THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Inequality of Opportunity in China

Dongjie Wu

BA, MIntEcon&F

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2018

The School of Economics

Abstract

This thesis makes use of advanced econometric tools to improve empirical frameworks and methods for measuring inequality of opportunity. In addition, we apply these empirical methods to measure inequality of opportunity in China — a developing country with rising economic inequality and rapid economic growth. A conventional approach to measuring inequality of opportunity classifies outcomes such as the distribution of income, wealth and health status into two sets of factors: those beyond individuals' responsibility ("circumstance") and those within individuals' responsibility ("effort"). Based on this framework, we model the correlation between circumstances and effort.

In Essays I and II, we use a model allowing heteroscedasticity between circumstances and a latent class model to identify two different ways that circumstances could be affected by effort. First, different types, which are groups with individuals sharing the same circumstances, could have different effort distributions. Second, circumstances can have different effects on income for different levels of effort. Using these models, we measure inequality of opportunity in China at both provincial and national levels and find a higher inequality of opportunity than the conventional approach identifies.

In addition, we examine educational inequality of opportunity using administrative data from a highly-ranked university. This study is different from most empirical literature in measuring inequality of opportunity in that graduate outcome, the outcome of interest, is a categorical variable. We use the multinomial regression model and stochastic dominance to study how graduates' family backgrounds affect their graduate outcomes. We find that those who are from a low-income or from rural family are disadvantaged in receiving high-quality higher education. Even though they are enrolled in a top university, they have less opportunity for a postgraduate degree.

Declaration by author

(All candidates to reproduce this section in their thesis verbatim)

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications during candidature

Peer-reviewed Papers:

Wu, D., & Rao, P. (2017). Urbanization and Income Inequality in China: An Empirical Investigation at Provincial Level. *Social Indicators Research*, *131*(1), 189-214.

Chen, D., Petrie, D., Tang, K., & Wu, D. (2017). Retirement saving and mental health in China. *Health Promotion International*. doi:10.1093/heapro/dax029.

Conference Abstracts:

Wu, D., Rao, P., Tang, K. K., & Trivedi, P. (2016). Sources of Income Inequality in China: Individual's Effort or Circumstances?. Presented at IARIW 34th General Conference, Dresden, Germany.

Publications included in this thesis

No publications included.

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Research Involving Human or Animal Subjects

No animal or human participants were involved in this research.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisory team: Prof. Prasada Rao, Prof. Pravin Trivedi and Assoc. Prof. KK Tang, for their continuous support of my Ph.D. study and related research. They offered valuable advice and guidance thoughout my Ph.D. journey.

In particular, I would like to thank my principal supervisor, Prof. Prasada Rao, who encouraged me to pursue this degree and who has mentored me for more than 5 years. During my Ph.D. study, he patiently encouraged me and helped me explore ideas about inequality of opportunity in China. It would not have been able to complete this thesis without Prof. Prasada Rao's help.

I am equally indebted to Prof. Pravin Trivedi. He generously offered his time and expertise to help me develop empirical models and polish this thesis. I have learned a lot from him about how to put empirical models into practice to solve real world problems.

Further, I would like to thank Assoc. Prof. KK Tang for his time, guidance and careful attention to detail. He gave me enormous support on improving my academic writing, asking good research questions and clarifying the implications of research findings.

I also gratefully acknowlege other members of the Ph.D. committee, Assoc. Prof. Xavi Ramos, Assoc. Prof. Peter Earl, Assoc. Prof. Jeffrey Kline, Dr Haishan Yuan and Dr Nicholas Rohde, for their time and insightful comments. Their valuable comments were instrumental in improving this thesis.

My sincere gratitude also goes to Prof. Nanak Kakwani and Prof. Lars Osberg for their insightful suggestions and comments. Their adivce led me to view my thesis from a different perspective.

Special thanks go to the Institute of Social Science Survey (ISSS) of Peking University, China and the School of Economics of Shandong University for providing me with high-quality datasets. It would be impossible to write this thesis without these valuable data.

I must also express my deep gratitude to my parents. Their continuous support and encouragement were crucial for me to overcome all challenges in the last four years.

I especially thank my wife Qi Zhang for her love and companionship. She listened to my presentation rehearsal on numerous occasions and patiently, gave me feedback on my thesis and supported me spiritually throughout my Ph.D. study. Last but not least, I would like to thank Nomi and Tom for their companionship. They brought me a lot happiness and relaxation in the most intensive days and nights of my Ph.D journey.

Financial support

This research was supported by an Australian Government Research Training Program Scholarship.

Keywords

inequality of opportunity, income inequality, educational inequality, shapley decomposition, finite mixture model, multinomial regression.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 140299 Applied Economics, 30%
ANZSRC code: 140219 Welfare Economics, 20%
ANZSRC code: 140301 Cross-section analysis, 50%

Fields of Research (FoR) Classification

FoR code: 1402 Applied Economics, 50%
FoR code: 1403 Econometrics, 50%

Contents

List of Figures

List of Tables

List of Abbreviations

IOP	Inequality of Opportunity
IOR	The Ratio of Inequality of Opportunity to Income Inequality
EOR	The Ratio of Inequality of Effort
IOL	The Level of Inequality of Opportunity
EOL	The Level of Inequality of Effort
FMM	Finite Mixture Model
FMMV	Finite Mixture Model with Variant Probabilities
MLE	Maximum-likelihood Estimation
MLD	Mean Log Deviation

Chapter 1

Introduction

Economic inequality has received significant attention from academics, the general public and policy makers. Reports show that the wealthiest 1% own more than 50% of the world's wealth by 2016 (Hardoon, 2017). However, research suggests that what concerns people is not economic inequality itself but the economic unfairness (Starmans et al., 2017) — when one is unfairly under-rewarded or over-rewarded due to "irrelevant" factors such as gender, ethnicity, parents' socioeconomic background, etc. Because of the aversion to economic unfairness, some scholars and politicians argue that policy interventions related to redistribution should be based on economic fairness rather than economic equality.

Researchers have designed measures of inequality of opportunity to show the level of economic unfairness and capture the extent to which economic inequality is due to irrelevant factors (Roemer, 1998). However, these measures could be potentially biased due to limited data and the use of simplified empirical methods. This thesis makes use of advanced econometric tools to improve empirical methods used in measuring inequality of opportunity. In addition, we apply these new empirical methods to measure inequality of opportunity in China — the most populous country in the world, where economic inequality has been rising amid rapid economic growth.

Rising economic inequality in China has been well-documented. Measured by the Gini coefficient, income inequality is reported to have increased from under 0.30 before 1980 to 0.55 in 2012 (Xie and Zhou, 2014). Educational inequality has also risen with income inequality. Children living in urban areas or having rich parents receive more years of schooling (Zhang et al., 2015) and are more likely to enter a college (Wang et al., 2013). In terms of wealth inequality, a report conducted by Beijing University using China Family Panel Studies (CFPS) found that the top one percent of households held one-third of total assets, while the bottom 25 percent held only one percent (ISSS, 2014).

These rising inequalities should be a great concern to the public if they are deemed "unfair", that is, if inequalities are determined by factors beyond individuals' control. For example, the Hukou system, a household administrative system in China that restrains rural population from accessing education and employment in urban areas (Liu, 2005), has been found to contribute significantly to rural-urban inequality. Moreover, regional disparities are reported to influence educational attainment (Hannum and Wang, 2006). To measure unfair inequality, we follow the framework of equality of opportunity developed by Roemer (1998).

In the literature on economic inequality, most measures of inequality refer to inequality of economic outcomes such as income, wealth and consumption. On the contrary, a measure of inequality of opportunity shows *the proportion of inequality of outcome that is due to irrelevant factors* — factors which are out of individuals' control. Roemer (1998) postulates that outcomes such as the distribution of income, wealth and health status are determined by two sets of factors: those beyond individuals' responsibility (denoted as "circumstances") and those within individuals' responsibility (denoted as "effort"). In order to achieve equal opportunity, he decomposes outcomes into circumstances and effort, and argues that the effect of "circumstances" on inequality of outcome should be eliminated. Based on this point of view, economists such as Roemer (1998) and Fleurbaey and Peragine (2013) have established conceptual and theoretical frameworks to measure inequality of opportunity.

Based on Roemer's theoretical concepts and framework, researchers have measured inequality of opportunity for different countries. Most of these studies focus on developed countries such as Norway (Almas et al., 2011), Sweden (Björklund et al., 2012), Italy (Checchi and Peragine, 2010), France (Lefranc et al., 2009) and U.S. (Pistolesi, 2009). Others measure inequality of opportunity in Latin America including Ferreira et al. (2003), Bourguignon et al. (2007), Ferreira and Gignoux (2008) and Paes de Barros et al. (2009). Few studies consider inequality of opportunity in China.

However, both the theoretical foundations and empirical methodologies currently in use have limitations. First of all, one may question whether inequality of outcome can be conceptually separated into circumstances and effort (Kanbur and Wagstaff, 2014). It is possible that one's effort is strongly influenced by circumstances. For example, a child from a single-parent family may exert less effort in school than other children. In addition, it is difficult to identify some factors such as luck, risk and talent as either circumstances or effort. In addition, most empirical studies treat parents' socioeconomic status as one's circumstances. It is equivalent to saying that one's circumstances depend on parents' effort (Kanbur and Wagstaff, 2014). Thus, implementing inequality of opportunity by eliminating parents' influence may weaken the role of the family (Roemer and Trannoy, 2015, pp. 56). In our study, we do not focus on the conceptual limitation of the literature. We take the view of Roemer (1998) that circumstances and effort are correlated with each other and the effects of circumstances on effort should also be eliminated in order to achieve equal opportunity. This view is also adopted in empirical research, e.g. Bourguignon et al. (2007) and Ferreira and Gignoux (2008).

The relationship between circumstances and effort is not just a conceptual issue. It has implication for empirical work. For example, circumstances and effort are assumed to be independent in most empirical literature. However, if they are correlated with each other, neglecting it could lead to a biased measure of inequality of opportunity. Another problem is that individual effort is difficult to observe. Therefore, most studies treat effort as an unobserved variable. This unobserved effort, combined with some unobserved circumstances, could also bias the measures of inequality of opportunity.

Due to these theoretical and empirical limitations, the concept and the measurement of equality of opportunity is still in development. In the thesis, we focus mainly on the empirical issues rather than the theoretical ones.

In Essay I, we examine unfair income inequality in contemporary China at both the national and the regional levels using data from the China Family Panel Study, which contains 33,600 individual observations for the years 2010 and 2012. Our empirical analysis includes zero-income observations using a Hurdle model and a Heckman model. In addition, we study the correlation between circumstances and effort by parameterizing heteroskedasticity between circumstances using maximum likelihood estimation. Shapley decomposition is implemented to identify the contributions of each of the identified "circumstances" and "efforts" to income inequality. We find that more than 20% of income inequality is due to the effect of circumstances on income through effort. This finding proves the significant effect of circumstances on effort. An estimation with no regards to this correlation could lead to an underestimated measure of inequality of opportunity. Within this model framework, we identify gender, geographic factors and parents' socioeconomic status as the three main factors contributing to unfair income inequality. At the regional level, as we move from low-income to high-income regions, fair within-region income inequality decreases significantly while unfair within-region income inequality increases slightly, with a small net effect on total income inequality. This implies that people suffer similar unfair within-region income inequality no matter in which region they live.

In Essay II, we propose an alternative approach which treats unobserved effort as a categorical latent variable. In the conventional empirical approach, effort is commonly treated as unobserved and captured by the residual in a model. Moreover, effort and circumstances are assumed to be independent of each other. In this essay, individual effort is assumed to be distributed around three levels: low, middle, and high. The assignment probability to each class is determined by a probabilistic function. These probabilities are estimated by a finite mixture model. Using this model, we allow the effects of circumstances on income to be different across levels of effort. Including these heterogenous effects, our estimates show a higher unfair income inequality than the estimates obtained by using a conventional approach. In addition, we find a substantial income gap between low-effort and high-effort levels and that inequality of opportunity is the highest if every individual exerts the middle-level effort.

In Essay III, we shift the focus from income to education by investigating the expansion of tertiary education in China. The number of graduates from Chinese higher education institutions has increased substantially in the last two decades. However, it raises the question as to whether the higher education sector provides equal opportunity to everyone amid the expansion. Our study examines whether individuals' circumstances could affect enrolment in tertiary education and graduates' outcomes using the adminstrative data from a highly-ranked university. Our study finds a widening rural-urban inequality in tertiary education. Urban students are overrepresented in this university and they have more choices after graduation than their rural counterparts. Another determinant is family income. A child from a low-income family background is less likely to enter this university and has fewer choices after graduation.

Through these three essays, this thesis contributes to the literature on inequality of opportunity in three aspects. Firstly, it provides an improvement on the econometric methodology for measuring inequality of opportunity. In the first essay, a heteroskedastic model is designed to capture the effect of circumstances on effort. In the second essay, we use a latent class model to capture the unobserved effort and to deal with the heterogenous effects of circumstances on income between levels of effort. These two empirical approaches shed light on how circumstances and effort are correlated and the role effort plays in inequality.

Secondly, this thesis enriches the literature of inequality of opportunity in China. Since most of the literature on inequality of opportunity focuses on developed countries and Latin American countries, this thesis makes use of Chinese datasets from the last decade and estimates the measure of inequality of opportunity in income and educational outcome in contemporary China.

Finally, this thesis extends the study in equal opportunity to access to education. Most studies measuring inequality of opportunity in education treat academic performances as the outcome. In the third essay, we extend the framework to examine youths' abilities to gain access to higher education. Our study fills the knowledge gap on how circumstances affect access to higher education in China.

The thesis comprises a comprehensive literature review and three essays. It proceeds as follows. Chapter 2 focuses on the current methodology of measuring inequality of opportunity and previous empirical work on measuring inequality of opportunity in China. Chapters 3 to 5 present three essays respectively. Chapter 6 offers a summary and conclusion draw from the thesis.

Chapter 2

Review of Literature

2.1 Introduction

Emerging only in the last two decades, equality of opportunity is a new and attractive approach to improve social justice. It has a distinctive meaning in terms of equality of outcome. In the traditional literature that examines economic inequality, most "inequalities" are referred to as inequality of outcome — inequality such as income inequality, wealth inequality and consumption inequality — while inequality of opportunity examines inequality of outcome in such a way that the allocation of justice and fairness should be considered. It argues that there could be inequality which is fair and equality which is unfair. The redistribution of inequality of outcome should eliminate the unfair part and retain the fair one.

This literature review introduces the origin and development of the concept of equality of opportunity, examines the establishment of theoretical frameworks and measurements, presents the empirical works on inequality of opportunity in different countries and discusses the limitation and scope of this area.

This literature review proceeds as follows. Section 2 introduces the philosophical background of equal opportunity. Section 3 examines theoretical models of equality of opportunity. Section 4 reports the current methodology used in measuring inequality of opportunity. Section 5 discusses the relationship between equality of opportunity and other economic concepts and ideas such as development, income inequality and intergenerational mobility. Section 6 provides empirical findings and measures inequality of opportunity in China. Finally, the last section discusses the limitation and scope of equality of opportunity.

2.2 Origins and Concepts of Equal Opportunity

The study on equal opportunity originates from the discussion of social choice, fair allocation and distributive justice. Harsanyi (1953, 1955) and Rawls (1971) stated that a “just” social choice should be made by a rational agent with a fair mind. Rawls (1971) assumed an “original position” so that “the veil of ignorance” is such an ”original position“ from which the person makes a social choice under uncertainty without knowing his own social status or circumstances in society (Rawls, 2009).

Under the assumption of “the veil of ignorance”, Rawls stated two principles of justice (Rawls, 2009, p. 266):

> First principle: Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.
>
> Second principle: Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.

The first principle assumes that every individual has a lexicographic preference on “primary goods”, goods that “every rational man is presumed to want” (Rawls, 1971, p. 92), such as rights, liberties, income and wealth, intelligence, health, etc. The lexicographic preference is that given two sets of “primary goods” that include intelligence, health, liberties, wealth, and etc., one only compare the next important good if the amount of the most important is equal. John Rawls claimed that “liberty” is the most important among these primary goods so that each person is to have an equal right for it.

The second principle prescribes social and economic inequalities by applying “the difference principle”(which is part (a) in the second principle) and “fair equality of opportunity”(which is part (b) in the second principle).

The difference principle obeys a maximin criterion which can be expressed as the following equation if a person’s preferences can be represented by a utility function.

$$\min[u^1(x), \dots, u^N(x)] > \min[u^1(y), \dots, u^N(y)] \tag{2.1}$$

where $u^i(x), i = \{1, \dots, N\}$ is the utility function for individual i to the social state x; and where $u^i(y), i = \{1, \dots, N\}$ is the utility function for individual i to the social state y. A rational individual will be risk averse so that he/she will prefer “the greatest benefit of the least advantaged”. This maximin criterion has been used in the theoretical framework of inequality of opportunity (Roemer, 1998).

In terms of the “fair equality of opportunity”, according to Rawls (2009, p. 57), there are two natural senses in which positions and offices are open equally to all: equality as careers open to talents and equality as fair equality of opportunity. The difference between these two senses is that the former fails to eliminate social contingencies, e.g.

family background, while the latter succeeds. Nonetheless, even for the latter sense, individuals' abilities may be affected by either all kinds of social conditions or their natural talents. Rawls asserts that "there is no more reason to permit the distribution of income and wealth to be settled by the distribution of natural assets than by historical and social fortune."

Rawls addresses the question of "moral desert" (Rawls, 2009, pp. 274); that is, whether the distribution of income and wealth according to some characteristics or conditions is morally deserved. He lists at least three different kinds of factors which are not deserved by individuals: natural talent, all kinds of social conditions and class attitudes which develop natural capacities; and other social conditions which contribute less to individual abilities. This question of "moral desert" affects the later debate on the responsibility cut between circumstances and effort in terms of measuring inequality of opportunity.

Although Rawls provides a constructive theory of distributive justice, his theory has been criticized by many academics. Harsanyi (1976) disagreed with the application of the *maximin principle* and the *difference principle* because of their violation of the assumption of rationality. He provided a counterexample of a person deciding between working in a low-paid job in New York and working in a well-paid job in Chicago. He assumed that the person was currently living in New York, and if he decided to work in Chicago, he had to take a plane with a very small possibility of being killed in a plane crash. According to the *maximin principle*, he would prefer to stay in New York to avoid the worst consequence. Then he further argues that people may face a lot of these worst possible outcomes. He goes on to say:

> If you took the maximin principle seriously then you could not ever cross a street (after all, you might be hit by a car); you could never drive over a bridge (after all, it might collapse); you could never get married (after all, it might end in a disaster), etc. If anybody really acted this way he would soon end up in a mental institution (Harsanyi, 1976, p. 595).

Thus, Harsanyi (1976) prefers the expected-utility maximization principle rather than the maximin principle or the difference principle proposed by Rawls.

Dworkin (1981a) provided an alternate approach which applies the expected-utility maximization. He claimed that the talents and abilities individuals have are pure luck ("brute luck" defined by Dworkins), which are out of individuals' control; while there are the other kinds of luck ("option luck") such as the lottery and gambling which are within individuals' control. Thus, the distributive justice problem can be solved by transferring "brute luck" to "option luck" by means of a hypothetical insurance system to transfer the "brute luck" to the "option luck".

In addition to "moral desert", another issue concerning distributive justice is whether individuals' taste or preference should be respected. A debate related to this issue is between *equality of resources* and *equality of welfare*. The former claims that individuals should share the same amount of resources which are available to be distributed (Dworkin, 1981a). The latter states that regardless of how many resources are provided, individuals should enjoy the same welfare. These two claims contradict each other because equality of resources might fail to balance equality of welfare and vice versa.

In order to deal with the incompatibility of equality of resources and equality of welfare, Arneson (1989) proposed an *equal opportunity for welfare*. He defines an opportunity as "a chance of getting a good if one seeks it" (Arneson, 1989, pp. 85). Then he claims that each individual should face "equivalent arrays of options" and these arrays of options should be *effective* such that one of the following is true (Arneson, 1989, pp. 86):

> (1) the options are equivalent and the persons are on a par in their ability to "negotiate" these options
>
> (2) the options are nonequivalent in such a way as to counterbalance exactly any inequalities in people's negotiating abilities
>
> (3) the options are equivalent and any inequalities in people's negotiating abilities are due to causes for which it is proper to hold the individuals themselves personally responsible.

The improvement of equal opportunity for welfare is that it considers individual's choices and options. Only welfare loss due to factors beyond a person's control will be compensated.

Although equal opportunity of welfare seems to solve the compensation difficulty in terms of taste and preference, the problem with inequality of welfare and equal opportunity of welfare is that people who have the same loss of resources (disabilities, for example) may suffer different welfare loss. In contrast, for resource-egalitarians, the same degree of disabilities yields the same level of resource loss. Thus, people who have the same degree of disabilities should be compensated at the same level. To deal with this welfare deficiency, Cohen (1989) proposed a broader concept of equal opportunity for welfare —*equal access to advantage*. He argued that both welfare-egalitarians and resource-egalitarians assume a number of dimensions of disadvantage and judge whether inequality due to the disadvantage is acceptable or not. The difference between them is that welfare-egalitarians focus on preference and welfare, while resource-egalitarians pay attention to the different kinds of resources individuals have. However, "the right cut is between responsibility and bad luck"(Cohen, 1989, pp. 922). He believes that disadvantage should be compensated only when it is out of a person's control, regardless of his "unfortunate resource endowment and unfortunate utility function". This responsibility-sensitive thinking on the problem of justice led later to the emergence of an economic literature dealing with the concept of inequality of opportunity.

In addition, Sen (1995) proposed *equality of capability* as an alternative way to deal with distributive justice and this concept has a relationship with inequality of opportunity. He defined a person's position in a social arrangement under two different perspectives: "the actual achievement" and "the freedom to achieve"(Sen, 1995, pp.31). Then, he proposed *functionings* consisting of a person's achievement, and *capability* as "the person's freedom to choose from possible livings"(Sen, 1995, pp. 40). Thus, the capability to achieve functionings is the real opportunity which represents people's freedom to achieve well-being. He argued that capability can better reflect the level of a person's freedom and opportunity than the "primary good" proposed by John Rawls and the "equality of resources" suggested by Ronald Dworkin because "primary good" and "equality of resources" are *means to freedom*, not freedom itself; though improving means to freedom can help a society to achieve better freedom. However, Sen's *equality of capability* also is difficult to apply in empirical research (one attempt is the Human Opportunity Index by Paes de Barros et al. (2009)). It is hard to quantify with the intention of measuring the real inequality of capability; furthermore, some capabilities are congenital and difficult to be equalized.

In summary, this section discusses the theoretical background of inequality of opportunity. The concept of inequality of opportunity originates from the idea that equalizing outcomes is not enough to achieve distributive justice. Political philosophers such as John Rawls, Ronald Dworkin, Amartya Sen, etc. propose a variety of theories and methodologies to improve distributive justice. These theories and debates have led to the emergence of researches on inequality of opportunity in economics.

2.3 Theoretical Models of Equality of Opportunity

The theoretical models of equal opportunity are based on two basic principles: *the compensation principle* and *the reward principle*. The compensation principle focuses on compensating inequalities due to circumstances given the same degree of effort, while the reward principle examines how individuals are rewarded under the same circumstances.

In terms of the compensation principle, individuals can be compensated before or after their outcome have been made. An ex-ante compensation principle defines as a compensation principle implemented before knowing one's actual outcome. The compensation is only based on circumstances that should not be the responsibility of individuals. An ex-post compensation principle reallocates the resources after knowing individual outcomes. In other words, the compensation is based on both circumstances and effort. Most literature use either the ex-ante or ex-post approach. However, the ex-ante approach is easier to implement because it only requires the information on circumstances.

In contrast, the reward principle values the inequality among those sharing the same circumstances. For example, liberal reward principle states that income due to pure effort should not be redistributed, while utilitarian reward focus on those who sharing the same circumstances should have the same income. However, the reward principle is incompatible with the ex-post compensation principles (Fleurbaey and Peragine, 2013).

In this section, we introduce several models developed based on these two principles.

2.3.1 A Model for Equal Opportunity Policy

Roemer (1998) proposed a model for an equal opportunity policy. He categorized the population into a finite set of types $\mathcal{T} = 1, 2, \ldots, T$ based on an individual's circumstances. Within a type t, he defined an achieved level of individual outcomes (which could be income and educational achievement) denoted by $u^t(e, \phi)$ where e is a measure of individuals' effort and ϕ is a unitary social policy in the set of social policies Φ. He assumed that the function u^t is strictly monotone increasing in e.

In order to identify the effect of circumstances on the function u^t, Roemer (1998) examined the relationship between circumstances and effort and decomposes the raw effort into the average effort for type t and the degree of effort $\pi = G^t_\phi(e)$ where G is the distribution function of effort in type t given the policy ϕ. He argued that the influence of circumstance is mainly due to the average effort for type t. An individual should only be accountable for his effort compared with others having the same circumstances. Formally, he defined:

$$v^t(\pi, \phi) = u^t(e^t(\pi), \phi) \tag{2.2}$$

where v^t is a function measuring the achieved level for a type t individual given his degree of effort π and the policy ϕ.

Given the limitation of resources and the feasible set of policies Φ, an optimal policy in line with equal opportunity can adopt the principle of justice proposed by Rawls (1971) in which he claimed that "social and economic inequalities are to be arranged to the greatest benefit of the least advantaged".

Therefore, Roemer (1998) suggested an approach to maximize the least advantaged:

$$\max_{\phi \in \Phi} \int_0^1 \min_t v^t(\pi, \phi) d\pi \tag{2.3}$$

A similar approach from Van De Gaer (1993) is:

$$\max_{\phi \in \Phi} \min_t \int_0^1 v^t(\pi, \phi) d\pi \tag{2.4}$$

The difference between these two approaches is that Roemer's (1998) approach chose the minimum value of v_t for each degree of effort π and Van De Gaer's (1993) approach calculated the average degree of opportunity (the aggregation of v_t for every possible effort in that type t) for each type t first and then chose to maximize the type with the least average degree of opportunity.

Table 2.1 shows an example with 2 types and 4 degrees of effort. Following Roemer's approach, the optimal policy should maximize the outcome of type 1 with 1 and 3 degree of effort and the outcome of type 2 with 2 and 4 degree of effort. Based on Van De Gaer's (1993) approach, the optimal policy should consider maximizing type 2's outcome instead of type 1's since type 2 has the lower average outcome. The two approaches are indifferent if one type is unambiguously disadvantaged compared with the other types (that is, for each π, there exists a $\hat{t}$ whose achievement function $v^{\hat{t}}(\pi) \leq v^{t}(\pi)$ for all $t \in \mathcal{T}$).

Table 2.1: An example with 2 types and 4 degrees of effort

	Type 1	Type 2
1	3	5
2	5	3
3	4	6
4	8	2

As the above example shows, one requires the information of degree of effort if applying Roemer's (1998) approach while for Van De Gaer's (1993) approach, one needs to know the average outcome for each type. However, both approaches do not consider the population for each degree of effort. For example, both approaches might be biased if one degree of effort in type t represents only one individual and the other comprises thousands.

2.3.2 A Fair Compensation Model

Bossert (1995) and Fleurbaey and Bossert (1996) postulated several principles on distinguishing the responsibility of an individual's income. They decomposed the sources of income into "relevant characteristics" and "irrelevant characteristics". The former are the individual characteristics whose influence to income should be considered "relevant" and the latter are the characteristics whose effect to income should be considered as "irrelevant". This decomposition is more general and broader than the concept of equal opportunity in which income (or other forms of outcome) should not be affected by circumstances (factors out of an individual's control) but by effort (factors within an individual's control). Effort could be one of the "relevant characteristics" but not vice versa. Besides, effort could be ordinal (high effort vs. low effort) while "relevant characteristics" could be nominal.

Based on Bossert's (1995) framework, one can derive two principles: Group Solidarity (GS) and Individual Monotonicity (IM). The former is deduced from the perspective of irrelevant characteristics — if one changes his irrelevant characteristics, everyone should change a similar amount of their income or keep the same income as before; while the latter is derived from the perspective of relevant characteristics — if one changes his relevant characteristics, no redistribution is needed.

Based on these two basic principles, researchers designed several redistribution mechanisms to identify situations where the allocation is fair and equal opportunity is satisfied. These redistribution mechanisms are designed to satisfy group solidarity or individual monotonicity. Besides, they can also be designed to satisfy Pareto efficiency or the no-envy principle. Using redistribution mechanisms, one can calculate the counterfactual income distribution when the allocation is fair (equal opportunity is satisfied) and then measure the distance between the real income and the counterfactual income.

However, there are some drawbacks when applying this approach. First, one needs a reference or benchmark vector for personal characteristics. These benchmarks are computed from specific redistribution mechanisms. In consequence, different benchmarks yield different results. Second, it takes several steps to calculate the measurements, which might increase the error. Besides, it is still not clear how to treat the error — as irrelevant or relevant characteristics. The last shortcoming is that this approach assumes the independence between the irrelevant and relevant characteristics. It neglects the situation when the relevant characteristics are affected by the irrelevant characteristics.

In this section, we briefly reviewed the allocation mechanisms and their implementation in empirical studies.

2.3.2.1 The Notation

The population in a given society is $N = 1, \ldots, n$, where $n \geq 2$. There are $r \in \mathbb{N}$ individual characteristics that are considered "relevant" and $s \in \mathbb{N}$ "irrelevant". Person i's irrelevant characteristics are described by $a_i^S \in \mathbb{R}^s$. The characteristics vector of $i \in N$ is $a_i = (a_i^R, a_i^S) \in \mathbb{R}^{r+s}$. A characteristics profile is given by $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^{n(r+s)}$, and it can be partitioned into $\mathbf{a}^R = (a_1^R, \ldots, a_n^R) \in \mathbb{R}^{nr}$ and $\mathbf{a}^S = (a_1^S, \ldots, a_n^S) \in \mathbb{R}^{ns}$.

The set of possible characteristics vectors is $\Omega = \Omega_R \times \Omega_S$, where $\Omega_R \subseteq \mathbb{R}^r$, $\Omega_S \subseteq \mathbb{R}^s$, and $\Omega_R, \Omega_S \neq \emptyset$.

Furthermore, each agent has his/her non-negative individual income $y_i \in \mathbb{R}_+$. Let the vector $\mathbf{y} = (y_1, \ldots, y_n) \in {\mathbb{R}_+}^n$. The set of possible income vectors is $\mathfrak{y} \subseteq {\mathbb{R}_+}^n$.

An economy e is a pair $(\mathbf{y}, \mathbf{a}) \in \mathfrak{y} \times \Omega$. Let $\mathcal{E}$ be the domain of economies. A complete and transitive binary relation $\succsim_O$ is defined to compare inequality of opportunity between two economies. $\sim_O$ represents two economies having the same inequality of opportunity.

If $\succsim_O$ is continuous on $\mathfrak{y} \times \Omega$, we can find a function $I : \mathfrak{y} \times \Omega \mapsto \mathbb{R}$ such that $I(e)$

represents the measurement of inequality of opportunity in economy e.

2.3.2.2 The Redistribution Mechanisms

The concept of equal opportunity states that inequality due to relevant characteristics is acceptable and inequality due to irrelevant characteristics is offensive. Therefore, (Fleurbaey and Bossert, 1996) purposed two basic principles for redistribution: *Group solidarity* (GS) based on irrelevant characteristics and *Individual monotonicity* (IM) based on relevant characteristics.

GS says that if one agent changes his irrelevant characteristics, the measurement of inequality of opportunity should not change when every agent's income increases or decreases by the same amount.

Axiom 2.3.1 (Group solidarity (GS)) $\forall \mathbf{a}, \hat{\mathbf{a}} \in \Omega, \forall k \in N$, $a_k^S = \hat{a}_k^S$ *and* $a_j = \hat{a}_j, \forall j \in N \setminus \{k\}$. $e \sim \hat{e}$ *if and only if* $\hat{y_k} - y_k = \hat{y_j} - y_j, \forall k, j$.

IM says that if one agent changes his relevant characteristics, the measurement of inequality of opportunity should not change when only this particular agent's income changes.

Axiom 2.3.2 (Individual Monotonicity (IM)) $\forall \mathbf{a}, \hat{\mathbf{a}} \in \Omega, \forall k \in N$, $a_k^S = \hat{a}_k^S$ *and* $a_j = \hat{a}^j, \forall j \in N \setminus \{k\}$. $e \sim \hat{e}$ *if and only if* $\hat{y_j} = y_j, \forall j$

Bossert (1995) proved that IM and GS are compatible only if the income is an additively separable function of relevant and irrelevant characteristics. He assumed that an income function is a mapping $f : \Omega \mapsto \mathbb{R}_{++}, a = (a^R, a^S) \mapsto f(a)$. In other words, an agent's income is determined by his or her *own* characteristics only. He also suggested a possible generalization that allows income to depend on the entire characteristics profile.

The redistribution mechanism might also aim to satisfy other allocation principles such as the Pareto efficiency and no-envy allocation rule. Foley (1967) defined no-envy allocation as follows.

Definition 2.3.1 (No Envy Allocation) *Given* $e \equiv (R, \Omega) \in \mathcal{E}$, *the allocation* $\mathbf{z} \in \mathbf{Z(E)}$ *is envy-free for e, written as* $z \in F(e)$, *if for each pair* $i, j \subset N, z_i R_i z_j$.

The no-envy allocation states that an allocation is said to be fair if nobody prefers anybody else's bundle over his own. The advantage of no-envy allocation is that "it treats economic agents symmetrically, is ordinal in nature, and is free of interpersonal comparisons of utility" (Pazner and Schmeidler, 1978). However, Pazner and Schmeidler (1978) showed that among all Pareto-efficient allocations, none can be found that is fair under the standard Arrow-Debreu production economies.

In order to have an allocation which never conflicts with Pareto efficiency, Pazner and Schmeidler (1978) defined an egalitarian-equivalent allocation.

Definition 2.3.2 (Egalitarian Equivalence) *Given* $e \equiv (R, \Omega) \in \mathcal{E}$*, the allocation* $\mathbf{z} \in \mathbf{Z(e)}$ *is egalitarian-equivalent for e, written as* $z \in E(e)$*, if there is* $z_0 \in Z(e)$ *such that* $zI(z_0, \dots, z_0)$

Pazner and Schmeidler (1978) explained the egalitarian-equivalent allocation as follows:

> "An allocation is said to be egalitarian-equivalent if there exists a fixed commodity bundle (the same for each agent) that is considered by each agent to be indifferent to the bundle that he actually gets in the allocation under consideration. It is shown that Pareto-efficient and egalitarian-equivalent allocations always exist under (even weaker than) the standard conditions on the economic environment."

Based on the fair allocation rule, Bossert (1995) defined a redistribution mechanism to eliminate the effects of "irrelevant" characteristics given that income is a function $f : \Omega^n \mapsto \mathbb{R}^N_{++}$ of individual's own characteristics.

Definition 2.3.3 *A redistribution mechanism is a mapping* $F : \Omega^n \mapsto \mathbb{R}^N_{++}, \mathbf{a} \mapsto F(\mathbf{a})$ *such that,*

$$\sum_{i=1}^{n} F_i(\mathbf{a}) = \sum_{i=i}^{n} f(a_i), \forall \mathbf{a} \in \Omega^n \tag{2.5}$$

If income functions f are additively separable in R and S, the following mechanism F^0 seems very plausible.

$$F^0_k(\mathbf{a}) \equiv g(a^R_k) + \frac{1}{n}\sum_{i=1}^{n} h(a^S_i), \forall \mathbf{a} \in \Omega^n, \forall k \in N. \tag{2.6}$$

Fleurbaey and Bossert (1996) found two alternative redistribution mechanisms to avoid the impossibility theorem and release the assumption of additive separability.

The egalitarian-equivalent mechanism F^{EE} is defined by

$$F^{EE}_k(\mathbf{a}) \equiv f(a^R_k, \tilde{a}^S) - \frac{1}{n}\sum_{i=1}^{n}[f(a^R_i, \tilde{a}^S) - f(a_i)], \forall \mathbf{a} \in \Omega^n, \forall k \in N \tag{2.7}$$

The conditionally egalitarian mechanism F^{CE} is defined by

$$F^{CE}_k(\mathbf{a}) \equiv f(a_k) - f(\tilde{a}^R, a^S_k) + \frac{1}{n}\sum_{i=1}^{n} f(\tilde{a}^R, a^S_i), \forall \mathbf{a} \in \Omega^n, \forall k \in N \tag{2.8}$$

Moreover, Fleurbaey and Bossert (1996) proposed two additional redistribution mechanisms — the *average egalitarian-equivalence* and *average conditionally egalitarian.* However, these two mechanisms require more information than F^{EE} and F^{CE}. In equation

(2.7) and (2.8), either the information of observed relevant characteristics or the observed irrelevant characteristics are required, but not both.

2.3.2.3 Econometrics Issue for the Equal Opportunity Approach

The income function f can be estimated by linear regression.

$$log(f(a)) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\boldsymbol{S}} a^S + \hat{\boldsymbol{\beta}}_{\boldsymbol{R}} a^R + \hat{\epsilon} \tag{2.9}$$

Devooght (2008) used F^{EE} to calculate the norm income — i.e. the income which achieves equal opportunity after redistribution and uses the distance measures proposed by Cowell (1985) to measure the distance between the observed income and the norm income.

There are several issues related to this model:

Error term

The EE and CE mechanisms are based on the assumption of the anonymous redistribution mechanism (Fleurbaey and Bossert, 1996), which implies that persons with identical characteristics should have the same pretax income. However, during the estimation, the error term will cause the violation of this assumption. Thus, this error term should be a new variable either assigned to a^R and a^S. Both Almas et al. (2011) and Devooght (2008) treated the error term as the irrelevant characteristics a^C. In Devooght (2008), the inequality due to irrelevant characteristics is 90-97.5% and it is 75% in Almas et al. (2011). The error term might contribute to the high inequality due to irrelevant characteristics.

