* involves at least one circumstance regarding parental information. Nevertheless the low estimates of *IOP* for the last row of Table 1.4 does not indicate that caste has no role to play in generating unequal opportunities in the Indian society, rather it is indicative of the fact that caste alone can not capture well the differences in other circumstances especially that of parental backgrounds. The most historically disadvantageous caste category of *SC/ST*s are indeed found to be clearly dominated by the relatively advantageous upper caste categories even in the twenty-first century and the caste premium enjoyed by the forward caste group is actually increasing over time as far as earning opportunity is concerned (*see* Chapter 2). So the practice of casteism surely adds an extra deep rooted level of hierarchy even in the social fabric of twenty-first century India, but taking caste

as the only responsible factor for quantifying *IOP* may be too coarse to account for the underlying unequal opportunity in the society.

### 1.4.3 Opportunity tree for contemporary India

Either of the non-parametric or the parametric approach uses a fixed model specification for analyzing *IOP*, where all the circumstances are given equal importance while estimating the resulting measures of *IOP* in India. However it is possible that caste may matter more in some part of the country with certain family backgrounds or educational opportunity is always less with lesser educated parents but even more when father is an agricultural worker. Neither of the non-parametric or the parametric measures have an answer to this question in the context of *IOP*. So to investigate the intertwining of our circumstances we adopt the regression tree approach that has been recently introduced in the literature by Brunori et al. (2018).

Because of our data structure we have to impute the information on parental backgrounds throughout our analysis. Although we computed the non-parametric and parametric estimates on multiply imputed data set for more precision, it is difficult to perform the same for the regression tree analysis as far as the drawing of opportunity tree is concerned. Since each imputed data set may generate slightly different opportunity trees depending on the imputed values of parental education and occupation, the interpretation of the multiple opportunity trees for a single outcome variable becomes rather complicated. We therefore pick a randomly chosen imputed data set and draw the opportunity tree for that single imputed data-set for the survey year of *2011-12*, separately for each of our outcome variables.

All the opportunity trees are drawn on the basis of the same set of circumstances as they are considered for the non-parametric and parametric analysis. So the opportunity tree for all outcome variables are therefore drawn on the basis of - (i) three categories of caste - General [Gen], Other Backward Classes [OBC] and Scheduled Castes/Scheduled Tribes [SCST] (ii) two categories of sex - male [M] and female [F] (iii) six categories of region - North [N], East [E], Central [C], North-East [NE], South [S], West [W] (iv) three categories of parental education - none of the parents have any formal schooling

[No], at least one have below primary schooling (considered as medium education) [Med] and at least one of them have above primary schooling (considered as high education) [High] (v) three categories of father's occupation - white collar [WC], blue collar [BC] and agriculture [Agr], where abbreviations in the square brackets are used to label the corresponding categories in the opportunity trees (Figures 1.1, 1.2, 1.3).

We submit this full set of circumstances to the program and let the algorithm choose the most relevant ones to draw out the opportunity tree, where the initial node represents the most important circumstance for the respective outcome. Unlike the non-parametric and parametric approaches, *types* in the regression tree are not all possible combination of the circumstances, rather each terminal node of the tree now correspond to a different *type* and is represented by the mean outcome of that *type*. *IOP* is then measured as the inequality between these *type*-mean outcomes. The major difference with the non-parametric and parametric analysis is that the regression tree traces out the most important interactions among the circumstances in a statistically significant way and estimates *IOP* only on the basis of those limited number of interactions which are chosen by the program as the most relevant ones. The opportunity tree is therefore able to produce an estimate of *IOP* that escapes the possible risk of over-fitting, arising from unregulated number of interactions. Indeed during *2011-12*, Table 1.5 shows that *IOP* in consumption is less than 7% when it is estimated using the regression tree algorithm. For the same year, unequal opportunity in wage and education are still estimated by the regression tree as about 32% and 23% of their corresponding total inequality, respectively. But for all the outcomes, *IOP* estimated by the regression tree are considerably lower than their corresponding non-parametric and parametric estimates[26].

The opportunity trees for MPCE, wage and education are presented in Figures 1.1, 1.2 and 1.3, respectively. First of all notice that except for MPCE, parental education has turned out to be the most important circumstance factor for all other outcomes, as denoted by the initial nodes of Figures 1.2 and 1.3. Whereas for *MPCE*, the most crucial circumstance is the occupational category of fathers and the average monthly

[26]Notice that although we draw the respective opportunity trees on the basis of a randomly chosen single imputed data-set, the same is not done for quantifying *IOP* under the regression tree approach. Similar to the non-parametric and parametric analysis, *IOP* is measured in the regression tree analysis using all the 20 imputed data-sets and by the index of mean log deviation.

| | Measures of relative *IOP* | | |
|---|---|---|---|
| | Regression tree | Parametric | Non-parametric |
| MPCE | 0.068 | 0.107 | 0.112 |
| Wage | 0.318 | 0.377 | 0.393 |
| Education | 0.225 | 0.248 | 0.267 |

Table 1.5: Different estimations of IOP (2011-12)[a]

[a]All IOP estimates are measured by the index of mean log deviation on multiply imputed data-sets.

consumption expenditure is always higher when fathers are engaged in non-agricultural jobs (Figure 1.1). Not only for consumption, having an agricultural family background is always less advantageous for all outcomes, whenever it matters,. Another common feature across most of the outcome variables is that the effect of some discriminatory social attributes like sex and caste, seem to be rather conditioned by differences in parental backgrounds. Figures 1.2 and 1.3 for example reflect that females have always less earning and educational opportunity than males, but even more so if parents have no or very little experience of formal schooling. Except for North-East India, women have less earning opportunity than men as well, if parents have no or medium level of formal schooling (Figure 1.2).

Similar to sex, the role of caste in the circumstance hierarchy also comes after parental attributes, but the forward caste premium is not limited to individuals with lesser educated parents only. In fact for educational opportunity, casteism has turned out to be rather relevant when parents are relatively more educated (Figure 1.3). However the forward *General* caste category has always better educational opportunity than the relatively disadvantageous caste groups of *OBC* and *SC/ST*, and even more so if their fathers are also engaged in white collar occupations. Similar to education, Figure 1.2 shows that the forward *General* caste categories also have an even better earning opportunity for most part of the country, when parents are comparatively more educated and fathers are in non-agricultural professions.

The geographical habitat have a distinguishing effect for all of our outcome variables, but its order of relevance varies across the outcomes. As compared to *MPCE* and wage, the region of residence seems to be relatively less important for generating unequal opportunity in education. While region of residence (zone) is the second most important

circumstance variable after parental backgrounds for consumption and income (Figures 1.1 and 1.2), it becomes relevant for education at a later stage (Figure 1.3). Residents of East and Central regions however seem to have lesser opportunity on average, in both consumption and education[27].

Opportunity tree for wage earning is however limited to the casual or regular wage earners who are necessarily non-self employed and the wage tree structure may well be very different for India with the inclusion of the self-employed workers. Nevertheless as far as regular/casual earning opportunity is concerned, the circumstance of region divides the country in two parts. Although father's occupation is selected as the next important circumstance after region for the whole country, the historically destitute caste categories of *SC/ST* almost always have a better earning opportunity than the relatively upper castes in the North-Eastern region, irrespective of their father's occupational background. For the rest of the country however, earning opportunity is always better for the forward *General* caste individuals, as the average wage of them is always higher than that of the lower castes. This seemingly counter-intuitive caste dynamics is rather a region specific feature of the North-Eastern part of the country, that embodies not more than 5% of the national population, but is often called the tribal hub of India for having a much higher concentration of the marginalized lower castes of *SC/ST*s (particularly *ST*s).

[27] On a separate note, East and Central India are also found as two of the worst performing regions in terms of educational opportunity for children as well (*see* 3).

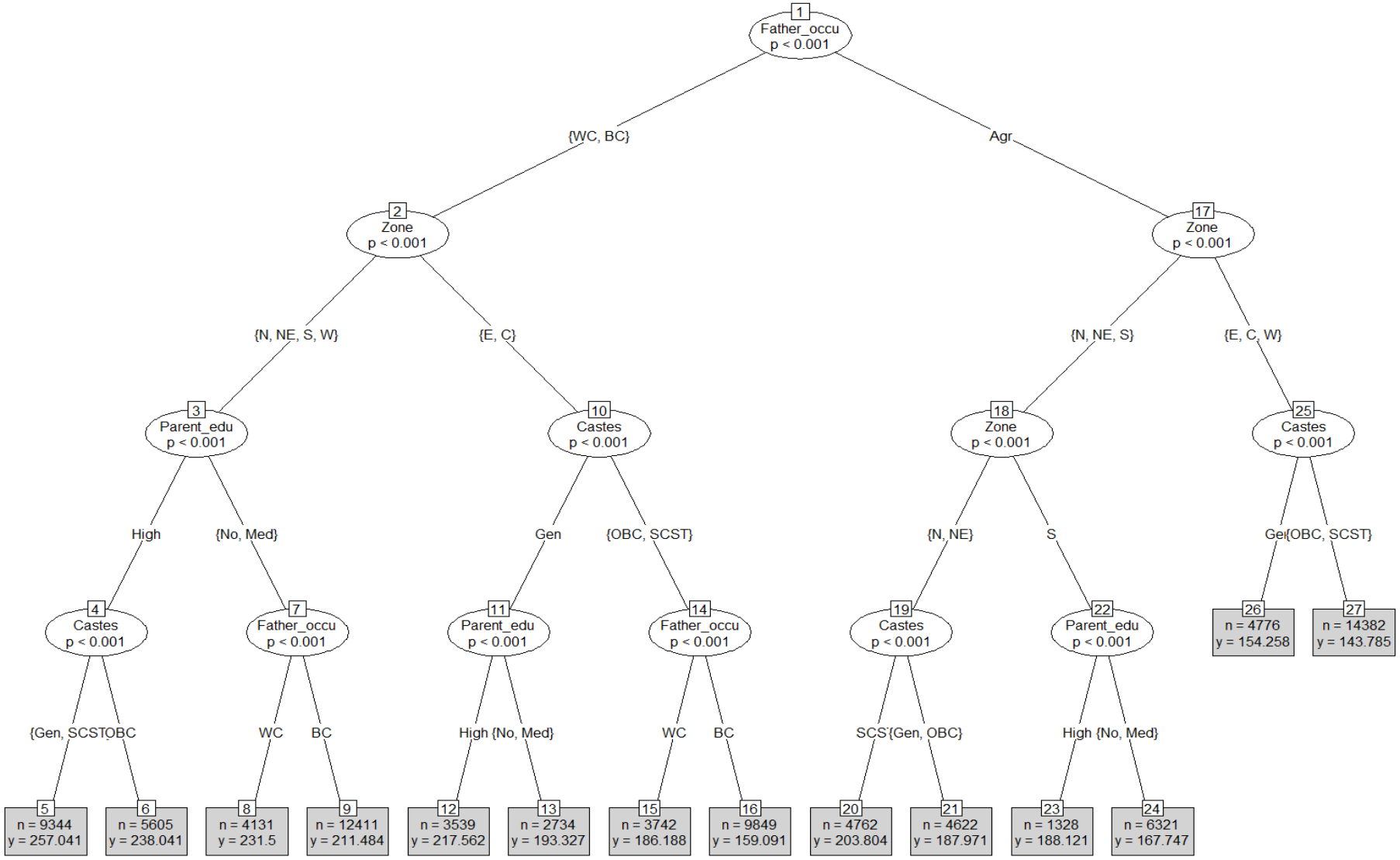



Figure 1.1: MPCE (*2011-12*)[a]

[a]'n' and 'y' denote the sample size and the mean MPCE in INR (Indian Rupee), respectively, for the corresponding terminal node. Parent_edu, Father_occu and zone represent the circumstances of parental education, father's occupation and region of residence, respectively.

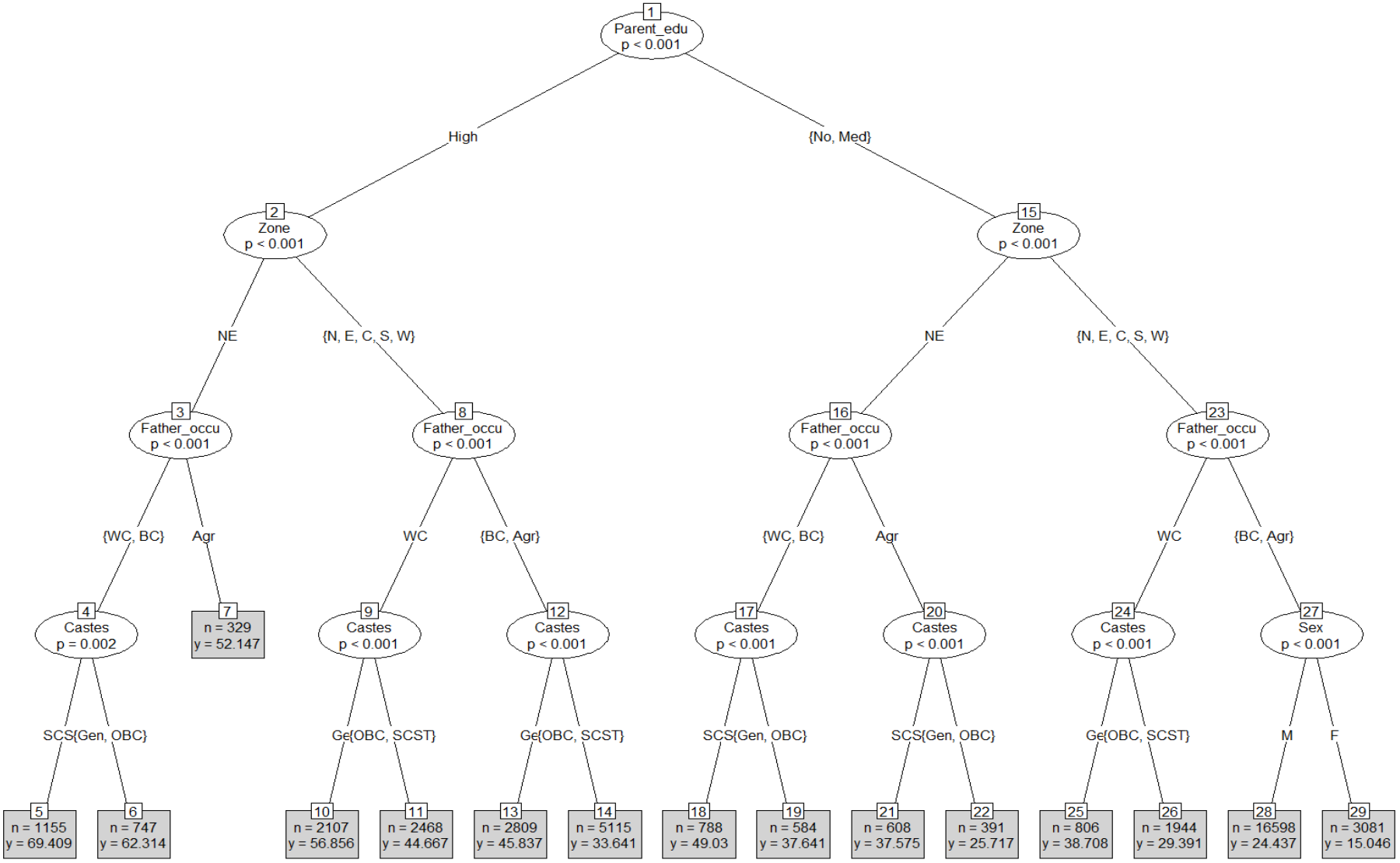



Figure 1.2: Wage (*2011-12*)[a]

[a]'n' and 'y' denote the sample size and the mean (daily) wage in INR (Indian Rupee), respectively, for the corresponding terminal node. Parent_edu, Father_occu and zone represent the circumstances of parental education, father's occupation and region of residence, respectively.

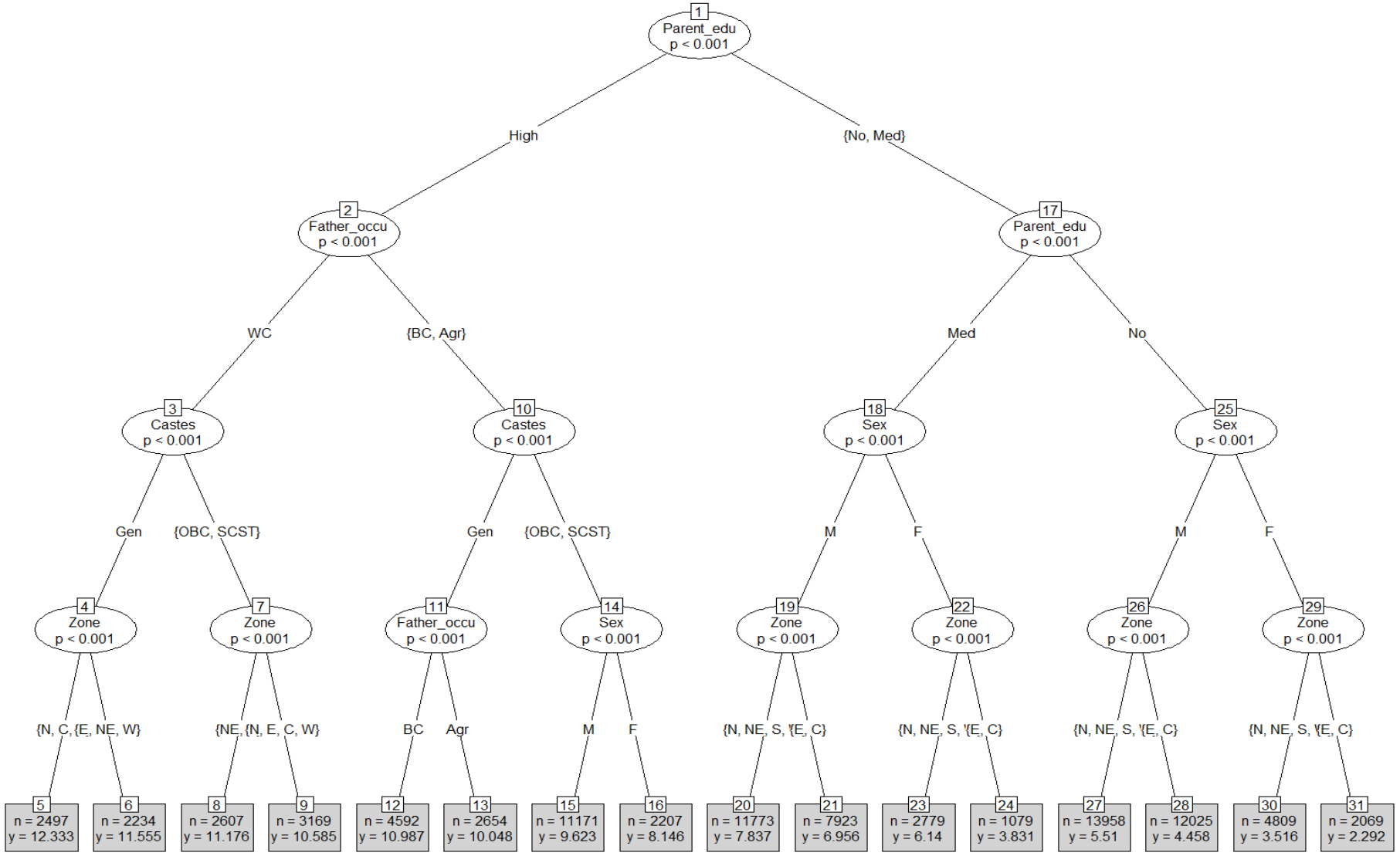



Figure 1.3: Education (*2011-12*)[a]

[a]'n' and 'y' denote the sample size and the mean year of education, respectively, for the corresponding terminal node. Parent_edu, Father_occu and zone represent the circumstances of parental education, father's occupation and region of residence, respectively.

## *1.5 Concluding remarks*

In this paper we estimate the amount of *IOP* for India in consumption expenditure, wage earning and education, using the last two survey years of the *NSS* data base. In our assessment of *IOP* in the Indian society, we consider a set of five circumstance factors comprising of caste, sex, region, parental education and father's occupation. Using the most widely used methodologies in estimating *IOP*, we found that 27-32% of educational inequality in due to unequal opportunity. Whereas during the time frame of *2004-12* earning opportunity in India is around 32-39%, which is higher than some of the most opportunity unequal countries in Latin America. However due to the selective reporting of wage data in *NSS*, our wage analysis is limited to the non-self-employed regular or casual workers of the country and excludes a substantial portion of self-employed working adults. On the other hand, both of the non-parametric and parametric methods estimate that the share of unfair inequality in consumption (MPCE) is around 11%. But consumption for being reported as the total monthly household consumption expenditure, may not be well responsive to changes in the individual circumstances and thereby has a chance to be underestimated.

Due to the structure of *NSS* information on parental attributes is provided for 'co-resident' households only where the offspring are enumerated along with their parents. The other social circumstances like caste, sex and region on the other hand, are non-missing for the entire sample. However we found the degree of *IOP* to be substantially underestimated if parental backgrounds are omitted from the set of circumstances, whereas this is not the case when caste is omitted. In fact *IOP* in India is estimated even lower than some of the developed countries while taking the social circumstances alone (caste, sex, region). In addition we also found that in spite of numerous evidence on caste discrimination in the Indian society, taking caste as the only circumstance factor is not enough as far as quantifying *IOP* is concerned. The hierarchical division of caste is therefore not able to capture well the differences in other omitted circumstances, especially that of parental backgrounds.

Similar to the extant literature, both of our non-parametric and parametric measures of *IOP* are based on all possible interactions of the five taken circumstances, while in

reality some of them may be more relevant than the others. To explore the intertwining of our circumstances we further provide the opportunity structure for India using the recently introduced approach of the regression tree analysis. We found parental education to be the most important circumstance for wage and education, whereas it is the occupational category of father that seems the most important source of unequal opportunity in consumption. Irrespective of the outcomes, individuals from agricultural family backgrounds are always worse off. In addition, the opportunity structure of India reveals the interesting way caste and sex are connected to parental education. Sex seems rather relevant when parents have lesser experience of formal schooling. Especially for earning and educational opportunity, females of poorly educated parents have significantly less opportunity than males. The forward caste premium on the other hand, is also prominent for higher educated families as well. The opportunity tree also brings forth the special case of the tribal part of India, the North-Eastern region, where the most historically disadvantageous caste categories ($SC/ST$) are actually found to be better off than the upper castes, at least in terms of regular/casual wage earning, which is never the case for the rest of the country.

# Appendices to Chapter 1

## *1.A Multiple imputation*

### 1.A.1 The algorithm of multiple imputation of chained equation

To impute parental education and father's occupation, we adopt a multivariate imputation approach, in particular, the sequential regression multiple imputation algorithm of Raghunathan et al. (2001). Two other popularly used method for multivariate imputation are multivariate normal imputation and sequential univariate monotone imputation. We could not use the former, as that is applicable in case of continuous imputed variables with multivariate normal distribution. The latter is a rather quick method, but is only applicable if the pattern of the missing data is monotone, which means, if missing values for one variable is completely nested in that of the other. For either rounds, father's occupation is missing for about 20% of the co-resident data points, because they are recorded against currently employed fathers. Whereas, parental education is almost non-missing or have very few missing values for all years. But, it is only one round, the missing pattern among these two variables are monotone. Hence we have to use the iterative multivariate imputation process.

The multivariate imputation algorithm draws the imputed values through a series of univariate regressions, or equivalently, through a series of chained equations and hence, is also called the multiple imputation of chained equations (*MICE*). The underlying imputation model specification takes all the variables as predictors except the one to be imputed. First, the variables to be imputed are ordered from the least to the highest (in terms of missing values) and then start imputing the variable for which missing information is minimum, using predictors without any missing value. The next ordered variable (with second least number of missing values) is then imputed using the non-missing predictors, as well as the imputed value of the first variable. The process continues till

the variable with highest number of missing value is imputed. Further, each imputation consists of multiple cycles or iterations to get more stable set of imputed values, based on which, the final vector of imputed values are drawn for the entire working sample. The algorithm is detailed in Raghunathan et al. (2001)[28]. For two imputed variables, the regression sequence is described as below.

Let $X_1$ and $X_2$ be the variables to be imputed with the fully specified vector of variables denoted by $Z$ and let $X_1$ be the variable with the least number of missing values (which in our case, is parental education for all rounds). In the first cycle, $X_1$ is regressed on $Z$ (*i.e.* $X_1 \rightarrow Z$) and the missing values in $X_1$ are imputed by simulated draws from the posterior distribution of $X_1$. Then $X_2$ is regressed on $Z$ along with the imputed values of $X_1$, (*i.e.* $X_2 \rightarrow X_1^m, Z$) and imputed values of $X_2$ are drawn similarly. In the cycles thereafter, each of $X_1$ and $X_2$ are regressed on the fully specified variables along with the previously imputed variables. Thus, in the second cycle, the prediction sequence is $(X_1 \rightarrow X_2^m, Z)$, $(X_2 \rightarrow X_1^m, Z)$ and so on. The cycles are continued (often upto 10 to 20 iterations) to converge to a set a stable imputed values $\{X_1^1, X_2^1\}$, that constitutes the first imputed data set. The entire process with the same number of iterations are then repeated $M$ times, to produce $M$ copies of the imputed data sets, with imputed variables $\{(X_1^1, X_2^1), \ldots, (X_1^M, X_2^M)\}$. The non-parametric and parametric measures of *IOP* are then estimated for each of these $M$ imputed data sets and the final estimate of *IOP* is then estimated as the average of all the imputed data sets [Rubin's rule (Rubin 1986)].

Notice that, after the first cycle, all the missing values are imputed. If the missing pattern is monotone, that is, if $X_2$ is missing only if $X_1$ is missing, there is no need of further iteration. Only cycle one is repeated $M$ times to produce multiple copies of the imputed data set. In that case the prediction sequence is like - $(X_1 \rightarrow Z)$; $(X_2 \rightarrow X_1^m, Z)$. Since $X_2$ is only missing when $X_1$ is missing, this sequence is enough to draw sensible imputed values for both the variables (Raghunathan et al. 2001). When missing pattern is arbitrary, iteration is needed so as to get a stable set of imputed values, that is repeatedly predicted by old and newly imputed values. Accordingly, we performed no iterations for year *2011-12*, for which missing pattern is detected as monotone. For the other round, we choose 20 iterations for each imputation.

[28] Also see Royston et al. (2011), Azur et al. (2011).

## 1.A.2 Imputation model and diagnostics

The variables to be imputed in our case, are - parental education and father's occupation, where the former is generated by combining father's and mother's education[29]. To reduce imputational rigor, we consider to impute the combined parental education, instead of imputing each of the father's and mother's education (much in the spirit of 'transform then impute' (Von Hippel 2009)). We estimate an ordered logistic regression as our imputation model, to estimate parental background with a broad range of covariates, that are not missing for the entire work sample. Following the literature (Rubin 1986, Little 1988, Schafer 1999), we include three broad set of covariates - (*i*) the *analysis model variables* (caste, sex, zone along with their all possible interactions), (*ii*) the *auxiliary variables* (household size, consumption expenditure, sector, religion, along with children age, age squared, education, occupation, sex, marital status, relation to head) and (*iii*) the *survey specific variables* (sub round, second stage stratum, first stage units[30]). Following Teyssier (2017), who have used MI for the same purpose of imputing parental background for Brazil, we include the sample weight as a predictor as well (along with the normal use of sample weights in the logit model). In addition, children wage and its interaction with age is also considered for the wage sample imputation. The imputation model does not have any claim of causality, but it should fit the data well. With highly significant model chi-square statistics for all rounds, Table 1.A.1 does not indicate that our chosen imputation model is a poor fit for any of the imputed variables.

| Year | Likelihood Ratio Chi-square | | | | Pseudo $R^2$ | |
|---|---|---|---|---|---|---|
| | Parental education | [p-value] | Father's occupation | [p-value] | Parental education | Father's occupation |
| ***Work sample*** | | | | | | |
| 2004-05 | 6439.4 | [0.000] | 6608.6 | [0.000] | 0.184 | 0.396 |
| 2011-12 | 2978.2 | [0.000] | 4118.6 | [0.000] | 0.181 | 0.418 |
| ***Wage sample*** | | | | | | |
| 2004-05 | 2646.0 | [0.000] | 2281.2 | [0.000] | 0.221 | 0.391 |
| 2011-12 | 1632.9 | [0.000] | 1779.4 | [0.000] | 0.215 | 0.388 |

Table 1.A.1: Imputation model check[a]

[a]We report McFadden $R^2$ in particular.

[29]In case of single-parent household, that constitute about 8% of the co-resident sample, parental education is the education of the single parent.

[30]Definition of these variables are provided in the *NSSO* data appendix A.

Across the rounds, 66-71% of our working sample have missing information on parental background that we needed to impute. Multiple imputation is a simulation based algorithm and hence, the power and precision of the multiply imputed values are likely to increase with the number of imputations, especially when missing data proportion is large. So far in the literature, there is no unequivocal rule to choose an *optimum* number of imputations. However, even with a high fraction of missing information, a number of literature often recommends that a modest number of imputation is good enough to generate statistically sound imputed values (Rubin 1986, Schafer 1999)[31]. As shown by Rubin (1986), the relative efficiency of an infinite number of imputations subject to a finite one, is $(1 + \gamma/m)^{-1/2}$, where $\gamma$ and $m$ are the fraction of missing information and the number of imputations, respectively[32]. In case of 70% missing information ($\gamma = 0.7$), the relative large sample efficiency is already 0.96 with 10 imputations, that increases to 0.98 for 20 imputations. Since in case of large degrees of freedom, each additional imputation adds little to the efficiency of the estimated parameter (Schafer & Olsen 1998), we choose to do 20 imputations for each of our imputed variables (for imputing around 68% missing data for a sample size of 0.1 million, on average). Further, in case of arbitrary missing pattern, each imputation is generated from a simulated draw of 20 iterations.

However, "a naive imputation is worse than doing nothing" (Little 1988, p 288). We have a total of 20 imputed data-set. For a randomly chosen imputation, Table 1.A.2 reports the distribution of the imputed variables in the observed data-set ('response'), the imputed data-set ('non-response') and the completed data-set ('response'+'non-response'), for both of our final working sample and the wage sub-sample. At a glance, father's occupation seem to have been imputed better, for it has similar distribution across all the data-sets. Whereas, more parents are pointed as having no formal education for the imputed data-set. But that does not mean a faulty imputation of parental education, and in fact, the difference in its distribution is indicative of a rather sensible imputation. The non-co-resident sample, who are, on average, 10 years older than the co-resident ones,

[31]Besides, in case of a complex imputation model with large number of variables and sample size, even a single imputation takes hours to complete, and so more, if it is iterative. The computational effort associated with the higher number of imputations in these cases, are often too prohibitively high to make little sense to increase the number of imputations for a marginal increase in efficiency (Allison 2003, Von Hippel 2005, Azur et al. 2011).

[32]Missing information, strictly speaking, is not the same as the number of missing data points. With high correlation between the missing variables and the observed covariates, $\gamma$ is actually lesser than the percentage of missing values (Graham et al. 2007). However, they are the same in the simplest setting.

are supposed to have older parents. Provided the substantial educational improvement over time for all generations, as is reflected by Table 1.B.1 and 1.B.2, older parents are more likely to be deprived of formal education, exactly as they are imputed. On the other hand, Table 1.B.1 also shows that occupational composition of the samples does not seem to be markedly different due to co-incidence. Provided low occupation mobility in India, this is likely to be true for parents as well[33]. Besides, as a robustness check, we found that the pattern of the distributions of the imputed values are similar for many other imputed data sets as well.

| | *2004-05* | | | *2011-12* | | |
|---|---|---|---|---|---|---|
| | obs. | imp. | comp. | obs. | imp. | comp. |
| ***Work sample imputation diagnostics*** | | | | | | |
| **Parental education** | | | | | | |
| No_schooling | 0.390 | 0.524 | 0.478 | 0.305 | 0.379 | 0.354 |
| Below primary | 0.286 | 0.246 | 0.260 | 0.280 | 0.263 | 0.269 |
| Above primary | 0.324 | 0.230 | 0.262 | 0.415 | 0.358 | 0.378 |
| **Father's occupation** | | | | | | |
| White collar | 0.119 | 0.105 | 0.109 | 0.198 | 0.194 | 0.195 |
| Blue collar | 0.322 | 0.359 | 0.348 | 0.353 | 0.408 | 0.393 |
| Agricultural | 0.559 | 0.536 | 0.542 | 0.449 | 0.398 | 0.412 |
| ***Wage sample imputation diagnostics*** | | | | | | |
| **Parental education** | | | | | | |
| No_schooling | 0.412 | 0.495 | 0.471 | 0.326 | 0.379 | 0.363 |
| Below primary | 0.278 | 0.238 | 0.250 | 0.270 | 0.247 | 0.254 |
| Above primary | 0.311 | 0.266 | 0.279 | 0.404 | 0.374 | 0.383 |
| **Father's occupation** | | | | | | |
| White collar | 0.115 | 0.097 | 0.101 | 0.181 | 0.124 | 0.138 |
| Blue collar | 0.405 | 0.443 | 0.435 | 0.473 | 0.489 | 0.485 |
| Agricultural | 0.480 | 0.460 | 0.464 | 0.346 | 0.387 | 0.377 |

Table 1.A.2: Imputation diagnostics[a]

[a]Where 'obs.', 'imp' and 'comp.' stand for *observed*, *imputed* and *completed* data set, respectively. For reporting the imputed and the completed data set, we choose one imputation at random (among 20 imputations).

[33]Also note from Table 1.B.2, that in *2011-12*, 56% of co-resident sample have their fathers working in agricultural sector, while 45% of them are in agricultural job themselves (Table 1.B.1).

## *1.B Additional tables and figures*

| | age | hhsize | %male | %rural | %SC/ST | %married | %noschool | %agri | %wage | N |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Working sample*** (total) | | | | | | | | | | |
| *2004-05* | 32.1 | 5.5 | 0.77 | 0.76 | 0.29 | 0.82 | 0.36 | 0.53 | 0.41 | 127002 |
| *2011-12* | 32.8 | 5.0 | 0.82 | 0.72 | 0.30 | 0.82 | 0.24 | 0.45 | 0.48 | 90574 |
| ***Non-response part*** (non-co-resident) | | | | | | | | | | |
| *2004-05* | 35.1 | 4.7 | 0.71 | 0.76 | 0.30 | 0.96 | 0.45 | 0.54 | 0.43 | 83201 |
| *2011-12* | 35.5 | 4.4 | 0.77 | 0.71 | 0.31 | 0.96 | 0.31 | 0.46 | 0.49 | 59592 |
| ***Response part*** (co-resident) | | | | | | | | | | |
| *2004-05* | 25.7 | 7.0 | 0.92 | 0.76 | 0.26 | 0.50 | 0.18 | 0.49 | 0.37 | 43801 |
| *2011-12* | 26.5 | 6.5 | 0.93 | 0.72 | 0.26 | 0.50 | 0.10 | 0.42 | 0.45 | 30982 |

Table 1.B.1: Summary statistics: working sample, response part and non-response part[a]

[a]Response part correspond to the co-resident sample for which parental information is provided in the data-set, whereas the non-response part are the non-co-resident samples for which parental backgrounds are needed to be imputed. Working sample is the union of the response and the non-response part. 'age' and 'hhsize' reports the mean age and household size of the respective sample. %male, %rural, %SC/ST, %married, %noschool, %agri and %wage reports the share of males, rural inhabitants, SC/STs, married individuals, samples without any formal schooling, samples engaged in agricultural jobs and samples who further have the information on wage data, respectively. The last column (N) reports the respective sample size.

| | age *father* | age *mother* | %noschool *father* | %noschool *mother* | %noschool *both* | edu_year *child* | edu_year *father* | edu_year *mother* | %dom_duty *mother* | %agri *father* |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Co-resident parents*** | | | | | | | | | | |
| *2004-05* | 54.0 | 48.8 | 0.47 | 0.75 | 0.45 | 6.6 | 3.8 | 2.1 | 0.60 | 0.63 |
| [61] | (0.07) | (0.06) | (0.00) | (0.00) | (0.00) | (0.03) | (0.03) | (0.02) | (0.00) | (0.00) |
| *2011-12* | 54.5 | 49.5 | 0.42 | 0.68 | 0.40 | 7.7 | 4.2 | 2.5 | 0.72 | 0.56 |
| [68] | (0.09) | (0.09) | (0.01) | (0.01) | (0.01) | (0.05) | (0.05) | (0.03) | (0.01) | (0.01) |

Table 1.B.2: Co-resident sample summary of parents[a]

[a]Standard errors are in parentheses and rounds in squared brackets. In particular, 'noschool father/mother' indicates fathers/mothers who are deprived of any formal schooling, whereas 'noschool both' means none of the parents have any formal schooling. 'edu_yr' abbreviates as the year of education. '%dom_duty mother' denotes the share of mothers who have reported not to be in the labor market for attending domestic duties and '%agri_father' are the share of fathers engaged in agriculture related jobs.

| | MPCE | | Wage | | Education | |
|---|---|---|---|---|---|---|
| | 2004-05 | 2011-12 | 2004-05 | 2011-12 | 2004-05 | 2011-12 |
| *Ref: General* | | | | | | |
| *OBC* | -0.157*** | -0.058*** | -0.037*** | -0.136*** | -0.082*** | -0.192*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *SC/ST* | -0.288*** | -0.063*** | -0.074*** | -0.146*** | -0.189*** | -0.278*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *Ref: Primary plus* | | | | | | |
| *Primary or below* | -0.066*** | -0.067*** | -0.345*** | -0.274*** | -0.437*** | -0.385*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *No schooling* | -0.109*** | -0.102*** | -0.498*** | -0.409*** | -1.029*** | -0.933*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *Ref: White collar* | | | | | | |
| *Blue collar* | -0.068*** | -0.075*** | 0.066** | -0.094*** | -0.129*** | -0.105*** |
| | (0.00) | (0.00) | (0.02) | (0.00) | (0.00) | (0.00) |
| *Agricultural* | -0.008 | -0.226*** | -0.001 | -0.303*** | -0.291*** | -0.208*** |
| | (0.53) | (0.00) | (0.29) | (0.00) | (0.00) | (0.00) |
| *Ref: North* | | | | | | |
| *East* | -0.293*** | -0.359*** | -0.338*** | -0.229*** | -0.283*** | -0.247*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *Central* | -0.214*** | -0.403*** | -0.313*** | -0.267*** | -0.188*** | -0.205*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *North-East* | -0.082*** | -0.179*** | -0.031 | -0.022 | 0.050* | 0.027 |
| | (0.00) | (0.00) | (0.20) | (0.39) | (0.03) | (0.37) |
| *South* | -0.088*** | -0.196*** | -0.197*** | -0.055** | -0.069*** | -0.009 |
| | (0.00) | (0.00) | (0.00) | (0.02) | (0.00) | (0.62) |
| *West* | -0.161*** | -0.281*** | -0.330*** | -0.299*** | 0.032 | 0.015 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.10) | (0.58) |
| *Ref: Male* | | | | | | |
| *Female* | 0.039*** | 0.010 | -0.403*** | -0.316*** | -0.576*** | -0.429*** |
| | (0.00) | (0.38) | (0.00) | (0.00) | (0.00) | (0.40) |
| Intercept | 5.36*** | 5.38*** | 3.48*** | 3.90*** | 2.36*** | 2.45*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |

Table 1.B.3: Reduced form OLS: for MPCE, Wage and Education[a]

[a]Standard errors are in parenthesis. (***, **, *) correspond to 1%, 5% and 10% level of significance, respectively.