Responsibility Cut

Devooght (2008) and Almas et al. (2011) considered the responsibility cut as a decision to be made by society. They provided empirical results with many possible cuts. For example, Devooght (2008) considered hours worked as the only responsibility variable in one reference group and then constructed several other groups by gradually moving variables from the irrelevant variables a^S to the responsibility variable a^R.

The benchmark or reference vector

To apply the Egalitarian Equivalence mechanism, one needs to calculate the reference or hypothetical irrelevant characteristics $\tilde{a^S}$. Thus, one needs to choose a fixed benchmark level of the compensation variables. Devooght (2008) suggested using the profile of the most disadvantaged. He explains:

> . . . anyone who works one hour extra ($\in a^R$) is entitled to keep the fruits of his additional effort as far as his extra income is independent of his compensation variables is taken to be the income one could earn if one were the most disadvantaged in terms of earning power or the least marginally productive

> member of the economy. If you earn more than the least advantaged for the same level of the responsibility variables, this is due to personal characteristics of which you are the lucky possessor, and for that very reason is open for redistribution (Devooght, 2008, pp. 289).

The Correlation between Relevant and Irrelevant Characteristics

This model requires relevant and irrelevant characteristics to be independent with each other.

2.3.3 An Opportunity Set Approach

Based on the definition of fairness proposed by Kolm (1973) and Thomson (1994), "equal opportunity" means that individuals in the society should face an identical opportunity set. Following this definition, Kranich (1996) developed an opportunity set approach. He assumed that each individual i faces a finite set of opportunities O^i and $\mathbf{O} = (O_1, \ldots, O_i, \ldots, O_I)$ is the set of all individuals' opportunity sets. He defined a *cardinality difference relation* $\succsim$ where $\mathbf{O} \succsim \mathbf{O}'$ means the difference of the cardinalities of opportunity sets between individuals in $\mathbf{O}$ is less than that in $\mathbf{O}'$.

Basically, his approach identifies how many options or opportunities each individual has. An "equal opportunity" society means that each individual is provided the same amount of opportunities.

However, this theoretical approach is difficult to implement empirically because researchers can only observe the choices individuals made rather than the options or opportunities available to choose. In Chapter 5, we use a multinomial regression model to study university graduate outcomes. This model assumes that every graduate has three options in their opportunity sets. Instead of measuring the cardinality of every graduate's opportunity set, we measure the probability of realising each option for each graduate. This approach is empirically more plausible than the opportunity set proposed by Kranich (1996).

2.3.4 The Stochastic Dominance Approach

Lefranc et al. (2008) and Lefranc et al. (2009) applied the stochastic dominance criteria to rank the opportunity sets offered by difference circumstances. In their approach, circumstances are treated as lotteries in such a way that inequality of opportunity is determined by either unequal returns to the lotteries (circumstances) or unequal risk of the lotteries. Assuming that individuals with circumstances c earn income y, the definitions of stochastic dominance are as follows:

Definition 2.3.4 *The circumstances c first-order stochastic dominance (FSD) the circumstance c', i.e. $c \succsim_{FSD} c'$ iff:*

$$F(y|c) \leq F(y|c'), \forall y \in \mathbb{R}_+ \tag{2.10}$$

Definition 2.3.5 *The circumstances c second-order stochastic dominance (SSD) the circumstance c', i.e. $c \succsim_{SSD} c'$ iff:*

$$\int_0^y F(x|c)dx \leq \int_0^y F(x|c')dx, \forall y \in \mathbb{R}_+ \tag{2.11}$$

The SSD is equivalent to generalized Lorenz dominance (GLD). Formally:

$$\forall y \in \mathbb{R}_+ c \succ_{SSD} c' \Leftrightarrow \forall \pi \in [0,1] GL_{F(\dot{|}c)}(\pi) \geq GL_{F(\dot{|}c')}(\pi) \tag{2.12}$$

where $GL_{F(\dot{|}c)}(\pi)$ is the value of the generalized Lorenz curve at π for the distribution of $F(\dot{|}c)$.

Based on the definition of stochastic dominance, equality of opportunity is defined as:

Definition 2.3.6 *Equality of opportunity is achieved if and only if $\nexists c, c' \in C$ such that $c \succsim_{SSD} c'$.*

In the definition of equality of opportunity, Lefranc et al. (2008) used the SSD criterion instead of FSD. This is because, in the empirical analysis, effort is usually difficult to observe. Given only the information of circumstances, the outcome is uncertain. The advantage of using the SSD criterion is that it takes not only the returns of the lottery but the risk or uncertainty into consideration.

Since the SSD criterion is equivalent to the generalized Lorenz curve, inequality of opportunity can be directly identified by comparing generalized Lorenz curves of different circumstances. However, if the Generalized Lorenz curves intersect with each other, it is uncertain which circumstances provide more opportunities. Another disadvantage of the stochastic dominance is that it does not provide "a quantification of how far those groups are from one another"(Ferreira et al., 2011). It cannot answer to which extent there is inequality of opportunity is in an economy.

Another limitation is that it cannot measure the relative importance of any factor. For example, suppose there are two factors —gender and parents' income—, should we find {male,rich} is the most advantageous type and {female,poor} is the least advantageous type through stochastic dominance, we still would not know which factors, gender or parents' income, have the greater effects on their level of advantage. As the number of categories expands, it is difficult to analyse data (Lefranc et al., 2008).

2.3.5 A Summary of the Theoretical Model

In this subsection, we compare the model of equality of opportunity of Roemer (1998) and Van De Gaer (1993) with the fair compensation model of Fleurbaey and Bossert (1996) and the stochastic dominance approach (Lefranc et al., 2008).

Table 2.2 lists the differences between these four models on five different aspects. Roemer (1998),Van De Gaer (1993) and Lefranc et al. (2008) use finite and discrete circumstance. They categorized different circumstances into finite types. When applying Roemer (1998) and Van De Gaer (1993) in an empirical study, one can use both non-parametric and parametric approaches. In contrast, Fleurbaey and Bossert (1996) allowed the circumstance to be a continuous variable so it might be better to use parametric approaches when applying Fleurbaey's and Bossert's (1996) model. Lefranc et al. (2008) applied non-parametric approaches with a stochastic dominance criterion.

The second difference between the model of equal opportunity policy and the fair compensation model is that the model of equal opportunity policy uses the degree of effort, the relative measure of effort, instead of the level of effort which is the absolute measure of effort. Roemer (1998) argued that the level of effort is correlated with circumstance since the difference in average of effort between types is out of the individual's control and only the degree of effort comparing with others in the same type should be considered as effort. Based on this argument, Fleurbaey's and Bossert's (1996) model might retain more bias from the correlation between effort and circumstance when measuring inequality of opportunity. Lefranc et al. (2008) and Van De Gaer (1993) did not require the effort being observed.

In the fourth column in Table 2.2, we list whether the model adopts an ex-ante or ex-post approach. In this context, ex-ante means inequality of opportunity is measured before knowing individuals' effort and ex-post means inequality of opportunity is measured after knowing individuals' effort. The former does not require the observation of effort while the latter does. Among four theoretical models, Lefranc et al. (2008) and Van De Gaer (1993) are ex-ante, which means one can measure inequality of opportunity without observation of effort.

In the fifth column, Table 2.2 shows that only Fleurbaey and Bossert (1996) require a reference vector. It is possible that choosing different reference vectors might influence the measurement of inequality of opportunities.

Furthermore, the two models of equal opportunity policy consider the redistribution with limited resources so the most disadvantaged group will be considered first. However, the fair compensation model redistributes under a pure allocation framework. The model first computes the ideal fair allocation with equal opportunity and then measures the distance between the reality and the counterfactual equality. Thus, the two models of equal opportunity policy are the second best approach and the fair compensation model is the first best approach.

Table 2.2: The Comparison of Theoretical Model

Model	Circumstance	Effort	The Approach	Reference Vecter	1st or 2nd Best
Roemer (1998)	finite/discrete	degree of effort	Ex-post	No	2nd
Van De Gaer (1993)	finite/discrete	degree of effort	Ex-ante	No	2nd
Fleurbaey and Bossert (1996)	could be continuous	level of effort	Ex-post	Yes	1st
Lefranc et al. (2008)	finite/discrete	degree of effort	Ex-ante	No	2nd

2.4 The Measurement of Inequality of Opportunity

2.4.1 The Outcome of Interest in Measuring Inequality of Opportunity

Equal Opportunity can be advocated in terms of many different inequalities of outcomes. A typical inequality of outcome is income inequality (Paes de Barros et al., 2009). Others are educational inequality (Ferreira and Gignoux, 2014), health inequality (Jusot et al., 2013). Most outcomes of interest are continuous variables. Some outcomes could be binary variables (Foguel and Veloso, 2014) or categorical variables (Dias, 2009). For a continuous outcome, a linear or log-linear specification is widely used. For a binary variable or a categorical variable, nonlinear models such as a logit model, an ordered probit or a multinomial model are more appropriate. In addition, one can use a stochastic dominance test to evaluate inequality of opportunity (Lefranc et al., 2008) between circumstances.

Given different outcomes of interest and different model specifications, the measures of inequality of opportunity might not be comparable. Measures can only be compared using the same outcomes of interest and the same model specifications.

2.4.2 Selection of Circumstances and Effort

Roemer's framework divides inequality of outcome into circumstances and effort. This raises questions about how to distinguish circumstances and effort. To clarify this question, first we need to define inequality of outcome; that is, what kind of inequality are we dealing with (just as "inequality of what" proposed by Sen (1995)). Second, factors affecting inequality of outcome should be identified, e.g. gender, IQ, parents' income, etc. The last step is that these factors should be distinguished with respect to the responsible factor (effort) and the non-responsible factor (circumstances).

In terms of "inequality of what", although different philosophers propose different kinds of inequality (e.g. inequality of welfare, inequality of resources, inequality of capability, etc.), most empirical literature uses earning inequality, wealth inequality and educational inequality as inequality of outcome. The outcome of interest has been discussed in the previous section. In this section, we focus on the last two steps.

2.4.2.1 Factors Contributing to Inequality

In considering economic inequality, including income inequality and wealth inequality, we may distinguish between four different types of factors that account for inequality of outcome: inheritable factors, personal choices and preferences, social background factors, and luck.

Inheritable factors

Inheritable factors are the genetic or non-genetic influences from parents. The genetic influence is the talent or innate ability inherited from parents. The non-genetic influence includes the family's cultural background and the intergenerational transfer of physical and human capital from parents to children. The former refers to *"meme"*, an analogy between cultural and genetic transmission proposed by Dawkins (2006). As mentioned by Becker and Tomes,

> Some children have an advantage because they are born into families with greater ability, greater emphasis on childhood learning, and other favourable cultural and genetic attributes. Both biology and culture are transmitted from parents to children, one encoded in DNA and the other in a family's culture (Becker and Tomes, 1986, pp. S4)

Thus, the cultural transmission from parents to children can be seen as another endowment determined when children are born.

Becker and Tomes (1986) emphasised intergenerational transfer through parents' human capital investment, and monetary and asset transfers. The rich may invest more in children's human capital, may save more money in old age (Carroll, 1998) and may leave larger bequests (Nardi, 2002). Although most literature concentrates on the role of parents' occupation and income, the role of intergeneration transfers such as human capital investment, expenditure on children, and bequest is neglected. Furthermore, the willingness to invest in children may vary. It is even more difficult to measure parents' willingness to invest in their own children.

Personal choices and preferences

Personal choices and preferences also can cause inequality of outcome among individuals. People face choices and uncertainties every day. Their differences in preferences and decisions may largely influence their outcomes. One important issue is *portfolio choice.* Rich households hold completely different portfolios compared with poor households. For example, rich households own more risky assets than poor households (Bertaut and Starr-McCluer, 2000). Studies of inequality of opportunity seldom include personal choices and preferences because it is hard to observe and measure personal choices and preferences.

Social background

The third type comprises the social background factors including the bias and discrimination in society, geographical disparity and government policy. Inequalities caused by social discrimination such as gender inequality, race inequality and regional disparities have been widely studied by researchers. More literature on inequality of opportunity will also consider these social background factors. Government policy also will affect inequality of opportunity. Although some redistribution policies do reduce inequality, these may not reduce inequality of opportunity.

Luck

The last factor is luck. The idea of luck referred to as "luck egalitarianism" was first proposed by Dworkin (1981b). There are four different kinds of luck defined by Dworkin (1981b): social background luck, genetic luck, brute luck and option luck. (Lefranc et al., 2009)

Social background luck is individuals' difference in social background beyond their control. Individuals cannot decide either the country or region they were born or their parents' income or occupation; however, these differences determine the circumstances of the rich or poor. A person born in a poor region will obviously have fewer opportunities than in a rich region.

Genetic luck is the difference in inheritance of innate characteristics and abilities from parents. Some people born with talents inherited from their parents. This genetic endowment may cause some people richer than others.

Brute luck is the luck that provides no choice to an individual. For example, if one is hit by a car on the street, it is totally out of that person's control. In contrast, option luck is the luck that provides choices to an individual such as lottery, gambling, etc. Dworkin (1981a) proposed a hypothetical insurance to hedge the bad luck—social background luck, genetic luck, brute luck. This insurance provides choices to individuals so that every kind of luck eventually becomes option luck.

Dworkin (1981a) argued that option luck provides choices to individuals, and hence individuals are responsible for their option luck. If every luck can be transferred into option luck by means of the hypothetical insurance, their resources could be equalized justly.

Since most studies just take social background luck (parents' income and occupation) into consideration, the measurement of inequality of opportunity excludes genetic luck and brute luck. Presumably, it is absorbed in the random component of the model.

2.4.2.2 Distinguishing Circumstances and Effort

After identifying the factors influencing inequality, we need to distinguish between factors that account for circumstance and those that do not. The question is referred to as "responsibility cut" in literature.

Table 2.3: A Literature Review on Responsibility Cut

Reference	Responsibility variable	Non-responsibility variable	Outcome
(Checchi and Peragine, 2010)	unobservable	the level of parents' education, region of birth, sex	individual annual earnings
(Bourguignon et al., 2007)	unobservable	Race, parental schooling, region of birth, father's occupational status	hourly earnings
(Zhang and Eriksson, 2010)	unobservable	Parental household income, Gender, parental education, parental occupational status, household size, region, urban or rural area	average household disposable income
(Björklund et al., 2012)	unobservable	Parental income, parental education, own IQ, number of siblings, body mass index, family structure	total market income before taxes
(Checchi et al., 2010)	unobservable	Parental education, parental occupation, gender, nationality, geographical location	post-tax individual earnings
(Pistolesi, 2009)	unobservable	age, parental education, father's occupation, ethnicity, region of birth	individual annual earnings
(Ferreira and Gignoux, 2008)	unobservable	gender, ethnicity, parental education, father's occupation, region of birth	household per capita income

Table 2.3 shows how literature allocates responsibility. There is no accepted standard in responsibility allocation. Consequently, it is hard to compare these empirical researches with each other. Besides, different literature may have different errors caused by different responsibility cuts.

In empirical research, there are three main ways to allocate the responsibilities between circumstances and effort. Most literature concerns family background and social background as circumstances. Some literature such as Björklund et al. (2012) considers genetic influence, such as IQ as circumstances. One literature (Lefranc et al., 2009) even takes luck into consideration.

Since most literature pays no attention to genetic influence, if we can estimate how these genetic factors influence inequality of opportunity, we may be able to estimate the bias in the empirical studies generated by ignoring genetic factors. One study conducted by Björklund et al. (2012) found that IQ is the most significant factor behind income inequality in Sweden. It accounts for 11.5% in total income inequality compared with the second factor, parent income (7.1%). If genetic effect widely varies from country to country, it is essential for researchers to control for it when measuring inequality of opportunity.

There is some evidence of genetic effects on effort. Mosing et al. (2014) studied 10,500 twins' music performances in Sweden. They found that the genetic factors affected both their willingness to practice and their skills. Music practice was substantially related to heredity (40% —70%). Moreover, their study excluded the possibility of the causal effect of music practice on musical ability. This result indicates that genetic variation affects both ability and inclination to practice.

Another study by Hambrick and Tucker-Drob (2014) shows a similar result. They looked for evidence for gene-environment correlation and interaction with respect to music accomplishment. They found that genetic factors played a more important role in music accomplishment compared with music practice. It is more likely that "genetic potentials for skilled performance are most fully expressed and fostered by practice".

2.4.3 Approaches to Measure Inequality of Opportunity

To measure inequality of opportunity, most of the literature starts from estimating counterfactual distributions. However, researchers apply different theoretical frameworks, make different assumptions and use different empirical models.

For example, some researchers use parametric estimates (Bourguignon et al., 2007) and others prefer non-parametric estimates (Checchi and Peragine, 2010). Ramos and Van de gaer (2016) discussed the pros and cons of the parametric approach. They argued that if non-parametric approach is applied, each group of data should contain a sufficient number of observations. Because circumstance is multi-dimensional, if more circumstances are added, number of types grows exponentially. As a consequence, the observation for each type would be very few. In addition, if one of the circumstances is continuous, the observation for each type would not be enough for non-parametric methodology as well. Another advantage of the parametric approach is that one can measure inequality caused by one circumstance (or a set of circumstances) by controlling other sets of circumstances. Thus, the partial effect of a particular circumstance is able to be measured.

Using a parametric approach (Bourguignon et al., 2007), the outcome of interest can be ideally modelled in the following equation:

$$y = f(\mathbf{c}, \mathbf{e}, u) \tag{2.13}$$

where y is the outcome of interest, $\mathbf{c}$ is the vector of circumstances, $\mathbf{e}$ is the vector of efforts and u is an error term capturing variation due to unobserved factors. Since circumstances and effort have limited categories, population can be grouped into types in which individuals share the same circumstances, and tranches in which individuals share the same effort.

The function f in equation (2.13) could be either linear or non-linear. Although a linear parametric approach would be more easily applied, it could underestimate inequality of opportunity compared with a nonlinear parametric approach (Hufe and Peichl, 2015). To show different methods used in measuring inequality of opportunity, we further assume a linear parametric model:

$$y = \alpha \mathbf{c} + \beta \mathbf{e} + u \tag{2.14}$$

where α, β are vectors of coefficients for vector $\mathbf{c}$ and $\mathbf{e}$.

Given the linear parametric model, counterfactual distributions can be generated based on the estimators of the coefficients in equation (2.14). Inequality of opportunity can be either measured directly by the counterfactual distributions or indirectly by the distance between the counterfactual distributions and the real distributions of y:

$$IOP_D = I(\tilde{y}) \tag{2.15}$$

$$IOP_I = I(y) - I(\tilde{y}) \tag{2.16}$$

where IOP_D is the direct measure of inequality of opportunity, IOP_I is the indirect measure of inequality of opportunity and $\tilde{y}$ is the counterfactual distribution given that effort is fixed. The difference between direct and indirect measure is that direct measure estimates inequality of opportunity IOP by only considering the contribution of circumstances, while indirect measure measures the distance between total inequality and equal opportunity by eliminating the inequalities due to differences in circumstances.

To estimate inequality of opportunity using direct and indirect measures, one should first determine how to compute the counterfactual distributions. The estimation of the counterfactual distributions depends on the theoretical frameworks and data limitations.

For example, effort is normally difficult to observe in datasets. In this case, one can use an ex-ante approach (Van De Gaer, 1993). The counterfactual distribution can be the fitted distribution given circumstances:

$$\tilde{y} = \hat{\alpha} \mathbf{c} \tag{2.17}$$

where $\hat{\alpha}$ is the estimators of an OLS model in which we assumed that effort is unobserved and in the error term u:

$$y = \alpha \mathbf{c} + u \tag{2.18}$$

The advantage of this approach is that only the information on circumstances are required. However, inequality of opportunity is underestimated using this approach because some unobserved circumstances could also be included in the error term. Researchers have found substantial measurement error due to unobserved circumstances. Using Monte Carlo simulations, Lara Ibarra and Martinez Cruz (2015) found that missing a relevant circumstance could lead to up to 80% downward bias. Balczar (2015) also found similar bias using the data of toddlers where effort plays no role.

To address the problem of underestimation, Niehues and Peichl (2014) used a panel data from Germany and the United States and estimated inequality of opportunity using a fixed-effects regression. In their fixed effect model, they assumed that all circumstances are unobserved and interpreted the time-invariant individual effect as the effect of all circumstances. In consequence, their results yielded an overestimated measure.

An alternative approach is to use a non-linear regression. Hufe and Peichl (2015) released the linearity assumption in a conventional framework and included the interaction terms of circumstances in the regression. Their results showed a 50% upwards correction of the downward bias. Donni et al. (2015) used a latent class approach to model the unobserved circumstances. Their method also partially corrected the downward bias.

Inequality of opportunity can be more precisely estimated if effort can be observed. In this case, an ex-post approach (Roemer, 1998) can be more appropriate to measure inequality of opportunity. A counterfactual distribution can be generated given circumstances and effort:

$$\tilde{y} = \hat{\alpha}\mathbf{c} + \hat{\beta}\mathbf{e} \tag{2.19}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimators of parameters in the model.

A direct measure given this counterfactual distribution is to replace each individual's income with the average of $\tilde{y}$ for the type each individual belongs to; that is, for each individual i in type j, suppose the number of individuals in type j is N_j,

$$y_i^d = \frac{1}{N_j} \sum_{1}^{N_j} \tilde{y}_i^j \tag{2.20}$$

where $y^d = \{y_1^d, \cdots, y_i^d, \cdots, y_N^d\}$ is the counterfactual distribution for the direct measure and $\tilde{y}_i^j$ is the outcome for individual i in j type.

An indirect measure is to replace individual's income with the average of $\tilde{y}$ for each tranche; that is, for each individual i in tranche k, suppose the number of individual in tranche k is N_k,

$$y_i^I = \frac{1}{N_k} \sum_{1}^{N_k} \tilde{y}_i^k \tag{2.21}$$

where $y^I = \{y_1^I, \cdots, y_i^I, \cdots, y_N^I\}$ is the counterfactual distribution for the indirect measure and $\tilde{y}_i^k$ is the outcome for individual i in k tranche.

After generating y_i^d and y_i^I, the direct and indirect measures can be estimated using equation (2.16).

One issue for the ex-post approach is that it requires the data on effort. However, effort is related to personal choices and preferences and is hard to be observed. In the second essay, we contribute to the literature by proposing a latent class model to capture the unobserved effort.

Another issue for the ex-post approach is that it is incompatible with the ex-ante approach. Thus, norm-based measures choose a weakened version of this principle by using *norm income.* (See Ramos and Van de gaer (2016) Almas et al. (2011) Devooght (2008)). In order to find the norm income for a particular individual, first we need to define a redistribution mechanism under which both ex-ante and ex-post compensation are satisfied to some extent. Then the norm income is computed based on the mechanism and treated as the income a particular individual should get. The distance from the real income to the norm income is the inequality of opportunity.

One distribution mechanism is called "the generalized proportionality allocation" developed by Bossert (1995) and Almas et al. (2011).

$$z_i^{GPP} = \frac{g(\mathbf{e}_i; \cdot)}{\sum_j g(\mathbf{e}_j; \cdot)} \sum_i y_i$$

where $g(\mathbf{e}; \cdot) = \frac{1}{n} \sum_j f(\mathbf{e}_i, \mathbf{c}_j)$

Another distribution mechanism was developed by Fleurbaey and Bossert (1996) and is called "the egalitarian equivalent allocation".

$$F_k^{EE}(\mathbf{c}, \mathbf{e}) := f(\mathbf{e}_k, \tilde{\mathbf{c}}) - f(\mathbf{c}, \mathbf{e}) \forall \mathbf{c}, \mathbf{e} \in \Omega^n, \forall k \in N.$$

Ramos and Van de gaer (2016) argued that norm-based measure should be used instead of the indirect measure because an indirect measure is only the decomposition of inequality of outcome which ignores inequality of opportunity itself. It answered the question of to which extent inequality of outcome is due to inequality of opportunity. In contrast, direct measure and norm-base measure are concerned with the inequality of opportunity directly.

2.4.4 The Partial Ordering of Inequality of Opportunity

Inequality of Opportunity can also be ranked using the partial ordering approach. On the one hand, Between groups of those who share the same circumstances, less inequality is preferred for an equal opportunity policy. On the other hand, within groups of those who exert the same effort, less inequality is preferred for an equal opportunity policy. Following either criteria, one can use a partial ordering such as the Lorenz or the generalized Lorenz partial ordering to rank difference policies (Peragine, 2004b). Several criteria on the basis of equality of opportunities can also be derived and examined using a partial ordering approach (Peragine, 2004a).

Rodrguez (2008) applies a partial equality-of-opportunity ordering to compare the degree of equality of opportunity among 12 countries. He found that Denmark dominates all other economies in terms of post-tax income.

The benefit of using the partial ordering approach is that one can compare income distribution rather than comparing the measure of income distribution.However, it cannot tell whether one economy dominates another when one inequality-of-opportunity curve cross the other.

2.4.5 Correlation between Circumstances and Effort

Effort could be shaped by circumstances. Jusot et al. (2013) summarized three different views on the correlation between circumstances and effort proposed by Roemer (1998), Barry (2005) and Swift (2005) respectively. They provided an example of education in which some students study hard under the influence of their parents.

Roemer (1998) argued that individuals should be responsible for factors within their control. This view is known as *the control view*. Based on this view, the influence of parents on children's efforts is out of the children's control and it should be considered as circumstances and cleaned from the effort variables.

In contrast with Roemer's view, Rawls (1971) and Dworkin (1981b) argued that individuals should be responsible only for their preferences and choices (known as *the preference view*). Based on this view, Barry (2005) stated that students' efforts are choices based on their own free will, even if made under the pressure of their parents; thus, this type of effort deserves respect.

Another view from Swift (2005) claimed that the pressure is an effort of parents which should be respected. Swift (2005, pp. 271) argued that the effort of parents on their offspring is "an interaction that we have reason to value and protect" and "preventing those interactions would violate the autonomy of the family".

Jusot et al. (2013) assessed whether the different views on the correlation affected the measurement of inequality of opportunity in health. Using the French Health, Health Care and Insurance Survey, they found that these three different views have little influence on the measurement of inequality of opportunity in health.

Based on Roemer's (1998) view, Bourguignon et al. (2007) developed a parametric approach to deal with the correlation problem between effort and circumstances. They used a matrix of coefficients to catch the effect of circumstances on effort which is vector-valued.

$$E_i = HC_i + v_i$$

where E_i is the effort of individual i, C_i is the circumstance of individual i, H is a matrix of coefficients and v_i is the white-noise. This approach models Roemer's view on the correlation between circumstances and effort. It can be easily implemented if both circumstances and effort can be observable.

For example, individual educational level is commonly regarded as a variable related to effort. Therefore, this variable can be regressed with circumstances variables to capture the indirect effect of circumstances on inequality of outcome through educational levels. For implementation of this approach, see Alain et al. (2010), Lazar (2013) and Deutsch et al. (2018).

In addition to Bourguignon's and Ferreira's (2007) model, with effort unobserved, Björklund et al. (2012) models the heteroskedasticity of effort across type so that the effort correlating with circumstances can be captured.

In our study, we propose two models in Chapter 3 and 4 to deal with the correlation issue given that effort variables are unobservable.

2.4.5.1 Human Opportunity Index

Paes de Barros et al. (2009) developed another approach called "Human Opportunity Index" to measure inequality of opportunity in different countries. In light of Human Development Index developed by Sen (1979), Paes de Barros et al. (2009) considered the concept of basic capability, e.g. the capability to attain water and electricity. By focusing on basic capability, this index is more applicable to developing rather than developed countries.

2.4.6 Measure Educational Inequality of Opportunity

The empirical approaches introduced in previous section can also be applied in measuring educational inequality of opportunity. For example, Ferreira and Gignoux (2014) study how circumstances contribute to the test score of the Program of International Student Assessment (PISA) which were conducted in 57 countries in 2006. They use a direct ex-ante approach similar to Paes de Barros et al. (2009) to estimate the lower-bound educational inequality of opportunity. The test score is used as a proxy of educational achievement.

However, when measuring educational inequality of opportunity, some outcome variables could be categorical variables. For example, Brunori et al. (2012) measure inequality of opportunity in the access to tertiary education in Italy. The access of tertiary is a binary variable. Zeng et al. (2014) look into gender inequality in education using educational attainment — a categorical variable — as an outcome variable.

In Chapter 5, we also use a categorical variable — the graduate choice — to study educational inequality in China. This variable is a multinomial variable. Our study shows how circumstances affect college graduates' choices in China.

2.4.7 Inequality Measures

Inequality measures such as the Gini coefficients, Entropy index and the dissimilarity index are widely used in empirical literature.

The Gini coefficients I_{Gini} can be derived from the relative Lorenz curve L which is the normalised cumulative income functional by the mean (Cowell, 2000):

$$L(F;q) \equiv \frac{C(F;q)}{\mu(F)} \tag{2.22}$$

where F is the distribution of income in the population q. C is the cumulative function and $\mu(F)$ is the mean of the distribution of income.

Thus, the Gini coefficient can be expressed as the normalised area between the Lorenz curve and the 45 degree line (Cowell, 2000):

$$I_{Gini}(F) \equiv 1 - 2\int_0^1 L(F;q)dq \tag{2.23}$$

Entropy indices such as the Theil index and the mean logarithmic deviation (MLD) index originate from the information-theoretic idea, which can be expressed in the following inequality index (Theil, 1967).

The Theil index can be expressed as:

$$I_{Theil}(F) \equiv \int \frac{x}{\mu(F)} \log(\frac{x}{\mu(F)})dF(x) \tag{2.24}$$

The MLD index can be expressed as:

$$I_{MLD}(F) \equiv -\int \log(\frac{x}{\mu(F)})dF(x) \tag{2.25}$$

They can be generalised into a single more flexible class—the generalised entropy (GE) family of measures (Cowell, 2000):

$$I^{\alpha}_{GE}(F) \equiv \frac{1}{\alpha^2 - \alpha}\int \log[(\frac{x}{\mu(F)})^{\alpha} - 1]dF(x) \tag{2.26}$$

where $\alpha \in (-\infty, +\infty)$ captures the sensitivity of a specific GE index to particular parts of the distribution. The Theil index and the MLD index are two special cases of the GE index. When $\alpha = 0$ or 1, the GE index becomes the MLD index and the Theil index respectively. Since positive and larger α entail more sensitivity to changes in the distribution that affect the upper tail, the Theil index is sensitive to the extreme rich while the MLD index is sensitive to the extreme poor.

Checchi and Peragine (2010) and Ferreira et al. (2011) prefer the MLD index rather than the Theil index and the Gini index since it is the only decomposable index which is path-independent (Foster and Shneyerov, 2000). A path-independent index makes sure that the sum of between-group inequality and within-group inequality is overall inequality.

An alternative is to apply the Shapley decomposition (Shapley, 1953) to any inequality index to get a path-independent measure. The Shapley decomposition provides flexibility to choose different inequality indexes with no violation of path independence.

Assume that a set of factors X_k indexed by $K = \{1, \dots, k, \dots, m\}$ with a characteristic function (the inequality index) $I : 2^K \to \mathbb{R}$. The set of factors can include circumstances such as gender, ethnicity and parents' socioeconomic status, and effort. Some researches (e.g. Zhang and Eriksson 2010 and Manna and Regoli 2012) decomposed only the predicted income for each circumstance. This approach shows how each circumstance contributes to total inequality of opportunity (Israeli, 2007). Björklund et al. (2012) take effort as one of the factors. Based on the Shapley decomposition, the factor k's contribution is determined by the Shapley value of k: $\phi_k(I)$ that can be calculated using the following equation:

$$\phi_k(I) = \sum_{S \subseteq K \setminus \{k\}} \frac{|S|!(m - |S| - 1)!}{m!} (I(S \cup \{k\}) - I(S)) \tag{2.27}$$

where S is the subset of K without k and $|S|$ is the number of factors in S. In this equation, $I(S \cup k) - I(S)$ is the marginal contribution of the factor k to total inequality and $\phi_k(I)$ can be interpreted as the average marginal contribution of all possible permutations in which factor k affects inequality jointly with other factors in the set S.

Using this equation, one can decompose total income inequality into inequality contributed by circumstances and effort. In Chapter 3, we use the Shapley decomposition method to decompose income inequality measured using the Gini coefficients. We show that, without using Shapley decomposition, the Gini coefficients could overstate inequality of opportunity.

2.5 The Economic Consequences of Inequality of Opportunity

Researchers have studied the economic consequences of inequality of outcome (such as income inequality and educational inequality). However, different strands of the literature draw different conclusions. For example, a common topic is the effect of income inequality on economic growth. Kuznets (1955) found an inverted U-shape relationship between income inequality and economic growth. However, other empirical findings suggested a positive (Li and Zou, 1998) or a negative (Clarke, 1995) relationship.

These contradictory findings might be due to the fact that different factors within income inequality could have the opposite effect to economic growth. Marrero and Rodriguez (2013) postulated that this opposite effect comes from inequality of opportunity. Using the data from the U.S., they found a negative relationship between inequality of opportunity and growth, and a positive relationship between inequality of effort and growth.

However, Ferreira et al. (2014) found that this negative relationship is not robust when using a cross-country meta-dataset. They used a dataset containing 118 household surveys and 134 Demographic and Health Surveys and found no evidence of the relationship between inequality of opportunity and growth.

In addition, Brunori et al. (2013) found an inverted U-shape curve of inequality of opportunity with economic development. Given the empirical findings, the relationship between inequality of opportunity and economic growth is also inconclusive.

Scholars also examined the relationship between inequality of opportunity, income inequality and intergenerational mobility. Brunori et al. (2013) showed that inequality of opportunity is positively related to income inequality and negatively related to intergenerational mobility. Corak (2013) postulates that high income inequality could result in high inequality of opportunity and low intergenerational mobility. He illustrated his idea using the Panel Study of Income Dynamics (PSID) from the U.S. For example, he showed that higher income families in the United States spend much more on their children purchasing things that promote the capabilities of their children such as books, high-quality child care, private schooling, etc. The expenditure gap grew during the period 1970 to 2010 as the income gap has increased since 1970. In this way, higher income parents transfer their economic advantage to their children, which leads to low intergenerational mobility and high income inequality.

In summary, inequality of opportunity is associated with income inequality and intergenerational mobility; however, its effect on economic growth still needs further investigation.

2.6 The Literature on Inequality of Opportunity in China

There are some recent researches on inequality of opportunity in China.

Using data from the China Health and Nutrition Survey (CHNS), Zhang and Eriksson (2010) measured inequality of opportunity in nine Chinese provinces from 1989 to 2006 using a parametric approach and found a substantial degree of inequality of opportunity due to parental income, and to father's and mother's type of employer which accounted for 23%, 19% and 20% of inequality of opportunity, respectively. However, parents' education has little influence on the earnings of their offspring, which implies parents' social connections have a more crucial effect on the earnings of their offspring than their intelligence.

They also found that the annual ratio of inequality of opportunity to income inequality from 1989 to 2006 in China ranged from 0.46 to 0.65. A similar study by Bourguignon et al. (2007) reported the ratio of inequality of opportunity to income inequality in Brazil in 1996 as 0.23. The difference between two studies is that the former used the Gini coefficients as the inequality measure and the sample size covered both rural and urban areas while the latter used the Theil index and the sample size covered only urban areas.

To estimate inequality of opportunity, Zhang and Eriksson (2010) assumed that income is a function of individuals' circumstances which can be observed and effort which is unobserved. Thus, after identifying the effect of circumstances, the remaining unexplained part of income is due to effort.

The model was specified as:

$$\begin{aligned} lnincome =& \alpha_0 + \alpha_1 gender + \alpha_2 Province + \alpha_3 Urbanarea + \alpha_4 Age + \alpha_5 Father's education \\ &+ \alpha_6 Mother's education + \alpha_7 Father's employer + \alpha_8 Mother's employer \\ &+ \alpha_9 ln(householdincome) + \alpha_{10} Householdsize + \alpha_{11} waveofsurvey + e \end{aligned} \tag{2.28}$$

In this equation, they included both parents' education attainment, occupation and household's income. The educational attainment was categorized into four levels—primary or less, middle school, high school or vocational school and College and above. Parental employment was classified into five different groups according to the type of employer they worked for—farming, collective, private enterprise, government or state-owned enterprise and foreign-owned enterprise. In addition, in the equation, they also considered the region of households—whether the province they lived in was inland or coastal; and whether households lived in urban or rural areas. The inclusion of region enabled the study to examine not only family background effect but also regional disparities.

Additionally, they applied sample restrictions on the CHNS data set. First, individuals lacking information about their parental background or their children's background were excluded. Second, they collected data only from the participants aged between 20 to 50 for the reason that participants under 20 years old may still receive education and those participants above 50 years old commonly have missing values for parents' income. Eventually, an unbalanced panel with 1287 valid observations during 1989-2006 was generated after data filtering.

However, the descriptive statistics provided show that 18.4% respondents' fathers worked as farmers and 57.8% worked in government or state-owned enterprises. For mothers, 23.2% are farmers and 48.0% worked for the government or state-owned enterprises. In this thesis we found that at least 50% of parents were farmers whether before or after data filtering. In contrast, the original data provided by CHNS[1] showed that 34.33% were farmers, fishermen and hunters, while Zhang and Eriksson (2010) found that 12% of the offspring were farmers.

Another statistic they reported was the percentage of births in urban areas which was

[1] See the CHNS codebook:
http://www.cpc.unc.edu/projects/china/data/datasets/longitudinal/codebook/jobs_00.pdf

51%. Since the study focused on the years between 1989 and 2006 and the average age they reported was 26.5 years old, the sample's year of birth ranges from 1962 to 1979 on average. According to the National Bureau of Statistics of China (NBS, 2013), the urban population between 1966 to 1979 ranged from 18% to 19%. Since the percentage of place of birth in urban areas is 51%, the current percentage of place of living in urban areas will be higher due to the urbanization after 1978. Although the authors claimed that inequality of opportunity is measured in a way that both urban and rural samples are included, their sample filtering excluded most rural residents and the results are therefore more representative of urban residents rather than the whole of China. This bias might be one explanation for the small proportion of parents' occupation being listed as a farmer.

Compared with Zhang and Eriksson (2010), Bourguignon et al. (2007) measured inequality of opportunity in Brazil in 1996 using only the urban sample. However, in 1996, the urban population in Brazil was around 80%. Their estimation of inequality of opportunity represents the majority of the population in Brazil. While in China, the urban population before 1980 was lower than 20%, the results from Zhang and Eriksson (2010) can only represent the minorities, half of whom were born in urban areas before 1980.

In terms of the methodology, the inequality of opportunity can be measured by $\frac{G(\widehat{lnincome})}{G(\widetilde{lnincome})}$ where $G(\widehat{lnincome})$ is the Gini coefficient of the fitted logarithm of income from equation (2.28) and $G(\widetilde{lnincome})$ is the Gini coefficient of the observed logarithm of income from the dataset.

Zhang and Eriksson (2010) also estimated the contribution of individual components in equation 2.28 to the overall inequality of opportunity. In the implementation, first, they drop the variable they would like to measure and do the regression again. Then, they get the fitted income without the dropped variable, let's say, the fitted income without $gender$—$\widehat{lnincome}_{nogender}$. Thus, the contribution of $gender$ to the overall inequality of opportunity can be measured by $\frac{G(\widehat{lnincome})-G(\widehat{lnincome}_{nogender})}{G(\widehat{lnincome})}$, where $G(\widehat{lnincome}_{nogender})$ is the Gini coefficient of predicted income without the variable $gender$. This approach might also be problematic since the estimators are estimated twice. A correlation between the omitted variable with remaining regressors may lead to two different estimators between regression with and without the omitted variable, which may increase the bias.

In addition, they used the Gini coefficient as the inequality measure with the log-linear model. Since the Gini coefficient cannot be decomposed, using $\frac{G(\widehat{lnincome})}{G(\widetilde{lnincome})}$ might lead to bias. In our study, we show results using the Shapley decomposition to decompose the Gini index and estimate inequality of opportunity.

2.7 A Summary of the Literature Review

The literature on inequality of opportunity established the theoretical and empirical foundation on decomposing inequality of outcome into circumstances and effort. It examines the sources of inequality of outcome by considering contributions of individual's circumstances to it.

Several theoretical and empirical issues still remain to be resolved. On a theoretical level, it is still not clear whether circumstances and effort can be distinguished and separated from inequality of outcome. For example, how should the individual's responsibility for their outcome be allocated? Are luck and genetic factors identified as circumstances or effort? How about different types of luck, such as brute luck and option luck (Dworkin, 1981b)? Although these questions are important to the theory of inequality of opportunity, they are not within the scope of the thesis.

Instead, the thesis mainly focuses on issues of the empirical implementation. We consider circumstances and effort, observed or unobserved, that may correlate with each other. Some researchers such as Jusot et al. (2013) and Bourguignon et al. (2007) tried to solve the correlation between circumstances and effort. Their methods are applicable under the assumption that circumstances and effort are both observable. This thesis proposed alternative approaches to the correlation issue given that efforts are unobserved.

In terms of the literature in China, Zhang and Eriksson (2010) measured inequality of opportunity in China during 1989-2006. However, the data might be out of date and represent a biased representation of the current situation in China. In addition, the method used in Zhang and Eriksson (2010) assumes independence of circumstances and effort and decomposes income inequality using the Gini coefficients which are path-dependent. This thesis investigates inequality of opportunity using a more contemporary dataset and improves the methodology by considering the correlation between circumstances and effort and using better inequality measures.

Chapter 3

Essay I: Measuring income inequality of opportunity in China

3.1 Introduction

The real per capita income in China has grown at an impressive rate in the last two decades, but so has income inequality. The Gini coefficient, an indicator of income inequality, rose from under 0.3 before 1980 to 0.55 in 2012 (Xie and Zhou, 2014); it is higher than in the U.S. (0.41) (The World Bank, 2016a) but similar to some Latin American countries such as Brazil (0.53) and Colombia (0.54). The increase in inequality was not due to a fall in the income levels of the poor [1], but due to the more rapid income growth of the rich (Li et al., 2013). This finding raises questions about the change in income distribution in China and its sources. In particular, what is the main source of the divergence in income growth between the poor and the rich?

The public is also aware of high income inequality in China. The International Social Survey Program (ISSP) (2009) surveyed perceptions of economic inequality in 38 countries in 2009. Most Chinese respondents tended to agree with the statement: "Income differences in China are too high" and the conceding rate is on par with the other 37 countries. More importantly, Chinese respondents have the lowest "feeling of procedural justice" Larsen (2016) among all respondents in the ISSP survey[2]. Most respondents strongly believed that socio-political connections and parents' socio-economic backgrounds were important for getting ahead in society.

Many researchers have studied income inequality and its determinants in China. They found that income inequality rose with regional disparities (Knight and Song, 1993, Wan

[1] Quite the opposite, the poverty rate decreased from 85% to lower than 11% during 1980-2012 (The World Bank, 2016b).

[2] ISSP asked respondents to what extent do "coming from a wealthy family", "having well-educated parents", "knowing the right people", "having political connections" and "giving bribes" are important to get ahead in society. (Larsen, 2016) combined these questions into a measure of perceptions on "procedural justice". This measure captures the extent to which people believe they used privileges to get ahead in society.

and Zhou, 2005), globalization (Wan et al., 2006), migration (Park and Wang, 2010), urbanization (Wu and Rao, 2017) and private ownership of assets (Li et al., 2013). These studies show the sources of income inequality from different perspectives; however, they focussed little on "procedural justice" and their findings might not explain people's perception of procedural unfairness in China.

In this paper, we try to fill this knowledge gap using the framework of equal opportunity (Roemer, 1998, Cohen, 1989, Arneson, 1989), in which society should only be concerned with inequality due to factors beyond individuals' responsibility ("circumstances") and acknowledge inequality due to factors within individuals' responsibility ("effort"). Inequality caused by circumstances is defined as "inequality of opportunity" (IOP). We discussed the framework of equal opportunity in Section 2.3 in depth.

If China has a higher IOP than other countries, it may explain the public perception of poor "procedural justice" in the country[3]. Implementing the theory of equal opportunity requires first a working definition of individual responsibilities or circumstances. In this essay, we define *observed* factors such as gender, ethnicity and parents' socioeconomic status as "circumstances" and effort as an *unobserved* factor. Furthermore, due to the very short timeframe of our dataset, circumstances are treated as *time-invariant* variables, while effort is *time-variant.*

In this paper, we used a representative dataset drawn from the China Family Panel Study (CFPS), which contains 33,600 individual observations for the years 2010 and 2012. We measured IOP at the national, regional, and provincial levels respectively. In addition, we investigated the relationship between the provincial gross regional product (GRP) and IOP. Since the data spans only three years, the emphasis is on the cross-regional variation in IOP. Although China as a whole is growing rapidly, the level and change of development and inequality differs vastly across various Chinese regions. The underdeveloped north-west of China has relatively high income inequality, while the highly developed south-east region has relatively low income inequality. Therefore, a cross-regional comparison in inequality can be used as a vehicle to assess how inequality might change with development and how IOP contributes to that change.

Since the data include samples with no income, we estimated the probability of earning positive income through a Heckman model and a hurdle model. In addition, Björklund et al. (2012) showed that the correlation between circumstances and effort is a source of IOP. They used a non-parametric approach to measure the heterogeneous effect of effort across circumstances. However, this approach only showed the effect of heterogeneity as a whole but failed to reveal the effect of each circumstance. Instead, we took a parametric approach and used maximum-likelihood estimation (MLE) to show the effect of each circumstance variable on heteroskedasticity.

[3] Some institutional features such as corruption are not considered in this paper but they may also explain the public perception of procedural fairness.

To better understand the roles of circumstances and effort in driving inequality in China, we conducted two decompositions. First, we followed Björklund et al. (2012) and applied the Shapley decomposition (Shorrocks, 2013) to identify the contributions of circumstances and effort to income inequality. This decomposition technique allows us to use a common inequality index —Gini coefficient— without violating path dependency (Foster and Shneyerov, 2000).

Second, we applied the Oaxaca decomposition (Oaxaca, 1973) to identify the differential effects of circumstances on income across the advantaged and disadvantaged groups (e.g. gender, ethnicity, etc.). The IOP measure only provides an overview of the unfair part of inequality, while the Oaxaca decomposition can reveal whether the higher income for the advantaged group is due to their better circumstances or bigger influence of their circumstances on income.

The main contribution of this paper is to evaluate IOP in China at both national and regional levels using representative, cross-sectional data. Taking advantage of heterogeneity of Chinese regions, this study sheds light on the contribution of circumstances to overall income inequality over various development stages. Moreover, this study is the first one to apply the hurdle model with the Shapley decomposition to include those who receive no income and to show the heterogeneous effect of each circumstance by implementing MLE.

We found that at the national level, circumstances accounted for at least 30% of the income inequality in China in 2010 and 40% in 2012. These figures rise by about 20% if we include heteroskedasticity between types as parts of IOP, which indicates a significant effect of circumstances on income through effort. Among this IOP, gender, geographic characteristics and parents' socioeconomic status are the three main factors for income inequality.

At the provincial level, GRP appears to have a negative relationship with income inequality and inequality of effort, but no discernible relationship with the level of IOP. As a result, the share of IOP in the overall inequality rises with the increase of GRP. Lastly, results from the Oaxaca decomposition show that getting rich does not require better circumstances per se but rather, the bigger influence of circumstances to income. In addition, the shares of IOPs in the overall inequality are similar across regions.

The rest of this paper is organized as follows. In section 3.2 we describe the approach to measuring inequality of opportunity. In section 3.3 we discuss the empirical strategies. Section 3.4 is the description of data. Section 3.5 shows the empirical results and section 3.6 is the conclusion.

3.2 Literature on Measuring Inequality of Opportunity

Equality of opportunity was first conceptualized by John Rawls, who argued that "offices and positions must be open to everyone under conditions of fair equality of opportunity" (Rawls, 1971, p. 302). Based on Rawls's argument, Roemer (1998) proposed a framework to measure IOP. He started with a classification of people based on types and tranches: those sharing same circumstances belong to the same type and those exerting the same level of effort share the same tranche.

Based on Roemer's division, IOP can then be measured *ex-post* or *ex-ante*. The ex-post IOP captures the within-tranche inequality, the inequality of a counterfactual income distribution where all tranches have the same mean income (Checchi and Peragine, 2010). Ex-post IOP is driven entirely by inequality between types conditional on effort. In other words, the ex-post approach defines equal opportunity as individuals with the same effort receiving the same outcome regardless of their types. On the contrary, the ex-ante IOP is the between-type inequality, the inequality of a counterfactual income distribution where everyone in a type has the same type-average income (Checchi and Peragine, 2010). This approach is based on a weaker definition of equal opportunity in which the inequality within tranches is allowed. Due to the strong definition and high data-demand of the ex-post approach, most literature uses the ex-ante approach. In this paper, we also use the ex-ante approach.

Based on an ex-ante approach, a common application of this framework is to assume that circumstances and effort are independent and measure how circumstances and effort contribute to total inequality respectively.[4] However, this method ignores the correlation between circumstances and effort. If the correlation exists, the measure of inequality of opportunity would be biased.

To address this issue, firstly, we need to clarify whether the correlation between circumstances and effort should be respected as individual effort or not. Roemer (1998) argued that individuals are not held responsible for the correlation between circumstances and effort because the correlation is also beyond the individual's *control*. In contrast, Barry (2005) believed that this correlation should be respected because it reflects individuals' *preferences* even though it could be shaped by circumstances. This paper adopts Roemer's view because the correlation could also be a significant source of income inequality, and studying the correlation between circumstances and effort is one objective in this paper.

Another problem related to the correlation between circumstances and effort is the observability of effort. Due to the difficulty in defining and observing effort, many researches assume effort as totally unobserved (e.g. Ferreira et al. 2011). If effort can be observed, one can use a structural equations model to identify the relationship between

[4] We discuss different approaches in measuring inequality of opportunity in Section 2.4 in detail.

circumstances and effort (See Bourguignon et al. 2007 and Jusot et al. 2013). If effort cannot be observed, one can capture the correlation by the type-specific variances (Björklund et al., 2012). If no correlation exists between circumstances and effort, the type-specific variance should be homogeneous. Since this paper assumes effort cannot be observed, we use the type-specific variances to identify the correlation between circumstances and effort.

Roemer's framework has widely been applied in empirical researches to measure inequality of opportunity in many countries. Paes de Barros et al. (2009) estimated the ex-ante IOP in seven Latin American countries[5]. They found that although Mexico has the highest overall income inequality, the contribution of IOP (20.8%) is the smallest among the seven countries. The biggest share of IOP (37.3%) belongs to Guatemala. Using longitudinal data, Pistolesi (2009) showed that rising income inequality in the U.S. during 1968-2001 was not driven by increases in IOP. In fact, they found that IOP in the U.S. had decreased from 43% to 20% over the period. Björklund et al. (2012) differ from many other studies by including individual IQ and body mass index as circumstances in a Swedish study and used the Shapley decomposition to decompose the effects of circumstances. They found that the share of IOP to total income inequality was less than 30%.

Only one research measured inequality of opportunity in China. Zhang and Eriksson (2010) estimated that IOP moved broadly in sync with overall income inequality during 1989 to 2006, with its share of overall income inequality ranging from 46% to 65%. They also found that the IOP was largely due to parental socio-economic status. However, due to lack of information about parents' socio-economic circumstances, most of their estimations were restricted to the urban population or state-owned enterprise workers which were mostly urban-based. Therefore, the study omitted the rural population, which accounted for 55% to 74% of the population during the sample period.[6] In addition, the study measured IOP using the Gini coefficient but did not correct for the bias caused by the coefficients *path-dependency* property (Foster and Shneyerov, 2000).

Some studies went beyond measuring IOP and examined the impact of IOP on development. Using data from 42 countries, Ferreira et al. (2014) found IOP to have a negative growth effect but the result is neither conclusive nor robust. In Marrero and Rodriguez (2013), IOP was found to have a negative growth effect in rich countries only, while both IOP and inequality of effort [7] enhanced growth in poor countries. In this paper, we examine the relationship between IOP and development at the provincial level in China.

Since the empirical researches introduced above used different approaches, inequality

[5] The seven countries are: Brazil, Colombia, Ecuador, Guatemala, Panama, Peru and Mexico

[6] China's rural population share has been declining steadily over time. Source of data: World Development Indicators.

[7] It is the counterfactual inequality after filtering out the effect of circumstances.

measures (e.g. Gini, Theil index, etc.), and definitions of circumstances and effort, one should be cautious about comparing their findings.

3.3 Measuring Inequality of Opportunity

3.3.1 The Model with Independent Circumstances and Effort

To measure IOP, we followed the approaches introduced by Checchi and Peragine (2010). We first partitioned an income profile according to circumstances and effort. Assuming that individuals' income y ($y_i \geq 0$) is determined by a finite set of exogenous and time-invariant circumstances $\mathbf{c}$ and one-dimensional continuous unobserved effort e, we proposed a function g:

$$y = g(\mathbf{c}, e) \tag{3.1}$$

where $\mathbf{c}$ is a set of variables concerned as circumstances with n finite values so that each value represents a *type* in which individuals have the same circumstances. Effort with m finite values is represented by e. We assume that each value of e represents a *tranche* in which individuals have the same effort.

This model also excludes the existence of random components or luck (Lefranc et al., 2008) and interaction between circumstances and effort. Therefore, the following two basic assumptions are satisfied given the non-observability of effort (Checchi and Peragine, 2010):

Assumption 3.3.1 *Function g is monotonically increasing in effort e.*

Assumption 3.3.2 *The conditional distribution of effort e is independent of circumstance* $\mathbf{c}$

The first assumption indicates that the more effort one exerts, the more income one earns, and the second assumption implies the independence between effort and circumstance. If equation (3.1) satisfies both assumptions, one can directly measure ex-ante IOP by computing the inequality of a counterfactual income distribution in which the contribution of effort has been eliminated (Ramos and Van de gaer, 2016). We denote this counterfactual income distribution as Y_c.

Therefore, IOP can be measured by Y_c. In this paper, we used two indexes introduced by Ferreira and Gignoux (2008): one for the absolute level of IOP —Inequality of Opportunity Level (IOL)— and the other for the share of IOP relative to total income inequality— Inequality of Opportunity Ratio (IOR). The former index is given by:

$$IOL = I(Y_c) \tag{3.2}$$

where the function $I : \mathbb{R}_+^N \mapsto \mathbb{R}_+$ is an inequality index such as Mean log deviation, Theil index and Gini coefficients.

The latter index is given by:

$$IOR = \frac{I(Y_c)}{I(Y)} \tag{3.3}$$

An alternative approach —the indirect measure— is to estimate a counterfactual income distribution Y_e by ruling out the contribution of circumstances and to measure inequality of opportunity by substracting inequality of Y_e from income inequality.

We defined the inequality of Y_e as the level of inequality of effort (EOL):

$$EOL = I(Y_e) \tag{3.4}$$

Based on the indirect measure, one can compute inequality of opportunity using the following equations:

$$IOL = I(Y) - EOL = I(Y) - I(Y_e) \tag{3.5}$$

$$IOR = 1 - EOR = 1 - \frac{I(Y_e)}{I(Y)} \tag{3.6}$$

where EOR is the ratio of inequality of effort.

To estimate the counterfactual income distribution, one can decompose the observed income distribution into two — a smoothed between-type distribution by replacing the within-type income with a type-average income, and a standardized within-type distribution that eliminates the differences in the type-average incomes (Foster and Shneyerov, 2000). Thus, Y_c and Y_e can be denoted as the following vectors (Checchi and Peragine, 2010):

$$Y_c = \{\mu(Y_1)\mathbf{1}_{N_1}, \ldots, \mu(Y_k)\mathbf{1}_{N_k}, \ldots, \mu(Y_n)\mathbf{1}_{N_n}\} \tag{3.7}$$

$$Y_e = \{\tilde{Y_1}, \ldots, \tilde{Y_k}, \ldots, \tilde{Y_n}\} \tag{3.8}$$

where $\mathbf{1}_{N_k}$ is the unit vector of length equal to type k's population and $\tilde{Y_k} = \frac{\mu(Y)}{\mu(Y_k)} Y_k$. This decomposition approach avoids the value of income in either counterfactual income distribution becoming negative.

To compute $\mu(Y_k)$ for each type k, we rely on a parametric model. Suppose e is unobserved; an individual's log income can be regressed on circumstances $\mathbf{c}$ using the following equation (Ferreira and Gignoux, 2008):

$$\ln y_i = \boldsymbol{\beta}\mathbf{c}_i + v_i \tag{3.9}$$

where β is a vector of coefficients and v_i is a normally distributed error term, i.e. $v|\mathbf{c} \sim$

$Normal(0, \sigma^2)$. The predicted value is $\mu(Y_k)$ when circumstances $\mathbf{c}_i$ corresponds to type k: $\mu(Y_k) = \bar{y_i} = \exp(\boldsymbol{\beta}\mathbf{c}_i + \frac{\sigma^2}{2})$. We prefer this parametric model to a non-parametric model because the measure would be more likely to be biased using a non-parametric model if there are a large number of types.

3.3.2 The Model with Circumstances and Effort Correlated

In reality, effort could be shaped by circumstances so that assumption 3.3.2 is likely to be violated. In this case, inequality of opportunity is comprised of inequality caused by circumstances directly and inequality caused by circumstances indirectly through effort.

For measuring the direct effect of circumstances to income inequality, one can measure the inequality of the expected income given circumstances. For each individual, the expected income conditional on circumstances can be computed using the following equation:

$$E(y_i|\mathbf{c}_i) = \mu + \boldsymbol{\beta}\mathbf{c}_i \tag{3.10}$$

Measuring the inequality of this conditional expected income, however, eliminates the degree of effort — that is, the deviation of the observed income from the expected income. Björklund et al. (2012) argued that this deviation can also be a part of inequality of opportunity if it varies across types. They identified the type-specific variances and account them for the indirect effect of circumstances to income inequality.

Björklund et al. (2012) separated the contribution of effort to income Y_e into a type-specific heterogeneous effort $\tilde{Y_e}$ and a standardized homogeneous effort u, in which $u_i = Y_{ei}\frac{\sigma_i}{\sigma_{ti}}$ and $\tilde{Y_e} = Y_e - u$. The type-specific effort $\tilde{Y_e}$ is determined by variances because each type has its own variance $\sigma_t^2 = Var[Y_e|\mathbf{c}]$. Therefore, the type-specific effort $\tilde{Y_e}$ is recognized as the correlation between circumstances and effort and considered as parts of circumstances, while the standardized effort u is the pure effort and should be respected.

This approach works well to identify the total contribution of the correlation. Björklund et al. (2012) found that the correlation accounts for around 4.3% to 6.4% of total income inequality. However, this approach failed to identify to what extent each circumstance contributes to total income inequality through the correlation between circumstances and effort.

To capture heteroskedasticity and identify the contribution of each circumstance, we use maximum-likelihood estimation(MLE). We identified the contribution of effort to income Y_e by rescaling individual income so that every type has the same average income[8]:

$$y_{ei} = \frac{\mu(Y)}{E(y_i|\mathbf{c_i})} y_i \tag{3.11}$$

[8] Björklund et al. (2012) measured effort by the residual $Y_e = Y - E(Y|\mathbf{c})$. We did not use this method because the residual could be negative, which could imply counterintuitive negative effort.

Then, we specified the skedasticity function of income with respect to circumstances **c**:

$$\sigma_i^2 = \exp(\mathbf{c}_i'\theta) \tag{3.12}$$

Under the assumption of normal distribution of income[9], the likelihood function is:

$$f(y|\mathbf{c}) = (\frac{1}{\sqrt{2\pi \exp(\mathbf{c}_i'\theta)}})^{n/2} \times \exp[-\sum_{i=1}^{n} \frac{(y - \mathbf{c}_i'\beta)^2}{2\exp(\mathbf{c}_i'\theta)}] \tag{3.13}$$

Using MLE, we can estimate both β and θ. The estimators of MLE allow us to further standardized Y_e with respect to its variance:

$$\tilde{y}_{ei} = y_{ei} \times \sqrt{\frac{Var(y_{ei})}{\hat{\sigma_i^2}}} \tag{3.14}$$

where $\tilde{y}_{ei}$ is the homogenized effort and is independent of circumstances, and $\hat{\sigma}_i^2$ is the estimator of σ^2 in Equation 3.12. Since the type-specific variances are parameterized, we can estimate and measure the contribution of each circumstance to income inequality through effort.

3.4 Empirical Strategy

In this section, we discuss the implementation of the econometric methodology. We used three models to measure inequality of opportunity. First, we used the multiple regression model in Equation (3.9) to estimate the expected income for each type. This model assumes linearity and heteroskedasticity.

Since more than 6% of the individuals in our dataset have zero income, we used the Heckman model (Heckman, 1979) and the lognormal hurdle model (Cragg, 1971) to account for the zero income individuals. The Heckman model assumes that individuals choose not to work when their reservation wage is greater than the market wage, while the hurdle model assumes that zero income is a result of choice without referring to reservation wage. Both of these models are superior to dropping zero income individuals from the sample.

The third model to be considered is the heteroskedastic model (Equations (3.12) and (3.13)) that extends the lognormal hurdle model by additional consideration of the correlation between circumstances and effort. After estimating these models, we implemented the Shapley decomposition, and measured inequality of opportunity not only for all circumstances combined but also for each circumstance as well. The last part of the analysis is based on the Oaxaca decomposition. It is used to examine whether the income gap be-

[9] If income is log-normally distributed, then y represents log income in Equation (3.13)

tween types is due to different levels of circumstances or different effects of circumstances on income.

3.4.1 The Heckman Model

The Heckman model (Heckman, 1979) assumes that people with zero incomes could have potentially earned wages. Wages could not be observed because their personal reservation wage is higher than the market wage. If income is determined by circumstances:

$$y = \mathbf{c}\boldsymbol{\beta} + u_1 \tag{3.15}$$

Income y can be observed if y is greater than the reservation wage y^*.

$$y - y^* = \mathbf{z}\boldsymbol{\gamma} + u_2 \tag{3.16}$$

where $\mathbf{z}$ is the vector of variables affecting the difference between the market wage and the reservation wage. Since equation (3.16) determines whether observations are selected, we refer to this equation as the *selection equation*. Accordingly, equation (3.15) is the *level equation*.

The errors that are correlated with each other are u_1 and u_2:

$$corr(u_1, u_2) = \rho \tag{3.17}$$

where ρ is the correlation coefficient between two errors. If $\rho \neq 0$, the estimator in equation (3.15) is biased. Equation (3.16) helps correct the selection bias. If $\rho = 0$, this model is equivalent to the hurdle model in which equations (3.15) and (3.16) are two independent equations.

In addition, the Heckman model and the hurdle model are different in that the hurdle model assumes no reservation wage. Since the measure of inequality of opportunity relies on observed incomes rather than counterfactual incomes or reservation wages, we used the results from the hurdle model to compute the measure of inequality of opportunity. Nevertheless, we still used the Heckman model as a tool to explore the effect of circumstances on labour participation and as a robust test of the measure of inequality of opportunity.

3.4.2 The Lognormal Hurdle Model

The lognormal hurdle model (Cragg, 1971) takes the zero-income observations into consideration without assuming a reservation wage. The model consists of a binary outcome model, which is used to account for zero versus positive incomes, and a nonlinear model, which deals with the positive income.

Assume that income y can be generated as

$$y = wy^* \tag{3.18}$$

where w is a binary variable that equals 1 if $y > 0$, and y^* is a continuous variable that equals y but it is observed only when $w = 1$. We assumed y^* to have a lognormal distribution:

$$y^* = \exp(\mathbf{c}'\boldsymbol{\beta} + u) \tag{3.19}$$

where $\mathbf{c}$ stands for circumstances and the error term $u|\mathbf{c} \sim Normal(0, \sigma^2)$. So the expectation of y^* given $\mathbf{c}$ is:

$$E(y^*|\mathbf{c}) = \exp(\mathbf{c}'\beta + \frac{\sigma^2}{2}) \tag{3.20}$$

To estimate the probability of receiving positive income w, we used the logistic model:

$$Pr(w = 1|\mathbf{c}) = \Lambda(\mathbf{c}'\boldsymbol{\gamma}) \tag{3.21}$$

where Λ is the logistic function and $\boldsymbol{\gamma}$ is the vector of coefficients for the circumstance variables in the logistic model.

Therefore, the between-type income distribution can be represented by the expectation conditional on circumstance variables:

$$y_c = E(y|\mathbf{c}) = \Lambda(\mathbf{c}'\boldsymbol{\gamma}) \times \exp(\mathbf{c}'\boldsymbol{\beta} + \sigma^2/2) \tag{3.22}$$

where y_c represents the expected income in each type.

To estimate the contribution of effort to income, we computed the within-type income distribution by rescaling the observed income until the income distribution in each type has the same mean as the overall income distribution:

$$y_e = \frac{y * \bar{y}}{y_c} \tag{3.23}$$

An alternative approach is to treat the residual of the model as income earned by effort (Ferreira and Gignoux, 2008). We did not use this approach because the residual would sometimes be negative, which might affect the computations of the Gini index (Chen et al., 1982).

To apply equation (3.22), we undertook the estimation in three steps. First, we estimated the binary part using a logistic regression and obtained the estimator $\hat{\boldsymbol{\gamma}}$. Second, we estimated the continuous part using a log-linear regression and obtained the estimator $\hat{\boldsymbol{\beta}}$. The last step was to estimate the predicted income $\hat{y}$. Since y is assumed to follow a log-normal distribution and the true distribution σ^2 could be unknown, we used Duan's (1983) *smearing estimate.*

If u is independent of $\mathbf{c}$, $E(y^*|\mathbf{c}) = E[\exp(u)]\exp(\mathbf{c}\boldsymbol{\beta})$. Let $\tau = E[\exp(u)]$. We can use the estimated smearing factor $\hat{\tau}$ to estimate τ. The value of $\hat{\tau}$ is:

$$\hat{\tau} = N^{-1}\sum_i \exp(\hat{u}_i) \tag{3.24}$$

where $\hat{u}_i$ is the residual of the log-linear regression.

3.4.3 The Shapley Decomposition

When measuring inequality of opportunity, we first decomposed income distribution Y into a smoothed distribution Y_c and a standardized distribution Y_e. the contribution of circumstances (IOL) and effort (EOL) can be computed based on Y_c and Y_e respectively using Equation (3.2) and (3.4). To ensure IOL is consistent with IOE, we require the sum of both indexes to be the total income inequality:

$$I(Y) = I(Y_c) + I(Y_e) \tag{3.25}$$

However, Equation (3.25) holds only if the inequality index $I()$ is path independent[10]. One path independent inequality measure is the mean log deviation (MLD). An alternative is to apply the Shapley decomposition (Shapley, 1953) to any inequality index. The Shapley decomposition provides flexibility to choose different inequality indices with no violation of path independence. Zhang and Eriksson (2010) used the Gini coefficients of $I(Y_c)$ to measure inequality of opportunity without the Shapley decomposition. In Section 3.6, we show that this method could lead to an overestimated IOL.

We used the Shapley decomposition to decompose income inequality by assuming a set of factors X_k indexed by $K = \{1, \ldots, k, \ldots, m\}$ with a characteristic function (the inequality index) $I : 2^K \to \mathbb{R}$. The set of factors can include both circumstances such as gender, ethnicity and parents' socioeconomic status, and effort. Some researchers (e.g. Zhang and Eriksson 2010 and Manna and Regoli 2012) regressed income with circumstances and decomposed the predicted income for each circumstance. This approach shows how each circumstance contributes to total inequality of opportunity but not income inequality (Israeli, 2007).

Deutsch and Silber (2007) demonstrates an alternative approach to decompose income inequality by population subgroups using the Shapley decomposition. If those population subgroups indicate differences between individuals' circumstances, decomposing income inequality into the between-group and within-group inequality can effectively measure

[10] *Path independence* holds when overall inequality is the sum of between-group inequality and within-group inequality (Foster and Shneyerov, 2000)

inequality of opportunity.

In this paper, we follow Björklund et al. (2012), taking effort as one of the factors. Based on the Shapley decomposition, the factor k's contribution is determined by the Shapley value of k: $\phi_k(I)$ that can be calculated using the following equation:

$$\phi_k(I) = \sum_{S \subseteq K \setminus \{k\}} \frac{|S|!(m - |S| - 1)!}{m!} (I(S \cup \{k\}) - I(S)) \tag{3.26}$$

where S is the subset of K without k, $|S|$ is the number of factors in S and ! represents factorial. In this equation, $I(S \cup k) - I(S)$ is the marginal contribution of the factor k to total inequality and $\phi_k(I)$ can be interpreted as the average marginal contribution of all possible permutations in which factor k affects inequality jointly with other factors in the set S.

For a factor in set S, we used the observed value of this factor; otherwise, we took the average of the observed value so that this factor has no effect on inequality.

Suppose we have k circumstances variables; the OLS model and the Heckman model would have $k+1$ factors (k circumstances and 1 effort variable) for the Shapley decomposition. We used the counterfactual income distribution Y_e as the effort variable. In Equation (3.26), Y_e is used when effort is in the set S and the mean of Y_e is used when effort is not in the set S.

The hurdle model would have $k+2$ factors (includes the factor of whether income is equal to zero or not). We used the predicted probabilities from the selection equation as the factor of whether an individual earns zero or positive income. If this factor is excluded in S, we used the average probabilities.

The number of factors would not increase when considering heteroskedasticity because we combined the contributions of direct and indirect effect of each circumstance. The difference between this model and the homoskedastic model is that this model uses the estimator from the heteroskedastic model. Both mean and variance change when a factor changes from observed values to fixed values.

One advantage of the Shapley decomposition is that the sum of the Shapley value of each factor is the total contribution of these factors to income inequality. In our study, Gini coefficients are used and the effort affected by the residuals of the model is taken into account. Therefore, the sum of the Shapley value for each factor measured by Gini coefficients is equal to total income inequality:

$$I(y) = \sum_k \phi_k(I) \tag{3.27}$$

3.4.4 The Oaxaca Decomposition

If the population is divided into two groups (e.g. female and male, urban and rural, minority and majority group or under-developed and developed region), circumstances may have different effects on income for each group. To study the group differences, we employed the Oaxaca decomposition (Oaxaca, 1973).

Considering two groups, A and B, income distribution for each group is denoted as Y_A and Y_B, and the difference between the means of the groups is:

$$R = E(Y_A) - E(Y_B) \tag{3.28}$$

where E() is the expected value of income distribution.

We decomposed the between-group difference to three components (Jann et al., 2008):

$$R = EN + CO + INT \tag{3.29}$$

where EN is the "endowments effect", CO is the contribution of differences in the coefficients and INT is the interaction effect of the former two.

In our model, we assumed that only circumstances can be observed. We specified the following model:

$$\ln Y = \mathbf{c}'\boldsymbol{\beta} + \epsilon \tag{3.30}$$

where $\boldsymbol{\beta}$ is the vector of coefficients and ϵ is the error term.

Using this model, the difference between the means of the groups R becomes

$$R = E(\mathbf{c}_A)'\boldsymbol{\beta}_A - E(\mathbf{c}_B)'\boldsymbol{\beta}_B \tag{3.31}$$

The first component EN,

$$EN = (E(\mathbf{c}_A) - E(\mathbf{c}_B))'\beta_B \tag{3.32}$$

captures the group differences in the predictors, i.e. whether the difference in income between groups is due to the difference in circumstances between groups.

The second component CO,

$$CO = E(\mathbf{c}_B)'(\beta_A - \beta_B) \tag{3.33}$$

is the difference in the contribution of coefficients. The level of contribution indicates the amount of inequality between groups coming from the effect of circumstances.

The third component INT,

$$INT = (E(\mathbf{c}_A) - E(\mathbf{c}_B))'(\beta_A - \beta_B) \tag{3.34}$$

is the interaction accounting for both the differences in endowments and coefficients.

The Oaxaca decomposition can also be applied to any number of groups to examine sources of income inequality (Deutsch and Silber, 2008).

3.5 Data Description

To measure inequality of opportunity in China, we used data from the China Family Panel Studies (CFPS). CFPS is a nationally representative annual longitudinal survey containing not only individual-level data but also household- and community-level data. It has been conducted since 2010 by the Institute of Social Science Survey (ISSS) of Peking University, China. As in 2016, this project has released data from the 2010 baseline survey, the 2011 maintenance survey, and the 2012 follow-up survey. Since the survey conducted in 2011 is a maintenance survey and the sample size is small relative to 2010 and 2012, we do not include the 2011 survey in our research.

CFPS covers 16,000 households with more than 33,000 adults and 8,900 children aged under 15 in 25 provinces/municipalities/autonomous regions in China. It is designed to record changes in the socioeconomic wellbeing of Chinese people, covering a variety of topics such as economic activities, educational attainment, family relationships and dynamics, migration, and physical and mental health.

The original sample sizes in 2010 and 2012 were 33,600 and 35,720 respectively in which 26,393 samples were recorded in both 2010 and 2012. Of these we focused on individuals aged between 21 to 60 because the labour participation rate outside this age range is relatively low. After filtering, the sample size was reduced to 19,736.

3.5.1 Micro-Level Data: Income and Circumstances

Table 3.1 and 3.2 present the summary statistics of variables used in the study. Male respondents make up 47% of the sample. Ethnicity is represented by the dummy variable “minority”. It is equal to 0 if an individual’s ethnicity is the majority group — Han; otherwise, it is equal to 1. The percentage of the minority group is around 8%. In addition, the average age of the respondents is 42.25. 90% of them are married. The percentage of members of the Chinese Communist Party (CCP) in the sample is 6% and 7% in 2010 and 2012 respectively.

We also included the number of siblings as one of the circumstance variables. Becker and Lewis (1974) studied the relationship between the number of children and children’s outcome such as educational attainment and socioeconomic status. An empirical study conducted in China (Li et al., 2008) also found a negative correlation between the family size and child outcome. In the dataset, the average number of siblings is around 3.

Another circumstance variable we used is the region of residence when the respondent was 12 years old. Two dummy variables were generated for the region of residence. One is whether the respondent held a non-agriculture Hukou at 12 years old and the other is whether the respondent lived in coastal provinces at that time. We used these variables because children are unlikely to change these circumstances through their own effort.

Hukou is a system for recording household registration in China. It divides households into an agriculture (rural) and a non-agriculture (urban) Hukou. The former lives in rural areas and is registered as a rural household and the latter lives in urban areas and is registered as an urban household. Due to the difficulty in changing a Hukou status from agriculture to non-agriculture, lots of rural immigrants hold an agriculture Hukou status even though they live in urban areas. These rural Hukou holders have lower educational attainment, are more likely to be unemployed, and are less likely to have employer-provided healthcare benefits (Liu, 2005).

Individuals normally have the same type of Hukou as their parents' before they grow up. In our sample, the percentage of individuals who held an urban Hukou status when they were 12 years old is 15%. We chose Hukou status instead of a place of residence because of the difficulty in changing the Hukou status (Wu and Treiman, 2004).

Coastal provinces are the provinces on the eastern coastline of China. This area is more developed than the inland area. We used a dummy variable to capture whether the respondents lived in the coastal provinces when they were 12 years old (coastal12). The data show that about 43% of the respondents grew up in a coastal province.