# Chapter 2

# Equalization of Opportunity across castes in India: A long-term analysis over 1983-2012

## *2.1 Introduction*

Since the late twentieth century, it is inequality of opportunity, that had become the more relevant 'currency of justice', rather than the outcome inequality in totality (Rawls 1971, Dworkin 1981*b*, Arneson 1989, Cohen 1989). The main shift in focus lies in drawing a line between fair and unfair inequality, where the latter is conceptualized as the *inequality of oppportunity*, generating exclusively from factors on which no individual has any control (Roemer 1993). In the doctrine of inequality of opportunity, *circumstances*, are defined as the inequality generating factors, that are beyond alteration by means of any subjective control, like race, sex, caste, ethnicity, birthplace or parental backgrounds. On the other hand, there are the *effort* factors, that the individual can presumably control and therefore can be considered as the legitimate source of inequality. Accordingly, from an ethical standpoint, social welfare policies, instead of targeting equal outcome for everyone in the society, should rather prioritize on *equalizing opportunities* across people from varying circumstances.

Most of the societies are generally auto-divided based on the predestined circumstances, generating a highly complex hierarchical social structure. In this heterogeneous social fabric, who belongs to which stratum is completely out of individual responsibility and is rather determined by fate. However, it is often seen that people belonging to certain

disadvantageous circumstances are always subject to doomed future, which is particularly true for India, a country that embodies a highly stratified society. People from poor, uneducated, lower caste families in India are often seen to overcrowd the bottom layer of the society with lesser economic advantage. However, social welfare in true terms should be inclusive, that excludes no one to get the fruits of development by virtue of providing equal opportunity to every marginalized people. The present work aspires to explore if the historically disadvantageous lower caste categories in India, have less consumption and wage earning opportunity than the upper caste groups. In addition, by covering a long time frame we further explore whether the existing unequal opportunities among the caste categories equalizes over time.

Inequality of opportunity ($IOP$) by definition, is the ethically objectionable part of the total outcome inequality that is generated by the circumstance factors only. Therefore the main methodological challenge to quantify $IOP$ is to isolate the unfair part of inequality, that exclusively exhibits inequality arising from differences in the chosen circumstances. Majority of the literature accomplish this by generating a suitable counterfactual outcome distribution, that by construction, smooth out any differences within the circumstances and differs only due to the differences between the circumstances. Most often this is executed by representing each circumstance groups by their mean outcome. Inequalities between the counterfactual distributions therefore quantify the amount of absolute $IOP$ in the society. Depending on the statistical model, the counterfactual distributions eventually generates the non-parametric and the parametric index of $IOP$. Bourguignon, Ferreira & Menéndez (2007), Checchi & Peragine (2010), Ferreira & Gignoux (2011), Marrero & Rodrìguez (2011), Björklund, Jäntti & Roemer (2012), are some of the notable works that provides the measures of $IOP$ for a number of developed and developing countries, either by the non-parametric or by the parametric index[1].

However, one drawback of both the non-parametric and parametric index approach is that individuals within each circumstance are represented by their mean outcome. This representation therefore implicitly assumes that individuals within a circumstance type are risk-neutral, which is rarely the case in reality. Lefranc, Pistolesi & Trannoy

[1] Also *see* Roemer & Trannoy (2013), Ramos & Van de Gaer (2012) for a more extensive overview and applications of the index approach in IOP. See Chapter 1 as well.

(2009) made a first attempt to show the existence of *IOP* without the risk-neutrality assumption, by comparing the entire distributions, conditional on different circumstances. Instead of resorting to a particular scalar index, they tested for *IOP* among the different circumstances, using the tools of stochastic dominance. However, this approach, what we call the *distributional dominance* approach, is only capable of testing the existence of *IOP* in a society, but is unable to compare two opportunity unequal societies.

To overcome this issue, Andreoli, Havnes & Lefranc (2019) introduced the concept of *equalization of opportunity*, that is able to rank different societies in terms of their existing *IOP* without assuming risk-neutrality within the circumstance types. Lefranc et al. (2009) concludes in favor of equal opportunity in the society if the circumstance specific distributions coincide with each other. However any gap in the aforementioned distributions is indicative of *IOP* in the society and therefore for the same set of circumstances, a reduction in this gap for another society should indicate lesser *IOP* in that society. Exploiting this concept, Andreoli et al. (2019) advances the distributional dominance approach by comparing the gap between the circumstance specific distributions under different social states and concludes in favor of *equalization of opportunity* upon finding a reduction in this gap. Therefore without any scalar index of *IOP* and eventually without masking the heterogeneous preferences towards risk within the circumstance types, Andreoli et al. (2019) provides a robust method to test for equalization of opportunities for a set of circumstances over different social states. This is the methodological set up adopted for the present work, to explore whether opportunity among the different hierarchical caste groups in India equalizes over the time period of *1983-2012*.

India is one of the very few countries, where the century old caste system is well embedded even to date. The origin of the caste system was found in the ancient Hindu text, where the society was divided in hierarchical occupational structure. Upper castes are supposed to be engaged in occupations, that are more pure in nature, like worshiping deities or serving the country as soldiers or traders. Whereas, the major occupation of the lower caste categories, is to serve the upper caste 'masters'. Although having its root in the occupational division, caste in its way became hereditary and is identified at birth. In Indian society, children inherit the caste of his/her father, that is not convertible for lifetime. That makes caste, a classic circumstance factor, in the context of *IOP*. Majority

of the historically disadvantageous caste categories in India, are still living under acute poverty with under-paid menial jobs, even today (Gang et al. 2017). However, caste is not the only source of hierarchy in the Indian society. Other circumstances, especially that of parental backgrounds, are repeatedly found to be one of the major source of *IOP* in several developed and developing countries. Even in the context of India, Chapter 1 finds parental education as one of the important source of unequal opportunity in consumption or wage. However for the present analysis we nevertheless choose caste as our only circumstance factor for the following reasons.

First of all, information on parental background is subject to data availability. We aspire to use the biggest micro-level data base for India, the National Sample Survey, which unfortunately does not have any direct provision of parental backgrounds. Instead parental attributes in this survey are only available for the co-resident households, where parents are enumerated along with at least one of their offspring. Therefore to incorporate parental backgrounds as our circumstances we have to limit the study of equalization of opportunity to the co-resident households, which we did not opt for, as in that case the analysis may suffer from selectivity issues. Secondly, even with the same data base, another option is to impute the information on parental backgrounds from the co-resident data points using the statistical method of multiple imputation. In fact following this approach Chapter 1 shows that the effect of caste is quite visible for generating unequal opportunity in consumption, wage and education, even after controlling for parental education and occupation. However, while the use of multiply imputed circumstances is suitable to the index approach it is not applicable to the distributional dominance approach, where opportunity equalization is analyzed on the basis of the entire circumstance specific distributions. So we prefer to analyze opportunity equalization based on a circumstance that is provided for all survey respondents. Third, we can expand our set of circumstances by including other factors, that, unlike the parental backgrounds, are available for the entire data set. Asadullah & Yalonetzky (2012) for example, analyzes educational opportunity in India using the same data base by considering sex and religion as their major circumstances as well. However, as casteism is majorly a Hindu phenomena, we did not use religion as another circumstance variable along with caste. Further, as we want to test opportunity equalization among the working adults, share of

working women are naturally under-represented in our sample due to low female labor force participation in India.

Provided the ample evidence of caste discrimination in India, we therefore choose to evaluate whether unequal opportunities among the different caste groups equalizes over time. In particular, we consider two outcome variables for our analysis, that of consumption expenditure and wage earning. Given the limited number of works on *IOP* in India, there is considerable scope for further work. The present work contributes in the literature in several ways. First, to our knowledge, this is the first study on India that analyzes equalization of opportunity between castes following the robust method of distributional dominance approach. Second, the present analysis covers a long time span of nearly three decades, from *1983* to *2012*. One of the major change in the policy regime over this time frame is, that the Government of India switched from a centrally interventionist policy regime to an open-market neo-liberal one, in the early nineties. Since this major economic reform, numerous studies have shown evidence of increasing consumption and income inequality in India. The present analysis will enrich this debate from the perspective of responsibility sensitive analysis of inequality by exploring the impact of the neo-liberal policy reform in equalizing opportunities among the caste groups. We found that the historically disadvantageous caste groups in India are always worse off as compared to the upper caste categories and in terms of earning opportunity, the forward caste premium enjoyed by the upper castes actually increases over time, especially since the economic reform. Consumption opportunity on the other hand reveals considerable equalization among the different caste categories, especially in the later phase of reform.

The remaining of the paper is organized as follows. A brief introduction to caste system in India is provided in section 2.2, which is followed by the theoretical background of the robust method of equalization of opportunity in section 2.3. Section 2.4 then provides details on our data base with the discussion on our main variables and sample selection criteria. Results on equalization of opportunities among the different caste groups are then discussed in section 2.5, first for the entire time frame of *1983-2012* and then for the selected time period with finer caste categorization. Finally, section 2.6 concludes.

## *2.2 Casteism in India*

The root of the caste system is found in the *Varnashrama dharma* of the ancient Hindu text, where the society was divided in four '*Varnas*' according to occupational hierarchy. In the top there are the *Brahmins* or priests, who are thought to form the most pristine layer of the society and are mostly engaged in teaching or worshiping deities. They are followed by *Kshatriyas* or soldiers, who in turn are followed by *Vaishyas* or traders. Among the four *Varnas*, the bottom layer was constituted of *Shudras* or servants, whose primary job is to serve the other three 'superior' *Varnas*. Outside these four *Varnas*, a destitute fifth category is often formed as the *Ati-shudras* or 'untouchables', who are considered as the most 'polluted' layer of the society for being engaged with 'impure' jobs like burning corps or manual scavenging and were banned from sharing any public property. The origin of untouchability is nevertheless debated and the practice of it is legally banned by the Untouchability Act of 1955. However the so called 'untouchables', who are also referred now as 'Dalits', still face considerable discrimination even in modern India.

All of these broad *Varnas* are further divided into thousands of sub-castes or *jatis* with intra-caste hierarchy, which are regrouped by the constitution of India into several caste categories for designating reservation status[2]. In 1950, the constitution of independent India lists 1108 castes as the *Scheduled Castes* (SC) and 744 tribal castes as the *Scheduled Tribes* (ST). Although the practice of untouchability is visible to both *SC* and *ST*, the former caste category is often referred as 'Dalits' (meaning oppressed or broken) and the latter as 'Adivasis' (meaning tribes). Together they constitute nearly 30% of the Indian population and are entitled to reservation in political assembly, education or public sector jobs since 1950, a few years after the country got independence from the British empire. In spite of that, more than one-fourth of the *SC* and *ST* households still live below the poverty line, which is considerably higher than the non-SC/ST households. Later around mid-eighties, more than two-thousand castes among the *non-SC/ST*s are further enlisted as the *Other Backward Classes* (OBC). *OBC*s are the relatively socially and economically backward class of the country other than the *SC/ST*s. In present India, the

[2]Although caste was historically originated as the four broad occupational *Varnas*, the term sub-castes and castes are often used equivalently.

caste categories consist of *SC*, *ST*, *OBC* and *General*, where the last category of 'General' comprises of all Indians who does not belong to any of the other three categories and are excluded from any caste based reservation policies for being considered as the most advantageous group of castes[3].

In proportion to their national population share, the *SC*s and *ST*s together are entitled to about 24% of reservation in several public sector jobs or higher educational institute. However, while the list castes to be included in the category of *SC/ST* are fixed by the constitution, it is not the same for the *OBC*s. Instead the eligible castes for the status of *OBC* are rather identified at the state-level, based on a host of nationally accepted criteria[4]. Since concentration of backward castes may vary with different states, it is possible for the reservation quota for *OBC* to vary across the states as well. Nevertheless, based on the report of the second backward class commission (the so called Mandal commission) the Supreme court of India set on average, a reservation benchmark of 27% for the *OBC*s in the early nineties. The basis of the criteria for the *OBC* status are often debated on the ground of lopsided political lobbying that eventually incites several riots in India since the announcements of reservations for the *OBC*s (Jaffrelot 2006, Gang et al. 2011). As a result, the list of *OBC*s are constantly updated and have a chance to vary over time as well, upon the legal inclusion of some hitherto deprived castes as *OBC*s.

In spite of taking several affirmative policies by the constitution of India since 1950, a substantial body of literature provides sufficient evidence of caste discrimination even in the modern Indian society. Lower castes are repeatedly found to be discriminated under several categories, like denial of employment, unusually long hour job, lower wage, under-representation in Government job and higher education institutes (Madheswaran & Attewell 2007, Thorat 2008, Deshpande & Ramachandran 2014). Other than direct discrimination, there is also evidence on how the caste system locked itself from getting broader opportunities. Munshi & Rosenzweig (2009) found the caste based mutual insur-

[3]For further account of the Indian caste system and its various social implications, *see* Dyson & Moore (1983), Ambedkar (2014), Roy (2017).

[4]Complete listing of *SC/ST* castes are available in article 341 and 342 of the Indian constitution. On the other hand, *OBC*s were first estimated based on certain criteria of economic backwardness provided by the second backward class commission, popularly known as the Mandal commission, that is overall approved by the constitution.

ance network to be responsible for the increasing rural-urban wage differential and low social mobility in rural India. Also from the perspective of inequality of opportunity the *SC/ST*s are often found to have much less educational or earning opportunity than the relatively advantageous *non-SC/ST*s, even after controlling for other discriminatory factors like sex, regional habitat or parental backgrounds (Asadullah & Yalonetzky (2012), Singh (2012*b*), Chaper 1). There are some evidence of improvements of the deprived caste groups of *SC/ST* in terms of lesser poverty, better education and higher occupation mobility over time (Papola 2012, Hnatkovska et al. 2012), a large amount of social gap with the non-SC/STs still persists even in the twenty-first century (Deshpande 2001, Kijima 2006).

## *2.3 Theoretical framework*

### 2.3.1 The compensation principle of equality of opportunity

Let $y_\pi \in \mathcal{R}$ denote an individual *outcome*, under a *social state*, $\pi$. The *outcome* can be thought of as any desirable economic advantage like income, consumption, standard of living, educational attainment, life expectancy, health index, job market access or any such thing for which the 'more is better' principle is applicable. Whereas, a *social state* is like an exogenous determinant frame for the realization of the outcome and can include, for example, different societies, policy regimes or time frames. For the present work, we consider two outcome variables separately, the monthly *consumption expenditure* and the weekly *wage earning*, both treated as continuous variables. In particular, we test for equalization of opportunity in India based on a single circumstance variable that of *caste*, between different social states, which corresponds to different *time periods* in our analysis.

As mentioned before, the analysis of *IOP* in principle, consider individual efforts as the legitimate source of inequality. Whereas, any inequality generated by the circumstance factors are ethically and morally objectionable. Therefore for a given level of effort, everyone should face identical outcome distributions irrespective of the differences in their individual circumstances. If we denote circumstance and effort factors by $c$ and $e$, respectively, then under an exogenous social state, $\pi$, the cumulative outcome distribution for a given level of circumstance and effort, can be expressed as $F_\pi(y|c, e)$. Let a *type*, refer

to the group of individuals sharing the same circumstances[5]. Hence, for a given effort level and for any pair of *types*, $(c, c')$, with $c \neq c'$, there exist equality of opportunity under social state, $\pi$, if for all $y$ -

$$F_\pi(y|c, e) - F_\pi(y|c', e) = 0 \tag{2.1}$$

The above condition however, rarely holds with equality and is therefore indicative of *IOP*. Therefore for a given level of effort, any non-zero difference between the circumstance specific outcome distributions should be compensated, so as to ensure equal opportunity in the society under the given social state, $\pi$. This is known in the literature as the *compensation principle* of equalizing opportunity (Roemer 1998, Ramos & Van de Gaer 2012). So when the above condition (2.1) hols as an inequality, then from the perspective of responsibility sensitive egalitarian justice, the policy-maker's objective should ideally be to bring down the left hand side of equation (2.1) to as close as zero, in order to establish an opportunity equal social platform.

Clearly, higher the gap in (2.1), more is the difference between the privileged and the disadvantageous types, and so higher is the *IOP* in the society for the exogenous social state, $\pi$. Therefore for the same set of types, a comparison of these gaps across different social states will eventually allow us to rank the exogenous social states in terms of the extent of their respective *IOP*. As compared to social state, $\pi$, a reduction in this gap for another social state, $\pi'$, indicates that the unethical advantage enjoyed by the privileged type is lesser under $\pi'$. In other words, we can say that opportunity *equalizes* if we move from social state $\pi$ to $\pi'$. However, without the unambiguous identification of the privileged type in any of the social state, the comparison of the social states in terms of their type specific opportunity gaps becomes futile. Testing for *equalization of opportunity* thus comes in two stages. The first stage is to identify the hierarchical ranking among the concerned types for each different social states. Given the ranking of the circumstance types, we can test further for the equalization of opportunity over different social states in the second stage, by ranking the social states themselves. The

[5] *Types* and *circumstances* are often used equivalently, particularly in case of analysis involving a single circumstance factor. Strictly speaking, types are all possible permutation of circumstances. If there are two circumstance variables (*e.g.* sex and race), each having two categories (male-female and black-white, *for example*), then we have four mutually exclusive *types*. See Ramos & Van de Gaer (2012).

methodology is discussed below.

### 2.3.2 Two-step method of equalization of opportunity

#### Ranking the *circumstances*: Inequality of opportunity

As a first step to test for opportunity equalization we need to get unambiguous ranking of all the concerned circumstance types, separately for each exogenous social states. This in effect is the test for the existence of *IOP* under each given social state, which is executed upon exploiting the notion of stochastic dominance. Although the use of stochastic dominance is not new to economics and finance, Lefranc, Pistolesi & Trannoy (2008, 2009) applies this concept for the first time in the literature of *IOP*[6]. The privileged type is identified as the one, the distribution of which dominates that of the other types at certain order of stochastic dominance.

The basic theoretical underpinning of stochastic dominance is provided in the expected utility theory. For any non-decreasing utility function, one distribution, $F(\cdot)$, yields unambiguously better return than another, $G(\cdot)$, if the former first order stochastically dominates the latter. This implies, $F(\cdot) \leq G(\cdot)$, with $F$ and $G$, being the cumulative distribution functions[7]. The same concept is applied in the set up of *IOP*, where $F(\cdot)$ and $G(\cdot)$, corresponds to the different cumulative distributions conditional on different types.

Consider any two types, $c$ and $c'$, such that $c \neq c'$. Therefore under an exogenous social state, $\pi$, and for a given level of effort, their respective type-specific distributions can be written by the cumulative distribution functions, $F_\pi(y|c,e)$ and $F_\pi(y|c',e)$, or equivalently, by the quantile functions $F_\pi^{-1}(p|c,e)$ and $F_\pi^{-1}(p|c',e)$, for all values of the cumulative population percentile, $p$, within the range of $[0,1]$. Then *type*, $c$, will be identified as the privileged type as compared to *type*, $c'$, if the outcome distribution corresponding to the former dominates that of the latter at order one ($c \succ_1 c'$), that is if

[6]For other applications of stochastic dominance in *IOP*, see for example, Peragine & Serlenga (2008), Trannoy, Tubeuf, Jusot & Devaux (2010). See Harris & Mapp (1986), Broske & Levy (1989) for its applications in other areas of economics and finance.

[7]*See* (Mas-Colell, Whinston & Green 1995, Chapter 6).

equation (2.1) holds in the form of following inequality -

$$c \succ_1 c' \iff \underbrace{F_\pi(y|c,e) \leq F_\pi(y|c',e) \quad \text{for all y}}_{\text{first order stochastic dominance}} \iff \underbrace{F_\pi^{-1}(p|c,e) \geq F_\pi^{-1}(p|c',e) \quad \text{for all} \quad p \in [0,1]}_{\text{first order inverse stochastic dominance}} \tag{2.2}$$

Therefore, the first order stochastic dominance of *type* $c$, over $c'$, indicates that the cumulative distribution corresponding to the former type should lie to the right of that of the latter. Equivalently, the above condition can also be concluded from the inverse stochastic dominance of *type* $c$ over $c'$, at order one, if the quantile distribution of the former type lies above than that of the latter[8]. Borrowing from Lefranc et al. (2009), we will say that there exist strong *IOP* in the society, under the social state, $\pi$, if the above condition is satisfied and further, *type* $c$, is enjoying an unethical privilege over *type*, $c'$.

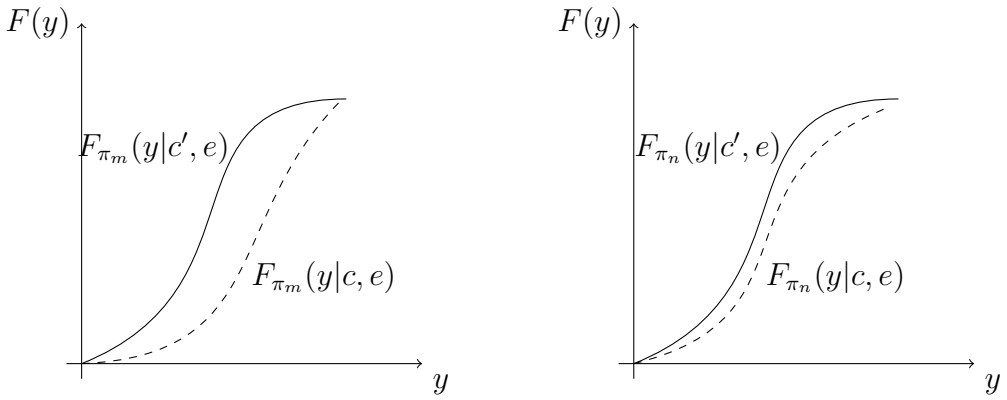
**Ranking the *social states*: Equalization of opportunity**

Once we have the unambiguous ranking of types within each of the exogenous social states we can proceed to rank the social states themselves, by applying the same concept of stochastic dominance, but in a *difference-in-difference* set up. The first difference measures the gap between the type-specific distributions for each given social state, whereas the second difference measures the gap between the social states in terms of the respective gaps in their type-specific distributions.

Figure 2.1 illustrates the basic concept of opportunity equalization across the different social states, for a pair of types $(c, c')$. The left and the right panel of the figure corresponds to the type-specific cumulative distributions under two different exogenous social states, $\pi_m$ and $\pi_n$, respectively. Notice that irrespective of the social states, *type* $c$, has always turned out to be the privileged one, as the cumulative distribution corresponding to this type always lies to the right of that of the other type, for either of the social states. But clearly, as compared to social state $\pi_m$, the privilege enjoyed by the advantageous type is less under social state $\pi_n$. Since the gap in the distributions between the advantageous and the disadvantageous type is lesser under social state $\pi_n$, we can say that the economic opportunity equalizes between those types, if we move from social state $\pi_m$ to

[8]The first and second order stochastic dominance is equivalent to the inverse stochastic dominance of the same order (Shorrocks 1983). However, the equivalence does not hold beyond the second order.

$\pi_n$.




(a) Social state $\pi_m$ (b) Social state $\pi_n$

Figure 2.1: A simple illustration of equalization of opportunity

However, it is possible that the privileged dominant type in social state $\pi_m$, turned out to be the disadvantageous one under social state $\pi_n$. Nevertheless, the principle of *IOP* holds for either social states and a fall in the gap between the type-specific distributions under social state $\pi_n$, is still indicative of opportunity equalization. Thus, from the perspective of responsibility sensitive egalitarian justice, the direction of dominance does not matter per se, what matters instead is the absolute gap between the types. Therefore as far as equalization of opportunity is concerned, we can claim for one if we see a fall in the absolute gap between the type-specific distributions for different social states.

For notational simplicity, let $F_{\pi}^{-1}$ and $F_{\pi}'^{-1}$, denote the distributions of $F_{\pi}^{-1}(p|c,e)$ and $F_{\pi}^{-1}(p|c',e)$, respectively. So for a pair of different exogenous social states, $\pi_m$ and $\pi_n$, the absolute gap between the type-specific distributions for a pair of types, $(c,c')$, can be expressed as -

$$\Gamma(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p) = |F_{\pi_m}^{-1} - F_{\pi_m}'^{-1}| \quad \text{for social state:} \quad \pi_m \tag{2.3a}$$

$$\Gamma(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p) = |F_{\pi_n}^{-1} - F_{\pi_n}'^{-1}| \quad \text{for social state:} \quad \pi_n \tag{2.3b}$$

Given the unequivocal first order dominance between the *types* $c$ and $c'$ for each of the exogenous social states (by equation 2.2), the right hand sides of equation (2.3) are always positive.

Further, let $\Delta_{(c,c')}^{(\pi_m,\pi_n)}$ denote the *difference in difference* between the type-specific distributions of two social states, $\pi_m$ and $\pi_n$, for the same pair of types, $(c, c')$. Hence as compared to social state $\pi_m$ the economic opportunity will equalize under social state $\pi_n$, if the extent of *IOP* between the same pair of types is lesser for the latter social state. Therefore the criteria for opportunity equalization while moving from social state $\pi_m$ to $\pi_n$, requires that for all $p \in [0, 1]$ -

$$\begin{aligned}\Delta_{(c,c')}^{(\pi_m,\pi_n)} &= |F_{\pi_m}^{-1} - F_{\pi_m}'^{-1}| - |F_{\pi_n}^{-1} - F_{\pi_n}'^{-1}| \geq 0 \\ &\Rightarrow \Gamma(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p) \geq \Gamma(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p) \end{aligned} \tag{2.4}$$

Provided the expression of equation (2.4) we can invoke the criteria of first order dominance as presented in equation (2.2), but in a difference-in-difference set up. Therefore we can say that equation (2.4) essentially means that the distribution of $\Gamma(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p)$ dominates that of the $\Gamma(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p)$ at first order inverse stochastic dominance. Since $\Gamma(\cdot)$ is nothing but the gap between the type-specific distributions, the above criteria of opportunity equalization between a pair of social states is also referred as the criteria of *gap curve dominance* in Andreoli et al. (2019). Provided unambiguous ranking among the pair of types, $(c, c')$, at order one, equation (2.4) provides a necessary and sufficient condition for opportunity equalization across a pair of social states.

However if in the first step a pair of types can not be ranked at dominance of order one for any of the social state, the above condition of equalization of opportunity is no longer sufficient. In that case, Andreoli et al. (2019) showed that upon further restricting the class of preferences, it is always possible to rank a pair of types by higher order inverse stochastic dominance so that a necessary and sufficient condition for opportunity equalization can always be formulated for that subset of preferences. In particular, if for all the concerned social states, $\pi_i \in \{\pi_m, \pi_n\}$, the distribution $F_{\pi_i}$ dominates that of $F'_{\pi_i}$, by order $k$ of inverse stochastic dominance, then for a subset of preferences, $\mathcal{R}^k \subset \mathcal{R}$, the general criteria of equalization of opportunity requires that, for all $p \in [0, 1]$ -

$$\Gamma^k(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p) \geq \Gamma^k(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p) \tag{2.5}$$

In the above expression, $\Gamma^k(\cdot)$ is the integrated cumulative distribution gap, defined for the social state $\pi$ as, $\Gamma^k(F_\pi^{-1}, F_\pi'^{-1}, p) = \Lambda_\pi^k(p) - \Lambda_\pi'^k(p)$, where $\Lambda_\pi^k(p)$ and $\Lambda_\pi'^k(p)$, are respectively, the distributions of $F_\pi^{-1}$ and $F_\pi'^{-1}$, integrated at order $k-1$[9].

Therefore if for any of the social states we fail to rank a pair of types $(c, c')$ by dominance of order one, we can not conclude on equalization of opportunity among the social states by condition (2.4), for the class of all rank-dependent preferences, $\mathcal{R}$. However, if for all the social states, types can still be ranked at order two, for example, we can proceed to test for opportunity equalization by condition (2.5), for the sub-class of risk-averse preferences, $\mathcal{R}^2 \subset \mathcal{R}$, and so on.

### 2.3.3 Empirical implementation

The empirical implementation therefore requires to perform two set of pairwise rank dominance tests, corresponding to the two steps described above. The first set consists of the dominance tests that can rank all possible pair of types, separately for each of the exogenous social states. Whereas, given the ranking of the types, the second set of tests consists of the dominance tests of the associated gap curves, for all possible pair of social states. Borrowing from Andreoli et al. (2019), we use the statistical inference set up of the inverse stochastic dominance test, for the empirical execution of the above mentioned tests for ranking types and social states.

Notice that each of the pairwise rank dominance tests, either for ranking types or for ranking social states, are essentially the test for dominance between a pair of distributions at certain order. Therefore for any pair of distributions to compare, say $A$ and $B$, the unequivocal dominance of $A$ over $B$ is always concluded on the basis of three simultaneous tests as - (i) testing that the associated distribution of $A$ dominates that of $B$ at some order, $k$ ($A \succ_k B$), (ii) testing the negation of reverse dominance, that is the corresponding distribution of $B$ does *not* dominate that of $A$ at the same order, $k$ ($A \not\prec_k B$) and (iii) testing for non-neutrality, that the two distributions of $A$ and $B$ does not coincide ($A \not\approx B$).

[9] In particular, define, $\Lambda_\pi(p) = F_\pi^{-1}$, $\Lambda_\pi^2(p) = \int_0^p F_\pi^{-1}(t)dt$, $\Lambda_\pi^3(p) = \int_0^p \int_0^p F_\pi^{-1}(t)dt = \int_0^p \Lambda_\pi^2(t)dt$ and so on. Whereas, $\Lambda_\pi'(\cdot)$ is the same, evaluated over the distribution of *type* $c'$.

Let $\phi_\pi$ denote the finite set of circumstances under *social state* $\pi$ and let $\Pi$ denote the full set of social states to compare. Therefore we conclude unambiguous first order dominance of *type* $c$ over $c'$, under social state $\pi$, if all of the following tests are significant -

$$c \succ_1 c' \iff \hat{F}_\pi^{-1} - \hat{F}_\pi'^{-1} > 0 \quad \forall (c, c') \in \phi_\pi \tag{2.6a}$$

$$c \not\succ_1 c' \iff \hat{F}_\pi^{-1} - \hat{F}_\pi'^{-1} \not< 0 \quad \forall (c, c') \in \phi_\pi \tag{2.6b}$$

$$c \not\approx c' \iff \hat{F}_\pi^{-1} - \hat{F}_\pi'^{-1} \neq 0 \quad \forall (c, c') \in \phi_\pi \tag{2.6c}$$

Where, $\hat{F}_\pi^{-1}$ and $\hat{F}_\pi'^{-1}$, are the empirical counterparts of the quantile distributions of types $c$ and $c'$, respectively, obtained from the sample under social state, $\pi$. The above batch of tests (2.6a, 2.6b, 2.6c) are then executed for all pairs of types, $(c, c') \in \phi_\pi$, within a social state $\pi$ and separately, for each of the exogenous social states, $\pi \in \Pi$. So the tests in (2.6) is to establish an unambiguous ranking order among all pair of types for each of the exogenous social states[10].

The empirical quantile distributions, eventually generates the empirical gap curves as, $\hat{\Gamma}(F_\pi^{-1}, F_\pi'^{-1}, p) = \hat{F}_\pi^{-1} - \hat{F}_\pi'^{-1}$, from equation (2.3). Therefore for the pairs of types, $(c, c')$, for which there is no ambiguity about the ranking of the types, we can conclude in favor of opportunity equalization in social state $\pi_n$ as compared to $\pi_m$, if all of the following tests of gap curve dominance are significant -

$$\hat{\Gamma}(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p) \succ_1 \hat{\Gamma}(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p) \iff |\hat{F}_{\pi_m}^{-1} - \hat{F}_{\pi_m}'^{-1}| - |\hat{F}_{\pi_n}^{-1} - \hat{F}_{\pi_n}'^{-1}| > 0 \tag{2.7a}$$

$$\hat{\Gamma}(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p) \not\succ_1 \hat{\Gamma}(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p) \iff |\hat{F}_{\pi_m}^{-1} - \hat{F}_{\pi_m}'^{-1}| - |\hat{F}_{\pi_n}^{-1} - \hat{F}_{\pi_n}'^{-1}| \not< 0 \tag{2.7b}$$

$$\hat{\Gamma}(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p) \not\approx \hat{\Gamma}(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p) \iff |\hat{F}_{\pi_m}^{-1} - \hat{F}_{\pi_m}'^{-1}| - |\hat{F}_{\pi_n}^{-1} - \hat{F}_{\pi_n}'^{-1}| \neq 0 \tag{2.7c}$$

The above tests of gap curve dominance, is executed for all pair of social states, $(\pi_m, \pi_n) \in \Pi$, so as to determine whether opportunity equalizes between the types, $(c, c')$, for different

[10]The estimates and null hypotheses associated to the tests in (2.6) are borrowed from Beach & Davidson (1983) and is provided in Apendix 2.A. In case of failure to rank types at order one, the test is reconstructed with the estimated integrals of the empirical quantile distributions, to test for higher order dominance. In particular, since the second order dominance is equivalent to the rank dominance of the generalized Lorenz curves (Shorrocks 1983), Beach & Davidson (1983) provides the estimation of the first integral of the empirical quantile functions, $\hat{\Lambda}_\pi^2$, from the corresponding generalized Lorenz functions, in case of second order dominance. Further, Andreoli (2018) provides details of inverse stochastic dominance for orders higher than two.

social states[11]. Details on the test-hypotheses of the corresponding dominance tests of (2.6) and (2.7), are provided in Appendix 2.A.

## *2.4 Data, variables and sample selection*

### 2.4.1 Data

For the present analysis of equalization of opportunity across castes in India, we have taken data from the National Sample Survey (*NSS*). This is the biggest nationally representative micro level database for India, collected by the National Sample Survey Organization (*NSSO*), India. Several important national level surveys have been regularly conducted by *NSSO* since world war II. We take the *employment-unemployment survey* in particular, that covers the whole country except some remote inaccessible area[12]. We have taken six consecutive rounds of the employment and unemployment survey of *NSS*, covering years *1983*, *1987-88*, *1993-94*, *1999-00*, *2004-05* and *2011-12*[13].

These rounds on average, survey 120000 households, enumerating over half-a-million Indian nationals. Initially we have to drop 500 to 2000 observations per round, to clean for valid age, sex, sector, caste specification, marital status and some other criterion, depending on different rounds. *NSS* provides several important household and demographic details, including information on caste and household expenditure. Wage data in *NSS* however, is truncated to the regular and casual wage earners, who are not self-employed. This eventually excludes 30-40% of the adult working population from our wage analysis. One of the most interesting feature of our data analysis is the chosen time frame. After following central interventionist policy for the first forty years of independence, India had gone through a major economic reform during the early nineties. Indian economy have seen a barrage of neo-liberal open-market Government policies during early to mid-nineties, along with a massive expansion of private and foreign investments. The impact of this neo-liberal reform is hugely debated, as both youth unemployment and inequality

[11]The null hypotheses associated to the tests in (2.7) are constructed by Andreoli et al. (2019) and is provided in Appendix 2.A. In case of types ranked by higher order dominance, the tests of (2.7) is simply reformulated with the associated empirical gap curves as, $\hat{\Gamma}^k(\hat{F}_\pi^{-1}, \hat{F}_\pi'^{-1}, p) = \hat{\Lambda}_\pi^k(p) - \hat{\Lambda}_\pi'^k(p)$.

[12]So conflict areas of Ladakh & Kargil districts of Jammu & Kashmir, some remote interior villages of Nagaland, few unreachable areas of Andaman & Nicobar Islands and those villages recorded as uninhabited by the respective population census, are kept out of these surveys.

[13]This means we have taken *Schedule 10.0* survey of *NSS*, for rounds 38 (1983), 43 (1987-88), 50 (1993-94), 55 (1999-00), 61 (2004-05) and 68 (2011-12). *See* NSSO data appendix A for further details.

shows a sharp increase since then, along with a slower pace of poverty reduction (Deaton & Dreze 2002, Himanshu 2007, Dev & Ravi 2007). By covering a time frame of *1983-2012*, we can enrich the debate by exploring the opportunity equalizing impact of the reform among the caste groups.

### 2.4.2 Main variables

**Circumstance:** As mentioned before, we analyze equalization of opportunity in India on the basis of castes. For analyzing opportunity equalization over the time frame of *1983-2012*, we consider two categories of caste. The lower category consists of the most historically disadvantageous caste groups together as *SC/ST*, whereas the rest of the population are considered as the upper caste category, whom we refer as the *non-SC/ST*. However, the *non-SC/ST*s did not benefit from any reservation policy until early nineties, before the classification of *OBC*s in India. As mentioned before, *OBC*s are the relatively socially and economically backward castes within the *non-SC/ST*s, who constitute over 40% of the national population and is entitled to a reservation quota of 27% since 1992. However, the provision of the *OBC* data for the present data base of *NSS* is only available from the survey year of *1999-00*. Therefore to provide a rather subtle and recent picture of opportunity equalization among castes in the country, we proceed to analyze the latest three survey years with finer categories of castes that resembles the caste categorization of modern India. In particular, over the time frame of *1999-2012*, we provide a separate analysis of equalization of opportunity among the three caste categories, namely, *SC/ST*, *OBC* and *General*, where the last category of '*General*' represents the most forward castes in the country, who are excluded from any caste based reservation benefits.

**Outcomes:** Caste based equalization of opportunity is examined for two outcomes, consumption and wage. Consumption is the monthly household per capita expenditure (*MPCE*), reported as the total monthly expenditure on selected durable and non-durable goods, incurred by the household over the month prior to the survey. *MPCE*, therefore, is reported for every household, which we divide by household size to get the individual level values. Our second outcome, that of wage, is selectively reported for the class of regular and casual wage earners, for multiple activities. Unlike *MPCE*, wage is reported as the weekly wage received or receivable over the past week prior to the date of the survey. We

consider the daily wage of the major activity pursued by the individual in the reference week. For this, we divide the total weekly wage by the number of days engaged in that major activity. We select 'major activity' as the one, on which maximum number of days had been spent by the individual. In case of equal number of days spent on more than one activity, we prioritize those having valid wage entry and occupation information. In particular, borrowing from Hnatkovska et al. (2012), we consider real consumption expenditure and wage earning as our outcome variables, upon dividing them by the state level absolute poverty lines[14].

### 2.4.3 Sample selection

For the present analysis we choose the adult working population as our sample. In particular, our sample consists of individuals aged between 18 to 60 years, who are not currently enrolled in any educational institution, have valid occupation information and are from single-headed as well as male-headed households. Their are several rationale for the sample selection criteria. First of all, we like to limit the age to the eligible working years. Since, 60 is the age of retirement for most of the jobs in India, we limit our sample to be aged between 18-60 years old. However, it is not uncommon for young adults to pursue higher studies. Since we do not want to analyze opportunity equalization for adults who are still in their formation period *per se*, we restrict our sample further to employed individuals who have reportedly finished their education and is not enrolled in any type of educational institution at the time of the survey. Finally, since both multi-headed and female-headed households are rare and subject to special constraints, we focus on single and male-headed households only. However, across all the rounds, over 90% household heads are male and 99% households are single-headed.

The above restrictions leave us about 0.13 to 0.18 million individuals as our working sample. As mentioned before, wage information in *NSS* data base is limited to the casual and regular wage only. Therefore our sample for the wage analysis (*wage sample*) is

[14]We use poverty lines, that can account for the differences in standard of living across the states of India. Besides, the measure of absolute poverty line is provided by the Planning Commission of India using data collected by the same survey, that of the National Sample Survey, the one we use for the present analysis. Another commonly used deflator is the consumer price index, which we did not use, as it was measured on the basis of a different survey and prior to 2011, the combined rural and urban price indices are not provided (instead, consumer price index used to comprise of multiple series like, urban non-manual labor, agricultural labor, rural labor and industrial workers).

further truncated to those who have valid wage information as well. Table 2.1 provides the sample summary statistics, where the upper panel corresponds to our working sample and the lower panel to our wage sample. Notice that our samples are predominantly rural married working males, who on average are 35 years old. Like the whole country, *SC/ST*s constitute nearly 30% of the working sample. However, over 75% of our working samples are male, although the male to female ratio in India is about 60 : 40. The over-representation of males in our sample is driven by the low female labor force participation in the country[15]. Further, two of the iconic feature of Indian economy is portrayed by the summary statistics. First, a clear improvement of education is prominent. While 56% of our working sample are deprived of any formal schooling in *1983*, the figure has fallen to 28% by *2012*. Secondly, a shrinking of the agriculture sector is also noticeable over this time, which is most natural for an emerging industrialized country like India.

The last but one column in Table 2.1 (%wage) shows the share of our working sample who have valid wage information, that eventually generates the respective sample sizes of our wage sample as reported in the last column of the lower panel of the table. For example, 46% of our working sample in *2011-12* have valid wage data, thereby shrinking the sample size for this survey year from 132552 to 58330 for wage analysis. Notice that only 4% of the working sample in *1987-88* have valid wage information, which is considerably low than all other rounds. This exceptionally low wage data in *1987-88* is in fact a result of unusually low rural wage observation for this round, which compels us to exclude this round from our wage analysis. Nevertheless, in terms of age, household size, sex or marital status, the selected wage sample is not very different from the working sample. However, not unnaturally, the non-self-employed regular salaried workers are relatively less rural agricultural laborers with comparatively better education. Further, the share of *SC/ST*s are marginally higher for of the wage sample, especially for the later rounds.

[15]In the said age bracket of 18-60 years, about 30% are working women, whereas over 60% females have reported not to be in the labor force for attending domestic duties.

| | age | hhsize | %male | %SC/ST | %rural | %married | %noschool | %agri | %wage | N |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Work sample*** | | | | | | | | | | |
| *1983* | 35.10 | 6.2 | 0.76 | 0.28 | 0.78 | 0.83 | 0.56 | 0.62 | 0.23 | 167609 |
| [38] | (0.04) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *1987-88* | 35.18 | 6.0 | 0.76 | 0.27 | 0.79 | 0.84 | 0.54 | 0.60 | 0.04 | 182816 |
| [43] | (0.03) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *1993-94* | 35.57 | 5.6 | 0.76 | 0.28 | 0.78 | 0.83 | 0.48 | 0.62 | 0.30 | 164496 |
| [50] | (0.04) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *1999-00* | 35.81 | 5.8 | 0.76 | 0.31 | 0.77 | 0.83 | 0.44 | 0.58 | 0.42 | 169724 |
| [55] | (0.03) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *2004-05* | 36.11 | 5.6 | 0.75 | 0.29 | 0.76 | 0.83 | 0.39 | 0.55 | 0.39 | 182191 |
| [61] | (0.04) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *2011-12* | 37.17 | 5.1 | 0.81 | 0.29 | 0.72 | 0.83 | 0.28 | 0.47 | 0.46 | 132552 |
| [68] | (0.06) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| ***Wage sample*** | | | | | | | | | | |
| *1983* | 34.85 | 6.0 | 0.78 | 0.27 | 0.64 | 0.81 | 0.49 | 0.52 | 1.0 | 40050 |
| [38] | (0.07) | (0.02) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *1987-88* | 35.35 | 5.6 | 0.79 | 0.21 | 0.35 | 0.82 | 0.38 | 0.31 | 1.0 | 9628 |
| [43] | (0.16) | (0.04) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | | |
| *1993-94* | 34.56 | 5.1 | 0.70 | 0.35 | 0.86 | 0.84 | 0.56 | 0.64 | 1.0 | 42059 |
| [50] | (0.06) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *1999-00* | 35.28 | 5.2 | 0.78 | 0.38 | 0.69 | 0.83 | 0.43 | 0.48 | 1.0 | 68627 |
| [55] | (0.05) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *2004-05* | 35.18 | 5.1 | 0.79 | 0.31 | 0.68 | 0.81 | 0.38 | 0.42 | 1.0 | 66297 |
| [61] | (0.06) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| *2011-12* | 36.03 | 4.8 | 0.83 | 0.33 | 0.65 | 0.81 | 0.27 | 0.34 | 1.0 | 58330 |
| [68] | (0.08) | (0.02) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.01) | | |

Table 2.1: Sample summary statistics [a]

[a]Standard errors are in parentheses and rounds in squared brackets. 'hhsize' is household size. %male, %SC/ST, %rural, %married, %noschool, %agri indicates the percentage share of our sample who are male, SC/ST, rural, married, have no formal schooling, are engaged in agriculture related jobs, respectively. Whereas %wage indicates share of our work sample who have valid wage data. The last column (N) reports the respective sample sizes.

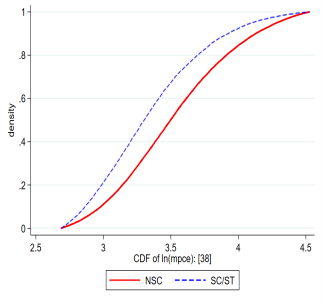
### 2.5.1 How far opportunity equalizes over castes in India

Our aim in this section is to find evidence of opportunity equalization, or the lack of it, among the privileged upper caste group, *non-SC/ST* (abbreviated as *NSC*), and the historically disadvantageous deprived caste group, *SC/ST*, over the time span of *1983-2012*. We focus on two aspects of standard of living in particular, that of consumption expenditure (MPCE) and wage earning.

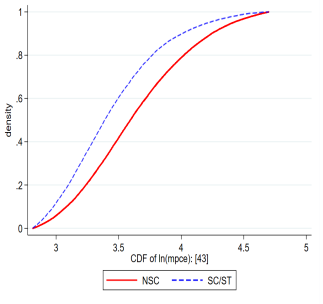
As a first step to test for opportunity equalization we need to confirm the dominant and the dominated caste groups, separately for all survey years. The caste specific *CDF*s in Figure 2.2 shows that irrespective of the outcome, the *CDF*s corresponding to the *NSC* lies always to the right of that of *SC/ST*, over the entire time frame. Therefore the visual plots are suggestive of first order dominance of *NSC* over *SC/ST*, for both *MPCE* and wage. The statistical confirmation of this fact is provided in Table 2.2, which reports the test-statics with the associated p-values, corresponding to the empirical test of first order dominance (2.6). Both for consumption and wage, Table 2.2 shows that the null of equality (NSC $\approx$ SC/ST) and that of the lower caste dominance (SC/ST $\succ_1$ NSC) are strongly rejected for all rounds. Whereas, the null of the upper caste dominance (NSC $\succ_1$ SC/ST) can not be rejected for any of the round. So in terms of consumption and wage, the non-scheduled caste group, *NSC*, is indeed more advantageous than the *SC/ST*s and remain so for nearly three decades.

However it is only *MPCE*, for which a convergence in gap between the upper and the lower caste is visible for the latest round in *2011-12*. Without any further test, we can at most claim for a possibility of equalization of opportunity in *MPCE* across the *NSC*s and the *SC/ST*s, during the time period of *2004-12*. In contrast, an increase in gaps between the caste specific distributions in Figures 2.2h and 2.2i, suggest a possible disequilization of the earning opportunity over the period of *1993-2000*. Other than that, almost none of the caste specific distributions gives a suggestive visual indication of opportunity equalization or the lack of it. We therefore proceed to test for the gap curve dominance to compare the extent of *IOP* in different survey years.

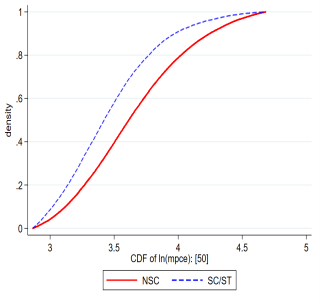
(A) Effect of caste on *MPCE*



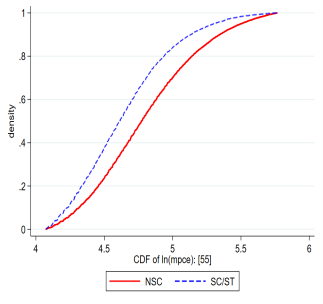

(a) 1983



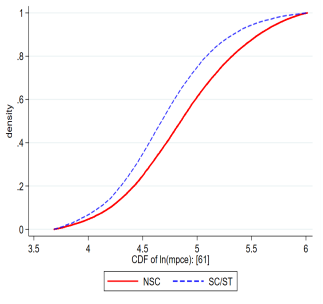

(b) 1987-88



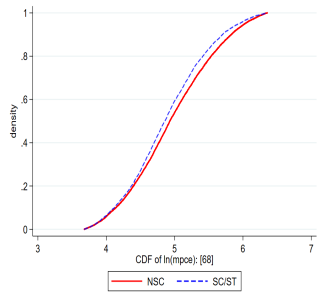

(c) 1993-94




(d) 1999-00

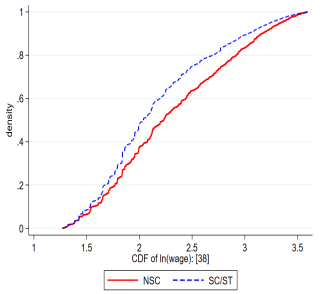



(e) 2004-05

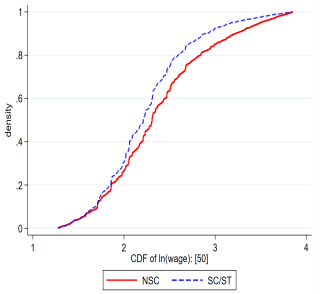



(f) 2011-12

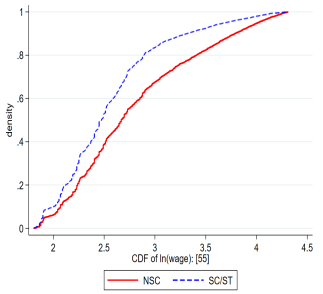
(B) Effect of caste on *Wage*



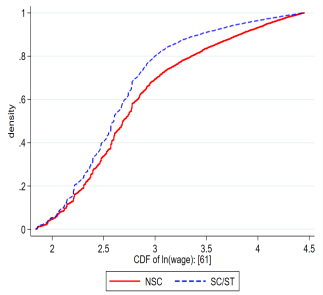

(g) 1983



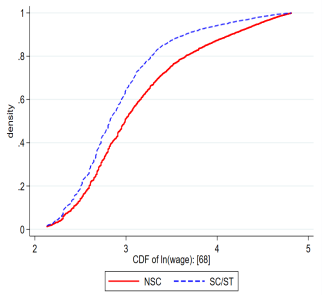

(h) 1993-94




(i) 1999-00




(j) 2004-05




(k) 2011-12

Figure 2.2: Caste specific cumulative distributions: MPCE and Wage

| | *1983* | *1987-88* | *1993-94* | *1999-00* | *2004-05* | *2011-12* |
|---|---|---|---|---|---|---|
| ***MPCE*** | | | | | | |
| NSC $\approx$ SC/ST | 5720.0 | 6793.4 | 6779.6 | 6054.9 | 4717.8 | 673.8 |
| (p-value) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| NSC $\succ_1$ SC/ST | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| (p-value) | (0.947) | (0.938) | (0.936) | (0.942) | (0.955) | (0.953) |
| SC/ST $\succ_1$ NSC | 5720.0 | 6793.4 | 6779.6 | 6054.9 | 4717.8 | 673.8 |
| (p-value) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| ***Wage*** | | | | | | |
| NSC $\approx$ SC/ST | 1373.2 | · | 2065.6 | 3943.4 | 2447.0 | 1946.5 |
| (p-value) | (0.000) | · | (0.000) | (0.000) | (0.000) | (0.000) |
| NSC $\succ_1$ SC/ST | 0.0 | · | 0.0 | 0.0 | 0.0 | 0.0 |
| (p-value) | (0.946) | · | (0.941) | (0.930) | (0.928) | (0.919) |
| SC/ST $\succ_1$ NSC | 1373.2 | · | 2065.6 | 3943.4 | 2447.0 | 1946.5 |
| (p-value) | (0.000) | · | (0.000) | (0.000) | (0.000) | (0.000) |

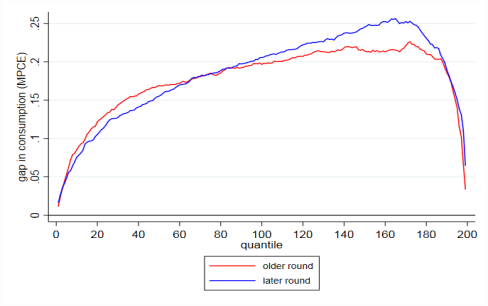
Table 2.2: Dominance test result for *IOP* between castes: All India[a]

[a]'X$\approx$Y' denotes the null of neutrality between X and Y, whereas 'X$\succ_1$Y' is null of first order dominance of X over Y. Selected (Kodde & Palm 1986) critical values - (1.642, 26.625) [10%]; (2.706, 29.545) [5%]; (5.412, 35.556) [1%], in the form of (lower bound, upper bound). Reject the null if larger than upper bound, accept if lower than lower bound, conclude on the basis of p-values otherwise.

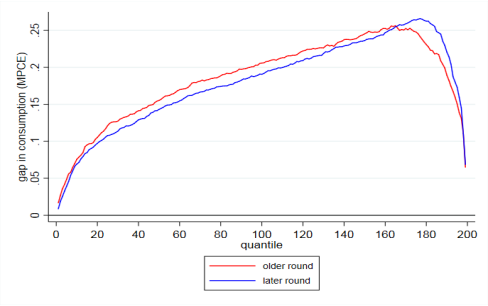
Figure 2.3 plots the respective gaps in the caste specific quantile functions for each pair of the adjacent survey years, where the left panel plots for *MPCE* and the right panel for wage[16]. Opportunity equalization across the caste groups requires that the gap curve correspond to the older round should lie above than that of the latter round. Figure 2.3 suggests that this criteria is satisfied for certain time brackets only, in case of both consumption and wage. The same is concluded from the statistical gap curve dominance test results as provided in Table 2.B.1 of Appendix 2.B, where each panel of *MPCE* and wage, reports the test-statistics corresponding to the empirical tests of (2.7) with the associated p-values. Therefore we can not say that India since 1983, have seen a consistent equalization or disequalization of opportunity in consumption or wage earning, among the *non-SC/ST* and *SC/ST*. Rather opportunity equalization not only remain sporadic over this time frame, but it affects consumption and earning differently, especially since 1993.

As mentioned before, over the time frame of *1983-2012*, one of the most interesting change in the Indian economy is the introduction of open-market neo-liberal economic

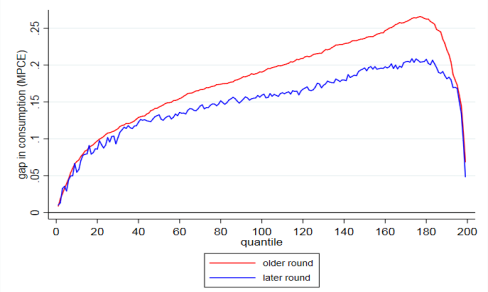
[16]Notice that, since the round corresponding to 1987-88 is left out of the wage analysis, the gap curves for adjacent survey years has less sub-figures for wage in Figure 2.3.



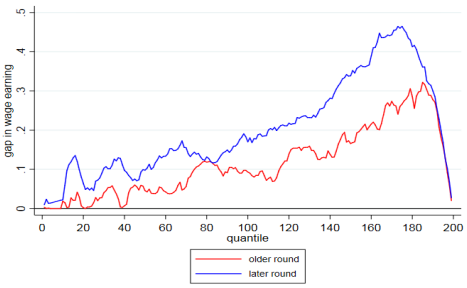

(a) MPCE: 1983-1987



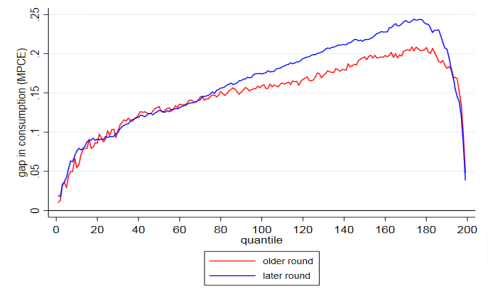

(b) MPCE: 1987-1993



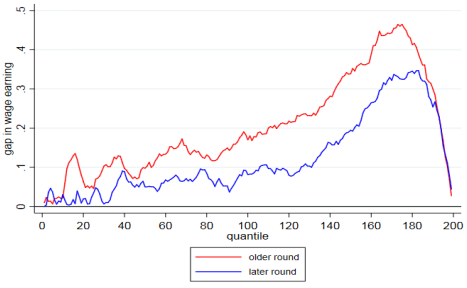

(c) MPCE: 1993-1999



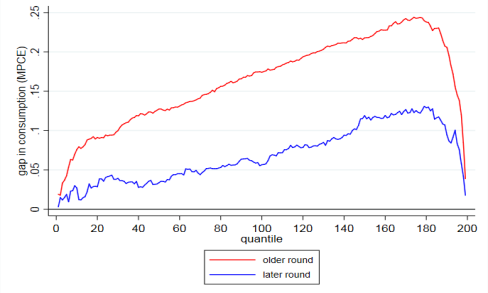

(d) Wage: 1993-1999



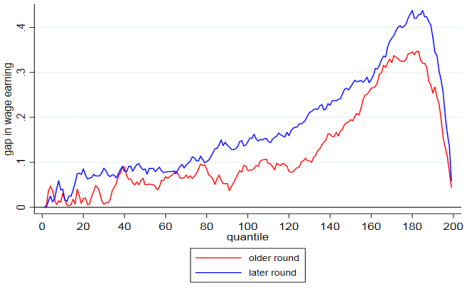

(e) MPCE: 1999-2004




(f) Wage: 1999-2004




(g) MPCE: 2004-12




(h) Wage: 2004-12

Figure 2.3: Gap curves: MPCE and Wage

reforms around early to mid-nineties. A host of liberalizing policies were initiated in the country since 1991, with expanding foreign investment and a shift in focus from the public to the private sector. We can therefore roughly divide our time frame in three policy-timezones, where the earlier three survey years (*1983, 1987-88, 1993-94*) can roughly be thought as the *pre-reform* period and that of the latest three (*1999-00, 2004-05, 2011-12*) as the *post-reform* period. The intermediate time period of *1994-99* can be denoted as the reform phase. Table 2.3 below provides a schematic representation of equalization or disequalization of opportunities among *non-SC/ST* and *SC/ST*, for all pairs of survey years, covering a decade prior to the economic reform to a decade after that, both for *MPCE* and wage[17].

| | *Reform*-10 | *Reform*-5 | *Reform phase* | | *Reform*+5 | *Reform*+10 |
|---|---|---|---|---|---|---|
| | *1983* | *1987-88* | *1993-94* | *1999-00* | *2004-05* | *2011-12* |
| ***MPCE*** | | | | | | |
| *1983* | - | ○ | ○ | ● | ⊘ | ● |
| *1987-88* | ● | - | ◉ | ● | ● | ● |
| *1993-94* | ◉ | ● | - | ● | ● | ● |
| *1999-00* | ○ | ○ | ○ | - | ○ | ● |
| *2004-05* | ⊘ | ○ | ○ | ● | - | ● |
| *2011-12* | ○ | ○ | ○ | ○ | ○ | - |
| ***Wage*** | | | | | | |
| *1983* | - | - | ◉ | ○ | ⊘ | ○ |
| *1993-94* | ○ | - | - | ○ | ○ | ○ |
| *1999-00* | ● | - | ● | - | ● | ● |
| *2004-05* | ⊘ | - | ● | ○ | - | ○ |
| *2011-12* | ● | - | ● | ○ | ● | - |

Table 2.3: *Equalization of opportunity across non-SC/ST and SC/ST*: A time-scape[a]

[a]Where, (○, ◉, ●) corresponds to p-values - ($p < 0.1, 0.1 < p < 0.5, 0.5 < p$), respectively, for the null of *row*-year dominates *column*-year at order one. Darker cells are therefore suggestive of stronger rejection of this null. ⊙ represents that the null of equality between the row and the column, can not be rejected at 5%. ⊘ represents inconclusive result for the corresponding cell, that is, neither of equalization, disequalization, or neutrality can be concluded. The column titled 'Reform±10' (likewise 'Reform±5') indicates 10 (likewise 5) years after/before neo-liberal economic reform.

Each cell of Table 2.3 reports the p-values corresponding to the null hypothesis that the gap curve corresponding to survey years in the row dominates that of the column. The darkest cells represent the failure to reject the associated null for p-values above

[17]To be specific, Table 2.3 represents the opportunity equalization tests over all pairs of survey rounds. The corresponding test results of gap curve dominance are provided in Table 2.B.1 of Appendix 2.B and Table 2.3 is constructed from the associated p-values as provided in Table 2.B.1.

0.5, whereas the lighter cells represent the same with lower p-values. Eventually, the lightest cells correspond to rejecting the aforementioned null for p-values lower than 0.1. For example in case of *MPCE*, the intersecting cell of 1987-88 as row-year and 1983 as column-year says that the opportunity gap in 1987-88 does dominate that of in 1983, which indicates a lack of opportunity equalization over the time period of 1983-88[18].

Notice that each panel of *MPCE* and wage in Table 2.3 can be read as two disjoint diagonals. While more darker cells in the upper diagonal is an indication of opportunity equalization over time, the same in the lower diagonal indicates the lack of it. Likewise, we see from Table 2.3 that, while opportunity among the castes mostly equalizes for *MPCE* over this time span, this is not the case for regular wage earners. Interestingly, the economic reform in the mid nineties, has opposite impacts on the two concerned outcomes. For *MPCE*, Table 2.3 shows a weak disequalization of opportunity in the pre-reform period (upto *1993*), follows by a much stronger equalization across the castes, thereafter. Whereas for the regular/casual wage earning, although the pre-reform period shows weak evidence of equalization, opportunity disequalizes strongly since *1993-94*. Even a phase of consumption disequalization in the first half of the post-reform period (*1999-2005*) was more than offset by an equalization phase thereafter (*2005-2012*), which is exactly the opposite for wage. Nevertheless during the post-reform period of *1999-2012*, opportunity equalizes both for *MPCE* and wage.

Therefore to paraphrase our results, we first of all find clear evidence of strong *IOP* between the *non-SC/ST*s and *SC/ST*s. The relatively advantageous *non-SC/ST*s are always found to be the dominant caste category as compared to the *SC/ST*s and remain privileged till date. But the caste gaps in consumption expenditure seem to converge substantially since the economic reform, indicating noticeable equalization of opportunity over this period. With almost similar sample, the non-parametric index of *IOP* also shows that 1% of consumption inequality is due to caste alone for the survey year of *2011-12* (Chapter 1). However, as compared to *1983*, the test of opportunity equalization shows that caste gaps among the regular/casual wage earners seem to have increased after three decades of time, specifically for the disequalizing effect on wage right at the time

[18]Further, this fact is also backed up by the diagonally opposite light shaded cell for *MPCE*, where 1983 is the row-year and 1987-88 is the column-year. The light shade of this cell says that the gap curves of 1983 does *not* dominate that of 1987-88.

of economic reform (during *1994-99*). Related dynamics of *IOP* over this time frame is reflected by the non-parametric index of *IOP* as well[19]. Consumption *IOP* increased about 26% in the pre-reform period and noticeably decrease for the latest survey year. Whereas wage *IOP* due to caste in *2012* is nearly three times larger than what it was thirty years ago, with a sharp increase in its value during *1999-00*.

**Equalization among castes: Why it differs for consumption and wage?**

Although consumption and wage are often analyzed side by side in many of the welfare analysis as two comparable standards of living, they generate very different results in terms of opportunity equalization among caste groups in India. A number of factors can be attributed to this difference. One of the main reason for this disparity lies in the reporting of the consumption and wage data in the *NSS* survey. Wage is reported only for the selected section of regular and casual wage earners, that excludes a substantial portion of working adults who are self-employed. Since the share of self-employed workers are higher for the lower castes, especially in rural India (Gang et al. 2008), the caste dynamics may very well be different after the inclusion of self-employed workers. The coverage of consumption data (*MPCE*) on the other hand, is not limited to certain section of the population, but unlike wage, it is reported as a household level measure and thereby masking the intra-household differences in consumption patterns.

The reported wage data may additionally be more heterogeneous because of treating all kinds of regular and casual workers similarly, although the nature of employment varies to a large extent among them. On one hand it includes the temporary casual labors who are mostly employed by the various transitory public work programs in rural India in exchange of a scanty short-term remuneration. On the other hand it is also reported for white collar professionals who have a steady flow of monthly income. The wage disparity across the castes may partially reflect this heterogeneity, as the share of upper castes are significantly higher in the white collar professions, whereas the deprived castes are rather concentrated in agriculture or impermanent low-skilled transitory jobs. Even though the share of *SC/ST*s in agro-based occupations does decrease over time it is

[19]Table 2.B.2 in Appendix 2.B reports the non-parametric ex-ante measure of *IOP* using the index of *mean log deviation*, adopted from Checchi & Peragine (2010).

considerably higher than that of the *non-SC/ST*s even today. Therefore the exclusion of a large section of self-employed agricultural *SC/ST*s may further aggravate the wage gaps among the castes. Further as pointed by Gang et al. (2011), depending on varying returns to physical and human capital, the income generating model can itself be very different between the *SC/ST*s and the *non-SC/ST*s. Indeed, lower return on human capital along with the presence of mutual caste based network is found to be responsible for generating a low level equilibrium trap for the deprived lower castes, who, by their choice of poor quality education often end up in low-paid traditional jobs (Munshi & Rosenzweig 2006). This aggravates the long sustaining wage gap between the caste groups.

The data on *MPCE* on the other hand is reported for all enumerated households as the total expenditure incurred by the household over the last month prior to the date of the survey. Hence members with different income but residing in the same household are reported to have identical consumption expenditure. Therefore one of the main drawback of the *MPCE* data is its inability to capture the intra-household consumption diversities, which could vary across the caste groups. Also *MPCE* includes the consumption of domestic production as well, that may generate better economies of scale for the rural agricultural *SC/ST*s, whose basic food consumption are often drawn from the household crop production. Nevertheless, *MPCE* is yet worth to consider as a proxy for the standard of living, because the coverage of it is not limited to certain households with selective characteristics. Several studies with the *NSS* data-base found that the average *MPCE* of the *SC/ST*s is significantly lower than that of the *non-SC/ST*s, as the share of person living below the poverty line is always higher for the former caste category (Deshpande 2001, Kijima 2006). We get a similar picture in consumption as well, that confirms a strong dominance of the *non-SC/ST*s over the *SC/ST*s even today, although by *2011-12* the upper caste premium enjoyed by the advantageous category does decrease considerably.

However, the consumption pattern among these caste groups may be very different. With a higher share of habitation below the poverty line, the destitute groups of *SC/ST*s are more likely to spend a large portion of their income on basic consumption items like food and clothing, instead of luxury goods. Whereas the difference in consumption will likely to be rather tuned with the inclusion of non-basic comfort goods that

typically signify affluence. Unfortunately the reported data of *MPCE* in the present *NSS* survey, that of the employment-unemployment survey, is not enough to capture the variability in consumption pattern across different castes. Instead, *NSS* conducts a separate national level survey dedicated exclusively to 'consumer expenditures' that provides much detailed data on individual consumption. Nevertheless, considering the importance of *MPCE*, the present schedule of the employment-unemployment survey do provide a mini-questionnaire on household *MPCE* that reports consumption on selected important durable and non-durable goods. Naturally this selection is more biased towards the inclusion of basic necessary commodities rather than luxury goods. Since most of the Government or non-Government subsidies are on basic goods, the partial reporting of *MPCE* eventually attenuates the consumption gap across different castes over this time span that witnessed the strengthening of several pro-poor subsidy schemes.

Therefore for the inherent characteristic differences in the *MPCE* and the wage data, the time-scape of these variables reflect quite different stories, especially since the neo-liberal economic reform in the early nineties. While opportunity mostly equalizes for *MPCE* since the launch of reforms around 1993, it disequalizes for the regular/casual wage earning. Along with an emphasis on the private sector, India has also seen some of the major pro-poor Government initiatives during the decade following liberalization. Our latest survey year of *2011-12* is the first national level survey after the implementation of the biggest rural public work program in India, the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA), by virtue of which rural unskilled workers are guaranteed hundred days of paid public work per year. The impressive opportunity equalization in *MPCE* for *2011-12* could have been attributed to this employment generation scheme, that is shown to have some positive impact on raising consumption expenditure of the poor (Bose 2017). Poorer households living below the poverty line also benefits from subsidized food grains provided by the Targeted Public Distribution System (TPDS) launched in 1997. Provided the larger share of food expenditure, this benefits the poor *SC/ST*s more than the affluent upper castes. Together, these pro-poor policies may have an equalizing impact on consumption expenditures among the different caste categories, particularly for the later survey years.

While the open-market policies in early nineties unleash new opportunities for the private sector employment, they also comes at the cost of shutting down of many public enterprises and thereby crunching opportunities for the illiterate low-skilled manual labor class of the country, who are often over-crowded by the deprived castes. Indeed, most of the disequalization of earning opportunity takes place in the first five years of reform. In fact liberalization of the Indian economy is not only accompanied by growing income inequality in India (Pal et al. 2007) but by increasing unequal earning opportunity as well (Chapter 1). Even the biggest the employment generation program of MGNREGA have shown little impact on increasing wage of working men of rural India (Zimmermann 2012). But delayed payments and lack of local administrative planning may grossly underestimate this impact where wage is reported only for the last week prior to the survey date.

### 2.5.2 Other backward classes: An account of post-reform India

For comparability across the survey years since 1983, our analysis so far was confined to the testing of equalization of opportunity between the *non-SC/ST*s and the *SC/ST*s. But there is huge intra-caste heterogeneity in terms of social and economic backwardness, especially among the *non-SC/ST*s who constitute about 70% of the total population. In fact as mentioned in section 2.2, modern India embodies a finer caste categorization since the classification of the 'Other Backward Classes' (*OBC*) in the mid-eighties. *OBC*s are identified as the relatively deprived section of the *non-SC/ST*s who embodies over 40% of the national population and are entitled to certain percentage of reservation in higher education, Government job or political assembly since the beginning of nineties. However, *NSS* provides data on *OBC* only since *1999-00*. Therefore in this section we focus our attention to the last thee survey years of *1999-00*, *2004-05* and *2011-12*, and test for opportunity equalization in *MPCE* and wage, among three categories of caste groups, namely, *General*, *OBC* and *SC/ST*.

There are several reasons to consider a finer caste categorization in the distributional analysis of equalization of opportunity. First of all, this is the caste categorization of modern India and is therefore able to generate more contemporary results on opportunity equalization among different castes. Secondly, *OBC*s benefit from some caste based

reservation quotas while *Generals* (abbreviated as *Gen*) are not. Taking the *non-SC/ST*s in one bracket therefore mix up a diverse class of caste groups where only some of them are beneficiaries of caste based affirmative policies. This may contaminate the associated results of opportunity equalization among castes. Third, the deprivation of the historically disadvantageous caste groups, *SC/ST*, are actually underestimated when the comparison group is *non-SC/ST* (*NSC*) rather than *General* (Azam 2012). Therefore the degree of *IOP* as well as the need of equalization is actually greater for the most deprived castes of *SC/ST* when compared to the most advantageous forward caste category of *General*[20]. Finally, due to limited data and the associated incompatibility for analysis over long time span, work on *OBC* is relatively rare. Covering more than a decade after the economic reform in India, we aim to fill this gap by providing a consistent evolution of the *OBC*s in the existing hierarchical social fabric of the country.

Table 2.4 provides the caste composition for the selected rounds in the post-reform period, both for our work and wage sample. Naturally the share of *SC/ST* is the same as before (*see* Table 2.1). But notice that the share of *OBC*s are not the same across rounds and is actually increasing over time. However as mentioned in section 2.2, unlike *SC/ST* there is no fixed national list of castes to be included as *OBC* and the sanctioned list of *OBC* is often determined at the state level. Besides, since the announcement of reservations for the *OBC*s around 1990, many of the hitherto deprived castes fought for the *OBC* status. Therefore, it is not unnatural to see an increase in the share of *OBC*s over time. However Table 2.4 shows that as compared to the work sample, *SC/ST*s are still little over-represented among the casual/regular wage earners as before, but *OBC*s are little under-represented.

Figure 2.4 and 2.5, provides the visual inspection of *IOP* and equalization of opportunity, among the three caste categories, for *MPCE* and wage respectively. The first panel of each figure plots the respective cumulative distributions for the different caste groups. Despite the heated debate regarding the classification of *OBC*, this panel provides a clear

[20]Figures 2.B.1 and 2.B.2 in Appendix 2.B draws the *CDF*s of *NSC* and *SC/ST* in the first panel, and that of *Gen* and *SC/ST* in the second panel, separately for *MPCE* and wage. The third panel of these figures plots the gaps in the distribution pairs of the other two panels, where the solid line corresponds to the gap between *NSC* and *SC/ST*, and the dotted line plots the same between *Gen* and *SC/ST*. Irrespective of the outcome, we always find the dotted gap curve to lie above the solid one, indicating the fact that *SC/ST*s are indeed more deprived when compared to the forward *General* castes.

| | ***Work sample*** | | | ***Wage Sample*** | | |
|---|---|---|---|---|---|---|
| | %Gen | %OBC | %SC/ST | %Gen | %OBC | %SC/ST |
| *1999-00* | 0.32 | 0.36 | 0.31 | 0.30 | 0.33 | 0.38 |
| *2004-05* | 0.30 | 0.41 | 0.29 | 0.31 | 0.37 | 0.31 |
| *2011-12* | 0.28 | 0.43 | 0.29 | 0.26 | 0.41 | 0.33 |

Table 2.4: Caste composition in post-reform India

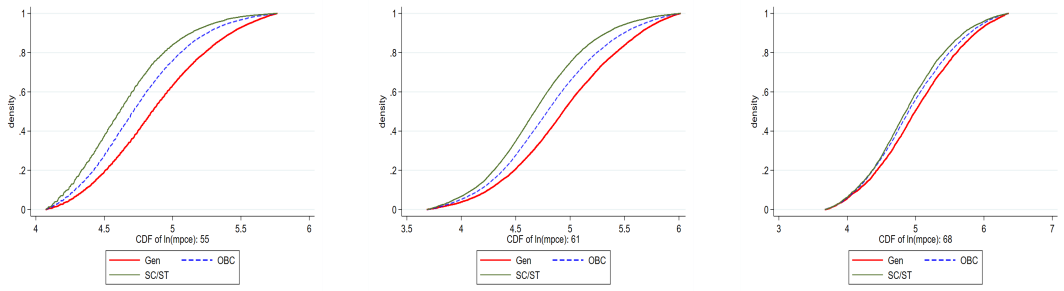
hierarchical order among the caste groups, that reveals *General* (Gen) as the most advantageous caste category till date. *OBC*s are visibly worse off than *Gen* and similar to Deshpande & Ramachandran (2014), in most of the cases the distribution of them is actually closer to that of the *SC/ST*s. However, *SC/ST*s remain the worst victim of casteism even until 2012. Especially for regular/casual wage earning, the forward category of *Gen* is way more advantageous than *SC/ST*, is spite of the over-representation of the latter in the wage sample. The same is statistically confirmed by the test of first order stochastic dominance, as shown by the first panel of Table 2.B.3 and 2.B.4 in Appendix 2.B. Irrespective of the outcome, the category *Gen* dominates both *OBC* and *SC/ST* at order one, whereas among the latter two, *OBC* is dominating *SC/ST*. However as before, the corresponding *CDF*s are only indicative of an opportunity equalization in *MPCE* for the latest survey year.

The entire time period of the analysis involving *OBC* (*1999-2012*) is in the post-reform phase of India. However similar to the previous analysis, it is better to refer the concerned time period in different policy-timezone. Likewise we designate the time span of *1999-2005* as the initial phase of reform and that of *2005-2012* as the later phase of reform. Figures 2.4 and 2.5 draws the respective gap curves and the second panel of Tables 2.B.3 and 2.B.4 in Appendix 2.B tabulates the corresponding results of the statistical test of gap curve dominance. First of all, Figures 2.4 and 2.5 shows that as compared to *1999-00* opportunity equalizes among most of the caste groups in *2011-12*, both for consumption and wage, which is similar to the trend we found in the previous analysis (*see* Table 2.3). However, by virtue of taking finer caste categorization we can further say that earning opportunity mostly equalizes in comparison with the upper-most caste group (*Gen*). While both *OBC*s and *SC/ST*s face lesser discrimination in *2011-12* as compared to *Gen*, over the same time period of *1999-2012* opportunity actually disequalizes between *OBC* and *SC/ST*. The statistical test results of Table 2.B.4 confirms that the null of

disequalization between the lower caste groups (*OBC* and *SC/ST*) can not be rejected over the post-reform time span of *1999-2012*.

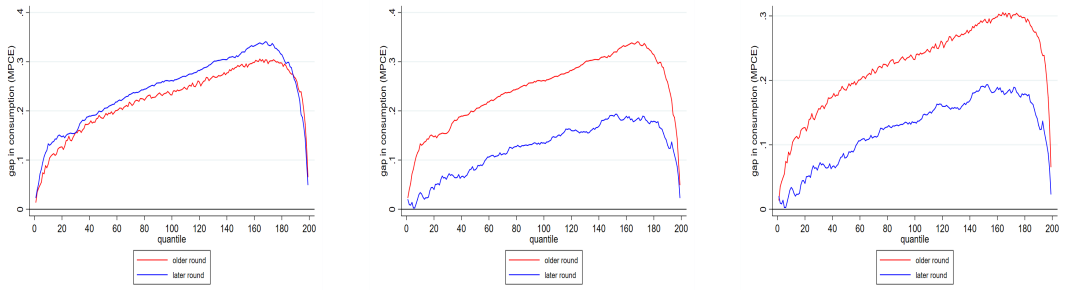
Similar to our previous analysis we see an opposite dynamics for consumption and wage equalization here as well. For *MPCE*, a small scale disequalization during the early phase of reform (*1999-2005*) had been more than offset by the equalization thereafter (*2005-12*). Whereas as compared to *General*, earning opportunity for both *OBC* and *SC/ST* had equalized during *1999-2005*, followed by a sizable disequalization thereafter. However the disequalization in the later phase of reform is not high enough to completely nullify the previous impact of equalization, so that over the entire post-reform period we find earning opportunity to equalize overall. However equalization of earning opportunity is not the same between the lower two caste categories of *OBC* and *SC/ST*, where a clear dominance of the former is already established over the latter (*see* first panel of Figure 2.5). The disequalization of earning opportunity among *OBC* and *SC/ST* in the later phase of reform is grave enough to annul the previous equalizing impact. However, unlike their respective gaps with the forward *General* caste category, the gap between themselves are much closer.