Table 3.2 shows respondents' *parents' socioeconomic status (SES) when respondents were 14 years old*[11], including parents' education level, parents' occupation status and parents' political affiliation. For all variables, we used only the higher of the values observed by parents.

Parents' education level is reported in eight levels in CFPS. We merged them into three levels, namely (1) low level: illiterate or semi-literate; (2) middle level: primary and junior high school; and (3) high level: senior high school or above. Of the parents surveyed, 38% and 21% had low and high levels of education, respectively.

Parents' occupation is divided into 8 big categories including 595 specific occupational codes in CFPS. We regrouped them into three levels: (1) low level: agricultural workers and workers in manufacture and transportation sectors; (2) middle level: professionals, clerks, technical staffs and other tertiary sector workers; and (3) high level: including the administrative/management positions, teachers for tertiary education, lawyers and high-rank military officers. On average, 79% and 9% of individuals reported low- and high-level of their parents' occupation, respectively.

[11] Parents' SES were only provided for respondents when they were 14 years old in CFPS

Table 3.1: Summary Statistics (Respondents)

Statistic	N	Mean	St. Dev.	Min	Max
Individual income(2010)	19,736	13,812.55	28,002.96	0.00	980,000.00
Individual income(2012)	19,736	17,297.48	32,215.31	0.00	1,809,000.00
Household income per capita(2010)	18,729	17,549.66	27,087.06	2.89	1,000,000.00
Household income per capita(2012)	19,248	23,470.16	28,641.32	0.45	918,924.00
Male	19,736	0.47	0.50	0	1
Minority	19,696	0.08	0.27	0	1
Age in 2010	19,736	42.25	10.79	21	60
Urban Hukou at age 12	19,625	0.15	0.35	0	1
Live in coastal province at age 12	19,736	0.43	0.50	0	1
Number of sibling	19,736	2.98	1.90	0	14
Married in 2010	19,736	0.90	0.30	0	1
CCP member in 2010	19,736	0.06	0.24	0	1
CCP member in 2012	19,736	0.07	0.25	0	1

[1] Income is the nominal value in Yuan.
[2] Income in 2012 is adjusted for inflation at the provincial level to 2010.
[3] CCP is the Chinese Communist Party.
[4] Household income per capita is shown in individual level.

Table 3.2: Summary Statistics (Respondents' Parents' SES)

Statistic	N	Mean	St. Dev.	Min	Max
Low Occupation	17,309	0.79	0.41	0	1
Mid Occupation	17,309	0.13	0.33	0	1
High Occupation	17,309	0.09	0.28	0	1
CCP member	19,736	0.16	0.37	0	1
Low Education	19,736	0.38	0.49	0	1
Mid Education	19,736	0.40	0.49	0	1
High Education	19,736	0.21	0.41	0	1

[1] All variables are defined when the respondents were 14 years old.
[2] All variables only record the higher value between parents.

A dummy variable indicating whether one of the parents was a member of the China Communist Party (CCP) when respondents' were 14 years old was used as a proxy for parents' political affiliation. 16% of the parents had at least one CCP membership.

Among the variables introduced above, we selected gender, ethnicity, Hukou at age 12, a coastal or inland province at age 12, number of siblings, parents' educational level, parents' occupational level and whether at least one parent was CCP member as circumstances. In total, the respondents were divided into 1331 types [12]. Most types had fewer than 10 samples. Therefore, using a non-parametric method to measure inequality of opportunity will result in a large upward bias.

For the dependent variables, we used the annual individual income because household income does not necessarily reflect how a person's upbringing affects his/her income when household members have different upbringings. Household income, household consumption and individual labour earnings have also been used in other studies (Ferreira and Gignoux, 2008). The annual individual income was 13,813 yuan on average in 2010 and 17,297 yuan in 2012. To construct the income variable, we first computed the labour income by summing up individual wages, awards and allowances provided by employers, income from working out of town and bonuses. Then we matched each individual to his household's business income (including agricultural and non-agricultural business income), property income, transfer income and other income (including gifts). The individual income is equal to labour income plus income from all sources of non-labour income divided by the square root of the family size[13]. The individual income increased by 30% from 2010 to 2012, which is mostly attributable to the increase in the household income per capita.

The household income per capita in 2010 was 17,550 yuan on average and rose to 23,470 yuan by 2012. Both the annual individual income and the household income were adjusted for inflation.

3.5.2 Macro-Level Data at the Regional Level in China

To measure inequality of opportunity at the regional level, we divided the whole dataset into 8 regions. Figure 3.1 shows the eight regions in colours, where the red region represents municipalities. Municipalities/autonomous regions and 25 out of 31 provinces in China are covered in the dataset. Generally, we grouped the regions by geographic proximity and development similarity.

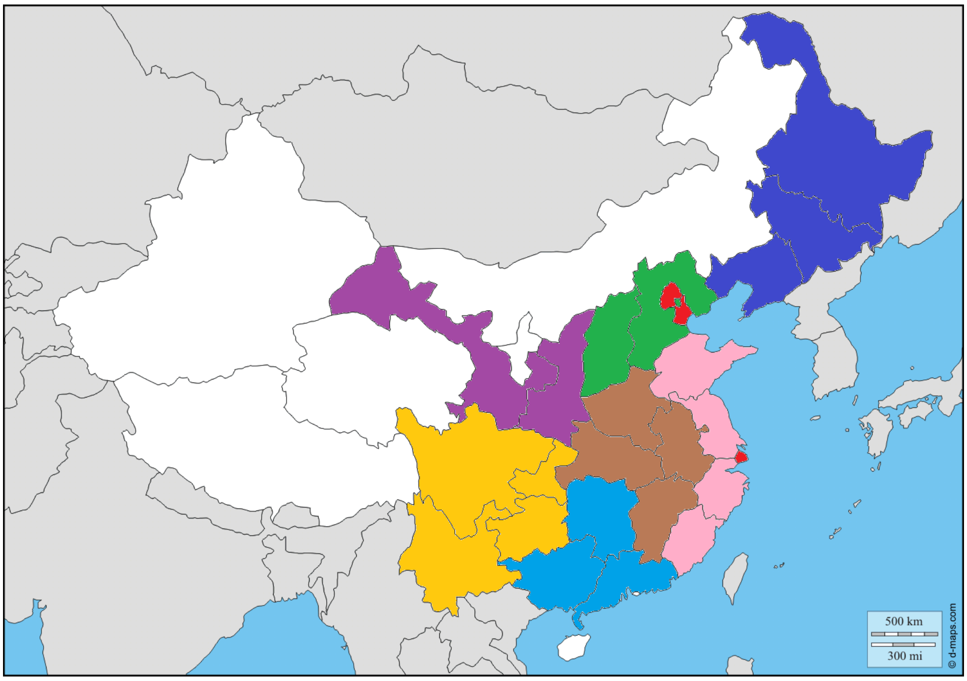
The gross regional product (GRP) and the growth rate per capita for each province are shown in Table 3.3. Generally, the east and metropolitan regions have higher per

[12] Those types with no observation are not counted

[13] Non-labour income and household income are rescaled by the square root to adjust the economies of scale within households. We also constructed income variables without the rescaling. We found that the results in this study are not affected by using income with different scales.

capita GRP compared with the remaining regions; while the west and north west have higher growth rates compared with the others.

Figure 3.1: The Regional Division of China



1. 25 provinces covered in the dataset are coloured in the map.
2. The detailed division for each province is presented in Table 3.3

3.6 The Results

The empirical results are presented in four sub-sections. The first sub-section explores how circumstances affect labour participation. The second sub-section presents inequality of opportunity at the national level; the third sub-section presents the provincial IOP and its relationship with GRP. Inequality of opportunity in regards to gender, ethnicity and Hukou status constitutes the final sub-section.

3.6.1 The Effect of Circumstances on Labour Participation

In our dataset, around 6.8% and 8.1% of the respondents received no income in 2010 and 2012 respectively. This raises the question to what extent do circumstances affect labour participation. One way to show the influence of circumstances is to compare circumstances between positive income and zero income observations. Table 3.4 shows the independent t-test between zero-income and positive income respondents. The coefficients represent the difference between the mean of the zero-income group and the mean of the positive income group. A positive coefficient suggests that a zero-income group is more likely to have individuals with a higher value of the relative variable. In terms of the

Table 3.3: Per Capita Gross Regional Product and Indices

		Per Capita GRP(Yuan)			Indices (preceding year=100)		
Province	Region	2010	2011	2012	2010	2011	2012
Fujian	East	40025	47377	52763	113.2	111.6	110.5
Jiangsu	East	52840	62290	68347	112	110.3	109.7
Shandong	East	41106	47335	51768	111.3	109.9	109.2
Zhejiang	East	51711	59249	63374	109.5	107.2	107.7
Tianjin	Metropolitan	72994	85213	93173	111.7	110.9	109.2
Shanghai	Metropolitan	76074	82560	85373	106.4	105	105.7
Beijing	Metropolitan	73856	81658	87475	104.8	103.8	104.9
Shanxi	Mid-North	26283	31357	33628	111.2	110.4	109.6
Hebei	Mid-North	28668	33969	36584	110.6	109.7	108.9
Anhui	Mid-South	20888	25659	28792	118.8	112.6	111.8
Hubei	Mid-South	27906	34197	38572	114.7	113.5	110.7
Jiangxi	Mid-South	21253	26150	28800	113.2	111.8	110.4
Henan	Mid-South	24446	28661	31499	112.6	112.5	110.1
Jilin	North	31599	38460	43415	113.6	113.5	111.9
Liaoning	North	42355	50760	56649	113.4	111.6	109.4
Heilongjiang	North	27076	32819	35711	112.6	112.2	110.1
Shaanxi	Northwest	27133	33464	38564	114.4	113.7	112.6
Gansu	Northwest	16113	19595	21978	111.6	112.3	112.2
Guangxi	South	20219	25326	27952	113.9	112	110.4
Hunan	South	24719	29880	33480	112.9	111.2	110.7
Guangdong	South	44736	50807	54095	109.5	108	107.4
Chongqing	West	27596	34500	38914	116.2	115.1	112.4
Sichuan	West	21182	26133	29608	115.7	115.9	112.3
Guizhou	West	13119	16413	19710	114.7	116.1	113.5
Yunnan	West	15752	19265	22195	111.6	112.9	112.3

Source: China Statistical Yearbook NBS (2013) significance, “Male”, “coastal province” and “high parents’ occupation” are significant in both years. “Urban Hukou status”, “number of siblings” and “high parents’ education” are significant in 2010, and “minority”, “mid parents’ education” and “mid parents’ occupation” are significant in 2012. From the sign of the coefficients of these variables, a zero-income individual is more likely to be a female Han (the major ethnic group), living in coastal provinces with an urban-Hukou status, having few siblings, and whose parents have a higher socio-economic status.

Table 3.4: Zero Income Vs Positive Income (the Independent t-test)

	(1) 2010		(2) 2012	
Male	-0.102***	(-7.15)	-0.124***	(-8.78)
Minority group	-0.00720	(-0.94)	-0.0255***	(-3.37)
Coastal province at age 12	0.0909***	(6.38)	0.0922***	(6.56)
Urban Hukou at age 12	0.0685***	(6.83)	0.0158	(1.59)
Mid education(Parents)	-0.000528	(-0.04)	0.0232*	(1.66)
High education(Parents)	0.0255**	(2.32)	0.0113	(1.04)
Mid occupation(Parents)	0.00624	(0.65)	0.0173*	(1.84)
High occupation(Parents)	0.0206**	(2.57)	0.0166**	(2.10)
Member of CCP(Parents)	0.00542	(0.50)	-0.00490	(-0.46)
Number of sibling	-0.257***	(-4.72)	-0.0767	(-1.42)

[1] *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

[2] The standard error is in the parentheses.

[3] The coefficients represent the mean difference between the zero-income group and the positive-income group. The null hypothesis for the independent t-test is the mean (of the percentage of male, for example) for the zero-income group is identical to the mean for the positive-income group.

Including zero income observations, we ran regressions based on the Heckman model and the hurdle model for the data in 2010 and 2012 respectively. The results for 2012 are shown in Table 3.5. Since the results for 2010 are similar to those for 2012, we show the results in Table 3.14 in the Appendix. In Table 3.5, columns (1) and (2) show the results from the hurdle model. Columns (3) and (4) are the results from the Heckman model. Columns (1) and (3) are the results from the selection equations, in which the estimators are presented in the form of marginal effect. Columns (2) and (4) are the results from the level equations.

In the selection equations (columns (1) and (3)), we added two variables —the interaction between male and marital status— as the exclusion restriction of the Heckman model because gender and marital status are likely to affect the reservation wage and the decision on labour participation. We did not include these two variables in the Hurdle model because this model does not require exclusion restrictions.

In general, compared to the Hurdle model, the Heckman model estimates a higher log income distribution. This implies that zero income observations could receive higher market wages than the average income level. This implication is also supported by the negative correlation (ρ in Table 3.5) between the errors of the selection and the level equation and the coefficients of the selection equation as well. The coefficients of the selection equation suggest that those with circumstances such as female, major ethnic group, urban Hukou, coastal provinces and higher parents' SES are more likely to have a higher reservation wage than the market wage. Most of these circumstances, however, contribute to a higher income.

Table 3.5: Hurdle Model vs Heckman Model (2012)

	Hurdle Model		Heckman Model	
	(1)	(2)	(3)	(4)
	Selection	Level	Selection	Level
Male	0.036***	0.595***	0.098***	0.488***
Minority	0.023***	−0.377***	0.248***	-0.442***
Urban Hukou at age 12	-0.001	1.164***	-0.284***	1.125***
Coastal Province at age 12	-0.025***	0.372***	-0.118***	0.443***
Mid education(Parents)	-0.005	0.255***	-0.074***	0.258***
High education(Parents)	-0.002	0.221***	-0.051	0.217***
Mid occupation(Parents)	-0.010	0.252***	-0.076**	0.281***
High occupation(Parents)	-0.017**	0.231***	-0.056	0.295***
Member of CCP(Parents)	0.009	0.103***	0.043	0.073**
Number of sibling	0.001	−0.046***	0.005	-0.052***
Married in 2010			-0.118***	
Constant	11.140***	8.173***	1.442***	8.544***
Observations	17176	15821	17176	
ρ			-0.952***	
σ			1.65***	
λ			-1.57***	

[1] *p<0.1; **p<0.05; ***p<0.01

[2] The coefficients of the selection equations are presented in the form of marginal effect.

The selection equation in the Hurdle model shows a similar result. The coefficients for male, minority, coastal province and high parents' occupational level are significant at the 5% level in 2012. Given their coefficient, we conclude that those who are male, the minority group, living in inland provinces or having lower parents' SES are more likely to earn positive income. This result is consistent with the results from the independent t-test as well as the Heckman model.

In the selection equation, the difference between the Hurdle model and the Heckman model is that the Heckman model indicates a higher marginal effect than the Hurdle model. For example, the Hurdle model suggests that rural and urban Hukou have similar probabilities of receiving zero income, while the Heckman model indicates that urban Hukou holders are 28.4% more likely to have a higher reservation wage than the market wage. Given a negative and significant ρ, we conclude that the OLS model (which is also the level equation in Hurdle model) could be biased and the assumption of the reservation wage gives a better explanation of the decision on the labour participation than the Hurdle model.

In the level equation, different circumstances have different impacts on income. Comparing parents' SES with other variables, we find that gender, ethnicity, whether one lives in a coastal province and Hukou status seem to contribute more to income inequality than parents' SES does. These demographic characteristics affect not only income but also the decision in labour participation. Parents' SES, on the other hand, affects a person's income earned but has less implication in their labour participation.

Among all circumstances, Hukou status has the largest impact on income and the labour participation, which might indicate a substantial rural-urban inequality in China. Urban Hukou holders are less likely to participate in the labour market but they could earn 112.5% more, if they work, than their rural counterparts. Given a similar impact of living in a coastal/inland province[14], we conclude that regional disparities can be the most important factor of income inequality in China. This finding is in line with other researches in regional inequality in China such as researches from Wan et al. (2006), Xie and Zhou (2014) and Wu and Rao (2017).

3.6.2 Inequality of Opportunity at the National Level

Considering the correlation between circumstances and effort, we parameterized heteroskedasticity and estimated the model using MLE with equation (3.13). Table 3.6 shows the results. The first two columns are the estimators of the mean and the last two are of the variance. Comparing this table with Table 3.5, the coefficients of the mean are similar to the level estimation of the hurdle model. Therefore, when computing the Shapley values, the results might be similar if replacing the coefficients of the mean from OLS with those from MLE.

As is shown in the last two columns in Table 3.6, heteroskedasticity has a significant effect on income inequality through a variety of factors. Male's income has a higher variation than female's in 2010 (8.8%) but not in 2012 (-0.3%). This difference in income variance might be due to a larger increase in income for low income males than high income males from 2010 to 2012.

Hukou status also has a large impact on income distribution. Those with a urban Hukou have 58.8% lower income variance in 2012, but are not different from their rural Hukou counterparts in 2010; while income differential is much higher in 2012 (117.2%) than 2010 (62.6%). This indicates that income grows much faster for those who hold urban Hukou status than those with rural Hukou status. This finding is also supported by other studies such as Liu (2005) and Afridi et al. (2015). For urban Hukou holders,

[14] For the convenience in computing IOR, we divided the 25 provinces in China into only two categories: inland and coastal. Therefore, the coefficient of inland/coastal dummy variable only roughly shows the regional disparity.

those with low income might benefit more than their high income counterparts so that the income distribution converges and the variance reduces by 58.8%.

In addition, living in the coastal province reduces the income variances from 17.8% in 2010 to 7.7% in 2012, while the income differential between inland and coastal province increases from 27.7% to 39.1%. This indicates that for residents in the coastal province, those with low income might benefit more than their high income counterparts.

In summary, from the heteroskedastic model, we find that variances are influenced mainly by gender, ethnicity, Hukou status and whether one lives in the coastal province. Those who are male, urban Hukou status and living in the coastal province with lower income are more likely to have higher growth of income from 2010 to 2012.

Table 3.6: The MLE with Type Heteroskedasticity at the National Level

	Mean		Variance	
	2010	2012	2010	2012
Constant	8.287^{***}	8.191^{***}	0.611^{***}	0.901^{***}
	(0.032)	(0.035)	(0.033)	(0.033)
Male	0.556^{***}	0.550^{***}	-0.088^{***}	0.003
	(0.022)	(0.024)	(0.023)	(0.023)
Minority	-0.196^{***}	-0.362^{***}	-0.007	0.088^{**}
	(0.040)	(0.046)	(0.042)	(0.042)
Urban Hukou at age 12	0.626^{***}	1.172^{***}	-0.054	-0.588^{***}
	(0.034)	(0.031)	(0.035)	(0.035)
Coastal Province at age 12	0.277^{***}	0.391^{***}	0.178^{***}	0.077^{***}
	(0.023)	(0.024)	(0.023)	(0.023)
Mid education(Parents)	0.199^{***}	0.232^{***}	0.037	0.014
	(0.025)	(0.027)	(0.026)	(0.026)
High education(Parents)	0.221^{***}	0.214^{***}	0.026	-0.014
	(0.034)	(0.035)	(0.035)	(0.034)
Mid occupation(Parents)	0.137^{***}	0.235^{***}	0.026	-0.052
	(0.036)	(0.036)	(0.037)	(0.037)
High occupation(Parents)	0.129^{***}	0.216^{***}	0.087^{**}	0.070
	(0.044)	(0.045)	(0.044)	(0.045)
Member of CCP(Parents)	0.132^{***}	0.099^{***}	-0.122^{***}	-0.054^{*}
	(0.030)	(0.032)	(0.032)	(0.032)
Number of sibling	-0.033^{***}	-0.044^{***}	-0.003	-0.019^{***}
	(0.006)	(0.006)	(0.006)	(0.006)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

The standard error is in the parentheses.

To implement the Shapley decomposition, we grouped 10 explanatory variables into 5 factors — gender, ethnicity, geographic characteristics, parents' SES and the number of siblings. Geographic characteristics comprise Hukou status and coastal province; and parents' socioeconomic status includes parents' educational level, occupational level and member of CCP.

Table 3.7 shows the decomposition of IOR at the national level. The first two columns are the decomposition of the results from the linear regression. Only observations with positive income are included. The third and fourth columns are the decomposition using the Heckman model. The fifth and sixth columns are the decomposition using the whole hurdle model with homoskedasticity. The last two columns are the decomposition using the hurdle model with heteroskedasticity.

Comparing the model without zero income with the first hurdle model, we found that the inclusion of zero income only slightly changed the Shapley value for each factor. In total, IOR decreased over 1% in both years if considering zero-income individuals. This slight decline might indicate that those who have advantages in circumstances might be more likely to receive zero income.

Although the Heckman model and the Hurdle model have different assumptions on labour participation, IOR from both models yields similar results. The results from these two models show the robustness of the Shapley decomposition when zero income observations are considered.

In terms of the contribution of each factor to total income inequality, gender and geographic characteristics are two main sources. They together contribute more than 20% of total income inequality if assuming homoskedasticity. This figure increases to more than 40% when the error is heteroskedastic. Parents' socioeconomic status accounts for around 5% to 7% for homoskedasticity and more than 9% for heteroskedasticity. This effect of parents' socioeconomic status on income might explain low income mobility and low intergenerational mobility. For example, researchers find that as the Chinese economy has developed rapidly in the last two decades, it became more difficult for those in the bottom to climb the economic ladder but much easier for those on the top to stay there (Chen and Cowell, 2015). This decrease in mobility might be related to the bigger impact of parents' socioeconomic status on children's income. Researchers find that the similarities in income, education and health between parents and children have increased in the last two decades (See Qin et al., 2016 and Eriksson et al., 2014).

Sibling number and ethnicity makes up less than 5% of total income inequality for homoskedasticity and around 5% to 8% for heteroskedasticity. Considering all factors, we found that IORs were 29.35% in 2010 and 40.15% in 2012 for the linear regression model. When we accounted for the samples with zero income, IORs reduced to 27.59% in 2010 and 37.97% in 2012. However, when we allowed for heteroskedasticity, IORs increased to around 50% in both years. These measures identify the lower bound of inequality of opportunity in China considering that our models do not capture some unobserved circumstances.

The difference in IORs between homoskedasticity and heteroskedasticity implies that circumstances also largely affect income inequality indirectly through effort. Among these differences in IORs, we found that around 4-5% of income inequality was due to the indirect effect of gender, about 8% to 12% was due to geographic characteristics and about 5% was due to parents' SES. Comparing the results over two periods, we found that the difference in IOR arose mainly from the geographic characteristics. The gap shrinks when heteroskedasticity is taken into account, which is in line with what we found in the results of MLE.

Table 3.7: The Shapley decomposition at the National Level

	OLS		Heckman		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012	2010	2012
Gender	9.34	8.71	7.35	8.34	8.95	8.39	12.41	12.57
Ethnicity	0.74	1.28	0.68	1.21	0.68	1.16	1.12	1.39
Geographic	11.27	20.33	12.54	19.02	10.47	19.02	21.61	23.99
Parents' SES	5.77	6.92	5.26	6.48	5.34	6.4	9.69	10.13
Sibling_number	2.23	2.89	2.76	2.81	1.96	2.61	3.82	5.93
Income: +/0					0.19	0.39	0.16	0.38
IOE	70.65	59.85	71.41	62.15	72.41	62.03	51.19	45.62
IOR	29.35	40.15	28.59	37.85	27.59	37.97	48.81	54.38

[1] OLS is the regression without zero-income. Heckman is the Heckman model. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.

[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.

[3] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.

[4] Income: +/0 is the contribution of probability to have a positive income.

[5] All values are presented in percentages.

[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

We also conducted sensitive analysis for the IOR measures shown in Table 3.7. The results are presented in the Appendix (Tables 3.15, 3.16, and 3.17). In the sensitive analysis, the first test changed the sibling number into a three-level variable: with no sibling, 1 sibling, and with 2 or more siblings. The number of types was reduced to 585. The second test dropped the types with less than 5 samples. The number of types was dropped to 534. The last test dropped the types with less than 10 samples. The number of types was further reduced to 296. The measures in the first two tests (Tables 3.15 and 3.16) are similar to our main results (Table 3.7). The measures slightly decreased when types with fewer than 10 samples were dropped. These results from the sensitivity analysis show that our main results in Table 3.7 are robust and not affected by the reduction of samples and types.

Zhang and Eriksson (2010) measured IOR in nine provinces in China from 1989 to 2006 excluding individuals with no income. Their results in IOR ranged from 46% in 1989 to 63% in 2006. Their paper uses Gini coefficients without the Shapley decomposition and treated the predicted income from the linear regression as inequality of opportunity. Since Gini coefficients are not path independent, the results could be biased.

Given that circumstances $\mathbf{c}$ is observable and effort e is unobserved, the relationship between income inequality and circumstances can be modelled by $y = \mathbf{c}\boldsymbol{\beta} + \epsilon$ where ϵ is the residual. Let the predicted income be $\hat{y} = \mathbf{c}\hat{\boldsymbol{\beta}}$, $IOR = I(\hat{y})/I(y)$[15]. Applying the same method as Zhang and Eriksson (2010), we found that IOR is 37.18% in 2010 and 53.64% in 2012. The results are higher than those from our method. Therefore, the method using Gini coefficients without Shapley decomposition could overstate IOR.

We further implemented Shapley decomposition on IOR computed from predicted income to identify the contribution of each factor. The results are shown in Table 3.8. The first two columns are the results using the Gini coefficients with Shapley decomposition only on circumstances (the same as Zhang's and Eriksson's (2010) method) and the last two are the results using the Gini coefficient with Shapley decomposition on both circumstances and effort. It shows that almost all factors record higher contribution under the former method.

Table 3.8: The Shapley Decomposition (Including or Excluding Effort)

	Excluding Effort		Including Effort	
	2010	2012	2010	2012
Gender	13.97	11.24	9.34	8.71
Ethnicity	0.98	1.75	0.74	1.28
Geographic	14.07	28.98	11.27	20.33
Parents' SES	6.36	8.10	5.77	6.92
Sibling Number	1.8	3.58	2.23	2.89
IOR	37.18	53.64	29.35	40.15

[1] In the first two columns, Shapley decomposition is used to decompose the counterfactual income distribution determined only by circumstances; while in the last two columns, Shapley decomposition is implemented to the observed income assumed to be influenced by both circumstances and effort.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] All values are presented in percentages, representing the contribution of the relative factor to total income inequality.
[5] IOR represents the proportion of income inequality due to circumstances.

Given the same method, IOR reduces from 63% in 2006 as reported in Zhang and Eriksson (2010) to around 50% in 2012 as reported in Table 3.8. In addition, Zhang and Eriksson (2010) estimated the contributions of each circumstance to total income. They found that parents' socioeconomic status is the most important factor in circumstances while our analysis shows that the most important factors are geographic characteristics including whether living in a coastal province and Hukou status.

Comparing this study to studies in other countries, we found that IOR in China is higher. If we only consider individuals with positive incomes and assume homoskedasticity across type, IOR in China measured by a direct ex-ante approach (29.35% in 2010 and 40.15% in 2012) is higher than most of the Latin American countries (20.8% - 37.3%

[15] The same method can also be found in Manna and Regoli (2012); if using variances as the inequality index, $I(\hat{y})$ is equal to the coefficient of determination R^2 (Israeli, 2007).

from Barros et al.'s (2009) study) and U.S. (around 20% in 2001 (Pistolesi, 2009)) with the same approach.

These differences may be due to different empirical methods and data used in these studies. For example,Pistolesi (2009) use ethnicity, birth region, parents education and father's occupation as circumstances variables. Paes de Barros et al. (2009) use gender, race, parent's occupation, parent's education and birth region to identify circumstances. Our study includes not only common variables like ethnicity, parent's education and occupation but also variables that are particularly relevant in the case of China such as the Hukou status, an indicator of residency in a coastal province, an indicator of political affiliations, and number of siblings. To some extent, the Hukou status and residency in a coastal province are similar measures compared to birth region in other studies. We found these variables contribute up to more that 1% of income inequality, which is much higher than the contribution of birth region for other countries.

Using Shapley decomposition and accounting for heteroskedasticity, Björklund et al. (2012) reported IOR to be greater than 30% in Sweden while we found that IOR is around 50% in China. Björklund et al. (2012) use parents' income, parents' education, IQ, number of siblings, BMI and family structure as circumstances. They found that the largest contribution of circumstances to inequality comes from IQ while our results show that it comes from the Hukou status. If IQ also has a significant contribution to income inequality in China, the IOR could be even higher.

The higher IOR in China might explain why Chinese respondents have the lowest "feeling of procedural justice" in the ISSP survey. However, international comparison should be conducted with caution because different studies use different methods.

3.6.3 Inequality of Opportunity at the Regional Level

Tables 3.9 and 3.10 show the measures of inequality of opportunity at the regional level using the hurdle model assuming homoskedasticity across types in 2010 and 2012 respectively (the results of regressions are presented in the Appendix from Tables 3.18 to 3.25). We did not include the heteroskedastic model because, at the regional level, most coefficients of the mean and variance from MLE were not significant. Since some regions contain all coastal provinces and some contain all inland provinces, we removed the "coastal province" dummy in the regressions.

In general, IOR varies from 19.77% to 28.36% in 2010 and from 26.55% to 33.91% in 2012 across regions. These figures are smaller than those at the national level. It is probably because regional disparity contributes to IOP at the national level. In particular, IOR is the highest in the mid-north region in 2010 and in the south region in 2012, while it is the lowest in the mid-south region in 2010 and in the west region in 2012. The differences between the highest and the lowest are around 9% in 2010 and 7.5% in 2012, which indicates a regional disparity in inequality of opportunity in China.

In terms of the Shapley decomposition, gender, Hukou and parents' socioeconomic status are three main sources of income inequality for all regions. This result is in line with that at the national level. However, the contributions of these three sources vary across regions. In the metropolitan region, gender accounts for 8.42% in 2010 and 4.36% in 2012 of total income inequality; while in the mid-north, it represents 13.28% in 2010 and 14.41% in 2012 of total income inequality. The difference between these two regions is more than 5%.

The contributions of Hukou and parents' SES also show huge difference across regions. Hukou status contributes approximately 10.61% of income inequality in 2010 in the northern west and around 11.33% in 2012; while it contributes only 1.02% to income inequality in the east in 2010. This negligible contribution might be due to the smaller rural-urban income gap in the east. In our dataset, the average income in the rural east region is 14,470 yuan and the average urban income is 20,230 yuan; while in the northern west the rural samples earn 7,344 yuan and the urban samples earn 21,800 on average. In terms of parents' SES, the south region is the highest for both years. It accounts for more than 9% in both periods. The lowest contribution is in the north (2.75% in 2010 and 4.16% in 2012).

To conclude, we find that regional disparities exist not only in income inequality but also in its sources. Rich regions like the metropolitan region have a lower level of income inequality but higher IOR; while poor regions have a higher level of income inequality but lower IOR. Specifically, gender, Hukou and parents' socioeconomic status are the three main sources of income inequality. Their contribution, measured by Shapley decompositions, varies from region to region, which indicates large regional heterogeneity in each source of income inequality.

3.6.4 Provincial Inequality and GRP per capita

We also estimated inequality of opportunity at the provincial level for 25 provinces in the dataset. The IORs at the provincial level are computed using the Shapley decomposition with the hurdle model. We assumed type homoskedasticity since the sample size for each province is too small to cover all types.

Figure 3.2 demonstrates the relationship between GRP per capita and inequality of opportunity measured by IOL and IOR at the provincial level. The upper two graphs show GRP per capita with the observed Gini coefficients and IOR respectively. Provinces with higher GRP per capita clearly have lower Gini coefficients but higher IOR. To interpret this difference, we graphed the level of inequality contributed by effort (EOL) alongside the level of inequality contributed by circumstances (IOL) in the lower panel. The graph shows that EOL has a negative relationship with GRP, dropping from around 50% when

Table 3.9: Inequality of Opportunity at the Regional Level (2010)

	Metropolitan	Mid-North	North	East	Mid-South	South	West	Northern West
GRP per Capita	74308.00	27475.50	33676.67	46420.50	23623.25	29891.33	20160.75	20126.00
Observed Gini	53.91	59.98	59.13	60.16	59.91	68.62	57.62	60.08
Gender	8.42	13.28	12.05	11.82	9.25	7.97	8.26	9.65
Ethnicity	0.05	3.17	0.67	0.05	0.63	0.68	1.72	0.43
Hukou	5.71	3.47	1.92	1.02	2.16	6.55	5.28	10.61
Parents' SES	9.76	6.46	2.75	10.34	4.55	10.07	4.84	3.87
Sibling_number	3.21	0.77	4.04	2.73	2.64	0.07	0.72	0.85
Income: +/0	0.65	1.21	0.59	0.66	0.53	1.11	0.51	-0.07
IOE	72.20	71.64	77.98	73.38	80.23	73.55	78.68	74.66
IOR	27.80	28.36	22.02	26.62	19.77	26.45	21.32	25.34

[1] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[2] Income: +/0 is the contribution of probability to have a positive income.
[3] All values are presented in percentages.
[4] GRP denotes the Gross Regional Product.
[5] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.
[6] Observed Gini is computed using data in the sample.

Table 3.10: Inequality of Opportunity at the Regional Level(2012)

	Metropolitan	Mid-North	North	East	Mid-South	South	West	Northern West
GRP per capita	88673.67	35106.00	45258.33	59063.00	31915.75	38509.00	28173.75	29137.00
Observed Gini	49.01	63.88	57.72	63.15	59.68	65.40	64.74	64.38
Gender	4.36	14.41	8.59	10.71	11.53	6.65	8.89	11.33
Ethnicity	0.02	1.63	0.29	0.13	0.46	0.03	2.71	0.83
Hukou	10.79	5.73	15.23	4.75	8.31	13.22	7.91	13.68
Parents' SES	8.29	9.46	4.16	7.02	7.61	9.49	5.67	5.38
Sibling_number	4.34	0.21	1.88	4.11	3.46	3.15	1.38	0.07
Income: +/0	2.41	0.85	0.97	0.73	0.18	1.37	-0.02	0.67
IOE	69.78	67.71	68.89	72.54	68.45	66.09	73.45	68.03
IOR	30.22	32.29	31.11	27.46	31.55	33.91	26.55	31.97

[1] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.

[2] Income: +/0 is the contribution of probability to have a positive income.

[3] All values except GRP and Gini are presented in percentages.

[4] GRP denotes the Gross Regional Product.

[5] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

[6] Observed Gini is computed using data in the sample.

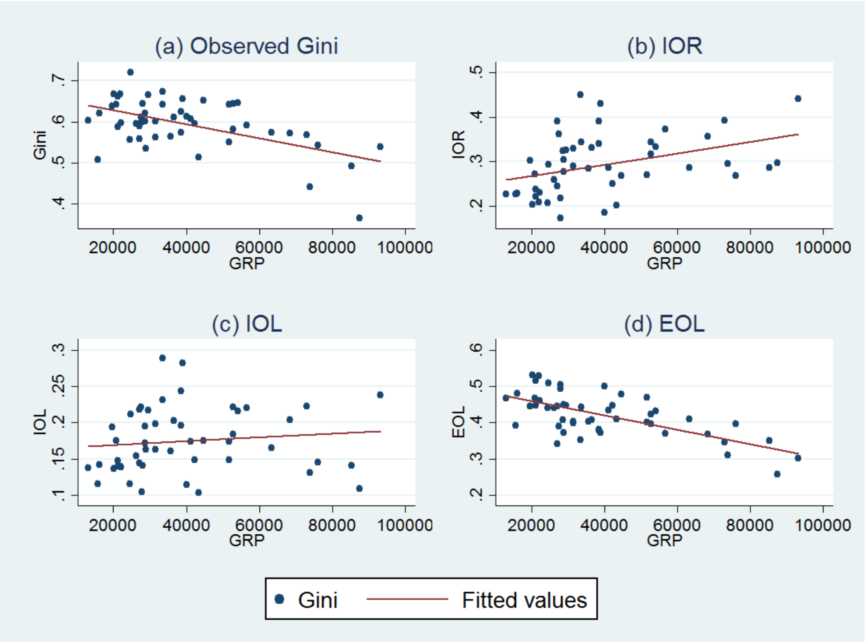
GRP per capita is lower than 20,000 Yuan to below 30% when GRP per capita is close to 100,000 Yuan if assuming type homoskedasticity, while IOL only increases slightly (from 20% to 22%).

To sum up, EOL clearly shows a decreasing trend; while IOL does not show a clear trend. This finding is in contrast to Marrero and Rodriguez (2013), who found that inequality of opportunity is negatively related to growth and inequality of effort is positively related.

The difference in the results might be due to the fact that our research focussed on a developing country— China, while Marrero and Rodriguez (2013) studied a developed country— the United States. At the early stages of development, an increase in effort (e.g. decision for rural residents to move to urban areas to find jobs) might make a huge difference in income, while at the later stage of development, the same amount of increase in effort might make no difference (e.g. for rural migrants, urban jobs are not as easy to find as 20 years ago.).

In summary, the results indicate that income inequality reduces from about 0.7 to 0.5 when GRP per capita rises from below 20,000 yuan to more than 90,000 Yuan. This reduction seems mostly due to the decrease in EOL, which might imply that a poor province has a more diverse distribution of effort or a bigger influence of effort on income inequality.

Figure 3.2: Provincial Inequality and GRP per capita



1. GRP is the Gross Regional Product per capita.
2. IOR is the proportion of income inequality due to circumstances.
3. IOL stands for the level of income inequality due to circumstances.
4. EOL represents the level of income inequality due to effort.
5. Source: GRP is collected from the China statistical yearbook (NBS, 2013).
6. Observed Gini, IOL, EOL and IOR are based on authors' calculation.