1-Distribution of MPCE



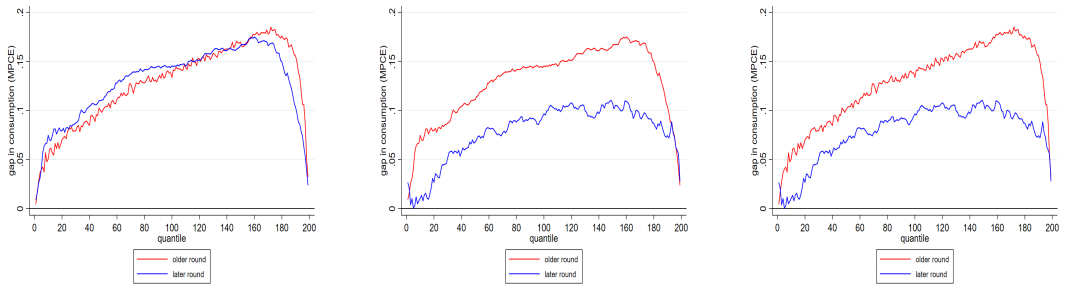

(a) 1999-00 (b) 2004-05 (c) 2011-12

2A-Gap curves: General Vs SC/ST



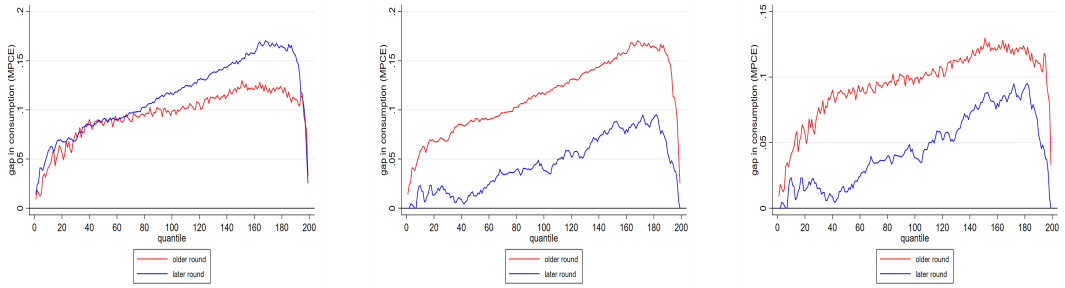

(d) 1999 to 2005 (e) 2005 to 2012 (f) 1999 to 2012

2B-Gap curves: General Vs OBC




(g) 1999 to 2005 (h) 2005 to 2012 (i) 1999 to 2012

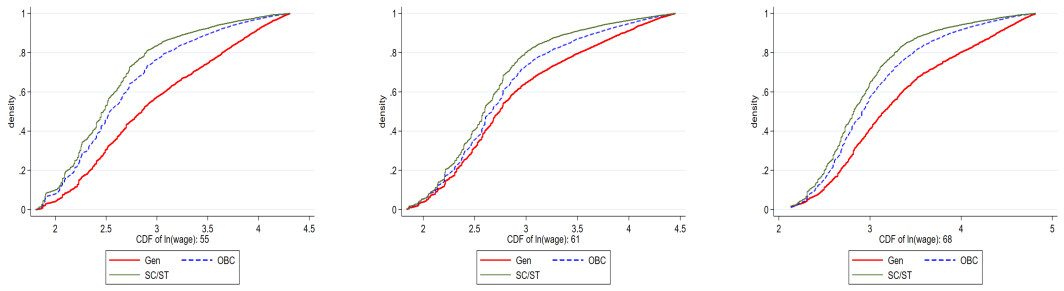
2C-Gap curves: OBC Vs SC/ST




(j) 1999 to 2005 (k) 2005 to 2012 (l) 1999 to 2012

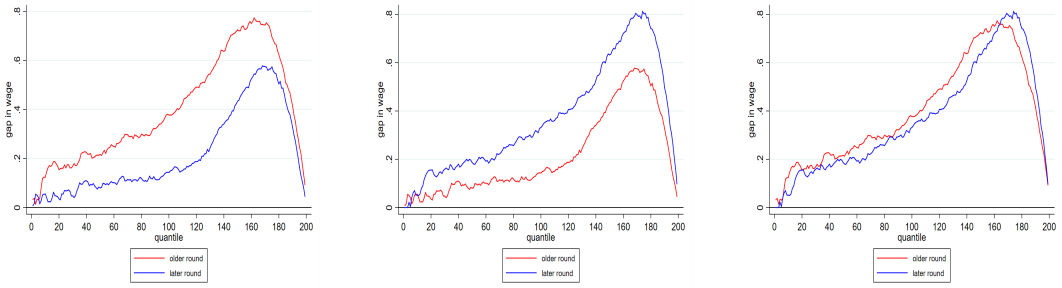
Figure 2.4: IOP and Equalization across castes: MPCE

1-Distribution of Wage



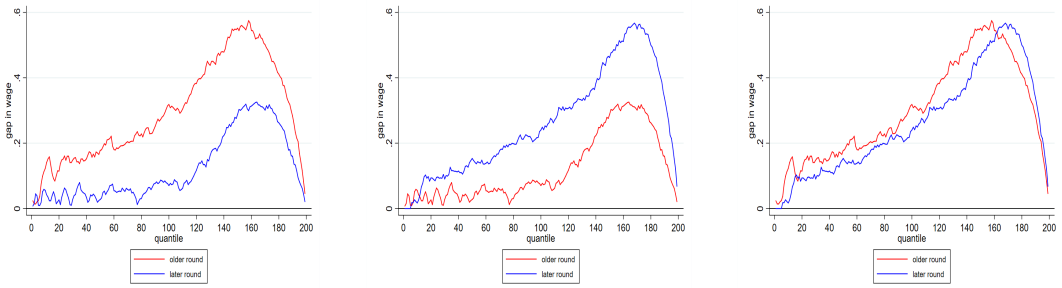

(a) 1999-00 (b) 2004-05 (c) 2011-12

2A-Gap curves: General Vs SC/ST



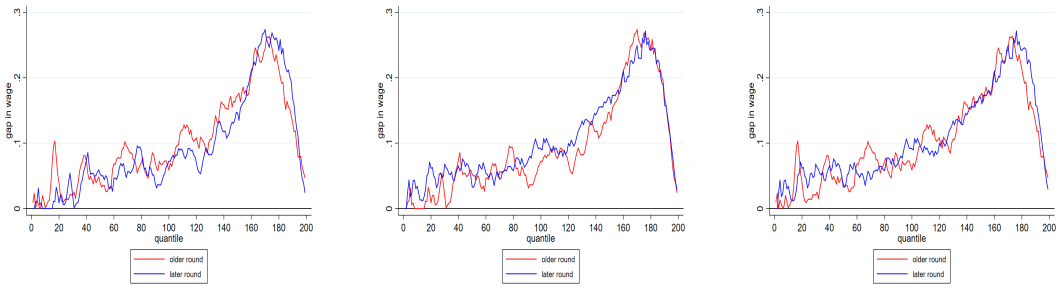

(d) 1999 to 2005 (e) 2005 to 2012 (f) 1999 to 2012

2B-Gap curves: General Vs OBC




(g) 1999 to 2005 (h) 2005 to 2012 (i) 1999 to 2012

2C-Gap curves: OBC Vs SC/ST




(j) 1999 to 2005 (k) 2005 to 2012 (l) 1999 to 2012

Figure 2.5: IOP and Equalization across castes: Wage

## *2.6 Concluding remarks*

Caste hierarchy is still a pertinent concern even for modern India. Since *1983*, the present analysis undoubtedly finds the historically disadvantageous caste category of *SC/ST* as the most dominated caste groups in India even until *2012*, in terms of both consumption (*MPCE*) and regular/casual wage earning. The relatively advantageous *non-SC/ST*s do enjoy an upper caste premium, although the extent of this undue advantage diminishes substantially for *MPCE*, especially over the time period of *2005-12*. The same can not said for regular/casual wage though. In fact, as compared to *1983* earning opportunity is actually lesser for the *SC/ST*s in *2011-12*.

India changes the centrally interventionist policy regime by adopting a host of neo-liberal open-market policies during early to mid-nineties. This phase of economic reform was much debated in the context of Indian economy as income and consumption inequality show a sharp increase since then. A separate post-reform analysis with finer caste categorization however reveals a relatively optimistic picture. We find economic opportunity to equalize among the caste categories for both *MPCE* and wage over the time span of *1999-2012*. Still a clear hierarchical order among the caste categories are very prominent even for the latest survey year (*2011-12*) that identifies the forward *General* caste category as the most advantageous group of castes who enjoys a substantial forward caste premium over the other deprived caste categories of *OBC* and *SC/ST*. Among the lower two caste categories, the dominance of *OBC*s over the *SC/ST*s identifies the latter as the worst victim of casteism in India with minimum economic opportunity. In fact the separate analysis on the post-reform period actually corroborates a disequalization of earning opportunity among the two deprived caste categories, although the distributions of *OBC* and *SC/ST* are much closer than that of the *General*. We also find the opportunity gap of the *SC/ST*s are higher when compared to the forward *General* caste category than when compared to the *non-SC/ST*s, which indicates that the degree of deprivation of *SC/ST* is indeed undervalued without the classification of the *OBC*s.

The dynamics of the equalization of caste based opportunity over time nevertheless differs for consumption and wage. Particularly since *1993-94*, while *MPCE* reveals a phase of equalization thereafter, wage reflects the lack of it. An elaborate temporal

analysis reveals that the deterioration of earning opportunity is particularly attributable to its huge disequalization impact right at the onset of neo-liberal economic reform (*1994-2000*). However, given the very different reporting of *MPCE* and wage data in *NSS*, these two outcomes are not really comparable *per se*. While the former is reported as the total monthly household expenditure for every enumerated households, the latter is only reported for the selected section of regular and casual wage earners who are not self-employed. Subject to that caveat we can still say that the lower caste regular/casual wage earners in *2012* have less earning opportunity than the relatively upper caste groups, which is not the case in terms of consumption expenditure. However, if we focus only on the last decade of twenty-first century, an overall improvement in both consumption and wage is visible among the caste groups.

# Appendices to Chapter 2

## *2.A Dominance test hypothesis*

All the dominance tests as mentioned in section 2.3.3, are implemented in the form of a discretization process, whereby the null hypotheses for each of the tests in (2.6) and (2.7), are constructed upon division of the entire outcome distributions in finite proportions. First, the respective test samples to be analyzed, are divided into $m$ population quantiles as, $0 < p_1 < \ldots < p_m < 1$, such that, for each pair of types, $(c, c') \in \phi_\pi$, the distributions of $\hat{F}_\pi^{-1}$ and $\hat{F}_\pi'^{-1}$, under social state $\pi$, can be represented by the sequence of $\left(\hat{F}_\pi^{-1}(p_1|c,e), \ldots, (\hat{F}_\pi^{-1}(p_m|c,e)\right)$ and $\left(\hat{F}_\pi'^{-1}(p_1|c,e), \ldots, (\hat{F}_\pi'^{-1}(p_m|c,e)\right)$, respectively. Similarly, the associated gap curves corresponding to the pair of types, $(c, c')$, can be represented by the sequence of $\left(\hat{\Gamma}(F_\pi^{-1}, F_\pi'^{-1}, p_1), \ldots, \hat{\Gamma}(F_\pi^{-1}, F_\pi'^{-1}, p_m)\right)$, for each of the social states, $\pi \in \Pi$. The order of dominance between the respective pair of distributions, is then tested on the basis of a joint dominance and neutrality test, at each point of the population quantiles, as described below.

To simplify the notations, let us write $\hat{F}_\pi^{-1}(p_i|c,e)$ and $\hat{F}_\pi'^{-1}(p_i|c,e)$, by $\hat{F}_\pi^{-1}(p_i)$ and $\hat{F}_\pi'^{-1}(p_i)$, respectively, for $i = 1, \ldots, m$. Further, let $\delta$ denote the sequence of difference vectors between a pair of distributions, among which the dominance test is to be performed, and let $\hat{\delta}$ be its empirical counterpart. Depending on the pair of distributions under concern, $\hat{\delta}$, in particular, can be defined by the following sequence of difference

vectors -

$$\text{Difference between types} \Rightarrow \begin{cases} & \hat{\delta}_\pi = \Big(\hat{\delta}_\pi(p_1), \ldots, \hat{\delta}_\pi(p_m)\Big) \\ \textit{where} & \hat{\delta}_\pi(p_i) = \hat{F}_\pi^{-1}(p_i) - \hat{F}_\pi'^{-1}(p_i) \end{cases} \tag{2.8a}$$

$$\text{Difference between gap curves} \Rightarrow \begin{cases} & \hat{\delta}_{(c,c')} = \Big(\hat{\delta}_{(c,c')}(p_1), \ldots, \hat{\delta}_{(c,c')}(p_m)\Big) \\ \textit{where} & \hat{\delta}_{(c,c')}(p_i) = \hat{\Gamma}(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p_i) - \hat{\Gamma}(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p_i) \end{cases} \tag{2.8b}$$

Where, $\hat{\delta}_\pi$ in equation (2.8a), is the sequence of difference vectors, denoting the difference in the distributions between *types* $c$ and $c'$, under the exogenous social state, $\pi$. Similarly, equation (2.8b) defines the sequence of difference vectors, $\hat{\delta}_{(c,c')}$, that denotes the difference between the gap curves in social states, $\pi_m$ and $\pi_n$, for the type-pair, $(c, c')$. Notice that the difference vectors render each of the tests in (2.6) and (2.7), to be executed as a joint equality and inequality test of $\delta = 0$ and $\delta > 0$[21].

To construct the null of the associated equality and inequality tests of (2.6), for ranking a pair of types within an exogenous social state, $\pi$, each of the $m \times 1$ sample vectors, corresponding to *types*, $(c, c')$, are stacked together in a $2m \times 1$ vector as $\hat{\Upsilon}_\pi = \Big((\hat{F}_\pi^{-1}(p_1), \ldots, \hat{F}_\pi^{-1}(p_m), \hat{F}_\pi'^{-1}(p_1), \ldots, \hat{F}_\pi'^{-1}(p_m)\Big)$. Let $\mathbf{R} = (I_m, -I_m)$, where $I_m$ is a $m$-dimensional identity matrix. Therefore, we can write the difference vector as, $\hat{\delta}_\pi = \mathbf{R}\hat{\Upsilon}_\pi$. Provided that $\hat{F}_\pi^{-1}$ and $\hat{F}_\pi'^{-1}$ are generated from independent process, we can write their respective asymptotic distributions as, $(\hat{F}_\pi^{-1} - F_\pi^{-1}) \sim \mathcal{N}(0, \frac{\Sigma_\pi}{n_\pi}))$ and $(\hat{F}_\pi'^{-1} - F_\pi'^{-1}) \sim \mathcal{N}(0, \frac{\Sigma_\pi'}{n_\pi'})$ (Beach & Davidson 1983, Andreoli 2018). If under social state, $\pi$, the respective sample sizes of the distributions of *types* $c$ and $c'$, are $n_\pi$ and $n_\pi'$, then the asymptotic distribution of $\hat{\delta}_\pi$ can be written as -

$$\hat{\delta}_\pi \sim \mathcal{N}(\mathbf{R}\Upsilon_\pi, \Omega_\pi) \quad \text{where,} \quad \Omega_\pi = \mathbf{R}\,\text{diag}\left(\frac{\Sigma_\pi}{n_\pi}, \frac{\Sigma_\pi'}{n_\pi'}\right)\mathbf{R}^T \tag{2.9}$$

where $\Upsilon_\pi$ is the corresponding population vector.

[21]The test of reverse dominance is simply rendered by writing the difference vectors as, $\hat{\delta}_\pi^r(p_i) = -\hat{\delta}_\pi(p_i) = \hat{F}_\pi'^{-1}(p_i) - \hat{F}_\pi^{-1}(p_i)$ and $\hat{\delta}_{(c,c')}^r(p_i) = -\hat{\delta}_{(c,c')}(p_i) = \hat{\Gamma}(F_{\pi_n}^{-1}, F_{\pi_n}'^{-1}, p_i) - \hat{\Gamma}(F_{\pi_m}^{-1}, F_{\pi_m}'^{-1}, p_i)$, for $i = 1, \ldots, m$, so that the reverse dominance test is equivalent to the inequality test of $\delta^r > 0$.

Andreoli et al. (2019) applied the asymptotic result of the above estimator (2.9), in the set up of gap curve dominance, for comparing equalization of opportunity between the *types*, $(c, c')$, across different social states. Let us consider two different social states as, $\pi = 0$ and $\pi = 1$, where the pair of *types*, $(c, c')$, are unequivocally ranked by the tests of (2.6), in each of the social states. Let, $n_0, n'_0, n_1, n'_1$, be the corresponding sample sizes for each type in each of the social state. To formulate the null hypotheses for the tests in (2.7), for ranking the social states, let us further stack $\hat{\Upsilon}_0$ and $\hat{\Upsilon}_1$, to get the $4m \times 1$ vector of $\hat{\Upsilon}_\Gamma = (\hat{\Upsilon}_0, \hat{\Upsilon}_1)$. If we write the $m \times 4m$ difference-in-difference matrix as, $\mathbf{R}_\Gamma = (\mathbf{R}, -\mathbf{R})$, then Andreoli et al. (2019) showed that the asymptotic distribution of $\hat{\delta}_{(c,c')}$ will be -

$$\hat{\delta}_{(c,c')} \sim \mathcal{N}(\mathbf{R}_\Gamma \Upsilon_\Gamma, \Omega_\Gamma) \quad \text{where,} \quad \Omega_\Gamma = \mathbf{R}_\Gamma \, \text{diag} \left( \frac{\Sigma_0}{n_0}, \frac{\Sigma'_0}{n'_0}, \frac{\Sigma_1}{n_1}, \frac{\Sigma'_1}{n'_1} \right) \mathbf{R}_\Gamma^T \tag{2.10}$$

where $\Upsilon_\Gamma$ be the corresponding population vector. The respective test statistics associated to the test of equality and dominance are described below[22].

**Testing equality:** The null and the alternative hypotheses for testing equality between a pair of distributions, divided in $m$ quantiles, associated to $\{p_1, \ldots, p_m\}$ are -

$$H_0 : \quad \delta = 0 \quad \text{against} \quad H_1 : \quad \delta \neq 0$$

Where, under this null hypothesis, the test-statistic, $T_E$, is a Wald type test-statistic and follows a Chi-square distribution with $m$ degrees of freedom. Thus[23] -

$$T_E = n \hat{\delta}^T \hat{\Omega}^{-1} \hat{\delta} \sim \chi^2_m$$

**Testing dominance:** For testing dominance between a pair of distribution, the null and alternative hypotheses are stated as -

$$H_0 : \quad \delta \in \mathbb{R}^+_m \quad \text{against} \quad H_1 : \quad \delta \notin \mathbb{R}^+_m$$

[22] In the test statistic, depending on the tests, (2.6) or (2.7), $\delta$ is either $\delta_\pi$ or $\delta_{(c,c')}$, whereas $\Omega$ can be either $\Omega_\pi$ or $\Omega_\Gamma$.

[23] Where $n$ denotes the respective total sample size, therefore $n = n_\pi + n'_\pi$ for ranking types and $n = n_0 + n'_0 + n_1 + n'_1$ for ranking social states.

The Wald test statistics with this positivity constraints, $T_D$, are shown to be asymptotically distributed as a mixture of $\chi^2$ distribution, say $\bar{\chi}^2$, by Kodde & Palm (1986) as follows[24] -

$$T_D = \min_{\delta \in \mathbb{R}^+_m} \{n(\hat{\delta} - \delta)^T \hat{\Omega}^{-1} (\hat{\delta} - \delta)\} \sim \bar{\chi}^2$$

where,

$$\bar{\chi}^2 = \sum_{j=0}^{m} w(m, m-j, \hat{\Omega}) \Pr(\chi^2_j \geq c)$$

where $w(m, m-j, \hat{\Omega})$ is the probability that $m-j$ elements of $\delta$ are strictly positive. Although $\bar{\chi}^2$ is not a fully tabulated distribution, Kodde & Palm (1986) provides the critical values of the test statistics, for some selected significance level. The null is accepted if it is lower than the lower bound and rejected if higher than the upper bound. In case the respective test-statistic value is between the upper and the lower bound, dominance is tested on the basis of $p$-values.

## *2.B Additional tables and figures*

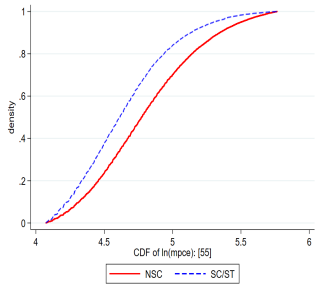
[24] Also *see*, Lefranc et al. (2009), Andreoli (2018), Andreoli et al. (2019).

| | | *1987-88* | *1993-94* | *1999-00* | *2004-05* | *2011-12* | *1993-94* | *1999-00* | *2004-05* | *2011-12* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ***MPCE*** | | | | | ***Wage*** | | | |
| *Neutrality* | | 147.6 | 167.1 | 45.8 | 93.1 | 1478.4 | 47.6 | 238.9 | 148.5 | 170.6 |
| | | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| *Equalization* | *1983* | 142.9 | 145.5 | 0.0 | 62.1 | 0.0 | 17.7 | 238.9 | 102.1 | 170.6 |
| | | (0.00) | (0.00) | (0.98) | (0.00) | (0.98) | (0.14) | (0.00) | (0.00) | (0.00) |
| *Disequalization* | | 1.2 | 8.6 | 45.8 | 30.6 | 1478.4 | 25.8 | 0.0 | 39.8 | 0.0 |
| | | (0.92) | (0.49) | (0.00) | (0.01) | (0.00) | (0.03) | (0.95) | (0.00) | (0.95) |
| *Neutrality* | | - | 16.6 | 132.8 | 27.7 | 1260.1 | - | - | - | - |
| | | - | (0.61) | (0.00) | (0.09) | (0.00) | - | - | - | - |
| *Equalization* | *1987-88* | - | 11.8 | 0.0 | 0.1 | 0.0 | - | - | - | - |
| | | - | (0.33) | (0.98) | (0.97) | (0.98) | - | - | - | - |
| *Disequalization* | | - | 4.5 | 132.7 | 27.1 | 1260.1 | - | - | - | - |
| | | - | (0.73) | (0.00) | (0.01) | (0.00) | - | - | - | - |
| *Neutrality* | | - | - | 145.9 | 35.9 | 979.3 | - | 152.9 | 53.3 | 124.9 |
| | | - | - | (0.00) | (0.01) | (0.00) | - | (0.00) | (0.00) | (0.00) |
| *Equalization* | *1993-94* | - | - | 0.0 | 0.3 | 0.0 | - | 152.9 | 34.2 | 124.9 |
| | | - | - | (0.98) | (0.95) | (0.98) | - | (0.00) | (0.00) | (0.00) |
| *Disequalization* | | - | - | 145.9 | 35.6 | 979.3 | - | 0.0 | 2.1 | 0.0 |
| | | - | - | (0.00) | (0.00) | (0.00) | - | (0.94) | (0.77) | (0.95) |
| *Neutrality* | | - | - | - | 61.8 | 769.6 | - | - | 103.0 | 27.6 |
| | | - | - | - | (0.00) | (0.00) | - | - | (0.00) | (0.09) |
| *Equalization* | *1900-00* | - | - | - | 61.4 | 0.0 | - | - | 0.0 | 0.7 |
| | | - | - | - | (0.00) | (0.98) | - | - | (0.95) | (0.87) |
| *Disequalization* | | - | - | - | 0.0 | 769.6 | - | - | 103.0 | 26.0 |
| | | - | - | - | (0.98) | (0.00) | - | - | (0.00) | (0.03) |
| *Neutrality* | | - | - | - | - | 883.5 | - | - | - | 79.7 |
| | | - | - | - | - | (0.00) | - | - | - | (0.00) |
| *Equalization* | *2004-05* | - | - | - | - | 0.0 | - | - | - | 79.7 |
| | | - | - | - | - | (0.97) | - | - | - | (0.00) |
| *Disequalization* | | - | - | - | - | 883.5 | - | - | - | 0.0 |
| | | - | - | - | - | (0.00) | - | - | - | (0.94) |

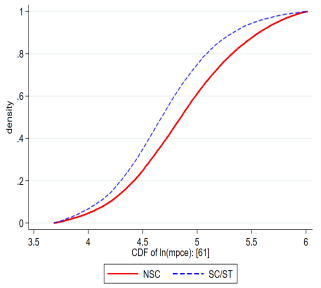
Table 2.B.1: *Equalization of opportunity* over all pairs of years: *MPCE* and *Wage*[a]

[a]Selected (Kodde & Palm 1986) critical values - (1.642, 26.625) [10%]; (2.706, 29.545) [5%]; (5.412, 35.556) [1%], in the form of (lower bound, upper bound). Reject the null if larger than upper bound, accept if lower than lower bound, conclude on the basis of p-values otherwise. 'Neutrality' tests for identical gap curves for different years. 'Equalization' tests the null that the gap curve corresponding to the older year dominates that of the latter year and 'disequalization' tests the reverse dominance.

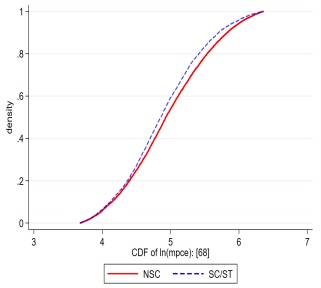
[A] CDFs: *NSC* and *SC/ST*




(a) 1999-00

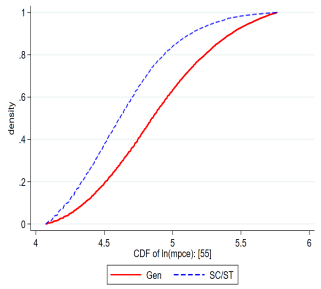



(b) 2004-05

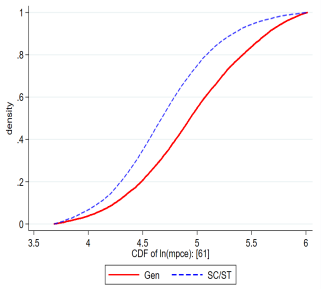



(c) 2011-12

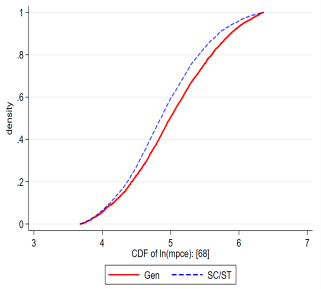
[B] CDFs: *Gen* and *SC/ST*




(d) 1999-00

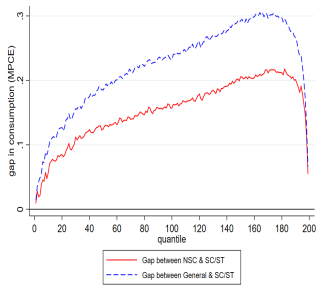



(e) 2004-05

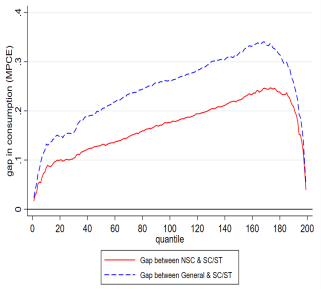



(f) 2011-12

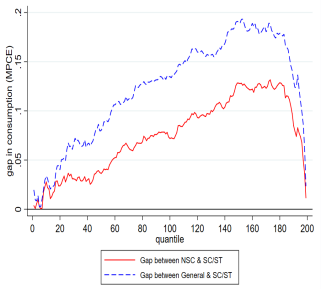
Gaps curves corresponding to [A] and [B]




(g) 1999-00



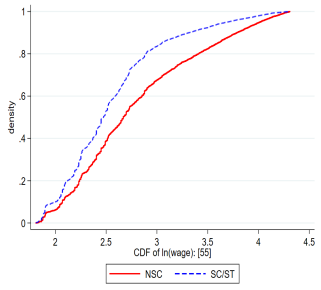

(h) 2004-05



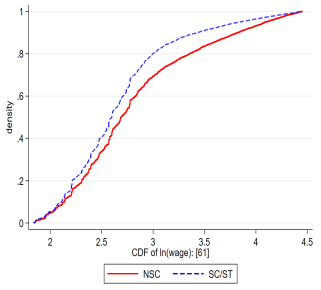

(i) 2011-12

Figure 2.B.1: Underestimation of *SC/ST* deprivation: *MPCE*

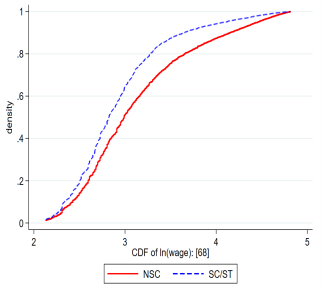
[A] CDFs: *NSC* and *SC/ST*




(a) 1999-00

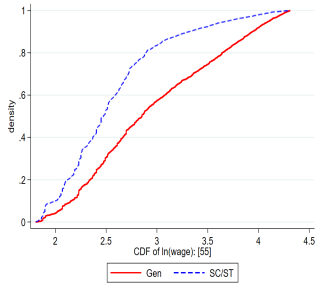



(b) 2004-05

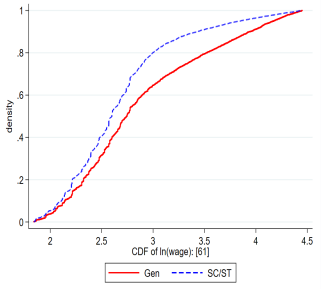



(c) 2011-12

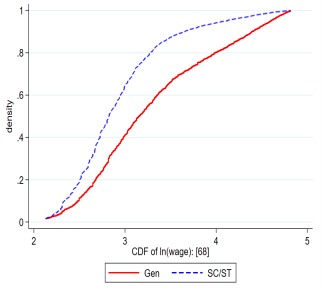
[B] CDFs: *Gen* and *SC/ST*




(d) 1999-00

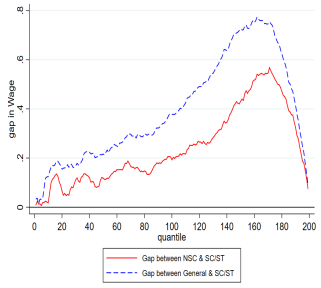



(e) 2004-05

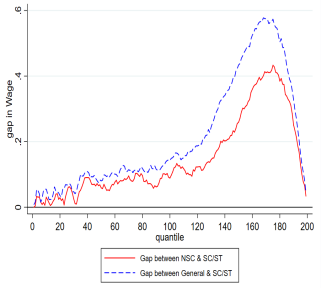



(f) 2011-12

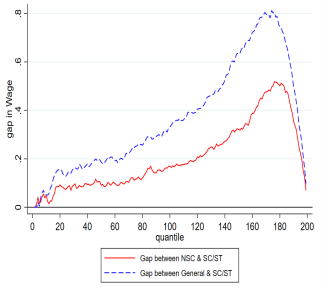
Gaps curves corresponding to [A] and [B]




(g) 1999-00




(h) 2004-05




(i) 2011-12

Figure 2.B.2: Underestimation of *SC/ST* deprivation: *Wage*

| | *MPCE* | $\%\triangle_{\text{mpce}}$ | *Wage* | $\%\triangle_{\text{wage}}$ |
|---|---|---|---|---|
| *1983* | 0.0336 | · | 0.0133 | · |
| *1987-88* | 0.0375 | ↑ 11.6 | · | · |
| *1993-94* | 0.0434 | ↑ 15.7 | 0.0197 | ↑ 48.1 |
| *1900-00* | 0.0380 | ↓ 12.5 | 0.0478 | ↑ 142.6 |
| *2004-05* | 0.0259 | ↓ 31.8 | 0.0193 | ↓ 59.6 |
| *2011-12* | 0.0042 | ↓ 83.7 | 0.0335 | ↑ 73.5 |

Table 2.B.2: Non-parametric *IOP* due to caste (non-SC/ST & SC/ST)[a]

[a] *IOP* for MPCE and wage are measured by the index of Mean Log Deviation, adopting the methodology of the relative non-parametric ex-ante measure of Checchi & Peragine (2010). $\%\triangle$ denotes the percentage change in *IOP* index over adjacent survey years, whereas '↑/↓' symbolizes for increase and decrease, respectively. All measures of *IOP* are estimated taking caste as the only circumstance, where caste has two categories, *non-SC/ST* and *SC/ST*.

| | ***Gen*** *Vs* ***SC/ST*** | | ***Gen*** *Vs* ***OBC*** | | ***OBC*** *Vs* ***SC/ST*** | |
|---|---|---|---|---|---|---|
| *Testing inequality of opportunity* | | | | | | |
| **1999-00** [1] | | | | | | |
| $\approx$ | 9211.3 | (0.000) | 3282.5 | (0.000) | 1900.1 | (0.000) |
| $\succ$ | 0.0 | (0.943) | 0.0 | (0.950) | 0.0 | (0.943) |
| $\prec$ | 9211.3 | (0.000) | 3282.5 | (0.000) | 1900.1 | (0.000) |
| **2004-05** [2] | | | | | | |
| $\approx$ | 6962.3 | (0.000) | 2023.7 | (0.000) | 1740.7 | (0.000) |
| $\succ$ | 0.0 | (0.959) | 0.0 | (0.964) | 0.0 | (0.951) |
| $\prec$ | 6962.3 | (0.000) | 2023.7 | (0.000) | 1740.7 | (0.000) |
| **2011-12** [3] | | | | | | |
| $\approx$ | 1042.7 | (0.000) | 456.1 | (0.000) | 285.7 | (0.000) |
| $\succ$ | 0.0 | (0.956) | 0.0 | (0.958) | 0.0 | (0.958) |
| $\prec$ | 1042.7 | (0.000) | 456.1 | (0.000) | 285.7 | (0.000) |
| *Testing equalization of opportunity* | | | | | | |
| **1999** to **2005** ([1] Vs [2]) | | | | | | |
| *Neutrality* | 49.6 | (0.000) | 32.7 | (0.026) | 313.3 | (0.000) |
| *Equalization* | 49.6 | (0.000) | 14.9‡ | (0.199) | 311.6 | (0.000) |
| *Disequalization* | 0.0 | (0.980) | 17.5‡ | (0.122) | 0.04 | (0.973) |
| **2005** to **2012** ([2] Vs [3]) | | | | | | |
| *Neutrality* | 1008.2 | (0.000) | 1142.4 | (0.000) | 933.5 | (0.000) |
| *Equalization* | 0.0 | (0.977) | 0.0 | (0.984) | 0.0 | (0.972) |
| *Disequalization* | 1008.2 | (0.000) | 1142.4 | (0.000) | 933.5 | (0.000) |
| **1999** to **2012** ([1] Vs [3]) | | | | | | |
| *Neutrality* | 784.5 | (0.000) | 984.1 | (0.000) | 696.8 | (0.000) |
| *Equalization* | 0.0 | (0.983) | 0.0 | (0.980) | 0.0 | (0.977) |
| *Disequalization* | 784.5 | (0.000) | 984.1 | (0.000) | 696.8 | (0.000) |

Table 2.B.3: Castes in post-reform India: MPCE[a]

[a] *Gen*, *OBC*, *SC/ST* are caste groups, ordered from the most to the least advantageous. '$\succ$' means the relatively advantageous caste category dominates the weaker caste category at order one, whereas '$\prec$' means the same in opposite order and '$\approx$' means equality in their respective distributions. Selected Kodde & Palm (1986) (lower bound, upper bound) critical values are - (1.642, 26.625) [10%], (2.706, 29.545) [5%], (5.412, 35.556) [1%]. Reject the null if larger than upper bound, accept if lower than lower bound, conclude on the basis of p-values otherwise. Columns marked with ‡ indicates that equalization of opportunity is inconclusive, in the sense that, no single test of neutrality, equalization or disequalization, can be rejected with $p > 0.5$.

| | ***Gen*** *Vs* ***SC/ST*** | | ***Gen*** *Vs* ***OBC*** | | ***OBC*** *Vs* ***SC/ST*** | |
|---|---|---|---|---|---|---|
| *Testing inequality of opportunity* | | | | | | |
| **1999-00** [1] | | | | | | |
| $\approx$ | 7079.6 | (0.000) | 3294.9 | (0.000) | 1701.8 | (0.000) |
| $\succ$ | 0.0 | (0.946) | 0.0 | (0.951) | 0.0 | (0.923) |
| $\prec$ | 7079.6 | (0.000) | 3294.9 | (0.000) | 1701.8 | (0.000) |
| **2004-05** [2] | | | | | | |
| $\approx$ | 2220.6 | (0.000) | 1161.8 | (0.000) | 1500.3 | (0.000) |
| $\succ$ | 0.0 | (0.930) | 0.0 | (0.935) | 0.0 | (0.921) |
| $\prec$ | 2220.6 | (0.000) | 1161.8 | (0.000) | 1500.3 | (0.000) |
| **2011-12** [3] | | | | | | |
| $\approx$ | 3676.6 | (0.000) | 1880.4 | (0.000) | 751.8 | (0.000) |
| $\succ$ | 0.0 | (0.939) | 0.0 | (0.947) | 0.0 | (0.913) |
| $\prec$ | 3676.6 | (0.000) | 1880.4 | (0.000) | 751.8 | (0.000) |
| *Testing equalization of opportunity* | | | | | | |
| **1999** to **2005** ([1] Vs [2]) | | | | | | |
| *Neutrality* | 158.7 | (0.000) | 390.2 | (0.000) | 24.2 | (0.190) |
| *Equalization* | 0.0 | (0.949) | 0.0 | (0.953) | 5.6 | (0.620) |
| *Disequalization* | 158.7 | (0.000) | 390.2 | (0.000) | 15.5 | (0.221) |
| **2005** to **2012** ([2] Vs [3]) | | | | | | |
| *Neutrality* | 159.5 | (0.000) | 327.4 | (0.000) | 33.2 | (0.023) |
| *Equalization* | 159.5 | (0.000) | 327.4 | (0.000) | 29.6 | (0.011) |
| *Disequalization* | 0.0 | (0.944) | 0.0 | (0.954) | 1.9 | (0.789) |
| **1999** to **2012** ([1] Vs [3]) | | | | | | |
| *Neutrality* | 40.1 | (0.003) | 38.3 | (0.005) | 23.9 | (0.197) |
| *Equalization* | 2.8 | (0.769) | 0.03 | (0.952) | 18.6 | (0.131) |
| *Disequalization* | 37.1 | (0.001) | 38.1 | (0.001) | 3.4 | (0.725) |

Table 2.B.4: Castes in post-reform India: Wage[a]

[a] *Gen*, *OBC*, *SC/ST* are caste groups, ordered from the most to the least advantageous. '$\succ$' means the relatively advantageous caste category dominates the weaker caste category at order one, whereas '$\prec$' means the same in opposite order and '$\approx$' means equality in their respective distributions. Selected Kodde & Palm (1986) (lower bound, upper bound) critical values are - (1.642, 26.625) [10%], (2.706, 29.545) [5%], (5.412, 35.556) [1%]. Reject the null if larger than upper bound, accept if lower than lower bound, conclude on the basis of p-values otherwise.

# Chapter 3

# Access to educational opportunity in the twenty-first century: An account of Indian children

## *3.1 Introduction*

The heterogeneous nature of inequality incited a scholarly debate in the late twentieth century to isolate the unethical part of inequality, by distinguishing between the *fair* and *unfair* inequality. Precisely, inequality was started to be assumed as the result of two broad classes of factors, one that are beyond any human control (the *circumstance* factor) and other that are subject to individual choice (the *effort* factors). Roemer (1993) quarantine the '*unfair*' part of inequality as *"inequality of opportunity"* (*IOP*), that is generated by the circumstance factors only. *Effort* factors on the other hand are considered as the legitimate source of inequality. Hence from an ethical standpoint no one in the society should face unequal distribution in any economic advantage due to differences in their fatalistic circumstances. This responsibility sensitive viewpoint is rather crucial for children who are often seen to face discriminatory access to basic schooling or health care due to differences in their habitation, parental backgrounds and many other circumstances on which they have no control. The aim of this paper is to see how far children have equitable access to basic educational opportunities in India, a country embodying world's one of the largest youth population.

Although there is no universal definition of circumstances and efforts, educational environment of the family has often turned out to be as one of the major impediment to

success. In fact parental education has often found out to be the main responsible circumstance in generating unequal opportunity in many economic advantages, for a number of developed and developing country. For example, as compared to gender, race, birthplace and family wealth, *IOP* in Brazil is found to be largely driven by parental education (Bourguignon et al. 2007). In fact, for some Latin American countries, mother's education sometimes seems to be more important than that of fathers (Ferreira & Gignoux 2011). For Italy (Checchi & Peragine 2010) and several other western countries[1], father's education and occupation was repeatedly shown to have non trivial effect on generating substantial degree of inequality of opportunity in the society (Lefranc et al. 2008). Dardanoni et al. (2005) explores how parental background affect individual outcome through separate channels for USA & Britain. As for developing countries, demanding data requirement often limit the study of inequality of opportunity in these countries. However, parental education again, shown as one of the main component of high inequality of opportunity for countries like Africa (Cogneau & Mesplè-Somps 2008) and India (Chapter 1).

While all of the above mentioned studies on *IOP* are on the working adult population of a society, Paes de Barros et al. (2009) analyzed unequal opportunity among children in their access to basic opportunities. Following the old adage of 'prevention is better than cure', they argue that many of the resulting inequality of opportunity can be traced back to the formative child-age period and therefore can be taken into account in the scheme of early childhood policy design. If all children in that society have equal access to basic opportunities irrespective of their background, a society will more prone to be equal from the standpoint of responsibility sensitive egalitarian justice. Following similar line of argument, access of children to basic opportunities, like basic education, health, water supply, electricity or sanitation, are investigated for selected countries of Latin America & Caribbean (Molinas Vega et al. 2011, Molina et al. 2013) and Sub-Saharan Africa (Dabalen et al. 2015). Notice that as compared to adults the set of circumstances is likely to be larger for children, as they can not be held responsible for a broad class of social and family backgrounds, like household earning, sanitation, location, parental education or occupation, along with their own caste, race, sex or religion. Nevertheless

[1]Belgium, France, Great Britain, Netherlands, Norway, Sweden.

parental background again turned to be the most important circumstance variable for generating unequal opportunity among children as well.

Right to education has been considered as one of the basic rights to be ensured across the world. India embodies one-fifth of the world population and is the largest in terms of youth population. In the complex stratified Indian society, fortune transmits from older to younger generation through several channels. The century old caste system is still functional even in the twenty-first century, with the lowest 'untouchable' castes forming the bottom layer in most of the cases. In spite of taking a number of affirmative policies for economically backward castes and classes, caste discrimination in the labor market is rampant (Thorat 2008, Madheswaran & Attewell 2007). Also access to basic immunization of children in India is found to be impaired by varying caste, religion and parental attributes (Singh 2011). The mutual caste based networks are often seen to dissuade the lower caste people from better quality schooling choice in their childhood that eventually leads to low-paid traditional jobs (Munshi & Rosenzweig 2009). Not only children, a very high degree of educational opportunity is also prominent for Indian adults as well. Chapter 1 found more than one-fourth of total educational inequality for working adults in India is due to unequal opportunity generating from differential caste, sex, region and parental backgrounds.

However literature on *IOP* in India is very thin especially for analyzing inequality of opportunity among children, with two notable exceptions. Singh (2011) and Singh (2012*a*) found considerable inequality of opportunity for children in their access to immunization and primary education, respectively. Other than that there is no work on India to our knowledge that explore unequal opportunity for children. The present work aim to fill this gap by analyzing access to educational opportunity for Indian children with more details, that encompasses unequal access to elementary as well as post-elementary schooling using the National Sample Survey, which is the biggest nationally representative micro-data on India and a database different from that of the few existing studies. Besides by virtue of taking recent survey years into account, our analysis delivers a more contemporary picture of Indian children over the time span of *2004-12*.

Borrowing from the methodological set up of Paes de Barros et al. (2009), we formulate children's access to basic education as a simple binary outcome variable that takes the value 1 if children have access to the outcome and 0 otherwise. In this set up, *IOP* is estimated as the unequal probability of access to basic education for children with varying circumstances, which is measured by the *dissimilarity index* (D-index). Being a probability measure the D-index typically lies between $[0, 1]$, although often expressed in terms of percentages. The closer the D-index towards 100% the higher is the *IOP* in the society. One of the particular advantage of using the D-index is its amicability to treat binary dependent variables in quantifying *IOP*, which is not the case with the other widely used non-parametric or parametric method. However the probability of access to basic opportunities not only depends on differences in individual circumstances of the children, but also on the distribution or availability of those basic facilities in the society. We therefore also adopt the *Human Opportunity Index* (HOI) from Paes de Barros et al. (2009) that accommodates both the extent of available opportunities for children in the society, as well as the equitable distribution of those opportunities among children with differential circumstances. In particular, *HOI* is often interpreted as an opportunity sensitive development index in the sense that, it increases with better provision of basic opportunities in the society but decreases with the unequal distribution of those opportunities across children from various circumstances.

Every children older than 6 years are supposed to be enrolled in some educational institution in India irrespective of their circumstances and with a steady continuity, should have finished the elementary eight years of schooling in their early teenage. Although we found lesser *IOP* in terms of early age school enrollment, it increases as the children ages and children from lesser educated parents are always less likely to finish the basic eight years of schooling on time. A similar pattern is reflected in terms of school attendance as well. The chances of attending school at the onset of schooling is not very differentiated for children with diverse backgrounds, but drops significantly for older children from lesser educated families, which is exactly the opposite trend as found in the Sub Saharan African countries (Dabalen et al. 2015). A separate analysis of post-elementary schooling reveals even higher *IOP* in access to education among the relatively older children. Together the observations indicate the pertinent problem of high rate of school drop outs in India even

in the twenty-first century, especially for the lesser educated rural agro-based families. In addition, a regional picture reveals that access to basic education for children seems less impaired by the differences in their circumstances in the Southern and Western part of the country.

The remaining of the paper is organized as follows. Section 3.2 presents a brief introduction to the schooling system in India. The theoretical and methodological background is then provided in section 3.3, followed by a detailed description of our database and sample selection in section 3.4. Section 3.5 presents our results, separately for elementary education at national and regional level, along with a brief discussion on the access to post-elementary education. Section 3.6 concludes.

## *3.2 School education in India*

Modern India follows a twelve year school education system in most part of the country. First eight years of schooling consists of elementary level education, followed by four years of secondary level schooling. The entire elementary education is further divided into the lower primary (grade 1 to 4) and upper primary (grade 5 to 8) education in most of the states. Similarly, secondary schooling is also divided in lower secondary (grade 9 and 10) and higher secondary education (grade 11 and 12). The lower primary education is more commonly known as the primary education, whereas secondary education usually refers to the lower secondary schooling. Beyond higher secondary schooling, education is pursued at college level and falls in the category of tertiary education.

India have seen considerable educational improvement since independence in 1947. While little more than 10% of the national population was literate according to the census of 1951, the literacy rate is more than 70% by 2011 census (Dutta & Sivaramakrishnan 2013). Nearly 95% of children aged between 6-14 years are enrolled in elementary schooling, according to the latest survey of Annual Status of Education report (ASER) in 2018. Further, the same report shows that the share of girl children without any formal schooling have dropped by 5% in 2018 as compared to 2006. Increase in female literacy is in fact rather prominent for rural India, albeit the male-female literacy gap is yet visible in the twenty-first century and is lesser in urban sectors of the country (Govinda &

Biswal 2006). The latest all India school education survey by National Council of Educational Research and Training (NCERT) shows impressive improvement in educational infrastructure of the country by 2009. Over 90% of the rural households have access to a primary school within 1 km of the location of their household, whereas over 80% have a secondary schooling facility within 3 km of their habitation. Further as compared to the *non-scheduled* upper castes, schooling facilities at least upto the lower secondary level have similar distribution for the historically disadvantageous poor caste groups of Scheduled Castes and Scheduled Tribes ($SC/ST$) as well, who live in relatively less developed villages.

According to the World Bank data, the Government expenditure on education is 3.8% of GDP in India during 2012, which is higher than all of its neighboring countries like Pakistan, Bangladesh, Sri Lanka and Nepal. Since the formation of the Constitution of India in 1950, several initiatives have been taken to improve the educational infrastructure of the country that is reflected by a consistent improvement of literacy rates in the country. One of the major Government initiative in the twenty-first century is the Right of Children to Free and Compulsory Education Act, also known as the Right to Education (RTE) act that came into effect in 2010. By virtue of the enactment of this act by the Supreme court of India, eight years of elementary education is not only made compulsory but all children aged between 6-14 years are entitled to free elementary school education. The RTE act comes with the additional compliance of adequate pupil-teacher ratio, toilets for girls, clean drinking water, library, prohibition of detention and no discrimination (Krishna et al. 2017). However less than 10% of Government schools are compliant with all the RTE norms (Rai 2014).

The free education mandate is for all Government sponsored schools that includes state and central Government schools as well as schools run by the local bodies like town municipalities. But over the first decade of the twenty-first century India has also seen an increase in the fee-charging private school enrollment, especially in the urban parts. Even 30% of the rural households are found to enroll their children in the private schools as well because of the questionable 'quality' of education in the free public schools (Muralidharan & Kremer 2006). Albeit more than 60% of Indian children are still enrolled in the free public Government schools according to the National Sample Survey in 2011-12.

In any type of schooling, all students upto the lower secondary level (grade 10) follow a compulsory syllabus that includes literature, science, history, geography and other school-specific courses, if any. Whereas at the higher secondary level (grade 11 and 12), students have to choose a core area that can be science, arts, humanities, commerce or other optional courses offered by the institute. Therefore not unexpectedly, the provision of higher secondary schooling is often limited as compared to the elementary and lower secondary schooling. Nearly 25% of the rural households do not have a higher secondary school within 8 km of their residence even in 2009 and the figure is nearly 50% for the *Scheduled Tribe* population who lives in relatively remote areas (NCERT 2016). In spite of ample evidence on higher return to secondary and tertiary education, less than half of the lower secondary schools in rural India have the provision of higher secondary education as well (ASER 2007).

## *3.3 Theoretical background*

### 3.3.1 Inequality of opportunity in childhood

The disposition of *IOP* is to ensure equal opportunity across all individuals in the society irrespective of their circumstances. Differential social or economic background that are beyond the hold of subjective responsibility should not be correlated with favorable or unfavorable economic returns. Let $z$ denote a desirable economic advantage with the corresponding distribution of $F(z)$ and let $(c, c')$ denote a pair of different circumstances such that $c \neq c'$. Then for a given level of effort, the principle of *IOP* by Roemer (1998) requires that the following condition must hold true to generate an opportunity equal society -

$$F(z|c) - F(z|c') = 0 \quad \text{when} \quad c \neq c' \tag{3.1}$$

The claim of the above condition is that the distribution of any advantage should be independent of uncontrollable circumstances, which however is never the case in reality and equation (3.1) therefore holds as an inequality. But the objective of a benevolent egalitarian policymaker should be to compensate for this inequality as much as possible so as to advance the society towards equality of opportunities. So strictly from the perspective of responsibility sensitive egalitarian justice, the society is not liable to compensate for

any opportunity gap created by the lack of education, as pursuing adequate education is often subject to individual choice and ability.

However education typically begins at childhood at the care of parents or other concerned family members. Even though consistent schooling is subject to children's own performance and aspiration, a large part of that comes from the inherent family characteristics (Heckman 2011). For example in hope of guaranteed jobs related to the caste based mutual networks, children from minority lower caste households in rural India are often found to prefer local language schools with questionable quality, even though they can afford a better one with more promising long-term benefits (Munshi & Rosenzweig 2006). Since the need and scope of persistent quality schooling is often shaped by family or social attributes, not only during childhood but also as young adults, one should not be held completely responsible for not pursuing enough schooling. Particularly children are surely not culpable if their environment or living condition makes them to prefer out adequate education in their formation period. In fact society should actually compensate for the resulting opportunity gap that is created by the insufficient human capital formation during childhood.

But childhood schooling is related to varying level of skill and cognitive formation and the lack of it can never actually be compensated *per se*. There is a body of literature showing evidence of positive impact of early child care intervention program on lifetime outcome for developing as well as developed countries[2]. Especially for human capital formation, the return to early intervention are actually more effective for the disadvantageous class of the society than compensating economic outcomes during their adulthood (Heckman 2006). To compensate for the opportunity gaps among the castes, Indian constitution provides reservations in several aspects for the backward class and castes in the society. But in spite of that, the deprived castes of *SC/ST*s are always under-represented in white collar Government jobs even today and one of the main reason for this is their lack of necessary skill or education required for the job. As a result those reserved 'seats' are often remain vacant at the cost of other potentially skilled applicants, which makes the system inherently inefficient.

[2]See for example, Currie (2001), Campbell et al. (2002), Heckman (2006, 2011), Golsteyn et al. (2014), Havnes & Mogstad (2015), Andreoli et al. (2019).

Indeed the share of *SC/ST*s in higher educational institute are far behind the declared reserved quota for them (Weisskopf 2004), whereas they are not so differentiated as compared to the *non-SC/ST*s in terms of primary school enrollment. One of the major reason for higher drop out rates among the disadvantaged marginalized people is that the effect of education is not immediately visible. However early intervention in childhood education is also found to make the children more future oriented to understand the value of education relatively early in their life that eventually generates long-term benefit (Golsteyn et al. 2014). Therefore Paes de Barros et al. (2009) propose that a more efficient way to equalize opportunities in the society is to avert the calamity from the beginning, so that every children have equal access to some *basic opportunities* in that society, irrespective of their circumstances.

Hence Paes de Barros et al. (2009) reformulate the principle of *IOP* as provided by equation (3.1) for children, in terms of equal probability of accessing an economic advantage across every children from diverse circumstantial backgrounds. In this context, the economic advantage, $z$, is called as the *basic opportunities*, that can be education, health care, sanitized residence, clean water, electricity or any such thing that can potentially hamper the 'childhood' of a child. Equal opportunity then indicates equal probability of access to the basic opportunities for children across varying circumstances and in our case, it is equal access to elementary/lower-secondary education for all eligible children in India irrespective of their individual circumstances.

### 3.3.2 Methodological framework

To measure access to basic educational opportunity for Indian children and to track the overall development of educational facility in the country, we adopt the method proposed by Paes de Barros, Ferreira, Molinas-Vega & Saavedra-Chanduvi (2009). They measure *IOP* in terms of the differences in the probability of access to basic education for children, conditional on different circumstances, by using the *Dissimilarity index* and the *Human Opportunity Index*[3]. The method is described as below.

[3]There is a body of literature that found evidence of *IOP* among children on several aspects like basic education, sanitation or health immunization, adopting the methodology of Paes de Barros et al. (2009). *See* for example Molinas Vega, Paes de Barros, Saavedra & Giugale (2011), Molina, Narayan & Saavedra-Chanduvi (2013) for selected Latin American countries, Dabalen, Narayan, Saavedra-Chanduvi & Hoyos Suarez (2015) for Sub-Saharan Africa, Singh (2011) for India (with a different database and

Let $y$ be the outcome variable that is dichotomous in nature, such that -

$$y = \begin{cases} 1, & \text{if children has access to a certain advantage} \\ 0, & \text{otherwise} \end{cases}$$

Provided the dichotomous nature of the variable, the empirical estimation of the outcome conditional on a set of circumstances is simply the estimated conditional probability of access, such as, $\hat{p}_i = E(y_i|\mathcal{C}_i)$. So assuming a linear relationship between the circumstance and outcome variables, $\hat{p}_i$ for individual $i$ can be estimated from the following logistic model for a set of $m$ circumstances, $\mathcal{C}_i \in \{c_{1i}, \ldots, c_{mi}\}$, as -

$$ln\left(\frac{P(y_i = 1|\mathcal{C})}{1 - P(y_i = 1|\mathcal{C})}\right) = \beta_0 + \beta_1 c_{1i} + \ldots + \beta_m c_{mi} \tag{3.2}$$

Therefore the estimated probability of access is given by -

$$\hat{p}_i = \frac{exp\left(\hat{\beta}_0 + \sum_{k=1}^{m} \hat{\beta}_k c_{ki}\right)}{1 + exp\left(\hat{\beta}_0 + \sum_{k=1}^{m} \hat{\beta}_k c_{ki}\right)} \tag{3.3}$$

Hence the average probability of access to a particular advantage across all the individuals in the society can be estimated from -

$$\bar{p} = \sum_{i=1}^{n} w_i \hat{p}_i \tag{3.4}$$

Where $w_i$ is the share of sample $i$ in the population.

Provided the estimates of $\{\hat{p}_i, \bar{p}\}$ inequality of opportunity can then be measured by the *Dissimilarity index* (D). Paes de Barros et al. (2008) showed that once $\{\hat{p}_i, \bar{p}\}$ are estimated, a consistent estimator for the D-index is[4] -

$$\hat{D} = \frac{1}{2\bar{p}} \sum_{i=1}^{n} w_i |\hat{p}_i - \bar{p}| \tag{3.5}$$

outcome than the present study), Saidi & Hamdaoui (2017) for Tunisia. For a broader cross-country level analysis see Hoyos & Narayan (2012) and Balcázar, Narayan & Tiwari (2015).

[4]The asymptotic variance of $\hat{D}$ is provided in Paes de Barros et al. (2008).

Due to the binary nature of the outcome variable one can write, $\bar{p} = E(y = 1|\mathcal{C})$. So when $\hat{p}_i = \bar{p}$, everyone in the society have equal access to the advantage irrespective of their circumstances and the $D$-index has a value of zero. Any positive $D$-index is therefore indicative of *IOP*, in the sense that the chance of getting access to the basic opportunity varies across differences in circumstances.

However *IOP* measured by the differential access probability gaps crucially depends on $\bar{p}$, that measures the average distribution of basic opportunities in the society and higher the $\bar{p}$ better developed is the society in the provision of basic opportunities. Paes de Barros et al. (2009) therefore further propose the *Human Opportunity Index* (HOI) that can take account of both the provision of basic opportunities in the society as well as their equitable distribution among children from different circumstances as follows -

$$HOI = \bar{p}\left(1 - \hat{D}\right) \tag{3.6}$$

HOI is therefore often interpreted as a responsibility sensitive development index, that along with measuring development also quantifies the degree of 'punishment' for development, due to the existing unequal opportunity in the society. Had there been equal access for children in the society ($D = 0$), HOI is then only determined by the provision of basic opportunities in the society ($\bar{p}$). But given $\bar{p}$, if there is unequal probability of access to basic opportunities for children with varying circumstances, HOI falls below $\bar{p}$. The presence of *IOP* in the society therefore takes a toll on development as measured by the 'penalty' of, $\bar{p}D$. Being a probability measure, both D and HOI lies between 0 and 1, but commonly expressed as percentages. While an increase in HOI is indicative of better development of the society overall, an increase in D indicates that the society is more opportunity unequal for the children.

So we can analyze the HOI index by dismantling it into two parts - the development part (captured by $\bar{p}$) and the *IOP* part (captured by $D$), where the former is also referred as the *coverage* of basic opportunities in the society. Notice that an increase in HOI can be brought about by either an increase in $\bar{p}$ or a fall in D or both. From the perspective of development, either of the above is an Pareto improvement in the sense that the society is anyway better in terms of more provision of basic opportunities or in terms of more

equal distribution of them or both. But from the perspective of *IOP* it is only the fall in the D-index that is of concern as far as opportunity equalization is the main objective of the society.

Since a change in HOI can be brought forward by either of $\bar{p}$ or $D$, we use the decomposability property of the HOI index to track how much of the change in the value of HOI is due to the change in coverage (the so called *scale effect*) and how much of it is due to the change in *IOP* (the so called *distribution effect*). Consider two different *social states*, $\pi \in \{0, 1\}$, (*e.g.* two countries or one country at two time frames/policy regimes). Then the difference in HOI between the social states, $\pi \in \{0, 1\}$, is -

$$\Delta_{HOI} = HOI^1 - HOI^0 = \bar{p}^1(1 - \hat{D}^1) - \bar{p}^0(1 - \hat{D}^0) \tag{3.7}$$

By adding and subtracting $\bar{p}^1(1 - \hat{D}^0)$, *equation* (3.7) can be decomposed into a scale and a distribution effect as following (Paes de Barros et al. 2008) -

$$\Delta_{HOI} = \underbrace{(1 - \hat{D}^0)(\bar{p}^1 - \bar{p}^0)}_{\text{scale effect}(\Delta_p)} + \underbrace{\bar{p}^1(\hat{D}^0 - \hat{D}^1)}_{\text{distribution effect}(\Delta_D)} \tag{3.8}$$

While the scale effect captures the change in the coverage of opportunities, the distribution effect captures how the distribution of the existing opportunities changes, while moving from *social state* $\pi = 0$ to $\pi = 1$.

It is worth mentioning, that both the D-index and the HOI are distribution sensitive as they give relatively higher weights to the distribution of the under privileged circumstances. However the Pareto-consistency of HOI do imply that an improvement in its value is nevertheless associated to an increased and likely more equitable set of opportunities in the society. But the impact is higher if the improvement in opportunity benefits the hitherto disadvantageous sector more (Paes de Barros et al. 2008).

## *3.4 Data, variables and sample selection*

### 3.4.1 Data

We use the biggest micro database for India for the present analysis, that of the National Sample Survey *NSS*. In particular we take two successive rounds of the employment-unemployment survey of *NSS* covering the survey years *2004-05* and *2011-12*. The survey covers the entire India except some remote inaccessible parts enumerating over half-a-million individuals[5]. These two survey years cover about 100,000 households from all over India enumerating 0.4 to 0.6 million Indian nationals. Initially, we have to drop about 1000 observations per round, to clean for valid age, sex, sector, caste specification and some other important attributes[6].

The survey provides detailed information on every household member on several important demographic aspects. For the present analysis with children we need a good deal of information regarding their parental background. However *NSS* has no direct provision of parental information in their questionnaire, instead the data is only available for households where the children is enumerated along with his/her parents for sharing the same households. For adults this may raise the issue of selectivity bias due to the adult inter-generational co-residence, but is not problematic for the present analysis, as children are likely to live with their parents under normal conditions. Apart from the focus on twenty-first century, another interesting aspect of our chosen time frame is that it can capture well the effect of a significant change in the Indian education policy for children. As mentioned before, education have legally been made free and compulsory for all Indian children aged between 6-14 years, by the latest amendment to the Right to Education Act (*RTE*, 2010). Although we can not estimate the direct effect of this policy due to the limitation of hard data related to the policy, our chosen time frame of *2004-12* is still reflective of the overall efficacy of this policy in equalizing educational opportunities across children from different circumstances.

[5]This means we have taken *Schedule 10.0* survey of *NSS* for the round of 61st (2004-05) and 68th (2011-12). Conflict areas of Ladakh & Kargil districts of Jammu & Kashmir, some remote interior villages of Nagaland, few unreachable areas of Andaman & Nicobar Islands and those villages recorded as uninhabited by the respective population census, are kept out of these surveys.

[6]*See* the data appendix A for further details about the *NSSO* employment-unemployment database, including data cleaning.

### 3.4.2 Definition of variables

#### Outcome variables

For the present analysis we concentrate on access to educational opportunity for Indian children. Since the level of school education varies with the age of the children, we execute our analysis on separate level of schooling suitable to different age cohorts, as shown by Table 3.1. We choose two broad batches of outcomes, one for analyzing access to elementary education and another for the post-elementary education (to be specific, lower secondary education). Each batch of outcomes consists of two kind of different outcomes, one that deals with the timely beginning and/or finishing of a certain level of schooling and another that takes into account the school attendance of children for age cohorts suitable to that level of schooling.

Table 3.1: List of outcome specifications

| Outcome | Sample age cohort | Grade specific criteria |
|---|---|---|
| **Elementary education** | | |
| Starting elementary education on time | 6-7 years | enrolled in *grade 1* or above |
| Completing elementary education on time | 14-15 years | finished *grade 8* or above |
| School-attendance: younger cohort | 6-10 years | attending any schooling from *grade 1* |
| School-attendance: older cohort | 11-15 years | attending any schooling from *grade 1* |
| **Post-elementary education** | | |
| Completing lower secondary education on time | 16-18 years | finished *grade 10* or above |
| School-attendance: adolescent cohort | 16-18 years | attending any schooling from *grade 1* |

Provided the grave importance of basic education for all children, our focus is rather skewed for the analysis of access to elementary education, which is the first eight years

of schooling. Borrowing from Dabalen et al. (2015) we consider the timely beginning and finishing of elementary level schooling as a proxy for the quality of elementary education. In particular we want to see whether all children of age 6-7 years have started formal schooling of grade-1 or above and whether those of age 14-15 years have reported to finish at least grade-8 of formal schooling, irrespective of their circumstances. To complement our analysis on elementary education we also consider the effect of circumstances on school attendance within the eligible age limit of elementary schooling (6-15 years) in two different age cohorts, the 'younger' cohorts consists of 6-10 years old children and that of the 'older' cohort consists of 11-15 years old.

As mentioned in section 3.2, since most of the elementary schools also have the provision of lower secondary level schooling up to grade-10, we further analyze to what extent unfavorable circumstances of the children determine their continuation of school education beyond the elementary level. For this we consider whether all children of age 16-18 years have finished at least grade-10 of formal schooling. Similar to the analysis of elementary education we additionally consider the effect of circumstances on school attendance for this group of children aged between 16-18 years, whom we call as the 'adolescent' cohort to distinguish them from the younger and older cohort mentioned before.

**Circumstance variables**

As compared to adults the set of circumstances for children is usually larger as children can not alter a number of household and social attributes they are born with. For the present analysis we choose a broad range of circumstances including parental backgrounds, social attributes along with some family characteristics. Table 3.2 lists our full set of circumstances along with their respective categories. Among parental attributes, father's and mother's education are considered separately having five categories in each. Three categories of father's occupation are considered as - white collar, blue collar and agricultural job. Due to the low rate of female labor force participation in India we can not take mother's occupational category, as only 30% of the enumerated children are son or daughter of an working mother. However, since household environment can be quite different for a child of working mother, we consider three categories of mother's employment status as well, namely, working mother, domestic mothers who chose not to work

for attending household duties and mothers out of labor force for any other reason. For either of the survey years considered for the present work, over 60% of Indian children have domestic mothers.

As for accounting differences in the social attributes we consider caste, religion and sex (male-female) of a children. The caste has the same categories as it is in modern India, that of the Scheduled Caste and Tribes taken together (*SC/ST*), the Other Backward Classes (*OBC*) and the *General* caste category. Among them *SC/ST*s are the most historically disadvantageous caste categories of India, whereas *General* category consists of all those who does not belong to either *SC/ST* or *OBC* and are excluded from any caste based reservation policies for being the most advantageous caste category in India. *OBC*s can be considered as the middle level caste category who are usually more advantageous than that of *SC/ST* but have less opportunity than the forward *General* caste category. Although over 70% of Indians are Hindu, the second largest religion of the country is Muslim, that makes India the country that embodies world's largest Muslim population. We therefore take three categories of religion as - Hindu, Muslims and all other than Hindu and Muslims.

In addition, subject to the availability of data we consider differences in family characteristics in terms of location of the household (rural-urban), household monthly consumption expenditure (MPCE) and number of siblings in the household. MPCE is the total expenditure incurred by the household over the past month prior to the date of the survey that includes expenditure on several important aspects including education. Unlike all other circumstances MPCE is treated as a continuous variable. Since education incurs some kind of opportunity costs it is often seen to depend on the presence of other siblings in the households. But as children of different age requires varying amount of parental care, we consider three kind of sibling related circumstances to better account for the sibling effect as - infant siblings (below 6 years), siblings who are typically in the school-going-age (6-18 years) and adult siblings (above 18 years).

Table 3.2: List of circumstance variables

1. **Father's education**
   - No formal schooling
   - Below primary schooling
   - Primary schooling
   - Middle level, below secondary schooling
   - Secondary or above level of education

2. **Mother's education**
   - No formal schooling
   - Below primary schooling
   - Primary schooling
   - Middle level, below secondary schooling
   - Secondary or above level of education

3. **Father's occupation**
   - White collar (professional & executives)
   - Blue collar (service & sales workers)
   - Agricultural (including hunting, fishing)

4. **Mother's employment status**
   - Working mothers
   - Domestic mothers (not working for attending domestic duties)
   - Mothers not in labor force for other reasons

5. **Caste**
   - General
   - Other Backward Class (*OBC*)
   - Scheduled Castes/ Scheduled Tribes (*SC/ST*)

6. **Religion**
   - Hindu
   - Muslim
   - Others

7. **Sectoral location**
   - Rural
   - Urban

8. **Sex**
   - Male
   - Female

9. **School-going-age siblings**(*6-18 years*)
   - None
   - At least one

10. **Infant siblings**(*below 6 years*)
    - None
    - At least one

11. **Adult siblings**(*above 18 years*)
    - None
    - At least one

12. **Household MPCE**
    - *Continuous variable*

### 3.4.3 Sample selection criteria

For the present analysis we take children of age 6-18 years who have valid information on all the circumstances as mentioned in Table 3.2 and are living with both of their parents where either of the parent is the household head. In particular we exclude four cases for which we did not take a child in the said age limit as our sample. First, we did not take grandchildren for selectivity issue, as this information is only available for the few selected households where the three generations (grandparents-parents-children) are living together. About 13-14% of children in the said age limit of 6-18 years are therefore excluded for being grandchildren. Second, as we want to see the effect of father's and mother's attributes separately, we rule out single-parent children for which education or occupation information of either of the parent will necessarily be missing. Besides living with both parents is the most common case in this age limit (6-18 years) and around 92% children in this age limit are living with both of their parents. Third, we can not take children who are brothers/sisters of the household head (and hence not son/daughter of the head) as information on parental backgrounds is not available for them[7]. Fourth, we exclude children who does not fit the social definition of a children for being married or parents themselves[8]. The third and fourth criteria together does not exclude more than 5% of the children in this age bracket.

However as mentioned in Table 3.1, our analysis of educational opportunity for children is segmented in several age cohorts depending on the specific outcome under analysis. For avoiding any confusion we therefore refer the 6-18 years old child sample as our 'pooled sample', selected as per the above mentioned criteria. The respective sample space specific to each of the outcomes are then rendered to different subsets of the 'pooled sample' that differs only on the basis of age cohorts. Table 3.3 provides the circumstance specific summary statistics of our samples for all different age cohorts as well as for the pooled child sample. It shows that the caste, religion, sex and sector (rural/urban) composition are very similar across all the age cohorts. The summary statistics portray India as

[7]In cases where an elder brother/sister of the sample-child is the head of the household and is living with their parents, parental information is still technically available for that sample. But in this case (when parents live in the household without sharing the headship) *NSS* reports information on father/mother/father-in-law/mother-in-law under a single code and thereby making it impossible to extract attributes of biological parents of the sample.

[8]The legal age of marriage in India is 18 for girls and 21 for boys.

Table 3.3: Circumstance specific summary statistics of child samples of different age cohorts[a]

| | %*OBC* | %*SC/ST* | %Hindu | %male | %rural | %WC_dad | %dom_mom | %inf_sib | %sch_sib | %adt_sib | %noedu_dad | %noedu_mom | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age cohort: **6-7 years** [used for: *Starting elementary education on time*] | | | | | | | | | | | | | |
| *2004-05* | 0.41 | 0.32 | 0.81 | 0.52 | 0.78 | 0.10 | 0.64 | 0.61 | 0.74 | 0.06 | 0.41 | 0.64 | 17935 |
| *2011-12* | 0.46 | 0.30 | 0.79 | 0.53 | 0.75 | 0.15 | 0.76 | 0.54 | 0.69 | 0.04 | 0.31 | 0.49 | 12023 |
| Age cohort: **14-15 years** [used for: *Finishing elementary education on time*] | | | | | | | | | | | | | |
| *2004-05* | 0.41 | 0.27 | 0.80 | 0.54 | 0.74 | 0.12 | 0.60 | 0.16 | 0.88 | 0.28 | 0.39 | 0.65 | 19831 |
| *2011-12* | 0.44 | 0.29 | 0.81 | 0.54 | 0.74 | 0.16 | 0.73 | 0.11 | 0.88 | 0.25 | 0.32 | 0.55 | 15111 |
| Age cohort: **6-10 years** [used for: *School attendance - younger cohort*] | | | | | | | | | | | | | |
| *2004-05* | 0.41 | 0.31 | 0.81 | 0.52 | 0.78 | 0.09 | 0.62 | 0.51 | 0.83 | 0.08 | 0.42 | 0.65 | 48876 |
| *2011-12* | 0.46 | 0.31 | 0.79 | 0.53 | 0.75 | 0.14 | 0.75 | 0.41 | 0.80 | 0.07 | 0.32 | 0.51 | 33200 |
| Age cohort: **11-15 years** [used for: *School attendance - older cohort*] | | | | | | | | | | | | | |
| *2004-05* | 0.41 | 0.28 | 0.80 | 0.53 | 0.75 | 0.11 | 0.60 | 0.22 | 0.91 | 0.21 | 0.39 | 0.64 | 48228 |
| *2011-12* | 0.44 | 0.30 | 0.80 | 0.54 | 0.74 | 0.15 | 0.73 | 0.15 | 0.89 | 0.18 | 0.32 | 0.53 | 35933 |
| Age cohort: **16-18 years** [used for: *Post-elementary education*] | | | | | | | | | | | | | |
| *2004-05* | 0.40 | 0.27 | 0.80 | 0.58 | 0.72 | 0.13 | 0.63 | 0.09 | 0.77 | 0.45 | 0.38 | 0.64 | 25911 |
| *2011-12* | 0.45 | 0.28 | 0.80 | 0.58 | 0.73 | 0.16 | 0.74 | 0.06 | 0.75 | 0.42 | 0.32 | 0.57 | 20985 |
| Pooled sample of age 6-18 years | | | | | | | | | | | | | |
| *2004-05* | 0.41 | 0.29 | 0.80 | 0.54 | 0.76 | 0.11 | 0.61 | 0.31 | 0.85 | 0.21 | 0.40 | 0.65 | 123015 |
| *2011-12* | 0.45 | 0.30 | 0.80 | 0.55 | 0.74 | 0.15 | 0.74 | 0.23 | 0.83 | 0.19 | 0.32 | 0.53 | 90118 |

[a] %X means the percentage share of circumstance 'X' in the sample. Some of the circumstances are abbreviated as follows - father in white collar job (WC_dad), mothers not working for attending domestic duties (dom_mom), infant siblings below 6 years old (inf_sib), sibling at school-going-age of 6-18 years (sch_sib), adult sibling above 18 years old (adt_sib), father without any formal education (noedu_dad) and mother without any formal education (noedu_mom). 'N' is sample size for the respective age cohorts.

majorly a Hindu country that is predominantly rural even in the twenty-first century, where nearly 70% of the children belong to the lesser advantageous caste categories of *OBC* or *SC/ST* and none of the age cohorts reflect any considerable male-premium as far as school education is concerned. Further for all age brackets, majority of the children are sons/daughters of non-working mothers, whereas hardly 15% of them have their fathers engaged in a white collar occupation. Although there is no sex bias for the present generation (our child samples), mothers are always more probable to be deprived of any formal education as compared to fathers. It is the composition of siblings that vary the most across different age cohorts. Not unexpectedly, the youngest cohort (6-7 years) have the least share of adult siblings whereas the oldest cohort (16-18 years) have the least share of infant siblings aged below six years.

## *3.5 Results and discussion*

### 3.5.1 Assessing the quality of elementary schooling in India

Provided the schooling structure in India as presented in section 3.2, all children by the age of 6 or at most 7 years should have started their formal schooling and therefore without failing, should have finished the basic eight years of elementary education in their early teenage or latest by the age of 15 years. This is clearly not the case as reflected by Table 3.4 which tabulates the share of children compliant with the above criteria. Nearly one-fourth of the children of age 6-7 years are yet to enroll in any kind of formal schooling in 2004-05 and in the same survey year, not even 50% children have reported to finish their elementary education even by the age of 15. Although both figures show an increase for the latest survey year of 2011-12, school attendance of children always show a decline with age. The aim of this section is to see how children's access to elementary education in the form of school attendance as well as timely beginning and finishing of it is affected by differences in their circumstances.

The logistic regression estimates of Table 3.A.1 and 3.A.2 in Appendix 3.A show that all of our circumstances have very significant effect on timely beginning and finishing of elementary education for children in the eligible age cohorts as well as on their school

| | Share of children | |
|---|---|---|
| Survey years → | 2004-05 | 2011-12 |
| Starting elementary by 6-7 yrs. | 76.4 | 84.9 |
| Finishing elementary by 14-15 yrs. | 49.6 | 63.7 |
| School attendance for 6-10 yrs. | 83.9 | 91.2 |
| School attendance for 11-15 yrs. | 78.7 | 89.6 |

Table 3.4: Share of children with access to basic opportunities: Elementary education[a]

[a]Share of children reports the percentage share of children of the respective age cohorts who satisfy the outcome specifications. For example during the survey year of 2004-05, 76.4% children of age 6-7 years have started elementary schooling and 83.9% children of age 6-10 years are currently attending any kind of formal schooling.

attendance[9]. While lower caste children are much less likely to start and finish elementary education on time in 2004-05, the situation have improved over the years especially in timely enrollment to formal schooling. Unfortunately *OBC* and *SC/ST* children for the older age cohort (11-15 years) are still less likely to attend school as compared to the forward *General* caste categories. Females are always less likely to attend schools as compared to male children even by 2011-12, but those who attend are almost equally likely to finish elementary education on time as males. Similar to other studies on education in India we also found Muslims to be significantly worse off than Hindus in all aspects of elementary schooling. Although rural children are always more likely to attend formal schooling as compared to the urban sector, they are much less likely to complete the basic eight years of elementary schooling on time. This may indicate that rural India albeit not suffering from inadequate number of schools, are yet to well-build the institutional infrastructure for good quality education.

Children from agricultural families are relatively worse off than those having fathers in white collar professions, although this situation has improved for 2011-12 especially for the younger kids at the onset of their schooling. Further, children of working mothers are always more probable to begin and complete elementary education on time. But the educational access is miserable for children when their fathers or mothers are deprived of any formal schooling. Interestingly children of lesser educated parents are more prone

[9]Following Dabalen et al. (2015) we did not take any interaction between the circumstances in the logistic regression. Taking interactions essentially means an increase in the number of circumstances, thereby lowering the HOI and increasing the D. Since no econometric analysis can take 'all' possible circumstances, it is simpler to consider circumstances without interaction and to interpret the results as the upper bound of HOI and the lower bound of D (Dabalen et al. 2015, Box 2.4).

to be enrolled in formal schooling on time when parents have less but some experience of schooling, probably because parents who experience some schooling themselves, aspire more for the education of their children. But unfortunately this aspiration does not hold long and children of lesser educated parents are always less likely to finish the basic eight years of schooling on time. While presence of infant siblings who demand more parental investment, is a barrier to education for the other children in the same household, having a sibling who is in his/her school-going age seem to boost school attendance for each other. Finally the positive coefficients on MPCE indicates that in spite of making elementary education free for all, probability of attending school is still a bit high for children from wealthy households.