3.6.5 The Role of Circumstances in Income Differential between Gender, Ethnicity and Hukou Status

This section shows the role of circumstances in income differential between gender, ethnicity and Hukou status respectively. We used Hukou status in one's childhood because it is determined by one's parents' Hukou status and beyond the individual's control.

Firstly, we grouped the dataset by gender, ethnicity and Hukou status. The means and standard deviations are listed in Table 3.11. Male's individual income is around 60% more than female's, while male's household income per capita is almost the same as female's. Therefore, using household income per capita might fail to reveal inequality of opportunity in gender.

Household income of the majority group is 27% more than that of the minority group in 2010, and a similar gap is observed for individual income. This gap enlarges in 2012 to around 45%, which might imply that minority group benefits little from the growth during 2010-2012.

Among gender, ethnicity and Hukou status, the income gap is the biggest for Hukou status. Individual income of urban Hukou holders is around 93% more than their rural counterparts in 2010. This gap increases to 156% in 2012. A similar gap is observed for household income, which indicates that family members in most households have the same Hukou status as each other.

Table 3.11: Income difference in gender, ethnicity and Hukou status

	(1)		(2)		(3)	
	Female	Male	Majority	Minority	Rural	Urban
HHincome(2010)	17614.7	17584.6	17898.7	14108.9	15367.7	30357.1
	(26598.9)	(27818.2)	(27470.1)	(23498.2)	(23792.5)	(39166.1)
HHincome(2012)	21294.3	21992.0	22175.8	15294.3	19280.7	35031.5
	(25087.9)	(28270.2)	(27222.3)	(17834.9)	(23905.1)	(35873.9)
INDincome(2010)	10755.7	17545.1	14289.3	10321.8	12278.1	23743.4
	(23338.6)	(32100.0)	(28438.6)	(22907.3)	(24838.9)	(40561.5)
INDincome(2012)	12483.2	20302.9	16644.3	10873.1	13117.5	33615.1
	(20274.7)	(37968.9)	(31066.8)	(17552.3)	(26298.8)	(42689.3)

[1] HHincome is the household income per capita.
[2] INDincome is the individual income per capita.
[3] "Majority" and "Minority" indicate ethnicity.
[4] "Rural" and "Urban" indicate Hukou status.
[4] The values in parentheses are the standard deviations.
[5] Source: CFPS and authors' calculation.

To illustrate how circumstances contribute to the income differential shown in Table 3.11, we use Oaxaca decomposition and the results in 2012 are shown in Table 3.12. In the table, "Advantage" represents the predicted income for the advantaged groups; while "Disadvantage" represents the predicted income for the disadvantaged groups. According to Table 3.11, we identify the group of male, urban Hukou holders and the major ethnic group (Han) as the advantage groups. We treat zero income observations as if they earn 1 yuan, transform income into log income and regress log income with other circumstances. Based on the predicted income from the regressions, we decomposed the expected income differential between groups into three components: endowments, coefficients and interaction.

We found that almost all the income differential between gender comes from coefficients. Given the same circumstances, a male earns 83.6% more income than a female in 2012. However, the effort of endowments are close to 0 and not significant, which means a male has no advantage in circumstances.

Income differential caused by circumstances between urban and rural Hukou holders are the largest among the three demographic variables. Urban Hukou holders earn 137.5% more income on average than their rural counterparts. Around 39% of this gap can be explained by the difference in circumstances and 85% is related to the effect of circumstances. In other words, rural Hukou holders could not only have disadvantage in circumstances (lower parents' SES, for example) but also earn less if they have the same circumstance as their urban counterparts. This finding is consistent with Afridi et al. (2015)'s experimental study on the impact of China's Hukou system on individuals' performance. They found that a rural Hukou reduced the academic performance of rural migrant students compared to their local urban counterparts.

Income gap caused by circumstances can also be found between different ethnic groups. Individuals in the major ethnic group earn 39.7% more income on average than other minorities. Around 38.5% of this difference in income is due to the difference in circumstances, while around 77.6% is due to the effect of circumstances. This indicates that individuals in the major ethnic group have advantage in circumstances and earn more income even though they have the same circumstances as minorities.

Since the results in Table 3.12 include zero income observations, one source of income differential could be from those zero income observations. In Table 3.13, we excluded zero income observations and did the Oaxaca Decomposition again. We found that the income gap becomes smaller (reduced by 20%) for gender but becomes larger for ethnicity (18% increase) and Hukou status (20% increase). This result indicates that female, urban Hukou holders and majority groups are more likely to choose not to work, which is consistent with what we found in the analysis of labour participation.

We implemented the Oaxaca decomposition using the data in 2010. The results are shown in Table 3.26 in the Appendix. Results are generally similar to those for 2012. The only difference is that in 2010, income differential between rural and urban Hukou holders was much lower than that in 2012.

In summary, we find that a male earns a higher income than a female given the same circumstances but has no advantage in other circumstances. This gap is partially due to the lower labour participation of females and partially due to the discrimination of

Table 3.12: Oaxaca Decomposition by Gender, Hukou Status and Ethnicity(2012)

	(1) Gender	(2) Hukou	(3) Ethnicity
Disadvantage	7.772***	8.020***	7.801***
	(0.0327)	(0.0240)	(0.0733)
Advantage	8.618***	9.395***	8.198***
	(0.0318)	(0.0763)	(0.0244)
Difference	-0.846***	-1.375***	-0.397***
	(0.0456)	(0.0799)	(0.0773)
Endowments	-0.0119	-0.538***	-0.153***
	(0.00884)	(0.0991)	(0.0212)
Coefficients	-0.836***	-1.174***	-0.308***
	(0.0451)	(0.108)	(0.0827)
Interaction	0.00185	0.337**	0.0640
	(0.00743)	(0.123)	(0.0392)
Observations	13690	13690	13690

[1] Advantage is the predicted income when the dummy variable listed in column is equal to 1.
[2] Disadvantage is the predicted income when the dummy variable listed in column is equal to 0.
[3] Standard errors in parentheses
[4] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

Table 3.13: Oaxaca Decomposition by Gender, Hukou Status and Ethnicity (2012 without Zero Income Observations)

	(1) Gender	(2) Hukou	(3) Ethnicity
Disadvantage	8.466***	8.597***	8.219***
	(0.0199)	(0.0151)	(0.0520)
Advantage	9.088***	10.15***	8.810***
	(0.0204)	(0.0314)	(0.0150)
Difference	-0.621***	-1.556***	-0.591***
	(0.0285)	(0.0349)	(0.0541)
Endowments	-0.0141	-0.319***	-0.213***
	(0.00964)	(0.0417)	(0.0195)
Coefficients	-0.608***	-1.042***	-0.425***
	(0.0268)	(0.0584)	(0.0568)
Interaction	0.000946	-0.195**	0.0464*
	(0.00511)	(0.0629)	(0.0270)
Observations	12760	12760	12760

[1] Advantage is the predicted income when the dummy variable listed in column is equal to 1.
[2] Disadvantage is the predicted income when the dummy variable listed in column is equal to 0.
[3] Standard errors in parentheses
[4] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

females in the labour market. However, for ethnicity and Hukou status, the disadvantageous groups have even higher job participation, which indicates higher inequality of opportunity between ethnic groups or different Hukou status for employed people than is shown in Table 3.12.

3.7 Conclusion

In this essay, we used the data from the CFPS and computed IOR in 2010 and 2012 respectively. Taking advantage of the heterogeneity of regional development in China, we grouped 25 provinces into 8 regions. At the national level, we found that IOR was around 30% in 2010 and 40% in 2012 if assuming homoskedasticity across types. This finding suggests that inequality of opportunity in China is at least 30% of income inequality in 2010 and 40% in 2012 given the observed circumstances in our dataset. Higher than the findings for most Latin American countries and U.S., this figure indicates a large proportion of income inequality is due to circumstances beyond individuals' control. IOR increases to around 50% when heteroskedasticity is accounted for. This increase might suggest the effect of circumstances on effort and an underestimated measure of IOP if heteroskedasticity is neglected.

We also found evidence of relationships between regional development and inequality. GRP, as a proxy for regional development, has a negative relationship with the observed income inequality measured by Gini coefficients but a positive relationship with IOR. More specifically, income inequality due to effort decreases when comparing the rich regions to the poor while income inequality due to circumstances does not show a clear pattern. As a result, the overall observed income inequality decreases with regional development. This result suggests that different regions have a similar level of within-region inequality of opportunity even though income inequality varies across region.

On the one hand, the results shed light on how income inequality is driven by circumstances and effort. On the other hand, the analysis highlights possible bias in the conventional approaches to inequality of opportunity. Application of the Heckman model and the hurdle model were attempts to correct the bias from the exclusion of the observations with zero income. These two models also demonstrate as to how circumstances affect labour participation. Although IORs change little after including zero-income observations, a larger sample size helps improve the robustness and the representativeness of the results.

MLE aims to correct another bias—type heteroskedasticity. Assuming a heteroskedastic model and using MLE, we show the indirect contribution of each circumstance to total income inequality. After taking account of type heteroskedasticity, IORs show a consistency over the two periods than the results without type heteroskedasticity. This consistency also suggests that some circumstances such as Hukou status contributed more indirectly to total income inequality in 2010 than that in 2012.

Applying these econometric techniques, we are able to relax the assumption of independence between circumstances and effort and measure the indirect effect of effort on circumstances through the heteroskedastic model. Other advanced models such as models for panel data can be applied in a future study when the data for additional years become available in CFPS.

We are also aware of the different role circumstances play in income for different cohorts. Noticeably, we find a huge income gap between rural and urban Hukou holders. This income differential between urban and rural Hukou holders is largely due to circumstances. Rural Hukou holders initially have more disadvantages in circumstances than their urban counterparts. They tend to earn much less income even though they have the same circumstances as their urban counterparts except for Hukou status.

These sources of income differentials between rural and urban Hukou holders indicate that living in a rural/urban area in one's childhood could have a long-term effect on one's income. The disadvantage for rural Hukou holders has already existed in the previous generation because their parents have a lower SES on average compared with their urban counterparts. Given that circumstances have a large influence on total income, the disadvantage in lower parents' SES might result in lower income and lower SES in the next generation. For example, Eriksson and Zhang (2012) show a high level of income similarity between siblings in China. Du et al. (2014) report high income similarity between parents and children. These studies show a low intergenerational mobility in China which might help explain high inequality of opportunity found in this study.

In Chapter 5, we investigated further in the effect of Hukou status on educational inequality.

Given the substantial income gap between rural and urban Hukou holders, future studies can explore how Hukou status affects intergenerational mobility and inequality of opportunity in the long run and whether the influence comes from living in a rural/urban area or the Hukou system itself.

3.8 Appendices

Table 3.14: Hurdle Model vs Heckman Model (2010)

	Hurdle Model		Heckman Model	
	(1)	(2)	(3)	(4)
	Selection	Level	Selection	Level
Male	0.030***	0.560***	0.116***	0.481***
Minority	0.001	−0.200***	0.035	-0.200***
Urban Hukou at age 12	-0.034***	-0.253***	-0.253***	0.706***
Coastal Province at age 12	-0.027***	-0.101***	-0.101***	0.337***
Mid education(Parents)	0.004	0.207***	0.028	0.195***
High education(Parents)	-0.001	0.219***	0.022	0.221***
Mid occupation(Parents)	0.008	0.140***	0.056	0.128**
High occupation(Parents)	-0.008	0.143***	0.001	0.172***
Member of CCP(Parents)	0.004	0.128***	-0.001	0.113***
Number of siblings	0.004***	−0.034***	0.012	-0.045***
Married			0.057	
Constant	10.339***	8.286***	1.411***	8.532***
	17176	15860	17176	
			-0.91***	
			1.518***	
			-1.381***	

[1] *p<0.1; **p<0.05; ***p<0.01

[2] The coefficients of the selection equations are presented in marginal effect.

Table 3.15: The Measures of Inequality of Opportunity at the National Level (2-level Sibling Number)

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	9.37	8.81	8.97	8.45	12.72	13.03
Ethnicity	0.75	1.30	0.69	1.18	1.18	1.49
Geographic	11.20	20.44	10.39	19.10	22.03	24.90
Parents' SES	5.99	7.30	5.54	6.73	10.07	10.72
Sibling_number	1.89	2.00	1.68	1.82	1.03	1.81
Income: +/0			0.21	0.46	0.20	0.45
IOE	70.79	60.16	72.52	62.26	52.76	47.60
IOR	29.21	39.84	27.48	37.74	47.24	52.40

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] Income: +/0 is the contribution of probability to have a positive income.
[5] All values are presented in percentage.
[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 3.16: The Measures of Inequality of Opportunity at the National Level (Dropping Types with less than 5 Samples)

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	9.75	9.16	9.34	8.83	12.67	12.88
Ethnicity	0.63	1.23	0.58	1.11	0.94	1.26
Geographic	10.74	18.69	9.99	17.46	20.97	23.32
Parents' SES	5.73	6.89	5.32	6.38	9.67	10.04
Sibling_number	2.35	3.14	2.08	2.84	3.84	6.04
Income: +/0			0.20	0.42	0.16	0.42
IOE	70.81	60.89	72.49	62.95	51.74	46.05
IOR	29.19	39.11	27.51	37.05	48.26	53.95

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] Income: +/0 is the contribution of probability to have a positive income.
[5] All values are presented in percentages.
[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 3.17: The Measures of Inequality of Opportunity at the National Level (Dropping Types with less than 10 Samples)

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	10.18	10.14	9.78	9.78	12.98	13.47
Ethnicity	0.38	0.98	0.34	0.88	0.89	1.16
Geographic	9.78	16.82	9.11	15.68	19.87	21.93
Parents' SES	5.55	6.54	5.17	6.07	9.14	9.26
Sibling_number	2.04	3.10	1.79	2.80	3.89	6.15
Income: +/0			0.27	0.52	0.22	0.52
IOE	72.08	62.41	73.54	64.27	52.99	47.52
IOR	27.92	37.59	26.46	35.73	47.01	52.48

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.

[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.

[3] Parents' SES is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.

[4] Income: +/0 is the contribution of probability to have a positive income.

[5] All values are presented in percentages.

[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 3.18: The Hurdle Model at the Regional Level (Metropolitan)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.574**	0.458***	1.316	0.236***
	(0.205)	(0.064)	(0.201)	(0.060)
Minority	0.870	−0.215	286,301.700	0.038
	(1.073)	(0.377)	(426.746)	(0.337)
Hukou at age 12	1.153	0.308***	2.911***	0.480***
	(0.223)	(0.070)	(0.251)	(0.065)
Mid education(Parents)	1.379	0.188**	0.723	0.310***
	(0.257)	(0.079)	(0.237)	(0.075)
High education(Parents)	0.737	0.505***	0.667	0.321***
	(0.288)	(0.098)	(0.300)	(0.092)
Mid occupation(Parents)	0.963	−0.035	1.214	0.145*
	(0.255)	(0.082)	(0.272)	(0.076)
High occupation(Parents)	0.909	0.068	1.447	0.120
	(0.327)	(0.104)	(0.360)	(0.097)
Member of CCP(Parents)	1.657*	0.162*	1.015	0.028
	(0.291)	(0.084)	(0.277)	(0.078)
Number of siblings	1.102	−0.046**	0.873**	−0.054***
	(0.063)	(0.019)	(0.056)	(0.018)
Constant	7.080***	9.383***	11.928***	9.815***
	(0.276)	(0.092)	(0.275)	(0.086)
Observations	1,484	1,373	1,484	1,370
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.19: The Hurdle Model at the Regional Level (Mid-North)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	2.324***	0.722***	2.318***	0.916***
	(0.198)	(0.067)	(0.206)	(0.075)
Minority	0.876	$-$1.144***	2.970*	$-$0.632***
	(0.348)	(0.140)	(0.596)	(0.152)
Hukou at age 12	0.541	0.815***	0.522*	1.162***
	(0.390)	(0.171)	(0.378)	(0.193)
Mid education(Parents)	0.956	0.366***	1.213	0.332***
	(0.211)	(0.078)	(0.220)	(0.087)
High education(Parents)	1.067	0.403***	1.184	0.588***
	(0.283)	(0.102)	(0.283)	(0.115)
Mid occupation(Parents)	1.692	0.019	0.668	0.347**
	(0.363)	(0.119)	(0.296)	(0.136)
High occupation(Parents)	3.216**	0.056	1.833	0.042
	(0.541)	(0.141)	(0.492)	(0.158)
Member of CCP(Parents)	0.850	0.187**	1.258	0.155
	(0.256)	(0.095)	(0.284)	(0.105)
Number of siblings	1.081	0.021	1.138**	0.007
	(0.055)	(0.020)	(0.058)	(0.022)
Constant	7.364***	7.851***	6.077***	7.666***
	(0.261)	(0.100)	(0.263)	(0.113)
Observations	1,881	1,745	1,881	1,755
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.20: The Hurdle Model at the Regional Level (North)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.450^{***}	0.620^{***}	1.954^{***}	0.477^{***}
	(0.125)	(0.055)	(0.145)	(0.054)
Minority	1.198	0.156^{*}	1.111	-0.079
	(0.230)	(0.095)	(0.246)	(0.094)
Hukou at age 12	0.677^{***}	0.180^{***}	0.973	0.820^{***}
	(0.137)	(0.067)	(0.160)	(0.064)
Mid education(Parents)	0.843	-0.030	0.759^{*}	0.121^{*}
	(0.150)	(0.065)	(0.167)	(0.063)
High education(Parents)	0.868	0.057	0.889	0.034
	(0.192)	(0.086)	(0.217)	(0.084)
Mid occupation(Parents)	1.015	0.161^{*}	0.716^{*}	0.149^{*}
	(0.178)	(0.084)	(0.192)	(0.082)
High occupation(Parents)	0.952	0.123	0.764	0.276^{***}
	(0.212)	(0.101)	(0.244)	(0.098)
Member of CCP(Parents)	0.926	0.084	1.507^{**}	0.032
	(0.160)	(0.075)	(0.196)	(0.073)
Number of siblings	1.072^{**}	-0.060^{***}	1.019	-0.029^{**}
	(0.033)	(0.014)	(0.035)	(0.014)
Constant	6.837^{***}	8.818^{***}	8.656^{***}	8.675^{***}
	(0.185)	(0.082)	(0.204)	(0.080)
Observations	2,675	2,367	2,675	2,433
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.21: The Hurdle Model at the Regional Level (East)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.292	0.657^{***}	1.443^{**}	0.691^{***}
	(0.164)	(0.075)	(0.161)	(0.085)
Minority	695,723.500	-0.193	911,550.300	0.427
	(482.469)	(0.487)	(480.666)	(0.554)
Hukou at age 12	0.597	0.343^{*}	0.533^{*}	1.084^{***}
	(0.338)	(0.186)	(0.328)	(0.213)
Mid education(Parents)	1.304	0.567^{***}	0.785	0.402^{***}
	(0.184)	(0.085)	(0.179)	(0.097)
High education(Parents)	1.007	0.271^{**}	1.002	0.373^{***}
	(0.247)	(0.120)	(0.259)	(0.135)
Mid occupation(Parents)	1.232	0.088	1.249	-0.170
	(0.301)	(0.130)	(0.280)	(0.149)
High occupation(Parents)	1.080	0.333^{**}	1.490	0.082
	(0.307)	(0.142)	(0.315)	(0.161)
Member of CCP(Parents)	1.074	-0.041	0.751	0.226^{*}
	(0.233)	(0.104)	(0.208)	(0.121)
Number of siblings	1.071	-0.042^{**}	0.950	-0.066^{***}
	(0.044)	(0.020)	(0.041)	(0.023)
Constant	5.725^{***}	8.459^{***}	9.438^{***}	8.331^{***}
	(0.208)	(0.101)	(0.214)	(0.114)
Observations	1,704	1,531	1,704	1,520

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.22: The Hurdle Model at the Regional Level (Mid-South)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.662***	0.495***	1.792***	0.660***
	(0.164)	(0.051)	(0.174)	(0.056)
Minority	1.623	-1.164^{***}	0.711	-0.999^{***}
	(0.728)	(0.194)	(0.536)	(0.217)
Hukou at age 12	1.099	0.269***	0.494***	0.820***
	(0.266)	(0.087)	(0.227)	(0.097)
Mid education(Parents)	0.827	0.074	0.742	0.169***
	(0.187)	(0.059)	(0.202)	(0.065)
High education(Parents)	0.768	0.222***	0.814	0.196**
	(0.230)	(0.076)	(0.250)	(0.083)
Mid occupation(Parents)	1.151	0.134	0.952	0.302***
	(0.265)	(0.084)	(0.256)	(0.091)
High occupation(Parents)	0.776	-0.092	0.648	0.204*
	(0.271)	(0.097)	(0.265)	(0.107)
Member of CCP(Parents)	1.000	0.188**	0.960	0.215***
	(0.225)	(0.074)	(0.226)	(0.082)
Number of siblings	1.057	-0.046^{***}	1.044	-0.059^{***}
	(0.048)	(0.015)	(0.051)	(0.016)
Constant	11.820***	8.486***	16.784***	8.341***
	(0.225)	(0.074)	(0.246)	(0.080)
Observations	2,791	2,618	2,791	2,634
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.23: The Hurdle Model at the Regional Level (South)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.729***	0.591***	1.510***	0.520***
	(0.153)	(0.069)	(0.127)	(0.078)
Minority	5.034**	0.346**	1.437	0.020
	(0.720)	(0.157)	(0.342)	(0.179)
Hukou at age 12	0.522***	0.711***	1.175	1.151***
	(0.207)	(0.112)	(0.203)	(0.123)
Mid education(Parents)	1.635***	0.353***	1.569***	0.425***
	(0.170)	(0.080)	(0.142)	(0.090)
High education(Parents)	1.501*	0.205**	1.539**	0.136
	(0.210)	(0.099)	(0.178)	(0.111)
Mid occupation(Parents)	1.555*	0.241**	0.878	0.302**
	(0.267)	(0.110)	(0.205)	(0.124)
High occupation(Parents)	0.616**	0.435***	0.476***	0.554***
	(0.247)	(0.133)	(0.212)	(0.152)
Member of CCP(Parents)	1.332	0.281***	1.654**	0.081
	(0.237)	(0.104)	(0.210)	(0.117)
Number of siblings	1.003	0.003	1.041	−0.058***
	(0.040)	(0.019)	(0.034)	(0.021)
Constant	5.940***	8.045***	3.484***	8.270***
	(0.202)	(0.100)	(0.170)	(0.113)
Observations	2,223	2,013	2,223	1,914
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.24: The Hurdle Model at the Regional Level (West)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.264	0.440***	1.429*	0.548***
	(0.176)	(0.060)	(0.202)	(0.071)
Minority	1.054	−0.133**	1.319	−0.246***
	(0.188)	(0.064)	(0.227)	(0.075)
Hukou at age 12	0.629	0.832***	0.589	1.209***
	(0.341)	(0.144)	(0.349)	(0.172)
Mid education(Parents)	1.414*	0.252***	0.760	0.281***
	(0.205)	(0.068)	(0.225)	(0.080)
High education(Parents)	1.152	0.085	0.743	0.204*
	(0.269)	(0.097)	(0.297)	(0.114)
Mid occupation(Parents)	0.640	−0.051	0.673	0.199
	(0.310)	(0.118)	(0.327)	(0.141)
High occupation(Parents)	0.602	−0.061	0.811	−0.132
	(0.384)	(0.152)	(0.428)	(0.179)
Member of CCP(Parents)	1.473	0.170*	1.129	0.138
	(0.297)	(0.094)	(0.305)	(0.113)
Number of siblings	1.069	0.019	1.135**	−0.028
	(0.047)	(0.016)	(0.056)	(0.019)
Constant	9.014***	8.220***	12.346***	8.076***
	(0.210)	(0.077)	(0.243)	(0.090)
Observations	2,084	1,939	2,084	1,973
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.25: The Hurdle Model at the Regional Level (Northern West)

	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.287	0.487***	2.260***	0.643***
	(0.271)	(0.052)	(0.216)	(0.058)
Minority	0.228***	−0.281*	0.390**	−0.941***
	(0.448)	(0.165)	(0.425)	(0.180)
Hukou at age 12	0.395**	1.097***	1.141	1.551***
	(0.375)	(0.112)	(0.399)	(0.121)
Mid education(Parents)	0.823	0.158***	0.896	0.260***
	(0.301)	(0.059)	(0.226)	(0.065)
High education(Parents)	1.069	0.170*	0.649	0.238**
	(0.457)	(0.089)	(0.305)	(0.099)
Mid occupation(Parents)	1.098	−0.055	0.863	0.080
	(0.495)	(0.107)	(0.359)	(0.119)
High occupation(Parents)	1.092	−0.057	1.209	0.041
	(0.542)	(0.122)	(0.440)	(0.135)
Member of CCP(Parents)	0.658	0.124*	0.724	0.094
	(0.347)	(0.075)	(0.259)	(0.084)
Number of siblings	1.305***	−0.015	0.990	0.002
	(0.084)	(0.014)	(0.053)	(0.016)
Constant	23.022***	8.122***	18.405***	7.887***
	(0.329)	(0.068)	(0.252)	(0.074)
Observations	2,334	2,274	2,334	2,222
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3.26: Oaxaca Decomposition by Gender, Hukou Status and Ethnicity(2010)

	(1) Gender	(2) Hukou	(3) Ethnicity
Differential			
Disadvantage	7.845***	8.133***	7.917***
(0.0317)	(0.0231)	(0.0758)	
Advantage	8.605***	8.753***	8.223***
(0.0309)	(0.0822)	(0.0235)	
Difference	-0.760***	-0.620***	-0.306***
(0.0442)	(0.0854)	(0.0793)	
Decomposition			
Endowments	-0.00358	-0.463***	-0.0859***
(0.00656)	(0.107)	(0.0177)	
Coefficients	-0.762***	-0.474***	-0.258**
(0.0442)	(0.110)	(0.0859)	
Interaction	0.00515	0.316**	0.0379
(0.00696)	(0.128)	(0.0392)	
Observations	13690	13690	13690

[1] Advantage is the predicted income when the dummy variable listed in column is equal to 1.

[2] Disadvantage is the predicted income when the dummy variable listed in column is equal to 0.

[3] Standard errors in parentheses

[4] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

Chapter 4

Essay II: Measuring Inequality of Opportunity in China using A Finite Mixture Model

4.1 Introduction

In the last two decades inequality of opportunity (Cohen, 1989, Arneson, 1989 and Roemer, 1998) has frequently been studied both theoretically and empirically in economics. Roemer (1998) proposed a framework to measure inequality of opportunity. He argued that inequality of outcome (such as income, health and educational equality) due to factors beyond individuals' responsibility ("circumstances") is unfair while inequality due to factors within individuals' responsibility ("effort") is acceptable (Roemer, 1998). Based on Roemer's framework, equal opportunity requires two basic principles: the *compensation principle* and the *reward principle* (Fleurbaey, 2008). The former demands that inequality due to circumstances should be compensated and the latter states that inequality due to effort should be fully respected and rewarded.

Applying these two principles, researchers have measured inequality of opportunity in different countries (see recent surveys such as Roemer and Trannoy, 2015; Peragine and Ferreira, 2015; and Ramos and Van de gaer, 2016). One can easily measure the contribution of circumstances to inequality of outcome if both circumstances and effort can be observed. However, data imperfections restrict researchers from observing all variables regarded as circumstances or effort. Data on circumstances such as gender, ethnicity and family background might be easy to access, but the list of circumstances could never be exhaustive. These unobserved circumstances can lead to a downward bias on measures of inequality of opportunity (Lara Ibarra and Martinez Cruz, 2015).

Effort variables are even more likely to be affected by lack of data. Effort related to individuals' preferences and choices are difficult to observe. With the absence of effort, one can only measure inequality of opportunity before knowing effort, i.e. using the *ex-ante* compensation principle. Measures relying on the information of effort, i.e. the *ex-post* compensation principle, cannot be used (Fleurbaey and Peragine, 2013).

To solve this problem, Roemer (1998) assumed that quantiles of income distribution conditional on circumstances represent the degree of effort. Although this assumption provides an effective way to identify the degree of effort, it could lead to biased results if circumstances and effort are not independent of each other.

In terms of the correlation between circumstances and effort, researchers agree that effort variables could be shaped by circumstances; however, they use different approaches to accounting for this correlation due to their different views in terms of what people should be responsible for. Roemer (1993, 1998) argued that individuals should be responsible for factors within their control (this view is known as *the control view*). Based on his view, preferences and tastes related to family backgrounds should also be considered as circumstances because these factors are out of individual control. If effort can be observed, one can measure the indirect effect of circumstances on income through effort as part of inequality of opportunity (Bourguignon et al., 2007). If effort cannot be observed, one can take account of different variances from the distribution of effort conditional on circumstances (Björklund et al., 2012).

In contrast with Roemer's view, Rawls (1971), Dworkin (1981a,1981b,) and Fleurbaey (2008) considered individual responsibilities as their preferences and choices (known as *the preference view*). Based on this preference view, Barry (2005) argued that preferences and tastes shaped by family backgrounds should be respected. Jusot et al. (2013) compared different views on the correlation between circumstances and effort and found that the correlation makes little difference to the measure of inequality of opportunity in terms of health inequality.

In this study, we propose an alternative model — a latent class model, in which each class corresponds to an unobserved level of effort. It can also be interpreted as a finite mixture model defined as a probability-weighted mixture of distributions. The finite mixture model (FMM) is used to estimate the average income in each level conditional on circumstances. More importantly, this model allows the effect of circumstances on income to be different across levels of effort. This heterogeneous effect of circumstances indicates that the unfair inequality represented by the contribution of circumstances to income inequality could vary with the level of individual effort. Low individual effort such as choosing not to participate in the labour force could substantially decrease the effect of circumstances on income. As a result, neglecting this individual effort could lead to a biased estimate of inequality of opportunity.

However, the latent class in FMM could also correspond to unobserved circumstances (Donni et al., 2015). To address this issue, we proposed and implemented a finite mixture model with varying probabilities (FMMV) —a generalised version of the finite mixture model where the classification is assumed to be determined by either circumstances or effort. A circumstance-determined latent class captures unobserved circumstances or effort correlated with circumstances, while an effort-determined latent class captures unobserved effort that is independent of circumstances. Comparing the results between FMMV and FMM, we can have a better understanding on whether the latent class in FMM corresponds to unobserved circumstances or effort.

In addition, we can obtain not only ex-ante but also ex-post measures to examine inequality of opportunity using FMM. Theoretically, inequality of opportunity can be measured by two different approaches — ex-ante and ex-post approaches: the former approach requiring effort to be observed and the latter allowing effort to be latent. Applying the finite mixture model, we can measure inequality of opportunity using both approaches even without the observation of effort.

Latent class models have been widely applied in both microeconomics and macroeconomics to handle inter-individual or time heterogeneity with different types of data such as time-series (Frühwirth-Schnatter, 2006), cross-section (Deb and Trivedi, 1997) and panel data (Deb and Trivedi, 2013). The advantage of this model is that it specifies data as a mixture of distributions so it has greater flexibility compared with other fully parametric models such as a linear regression model.

In addition, a conventional model naturally assumes that the effects of circumstances on income are homogeneous. However, in reality, this might not be the case. A low-income individual, having deliberately chosen a low level of effort (e.g. short working hours), might not feel the disadvantage from his/her circumstances. A more common complaint on inequality of opportunity could be from one with a high level of effort but low income due to his/her disadvantageous profiles. This raises an important question about whether those with high effort may also be exposed to a higher inequality of opportunity.

This question might be addressed by using a finite mixture model. Since income distribution is parameterized in each component, a counterfactual income distribution can be estimated for each level of effort so that inequality of opportunity can be measured by comparing counterfactuals.

Our results showed a better performance in terms of the information criteria and goodness of fit when using a finite mixture model rather than a conventional linear regression model. We found evidence of heterogeneous effects of circumstances on income across effort levels. In other words, sources of income inequality for each effort level are different. Inequality of opportunity could be significantly underestimated when ignoring this heterogeneous effect. FMM shows results similar to those obtained with FMMV assuming an effort-determined latent class, which implies that the latent class in FMM captures factors associated with individual effort even though effort is assumed to be unobserved. Moreover, through examining the counterfactual income distribution for each level of effort, we found that inequality of opportunity is the highest if everyone exerts a middle level of effort.

The rest of this article is organised into the following sections. In section 4.2 we briefly review the literature on conventional approaches to measuring inequality of opportunity. Section 4.3 introduces model specifications and econometric methodologies. Section 4.4 describes the data used in this study, while section 4.5 presents our results. Section 4.6 draws conclusions and discusses their implications.

4.2 The Conventional Approach to Measuring Inequality of Opportunity

To measure inequality of opportunity, one can assume a population N with two sets of individual characteristics: a vector of *circumstances* $\mathbf{c}$ and a vector of *efforts* $\mathbf{e}$. As a working definition, circumstances and efforts are sets of variables with finite values. In other words, circumstances $\mathbf{c}$ belong to a finite set $\Omega = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ and efforts $\mathbf{e}$ belong to a finite set $\Theta = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$. Most of the literature defines a group of individuals with the same circumstances as a *type* and a group of individuals with the same effort as a *tranche*. Hence, there are n types and m tranches.

Assume that an individual's income y is generated by circumstances, efforts and a random term ϵ (assumed to be independently and identically distributed) by a function $g : \Omega \times \Theta \mapsto \mathbb{R}$. The function $g(\cdot)$ is treated as a linear function in most empirical researches. Given that ϵ is additively separable in the g function, we can express the data generation process in the following equation.

$$y = g(\mathbf{c}, \mathbf{e}) + \epsilon \tag{4.1}$$

where $E(\epsilon|\mathbf{c}, \mathbf{e}) = 0$.

Based on this equation, one can compute expected outcome:

$$E(y|\mathbf{c}, \mathbf{e}) = g(\mathbf{c}, \mathbf{e}) \tag{4.2}$$

Two approaches are commonly applied to measure inequality of opportunity: the *ex-post* and *ex-ante* approaches. The ex-post approach eliminates between-tranche inequality — the inequality of a counterfactual distribution in which each individual's outcome is replaced by the expected outcome given one's effort (Checchi and Peragine, 2010). In contrast, the ex-ante IOP measures between-type inequality — the inequality of a counterfactual distribution where individual's outcome is replaced by the expected outcome given one's circumstances (See Van De Gaer, 1993 and Checchi and Peragine, 2010).

Using the ex-ante approach, one can measure inequality of opportunity by a distribution of expected outcome conditional on circumstances $E(y|\mathbf{c})$:

$$IOP_{ante} = I(E(y|\mathbf{c})) \tag{4.3}$$

where $I(\cdot)$ is an inequality index such as Gini coefficient, Theil index or the mean log deviation (MLD).

Using the ex-post approach, inequality of opportunity can be measured by a standardised distribution of expected outcome where the difference between tranches is eliminated:

$$IOP_{post} = I[\frac{y * \bar{\mu}}{E(y|\mathbf{e})}] \tag{4.4}$$

where $\bar{\mu} = \frac{1}{N}\sum_{i \in N} y_i$ is the mean of expected outcome y, and $E(y|\mathbf{e})$ is the expected income given efforts $\mathbf{e}$. Therefore, $\frac{y}{E(y|\mathbf{e})}$ is an element-wise rescaling for each individual's income.

One can also measure inequality of opportunity relative to the observed inequality:

$$IOR = \frac{IOP}{I(y)} \tag{4.5}$$

This measure shows the proportion of income inequality due to inequality of opportunity.

The difference between the ex-ante and ex-post approach is that the ex-post approach requires information on effort while the ex-ante approach can measure inequality of opportunity without knowing effort. This difference makes the ex-post approach more data-demanding.

If effort is observed, one can also use a measure named *direct unfairness* proposed by Fleurbaey and Schokkaert (2009). In this measure, a reference value for effort $\tilde{\mathbf{e}}$ is chosen. Hence, the measure of inequality of opportunity represents the inequality of a counterfactual distribution if every individual exerts the reference level of effort $\tilde{\mathbf{e}}$, and

$$IOP_{DU} = I(E(y|\mathbf{c}, \tilde{\mathbf{e}})) \tag{4.6}$$

Due to the difficulty in observing effort, most literature used the ex-ante approach. Therefore, a common practice of measuring inequality of opportunity in terms of income inequality is based on the following linear regression (Ferreira et al., 2011):

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon \tag{4.7}$$

where y_i is individual income, $\mathbf{x}_i$ is a set of observed individual circumstances and ϵ is the error term. In this model, one treats the error term as the contribution of effort to income.

Once the parameters $\boldsymbol{\beta}$ are identified, the level of inequality of opportunity can be measured using the following equation (Ferreira et al., 2011).

$$IOP_{ante}^{ols} = I(E(y|\mathbf{x})) = I(\mathbf{x_i}'\hat{\boldsymbol{\beta}}) \tag{4.8}$$

where $\hat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$.

However, this model relies on the following four assumptions.

Assumption 4.2.1 *Circumstances* $\mathbf{x}$ *have a linear relationship with income* y [1]

A non-linear relationship between circumstances and income could violate this assumption. Hufe and Peichl (2015) found that the measure of inequality of opportunity increases by 11% after considering the non-linearity between circumstances and income. In this study, we add some interaction terms to capture possible non-linearity between circumstances and income.

Assumption 4.2.2 *Observed circumstances* $\mathbf{x}$ *and unobserved effort included in the residual* ϵ *should be independent of each other.*

This assumption could be an issue because effort is likely to be shaped by circumstances. This raises the question of whether the indirect effect of circumstances on individual income through effort should be included in the measure of inequality of opportunity. Scholars have different views on this question. Roemer (1998) argued that the part of effort determined by circumstances should also be considered as circumstances and compensated; while Barry (2005) believed that this part of effort should be rewarded. Another view, that of Swift (2005) emphasizes the autonomy of a family. He argued that interactions within families should be respected. If some parts of effort are determined by family-related circumstances such as family backgrounds, that part of circumstances should be respected and rewarded. Jusot et al. (2013) compared the contribution of circumstances and effort to health inequality based on these three different views and found little difference in the contributions of circumstances and effort.