Table 3.5 reports the HOI index as a measure of overall development as well as the D-index which quantifies the degree of *IOP* in access to elementary education for children in India. The left panel of the table reports the indices for the timely beginning and finishing of elementary education for children at eligible age brackets, whereas the right panel reports the same for school attendance at two different age cohorts. For any outcome, the first panel of Table 3.5 provides the index of HOI as estimated from equation (3.6) and the second panel reports the availability of the 'basic opportunity' or the *coverage* of it as presented by their estimated average probability of access ($\bar{p}$). The third panel reports the D-index that is estimated from equation (3.5) and is interpreted as the degree of *IOP* for each of the respective outcomes. Therefore an opportunity equal development of the society should be reflected by an increase in HOI and a fall in the D-index.

First of all notice that the outcome of school attendance is always better than either starting or finishing elementary education on time, which is similar to the Sub-Saharan African and the Latin American countries. As compared to starting and finishing elementary schooling at the right age, the outcome of school attendance always have a higher HOI and lower *IOP* (D-index) for either of the younger or the older age cohorts. HOI is always above 70% for starting formal schooling by the age of 6-7 years along with a relatively low *IOP*, which not only indicates good availability of elementary schools in India but also its wide spread access to children from varying circumstances. Further, comparatively better access to the timely starting of elementary education irrespective of individual circumstances, is also complemented by higher HOI and lower *IOP* in school

| Outcomes → | **Start** on time | **Finish** on time | **School attendance** younger cohort | older cohort |
|---|---|---|---|---|
| *Age cohorts →* | *6-7 years* | *14-15 years* | *6-10 years* | *11-15 years* |
| | | Overall development: *HOI* | | |
| *2004-05* | 71.23 | 39.71 | 79.18 | 72.30 |
| | [0.592] | [0.534] | [0.331] | [0.359] |
| *2011-12* | 81.47 | 56.01 | 88.41 | 85.85 |
| | [0.817] | [0.868] | [0.413] | [0.443] |
| | | Distribution of basic opportunities: *Coverage* ($\bar{p}$) | | |
| *2004-05* | 76.39 | 49.64 | 83.88 | 78.75 |
| | [0.486] | [0.481] | [0.258] | [0.276] |
| *2011-12* | 84.91 | 63.66 | 91.19 | 89.67 |
| | [0.615] | [0.713] | [0.297] | [0.315] |
| | | ***IOP*** as a penalty for development: *D-index* | | |
| *2004-05* | **6.76** | **20.01** | **5.60** | **8.20** |
| | [1.198] | [1.871] | [0.614] | [0.700] |
| *2011-12* | **4.05** | **12.01** | **3.04** | **4.25** |
| | [1.117] | [2.115] | [0.713] | [0.793] |

Table 3.5: Elementary education: HOI & IOP[a]

[a]Standard errors in square brackets.

attendance for the younger cohort. A further improvement is also noticeable in either of these outcomes after the mandate on free elementary education by the RTE act of 2009. By 2011-12, both school attendance of the younger cohort and on-time starting of elementary schooling has a HOI over 80% with a low *IOP* of 3-4%.

But in spite of the impressive starting, HOI shows a significant drop in its value with a sharp rise in *IOP* as far as the timely finishing of elementary education is concerned. The average access probability (*coverage*) to finish eight years of elementary schooling in India is 49% in 2004-05, which is marginally higher than a handful of Latin American countries (like Brazil, El Salvador, Nicaragua, Guatemala) (Paes de Barros et al. 2008) but higher than most of the Sub-Saharan African countries (Dabalen et al. 2015). However the *IOP* for completing the basic eight years of elementary school education latest by the age of 15 is over 20% during 2004-05, which is higher than most of the Latin American countries (Paes de Barros et al. 2009). Notice also the deterioration of either of the HOI and the D-index for school attendance among older cohort of 11-15 years old children, which is consistent with the low HOI in finishing elementary education on time. Together this suggests that the enthusiastic beginning of schooling is rather short-lived in India and the timely completion of elementary education is even more crippled for children from adverse backgrounds. This trend is in contrary to the Sub-Saharan African countries however, that exhibits an improvement in school attendance with age so that there is lesser *IOP* and better HOI as far as finishing primary education on time is concerned as compared to its timely beginning (Dabalen et al. 2015). Nevertheless an improvement in overall elementary education is evident in India across all the age cohorts over the time span of *2004-12*.

Table 3.6 shows that from 2004 to 2012 HOI improves by more than 10% in terms of timely start and completion of elementary education. Also HOI for school attendance in 2011-12 is 9-13% higher than it was in 2004-05. However as mentioned in section 3.3.2, the change in HOI can be brought about by two factors - the increased availability of opportunities or the scale effect and the decreased *IOP* in the available opportunities or the distribution effect. Similar to the extant literature we also found that majority of the change in HOI is brought about by the increased coverage of basic opportunities in the society. However Table 3.6 shows that nearly one-fourth of the improvement in

HOI can be attributed to the abated *IOP* in the society and it is the timely completion of elementary education for which the amelioration of HOI as well as *IOP* is maximum. The scale effect on the other hand is maximum for the timely beginning of elementary schooling which could be brought about by the establishment of new elementary schools as an effect of the free education policy.

| Time frame: *2004-12* | ΔHOI | Scale effect | Distribution effect |
|---|---|---|---|
| Change in → | overall development | availability of opportunities | inequality of opportunities |
| *Starting on time* | 10.23% | 77.54% | 22.46% |
| *Completing on time* | 16.31% | 68.76% | 31.24% |
| *School attendance: younger cohort* | 9.22% | 74.69% | 25.31% |
| *School attendance: older cohort* | 13.55% | 73.91% | 26.08% |

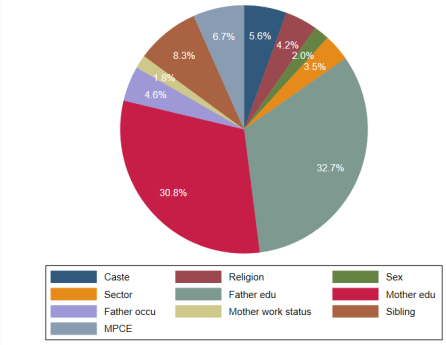
Table 3.6: Improvement in elementary education: Scale and distribution effects of HOI[a]

[a]ΔHOI indicates percentage change in HOI from 2004-05 to 2011-12. Thus in terms of starting elementary schooling on time, the value of HOI in 2011-12 is 10.23% higher than that of 2004-05 and 77.54% of this change in HOI is due to the scale effect whereas 22.46% is due to the distribution effect.

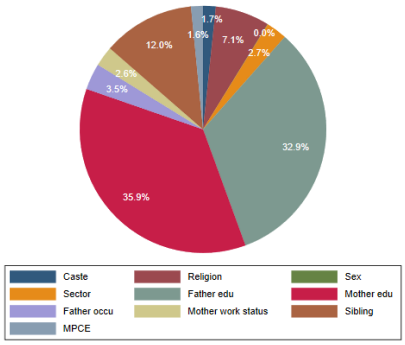
A relatively low D-index of 3-7% reflects that access to free and compulsory elementary education seem to be more equal for all children at the onset of schooling. However a higher D-index of 12-20% indicates that the differential circumstances do play a role in the continuation of elementary schooling so as to finish elementary education on time within the eligible age limit of free education. We therefore proceed to estimate the relative importance of different circumstances in the resulting D-index, by the Shapley value decomposition. This is a concept from the co-operative game theory, applied in the context of distributional analysis by Shorrocks (2013). To estimate Shapley decomposition of the D-index, the power set of all the circumstances are formed first. Then for each circumstance, its marginal contribution is estimated as the difference between the D-index for all the sets where that particular circumstance is included and those where it is not. The Shapley value is the average of all such marginal contribution and is often expressed in percentages that sum up to 100 for all circumstances. Figures 3.1 represents the Shapley decomposition pie diagrams for all of our concerned outcomes.

First of all notice that for all outcomes father's and mother's education are the most important circumstances that together account for more than 60% of the resulting *IOP* (D-index). In fact for the timely beginning and finishing of elementary schooling, mother's education matters more than that of father's especially for the latest survey year. Not

Starting elementary education on time

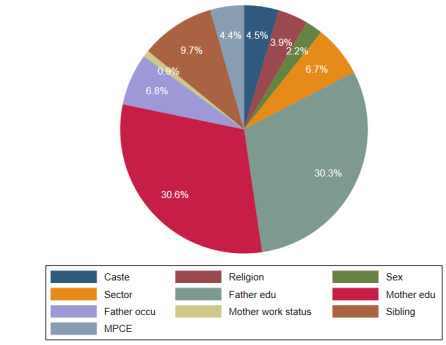



(a) 2004-05

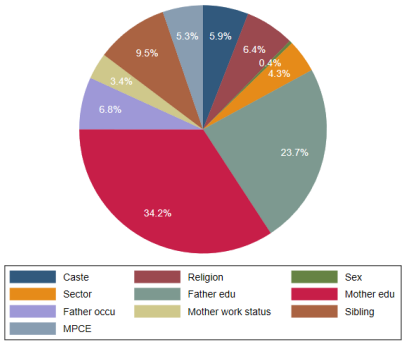



(b) 2011-12

Completing elementary education on time

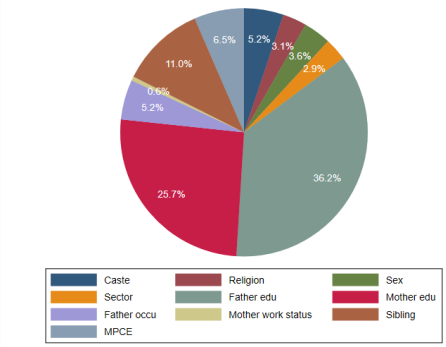



(c) 2004-05

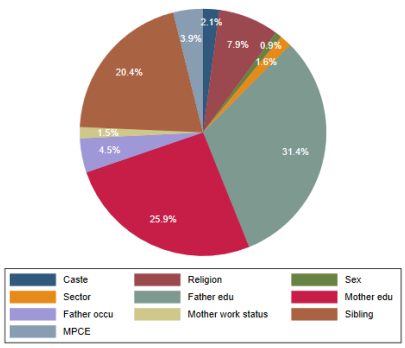



(d) 2011-12

Elementary school attendance: younger cohort (6-10)

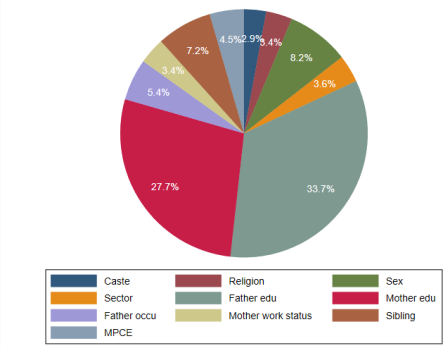



(e) 2004-05

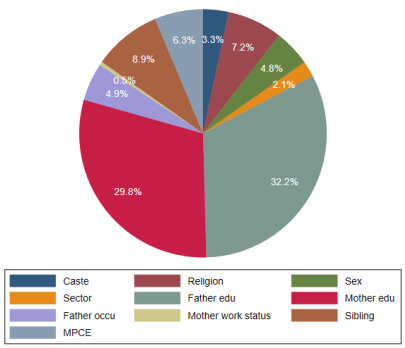



(f) 2011-12

Elementary school attendance: older cohort (11-15)




(g) 2004-05




(h) 2011-12



Figure 3.1: Shapley decomposition of the D-index: Elementary education

unexpectedly, after parental education the most contributing circumstances are the sibling composition of the household and father's occupation. Contribution of the social backgrounds of caste, sex, religion or sector (rural-urban) have relatively lesser impact on the resulting *IOP* but their relative importance changes over time. While the male premium for access to elementary education reduces over time across all age cohorts, the forward caste premium decreases only for the relatively younger cohorts. The role of religion on the other hand seems rather protruded in 2011-12 as compared to 2004-05. Still by 2011-12 more than 10% of *IOP* comes from the differences in caste, sex, religion or sectoral habitat.

To summarize, the situation of elementary education in India has found to be substantially improved over the time span of 2004-12 which could be indicative of the efficacy of the free elementary education policy. The average HOI for school attendance is around 90%, which is higher than most of the Sub-Saharan African countries. An increasing coverage and scale effect for the later round indicates infrastructural development in terms of provision of basic education in the country. Whereas the low D-index of around 3-4% reflects that the provision of more schooling is not restricted to children from advantageous background, especially at the onset of formal school education. However the impressive starting of formal schooling is not backed up enough by regular continual attendance, which leads to a much lower HOI and a considerable higher D-index in terms of completion of elementary education on time. Thus the legal mandate on compulsory elementary schooling seems to be mostly beneficial at an early age and the problem of drop outs is still persistent even after making basic elementary education free for all children. Like most other studies parental education turned out to be the most important circumstance for India as well. Interestingly, while lesser educated parents are more likely to enroll their children to formal schooling, it is the offspring of the more educated ones who have a higher chance to finish basic education on time.

### 3.5.2 Regional variation in basic educational opportunity

India is a large country with a good amount of regional variation, not only in culture, language, caste or religious dynamics, but also in terms of educational development. While Kerala (a Southern state) with a 100% literacy rate stands as the best performing state

in India in terms of education, Bihar (an Eastern state) remain consistently underdeveloped that still shows a literacy rate of below 50% even in the 2001 census (Ghosh 2006, Gourishankar & Sai Lokachari 2012). It is therefore worth a separate analysis, to see how differently the regions perform in terms of child access to elementary education. We use the most commonly used regional partition of India as - *North*, *East*, *Central*, *North-East*, *South* and *West*[10].

For the survey years used for the present analysis, majority of the population is from Central, Eastern and Southern region, whereas North-East and North have the smallest share. All regions are majorly Hindu regions, but the non-Hindu population share is relatively higher for the North and the North-East. Caste composition is even more varying across regions. South has relatively more concentration of *OBC*s, while North-East is the tribal hub of India with the largest share of *SC/ST*s who are mostly from rural sector. Proportion of female school goers are almost similar for all regions with a little higher share in South which is known for its phenomenal progress of female literacy. Mothers are always way less educated than fathers in all over India, but West has relatively more children with high educated fathers occupied in white collar type jobs. Since circumstance specific composition varies across regions, it is possible that a particular circumstance that is advantageous to one region may be a cause of hindrance to other. We nevertheless consider the same set of circumstances for comparability in our regional analysis as well.

However neither of the HOI and the D-index are sub-group consistent, in the sense that it is possible to have a non-increasing HOI for the nation even though most of the regions have an increasing HOI and others remain the same. An alternative is to use the geometric HOI that is calculated from the geometric mean of the circumstance specific coverage for each region. But geometric HOI may have a very low value in case of diverge and large number of circumstances, particularly if coverage of a specific circumstance is relatively low for one region (Dabalen et al. 2015). Also unlike HOI, the geometric HOI

[10]State wise composition: Jammu & Kashmir, Himachal Pradesh, Punjab, Haryana and Uttarakhand - constitutes *North*; Bihar, Jharkhand, Orissa, West Bengal - constitutes *East*; Uttar Pradesh, Rajasthan, Madhya Pradesh, Chattisgarh - constitutes *Central*; Sikkim, Arunachal Pradesh, Assam, Nagaland, Meghalaya, Manipur, Mizoram, Tripura - constitutes *North-East*; Karnataka, Andhra Pradesh, Tamilnadu, Pondichery, Kerala, Lakshadeep - constitutes *South* and Gujrat, Daman & Diu, Dadra & Nagar Haveli, Maharashtra, Goa - constitutes *West*.

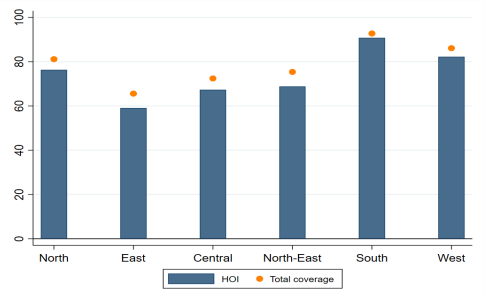
is not amicable to intuitive graphical interpretation. For this reason we consider HOI for the regional analysis as well taking into account the same group of circumstances for each region. We therefore refrain from comparing the regional analysis to the national one and interpret the regional results strictly on their own, not in comparison with a national yardstick.

Figure 3.2 gives the HOI and the associated coverage for all outcomes associated to elementary education, namely, timely start of elementary schooling by the age of 6-7 years, timely finishing of the same by the age of 14-15 years and school attendance for different age cohorts (younger cohort of 6-10 years and older cohort of 11-15 years). The first two panel corresponds to the timely starting and completion of elementary schooling and the latter two panels are for school attendance at different age cohorts. In each diagram the vertical bar corresponds to the value of HOI and the dot represents the coverage of the outcome for that region[11]. The gap between the dot and the bar is therefore the penalty for development due to the existing *IOP* in the society. So higher this gap larger is the degree of IOP for that outcome in that particular region. First of all notice that there exist IOP for almost all regions as the top of the bar is always below the dot. Secondly regional variation is less in case of school attendance than in case of timely start and finishing, which is a similar trend found in the Latin American and Sub-Saharan African countries as well.

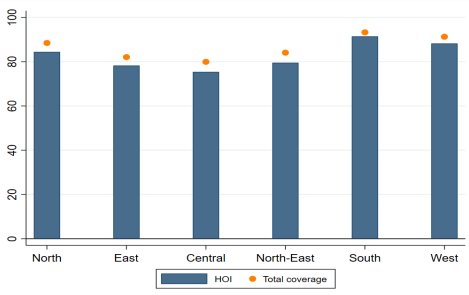
Although both HOI and the coverage increases over time in all aspects of elementary education across all regions, Central and East are two of the worst performing region both before and after the implementation of the free education policy. Southern India on the other hand not only stands better in terms of timely beginning and completion of basic elementary education, *IOP* is minimum for South as well (as reflected by the minimum gap between the bar and the dot). As far as successful completion of basic elementary schooling is concerned, South and West were equally good in 2004-05 but the improvement over time is rather pronounced for the former region. Whereas East has the lowest HOI as well as the lowest coverage for starting elementary education at the right age that is closely followed by Central India. But it is the Central region that stands worst in terms of timely finishing of elementary education, indicating that the problem

[11]The figure is drawn from Table 3.A.4 in appendix 3.A.

Staring elementary schooling on time (6-7 yrs.)

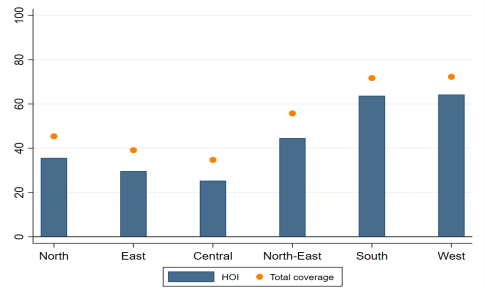



(a) 2004-05

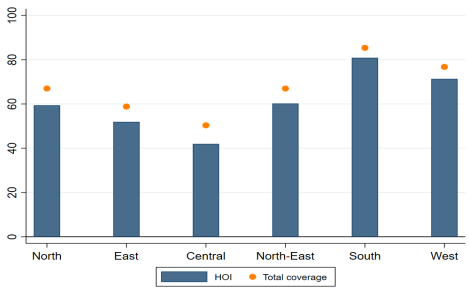



(b) 2011-12

Finishing elementary schooling on time (14-15 yrs.)

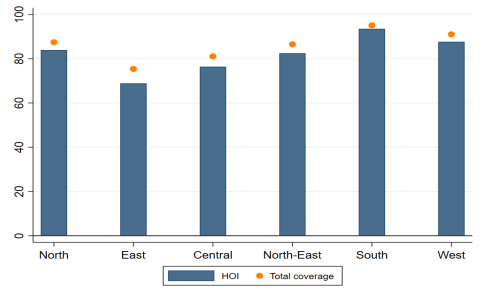



(c) 2004-05

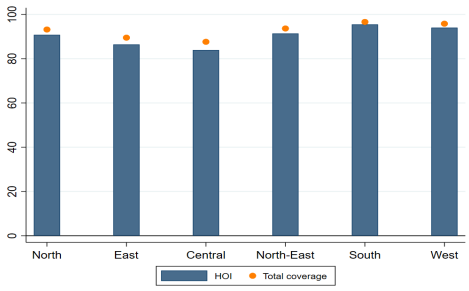



(d) 2011-12

School attendance: younger cohort (6-10 yrs.)

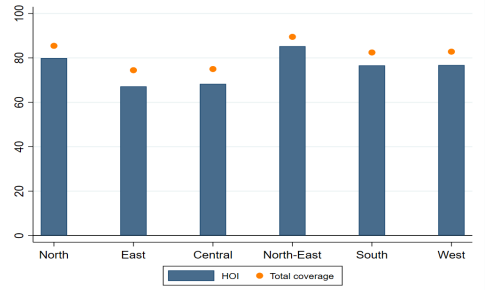



(e) 2004-05

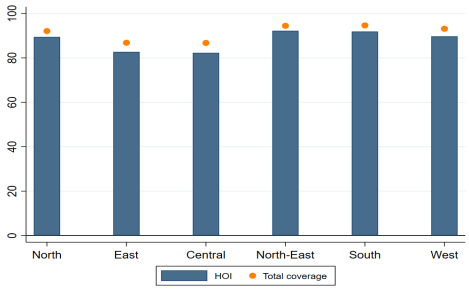



(f) 2011-12

School attendance: older cohort (11-15 yrs.)




(g) 2004-05




(h) 2011-12

Figure 3.2: Elementary education: Regional variation in HOI

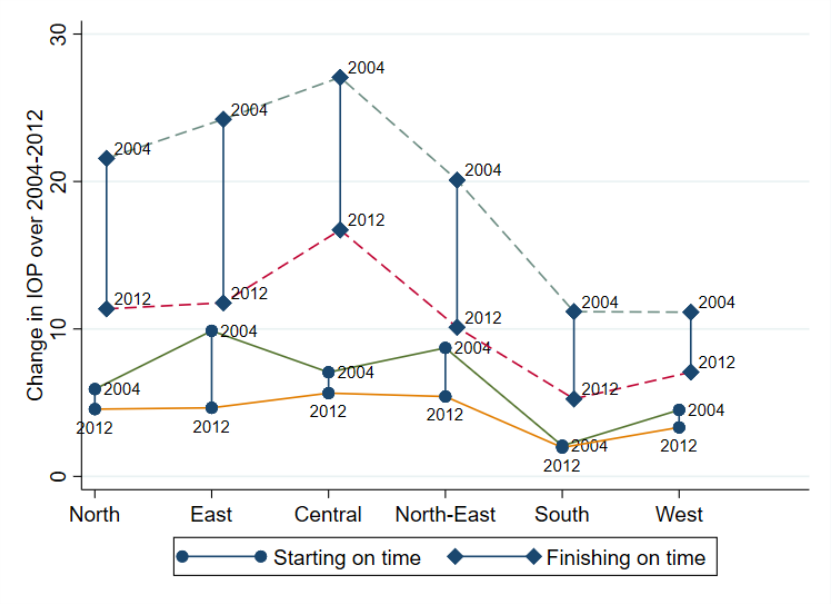
of school drop-outs may be higher for this part of the country.

Figure 3.3 plots the value of *IOP* as represented by their respective D-indices for all outcomes and across every region. The upper panel plots the regional D-indices for timely starting and finishing of elementary schooling, whereas the lower panel plots the same for school attendance at different age cohorts[12]. Each panel therefore plots two outcomes that are differentiated by the diamond and round headed points. While the vertical distance between the similar points represents the change in *IOP* over time within a specific region, the horizontal line connecting the point-heads trace out a relative ranking of the regions in terms of *IOP* within each survey years. Once again the regional variation in *IOP* is rather noticeable in the upper panel for the beginning and completion of elementary schooling on time and while South stands the best performing region overall, East and Central are the worst performing ones, also in terms of *IOP*.

East has the highest *IOP* in timely beginning of elementary schooling with relatively large differences as compared to the other regions and during 2004-05 it also stands out as a rather opportunity unequal region in terms of school attendance as well. But it is the same region that improves the most over time in almost all aspects of quality elementary education (as the vertical distance between similar point-heads are maximum for East mostly). Central part of India however did not improve as much as East and by 2011-12, it is the region for which *IOP* is maximum for most of the outcomes. South is the best performing region in terms of *IOP* as well followed closely by West. Both of these regions were quite close in terms of school attendance at older cohort and timely completion of elementary schooling, although it is South that better equalizes educational opportunity over time as compared to the West. The upper panel of Figure 3.3 reflects that for either of the survey rounds *IOP* is considerably higher for the timely completion of elementary education than the beginning of it, for all regions. Except for North-East, the second panel shows that IOP in school attendance is higher as well for older school going children. Together these indicate that in almost all parts of India, while children from varying circumstances get a relatively more enthusiastic beginning by timely enrollment in schools, the differences in their social and family backgrounds started to impair the

[12]Notice that the scale of *IOP* measures for the two panels are different. Hence the graphs within the panel and not across the panel, are immediately comparable. While comparing the visuals between the two panels one needs to mind the range of y-axis.

Quality of elementary education: staring and finishing on time




School attendance: younger and older cohort

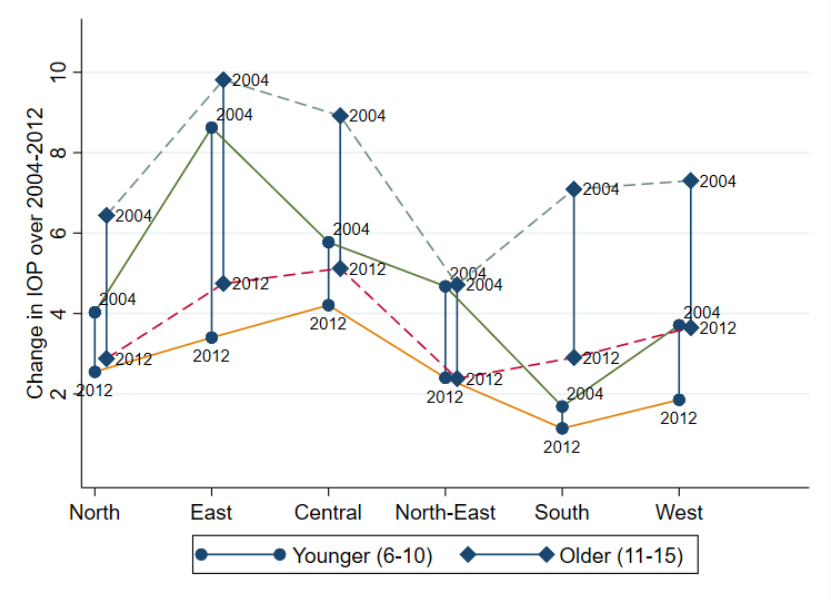



Figure 3.3: Elementary education: Changes in *IOP* over time and regions

regular continuation of formal education as the children gets older.

### 3.5.3 An anecdote on post-elementary schooling

Although the completion of basic primary or elementary education is crucial for the cognitive formation of a child, a steady continuation to secondary education is often more beneficial in terms of higher labor market returns, lower poverty or even lesser infant mortality rates for many developing countries including India (Duraisamy 2002, Tilak 2007, Peet et al. 2015, Sánchez & Singh 2018). In fact unsuccessful transition to secondary education is found to have a negative spillover effect in terms of finishing elementary education for children in the society as well (Lewin 2011). In this section we therefore provide a brief analysis of access to lower secondary education for the relatively older children in India, for the same set of circumstances. In particular we analyze whether all children on or above 16 years have completed their lower secondary education (grade-10) latest by their adulthood, irrespective of the differences in their individual circumstances. Further like the case of elementary education, the present discussion of post-elementary schooling is also complemented by analyzing the school attendance profile for children in an 'adolescent' age cohort of 16-18 years.

As compared to elementary education, pursuing secondary education is rather subject to individual choice and performance and therefore strictly speaking, should not be a subject of study for the responsibility sensitive opportunity analysis. But social and family backgrounds are often found responsible for shaping these 'choices' since childhood that eventually renders in lack of motivation to continue education above the basic minimum. Besides as discussed in section 3.2, most of the private and public schools that provides elementary schooling in India have the provision of lower secondary education (up to grade 10) as well. Hence even though the free education mandate is on the first eight years of elementary schooling, the additional cost for pursuing two more years of lower secondary education is not prohibitively high in most cases. On the other hand, because of the need of advanced infrastructure with specialized teachers and other educational instruments in the higher secondary education, there are relatively fewer schools in India that cover the entire elementary and secondary education[13].

[13]As mentioned in section 3.2, unlike the lower secondary education (grades-9,10) higher secondary schooling (grades-11,12) is provided on advanced core areas of a subject (*e.g.* science, arts, humanities,

So every children enrolled in an elementary school on time should have finished their lower secondary education (grade-10) by the age of 16, without failing. Provided the higher rate of school drop-outs for older children we relax the age limit till adulthood. In spite of that only 35% of all children aged between 16-18 years have reported to finish their lower secondary education in 2004-05, that increased to 54% by 2012. Further the school attendance profile shows that less than half of the children in this age cohort are attending any formal schooling at all in 2004-05 and even by 2012 more than two-third are out of any formal educational institute before they reach their adulthood. Table 3.7 reports the HOI and *IOP* for outcomes related to the lower secondary education along with their respective coverage ($\bar{p}$).

| Outcomes → | Finishing lower secondary on time (*16-18 yrs.*) | | School attendance for adolescent children (*16-18 yrs.*) | |
|---|---|---|---|---|
| | Overall development: *HOI* | | | |
| *2004-05* | 25.03 | [0.396] | 38.04 | [0.475] |
| *2011-12* | 44.06 | [0.697] | 57.94 | [0.757] |
| | Distribution of basic opportunities: *Coverage* | | | |
| *2004-05* | 34.89 | [0.398] | 47.95 | [0.427] |
| *2011-12* | 53.53 | [0.595] | 66.10 | [0.604] |
| | ***IOP*** as a penalty for development: *D-index* | | | |
| *2004-05* | **28.24** | [2.288] | **20.69** | [1.791] |
| *2011-12* | **17.69** | [2.117] | **12.34** | [1.792] |

Table 3.7: Post-elementary education: HOI & IOP[a]

[a]Standard errors in square brackets.

Table 3.7 shows that the HOI are much lower and the D-indices are considerably higher for the lower secondary education as compared to the elementary schooling. The average probability of access (coverage) to lower secondary education is only a little higher than 50% even by 2012. Further a lower coverage of school attendance also reflects that most of the children are more likely to drop out from formal schools as they approach adulthood.

commerce etc.) that demands specialized instructors and instruments. Over one-fourth of the rural households in India does not have a higher secondary school within 8 km of their residence and the figure is nearly 50% for the destitute castes of *ST*s who live in relatively remote areas.

A high D-index od 28% in 2004-05 reflects that the probability of finishing lower secondary level schooling is rather handicapped for children from adverse circumstances. In spite of a fall in *IOP* for the latest survey year of 2011-12, a D-index of more than 17% indicates that children from disadvantageous circumstances are yet less probable to finish ten years of formal schooling even by their adulthood.

Nevertheless the improvement in school attendance for children in this age bracket for the latest survey year indicates the possibility of a positive spillover effect of the mandate on free elementary education in terms of continuing school education even beyond the elementary level. However Table 3.8 shows that similar to elementary education, the improvement in HOI for the lower secondary schooling as well is majorly driven by the scale effect. Over the time frame of *2004-12* HOI increases over 19% in terms of finishing lower secondary education before adulthood and nearly 30% of this improvement is driven by a lesser *IOP* among the children in this age bracket (16-18 years).

| Time frame: *2004-12* | ΔHOI | Scale effect | Distribution effect |
|---|---|---|---|
| *Finishing lower secondary education in childhood* | 19.03% | 70.29% | 29.71% |
| *School attendance of older children (above 16 yrs.)* | 19.91% | 72.28% | 27.72% |

Table 3.8: Amelioration in post-elementary schooling for older children

| Outcomes → | Finish lower secondary | | School attendance | |
|---|---|---|---|---|
| Age cohorts → | 16-18 years | | 16-18 years | |
| | 2004-05 | 2011-12 | 2004-05 | 2011-12 |
| Social attributes | 16.9% | 20.6% | 14.7% | 18.8% |
| Parents' education | 59.3% | 58.2% | 62.3% | 60.9% |
| Parents' working status | 12.2% | 7.8% | 13.9% | 11.0% |
| Other family attributes | 11.5% | 13.4% | 9.1% | 9.4% |

Table 3.9: Post-elementary education: Shapley decomposition of circumstances[a]

[a]Where social attributes include - caste, religion, sex, sector (rural-urban); parents' education includes - father's and mother's education; parents' working status includes - father's occupational category and mother's working status (whether working or not); other family attributes include - sibling and household MPCE. See Figure 3.A.1 in Appendix 3.A for the Shapley decomposition of all different circumstances separately.

The Shapley decomposition of circumstances as reported in Table 3.9 shows that similar to the elementary education, parental education alone constitutes of nearly 60% of *IOP*

in access to lower secondary education as well[14]. However as compared to finishing elementary education on time, social attributes of caste, religion, sex or sector (rural-urban) matter more for the completion of lower secondary schooling before adulthood. Also as expected, working profile of parents seem to matter more for school attendance of relatively older children and the associated logistic regression results of Table 3.A.3 in Appendix 3.A shows that children with father's in blue collar jobs are almost equally worse off than those with agricultural fathers, as compared to children with white collar job fathers. But while at a younger age nearly one-fourth of *IOP* in school attendance is due to differences in the family attributes like sibling composition and household MPCE, the effect of these circumstances decreases as the children ages. For school attendance in particular, the contribution of sibling composition alone is found to be 20.4% for the younger cohort (6-7 years), 8.9% for the older cohort (11-15 years) and 5.6% for the adolescent cohort (16-18 years), for the latest survey year of 2011-12 (*see* Figures 3.1 and 3.A.1). But in spite of the overall improvement in both elementary and lower secondary education in India, the effect of caste and religion seem to aggravate over time in access to education for relatively older children.

## *3.6 Concluding remarks*

We analyze access to educational opportunity for children aged between 6-18 years in the twenty-first century India. To distinguish the differential effect of age on schooling we consider a set of six different outcomes for different age cohorts. Starting and finishing elementary education concerns with children aged below 16 years, whereas completing lower secondary education is analyzed for all on or above 16 years. Either of the analysis of elementary and post-elementary schooling is further complemented by the outcome of school attendance at different age cohorts. Over the time frame of *2004-12* we found a clear improvement in the overall educational development of the country as well as a fall in *IOP* at all levels of children education. This could have been attributed to the mandate on free compulsory education policy launched in 2009, that not only improves elementary

[14]We abbreviated the circumstances of caste, religion, sex and sector as 'social attributes'; that of father's and mother's education as 'parents' education'; that of father's occupation category and mother's employment status (whether working mother or not) as 'parents' working status' and that of sibling and household MPCE as 'other family attributes'. Figure 3.A.1 in Appendix 3.A gives the Shapley decomposition for each separate circumstances for finishing lower secondary education and school attendance for 16-18 years old children.

education but also have its indirect positive effect on continuing school education beyond the free elementary level.

Similar to many other countries, literacy and the cognitive formation at the onset of childhood is the primary focus of educational policy in India as well. We therefore emphasize more on the analysis of basic elementary education for 6-15 years old children. Along with school attendance at different age cohorts (6-10 years and 11-15 years), we consider the timely beginning (by 6-7 years) and finishing (by 14-15 years) of eight years of basic schooling, as a proxy for the quality of elementary education. Similar to the Latin American Countries (LAC) and the Sub-Saharan African (SSA) countries, school attendance for younger Indian children (6-10 years) has an impressively high HOI over 80%, along with lower IOP as reflected by a lower D-index of 3-5%. Especially in terms of early age school attendance, India is better than most of the SSA countries and rather close to the best-performing countries in Latin-America (Paes de Barros et al. 2009). Further, our regional analysis shows that this improvement is a pattern for all parts India. However South and West turns out to be the best regions in terms of school education access of all children, while East and Central India are the worst.

But both HOI and IOP deteriorates as the children ages. D-index in 2004-05 was as high as 20% in terms of completion of elementary education by 14-15 years, while it has a very low value as far as beginning formal schooling by 6-7 years is concerned. This is consistent with lower HOI for school attendance for the older cohort (11-15 years) than that of the younger (6-10 years). In the similar age group and time frame, Dabalen et al. (2015) found that HOI in attendance for several SSA countries improves with age along with an improvement in its value for the completion of primary education as compared to the timely beginning of it. While HOI for the timely start of primary education by 6-7 years was less than 50% for almost all of the SSA countries (except Malawi and Zimbabwe), it is more than 70% for India. But in terms of timely completion, India lags behind some of the best-performing countries of SSA (like Zimbabwe). Also contrary to the SSA countries, HOI for school attendance of older cohort in India (11-15 years) is less than that of younger ones. Therefore in terms of quality of basic education, India depicts an opposite pattern than that of the SSA countries. While access to education was poor at the onset of schooling for SSA, it was much stronger for India at the beginning that

decays with age.

Since labor market return have been repeatedly found to be determined by secondary and tertiary education, we proceed to analyze post-elementary education for relatively older children who are approaching their adulthood. Deterioration of access to education is even more pronounced in case of lower secondary level education (grade 10) for children aged between 16-18 years. However a substantial improvement in completion of grade 10 schooling by the age of 16-18 years and school attendance of children of the same age is evident over time. This indicates that the legal enforcement of free elementary education may have an indirect positive impact at the level of lower secondary schooling as well. Although child sample corresponding to the post-elementary analysis are outside the age limit for free-education, a guaranteed schooling up to grade 8, indeed improves a further continuation till grade 10. This is normal, given that most of the private and public schools in India that provides elementary education, also has the provision of lower secondary. But overall, educational scenario both in terms of infrastructural development and opportunity equalization are mostly concentrated at the onset of school education that fades away gradually as the children ages. This suggests that India has long way to go in analyzing and arresting school dropouts.

The Shapley decomposition of different circumstances for all outcomes reveal that like most of the countries parental education is the most important determinant for the educational attainment of the children. However while lower educated parents are more prone to enroll their children in elementary schooling at the right age, this aspiration dies with age and in terms of completion and it is always the children of high educated parents who are more likely to finish. Number of siblings and father's occupation stand out to be the next important circumstances and children of white collar job fathers with no dependent infant sibling are almost always more advantageous especially for attending school at an early age. In terms of quality elementary education, rural India improves non-trivially so as to compensate the gap with the urban privileged schooling. But in case of lower secondary education, the relative disadvantages of rural households remains high. Sex of a child have much little role to play in terms of overall access to schooling, although older girl children are still less likely to attend schools than boys. However the effect of caste and religion in accessing school level education is relatively low, but is

found to be rather pronounced for the latest year especially as the children gets older.