Assumption 4.2.3 *No unobserved circumstances exist in the model.*

In Equation (4.7), $\mathbf{x}$ is unlikely to capture all circumstances. The contribution of unobserved circumstances could be included in the residual and treated as the contribution of effort. Hence, the measure of inequality of opportunity could be downwardly biased. Using Monte Carlo simulations, Lara Ibarra and Martinez Cruz (2015) found that the omission of a relevant circumstance can cause up to an 80 percent underestimation on inequality of opportunity.

Assumption 4.2.4 *The error term* ϵ *should be homoskedastic, i.e.* $Var(\epsilon|x) = \sigma^2$

A heteroskedastic error term could indicate heterogeneity in type-specific effort distribution or heterogeneity in the effect of circumstances on individual income. Björklund

[1] Income could also be log-normally distributed. In this case, y represents log income and circumstances could have a linear relationship to log income.

et al. (2012) found that the heterogeneous type-specific variance is an important source of inequality of opportunity. Waltenberg and Vandenberghe (2007) captured the heterogeneous effect of circumstances on individual income using a conditional quantile regression model.

4.3 Measuring Inequality of Opportunity: A Latent-Class approach

4.3.1 Model Specification and Assumptions

Assume there are N individuals where $\mathcal{N} = \{1, \ldots, N\}$ represents the set of population. Due to the difficulty of observing effort, we assume a vector of circumstances $\mathbf{c}$ and define effort as a categorical variable with m finite values. Let $\mathbf{\Omega}$ denote the finite set of circumstances and let $\mathcal{J} = \{1, \ldots, m\}$ represent the finite levels of effort. By working definition, circumstances $\mathbf{c}$ are exogenous and time-invariant variables. Level of effort could be time-variant and influenced by circumstances[2].

We further assume a vector of probability to exert each level of effort. Given that j is a level of effort such that $j \in \mathcal{J}$, for all $j \in \mathcal{J}$, π_i^j represents individual i's probability to exert j level of effort.

One can generate income y^j given a level of effort j for all $j \in \mathcal{J}$ by the function $g^j : \Omega \mapsto \mathbb{R}_+$:

$$y^j = g^j(\mathbf{x}) + \epsilon^j \tag{4.9}$$

where y^j represents income received if exerting j level of effort. Since function g^j varies across j, the effect of circumstances on income is heterogeneous across levels of effort. Therefore, we assume the homoskedasticity of ϵ_j instead of ϵ in Assumption 4.2.4.

Assumption 4.3.1 *The error term ϵ^j should be homoskedastic, i.e. $Var(\epsilon^j|\mathbf{x}) = {\sigma^j}^2$ for all $j \in \mathcal{J}$.*

One can compute the expected value of y^j using the following equation.

$$E(y^j) = g^j(\mathbf{x}) \tag{4.10}$$

For the convenience of interpretation and comparison, we restrict the label of effort in an increasing order so that a bigger index represents a higher level of effort.

Assumption 4.3.2 *For all $j, l \in \mathcal{J}$, $j > l$ (or $j < l$) indicates that j (or l) is a higher level of effort than l (or j).*

[2] Since we use a cross-sectional model in this chapter, we specify this model with no regard to time.

Therefore, if two individuals possess the same circumstances, higher effort should result in higher income, and vice versa:

Assumption 4.3.3 *For all $i, k \in \mathcal{N}$ and $i \neq k$, given $\mathbf{x}_i = \mathbf{x}_k$, $y_i^l > y_k^j \Leftrightarrow l > j$ for $l, j \in \mathcal{J}$.*

Suppose y_i^j is individual i's income given j level of effort and π_i^j is the probability that individual i exerts j level of effort; for any individual $i \in N$, one can compute an individual's expected income given circumstances:

$$E(y_i|\mathbf{x}) = \sum_{j \in \mathcal{J}} \pi_i^j E(y^j|\mathbf{x}) \tag{4.11}$$

Similarly to the conventional approach, one can measure inequality of opportunity using the ex-ante, the ex-post or the direct-unfairness approach.

Derived from Equation 4.3, inequality of opportunity using the ex-ante approach is:

$$IOP_{ante} = I(E(y|\mathbf{x}) = I\{\sum_{j \in \mathcal{J}} \pi^j E(y^j|\mathbf{x})\} \tag{4.12}$$

where $E(y|\mathbf{x})$ is the vector of expected incomes given circumstances.

Based on Equation 4.6, inequality of opportunity using the direct-unfairness approach is that for any $j \in \mathcal{J}$, we set $\pi^j = 1$ and $\pi^l = 0$ for all $l \neq j \in \mathcal{J}$:

$$IOP_{DU}^j = I\{E(y^j|\mathbf{x})\} \tag{4.13}$$

This measure examines the level of inequality of opportunity given that every individual exerts j level of effort.

Although effort cannot be observed, one can estimate individuals' level of effort based on the probabilities π. Individuals are more likely to be exert j level of effort if $\pi_i^j > \pi_i^l$ for all $l \neq j, l \in \mathcal{J}$. We define that $\tilde{j}_i = \arg\max_{j \in \mathcal{J}} \pi_i^j$ is a classification variable which identifies the level of effort each individual exerts. This classification variable can partition individual income based on levels of effort. If the latent class captures all variations of income due to effort, i.e. the within class variation is only due to circumstances, the latent class can represent the tranche defined in the conventional approach. Alternatively, we define the latent class as "effort group".

Whether or not a latent class represents a tranche, the ex-post measure can be used to eliminate the between-class inequality.

Defining that individuals exerting j level of effort are in the set $\mathcal{N}^j = \{1, \ldots, N^j\}$, we can measure the ex-post inequality of opportunity using:

$$IOP_{post} = I[\frac{y\mu}{\tilde{y}}] \tag{4.14}$$

where IOP_{post} is ex-post inequality of opportunity, μ is the mean of y, $\tilde{y}$ is the counterfactual income distribution replacing each individual income by the mean income of the given identified class.

Decomposition approaches can also be used to measure the ex-post inequality of opportunity. Since the latent class partitions data into several effort groups, one can make use of the entropy index to decompose overall income inequality into within-group and between-group inequality.

For the Theil index, the overall income inequality can be decomposed using the following equation:

$$T = T_w + T_b = \sum_{j\in J} s_j T_j + \sum_{j\in \mathcal{J}} s_j \ln \frac{\mu_j}{\mu} \tag{4.15}$$

where T is the Theil index for overall income inequality, T_w and T_b are within-class Theil index and between-class Theil index respectively, s_j is the income share of class j, μ_j is the average income of class j and μ is the average income for the entire population.

If the latent class represents the effort group, the between-class inequality measured by T_b should represent inequality due to effort. The within-class inequality T_w captures inequality due to circumstances and inequality due to variation in effort within each effort group. Therefore, T_w indicates the upper-bound of inequality of opportunity. It gets closer to the real inequality of opportunity when the latent class captures more variation in effort.

Similarly, Mean Log Deviation (MLD) can also be used in the decomposition of income inequality:

$$L_T = L_w + L_b = \sum_{j\in J} r_j L_j + \sum_{j\in J} r_j \ln \frac{\mu}{\mu_j} \tag{4.16}$$

where L is the MLD index for overall income inequality, L_w and L_b are within-class MLD index and between-class MLD index respectively, and r_j is the population share of class j.

4.3.2 The Finite Mixture Model

To estimate g^j and π_i^j in Equation (4.11) , we used the finite mixture model (the latent class model[3]). We assumed that effort is a latent variable h with m finite values and each i has the probability π_i^j on each class j so the average probability for being in class j is π^j. If a sample is large enough, π^j can also represent the population proportions of class j.

Formally, individual income is determined by circumstances $\mathbf{x}$ and class j:

[3] We used the finite mixture model (FMM) and the latent class model (LCM) interchangeably: the number of mixture components is equal to the number of latent classes

$$y_i^j \;=\; x_i'\beta^j + \epsilon_i^j \tag{4.17}$$

$$y_i \;\sim\; \sum \pi_i^j \mathcal{N}(\mu_j, \sigma_j^2) \tag{4.18}$$

For π_i^j, we first identified the mean of π_i^j — π^j where $\pi^j = E[\pi_i^j] = \frac{1}{N}\sum_{i=1}^N \pi_i^j$, $\sum \pi_j = 1$ and $1 > \pi_j > 0$. Let $\theta_j = (\beta_j, \sigma_j^2, \pi_j)$ be the parameters of the model. To be consistent with Assumptions 4.3.2 and 4.3.3, we defined the order of the components as $\mu_{j-1} < \mu_j < \mu_{j+1}$.

Given each $j \in J$, the distribution of income conditional on $\mathbf{x}$ and j can be represented by $f_j(y|\mathbf{x};\theta_j)$ where $\theta_j = (\mu_j, \sigma_j)$ is the vector of parameters from the normal distribution of tranche j with the class-specific mean $\mu_j = \mathbf{x}_i\boldsymbol{\beta}_j$ and variance σ_j^2. Therefore, the overall distribution of income $f(y|\mathbf{x})$ given $\mathbf{x}$ is the weighted sum:

$$f(y|\mathbf{x}) = \sum_{j=1}^{m} \pi^j f_j(y|\mathbf{x};\theta_j) \tag{4.19}$$

Instead of satisfying Assumption 4.2.1 to Assumption 4.2.4 in the conventional model, this model can relax Assumptions 4.2.1 and 4.2.4 so it should satisfy Assumption 4.2.2, 4.2.3, 4.3.1,4.3.2 and 4.3.3.

To compute the conditional expectation of individual income using Equation (4.11), the probabilities π_i^j should also be estimated. To estimate π_i^j, we used the posterior probability of being in class j:

$$\pi_i^j = Pr(i \in j|\theta, y_i) = \frac{\pi^j f_j(y_i|\mathbf{x}_i, \theta_j)}{\sum_{j\in J}\pi^j f_j(y_i|\mathbf{x}_i, \theta_j)}, \forall j \in J \tag{4.20}$$

In this equation, a higher prior probability could result in a higher posterior probability, and the posterior probability of class j also depends on the expected income in that class. Higher expected income in one class indicates higher probabilities for individuals of being in that class. This effect of expected income on individual's probabilities is consistent with what we assumed in Assumption 4.3.3, namely that a higher income indicates a higher level of effort.

Therefore, the conditional expectation of individual income is a weighted sum of the conditional mean of each class:

$$E(y_i|\mathbf{x}_i) = \sum_{j=1}^{m} \pi_i^j \mu_j \tag{4.21}$$

where $\sum_{j=1}^m \pi_i^j = 1$, $\mu_j = E_j(y_i|\mathbf{x}_i)$.

To estimate the parameters $\theta = \{\pi^1, \ldots, \pi^m, \theta_1, \ldots, \theta_m\}$, we used gradient-based optimization methods (Deb, 2008) such as the Gauss-Newton or Newton-Raphson method to maximize the log-likelihood, which has the following functional form:

$$l(\theta) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \pi^j f(y_i|\mathbf{x}_i;\theta_j) \tag{4.22}$$

Using Equation (4.21), one can compute a counterfactual income distribution based on the estimator of θ for each class. Then, inequality of opportunity for each effort group can be measured, which sheds light on whether individuals with a high level of effort suffer a higher inequality of opportunity.

Our study attempted a finite mixture model with two to six components and used the information criterion (AIC and BIC) to determine which model performed better. In either case, the number of effort groups or levels of effort are assumed to be the number of components. For each component, we estimated the counterfactual income distribution, which is the income distribution for a given level of effort.

We also used the posterior probability to categorise individual income into different levels of effort. We compared the posterior probability for each individual in each class and classified individual income into the class with the highest posterior probability. Then using Equation (4.14), we removed the between-class inequality and computed the ex-post inequality of opportunity.

To measure ex-ante inequality of opportunity, we used the expected income computed by Equation (4.21) and computed the measures using Equation (4.12).

4.3.3 A Finite Mixture Model with Varying Prior Probabilities

One issue related to a finite mixture model is whether the latent class represents effort. The components in the model could also be a proxy for other variables such as unobserved circumstances, individual intelligence, etc. To address this issue, we used the generalized finite mixture model with varying prior probabilities (FMMV) (Jacobs et al., 1991) in which the prior probabilities π_i^j are parametrized. The model can be defined as the following mixture distribution,

$$f(y|\mathbf{x}) = \sum_{j=1}^{m} \pi_i^j(\mathbf{z}_i) f_j(y|\mathbf{x};\theta_j) \tag{4.23}$$

where π_i^j is a logistic transformation of a function of $\mathbf{z}$:

$$logit(\pi_i^j) = \alpha_j + \mathbf{z}_i\boldsymbol{\gamma}_j \tag{4.24}$$

This model allows the prior probability π_i^j to depend on some exogenous variables $\mathbf{z}_i$. We used two different specifications: one captures the influence of circumstances (denoted as $\mathbf{z}^c$), and the other represents the impact of effort (denoted as $\mathbf{z}^e$). We named the former specification as Finite Mixture Model 1 with Varying Probability (FMMV1) and the latter as Finite Mixture Model 2 with Varying Probability (FMMV2). In FMMV1, the circumstance-determined latent class captures the effect of unobserved circumstances or unobserved effort due to circumstances; while in FMMV2, the effort-determined latent class captures the pure effect of unobserved effort. In terms of the effort variables in FMMV2, since they can be endogenous and influenced by circumstances, we can express the relationship between each effort variable and circumstance in the following equation.

$$\begin{aligned} logit(\pi_i^j) &= \alpha^j + \mathbf{z}_i^e \gamma^j & (4.25) \\ \mathbf{z}_i^e &= \phi \mathbf{X}_i + \mathbf{v}_i^e & (4.26) \end{aligned}$$

where $\mathbf{X} = [\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x}]'$ and $\mathbf{v}_i^e$ is the residual for $\mathbf{z}_i^e$, which is independent of circumstances variables.

To let π_i^j only represent the effort level independent with circumstances, we set $\phi = 0$. Therefore, the logistic transformation can only transform the residuals of effort variables.

$$logit(\pi_i^j) = \alpha^j + \mathbf{v}_i^e \gamma^j \qquad (4.27)$$

4.4 Data Description

To measure inequality of opportunity in China, we use the data from China Family Panel Studies (CFPS). CFPS is a nationally representative annual longitudinal survey containing not only individual-level data but also household- and community-level data. It has been conducted since 2010 by the Institute of Social Science Survey (ISSS) of Peking University, China. By 2016, this project has already published its survey including three years— 2010 baseline survey, 2011 maintenance survey and 2012 follow-up survey. Since the survey conducted in 2011 is a maintenance survey, the sample size is small relative to 2010 and 2012. We will not include the 2011 survey in our research.

CFPS covers 16,000 households with more than 33,000 adults and 8,900 youths in 25 provinces, municipalities or autonomous regions in China. Its main purpose is to record the changes in the socio-economic wellbeing of Chinese people, covering a variety of topics such as economic activities, educational attainment, family relationships and dynamics, migration, and physical and mental health. The design of CFPS was inspired by authoritative panel studies in other countries such as the Panel Study of Income Dynamics (PSID) so that the international comparison becomes possible.

We used the China Family Panel Study (CFPS) dataset in 2010 and 2012 and selected individuals aged 21 to 60[4]. The initial sample sizes in 2010 and 2012 were 33,600 and

[4] There is a degree of repetition between Section 4.4 and 3.5. This information is repeated for the benefit of the readers who could be interested only in this essay

35,720 respectively, in which 26,393 subjects have records for both 2010 and 2012. Then we selected individuals aged between 21 to 60 years because the labour participation rate outside this age range is relatively low. After filtering, the sample size was reduced to 19,736.

Tables 4.1 and 4.2 present the summary statistics of the variables we used in the study. Male respondents make up 47% of the sample. Ethnicity is represented as the dummy variable for being in the minority groups in China. It is 0 if an individual's ethnicity is the majority group —Han; otherwise, it is 1. The percentage of the minority group is around 8%. In addition, the average age of the respondents is 42.25.

We also include the number of siblings as one of the circumstance variables. Becker and Lewis (1974) studied the relationship between the number of children and children's outcomes such as educational attainment and socioeconomic status. An empirical study conducted in China (Li et al., 2008) also found a negative correlation between family size and child outcome. In the dataset, the average number of siblings is around 3.

Another circumstance variable we used is regions of residence when the respondents were 12 years old. Two dummy variables were generated as the measurement of regions of residence. One is whether individuals held a non-agriculture Hukou status at age 12 and another is whether individuals lived in coastal provinces at age 12.

Hukou is a system for recording household registration in China. It divides households into agriculture (rural) and non-agriculture (urban) Hukou. The former live in rural areas and are registered as rural households and the latter live in urban areas and are registered as urban households. Since there is difficulty in changing the Hukou status from agriculture to non-agriculture, many rural immigrants hold agriculture Hukou even though they live in urban areas. Individuals normally have the same Hukou status as their parents before they are grown-up. In our sample, the percentage of individuals who held non-agriculture (urban) Hukou when they were 12 years old is 15%.

Coastal provinces are the provinces in Eastern China along the coast. This area is more developed than the inland area. We used a dummy variable to capture whether an individual lived in the coastal province when he/she was 12 years old (coastal12). The data showed that about 43% of the respondents lived in the coastal province when they were 12 years old.

The annual individual income is 10,575 yuan on average in 2010 and 13,412 yuan in 2012. To construct the income variable, we first computed the labour income by summing individual wages, awards, allowances, incomes of working out of town and bonuses. Then we matched each individual to his household's business income (including agricultural and non-agricultural business income), property income, transfer income and other income (including gifts). The individual income is equal to labour income plus all sources of non-labour income divided by the family size. In addition, we took the provincial-level inflation rate into account. We used the Consumer Price Index (CPI) in Chinese Statistical Yearbooks and adjusted income in 2012 to the same price level as income in 2010. In this study, we used the average income of the years 2010 and 2012 as the dependent variable because the average income eliminates the variation over time. It shows individual's income over a longer period.

In Table 4.1, we also reported variables related to effort. Around 6% of respondents are members of the Chinese Communist Party (CCP). On average, respondents had 6.54 years of education, 61% of them were employed and 30% of them migrated from rural to urban areas. They spent 34.98 hours on average on work and study per week.

Table 4.1: Summary Statistics (Respondents)

Statistic	N	Mean	St. Dev.	Min	Max
Average Income	19,736	11,993.64	20,389.70	0.00	864,893.10
Circumstances					
Male	19,736	0.47	0.50	0	1
Minority	19,696	0.08	0.27	0	1
Age	19,736	42.25	10.79	21	60
Urban Hukou at age 12	19,625	0.15	0.35	0	1
Live in Coastal Province at age 12	19,736	0.43	0.50	0	1
Number of Siblings	19,736	2.98	1.90	0	14
Income from other Households	18,729	12,271.25	20,694.68	1.00	697,459.90
Effort					
CCP Member	19,736	0.06	0.24	0	1
Years of Education	19,732	6.54	4.84	0.00	22.00
Employed	19,134	0.61	0.49	0	1
Migrant	19,625	0.30	0.46	0	1
Hours spent on Work and Study per Week	19,736	34.98	26.30	0.00	140.00

[1] Average income is the average income between 2010 and 2012.
[2] CCP is the Chinese Communist Party.

Table 4.2 shows respondents' *parents' background when respondents were 14 years old* including parent's education level, parent's occupation status and parent's political affiliation. For all variables, we used the higher value of the two parents.

In terms of parents' education level, it is reported in eight levels in CFPS. We merged these levels into three levels. (1) Low level: illiterate or semi-literate ; (2) Middle level: primary and junior high school; and (3) High level: senior high school or above. The percentage of parents with the low-level education was 38%, and the high level made up 21%.

As far as parents' occupations are concerned, they include eight large categories including 595 specific occupational codes in CFPS. We regrouped them into three levels: the low level included agricultural workers and workers in manufacture and transportation sectors; the middle level included professionals, clerks, technical staffs and other tertiary sector workers; and the high level included the administrative/management positions,

Table 4.2: Summary Statistics (Respondents' Parents)

Statistic	N	Mean	St. Dev.	Min	Max
Low Occupation	17,309	0.79	0.41	0	1
Mid Occupation	17,309	0.13	0.33	0	1
High Occupation	17,309	0.09	0.28	0	1
CCP member	19,736	0.16	0.37	0	1
Low Education	19,736	0.38	0.49	0	1
Mid Education	19,736	0.40	0.49	0	1
High Education	19,736	0.21	0.41	0	1

[1] All variables are respondents at 14 years of age.
[2] All variables only account the higher value between parents.

teachers in tertiary education, lawyers and high-rank military officers. On average, 79% of the individuals reported a low status of occupation for their parents and 9% a high status.

Furthermore, we generated a dummy variable equal to 1 when one of the parents was a member of China Communist Party (CCP) when respondents were 14 years old as a proxy for parents' political affiliation. The percentage of members of the CCP was 16%.

For the OLS and FMM models, we selected gender, ethnicity, Hukou status at age 12, living in the coastal province at age 12, number of siblings, parents' educational level, parents' occupational level, and parents' political affiliation as circumstances variables. In terms of the FMMV1 model, we added income from other households to the circumstances variables in FMM to prevent the explanatory variables for effort from being identical to the explanatory variables for log income. In FMMV2, we used the residuals of variables related to effort.

4.5 Results

4.5.1 Results from OLS and Finite Mixture Models

In this section, we present the estimation results from the OLS and four different specifications of a finite mixture model. Table 4.3 shows the estimation results from OLS and FMM[5]. In the first six rows, we presented the estimated average income in year 2010 and 2012 for each component in logarithmic form[6]. Given the number of classes, all FMM models significantly separate given data into classes.

For comparing these models, we presented log-likelihood and information criteria in the last three rows. Obviously, OLS performs worse than any type of finite mixture models in terms of log-likelihood and information criteria. We also used the information

[5] FMM is estimated using Stata module "fmm" from Deb (2008) in Stata version 14.

[6] We took the average of income for year 2010 and 2012 first, and then transformed the average income into the logarithmic form.

Table 4.3: OLS vs. FMM

	(1) OLS	(2) 2FMM	(3) 3FMM	(4) 4FMM	(5) 5FMM	(6) 6FMM
Mean						
Comp.1	9.276***	8.038***	7.702***	7.642***	7.336***	7.331***
Comp.2		9.533***	9.468***	9.451***	9.800***	9.804***
Comp.3			9.824***	9.807***	9.437***	9.434***
Comp.4				3.532***	0.953***	4.796***
Comp.5					10.027***	10.019***
Comp.6						10.468***
π_1		0.210	0.174	0.164	0.133	0.133
π_2		0.790	0.534	0.520	0.119	0.118
π_3			0.292	0.307	0.534	0.533
π_4				0.009	0.008	0.008
π_5					0.206	0.204
ll	-2.97e+04	-2.87e+04	-2.83e+04	-2.83e+04	-2.81e+04	-2.81e+04
aic	59491.554	57476.080	56769.411	56707.641	56412.463	56344.184
bic	59599.784	57715.733	57132.756	57194.678	57023.192	57078.605

[1] $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.
[2] The dependent variable for all models is the average income for year 2010 and 2012 in the logarithmic form.
[3] Source: Authors' Analysis.

matrix test (Cameron and Trivedi, 1990) to check for normality and homoskedasticity assumption. The test rejects both normality and homoskedasticity in OLS model. This rejection indicates a violation of Assumption 4.2.4 and heteroskedasticity in OLS. Therefore, a measure of inequality of opportunity using OLS could be an underestimate as it does not account for heteroskedasticity.

Column (2) to (6) show the results from FMM with two to six components. Based on log likelihood and information criteria, we concluded that 3FMM preforms significantly better than 2FMM; while 4FMM, 5FMM and 6FMM performs slightly better than 3FMM. However, the prior probability of component 4 in 4FMM, 5FMM and 6FMM is around 0.008. This suggests that respondents are highly unlikely to be in component 4. Therefore, we prefer 3FMM models. Since we are mainly interested in a predictive mean, a more parsimonious model is to be preferred.

Table 4.4 shows the estimation results from OLS and 3FMM with constant and varying prior probabilities respectively. Columns (4) and (5) show the results from two FMMV models with three components. Both models perform significantly better than 3FMM and OLS. The Akaike's information criterion (AIC) drops from 56,769 in 3FMM to 53,025 in FMMV1 and to 51,520 in FMMV2. Comparing the two FMMV models, we find that FMMV2 performs slightly better than FMMV1.

Table 4.4 also shows the mean and variance of each component. In general, means of log income of two FMM models range from around 8 to 10 and means of the two FMMV models range from about 6.7 to 10.1. Among the three components, the highest value is

Table 4.4: OLS vs. FMM (Including Varying Probability Model)

	(1) OLS	(2) 2FMM	(3) 3FMM	(4) FMMV1	(5) FMMV2
Mean					
Comp.1	9.276^{***}	8.038^{***}	7.702^{***}	6.713^{***}	7.685^{***}
Comp.2		9.533^{***}	9.468^{***}	8.942^{***}	9.403^{***}
Comp.3			9.824^{***}	10.041^{***}	10.054^{***}
Variance					
Comp.1		1.788^{***}	1.956^{***}	2.028^{***}	1.718^{***}
Comp.2		1.022	0.933^{**}	1.127^{***}	0.866^{***}
Comp.3			0.591^{***}	0.585^{***}	0.642^{***}
π_1		0.210^{***}	0.174^{***}		
π_2		0.790^{***}	0.534^{***}		
Loglikelihood	-2.97e+04	-2.87e+04	-2.83e+04	-2.64e+04	-2.57e+04
AIC	59491.554	57476.080	56769.411	53025.023	51519.852
BIC	59599.784	57715.733	57132.756	53601.044	51958.817

[1] $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.
[2] The dependent variable for all models is the average income for year 2010 and 2012 in the logarithmic form.
[3] Source: Authors' Analysis.

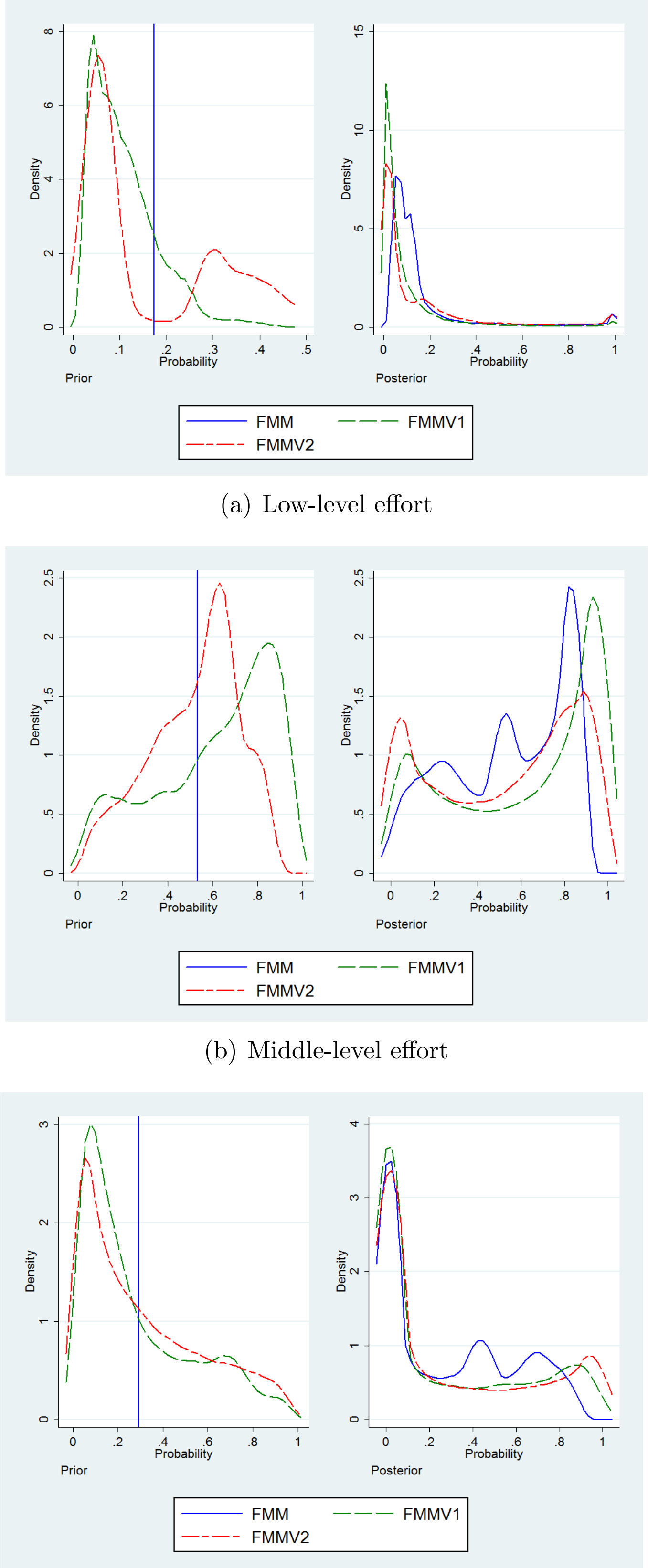
the mean of component 3 and the lowest is that of component 1. This result is consistent with the labelling convention we are using. Therefore, component 1 represents a low-effort level and component 3 represents a high-effort level according to the "monotonic tranche" assumption (Assumption 4.3.3). In terms of the variance, component 1 has the largest variance, and component 3 has the smallest for all 3-component models. As a result, the distribution of component 1 might overlap with the distributions of other components. However, variances are smaller in FMMV2 models than in 3FMM and FMMV1, which indicates that the overlap among the three components would be smaller in FMMV2 models. Therefore, we conclude that FMMV2 models appear to offer a better separation.

Given results from FMM models, we computed the prior and posterior probabilities for each individual in terms of both 3FMM and FMMV models. The distributions of the prior and posterior probabilities are presented in Figure 4.1. We find that distributions of the posterior probabilities are more polarized towards 0 or 1 compared to distributions of prior probabilities, which brings more certainties in predicting the level of effort for each individual. Therefore, we choice posterior probabilities to identify individual's level of effort rather than prior probabilities.

Using posterior probabilities for FMM models, we estimated the kernel density distributions of fitted log-income of the OLS model, three different FMM models and observed income (see Figure 4.2). The observed log-income distributes from roughly 5 to 12 in both years with two humps. This bimodality is consistent with a mixture of distributions.

Comparing the estimations of the four models, the estimation of the OLS model is highly concentrated around the mean. It poorly predicts either the bimodality or the two

Figure 4.1: The Kernel Distribution of the Prior and Posterior Probabilities by Levels of Effort

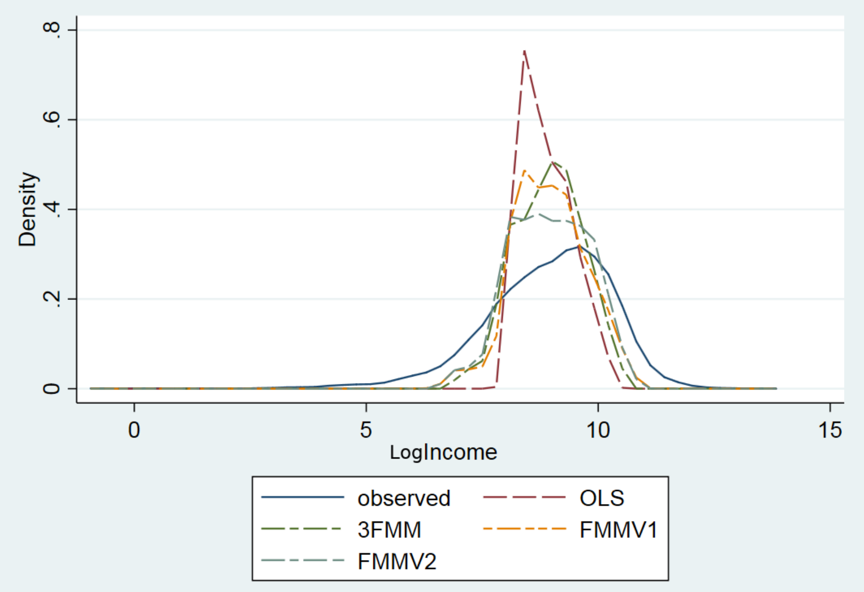


(a) Low-level effort

(b) Middle-level effort

(c) High-level effort

Figure 4.2: Comparing Predicted and Observed Log-Income



tails of the observed distribution. The three FMM models fit better to the bimodality of observed income compared to the OLS model. In terms of the three FMM models, FMMV2 has better goodness of fit than the other two FMM models.

This difference between OLS and FMM models shows that identifying levels of effort and allowing the effect of circumstance on income to be different across levels of effort significantly increase the goodness of fit. The highest goodness of fit appears in FMMV2 when the latent class is more closely determined by pure individual effort.

For three FMM models, we also examined determinants of the prior probabilities. Prior probabilities predict the likelihood of exerting different effort levels before knowing income. In 3FMM, prior probabilities are fixed. This model identifies that every individual has a 53.4% probability of belonging to the middle-effort level and a 17.4% probability of belonging to the low-effort level. In the two FMMV models, prior probabilities were parametrized by variables related to effort.

Table 4.5 shows the estimators for prior probabilities in the two FMMV models. The high-effort level is the baseline. In FMMV1, the independent variables are considered as observed circumstances. We find that female or minorities with rural Hukou are more likely to be in the low effort level. Individuals whose family members have higher income are more likely to be in the low level. The coefficients of parents' socio-economic status are not significant in the low level but negative and significant in the middle level, which means that individuals whose parents have a higher socioeconomic status are more likely to be in the high level.

In FMMV2, the independent variables are considered as pure effort. All coefficients are negative and significant, which indicates that individuals are more likely to be in the high-effort level if they exert more effort via CCP membership, years of education,

Table 4.5: The Estimators for Prior Probabilities in FMMV models

	FMMV1			FMMV2	
Effort Level	Low	Middle		Low	Middle
Female	-0.214	0.305	CCP Membership	-1.247***	-0.723***
Minority group	-1.849*	-0.018	Years of Education	-0.212***	-0.212***
Rural Hukou at age 12	0.317	1.610***	Employed	-3.382***	-1.467***
Female and Minority	0.88	0.625	Migration	-0.839***	-1.659***
Rural Hukou and Minority	2.753***	1.477***	Hours spent	-0.010***	-0.014***
Rural Hukou and Female	1.528***	0.903**			
Living in coastal province	0.134	-0.973***			
Mid-educated Parents	0.011	-0.647***			
High-educated Parents	0.229	-0.449***			
Mid Occupation (Parents)	0.133	-0.667***			
High Occupation (Parents)	0.168	-0.850***			
Parents CCP Membership	-0.274	-0.187			
Number of sibling	0.044	0.234***			
Log-income from other families	0.203***	0.271***			
Constant	-1.860***	-0.778*		-0.544***	0.940***

[1] $^{*}p < 0.10,^{**} p < 0.05,^{***} < 0.01$

[2] Source: Authors' Analysis

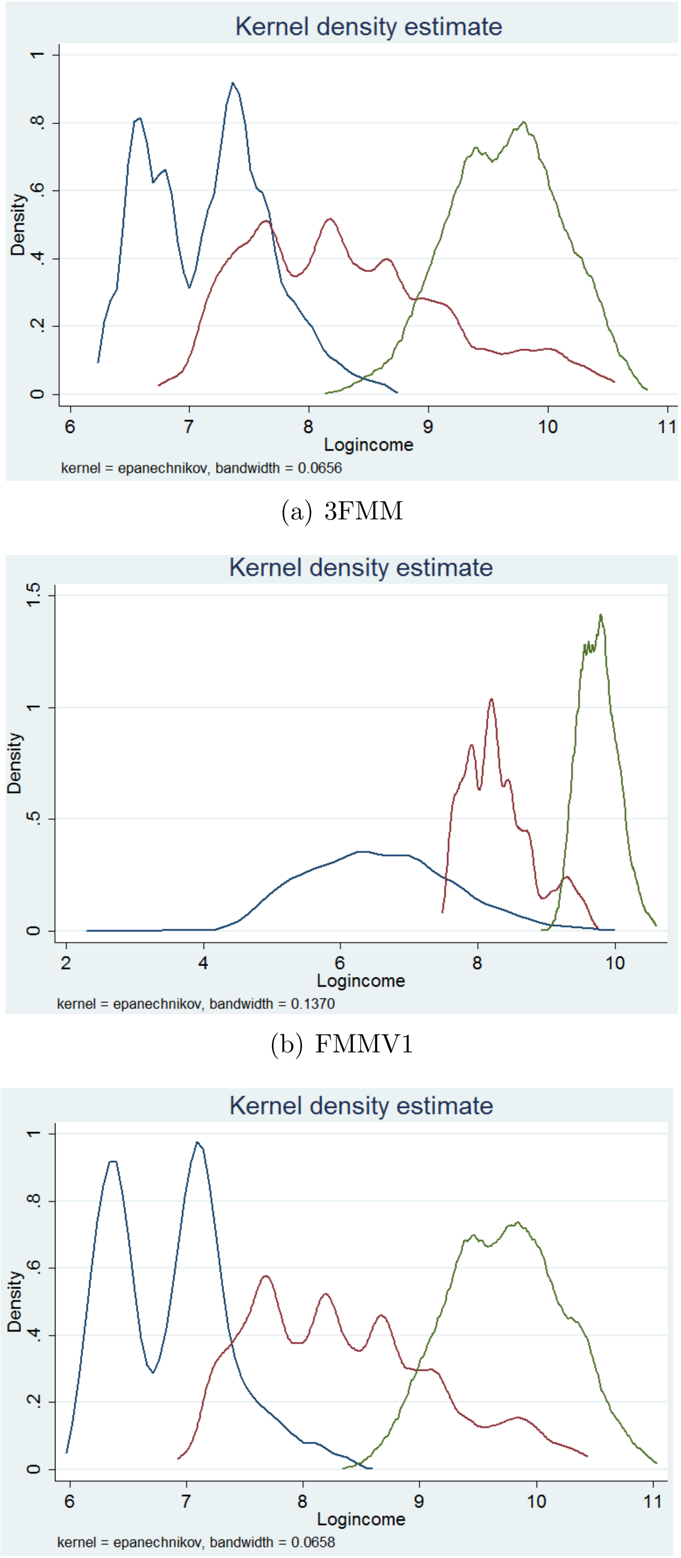
getting employed, migration, and hours spent on work and study.