# Appendices to Chapter 3

## 3.A Additional tables and figures

| Outcomes → | Starting elementary (*6-7* yrs.) | | | | Finishing elementary (*14-15* yrs.) | | | |
|---|---|---|---|---|---|---|---|---|
| | *2004-05* | | *2011-12* | | *2004-05* | | *2011-12* | |
| *Ref: Non-SC/ST* | | | | | | | | |
| OBC | -0.24*** | [0.00] | -0.04*** | [0.00] | -0.04*** | [0.00] | 0.13*** | [0.00] |
| SC/ST | -0.36*** | [0.00] | 0.01*** | [0.00] | -0.33*** | [0.00] | -0.17*** | [0.00] |
| *Ref: Hindu* | | | | | | | | |
| Muslims | -0.40*** | [0.00] | -0.37*** | [0.00] | -0.56*** | [0.00] | -0.47*** | [0.00] |
| Others | -0.18*** | [0.00] | -0.30*** | [0.00] | 0.07*** | [0.00] | -0.05*** | [0.00] |
| *Ref: Male* | | | | | | | | |
| Female | -0.17*** | [0.00] | 0.02*** | [0.00] | -0.31*** | [0.00] | -0.06*** | [0.00] |
| *Ref: Urban* | | | | | | | | |
| Rural | 0.01*** | [0.00] | -0.09*** | [0.00] | -0.20*** | [0.00] | -0.04*** | [0.00] |
| *Ref: Secondary or more (Father)* | | | | | | | | |
| Below secondary | 0.13*** | [0.00] | 0.27*** | [0.00] | -0.05*** | [0.00] | -0.19*** | [0.00] |
| Primary | 0.06*** | [0.00] | 0.40*** | [0.00] | -0.49*** | [0.00] | -0.38*** | [0.00] |
| Below primary | -0.38*** | [0.00] | 0.34*** | [0.00] | -0.65*** | [0.00] | -0.40*** | [0.00] |
| No schooling | -0.59*** | [0.00] | -0.33*** | [0.00] | -1.04*** | [0.00] | -0.65*** | [0.00] |
| *Ref: Secondary or more (Mother)* | | | | | | | | |
| Below secondary | 0.00 | [0.24] | 0.27*** | [0.00] | -0.32*** | [0.00] | -0.35*** | [0.00] |
| Primary | 0.04*** | [0.00] | -0.20*** | [0.00] | -0.94*** | [0.00] | -0.96*** | [0.00] |
| Below primary | 0.24*** | [0.00] | -0.44*** | [0.00] | -0.99*** | [0.00] | -0.65*** | [0.00] |
| No schooling | -0.71*** | [0.00] | -0.94*** | [0.00] | -1.65*** | [0.00] | -1.43*** | [0.00] |
| *Ref: White collar (Father)* | | | | | | | | |
| Blue collar | -0.29*** | [0.00] | -0.02*** | [0.00] | -0.38*** | [0.00] | -0.42*** | [0.00] |
| Agriculture | -0.34*** | [0.00] | 0.24*** | [0.00] | -0.40*** | [0.00] | -0.29*** | [0.00] |
| *Ref: Working (Mother)* | | | | | | | | |
| Domestic | -0.26*** | [0.00] | -0.17*** | [0.00] | -0.29*** | [0.00] | -0.37*** | [0.00] |
| Not in labor force | -0.43*** | [0.00] | 0.63*** | [0.00] | 0.12*** | [0.00] | -1.31*** | [0.00] |
| *Ref: No sibling at school-going-age* | | | | | | | | |
| At least one | -0.02*** | [0.00] | 0.21*** | [0.00] | -0.27*** | [0.00] | -0.19*** | [0.00] |
| *Ref: No infant sibling* | | | | | | | | |
| At least one | -0.24*** | [0.00] | -0.32*** | [0.00] | -0.79*** | [0.00] | -0.73*** | [0.00] |
| *Ref: No adult sibling* | | | | | | | | |
| At least one | -0.13*** | [0.00] | -0.03*** | [0.00] | -0.15*** | [0.00] | 0.06*** | [0.00] |
| *MPCE* | 0.19*** | [0.00] | -0.02*** | [0.00] | 0.10*** | [0.00] | 0.12*** | [0.00] |
| Intercept | 1.86*** | [0.00] | 2.68*** | [0.00] | 2.68*** | [0.00] | 2.21 *** | [0.00] |

Table 3.A.1: Logistic regression: *Starting and finishing elementary education on time*[a]

[a] *p*-values are in square brackets. (***, **, *) are for 1%, 5% and 10% level of significance, respectively.

| Outcomes → | Younger cohort (*6-10* yrs.) | | | | Older cohort (*11-15* yrs.) | | | |
|---|---|---|---|---|---|---|---|---|
| | *2004-05* | | *2011-12* | | *2004-05* | | *2011-12* | |
| *Ref: Non-SC/ST* | | | | | | | | |
| OBC | -0.29*** | [0.00] | -0.17*** | [0.00] | -0.07*** | [0.00] | -0.11*** | [0.00] |
| SC/ST | -0.42*** | [0.00] | -0.14*** | [0.00] | -0.22*** | [0.00] | -0.32*** | [0.00] |
| *Ref: Hindu* | | | | | | | | |
| Muslims | -0.38*** | [0.00] | -0.50*** | [0.00] | -0.43*** | [0.00] | -0.60*** | [0.00] |
| Others | -0.10*** | [0.00] | -0.31*** | [0.00] | 0.09*** | [0.00] | 0.04*** | [0.00] |
| *Ref: Male* | | | | | | | | |
| Female | -0.29*** | [0.00] | -0.06*** | [0.00] | -0.63*** | [0.00] | -0.42*** | [0.00] |
| *Ref: Urban* | | | | | | | | |
| Rural | 0.10*** | [0.00] | 0.03*** | [0.00] | 0.08*** | [0.00] | 0.32*** | [0.00] |
| *Ref: Secondary or more (Father)* | | | | | | | | |
| Below secondary | 0.05*** | [0.00] | 0.17*** | [0.00] | -0.26*** | [0.00] | -0.23*** | [0.00] |
| Primary | 0.07*** | [0.00] | 0.23*** | [0.00] | -0.43*** | [0.00] | -0.74*** | [0.00] |
| Below primary | -0.30*** | [0.00] | 0.13*** | [0.00] | -0.67*** | [0.00] | -0.67*** | [0.00] |
| No schooling | -0.89*** | [0.00] | -0.70*** | [0.00] | -1.19*** | [0.00] | -1.28*** | [0.00] |
| *Ref: Secondary or more (Mother)* | | | | | | | | |
| Below secondary | -0.09*** | [0.00] | 0.21*** | [0.00] | -0.94*** | [0.00] | -0.51*** | [0.00] |
| Primary | 0.16*** | [0.00] | 0.01*** | [0.00] | -1.41*** | [0.00] | -0.85*** | [0.00] |
| Below primary | 0.21*** | [0.00] | -0.37*** | [0.00] | -1.72*** | [0.00] | -1.02*** | [0.00] |
| No schooling | -0.84*** | [0.00] | -0.78*** | [0.00] | -2.41*** | [0.00] | -1.75*** | [0.00] |
| *Ref: White collar (Father)* | | | | | | | | |
| Blue collar | -0.32*** | [0.00] | -0.11*** | [0.00] | -0.31*** | [0.00] | -0.39*** | [0.00] |
| Agriculture | -0.48*** | [0.00] | 0.21*** | [0.00] | -0.33*** | [0.00] | -0.35*** | [0.00] |
| *Ref: Working (Mother)* | | | | | | | | |
| Domestic | -0.19*** | [0.00] | -0.08*** | [0.00] | 0.16*** | [0.00] | 0.01*** | [0.00] |
| Not in labor force | -0.02*** | [0.00] | 0.63*** | [0.00] | -0.25*** | [0.00] | -0.37*** | [0.00] |
| *Ref: No sibling at school-going-age* | | | | | | | | |
| At least one | 0.22*** | [0.00] | 0.23*** | [0.00] | 0.18*** | [0.00] | 0.34*** | [0.00] |
| *Ref: No infant sibling* | | | | | | | | |
| At least one | -0.41*** | [0.00] | -0.72*** | [0.00] | -0.30*** | [0.00] | -0.36*** | [0.00] |
| *Ref: No adult sibling* | | | | | | | | |
| At least one | -0.25*** | [0.00] | -0.21*** | [0.00] | -0.24*** | [0.00] | -0.20*** | [0.00] |
| *MPCE* | 0.24*** | [0.00] | 0.10*** | [0.00] | 0.15*** | [0.00] | 0.26*** | [0.00] |
| Intercept | 2.33*** | [0.00] | 2.94*** | [0.00] | 3.92*** | [0.00] | 3.23*** | [0.00] |

Table 3.A.2: Logistic regression: *School attendance below 16 years*[a]

[a] *p*-values are in square brackets. (***, **, *) are for 1%, 5% and 10% level of significance, respectively.

| Outcomes → | Finishing lower secondary (*16-18* yrs.) | | | | School attendance (*16-18* yrs.) | | | |
|---|---|---|---|---|---|---|---|---|
| | *2004-05* | | *2011-12* | | *2004-05* | | *2011-12* | |
| *Ref: Non-SC/ST* | | | | | | | | |
| OBC | -0.05*** | [0.00] | -0.08*** | [0.00] | -0.07*** | [0.00] | -0.09*** | [0.00] |
| SC/ST | -0.30*** | [0.00] | -0.52*** | [0.00] | -0.08*** | [0.00] | -0.37*** | [0.00] |
| *Ref: Hindu* | | | | | | | | |
| Muslims | -0.48*** | [0.00] | -0.68*** | [0.00] | -0.42*** | [0.00] | -0.66*** | [0.00] |
| Others | 0.09*** | [0.00] | -0.31*** | [0.00] | 0.06*** | [0.00] | -0.16*** | [0.00] |
| *Ref: Male* | | | | | | | | |
| Female | -0.07*** | [0.00] | -0.03*** | [0.00] | -0.31*** | [0.00] | -0.16*** | [0.00] |
| *Ref: Urban* | | | | | | | | |
| Rural | -0.31*** | [0.00] | -0.11*** | [0.00] | -0.18*** | [0.00] | 0.03*** | [0.00] |
| *Ref: Secondary or more (Father)* | | | | | | | | |
| Below secondary | -0.52*** | [0.00] | -0.57*** | [0.00] | -0.51*** | [0.00] | -0.48*** | [0.00] |
| Primary | -0.89*** | [0.00] | -0.88*** | [0.00] | -0.98*** | [0.00] | -0.73*** | [0.00] |
| Below primary | -1.05*** | [0.00] | -0.96*** | [0.00] | -1.22*** | [0.00] | -1.11*** | [0.00] |
| No schooling | -1.37*** | [0.00] | -1.27*** | [0.00] | -1.41*** | [0.00] | -1.28*** | [0.00] |
| *Ref: Secondary or more (Mother)* | | | | | | | | |
| Below secondary | -0.66*** | [0.00] | -0.31*** | [0.00] | -0.60*** | [0.00] | -0.79*** | [0.00] |
| Primary | -0.99*** | [0.00] | -0.50*** | [0.00] | -1.04*** | [0.00] | -1.00*** | [0.00] |
| Below primary | -1.19*** | [0.00] | -0.59*** | [0.00] | -1.21*** | [0.00] | -1.16*** | [0.00] |
| No schooling | -1.63*** | [0.00] | -1.09*** | [0.00] | -1.59*** | [0.00] | -1.42*** | [0.00] |
| *Ref: White collar (Father)* | | | | | | | | |
| Blue collar | -0.34*** | [0.00] | -0.27*** | [0.00] | -0.39*** | [0.00] | -0.49*** | [0.00] |
| Agriculture | -0.35*** | [0.00] | -0.28*** | [0.00] | -0.35*** | [0.00] | -0.42*** | [0.00] |
| *Ref: Working (Mother)* | | | | | | | | |
| Domestic | -0.05*** | [0.00] | -0.16*** | [0.00] | 0.18*** | [0.00] | 0.19*** | [0.00] |
| Not in labor force | -0.06*** | [0.00] | -0.29*** | [0.00] | -0.11*** | [0.00] | -0.09*** | [0.00] |
| *Ref: No sibling at school-going-age* | | | | | | | | |
| At least one | -0.30*** | [0.00] | -0.34*** | [0.00] | 0.23*** | [0.00] | 0.04*** | [0.00] |
| *Ref: No infant sibling* | | | | | | | | |
| At least one | -0.85*** | [0.00] | -0.98*** | [0.00] | -0.36*** | [0.00] | -0.76*** | [0.00] |
| *Ref: No adult sibling* | | | | | | | | |
| At least one | -0.08*** | [0.00] | -0.03*** | [0.00] | -0.07*** | [0.00] | -0.11*** | [0.00] |
| *MPCE* | 0.23*** | [0.00] | 0.16*** | [0.00] | 0.26*** | [0.00] | 0.07*** | [0.00] |
| Intercept | 1.10*** | [0.00] | 1.79*** | [0.00] | 0.98*** | [0.00] | 2.85*** | [0.00] |

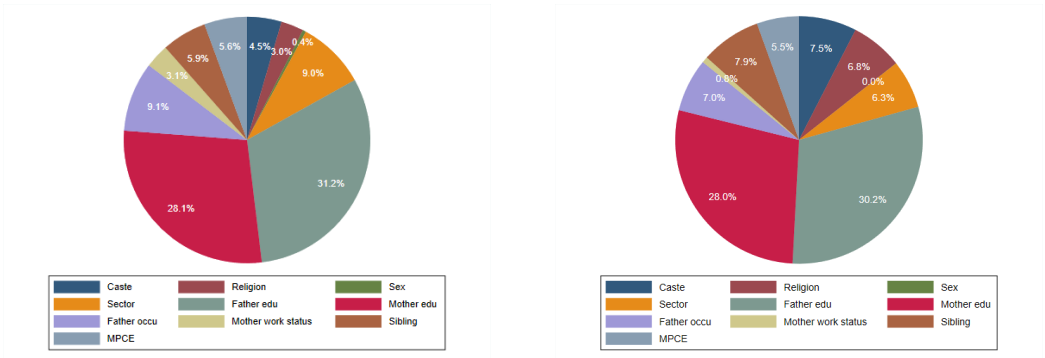
Table 3.A.3: Logistic regression: *Post-elementary education*[a]

[a] $p$-values are in square brackets. (***, **, *) are for 1%, 5% and 10% level of significance, respectively.

Table 3.A.4: HOI and *IOP* for elementary education: Regional measures

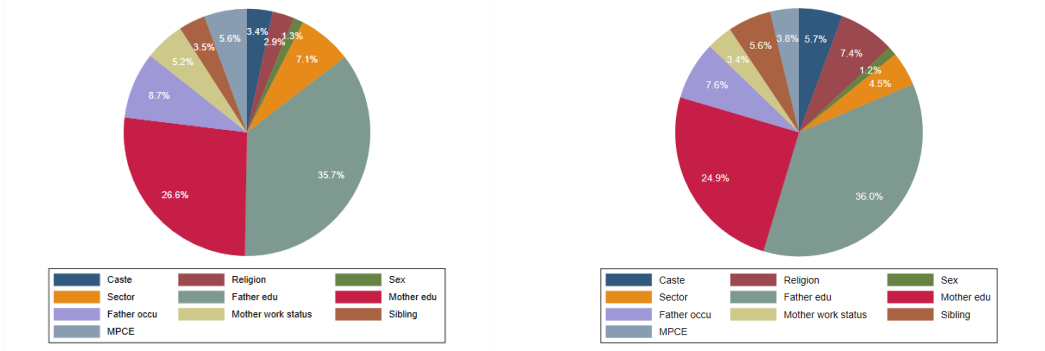
| | D-index | | Coverage | | HOI | |
|---|---|---|---|---|---|---|
| | *2004-05* | *2011-12* | *2004-05* | *2011-12* | *2004-05* | *2011-12* |
| **Starting elementary on time (6-7 yrs.)** | | | | | | |
| North | 5.9 | 4.6 | 81.2 | 88.5 | 76.4 | 84.4 |
| East | 9.9 | 4.7 | 65.6 | 82.1 | 59.1 | 78.3 |
| Central | 7.1 | 5.7 | 72.4 | 79.9 | 67.3 | 75.4 |
| North-East | 8.7 | 5.4 | 75.4 | 84.1 | 68.8 | 79.5 |
| South | 2.1 | 1.9 | 92.7 | 93.3 | 90.8 | 91.5 |
| West | 4.5 | 3.3 | 86.1 | 91.3 | 82.2 | 88.3 |
| **Finishing elementary on time (14-15 yrs.)** | | | | | | |
| North | 21.6 | 11.4 | 45.4 | 66.9 | 35.6 | 59.4 |
| East | 24.2 | 11.8 | 39.1 | 58.8 | 29.6 | 51.9 |
| Central | 27.1 | 16.7 | 34.7 | 50.4 | 25.3 | 41.9 |
| North-East | 20.1 | 10.1 | 55.7 | 66.9 | 44.5 | 60.2 |
| South | 11.2 | 5.3 | 71.7 | 85.4 | 63.7 | 80.9 |
| West | 11.1 | 7.1 | 72.3 | 76.8 | 64.2 | 71.3 |
| **School attendance: younger cohort (6-10 yrs.)** | | | | | | |
| North | 4.0 | 2.6 | 87.5 | 93.2 | 83.9 | 90.8 |
| East | 8.6 | 3.4 | 75.4 | 89.5 | 68.9 | 86.5 |
| Central | 5.8 | 4.2 | 81.1 | 87.6 | 76.4 | 83.9 |
| North-East | 4.7 | 2.4 | 86.5 | 93.7 | 82.5 | 91.4 |
| South | 1.7 | 1.2 | 95.1 | 96.6 | 93.5 | 95.5 |
| West | 3.7 | 1.9 | 91.0 | 95.8 | 87.6 | 94.0 |
| **School attendance: older cohort (11-15 yrs.)** | | | | | | |
| North | 6.5 | 2.9 | 85.4 | 92.1 | 79.9 | 89.5 |
| East | 9.8 | 4.8 | 74.5 | 86.8 | 67.2 | 82.7 |
| Central | 8.9 | 5.1 | 75.0 | 86.7 | 68.3 | 82.3 |
| North-East | 4.7 | 2.4 | 89.4 | 94.5 | 85.3 | 92.2 |
| South | 7.1 | 2.9 | 82.5 | 94.6 | 76.6 | 91.9 |
| West | 7.3 | 3.7 | 82.8 | 93.1 | 76.7 | 89.7 |

Finishing lower secondary education on time




(a) 2004-05 (b) 2011-12

School attendance: adolescent cohort (16-18)




(c) 2004-05 (d) 2011-12

Figure 3.A.1: Shapley decomposition of the D-index: Lower-secondary education

# Appendices

# Appendix A

# National Sample Survey

## *A.1 Foreward*

The National Sample Survey (NSS) is the oldest and one of the richest micro level database for India. Since the second World War, the Indian NSS is in fact marked as the first modern household survey in the world (Deaton 1997). The survey was first conducted under the supervision of Dr Prasanta Chandra Mahalanobis, the founder of the Indian Statistical Institute at Kolkata (then, Calcutta). All sampling and data processing activities were executed by the Indian Statistical institute since 1970, before this responsibility was brought under a single Government organization called the National Sample Survey Office (NSSO). This is the oldest organization in India that is responsible for the collection and publication of large scale national level surveys on multiple areas.

NSSO is now under the Ministry of Statistics and Program Implementation (MOSPI) that was formed as a single united ministry in 1999, upon merging of the two associated ministries, the Ministry of Statistics and the Ministry of Program Implementation. Two major 'statistical wing' of this ministry is the NSSO and the CSO or the Central Statistical Office. While CSO is responsible for all 'statistical activities' in the country (like estimation of quarterly GDP or conducting economic census), NSSO conducts several national, state and even district level 'socio-economic surveys'.

In the nomenclature of NSSO, it supervises certain *thick* sample surveys covering the whole country, as well as few other *thin* sample surveys that covers only selected part of India. The consumer expenditure and the employment-unemployment survey are some of the most important *thick* sample *household surveys* of NSSO, whereas surveys in unor-

ganized manufacturing, unorganized services and informal non-agricultural enterprise are some of the important *thick* sample *enterprise surveys*. The *thin* sample surveys are usually designed for certain target population and constitutes of particulars of slum, village facilities, land and livestock or living conditions of the tribal population in India. The present thesis uses the thick sample survey of the 'Employment-Unemployment Survey', the details of which are described below.

### A.1.1 Data: Coverage and scope

For the present thesis of inequality of opportunity in India, our data is drawn from the *Employment and Unemployment Survey* conducted by NSSO (*Schedule 10.0* survey in the nomenclature of NSSO). This quinquennial survey is one of the major thick sample survey of NSSO and its main focus is the labor market situation in India. It therefore provides records on several key characteristics of employment and unemployment patterns in India, both at the national and state level. The first employment-unemployment survey was conducted during the survey year 1972-73 that corresponds to the 27th round of NSS. Since then this survey have been conducted more or less regularly, once in every five years.

So far 9 such *rounds* of employment and unemployment survey have been conducted, namely, *27th* (September 1972-October 1973), *32nd* (July 1977-June 1978), *38th* (January 1983-December 1983), *43rd* (July 1987-June 1988), *50th* (July 1993-June 1994), *55th* (July 1999-June 2000), *61st* (July 2004-June 2005), *66th* (July 2009-June 2010) and *68th* (July 2011-June 2012). The only exception is the last round (*68*), that have been surveyed within two years of the previous one (*66*). The decision to conduct another large scale employment and unemployment survey shortly after the 66th round could be triggered by the fact that (2009-10) being a drought year, may have chances to bring out unnatural estimates of the general employment-unemployment scenario of India.

However data from only the *38th* round are disseminated by NSSO at present. For the present analysis of inequality of opportunity in India, we have taken data from six consecutive rounds of employment and unemployment (*Schedule 10*) survey, they are, 38 (1983), 43 (1987-88), 50 (1993-94), 55 (1999-00), 61 (2004-05) and 68 (2011-12). The

surveys on average cover 120,000 households, enumerating 0.4 to 0.6 million individuals. Total number of villages and urban blocks, households and individual surveyed for each round are provided in Table A.1. As for geographical coverage, all rounds cover the whole country except some inaccessible pockets. In particular, conflict areas of Ladakh & Kargil districts of Jammu & Kashmir, some remote interior villages of Nagaland in the North-East India, few unreachable areas of Andaman & Nicobar Islands in the Bay of Bengal located in the Eastern part of India and those villages recorded as uninhabited by respective population census are left out of the NSS thick sample coverage.

| Round | Year | Villages/Urban blocks | Household | Individual |
|---|---|---|---|---|
| 38 | 1983 | 12210 | 120921 | 623494 |
| 43 | 1987-88 | 12974 | 129194 | 667848 |
| 50 | 1993-94 | 11653 | 115409 | 564740 |
| 55 | 1999-00 | 10173 | 120578 | 596686 |
| 61 | 2004-05 | 12601 | 124680 | 602833 |
| 68 | 2011-12 | 12737 | 101724 | 456999 |

Table A.1: Survey summary[a]

[a]figures are from *NSS* reports of concerned rounds

The employment-unemployment survey, as mentioned before, is a thick sample 'household survey' of NSSO. The name 'household survey' attributes to the sampling unit. Unlike the enterprise surveys, the sampling design in the household surveys chooses sample households to conduct the associated survey and once the households are selected, all members of the selected households are enumerated in more details, depending on the particular aim of the survey. For example, while the consumer expenditure survey records consumption details of each member of the selected sample households, the employment-unemployment survey is rather focused on the occupation related details of the individuals. In almost all thick sample household surveys, NSSO provides two common 'blocks' of records. The 'household block' is reserved for the household characteristics like religion, social group (caste categories) or information on land holding. Whereas the 'individual block' records the demographic particulars like age, sex, relation to head, education level, school attendance, marital status or working status of each member of the household, whose records are already provided in the 'household block'.

As the name suggests, the employment-unemployment survey records the individual activity particulars in more details. In order to record information on both long-term and short-term employment, NSSO used three *reference frames* to record the working status of an individual. The current daily activity status (CDS) records the working activity prior to the day of the survey. Whereas the current weekly activity status (CWS) and the usual principal activity status (UPS) records the same for the last week and year respectively, prior to the date of the survey. Depending on their CWS or usual status, individuals are further marked as 'working' or 'non-working'. While details on job category including their specific occupation and industry codes are provided for individuals who are given the 'working' status, a follow up questionnaire records the details of the 'unemployed' respondents as well. Among the 'working' individuals, wage information however, is only recorded for the regular and casual wage earners, who are necessarily *not* self-employed. In addition, rounds of 38, 43 and 55 of the employment-unemployment survey provide records of migration particulars as well.

## *A.2 Details of survey rounds*

### A.2.1 Sampling frame

The employment-unemployment survey of NSS follows a complex multi-stage sampling design that is kept more or less similar over the different survey years, so that the data across the rounds are comparable. As mentioned in Table A.1, this survey of NSSO enumerates households both from the villages and the urban blocks, that consists the *First Stage Units* (FSU) of the respective survey. In particular, 'villages' are taken as per the respective national census and 'urban blocks' on the other hand are determined from the NSSO Urban Frame Survey[1]. Since each member of the selected households are the potential respondents, NSSO considers the households as the *Ultimate Stage Units* (USU) of the survey.

To facilitate the large scale thick sample surveys, the entire time frame of the survey (usually a year) is divided into four independent *sub-rounds* consisting of three months in each. Except for the 38th round, all the other rounds considered in this thesis are

[1]For the rounds covered here, three census have been conducted in the same time frame, in the years of 1981, 1991, 2001.

conducted over an agricultural crop year in India, that usually spans from July to June. Accordingly the 1st, 2nd, 3rd and 4th sub-rounds in those survey years constitutes the months of July to September, October to December, January to March and April to June, respectively[2]. For an uniform sample spread over the entire survey period, an equal number of villages/urban blocks are allocated to each sub-rounds. The selection of total number of FSUs are executed in the form of two independent sub-samples further. The additional sub-sample and sub-round wise sampling is to help the large data-base to be investigated from multiple independent dimensions, so that the margin of uncertainty (if any) in the combined final sample can be traced back by comparing samples from the sub-rounds or sub-samples.

The complex sample design adopted by most of the thick sample surveys of NSSO, can broadly be categorized into two chunks - (i) stratification and allocation of the total FSUs (villages/urban blocks) and (ii) final selection of the FSUs for the survey in particular. The sampling design for the selected rounds of the employment-unemployment survey considered here are similar, except for the 55th round (1999-00), that adopts a different frame of the so-called 'round-sampling scheme'. The common sampling frame as well as the particular sampling scheme of the 55th round are discussed below.

**Stratification and allocation of FSU**

At present, India is divided into twenty-nine *States* and seven *Union Territories* (UT), that are further divided into several *districts* for administrative purpose. NSSO nevertheless have its own stratification scheme for the purpose of survey, that assures an unbiased allocation of villages/urban blocks covering all parts of India. At this stage of sampling, the whole country is therefore divided into several state and district level strata, and the total number of FSUs are allocated in a way that ensures more or less equal number of FSUs for each separate stratum. Once the FSUs are uniformly allocated, the villages/urban blocks are then *selected* by some specific sampling rule in the next stage. All members of the selected FSUs are then enumerated according to the specific survey questionnaire.

[2]For the 38th round, that spans from January to December of 1983, the 1st, 2nd, 3rd and 4th sub-rounds consists of January to March, April to June, July to September and October to December, respectively.

The allocation of total FSUs is executed in several independent and simultaneous phases. A basic binary stratification is formed first, whereby each state/UT, as well as each district is divided into two separate stratum - rural and urban. The only exception to the binary stratification is for villages/urban blocks with more than one million population (as per census) where a separate third or fourth strata is formed as well, based on population. At the initial level, FSUs are allocated to each state/UT on the basis of their respective population as per the corresponding census. So naturally larger number of FSUs are allotted to the relatively populous states of the country. The allotted FSUs within each state are then simultaneously allocated to the rural and urban parts of that state as well, such that either sectors of the state have roughly equal number of allotted FSUs[3].

To ease the process of uniform allocation, NSSO divides each state into the rural and urban *sector*, as well as each district into the rural and urban *stratum*. To keep parity between the rural and urban sample size (in terms of allotted FSUs), NSSO impose some further allocation rules. Allocations are made in a way such that a minimum of 16 FSUs have been allotted to each state/UT, with a minimum of 8 FSUs in each of the rural and urban sector of that state/UT. Therefore a minimum of 8 villages and 8 urban blocks have to be allocated for each state/UT. The allocation further needs to make sure that each of the rural and urban stratum within a district have a minimum allocation of 4 FSUs as well. Together, the final allocation in a state/UT ensures to allocate a minimum of 8 FSUs for each sector (rural-urban) such that each stratum (rural-urban) of each district of that state have a minimum allocation of 4 FSUs[4]. This allocation then provides an organized ground for selecting the sample FSUs for the survey.

**Selection of FSU or villages/urban blocks**

For the purpose of selecting the sample FSUs from the total number of allocated FSUs, each of the rural and urban stratum of a district are further divided in suitable number

[3]India is predominantly rural where the rural-urban population ratio is nearly 3 : 1. Therefore to keep parity between the rural and the urban sample size, the allocation is done with relatively more weightage to the urban sector. However states with relatively large urban sector (like Maharastra, Tamil Nadu) are exempted from this allocation rule, where the FSUs are allotted in a way such that the rural and urban sectors of these states have roughly equal number of samples (villages/urban blocks).

[4]Depending on the sample size of the district, the number of minimum FSUs are further adjusted to a multiple of 4, if needed.

of *sub-stratum*. Let '$r$' and '$u$' be the number of FSUs allotted to the rural and the urban stratum, respectively. In that case a total of $r/2$ and $u/2$ number of sub-strata are formed for each of the rural and the urban stratum respectively, such that each sub-stratum comprises of more or less equal number of villages/urban blocks.

Provided the minimum allocation of 4 FSUs in each stratum, a minimum of 2 sub-strata is therefore formed for each of the rural and the urban stratum. In this case, all the 4 FSUs (with 2 FSUs from each sub-stratum) are selected for surveying. In case of more than 2 FSUs formed within a sub-stratum, 2 FSUs are selected by the probability proportional to size with replacement (PPSWR) for the rural part and by the simple random sampling without replacement (SRSWOR) for the urban part. This selection rule is universal to all rounds except for the 55th round, where sample villages and urban blocks are selected by the systematic circular sampling.

**Stratification of large FSU and selection of hamlet-groups/sub-blocks**

In case of large FSUs with relatively more population, the selected FSUs are further sub-grouped as an intermediate stage of sampling procedure. Large villages/urban blocks with population 1200 or more, are divided into certain number of *hamlet-groups* (*hg*) for the rural sector and *sub-blocks* (*sb*) for the urban sector[5]. For FSUs with population less than 1200, no hamlet-groups/sub-blocks (hg/sb) is formed and in that case an hg/sb number of 1 is assigned to that FSU. But in case of large FSUs, 2 hg/sb are selected by SRSWOR, whenever they have been formed. This intermediate stage of sampling therefore takes care of the large FSUs with huge population by selecting a portion of that respective village/urban block in terms of suitable number of sub-groups (hg/sb). Once the FSUs are selected for the survey, the next step is to select the households as the Ultimate Stage Units (USU) of the survey, which is done independently for each of the selected hg/sb.

**Selection of USU or households**

All households listed in the selected FSU or hg/sb are stratified into two to three *second stage strata* (SSS) based on the relative affluence of the respective household. The top

[5]The criteria for the suitable number of *hg/sb* is mainly based on population and the detailed criteria is provided in the final reports of NSSO for each round.

layer of SSS is formed by the most affluent households and the last layer constitutes of the least affluent ones. As expected, the criteria of 'affluence' is different for the rural and the urban sector, where for the latter affluence is determined mostly on the basis of monthly per capita expenditure of the household. The level of affluence in the rural sector on the other hand, are based upon the possession of certain luxury goods (like television, refrigerator, tractor for example)[6].

Certain number of households are then selected by specific sampling criteria from each of the SSS, depending on the household density in each SSS as well as adjusting for shortfalls in the required number of sample households. For the last two rounds of 61 and 68 for example, three SSS are formed and a total of 10 households are selected by SRSWOR from each selected FSUs, such that (2, 4, 4) households are selected from each of the (SSS-1, SSS-2, SSS-3) respectively, when no hg/sb is formed. For large FSUs for which two hg/sb are selected, (1, 2, 2) households are selected from each of the (SSS-1, SSS-2, SSS-3) respectively, separately for either hg/sb. All members of the selected households are then surveyed based on the specific questionnaire.

**The special case of round 55**

Because of the experimental adaptation of the *round sampling scheme*, the sampling design of the 55th round is a bit different from that of the other rounds. The purpose of this special scheme is to facilitate the resurveying of some of the selected households as a follow-up, for a selected portion of the same employment-unemployment questionnaire. One of the major structural difference of the 55th round is that, it draws on two independent sub-samples from each of the four sub-rounds. Therefore unlike the other rounds, the 55th round have a total of 8 sub-samples instead of 2.

Further, the stratification of the selected FSU into suitable number of hg/sb in this round, is not purely based on the size of the population but also on the relative concentration of the number of non-agricultural enterprises in the associated FSU. Unlike the other rounds, hg/sb with the maximum concentration of enterprises is selected with certainty, which are grouped as *segment-1* in this round. From the remaining hg/sb, 2

[6]Details of SSS criteria differs with rounds and are provided in the respective NSS employment-unemployment survey reports for each round.

more hg/sb will be selected by systematic circular sampling, which is called as *segment-2*. So the further stratification of the hg/sb into SSS in this round, is to stratify the above mentioned *segments* that consists of both households and enterprises. All households in the selected *segment* are stratified into two *SSS* based on the relative affluence. In rural areas, households are considered affluent upon possession of certain luxury or semi-luxury assets like motorcar, television, jeep, tractor, van, bus or telephone, and in the urban areas affluence is determined from monthly per capita expenditure similar to the other rounds. But in addition, all enterprises in each *segment* are also separately stratified into suitable number of *SSS* based on the broad industry groups and enterprise class.

**Multipliers**

Multipliers or sampling weights are crucial to any survey data to get the population projection of the sample estimates. As mentioned before, the total sample size in any thick sample survey of NSS is drawn in the form of two independent sub-samples (technically called interpenetrating sub-samples) and four independent sub-rounds as well. Accordingly NSSO provides two kind of sampling weights - the sub-sample multiplier and the sub-round multiplier. In order to ease the process, these multipliers are computed in a manner such that simple aggregation can generate the sample weights for the total surveyed sample. Along with the data for each round, NSSO provides instructions on how to combine the two provided multipliers if the analysis is based on the total surveyed sample. On the other hand if certain sub-round or sub-sample are the main subject of statistical analysis, then the respective sub-sample or sub-round specific sample weights can be used.

## *A.3 Data cleaning*

### A.3.1 Data processing

**Data availability:** Data provided by the NSSO socio-economic surveys are confidential micro level database for India and can be purchased from the Deputy Director General of the computer center in the Ministry of Statistics and Program Implementation, Government of India[7]. However at present, NSSO disseminates data for any survey only since

[7]Details of data purchase is available here: http://mospi.nic.in/data-dissemination

the 38th round, corresponding to the survey year 1983.

**Data format:** Most of the NSS data are usually supplied in fixed format ASCII files. Separate ASCII files are provided for the total number of state/UT for each round. Depending on the round specific questionnaire, each round of the employment-unemployment schedule (the one used for this thesis) is divided into several 'blocks'. Since 'blocks' are separated by data on different batch of variables, it may be sufficient to extract only certain 'blocks' that are relevant for the research question at hand. Some 'blocks' (for example, blocks - 0, 1, 2, 10, 11) are nevertheless common to any thick sample household surveys that records particulars of field operation and remarks of the supervising officers. Data is recorded for blocks 0 to 9 and for some rounds, set of different 'blocks' are sometimes separated by different 'levels' as well.

**Data extraction:** Along with the data, NSSO also provides a detailed layout of the fixed format data files, as well as the original questionnaire (in English) used for the survey. The ASCII data files can be transferred to the necessary statistical software by creating a suitable data dictionary using the layout provided with the data. However this straightforward but tedious process of data extraction is no longer needed for working with most of the NSS data lately. Since 2016, NSSO provides a specific 'toolkit' from the International Household Survey Network (IHSN), along with the purchased data. This micro-data management toolkit is developed by the World Bank and others for addressing technical issues regarding large sample surveys and to archive data in an internationally comparable format. By virtue of this toolkit one can readily transfer any part of the NSS survey data to a number of statistical software like STATA, SAS or SPSS. We only need to transfer the selected 'blocks' that are relevant for the present analysis of inequality of opportunity, as discussed below.

**Relevant blocks:** For analyzing inequality of opportunity in India, we are primarily interested in household and demographic particulars of the respondents along with the individual specific activity details. Accordingly for each different round we need to extract data from three 'blocks', that we can call as the household block, the individual block and the activity block. All the three blocks as mentioned above have different number of observations as the unit of reporting is different for each of them. The household block

consists of household particulars like religion, social group, total member of the household or land holding and therefore each observation in this block correspond to a household. The individual block on the other hand, records the demographic particulars (age, sex, education for example) of each member of the household and accordingly each observation in this block correspond to an individual. As mentioned before, the employment-unemployment schedule reports individual activity in more details and therefore each individual is allowed to report more than one principal activities[8]. The activity block reports the details of each of these activities pursued by an individual and each entry in this block therefore correspond to an activity. After extracting these relevant 'blocks' the next step is to merge them suitably as a single master-data, so that the each observation in the constructed master-data correspond to an individual along with his/her necessary demographic, household and activity details, as described in the following section.

### A.3.2 Constructing master-data

As far as analyzing inequality of opportunity is concerned, we need the household and the individual blocks for getting as may as 'circumstance variables' as possible (caste, sex, region etc.) along with some of the most important 'outcome variables' (consumption expenditure, education or school attendance). Whereas the main purpose of the activity block is to get the employment status of an individual (whether employed, unemployed or not in the labor force for attending domestic duties etc.), as well as data on the particular occupation classification codes for employed individuals and most importantly, for getting information on wage who are working in a casual or regular wage-earning occupation. We therefore extract the above mentioned blocks into STATA and proceed to construct the master-data for each round separately. The main challenge in this step is to merge these uneven blocks in a suitable way, such that each entry in the master-data corresponds to an individual along with all the necessary information from all blocks.

Therefore a necessary first step in construction of the master-data is to generate a unique identification number (id) for each block, so that the household-id uniquely iden-

[8]The activity that had been pursued by majority of the year and/or week are reported as the principal activities. Data on additional subsidiary activities for the selected eligible individuals are also provided in a separate block, which we did not consider in the present analysis of inequality of opportunity in India.

tify each different households (in the household block), the individual-id does the same for each person interviewed (in the individual block) and the activity-id uniquely identify each activity reported (in the activity block). The unique identification variables are generated by grouping (`group`) the *key identifying variables*, that is provided for each 'block' of the employment-unemployment data schedule[9]. Clearly, the total number of household-id should be less than that of the individual-id which in turn is less than the aggregate activity-id. We need to merge these three blocks based on their respective unique identification numbers, so that the total number of observations in the master-data correspond to the total number of individual-id.

We start by compressing the activity block to individual level, so that each observation of the compressed activity block is able to identify one activity for each unique individual. Activity block records particulars of daily activity/activities of each individual for each day of the reference week. Further for each day, one individual is allowed to pursue either one activity with full intensity (four hours or more) or two activities with half intensities each. An entry of 1.0 corresponds to a full intensity activity and 0.5 to a half intensity one. The total number of days engaged in all the activities in the reference week are then reported by summing these intensities. Borrowing from Hnatkovska, Lahiri & Paul (2012), we choose to keep the 'main' activity of the respondent and discard the others. We choose the 'main' activity as the one that is reportedly pursued by the individual for the maximum number of days in the reference week. However, some individuals may still have more than one 'main' activity, if equal number of days are spent by him/her on two or three activities simultaneously, in the same reference week. In case of multiple 'main' activities, we prioritize the wage earning activity along with a valid occupation code. Further duplicates in the 'main' activity, if remains, we prioritize the activity that yields higher wage. We compress the activity block by removing all other activities that are not the 'main' activity of the respondent, so that in the compressed activity block each unique individual now has exactly one 'main' activity.