To show how circumstances affect income differently for each level of effort, we estimated kernel density distributions of three components for each FMM models. The results are shown in Figure 4.3. The log income distribution of each component is the counterfactual log income assuming all individuals in the sample are in that component. In the 3FMM and FMMV2 models (Figure 4.4(a) and 4.4(c)), component 2 overlaps with component 1 and 3; while in the FMMV1 model, component 1 distributes from around 2 to 10, which overlaps with component 2 and 3. This is because the latent class is assumed to be determined by circumstances in FMMV1 while for FMMV2 the latent class is determined by pure effort. Since 3FMM is closer to FMMV2, the latent classes for FMM are more likely to be determined by pure individual effort.

Given the results from 3FMM and FMMV2, we can conclude that individuals with the middle level of effort could have the highest inequality of opportunity — the highest effect of circumstances on income. The estimated distribution for these two models are bimodal for the low-level effort, multimodal for the middle-level effort and close to unimodal for the high-level effort, which indicates that circumstances affect income differently across levels of effort.

For FMMV1, given that the latent class is determined by observed circumstances, the lowest level has the highest inequality of opportunity, which indicates that circumstances could have more effect on income if individuals exert the low level of circumstance-correlated effort — the effort that is driven by their own circumstance.

Figure 4.3: Kernel Density Estimation of Three Components



(a) 3FMM

(b) FMMV1

(c) FMMV2

To more specifically show heterogeneous effects of circumstances on income across effort levels, we present the estimators of circumstance variables for 3FMM and FMMV2 in Table 4.6. In general, the results from 3FMM and FMMV2 are similar.

Table 4.6: The Heterogenuous Effect of Circumstance Variables on Income Across Effort Levels

		3FMM			FMMV2	
Effort Level	Low	Middle	High	Low	Middle	High
Female	-0.743***	-0.315***	-0.196***	-0.759***	-0.360***	-0.223***
Minority group	-0.572	-0.854**	0.154*	-0.524	-0.806**	-0.053
Rural Hukou at age 12	-0.38	-1.107***	-0.012	-0.661***	-1.029***	-0.198***
Female and Minority	0.385	0.522***	-0.288**	0.2	0.498***	-0.192*
Rural Hukou and Minority	0.483	0.001	-0.306***	0.456	0.07	-0.069
Rural Hukou and Female	-0.086	-0.682***	-0.372***	-0.016	-0.614***	-0.367***
Living in coastal province	-0.196*	0.507***	0.394***	-0.214**	0.452***	0.387***
Mid-educated Parents	0.222**	0.289***	0.236***	0.138	0.307***	0.250***
High-educated Parents	0.295**	0.273***	0.217***	0.234*	0.255***	0.251***
Mid Occupation (Parents)	0.19	0.200***	0.200***	0.146	0.198***	0.183***
High Occupation (Parents)	0.157	0.230***	0.211***	0.126	0.194***	0.219***
Parents CCP Membership	0.409***	0.064	0.105***	0.308***	0.076**	0.120***
Number of sibling	0.015	-0.037***	-0.084***	0.024	-0.030***	-0.077***
Constant	7.702***	9.468***	9.824***	7.685***	9.403***	10.054***

[1] $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

[2] Source: Authors' Analysis

More importantly, there are significant differences in the effect of circumstances across effort levels. Some circumstances such as being in minority groups, having rural hukou, female in minority groups, rural female and living in a coastal province could have huge impacts on income in the middle level of effort compared to other levels of effort. This might explain the reason why inequality of opportunity is the highest for the middle-level effort.

For the low level of effort, circumstances like minority groups, rural hukou, female in the minority groups, rural female and parents' occupation have insignificant effects on income; while female and parents CCP membership have the highest impact on income compared to other effect levels. This suggests that the sources of inequality of opportunity for the low level of effect are limited to several circumstances such as gender and political affiliation, which explains the bimodal feature of income distribution presented in 4.3 for this level of effort.

For the high level of effort, although most circumstances have significant effect on income, their effect is relatively small compared to other effort levels. As a result, the income distribution for this level of effort is more concentrated to its mean.

Based on the estimators of the three FMM models, we conducted joint tests to test heterogeneity between class for each circumstance variable. The null hypothesis is that the estimators for a given variable are identical across all classes. The results are presented in Table 4.7. The heterogeneous effect mainly occurs for the variables related to regions such as Hukou status and living in coastal province. The effect of parents' socioeconomic background is homogeneous between classes.

Table 4.7: The Joint Test of Estimators of Circumstance Variables

	3FMM	FMMV1	FMMV2
Female	0.2052	0.156	0.1658
Minority group	0.0373 **	0.1228	0.0362 **
Rural Hukou at age 12	0***	0***	0***
Female and Minority	0.2513	0.0098***	0.0002***
Rural Hukou and Minority	0***	0.5601	0.67
Rural Hukou and Female	0.1357	0.0676*	0.0367 **
Living in coastal province	0***	0.0087***	0***
Mid-educated Parents	0.0831*	0.1876	0.2882
High-educated Parents	0.7826	0.7441	0.6952
Mid Occupation (Parents)	0.8258	0.5871	0.8959
High Occupation (Parents)	0.9499	0.976	0.4095
Parents CCP Membership	0.0443 **	0.6434	0.2292
Number of sibling	0.0039***	0.5893	0***

[1] $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

[2] The null hypothesis is that the estimators for given variable (e.g. female) are identical for all classes. For a three-class model, given estimators $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$. The null hypothesis is: $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3$.

[3] Figures in the table are p-values of the joint test for each variable.

[4] A significant value represents a rejection of the null hypothesis that estimators are the same between classes.

[5] Source: Authors' Analysis

The results from the joint test statistically prove the heterogeneous effect of circumstances on income across levels of effort. This heterogeneous effect is taken into account using the FMM models to measure inequality of opportunity.

4.5.2 Within- vs. Between-Tranche Inequality

In this section, we report within- and between-tranche inequality. Within-tranche inequality shows income inequality among individuals exerting the same effort level. Between-tranche inequality implies income inequality for different effort levels.

Table 4.8 shows predicted income for 3-component models. For 3FMM and FMMV1 models, the expected income for the high level of effort is about six times as high as that for the income for the low level. In terms of the FMMV2 model, the difference in income between the high and the low level is even larger (more than 15 times). This significant difference indicates high between-tranche inequality and shows that FMMV2 captures more between-tranche inequality than the other two, but results are broadly consistent across the three models.

Importantly, the results raise an additional question: whether the difference in effort causes income gap of more than ten times. One reason for such a large income gap could be due to the presence of extreme income observations. We conducted a robustness test by dropping observations with extreme income. The results are similar to those obtained when including extreme income. Therefore, a presence of extreme incomes is unlikely to be the source of large income gap.

In addition, the level of effort is always associated with an individual's decisions and choices. For example, an individual from a high-income family member might have more flexible working hours. A decision concerning working or not can make a huge difference in income. In FMMV models (Table 4.5), we find that if other family members earn no income, individuals are highly likely to be classified into the high level of effort.

Table 4.8 also reports standard deviations (Column 2) and Gini coefficients (Column 3). Gini coefficients show inequality in the counterfactual income distribution assuming every individual exerts the same level of effort (the within-tranche inequality).

We find that component 2 has the highest within-tranche inequality in FMM and FMMV2; while component 1 has the highest within-tranche inequality in FMMV1. The difference in the highest within-tranche could be due to the fact that the class variable is assumed to be determined by circumstances in FMMV1, whereas the class variables in 3FMM and FMMV2 are more likely to indicate the pure effect of effort.

Table 4.8: Predicted Income for Three-Components Model

	Mean	St. Dev	Gini	Min	Max
3FMM					
Comp.1	3948.445	2214.317	0.29	1445.066	15537.1
Comp.2	10038.94	10709.47	0.49	1344.219	61322.71
Comp.3	23200.88	10964.22	0.26	4583.262	67827.6
FMMV1					
Comp.1	3530.331	5001.106	0.56	31.58417	52347.34
Comp.2	8221.551	4966.052	0.30	3120.518	29831.92
Comp.3	23960.5	6941.157	0.16	9947.511	53545.4
FMMV2					
Comp.1	1714.289	1016.217	0.29	644.4515	7762.05
Comp.2	9342.952	9101.504	0.45	1565.19	52502.21
Comp.3	29644.51	15170.18	0.28	6448.702	95158.84

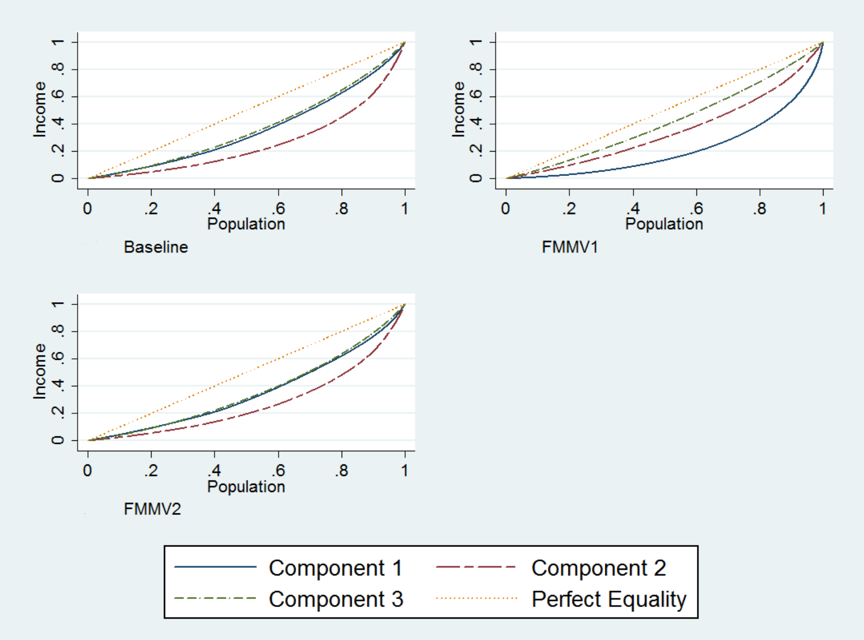
[1] Results are generated using log-normal transformation assuming homoskedasticity.

[2] St. Dev is standard deviation. Gini is Gini coefficient within components.

[3] Source: Author's calculation.

We also use the Lorenz curve to show the within-tranche inequality. Figure 4.4 shows the Lorenz curve of the counterfactual income distribution estimated by three component finite mixture models. The line "Component 1 (2 or 3)" represents counterfactual income distribution if one exerts the low (middle or high) level of effort. In general, income inequality is lowest if every individual exerts high-effort level in all models. Individuals with middle-effort level suffer the highest level of income inequality in FMM and FMMV2. This implies that circumstances have the largest impact on income for those who exert the middle level of effort.

Figure 4.4: Lorenz Curve of Predicted Income



4.5.3 Ex-ante and Ex-post Inequality of Opportunity

To compute inequality of opportunity, we first classified the sample into three tranches based on posterior probabilities. Table 4.9 shows the results of the classification. More than 62.94% is classified as the middle-effort level in 3FMM. The figure reduces to 61.89% in FMMV1 and 56.33% in FMMV2.

Table 4.9: Estimated Proportions of Latent Classes (Percentage)

Low	Middle	High
	3FMM	
6.94	62.94	30.13
	FMMV1	
3.98	61.89	29.74
	FMMV2	
9.10	56.33	31.95

[1] Figures in the table are represented in percentages.
[2] Source: Author's calculation.

Table 4.10 shows the results for ex-ante and ex-post inequality of opportunity using OLS and three different specifications of 3-component finite mixture models. We also used 5-component finite mixture models as a robustness check. Comparing the results obtained with a 3-components to a 5-components finite mixture model, we found that results are similar.

For Table 4.10, we use three different inequality measures: Gini coefficients, Theil index and MLD index. In general, the observed income inequality (Row 3 in Table 4.10) is around 0.65 in Gini coefficients, 0.84 in Theil and 1.05 in MLD.

Table 4.10: The Measures of Inequality of Opportunity (3 Components)

	OLS	3FMM	FMMV1	FMMV2
Gini Coefficient				
Ex-ante IOL	0.35	0.45	0.53	0.53
Observed Gini	0.65	-	-	-
Ex-ante IOR	0.54	0.69	0.82	0.82
Theil				
Ex-ante IOL	0.21	0.34	0.47	0.47
Ex-post IOL		0.77	0.44	0.45
Observed Theil	0.84	-	-	-
Ex-ante IOR	0.25	0.40	0.56	0.56
Ex-post IOR		0.92	0.52	0.54
MLD				
Ex-ante IOL	0.20	0.35	0.50	0.50
Ex-post IOL		0.76	0.40	0.43
Observed MLD	1.05	-	-	-
Ex-ante IOR	0.19	0.33	0.48	0.48
Ex-post IOR		0.72	0.38	0.41

[1] IOL is the absolute level of inequality of opportunity.

[2] Ex-ante IOL uses predicted income of each model; ex-post IOL removes the between-class inequality.

[3] IOR is the relative level of inequality of opportunity; it is the ratio of IOR to observed inequality.

[4] Source: Authors' own calculation based on CFPS data.

The first row shows results from the absolute values of ex-ante inequality of opportunity in Gini coefficients. We find that FMMV models capture more inequality than 3FMM model. IOL is 0.53 using the FMMV2 model and the FMMV1 model. The difference between the four models is similar for the Theil and MLD indices. However, the Gini and entropy indices differ in that Gini coefficients capture more inequality of opportunity than the entropy indices. This is because the Gini coefficients do not have a path independence property (Foster and Shneyerov, 2000) [7] so that the sum of ex-ante IOL and inequality due to effort is larger than the overall income inequality. Since decomposing income inequality in the Gini index requires different approaches such as the

[7] If one decomposes income distribution into two — a smoothed between-group distribution and a standardized within-group distribution, a "path independent" measure keeps the sum of inequality of these two distributions equal to the inequality of the total income distribution

Shapley decomposition, we only use the Theil and the MLD index to compute the ex-post inequality of opportunity in this essay.

For the ex-post measure, we find that the within-class inequality is the highest for 3FMM. This result indicates that the latent class captures less variation in effort compared to FMMV1 and FMMV2. In FMMV1, the ex-post measure captures 38% of the within-class inequality, using the MLD and 52% using the Theil index. Both figures are significantly lower than those obtained in 3FMM. This indicates that between-class inequality is much higher if the latent class represents factors related to observed circumstances. This result also suggests that inequality of opportunity could be highly underestimated if unobserved circumstances and circumstances-related effort are omitted in the measure.

In FMMV2, the ex-post measure captures 41% of the within-class inequality when using the MLD and 54% when using the Theil index. Since the latent class is determined by pure effort in FMMV2, between-class inequality fully represents inequality due to differences in effort. Therefore, this measure could be the upper bound of inequality of opportunity and it could also be a more accurate measure of inequality of opportunity compared to 3FMM and OLS.

One issue for our results is the incompatibility between the ex-ante and ex-post measures (Fleurbaey and Peragine, 2013). On the one hand, the ex-ante measure is computed by the between-type income inequality. If there are some unobserved circumstances, this measure could be underestimated. On the other hand, the ex-post measure is computed by eliminating the between-class income inequality and it could be overestimated because the variation in effort cannot be fully captured due to data limitations. However, our results show that the ex-ante measure, the one supposed to be underestimated, is larger than the ex-post measure in the two FMMV models, the one supposed to be overestimated. This inconsistent result might relate to the incompatibility between the ex-ante and ex-post measure.

In spite of the incompatibility, both ex-ante and ex-post measures show that IOR is more than 40% when using the MLD in FMMV2 models, which suggests that inequality of opportunity in China accounts for more than 40% of total income inequality in China.

4.6 Conclusion

Since inequality of opportunity requires a decomposition of inequality of outcome in terms of circumstances and effort, one needs to know whether each factor should belong to circumstances or effort when applying this idea. However, scholars have different views and have reached little agreement. Barry (2005) stated that effort should be respected even though it is determined by circumstances; while Roemer (1998) argued that the part of effort determined by circumstances is beyond the individual's control and should be eliminated.

Another difficulty is data imperfection. The lack of appropriate data explains that most of the literature treated circumstances such as gender, ethnicity, birthplace, the educational attainment of parents and the main occupation (Paes de Barros et al., 2009) as exogenous variables, while treating effort as unobserved. However, unobserved circumstances and effort lead to biased measures. As a result, most of the literature claims that their measures are "the lower bound" of inequality of opportunity.

To address these issues, we assume that effort is an unobserved categorical variable — a latent class, in other words. Using a finite mixture model, we estimated each individual's effort level according to the latent class to which he/she is likely to belong to. Although this classification only roughly shows every individual's level of effort (three levels of effort in our cases), it does not rely on a particular set of effort variables. Bourguignon et al. (2007) used the individual's own schooling attainment and a migration dummy as effort variables. Jusot et al. (2013) used health-related behaviours toward smoking, obesity and vegetables consumption as effort variables. In this paper, the latent class assumed as effort does not represent any specific variables. As a result, when we use the ex-post approach, the measures are more likely to represent inequality when the overall effect of effort is removed.

One issue for a latent class model is that it is difficult to know exactly what a latent class represents. Our FMM models identify three latent classes which might represent either different levels of effort or even unobserved circumstances. To clarify the identification of latent classes, we used the finite mixture model with varying prior probabilities that allows the prior probabilities (the latent class) to be influenced by a set of factors $\mathbf{z}$. These factors can be circumstances or effort variables. Our results showed that the identification of latent classes in the FMM model with constant prior probabilities is similar to that in the FMMV2 model in which the latent class is determined by pure effort. This result indicated that a latent class model is effective in identifying different levels of effort even though effort is unobserved.

With the ability to identify effort without observing it, this model makes the application of the ex-post approach feasible. Most literature uses the ex-ante approach because of the difficulty of observing effort. However, inequality of opportunity given the ex-ante approach shows only one particular perspective — the inequality when effort has not been exerted. Relying on this approach provides no insight into the role effort plays in income inequality and how effort interacts with circumstances. In this paper, we try to address these questions using the latent class model.

Our findings show a lower IOR based on the ex-post measure than the one obtained on the basis of an ex-ante measure. These differences between the ex-ante and ex-post measure have been addressed by other researches (Fleurbaey and Peragine, 2013). Our results confirmed that when we used more advanced models, we captured more variations so that the ex-ante measure increased, and the ex-post measure decreased. The ex-post measure turned out to be lower than the ex-ante measure.

Benefiting from the estimate of effort levels, our finding confirms the heterogeneous effect of circumstances on income across different levels of effort. Inequality of opportunity turns out to be the largest if everyone exerts the middle level of effort. This finding implies that the group with the middle level of effort should be compensated more than other groups.

This idea of categorizing effort into several classes has also some limitations. The finite mixture model cannot support too many components. In this paper, we also measured inequality of opportunity given five components. Although the results are similar to those obtained in a 3-components model, we do not know whether the results would change for 10 components. Future research could develop alternative empirical tools to overcome this obstacle.

Acknowledgement

Preliminary versions of this article were presented in the international workshop — recent advances in productivity and efficiency analysis: theory and practice — organized by the Centre for Efficiency and Productivity Analysis at the Univeristy of Queensland. We thank participants at this workshop for suggestions and discussions. We are also indebted to Prof. Nanak Kakwani, Assoc. Prof. Xavi Ramos, Dr Haishan Yuan and Dr Nicholas Rohde for their insightful comments.

Chapter 5

Essay III: Inequality of Opportunity in the Access to Higher Education in China: Evidence from a High-Ranked University

5.1 Introduction

People have more opportunities to attend college in contemporary China due to the rapid expansion of tertiary education in the last three decades. The number of graduates from higher education institutions with a bachelor's degree increased from 805,000 in 1995 to 6,809,000 in 2015 (NBS, 2016). At the macro-level, the expansion of tertiary education plays an important role in economic growth. Whalley and Zhao (2013) found that 39.4% of economic growth in China during 1999-2008 was contributed by growth in human capital stock.

At the micro level, tertiary education brings high returns for individuals. Awaworyi and Mishra (2014) conducted a meta-analysis and found that college education and above in China leads approximately to a 14% increase in individual income, more than other levels of educations do.

However, the opportunity to receive tertiary education might not be equally distributed. If the opportunities provided by the expansion favour those with a certain gender, ethnicity and family background, this unequal distribution of opportunities might drive educational inequality and lower economic and social mobility.

This paper uses the framework of equal opportunity (Roemer, 1998, Cohen, 1989 and Arneson, 1989) to examine the fairness of the expansion of tertiary education in China. This framework has been widely used in studying income (Checchi and Peragine, 2010 and Paes de Barros et al., 2009) and health (Jusot et al., 2013) inequality. It regards inequality due to factors beyond individuals' responsibilities ("circumstances") as unfair and accepts inequality due to factors within individuals' responsibilities ("effort"). Applying this framework, we examine to what extent college admission and graduate outcomes are determined by circumstances.

Some researchers have used the framework of equal opportunity to study educational inequality. Ferreira and Gignoux (2014) developed a measure to capture the effect of circumstances on students' academic achievements. This measure has been used to examine inequality of opportunity in education in different regions and countries, e.g. Gamboa and Waltenberg (2012) for Latin America and Salehi-Isfahani et al. (2014) for the Middle East and North Africa. Nevertheless, students' academic achievements are one objective in achieving equal opportunity in education. The access to education is also critical to educational equality of opportunity.

Brunori et al. (2012) examines the access to the tertiary education in Italy base on the framework of inequality of opportunity. In their paper, the educational outcome is defined as a binery outcome with intrinsic ordering (e.g. unemployed after graduation, and work or studied in a university after graduated from a high school). Different from Brunori et al. (2012), we use a multinomial outcome to identify graduate outcomes with no intrinsic ordering.

Researchers have found inequality of opportunity in terms of the access to education in China. For example, children from rural households receive fewer years of schooling than their urban counterparts (Zhang et al., 2015). Rural girls have been found disadvantaged in the access of middle school and high school (Connelly and Zheng, 2003).

Inequality of opportunity could be more obvious in college access. Li et al. (2015) found that rural youth from poor counties were seven times less likely to access any college than urban youth in 2003 and the gap was even larger in access to elite colleges.

In China, college admission is mostly determined by the College Entrance Examination (CEE). It is designed to provide equal opportunity to all high school graduates. However, researchers find an overrepresentation of certain characteristics (e.g. male, urban, etc.) of students in colleges. Wang et al. (2013) showed that rich, Han (the major ethnic group in China) and urban males are overrepresented in colleges. Zeng et al. (2014) found that a female's opportunity to attain college is significantly lower than a male in rural areas.

Although the expansion of tertiary education brings more opportunity to high school graduates, it also increases graduate unemployment (Li et al., 2014). To reduce graduate unemployment, the government expanded postgraduate enrollments. The numbers of entrants for postgraduate degrees rose from 51,053 in 1995 to 645,055 in 2015 NBS (2016), more than 10 times during the two decades, compared to the 8.5 times increase in bachelor degree places. Meanwhile, some college graduates choose to go overseas for a postgraduate degree. Therefore, the expansion of postgraduate study raises the question of whether college graduates have equal opportunity of gaining access to postgraduate study.

This research studies educational inequality of opportunity in China during the expansion of tertiary education and focuses on inequality of access to higher education in two aspects: college admission and graduates' whereabouts. Using the administrative data on economic graduates from a highly-ranked university in China during the period 2008-2015, we investigate students' demographic characteristics to see which individual characteristics were overrepresented and examine the effect of circumstances on graduates' CEE scores. If the enrolment of a university truly depends on the CEE scores, circumstances variables are expected to be uncorrelated with CEE scores. Deviation from this expectation could be evidence of inequality of opportunity, or some policies such as affirmative actions that target particular groups. We also study what factors influence graduates' whereabouts when seeking employment, or when continuing a domestic or oversea postgraduate degree. Opportunity is equal when graduates' whereabouts are independent of circumstances after controlling for their academic achievement.

In this study, we particularly focus on the difference in college admission and graduates' whereabouts between rural and urban areas. Researchers reported gaps between rural and urban areas in income (see Kanbur and Zhang, 1999, Sicular et al., 2007 and Whyte, 2010) and education (Qian and Smyth, 2008). These gaps are one of the main determinants of rising income inequality in contemporary China (Xie and Zhou, 2014). To study the rural-urban difference, we make use of students' Hukou status.

Hukou is a household registration system in China. Households' permanent residential address, referred to as "Hukou address", is recorded by the government. These addresses are divided into agricultural and non-agricultural Hukou based on whether the Hukou address is in rural or urban areas.

This division causes some restrictions on the rural population. For example, rural households could have limited access to education and employment (Liu, 2005) in urban China. Their children are only allowed to take college entrance exams in their domicile based on *Hukou.* These restrictions prevent rural householders from bringing their children to urban areas to receive education even though they find jobs there, and this might be one reason (Biao, 2007) that more than 20 million children (NBS, 2016) were left behind in rural areas by 2015. Due to these restrictions, the Hukou system has a structural effect on income inequality (Whalley and Zhang, 2007) and educational inequality (Afridi et al., 2015) between rural and urban areas. In this study, we examine how the Hukou system affects college admission and graduates' whereabouts using the administration data from one university in China.

Our study confirms that there is a widening rural-urban inequality in tertiary education. The proportion of rural students decreases dramatically between 2008 and 2015 in the examined university. Rural graduates also have fewer opportunities after their graduations than their urban counterparts. The household income gap between rural and urban areas might explain this rural-urban inequality. A higher household income can positively affect a student's chance to enrol in a highly-ranked university even though the student has a lower entrance exam score. It can also provide more opportunities for students after graduation.

The rest of this essay is presented in the following sections, as follows. Section 5.2 introduces the tertiary education system in China. Section 5.3 shows the estimation strategy. Section 5.4 describes the data used in this study. In section 5.5, we present and discuss the result and section 5.6 is the conclusion.

5.2 The College Admissions System in Contemporary China

The tertiary education system in China was reformed and started to recover in 1978 after the Cultural Revolution. It relies on the College Entrance Examination (CEE), known as *gaokao*; that is a decisive test for entrance into almost all tertiary education institutions. In this section, we explain the admission system including CEE and other related policies that affect students' opportunities to receive tertiary education in China.

In China, every child should receive nine-year compulsory education including five years in elementary school and four years in junior middle school. After graduating from junior middle school, graduates can choose to work, learn a specific skill in a technical or vocational school, or study in a high school. Those who enrolled in a high school would take part in CEE in the final year of their study in the high school.

In general, CEE comprises "3 + X" subjects, where "3" refers to three compulsory subjects, including "Chinese, Mathematics and English" and "X" stands for a subject students can choose by themselves. In the high school, students are required to choose between two courses of study: art or science. Those who choose the art stream can select a major from Political Sciences, History, and Geography while those who choose science can select a major from Physics, Chemistry, and Biology. This difference in majors could result in a university having different entry score requirements for students from art and science. In our dataset, students are classified as art or science students, depending on which subject they choose in the CEE.

Students can obtain an exemption or bonus points in the CEE. For example, some students could receive an exemption and a direct admission without the CEE due to their outstanding academic performances in high school. Minorities, foreign nationals, and those who have achieved distinguished results in science and technology competitions, sports competitions and art competitions can also receive additional points in the CEE.

After students take the CEE, they receive their CEE scores within several days. After they receive their CEE scores the students complete the application form with a list of ordered preference. The application form is classified into several tiers[1]. Students are aware of the cutoff score for each tier; however, they do not know the exact entry score for each university. Therefore, they face uncertainty when applying to universities.

Although the purpose of the CEE is to provide a nationwide standardized and merit-

[1] The first tier comprises key universities; the second tier is regular universities; the third is vocational universities.

based exam, the content and form of the exam varies from province to province. For example, the maximum score of the CEE is normally 750. Some provinces and municipalities such as Jiangsu, Shanghai, Zhejiang and Hainan Provinces have different maximum scores ranging from 480 to 900. This is because these provinces have permission to customize their own exams. Until now, 16 provinces and municipalities are permitted by the Ministry of Education to adopt customized exams.

In addition, universities set a different quota for each province. They usually give a large quota to students from their home province. For example, around 57% of students in our dataset come from the home province of the university investigated. Due to regional disparities in the CEE and the quota, we only surveyed students from the home province to analyze the impact of individual characteristics on the CEE score.

In the last decade, going abroad to study has become an increasingly popular choice among students. The Chinese Ministry of Education reports that fewer than 200,000 students went abroad in 2008 and this figure has increased to more than 500,000 in 2015. In our study, we consider an overseas postgraduate degree as one of three main options of graduates.

5.3 Data

The data used in this paper come from student administrative records in one of the Project 985 universities[2] in China. The data spans the years 2008 to 2015. This university is one of the top universities in China and it is the best university in its home province. The dataset includes graduates' academic performance, family backgrounds and their whereabouts immediately after graduation.

Table 5.1 presents summary statistics of the data. The data contain 2,565 observations of students in the economics faculty who graduated during 2008 to 2015. Females constitute 57% of all graduates. Ethnic minority accounts for only 4% of the graduates and 57% of students are from the home province; that is, the province where the university is. Graduates who have a single parent make up only 3%. Among those graduates, the majority of them (63%) holds an urban Hukou.

Regarding family background, 1,998 out of 2,565 report their father's annual income, with an average income of 27,335.6 yuan and 1,775 report their mother's income, with an average of 20,847.86 yuan. Household income per capita is reported by 2,354 graduates. We took into account economies of scale and computed household income per effective member, which on average was equal to 27,777.11 yuan. The number of missing variables was the lowest for household income. We therefore decided to use household income per effective member[3] instead of parents' income as one of the proxies for family background.

[2] Project 985 is a project for both central and local governments to provide funds to universities. There are 39 universities sponsored by Project 985; most of them are top universities in China.

[3] The Household income per effective member equals to the total household income divided by the

Table 5.1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Father's income	1,998	27,335.6	32,137.1	0	600,000
Mother's income	1,775	20,847.86	21,803.39	0	360,000
No. family members	2,565	3.62	0.95	2	9
Imputed No. family members	2,565	0.03	0.18	0	1
HHincome	2,278	27,777.11	39,749.4	0	1,247,077
GPA	2,554	83.11	5.34	29.00	94.74
CEES	2,338	611.50	56.16	168.50	785.00
Year of graduation	2,565			2,008	2,015
Female	2,565	0.57	0.49	0	1
Ethnic Minority	2,565	0.04	0.19	0	1
Single parent	2,565	0.03	0.18	0	1
Home Province	2,565	0.57	0.49	0	1
Urban Hukou	2,565	0.63	0.48	0	1
Self-reported poverty	2,565	0.18	0.39	0	1
Major in science	2,565	0.51	0.50	0	1
Direct admission	2,565	0.004	0.06	0	1
Unemployed	1,321	0.12	0.32	0	1
Failed postgraduate exam	783	0.09	0.28	0	1

[1] Home Province is whether the graduates' households are in the same province as the university.
[2] Imputed No. family members is a dummy for the missing value of No. family members.
[3] GPA is Grade Point Average during undergraduate study. CEES is Chinese Entrance Exam Score ranging from 0 to 750.
[4] Self-reported poverty is whether the graduates report that their families are in poverty.
[5] Major in science is whether the graduates majored in science when they were in high school.
[6] Direct admission is a dummy for those who were directly admitted by the university.
[7] Failed postgraduate exam is a dummy for those who failed the entrance exam for postgraduates.
[8] HHincome is household income per effective member.

In China, high school students usually have no income because getting in a top university is highly competitive, which left students with no time for any part-time work. It is quite rare even rarer for a university student to acquire a part-time job. This is because most families provide financial supports to their children until they graduate. Therefore, we assume that household income is exogenous. However, it is possible that to support children for studying in a good university, parents could work harder.

For the number of family members, we treated the missing value as "3" if the graduate was not from a single parent family and "2" if the graduate was from a single parent family because most families have only one child due to the One Child Policy. However, this treatment might understate the number of family members, especially among those from rural areas because rural parents are more likely to have more than one child than their urban counterparts. The average number of family members is 3.62.

In addition, we also collected the information about parents' occupation as shown in Table 5.2. Based on the administrative data, we divided parents' occupation into 11 categories: administrator (those who have administrative jobs such as business executives and government officers), civil servant, farmer, manual worker, other professionals (those who do non-manual work other than teaching), other teacher (those not teaching in university), passed away, retired, self-employed, unemployed and university teacher. Within these 11 categories, farmers and other professionals are the two biggest groups accounting for around 30% of the population; administrators are comprised of 14.7% males and 6.82% females; teachers (including university and non-university teacher) account for more than 10% of the population.

In terms of graduates' performances, the dataset has information on GPA obtained by students at universities, Chinese Entrance Exam Score (CEES) and their major in high school. On average, the students got 83.11 out of 100 for their GPA and 611.50 out of 750 for their CEES[4]. 51% graduates choose science in high school.

Table 5.3 shows graduates' major in university. Since the administrative data are only from the School of Economics of the University, the majors shown in Table 5.3 are provided by the School of Economics. The most popular major is finance with 34.22% of the population, and the least popular is insurance accounting for 5.59% of the population.

The dataset also has details for graduates' whereabouts. In the last row of Table 5.1, unemployed is a dummy for the graduates who did not get the job they had applied for or did not pass the entrance exam if they chose to take the entrance exam for a postgraduate degree. The percentage of those unemployed graduates is 10%.

Table 5.4 shows graduates' choices regardless of the outcomes of their choices. For example, some graduates may end up being unemployed even if they decided to look for a job. We put those graduates in the same group as those who successfully found a job.

square root of the number of members in the household

[4] Some students can be awarded extra marks due to their special talent.

Table 5.2: Parents' Occupation

	(1)		(2)	
	Father		Mother	
	No.	Percentage(%)	No.	Percentage(%)
Administrator	348	14.70	150	6.82
Civil Servant	112	4.73	69	3.14
Farmer	596	25.18	680	30.91
Manual Worker	144	6.08	97	4.41
Other Professionals	701	29.62	712	32.36
Other Teacher	211	8.91	257	11.68
Passed Away	24	1.01	9	0.41
Retired	14	0.59	37	1.68
Self-employed	122	5.15	102	4.64
Unemployed	19	0.80	31	1.41
University teacher	76	3.21	56	2.55
Total	2367	100.00	2200	100.00

[1] Administrator is a occupation category for business executives and government officers.

[2] Other Professionals are positions for non-manual work other than teachers.

[3] Other teacher is a category for teachers exluding those who are teaching in university.

Table 5.3: Graduates' Major in School of Economics

	No.	Percentage(%)
Public Finance	367	14.55
Economics	208	8.25
Finance	863	34.22
Financial Engineering	333	13.20
Insurance	141	5.59
International_Trade	399	15.82
Mathematical Finance	211	8.37
Total	2522	100.00

[1] These majors listed above belong to the School of Economics in the univerity we studied.

In Table 5.4, half of the graduates choose to work; 30.53% graduates decide to continue their postgraduate studies in China and 13.80% to do their postgraduate study overseas.

Table 5.4: Graduates' Choices

	No.	Percentage(%)
Domestic postgraduate	783	30.53
Not Sure	107	4.17
Oversea postgraduate	354	13.80
Work	1321	51.50
Total	2565	100.00

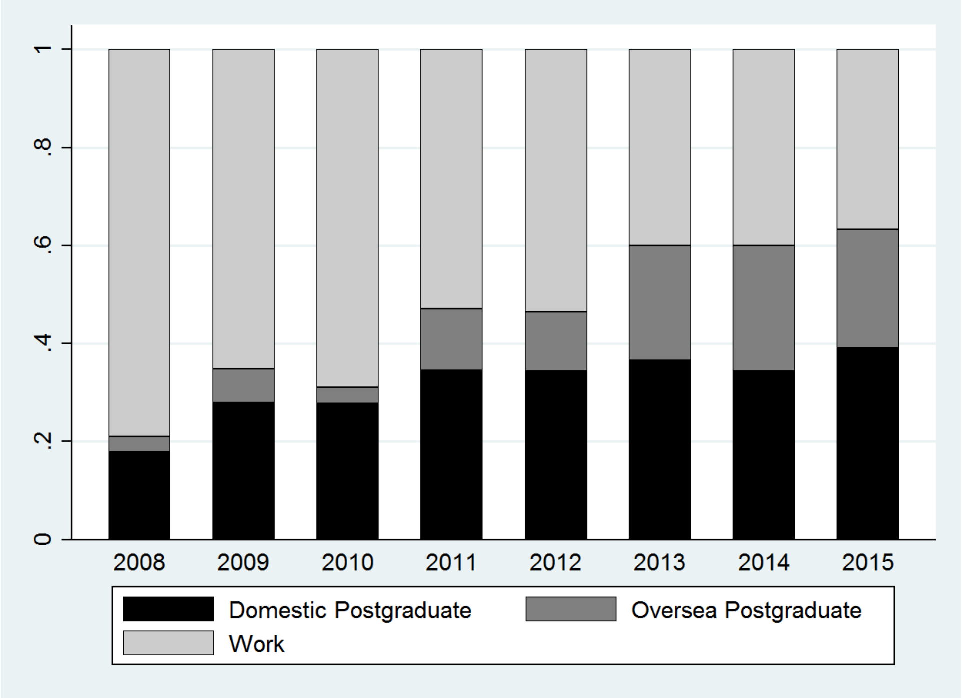
[1] Graduates choosing domestic postgraduate include those who had been successfully enrolled in universities in China for postgraduate programs and those who had failed National Postgraduate Entrance Examination.