We then merge (`merge`) the household and the individual block appropriately, so that this intermediate merged data does *not* have any duplicate individual-id but does have duplicate household-id. Each individual in the intermediate merged data so far have

[9]Texts in the parenthesis refers to the associated particular STATA (`command`) used.

their respective demographic particular along with the details on their corresponding households. The master-data is constructed upon merging the intermediate merged data to the compressed activity block, so that every observation in the master-data corresponds to an individual who have detailed information on his/her household characteristics, demographic details as well as activity particulars[10].

The construction of the master-data is only complete after repeatedly checking the summary statistics and distribution of several important variables, separately for the 'block' as well as in the merged data. For example, a correct merging is not supposed to alter the share of rural households in India and therefore should have the same value in the household block as well as in the master-data. Besides the master-data should also be checked for duplicate identification variables. Notice that the master-data for each round should not have any duplicate individual-id or activity-id, but many duplicate household-id. However as we did not take into consideration any survey specific reporting errors in our construction of the master-data, this is needed to be 'cleaned' for preparing this data-set for the purpose of empirical analysis.

### A.3.3 Preparing master-data for analysis

Unlike experimental data, where data entry is methodical and generated by controlled experiment, survey data is likely to be more 'messy' and needed to be 'cleaned' to prepare the data-set for any empirical analysis. This is particularly relevant for large sample survey database like NSS where the sample size often exceeds half a million observations and therefore data cleaning is crucial to our analysis to prepare the database void of any inconsistencies. Besides detailed information on a number of variables in the employment-unemployment survey also comes at the cost of numerous reporting errors due to lengthy questionnaire. There is no rule of thumb for data cleaning and that depends on the particular data structure instead[11]. For the present schedule of the employment-unemployment

[10] Except for rounds 38 and 43, daily activity particulars are reported for every enumerated individuals in all other rounds. For the 38th and 43rd round however, this is reported only for individuals in the labor force only (who are employed or unemployed). Therefore these rounds need some special treatment while merging for constructing the respective master-data. In particular, we need to make sure that each observation in the merged master-data for these rounds (38, 43) correspond to an unique individual with the necessary household and demographic details, as well as valid activity details if the respondent is in the labor force.

[11] *See* Deaton (1997), Hellerstein (2008) for some basic guidelines regarding data-cleaning of the large survey data-sets.

survey for various survey rounds, we extensively exploit the data cleaning strategies of Hnatkovska et al. (2012) who uses the same database for a different empirical exercise.

We consider six consecutive rounds of the NSS employment-unemployment survey as - 38 (1983), 43 (1987-88), 50 (1992-93), 55 (1999-00), 61 (2004-05) and 68 (2011-12), where the respective survey years are in parenthesis. First of all we make sure that all these rounds have uniform state codes over the span of 29 years from 1983 to 2012. India has 35 states/union territories at present, but some states were not yet generated in earlier rounds. For example, Chattisgarh, Jharkhand and Uttaranchal were created in 2000, whereas Daman & Diu was together with Goa before 1988. Accordingly we generate uniform state codes by moving Chattisgarh to Madhya Pradesh, Jharkhand to Bihar, Uttaranchal to Uttar Pradesh and Daman & Diu to Goa. For the purpose of regional analysis (whenever needed), we regroup the states in six broad regions as - North, East, Central, North-East, South, East[12].

One of the common problem in any survey database comes from the reporting of the missing data. Also for the present NSS survey (*schedule 10*), the reporting of missing information varies with different variables. Although data is missing for many of our necessary variables including education, occupation or wage, we prioritize those variables first, that are supposed to be strictly non-missing for any individual and is crucial to our analysis. We thereby drop observations with missing and miscoded values in age, sex, sector, marital status and social group (caste). Therefore an individual is considered as an 'incorrect' unit if the any of the above mentioned specification is missing for him/her and is therefore dropped. On this account we have to drop about 200 to 400 observations on average, across the rounds.

With the so-called 'correct' reporting units, we proceed then to clean for their respective occupation categories, as we need this information later for filtering our wage information as well as to construct our adult working samples. NSSO reports occupation categories

[12]State wise composition: Jammu & Kashmir, Himachal Pradesh, Punjab, Haryana and Uttarakhand - constitutes *North*; Bihar, Jharkhand, Orissa, West Bengal - constitutes *East*; Uttar Pradesh, Rajasthan, Madhya Pradesh, Chattisgarh - constitutes *Central*; Sikkim, Arunachal Pradesh, Assam, Nagaland, Meghalaya, Manipur, Mizoram, Tripura - constitutes *North-East*; Karnataka, Andhra Pradesh, Tamilnadu, Pondichery, Kerala, Lakshadeep - constitutes *South* and Gujrat, Daman & Diu, Dadra & Nagar Haveli, Maharashtra, Goa - constitutes *West*.

of each working individual by the three digit codes of National Classification of Occupation (NCO). The complete list of NCO is the responsibility of the Ministry of Labor and Employment, Government of India, and is constructed in a way so that the occupation classification in India is aligned with the International Standard Classification of Occupation (ISCO) as constructed by the International Labor Organization (ILO). The first NCO coding in India was made in 1948, followed by subsequent changes in 1958, 1968, 2004 and 2015, to keep abreast with the associated changes in ISCO. Except for the latest round (68), all other rounds considered here have used the NCO-1968 classification which corresponds to the ISCO-66 classification. But to account for many drastic changes in the labor market and to keep the NCO classification perfectly aligned to ISCO-88, the 68th round of the employment-unemployment survey have used the NCO-2004 classification with quite a few changes in the occupation codes. However in order to preserve comparability across all the rounds we convert the new NCO code (NCO-2004) to the old one (NCO-1968), with the help of a concordance table provided by NSSO[13].

Prior to converting to an uniform NCO classification for all rounds, we consider the valid occupation lists for each round separately. The NCO codes are numeric three digit codes that are reported separately for both weekly and yearly occupation of each respondent. But in addition to the conventional numeric NCO, NSSO also provides some non-numeric NCO codes corresponding to individuals whose particular occupation can not be classified by the three digit numeric NCO codes. Since NCO provides an extensive list of numerous kind of occupations, we did not consider the (small) set of individuals with unidentified occupation category and therefore discard ones with a non-numeric NCO code. On average, 300 to 500 observations are dropped further on this account, across all rounds[14].

For each rounds we further regroup the occupation codes into three broad occupation categories as - white collar, blue collar and agricultural occupation. The three digit coding of NCO uses a hierarchical occupation family structure, where the centennial

[13]The concordance table does not provide a one-to-one correspondence between the old and the new NCO codes and rather be used as a guideline for this conversion.

[14]Examples of non-numeric NCO are X99, X01, X09, X02, X10. We did this by exploiting the `sieve` command of STATA and then generate a new variable indicating the 'length' of such numeric NCO codes. We simply then drop (`drop`) observations with 'length' equal to one or two; thus keeping only the three digit valid numeric NCO codes or no NCO codes (but not 'wrong' NCO codes).

digit corresponds to an occupation family and the next two digits stand for the specific jobs within that occupation family. We therefore regroup our occupation categories based on the centennial digit and our exact occupation mapping is provided in Table A.2. Since occupation codes for the 68th round is different than that of the other rounds, we next proceed to convert the NCO-2004 codes of this round to NCO-1968 ones (used by the other rounds), as far as practicable.

| Centennial NCO code | Occupation description | Our category |
|---|---|---|
| 0-1 | Professional technician and associate professionals | white collar |
| 2 | Legislators, senior officials and managers | white collar |
| 3 | Clerks and others | white collar |
| 4 | Service workers | blue collar |
| 5 | Sales workers | blue collar |
| 6 | Farmers, fisherman, hunters, loggers and others | agricultural |
| 7-8-9 | Production and related workers, transport equipment operators and laborers | blue collar |

Table A.2: Occupation coding[a]

[a]Description of the occupation family corresponding to the centennial digits, are from the NCO code list provided by NSSO.

As mentioned before, the NCO code was updated for the 68th round with some notable changes, in order to keep parity with the changes in the ISCO codes. For example, to emphasize on skills, agriculture related managerial jobs in this round belong to the white collar category instead of the agricultural category. Conductors, guards, 'daftary', peon, draftsman, astrologers and palmists are identified as blue collar workers in NCO-2004 instead of white collar ones. On the other hand, safety and quality inspectors, finance and sales associated professionals, customs and border workers, police inspectors and detectives are now classified as white collar laborer instead of blue collar ones. It is not possible to take care of all the changes and we therefore tried to make the new NCO-2004 codes as comparable as possible to the one used by the other rounds (NCO-1968). Accordingly, we convert astrologers and palmists (515), travel attendants and guides (511), personal care and related workers (513) to the white collar job category, finance and sales associated professionals (341), safety and quality inspectors (315), police inspectors and detectives (345) to the blue collar job category and agriculture, fishery and related laborers (920) to the agricultural job category to have comparable occupation structure across all rounds[15].

[15]Numeric figures in parentheses are the corresponding three digit NCO-2004 codes.

Since most of the survey data suffers from various kind of response errors or bias in the reporting of the income-expenditure data, those variables demand special attention while cleaning. Once the occupation data is 'cleaned' satisfactorily, we therefore consider the cleaning of two of the most important variables in our analysis, wage earning and consumption expenditure, as follows.

NSSO conducts an independent thick sample survey of the Consumer Expenditure Survey, that provides detailed records of individual consumption. However, considering the importance of consumption expenditure, the employment-unemployment survey also provides this information, not for all individuals but for all the surveyed households. Therefore consumption is reported in this survey (schedule 10) as the monthly consumption per capita consumption expenditure (MPCE), that is the total expenditure incurred by the household over the last month prior to the date of the survey. The reporting of the MPCE however differs across the rounds of the employment-unemployment survey considered here.

Prior to the 55th round, the two major thick sample surveys of NSSO, that of the Consumer Expenditure Survey (schedule 1) and the Employment-Unemployment Survey (schedule 10), are conducted on the same set of households. Since the schedule-1 survey also provides data on the monthly household consumption, the value of MPCE is simply copied in the employment-unemployment survey from the detailed consumer expenditure survey, for the same reference month. The main purpose of conducting two of the most important thick sample surveys over the same set of sample households is to exploit the economies of scale and to reduce the survey expenditure. However for the simultaneous enumeration of both the schedule-1 and schedule-10 survey, interview time for each households increases considerably that eventually affects the accuracy of either survey. Therefore since the 55th round, these two very important thick sample surveys are canvassed on independent set of households.

Considering the importance of MPCE, it is nevertheless reported in the employment-unemployment survey anyway. However MPCE for the later three rounds (55, 61, 68) is reported differently. Since the 55th round, a separate mini proforma worksheet to collect information on the household consumption expenditure on thirty different items have

been appended to the employment-unemployment questionnaire, to have the records of MPCE in the schedule-10 survey. The proforma worksheet is prepared to cover the most important items of household consumption such that the total interviewing time on this worksheet does not exceed 15 minutes. The important items are selected in a way to keep parity with the consumption expenditure survey and varies a little over different rounds. MPCE is reported in Indian Rupee (INR) and we generate an individual level value of MPCE upon dividing it by the respective household size[16]. Except for the 55th round, MPCE is always reported for a monthly recall period that records the consumption expenditure of all the listed items that is incurred by the household over the month prior to the survey date. For the 55th round however, certain items regarding expenditure on health and education are experimentally recorded for the last year prior to the date of the survey. This experimentation with recall period is sometimes held responsible for unusual results, particularly for a low consumption inequality, in this round (Dreze et al. 1999).

Wage in the NSS employment-unemployment survey is reported in INR against selected individuals only, who are identified to be occupied in the regular or casual wage earning jobs. Regular jobs are defined as any profession for which the employee gets a steady monthly remuneration. Casual jobs on the other hand includes several short-term works (Government or non-Government) for which the payment is usually made on a daily or weekly basis. Either of the regular or casual jobs however, are necessarily concerned with employees working in other's enterprises. Information on wage is therefore not available for the self-employed workers, who constitute about 35-40% of the total working individuals. Unlike MPCE, wage is always reported with a weekly recall period that records the weekly wage of an activity received or receivable over the reference week. Wage is therefore reported against each regular/casual wage earning activity and one individual may have more than one wage data reported against him/her, prior to the construction of the master-data. However in the master-data, wage is reported only for the 'main' activity, by construction. Since every employed individuals including the regular/casual workers, must have an occupation code, wage data is further filtered based on the valid weekly and yearly occupation codes of NCO.

[16]For some rounds MPCE is reported in *paise* and therefore needed to be divided by 100 to have its value in INR.

As mentioned before, the schedule-10 survey of NSSO provides three kinds of 'activity status' based on different reference period, namely, the current weekly activity status (CWS), the current daily activity status (CDS) and the usual principal activity status (UPS). The reporting of these activity status however varies a little across rounds. Except for rounds 38 and 43, the weekly as well as the yearly activity status (CWS and UPS, respectively) are determined on a priority cum major time criterion as described below.

To determine the usual principal activity status (UPS), all individuals are first divided into two mutually exclusive groups - (i) those who are in the labor force (LF) and (ii) those who are not in the labor force (NLF). A person is in the labor force by definition, if he/she is either employed or available to work if unemployed. Therefore individuals in LF includes both employed (E) and unemployed (UE) person. Respondents marked as NLF on the other hand, are neither E nor UE and are reportedly not available to a job opportunity. Therefore person in NLF includes those who are not in the labor force due to attending domestic duties or pursuing higher studies or being physically disabled or retired. Based on which group the individual belongs for the majority of the reference year, a person is identified as either LF or NLF. Among the individuals who are in the LF, one is considered as E or UE depending on whether the person remain employed or not for majority of the reference period. The criterion therefore prioritizes those who are in the labor force and identifies one as employed, based on the major time criterion.

Each individual is therefore identified as E, UE or NLF, based on their respective UPS. The employment/unemployment status is not however determined from the CDS, although the weekly activity status of an individual is provided by NSSO based on their respective CWS. The job details of each employed individuals (E) are then recorded further including their respective NCO codes. Accordingly weekly and yearly NCO codes are assigned, based on their weekly and yearly activity status. Notice that all employed individuals may not necessarily have both kind of NCO codes as a person may be considered as E under the weekly reference frame, but may not be so over the reference year. Further if an employee has recently changed his/her job, the weekly and yearly NCO for the same person may be different[17].

[17]However, a person marked as UE should have no NCO codes. At this point we therefore crosschecked whether the NCO codes correspond to the employed individuals only and drop any observation for whom the weekly (yearly) NCO code has an entry even when the person is reportedly unemployed over the

The criteria for determining the usual activity status (UPS) is a bit different for the earliest two rounds considered here, 38 and 43. Unlike the other rounds, the yearly employment status in these rounds is based on only the major time criterion and thereby does not prioritize individuals in the labor force. The individuals in these rounds are therefore classified in any of the three groups, E, UE or NLF, based on which group the person belong for the majority of the reference year. Accordingly usual yearly NCO is provided for all those individuals who are identified as E by this major time criterion. But weekly activity status (CWS) for these rounds (38 and 43) are determined in a similar fashion by the priority cum major time criterion, where person in LF are prioritized over those in NLF. Therefore to keep parity across all rounds, we consider the weekly NCO codes for rounds 38 and 43, whereas use usual yearly NCO codes for the rest of the rounds. We therefore consider the wage data as valid if the corresponding yearly NCO is non-missing for the corresponding individual as well, for rounds 50, 55, 61 and 68. We nevertheless have to exclude the 43rd round from any analysis involving wage as this round has unusually low rural wage observations. But for the 38th round we filter the wage data as the additional imposition of corresponding non-missing weekly NCO code[18]. Since wage in our master-data correspond to the main activity that have reportedly been pursued for the maximum number of days in the reference week, we calculate the daily wage as well upon dividing the weekly wage by the reported number of days engaged in that wage earning activity.

Finally, following Hnatkovska et al. (2012), both wage and expenditure data are converted to real terms after dividing by the state level absolute poverty lines, taking 1983 rural Maharashtra as the base[19]. Poverty lines are estimated separately for the rural and

week (year).

[18]Consider for example a person, who is reported as, say, E for 4 months, UE for 3 months and NLF for 5 months, in the reference year. Then his UPS would be NLF before round 50, but E since round 50. Accordingly, this particular person have no usual yearly NCO listed against him in rounds 38 and 43, but have a valid one since the 50th round. Because CWS was determined on a similar *priority come major time criterion* for all individuals prior to round 50, this person in example still have a valid weekly NCO against him (that is determined from his CWS). So while considering individuals with valid occupation, we actually consider valid weekly NCO for rounds 38 and 43 (as a proxy for usual NCO) and valid usual NCO for all the other rounds.

[19]In particular, we did not use the consumer price index (CPI) for converting the income-expenditure variables to their real values for two reasons. First of all CPI is estimated from the census data conducted by CSO, whereas poverty lines are estimated based on the consumption data collected by NSSO. Secondly, CPI does not have an aggregate level rual and urban index prior to 2011 and is instead provided for multiple series like urban non-manual labor, agricultural labor, rural labor and industrial workers.

urban sector of each state, and are made publicly available by the Planning Commission of India, since the sixth five year plan (1980). Poverty lines are estimated under a Government appointed *expert group*. So far over the course of time, three different expert groups have estimated poverty lines in India, namely, the Lakdawala group (1993), the Tendulkar group (2004) and the Rangarajan group (2012). Except for the 68th round, we have the Lakdawala estimates of poverty lines for all other rounds. Poverty lines for the latest round (68) however is estimated by the Tendulkar group, which mainly differs from the Lakdawala estimation strategy in their underlying 'consumption basket' to estimate the poverty line. While the Lakdawala methodology is focused on the minimum per capita expenditure, the Tendulkar estimates concentrate on the minimum calorie intake, to determine the basic 'consumption basket' for measuring the poverty line. So borrowing from Hnatkovska & Lahiri (2013), we convert the Tendulkar poverty line estimates to the Lakdawala estimates for the 68th round, using the poverty line estimates for round 61, for which both of the Lakdawala and Tendulkar estimates are provided.

NSSO provides education in several categorical codes, that we convert to suitable years of education for the purpose of analysis. We first regroup the given education codes into five broad categories as - (i) without formal schooling (ii) below primary schooling (iii) up to primary schooling (iv) above primary but below (lower) secondary schooling (correspond to middle level schooling or elementary schooling) and (v) lower secondary schooling or more. We assign 1, 2, 4 and 8 years of schooling to the first four categories. For the last category we assign different year of education depending on the round specific education information available for the post secondary schooling, which are then updated suitably on account of availability of information on additional technical education (for example, certain diploma or certificate courses that are popularly pursued in the country). 10-12 years of education are assigned to the lower and higher secondary level education, although not all rounds have separate information on higher secondary education. Similarly 15 and 16 years of education is attached to three-year and four-year graduate degrees. Table A.3 reports the basic mapping of our years of education to the respective education codes as provided by the data-set. Following Hnatkovska et al. (2012) 1-4 years of education are added to the basic year of education as reported in Table A.3, whenever certain level of technical education is available.

| Education code | Year of education |
|---|---|
| Without formal schooling | 1 |
| Below primary | 2 |
| Up to primary | 4 |
| Above primary but below (lower) secondary | 8 |
| Lower and higher secondary | 10-12 |
| Graduate | 15-16 |

Table A.3: Mapping year of education to NSS education codes

The cleaning is finalized after calculating the sample weights, separately for each round. Since we use the total surveyed sample in our analysis we calculate the combined multiplier as per the instruction of NSSO, that takes into account all sub-samples as well as all sub-rounds. Table A.4 reports the exact number of observations that we need to drop in order to prepare the master-data for our empirical analysis. Not unnaturally, older rounds are more 'noisy' and cleaning requires to drop more than 2000 individuals. The latest rounds in the twenty-first century seems to improve in terms of data reporting as implemented by fewer number of dropped observations.

| Round | Year | Obs. in cleaned data | Obs. dropped |
|---|---|---|---|
| 38 | 1983 | 621204 | 2290 |
| 43 | 1987-88 | 665221 | 2627 |
| 50 | 1993-94 | 563075 | 1665 |
| 55 | 1999-00 | 594786 | 1900 |
| 61 | 2004-05 | 602241 | 592 |
| 68 | 2011-12 | 456502 | 497 |

Table A.4: Survey summary in cleaned data

The master-data thus cleaned for all rounds are then sorted over all the key identifying factors and are sometimes appended across necessary rounds to have our combined or *pooled master-data* composed of all the required cleaned rounds. In this thesis we use rounds 61 and 68 for Chapter 1 and 3, whereas use all the six rounds for Chapter 2. Each of these chapters use different sample selection criteria, which are mentioned in the data description section of the associated chapters.

# Bibliography


Allison, P. D. (2000), 'Multiple imputation for missing data: A cautionary tale', *Sociological methods & research* **28**(3), 301–309.

Allison, P. D. (2003), 'Missing data techniques for structural equation modeling.', *Journal of abnormal psychology* **112**(4), 545.

Alon, S. (2009), 'The evolution of class inequality in higher education: Competition, exclusion, and adaptation', *American Sociological Review* **74**(5), 731–755.

Ambedkar, B. R. (2014), *Annihilation of caste*, Navayana publishing (Indian edition), London, New York.

Andreoli, F. (2018), 'Robust inference for inverse stochastic dominance', *Journal of Business & Economic Statistics* **36**(1), 146–159.

Andreoli, F., Havnes, T. & Lefranc, A. (2019), 'Robust inequality of opportunity comparisons: Theory and application to early-childhood policy evaluation', *The Review of Economics and Statistics* .

Arneson, R. (1989), 'Equality of opportunity and welfare ', *Philosophical Studies* **56**, 77–93.

Asadullah, M. N. & Yalonetzky, G. (2012), 'Inequality of educational opportunity in India: Changes over time and across states', *World Development* **40**(6), 1151–1163.

ASER (2007), 'Annual Status of Education Report (Rural)', *Pratham* .

Azam, M. (2012), 'A distributional analysis of social group inequality in rural india', *Journal of International Development* **24**(4), 415–432.

Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. (2011), 'Multiple imputation by chained equations: what is it and how does it work?', *International journal of methods in psychiatric research* **20**(1), 40–49.

Balcázar, C. F., Narayan, A. & Tiwari, S. (2015), 'Born With a Silver Spoon: Inequality in Educational Achievement across the World ', *Policy Research Working Paper* .

Beach, C. M. & Davidson, R. (1983), 'Distribution-free statistical inference with lorenz curves and income shares', *The Review of Economic Studies* **50**(4), 723–735.

Björklund, A., Jäntti, M. & Roemer, J. (2012), 'Equality of opportunity and the distribution of long-run income in Sweden ', *Social choice and welfare* **39**, 675–696.

Bose, N. (2017), 'Raising consumption through india's national rural employment guarantee scheme', *World Development* **96**, 245–263.

Bourguignon, F., Ferreira, F. H. & Menéndez, M. (2007), 'Inequality of opportunity in Brazil ', *Review of income and wealth* **53**(4), 585–618.

Broske, M. & Levy, H. (1989), *The Stochastic Dominance Estimation of Default Probability. In: Fomby T.B., Seo T.K. (eds) Studies in the Economics of Uncertainty*, Springer, New York, NY.

Brunori, P., Ferreira, F. & Peragine, V. (2013), *Inequality of Opportunity, Income Inequality, and Economic Mobility: Some International Comparisons*, Paus E. (eds) Getting Development Right; Palgrave Macmillan, New York.

Brunori, P., Hufe, P. & Mahler, D. G. (2018), 'The roots of inequality: Estimating inequality of opportunity from regression trees.'.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J. & Miller-Johnson, S. (2002), 'Early childhood education: Young adult outcomes from the abecedarian project', *Applied developmental science* **6**(1), 42–57.

Checchi, D. & Peragine, V. (2010), 'Inequality of opportunity in Italy ', *Journal of economic inequality* **8**, 429–450.

Checchi, D., Peragine, V. & Serlenga, L. (2010), 'Fair and unfair income inequalities in Europe', *IZA discussion paper No. 5025* .

Cogneau, D. & Mesplè-Somps, S. (2008), 'Inequality of opportunity for income in five countries of Africa', *John Bishop, Buhong Zheng (ed.); Inequality and Opportunity:*

*Papers from the Second ECINEQ Society Meeting, Emerald Group Publishing Limited* **16**, 99–128.

Cohen, G. A. (1989), 'On the currency of egalitarian justice ', *Ethics* **99**, 906–944.

Currie, J. (2001), 'Early childhood education programs', *Journal of Economic perspectives* **15**(2), 213–238.

Dabalen, A., Narayan, A., Saavedra-Chanduvi, J. & Hoyos Suarez, A. (2015), *Do African Children Have an Equal Chance? A Human Opportunity Report for Sub-Saharan Africa*, The World Bank, Washington, DC.

Dardanoni, V., Fields, G. S., Roemer, J. E. & Sanchez Puerta, M. L. (2005), *How demanding should equality of opportunity be, and how much have we achieved? , in S. Morgan, D. Grusky and G. S. Fields (Eds.), Mobility and inequality: Frontiers of research from sociology and economics (pp. 59-82).*, Stanford University Press., Palo Alto, CA.

Deaton, A. (1997), *The analysis of household surveys: a microeconometric approach to development policy*, World Bank Publications.

Deaton, A. & Dreze, J. (2002), 'Poverty and Inequality in India: A Re-Examination ', *Economic and Political Weekly* **37**(36), 3729–3748.

Deshpande, A. (2001), 'Caste at birth? redefining disparity in india', *Review of Development Economics* **5**(1), 130–144.

Deshpande, A. & Ramachandran, R. (2014), 'How backward are the other backward classes? Changing contours of caste disadvantage in India', *Centre for development economics, Delhi Schhol of Economics wp no. 233* .

Dev, S. M. & Ravi, C. (2007), 'Poverty and Inequality: All-India and States, 1983-2005', *Economic and Political Weekly* **42**(6), 509–521.

Dreze, J., Sen, A. et al. (1999), 'India: Economic development and social opportunity', *OUP Catalogue* .

Duraisamy, P. (2002), 'Changes in returns to education in india, 1983–94: by gender, age-cohort and location', *Economics of Education Review* **21**(6), 609–622.

Dutta, S. & Sivaramakrishnan, L. (2013), 'Disparity in the literacy level among the scheduled and non-scheduled population: Indian scenario in the 21st century', *Transactions* **35**(2), 185–200.

Dworkin, R. (1981*a*), 'What is equality? Part 1: Equality of resources ', *Philosophy & public affairs* **10**, 283–345.

Dworkin, R. (1981*b*), 'What is equality? Part 1: Equality of welfare ', *Philosophy & public affairs* **10**, 185–246.

Dyson, T. & Moore, M. (1983), 'On Kinship Structure, Female Autonomy, and Demographic Behavior in India', *Population and Development Review* **9**(1), 35–62.

Ferreira, F. H. & Gignoux, J. (2011), 'The measurement of inequality of opportunity: theory and an application to Latin America ', *The review of income and wealth* **57**(4).

Ferreira, F. H. & Peragine, V. (2015), 'Equality of Opportunity: Theory and evidence ', *Policy research working paper* (WPS 7217, Washington, D.C: World Bank Group).

Gang, I. N., Sen, K. & Yun, M.-S. (2008), 'Poverty in rural india: caste and tribe', *Review of Income and Wealth* **54**(1), 50–70.

Gang, I. N., Sen, K. & Yun, M.-S. (2011), 'Was the mandal commission right? differences in living standards between social groups', *Economic and Political Weekly* **46**(39), 24.

Gang, I. N., Sen, K. & Yun, M.-S. (2017), 'Is caste destiny? occupational diversification among dalits in rural india', *The European Journal of Development Research* **29**(2), 476–492.

Ghosh, M. (2006), 'Economic growth and human development in indian states', *Economic and Political Weekly* pp. 3321–3329.

Golsteyn, B. H., Grönqvist, H. & Lindahl, L. (2014), 'Adolescent time preferences predict lifetime outcomes', *The Economic Journal* **124**(580), F739–F761.

Gourishankar, V. & Sai Lokachari, P. (2012), 'Benchmarking educational development efficiencies of the indian states: a dea approach', *International Journal of Educational Management* **26**(1), 99–130.

Govinda, R. & Biswal, K. (2006), 'Mapping literacy in india: who are the illiterates and where do we find them', *Background paper prepared for the Education for All Global Monitoring Report* .

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007), 'How many imputations are really needed? some practical clarifications of multiple imputation theory', *Prevention science* **8**(3), 206–213.

Harris, T. R. & Mapp, H. P. (1986), 'A Stochastic Dominance Comparison of Water-Conserving Irrigation Strategies ', *American Journal of Agricultural Economics* **68**(2), 298–305.

Havnes, T. & Mogstad, M. (2015), 'Is universal child care leveling the playing field?', *Journal of public economics* **127**, 100–114.

Heckman, J. J. (2006), 'Skill formation and the economics of investing in disadvantaged children', *Science* **312**(5782), 1900–1902.

Heckman, J. J. (2011), 'The economics of inequality: The value of early childhood education.', *American Educator* **35**(1), 31.

Hellerstein, J. M. (2008), 'Quantitative data cleaning for large databases', *United Nations Economic Commission for Europe (UNECE)* .

Himanshu (2007), 'Recent Trends in Poverty and Inequality: Some Preliminary Results', *Economic and Political Weekly* **42**(6), 497–508.

Himanshu (2018), 'Widening Gaps: India Inequality Report 2018', *Oxfam India* .

Hnatkovska, V. & Lahiri, A. (2013), 'The rural-urban divide in India ', *Int Growth Centr Work Pap* .

Hnatkovska, V., Lahiri, A. & Paul, S. B. (2012), 'Caste and labor mobility ', *Applied economics* **4**(2).

Hnatkovska, V., Lahiri, A. & Paul, S. B. (2013), 'Breaking the Caste Barrier: Intergenerational Mobility in India', *Journal of human resources* **48**(2), 435–473.

Hothorn, T., Hornik, K. & Zeileis, A. (2006), 'Unbiased recursive partitioning: A conditional inference framework', *Journal of Computational and Graphical statistics* **15**(3), 651–674.

Hoyos, A. & Narayan, A. (2012), 'Inequality of opportunities among children: How much does gender matter? ', *World Bank Development Report* .

Jaffrelot, C. (2006), 'The impact of affirmative action in india: More political than socioeconomic', *India Review* **5**(2), 173–189.

Jong-Sung, Y. & Khagram, S. (2005), 'A comparative study of inequality and corruption', *American sociological review* **70**(1), 136–157.

Kijima, Y. (2006), 'Why did wage inequality increase? evidence from urban india 1983–99', *Journal of Development Economics* **81**(1), 97–117.

Kodde, D. A. & Palm, F. C. (1986), 'Wald criteria for jointly testing equality and inequality restrictions', *Econometrica: journal of the Econometric Society* pp. 1243–1248.

Krishna, A., Sekhar, M., Teja, K. & Swamy, B. (2017), 'Primary education during pre and post right to education (rte) act 2009: An empirical analysis of selected states in india', *International Journal of Humanities and Social Science Invention* **6**(11), 41–51.

Lefranc, A., Pistolesi, N. & Trannoy, A. (2008), 'Inequality of opportunities vs. inequality of outcomes: Are Western societies all alike?', *Review of income and wealth* .

Lefranc, A., Pistolesi, N. & Trannoy, A. (2009), 'Equality of opportunity and luck: definitions and testable conditions, with an application to income in France (1979-2000)', *Journal of public economics* **93**, 1189–1207.

Lewin, K. M. (2011), 'Expanding access to secondary education: Can india catch up?', *International Journal of Educational Development* **31**(4), 382–393.

Little, R. J. (1988), 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics* **6**(3), 287–296.

Madheswaran, S. & Attewell, P. (2007), 'Caste discrimination in the Indian urban labor market: Evidence from the National Sample Survey', *Economic and Political Weekly* **42**(41), 4146–4153.

Marchenko, Y. V. & Eddings, W. (2011), 'A note on how to perform multiple-imputation diagnostics in stata', *College Station, TX: StataCorp* .

Marrero, G. A. & Rodrìguez, J. G. (2011), 'Inequality of opportunity in the United States: trends and decomposition', *Research on Economic Inequality* **19**, 217–216.

Mas-Colell, A., Whinston, M. D. & Green, J. R. (1995), *Microeconomic Theory*, Oxford University Press.

Molina, E., Narayan, A. & Saavedra-Chanduvi, J. (2013), 'Outcomes, Opportunity and Development: Why Unequal Opportunities and Not Outcomes Hinder Economic Development ', *Policy Research Working Paper; 6735* .

Molinas Vega, J., Paes de Barros, R., Saavedra, J. & Giugale, M. (2011), *Do our children have a chance?; The 2010 Human opportunity report for Latin America and the Caribbean*, The World Bank, Washington, DC.

Munshi, K. & Rosenzweig, M. (2006), 'Traditional institutions meet the modern world: Caste, gender and schooling choice in a globalizing economy ', *American economic review* **96**(4), 1225–1252.

Munshi, K. & Rosenzweig, M. (2009), 'Why is mobility in India so low? Social insurance, inequality, and growth. ', *National Bureau of Economic Research, No. w14850.* .

Muralidharan, K. & Kremer, M. (2006), 'Public and private schools in rural india', *Harvard University, Department of Economics, Cambridge, MA* .

Narayan, A., der Weide, R. V., Cojocaru, A., Lakner, C., Redaelli, S., Mahler, D. G., Ramasubbaiah, R. G. N. & Thewissen, S. (2018), *Fair Progress? Economic Mobility across Generations around the World*, The World Bank, Washington, DC.

NCERT (2016), 'All India School Education Survey in 2009', *National Council of Educational Research and Training* .

NSSO (2008), 'NSS Report No. 533: Migration in India: July, 2007-June, 2008', *National Sample Survey Organization, Ministry Of Statistics and Program Implementation (MOSPI), Govt. of India* .

Paes de Barros, R., , Molinas-Vega, J. & Saavedra-Chanduvi, J. (2008), ' Measuring Inequality of Opportunity for Children ', *World Bank document* .

Paes de Barros, R., Ferreira, F., Molinas-Vega, J. & Saavedra-Chanduvi, J. (2009), *Measuring Inequality of Opportunity in Latin America and the Caribbean*, The World Bank, Washington, DC.

Pal, P., Ghosh, J. et al. (2007), 'Inequality in india: A survey of recent trends', *DESA Working Paper No. 45* .

Papola, T. S. (2012), 'Social exclusion and discrimination in the labor market', *Institute for Studies in Industrial Development, wp no. 2012/04* .

Peet, E. D., Fink, G. & Fawzi, W. (2015), 'Returns to education in developing countries: Evidence from the living standards and measurement study surveys', *Economics of Education Review* **49**, 69–90.

Peragine, V. & Serlenga, L. (2008), 'Higher education and equality of opportunity in Italy', *Research on Economic inequality* **16**, 67–97.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey methodology* **27**(1), 85–96.

Rai, A. (2014), 'Implementation of the rte act: Rte forum's stocktaking report', *Social Change* **44**(3), 439–449.

Ramos, X. & Van de Gaer, D. (2012), 'Empirical approaches to inequality of opportunity: Principles, measures and evidence ', *IZA discussion paper no. 6672* .

Rawls, J. (1971), *A theory of justice*, Cambridge: Harvard University Press.

Roemer, J. (1993), 'A Pragmatic Theory of Responsibility for the Egalitarian Planner', *Philosophy & Public Affairs* **22**, 146–166.

Roemer, J. (1998), *Equality of Opportunity*, Harvard University Press, Cambridge, MA.

Roemer, J. E. & Trannoy, A. (2013), 'Equality of Opportunity', *Cowles foundation discussion paper no. 1921* .

Roy, A. (2017), *The Doctor and the Saint: Caste, Race, and Annihilation of Caste, the Debate Between BR Ambedkar and MK Gandhi*, Haymarket Books.

Royston, P., White, I. R. et al. (2011), 'Multiple imputation by chained equations (mice): implementation in stata', *J Stat Softw* **45**(4), 1–20.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Rubin, D. B. (1986), 'Basic Ideas of Multiple Imputation for Nonresponse', *Survey Methodology, Statistics Canada* **12**(1), 37–47.

Saidi, A. & Hamdaoui, M. (2017), 'On measuring and decomposing inequality of opportunity in access to health services among Tunisian children: a new approach for public policy ', *Health and Quality of Life Outcomes* **15**(213).

Salehi-Isfahani, D., Hassine, N. B. & Assaad, R. (2014), 'Equality of opportunity in educational achievement in the middle east and north africa', *The Journal of Economic Inequality* **12**(4), 489–515.

Sánchez, A. & Singh, A. (2018), 'Accessing higher education in developing countries: Panel data analysis from india, peru, and vietnam', *World Development* **109**, 261–278.

Schafer, J. L. (1999), 'Multiple imputation: a primer', *Statistical methods in medical research* **8**(1), 3–15.

Schafer, J. L. & Olsen, M. K. (1998), 'Multiple imputation for multivariate missing-data problems: A data analyst's perspective', *Multivariate behavioral research* **33**(4), 545–571.

Sen, A. (1980), *Equality of what?*, The Tanner lectures on human values, S.McMurrin (ed.), University of Utah press, Salt Lake City.

Shorrocks, A. F. (1983), 'Ranking income distributions', *Economica* **20**(197), 3–17.

Shorrocks, A. F. (2013), 'Decomposition procedures for distributional analysis: a unified framework based on the shapley value', *The Journal of Economic Inequality* **11**(1), 99–126.

Singh, A. (2011), 'Inequality of Opportunity in Indian Children: the case of Immunization and Nutrition', *Population Research and Policy Review* **30**(6), 861–883.

Singh, A. (2012*a*), 'Inequality of Opportunity in Access to Primary Education: The Case of Indian Children ', *Population Review* **51**(1), 50–68.

Singh, A. (2012*b*), 'Inequality of opportunity in earnings and consumption expenditure: The case of Indian men', *The review of income and wealth* **58**(1), 79–106.

Teyssier, G. (2017), 'Inequality of opportunity: New measurement methodology and impact on growth', *Seventh ECINEQ Meeting, New-York City (mimeo)* .

Thorat, S. (2008), 'Labor market discrimination: Concepts, forms and remedies in the Indian situation', *The Indian journal of labor economics* **51**(1), 31–52.

Tilak, J. B. (2007), 'Post-elementary education, poverty and development in india', *International Journal of Educational Development* **27**(4), 435–445.

Trannoy, A., Tubeuf, S., Jusot, F. & Devaux, M. (2010), 'Inequality of opportunities in health in france: a first pass', *Health economics* **19**(8), 921–938.

Von Hippel, P. T. (2005), 'Teacher's corner: How many imputations are needed? a comment on hershberger and fisher (2003)', *Structural Equation Modeling* **12**(2), 334–335.

Von Hippel, P. T. (2009), '8. how to impute interactions, squares, and other transformed variables', *Sociological methodology* **39**(1), 265–291.

Weisskopf, T. E. (2004), 'Impact of reservation on admissions to higher education in india', *Economic and Political Weekly* pp. 4339–4349.

Zimmermann, L. (2012), 'Labor market impacts of a large-scale public works program: evidence from the indian employment guarantee scheme'.