[2] Graduates choosing oversea postgraduate include those who reported themselves studying oveaseas or intending to study oveaseas.

[3] Graduates choosing work include those who reported their new jobs or reported themselves looking for jobs.

Figure 5.1 is the stacked bar chart of graduates' choices by year and by percentage. The proportion of graduates doing overseas postgraduate increased sharply after 2012.

Figure 5.1: The Stacked Bar Chart of Graduates' Choices by Year by Percentage



Source: Authors' calculation.

In summary, this section gives a brief introduction on the administrative data we used. It provides an overview of graduates' family backgrounds, academic performances and their choices after graduates. These data are used in this paper for studying how students' family backgrounds affect their educational outcomes.

5.4 Estimation Strategy

We use a multinomial model to identify the relationship between individuals' choices and their individual characteristics. The probability that individual i chooses option j is (McFadden et al., 1973):

$$\pi_{ij} = Pr(Y_i = j) = \frac{\exp(\mu_j + \beta_j \mathbf{X}_i)}{1 + \sum_{j=1}^{J} \exp(\mu_j + \beta_j \mathbf{X}_i)} \tag{5.1}$$

The probability, based on the theoretical framework for discrete choice, is the standardised utilities. We make use of the criterion of the second stochastic dominance to check whether equal opportunity prevails between rural and urban students, and between different parents' occupations.

Using administrative data from the university, we identified three types of graduates' choices: work; domestic postgraduate; and overseas postgraduate. These choices can be decided by graduates in three different ways. First, graduates can view the three choices independently. In this case, we adopted a multinomial regression model and tested the assumption of Independence of Irrelevant Alternatives (IIA) using the Small-Hsiao test and the Hausman test with one model including all variables except parents' occupation and the other including all variables except mother's occupation and major.

Graduates can also make their decision on two choices first, given some choices having similar attributes. For example, graduates are supposed to be in China if they choose work or a domestic postgraduate. Thus, they may first decide whether they would like to stay at home or go overseas. Alternatively, graduates would study if they choose a domestic postgraduate or an overseas postgraduate course. Therefore, they may first decide whether they would like to study or work. In both cases, we use a nested logit model.

If three choices are independent, the multinomial regression model should satisfy the Independence of Irrelevant Alternatives (IIA) assumptions; that is, the exclusion of one alternative should not affect the choices between the other two alternatives. We use the Hausman test and the Small-Hsiao test.

The Hausman test of IIA (Hausman and McFadden, 1984) is defined as:

$$H_{IIA} = (\hat{\beta_R} - \hat{\beta_F^*})'[\hat{Var}(\hat{\beta_R}) - \hat{Var}(\hat{\beta_F^*})]^{-1}(\hat{\beta_R} - \hat{\beta_F^*}) \tag{5.2}$$

where $\hat{\beta_F}$ is the estimate of the full model with all three choices, $\hat{\beta_F^*}$ is a subset of $\hat{\beta_F}$ in which the coefficients are dropped if their corresponding choice have not been estimated in a restricted model and $\hat{\beta_R}$ is the estimate of a restricted model with two choices. H_{IIA} is asymptotically distributed as a chi-square if IIA is true.

An alternative way to test IIA is the Small-Hsiao test (Small and Hsiao, 1985). The statistic of the Small-Hsiao test is from a Likelihood Ratio test which is:

$$SH = -2[L(\hat{\beta}_u^{S_1 S_2}) - L(\hat{\beta}_r^{S_2})] \tag{5.3}$$

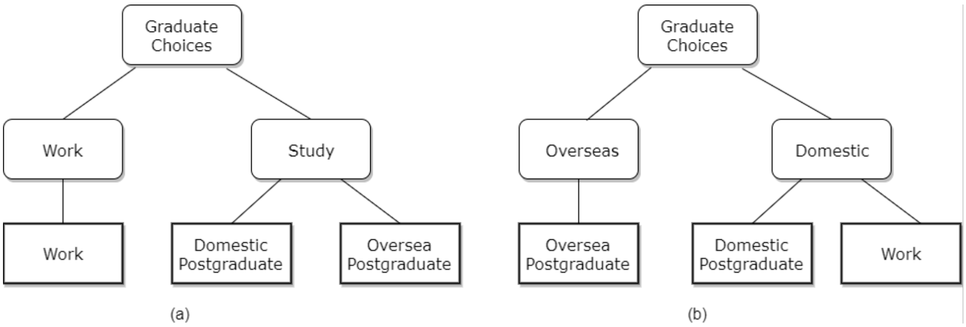
where

$$(\hat{\beta}_u^{S_1 S_2}) = (\frac{1}{\sqrt{2}})\hat{\beta}_u^{S_1} + [1 - (\frac{1}{\sqrt{2}})]\hat{\beta}_u^{S_2} \tag{5.4}$$

The data are randomly divided into two samples S_1 and S_2. $\hat{\beta}_u^{S_1}$ is the estimates of multinomial regression for sample 1. $\hat{\beta}_u^{S_2}$ is the estimates of multinomial regression for sample 2. $\hat{\beta}_r^{S_2}$ is the estimates of the restricted model for sample 2 in which an alternative is excluded. The statistic of the Small-Hsiao test is shown to be asymptotically distributed as a chi-square if IIA is true.

Alternatively, when two choices are correlated with each other, we use the nested logit model. Assume that in the set of all choices J, a nest n contains choices which are correlated with each other with a correlation coefficient $1 - \lambda$. There are two nests in J such that the two nests are mutually exclusive.

Figure 5.2: The Alternative Tree Structures for Nested Logit Models



Note: Figure (a) shows that individuals first make choices between work and study, while figure (b) shows that individuals first make choices between domestic and overseas.

Figure 5.2 shows the alternative tree structures for nested logit models. Individuals could make sequential decisions first between work and study or between domestic and overseas, and then they choose the specific alternative within each nest. Given three choices, the nested logit model releases the assumption of IIA. The probabilities between the nests can be decomposed into the probabilities between nests and the conditional probabilities given nest containing j (Koppelman and Bhat, 2006):

$$\pi_{ij} = \pi_n \pi_{ij|n} = Pr(j \in n)Pr(Y_i = j | j \in n) \tag{5.5}$$

where the within-nest probability is determined by factors X that vary within nests — factors affecting choosing between domestic study and oversea study within "study" or factors affecting choosing between work and domestic study within “domestic”:

$$\pi_{ij|n} = \frac{\exp(\alpha_0 + \boldsymbol{\alpha_j} X_i)}{\sum_{j \in n} \exp(\alpha_0 + \boldsymbol{\alpha_j} X_i)} \tag{5.6}$$

The between-nest probability is determined by factors Z that vary across nests — factors determining choosing between work and study or factors determining choosing between domestic and overseas:

$$\pi_n = \frac{\exp(\gamma_0 + \gamma_j Z_i + \lambda IV_n)}{\sum_{n \in J} \exp(\gamma_0 + \gamma_j Z_i + \lambda IV_n)} \tag{5.7}$$

where IV_n is the inclusive value of nest n:

$$IV_n = \ln \sum_{j \in n} \exp(\alpha_0 + \boldsymbol{\alpha_j} X_i) \tag{5.8}$$

Using the nested logit model, we can not only release the assumption of IIA but also examine the hypothesis of graduates choosing their whereabouts based on a sequential decision-making process.

5.5 Results

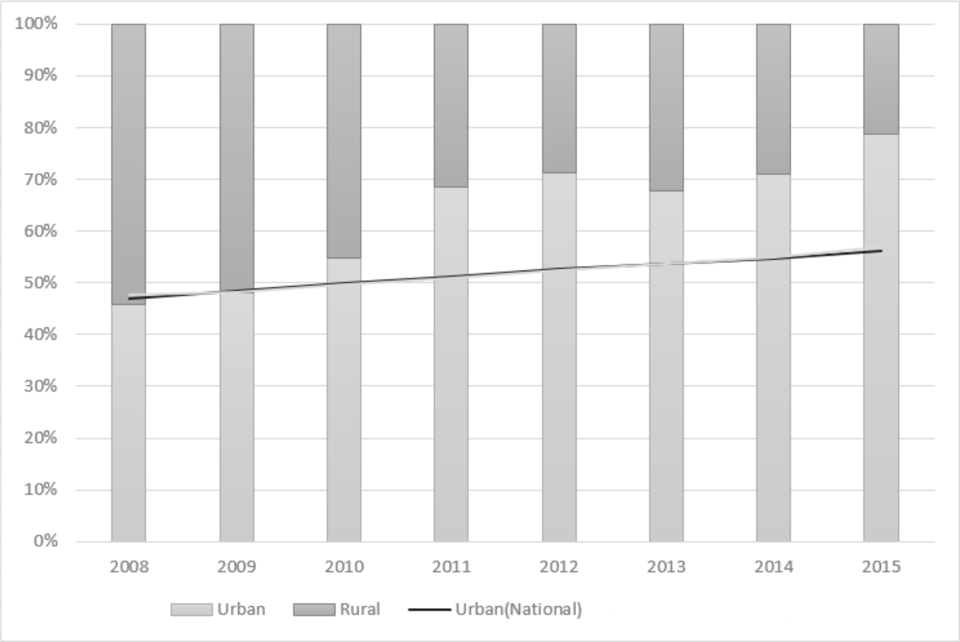
5.5.1 Rural vs. Urban Graduates

In this section, we examine the impact of Hukou status on graduate outcomes. First, we show the percentage of rural and urban Hukou by year in Figure 5.3. The rural Hukou graduates decrease dramatically from more than 50% in 2008 to less than 25% in 2015; while the percentage of rural residents, i.e. the urbanization rate in China only sightly declines from more than 50% to less than 50% during 2008-2015 (NBS, 2016). The sharp decrease in the proportion of rural graduates implies fewer opportunities for rural students to be admitted to this highly-ranked university.

We also compare GPA, entrance exam score, household income and numbers of family among rural and urban graduates (Table 5.5). Rural graduates on average live in households with 4.178 members which is more than the average among urban graduates (3.3 members). Although rural graduates had slightly higher CEES than urban graduates, their GPA scores are lower than their urban counterparts. The difference is not significant in CEES but significant in GPA.

Figure 5.4 shows the difference in average GPA and CEES between urban and rural graduates during 2008-2015. In this figure, a positive value represents a higher average score for urban graduates. The blue line represents the trend of the difference in GPA. The red line represents the difference in CEES including both home and other provinces.

Figure 5.3: Graduates' Hukou Status by Year (Percentage)



Note: 1.The annual data referring to the percentage of urban residents in the home province of the university and in China from 2008 to 2015 is from annual editions of NBS.
2. The stacked chart is based on authors' calculations.

Table 5.5: Rural vs. Urban Graduates

	(1) Rural	(2) Urban	(3) Difference
GPA	82.63	83.67	-1.034***
	(4.993)	(5.050)	(-4.90)
CEES	614.4	610.4	4.073
	(51.15)	(57.63)	(1.68)
HHincome	12432.4	36473.2	-24040.77***
	(15577.5)	(46141.3)	(-14.49)
No. Family	4.178	3.300	0.878***
	(1.076)	(0.695)	(24.54)

[1] "GPA" is Grade Point Average, ranging from 0 to 100.

[2] "CEES" is the score from the Chinese Entrance Exam, ranging from 0 to 700.

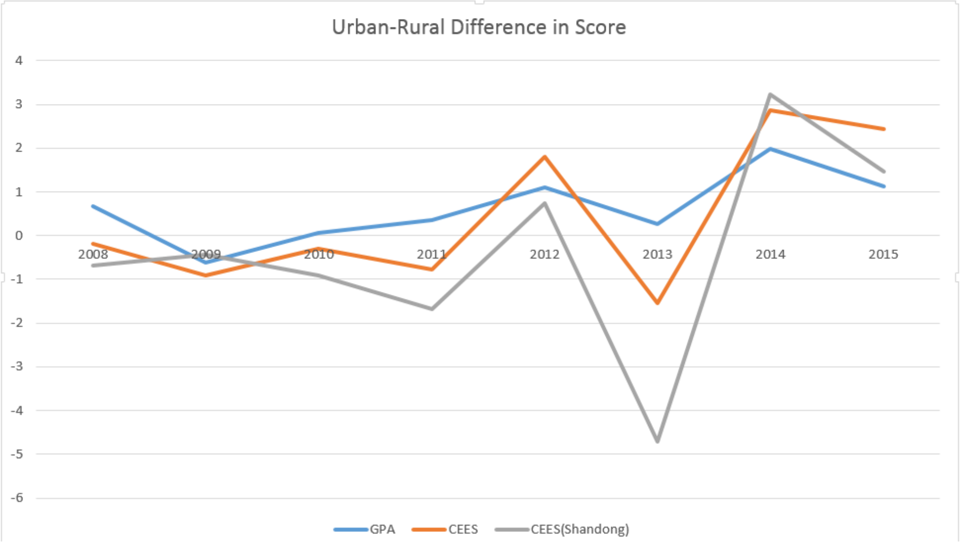
[3] "HHIncome" represents household annual income per effective household member and "No. Family" is the number of family members.

[4] For column (1) and (2), the coefficients represent the mean values and the standard errors are in parentheses.

[5] For column (3), the coefficients represent the difference of mean between rural and urban graduates. t statistics in parentheses: *p<0.1; **p<0.05; ***p<0.01.

The grey line represents the difference in CEES excluding other provinces. In general, CEES and GPA have similar trends during the period. This indicates that a student with a higher CEES is more likely to have a higher GPA in the university as well. In addition, the score gap between urban and rural graduates increased during 2008-2015 except in year 2013 even though the proportion of rural graduates decreased during that period.

Figure 5.4: Difference in GPA and CEES between Urban and Rural Graduates



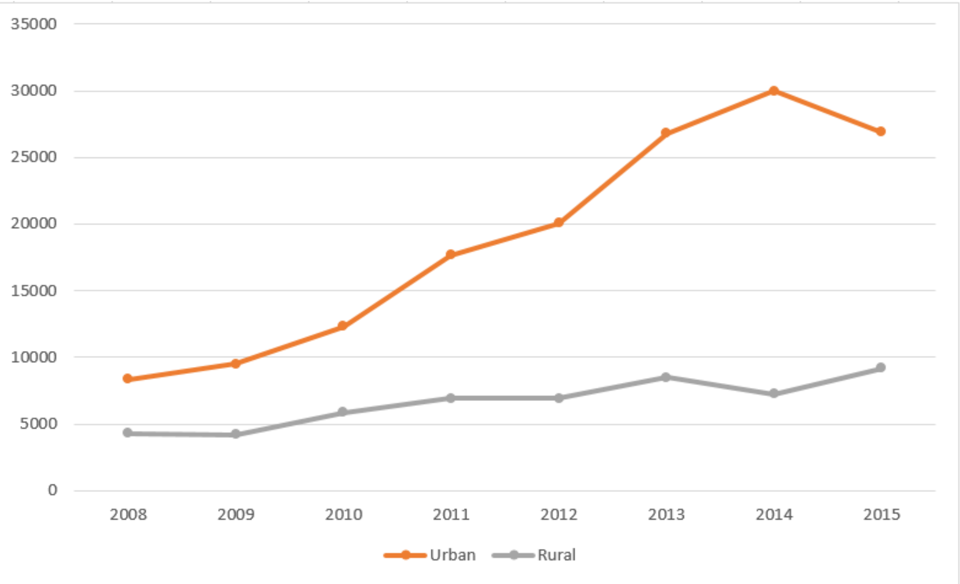
Note: A positive value represents a higher average score for urban students.
Source: Authors' calculation.

The biggest difference is in household income. Urban graduates' households earn more than three times as much as their rural counterparts. Figure 5.5 shows the change in income gap between rural and urban households by year. In 2008, urban graduates' households earned around two times as much as their rural counterparts. This gap became more than three times higher in 2015.

The difference between rural and urban graduates can also be found in their whereabouts. Figure 5.6 shows graduates' whereabouts by Hukou status during 2008-2015 excluding those who were not sure about their whereabouts. In 2008, the percentage of graduates who continued their studies in China is around 20% and 15% for urban and rural graduates respectively. The percentage of overseas postgraduates is less than 5% for both Hukou statuses. The percentage of domestic postgraduates increases at the same pace for students from rural and urban areas during 2008-2015. In contrast, the percentages of overseas postgraduates only increase for graduates from urban areas. During the whole period, only 2.13% of rural graduates choose to undertake a postgraduate course overseas compared to 21.39% urban graduates who choose an overseas postgraduate course. This difference in whereabouts could indicate more opportunities available for urban graduates than their rural counterparts.

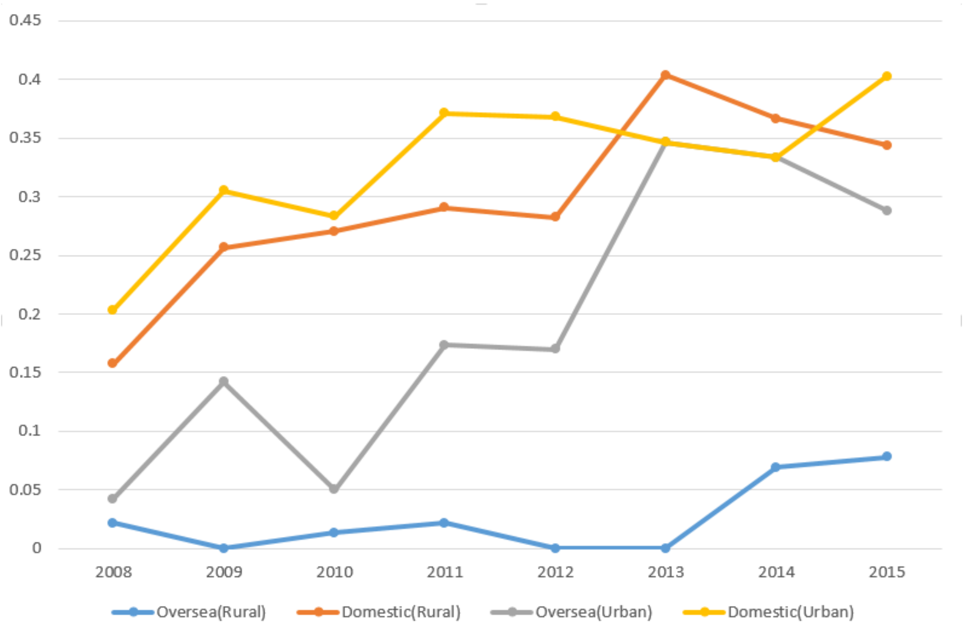
Given all these differences between rural and urban graduates, we use multivariate regression analysis in the next two sections to examine whether graduates' academic

Figure 5.5: Graduates' Household Income by Hukou Status



Source: Authors' calculation.

Figure 5.6: Postgraduates by Hukou in Graduates' Whereabouts During 2008-2015 (Percentage)



Note: Those who were not sure about their whereabouts are excluded in this figure.
Source: Authors' calculation.

performance and whereabouts are affected by their Hukou and other factors.

5.5.2 Regression Analysis on College Admission

To study the relationship between circumstances and college admission, we regress the CEE score (CEES) on circumstances. Suppose that before knowing the entry score of this university, there are N students applying for this university. Given those students' circumstances — the vector $\mathbf{c}$, the relationship between circumstances and students' CEES can be described in the following equation:

$$CEES_i = \alpha \mathbf{c}_i + \epsilon \tag{5.9}$$

where α is a coefficient for circumstance variables and ϵ is an error term. Since the university sets an entry score for admission, students' CEES should not be dependent on their circumstances. Therefore, the coefficients should be equal to 0.

However, circumstances could affect students' CEES in the following three ways. First, circumstances could affect CEES if the university sets different entry scores for different students. For example, students could have a lower entry score if they belong to a minority group. In addition, the entrance scores for students from art and science can be different.

The second way is that students' circumstances could directly affect their scores. For example, a gender gap has been found in PISA test scores (González de San Román and De La Rica, 2012). If distributions for each gender have the same variance but different means for all applicants, the difference might also be found after admission. To check this type of influence attributable to students' circumstances, we also regress students' GPA on their circumstances. GPA represents the academic performances after admission. It has no relationship with the admission and the selection problem. Therefore, the relationship between circumstances and GPA is more likely to be the direct influence of circumstances on academic performances.

The last way is that students with a certain circumstance might be more risk-taking than other students. If two distributions for two different circumstances have the same mean but different variances, a higher average score for the higher-variance distribution could be found after admission. Students have been given their scores before applying to colleges but they have not been given the entry score of each university until a couple of days after they submit the college application form. Students with a lower score could face more uncertainties when they apply to a top university. In consequence, their circumstances might play a role in their risk-taking preferences.

Table 5.7 shows the results of regressing GPA and CEES on a number of socioeconomic variables. We use GPA as dependent variables in Columns (1) to (3) and CEES as dependent variables in Columns (4) to (6). Since each province has different exam questions and the university gives different quotas to different provinces, we only use the observations from the home province in Columns (4) to (6). The summary statistics for the home province is presented in Table 5.6. We rescale CEES from 0 to 100, the same scale as the GPA.

Table 5.6: Summary Statistics for the Home Province

Statistic	N	Mean	St. Dev.	Min	Max
Father's income	1156	27126.42	29174.21	0	600000
Mother's income	1038	21317.98	19871.03	0	360000
No. family members	1468	3.52	0.86	2	9
Imputed No. family members	1468	0.04	0.2	0	1
HHincome	1352	15678.61	24546.92	0	1247077
GPA	1463	84.32	4.74	29	94.74
CEES	1272	630.18	37.06	168.5	785.00
Year of graduation	1468			2008	2015
Female	1468	0.54	0.5	0	1
Ethnic Minority	1468	0.01	0.11	0	1
Single parent	1468	0.03	0.16	0	1
Urban Hukou	1468	0.62	0.48	0	1
Self-reported poverty	1468	0.16	0.37	0	1
Major in science	1468	0.51	0.5	0	1
Direct admission	1468	0	0.07	0	1
Unemployed	669	0.12	0.32	0	1
Failed postgraduate exam	525	0.06	0.24	0	1

[1] Home Province is whether the graduates' households are in the same province as the university.
[2] Imputed No. family members is a dummy for the missing value of No. family members.
[3] GPA is Grade Point Average during undergraduate study. CEES is Chinese Entrance Exam Score ranging from 0 to 750.
[4] Self-reported poverty is whether the graduates report that their families are in poverty.
[5] Major in science is whether the graduates majored in science when they were in high school.
[6] Direct admission is a dummy for those who were directly admitted by the university.
[7] Failed postgraduate exam is a dummy for those who failed the entrance exam for postgraduates.
[8] HHincome is household income per effective member.

Due to the correlation between graduates' Hukou status and their household income and parents' occupation, we exclude household income and parents' occupation in Columns (1) and (4), add household income in Columns (2) and (5) and include both household income and father's occupation in Columns (3) and (6). In addition, we also include year dummies in each regression to control for time varying heterogeneity such as cohort effects and changes in university policies or practice.

In general, we find that graduates are more likely to achieve a higher GPA if they are female, belong to the major ethnic group, come from the home province or have a higher CEES. For CEES, females belonging to the major ethnic group are also likely to have a higher score.

The average entry score of the minority is around 10 points (out of 100) lower than the majority. This is due to China's affirmative action policies giving preferential treatment to

Table 5.7: The Regression Analysis on GPA and CEE Scores

	(1) GPA	(2) GPA	(3) GPA	(4) CEES	(5) CEES	(6) CEES
Female	2.809***	2.753***	2.716***	0.534*	0.636**	0.599**
	(0.189)	(0.197)	(0.203)	(0.291)	(0.307)	(0.300)
Minority	-2.257***	-1.970***	-1.859***	-9.603*	-9.453*	-10.34**
	(0.532)	(0.540)	(0.556)	(4.995)	(4.967)	(5.249)
Single	0.314	0.418	2.059**	-0.551	-1.145	-0.130
	(0.519)	(0.536)	(0.966)	(1.762)	(1.841)	(1.062)
Home Province	2.277***	2.292***	2.234***			
	(0.206)	(0.215)	(0.221)			
Urban Hukou	0.0252	-0.158	0.00137	-0.836**	-0.334	-0.186
	(0.221)	(0.255)	(0.336)	(0.349)	(0.380)	(0.497)
CEES	0.105***	0.109***	0.104***			
	(0.0181)	(0.0196)	(0.0203)			
No. Family	0.0281	0.0394	0.0116	0.173	0.0476	0.0452
	(0.114)	(0.116)	(0.119)	(0.209)	(0.219)	(0.204)
Missing No. Family	0.222	0.545	0.712	-1.668*	0.979	0.664
	(0.600)	(2.109)	(2.186)	(0.912)	(1.179)	(1.147)
Science	0.237	0.227	0.260	1.675***	1.678***	1.792***
	(0.203)	(0.210)	(0.215)	(0.268)	(0.275)	(0.281)
Log HHincome		0.184	0.115		-0.659***	-0.582**
		(0.125)	(0.143)		(0.156)	(0.198)
Retired			-2.811**			-1.123
			(1.220)			(2.289)
Constant	68.99***	66.99***	68.40***	86.36***	92.32***	90.91***
	(1.631)	(2.131)	(2.376)	(1.072)	(1.857)	(2.657)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Major	Yes	Yes	Yes	No	No	No
Father's Occupation	No	No	Yes	No	No	Yes
No. Significant Occupation			1		0	
Observations	2229	2068	1964	1243	1168	1126
Adjusted R^2	0.264	0.264	0.260	0.130	0.131	0.137

[1] Standard errors in parentheses.

[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

[3] All observations are used in Columns (1) to (3). Only observations from Home Province Province are used in Columns (4) to (6).

[4] GPA is Grade Point Average during undergraduate study. CEES is Chinese Entrance Exam Score transformed to 0 to 100.

minorities. Also, difference in CEES between science and art majors is due to universities in China setting different entry scores for science and arts students.

Interestingly, household income has a negative effect on CEES even though there is no policy for the university to explicitly give special considerations to those with lower household income. In addition, the coefficient of Hukou status is significant if household income is excluded in the model. This implies that the effect of Hukou status can be explained by the difference in household income. Different from CEES, GPA is not affected by Hukou status and household income. The coefficients of Hukou status and household income are not significant in all cases. Regarding father's occupation, only the coefficient of being a university teacher is significant, which means that graduates whose father is a university teacher are more likely to have a higher GPA.

Given these results, we cannot conclude that students with a higher household income have a low CEES on average because GPA does not show the same relationship with household income. This negative relationship might be due to the fact that students from rich families might have better options than those from poor families given the same CEES. For example, if students from rich families fail to enter the top university, they can choose to go overseas, paying more money to buy better education; while students from poor families cannot afford to do the same and thereby have to accept going to a low-rank university. Therefore, the negative relationship appears if students with better household economic conditions are more likely to apply to top universities given a lower CEE score.

To check this, we ran a quantile regression on CEES at 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles. The results are shown in Table 5.8. We found that the lower the quantile, the bigger the negative effects of household income. This could imply that the closer the students' CEES is to the entry score, the more risk averse the students are if they have a low household income. This result is an indirect evidence of differences in risk-taking performances for students with different household income. This explanation could be more concrete if we can get data from those who fail to be admitted to this university.

One possible explanation of the differences in risk-taking performances is the difference in family expectations given different economic background. Ashraf et al. (2017) find that both poor and rich families have high expectations on the education of their children. However, it is still not clear how different these expectations are between poor and rich families. Studies on a household survey might be necessary to find out these differences.

In summary, we found that circumstances could affect students' admission in a highly-ranked university. Some circumstances, such as belonging to a minority group, are due to the government's affirmative action policies. Some circumstances directly affect students' CEES scores. For example, females have higher scores on average than males. Most interestingly, students with high household income might have higher risk-taking performance when applying to a highly-ranked university.

Table 5.8: The Quantile Regression on CEES

	(1) 0.1Q	(2) 0.3Q	(3) 0.5Q	(4) 0.7Q	(5) 0.9Q
lninchh_percap	-0.659*	-0.484***	-0.412***	-0.351***	-0.0505
	(0.397)	(0.108)	(0.0925)	(0.114)	(0.0897)
Observations	1168	1168	1168	1168	1168

[1] Standard errors in parentheses.
[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.5.3 Regression Analysis on Graduates' Whereabouts

5.5.3.1 Results from the Multinomial Regression Model

Table 5.9 is the result of a multinomial logit regression model. The baseline choice is work and those who were not sure about their whereabouts are excluded in the regression analysis. The coefficients reported in Table 5.9 are the odd ratio, that is the ratio of the probability of choosing domestic postgraduate (or oversea postgraduate) to that of choosing working.

We also conducted Small-Hsiao and Hausman tests and found that they were not significant in both models, which indicates that the IIA cannot be rejected. Therefore, these choices are likely to be independent of each other.

Table 5.9: Graduates' Whereabouts (Multinominal Regression Model)

	(1)	(2)	(3)	(4)	(5)
Domestic postgraduate					
Urban Hukou	1.523***	1.457***	1.212	1.442*	1.745**
Log HHincome		1.042	0.955	1.021	1.222
Oversea Postgraduate					
Urban Hukou	6.872***	3.212***	1.518	1.959*	1.934*
Log HHincome		3.175***	2.685***	2.594***	4.442***
Year Dummy	Yes	Yes	Yes	Yes	Yes
Mother's Occupation	No	No	Yes	No	No
Father's Occupation	No	No	No	Yes	Yes
Observations	2452	2271	2008	2155	2080
Pseudo R^2	0.239	0.273	0.286	0.289	0.656
AIC	3708.2	3274.3	2942.3	3107.1	5505.4
Loglikelihood	-1822.1	-1603.2	-1417.1	-1499.5	-2702.7

[1] Exponentiated coefficients.
[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.
[3] Coefficients except Hukou status and household income are omitted in this table.
[4] Column (1) represents the baseline model. Column (2) to (4) show the results from the alternative models where variables such as household income and parents' occupation are included. Column (5) is the model using the inverse propensity score weight.

In Column (1) we exclude household income and parents' occupation and add household income in Column (2). We use father's occupation and mother's occupation in

Columns (3) and (4) respectively.

Comparing the R-squared statistics, we find that including household income significantly improves the R-squared statistics. This implies that household income is crucial for understanding graduates' whereabouts. In contrast, including parents' occupation slightly increases the R-squared statistics. In addition, we find that Columns (3) and (4) have similar R-squared statistics, which means that father's and mother's occupation help explain similar variation of graduates' whereabouts. Given the log likelihood and Akaike information criterion (AIC), we find that the model with household income and mother's occupation performs better than other models.

Regarding graduates' Hukou status, urban Hukou graduates are more likely to choose doing postgraduates than their rural counterparts. The coefficients of urban Hukou are significant at the 5% level in Columns (1) and (2) but not in Columns (3) and (4). This might be due to the correlation between urban Hukou and other circumstances.

Since Hukou status is correlated with other circumstances, the difference in graduates choices between the two Hukou statuses does not reflect the pure effect of Hukou status. If we consider the effect of Hukou status as a treatment effect, respondents in the dataset are not randomly assigned to two groups classified according to their Hukou status. Therefore, to show the pure effect of Hukou status on graduates' whereabouts, we weighted the regression by the inverse propensity score. We calculated the propensity score based on the following logistic regression:

$$P(hukou = 1|\mathbf{z}) = \sigma(\boldsymbol{\theta}'\mathbf{z}) \tag{5.10}$$

where *hukou* is a dummy variable for Hukou status. It equals 1 when Hukou status is urban. The logistic function is $\sigma()$, $\boldsymbol{\theta}$ is the coefficients and $\mathbf{z}$ is a vector of variables related to Hukou status. We use female, minority, the number of family members, father's occupation and household income per effective members in $\mathbf{z}$.

The predicted probability of urban Hukou given $\mathbf{z}$ is the propensity score p. Then the inverse propensity score weight is:

$$ipw = \begin{cases} \frac{1}{p} & \text{if } hukou = 1 \\ \frac{1}{1-p} & \text{if } hukou = 0 \end{cases} \tag{5.11}$$

The results from the weighted regression are presented in column (5).

After weighting the regression by the inverse propensity score, the coefficients of urban Hukou become significant again. According to this weighted regression, we find that urban Hukou graduates are 1.745 times more likely to choose domestic postgraduates and 1.934 times more likely to choose oversea postgraduates than their rural counterparts respectively. This indicates that Hukou status plays an important role in the choice of a postgraduate degree.

Household income also affects graduates' choices. The results show that a 1% increase in household income would increase the probability of choosing oversea postgraduates by at least 2.594 times, although this increase is not significant and has no impact on the probability of choosing a domestic postgraduate course. This difference suggests that a higher household income brings a graduate more opportunity to pursue an oversea study but has no effect on the choice of domestic postgraduate.

Since there are only 12 rural graduates who chose to study overseas, we did regressions on rural and urban graduates respectively and merged domestic and oversea postgraduate into one category. The results of graduates choosing postgraduate are shown in Table 5.10. For urban graduates, we assume either a three-choice model using a multinomial regression (Column (1) and (2)) or a two-choice model using a logistic regression (Column (3) and (4)); while for rural graduates (Column (5) and (6)), we only assume a binary choice between work and postgraduate study. Since we focus on choices between work and postgraduate study in Table 5.10, we omit the results from oversea postgraduate study in column (1) and (2).

The results in Table 5.10 show that for urban graduates, the two-choices model has lower AIC and loglikelihood than the three-choices model. This indicates that the two-choices model performs better than the three-choices model. Comparing the explanatory variables, we find that some factors influencing a postgraduate study are different between rural and urban graduates. The effect of GPA on a postgraduate degree is higher for rural than urban graduates. Rural graduates are 38% to 41% more likely to choose a postgraduate if their GPAs increase by 1%. In contrast, the figures for urban graduates are around 26%. In addition, the number of family members only affects urban graduates but not rural graduates. Urban graduates are 22% to 27% less likely to choose a postgraduate if they have one or more family members.

Household income and parents' occupations also have a different impact on choosing a postgraduate degree depending on whether graduates are rural or urban. Rural graduates' whereabouts are more related to the father's occupation. Rural graduates are more likely to choose a postgraduate if the father is a teacher. Given the father's occupation, the coefficient of household income becomes not significant. For urban graduates, household income is still a significant factor even though father's occupation is given.

Mother's occupation is more related to urban graduates' whereabouts than rural graduates' whereabouts. Some occupations such as administrator, professionals and teachers including university teachers are highly significant and positively related to the choice to undertake a postgraduate course. These effects are not found in urban graduates.

In addition, some occupations of urban parents are not found among rural counterparts. No rural parent is a civil servant, university teacher or retired. This might indicate that rural parents have a lower social status and welfare than their urban counterparts.

In summary, given the results from the multinomial regression and the logistic regres-

Table 5.10: Graduates' Choosing Domestic Postgraduate (Rural vs. Urban)

	(1) MNL Urban	(2) MNL Urban	(3) Logit Urban	(4) Logit Urban	(5) Logit Rural	(6) Logit Rural
Female	0.943	0.889	0.979	0.931	0.873	0.860
Minority	0.917	1.034	1.367	1.514	0.974	1.095
Single	1.179	0.395	0.993	0.524	0.446	1.425
Home Province	1.273	1.167	1.315**	1.228	0.99	1.055
GPA	1.357***	1.377***	1.267***	1.265***	1.418***	1.381***
No. Family	0.805*	0.88	0.730***	0.773**	1.029	0.990
Missing No. Family	6663004.4				3.748	2.493
Log HHincome	0.991	0.777*	1.351***	1.215*	1.164	1.280*
Parents Occupation	Father	Mother	Father	Mother	Father	Mother
Administrator	1.167	2.672*	1.238	2.600*	1.109	0.598
Civil Servant	0.71	1.739	1.062	2.182		
Manual Worker	1.092	2.249	0.855	1.943	1.039	1.583
Other Professionals	1.222	3.331**	1.165	2.905**	0.842	1.228
Other Teacher	1.339	4.108***	1.033	2.873**	2.931**	1.471
Passed Away	7.92E-008	4.172	0.169	2.834	2.577	1.014
Retired	3.096	3.129	1.835	3.086*		
Self-employed	1.606	1.967	1.058	1.357	0.979	0.559
Unemployed	0.246	0.805	0.24	0.862	0.588	3.577
University teacher	1.553	12.48***	2.078	10.24***		
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1378	1270	1377	1270	775	735
Pseudo R^2	0.245	0.239	0.253	0.254	0.274	0.257
AIC	2303.6	2161.5	1463.1	1344.7	738.6	721.3
ll	-1099.8	-1030.7	-706.6	-647.4	-346.3	-337.6

[1] Exponentiated coefficients.

[2] "MNL" stands for the multinomial logit regression and "Logit" represents the two-choice logit regression.

[3] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

sion, we find that Hukou status and parents' socioeconomic status have large effects on graduates' whereabouts. A urban graduate with rich parents or parents who are teachers is more likely to study towards a postgraduate degree. In addition, an oversea postgraduate is different from a postgraduate in the home country. An oversea postgraduate heavily relies on household income while a domestic postgraduate does not and urban graduates have more opportunities to study overseas than their rural counterparts.

5.5.3.2 Results from the Nested Logit Model Given the Whereabouts are Correlated

Table 5.11 shows the results from the nested model. Graduates first make choices between nests — work and study (Columns (1) and (3)), or staying in China and going oversea (Columns (2) and (4)), and the final within-nest decision is made based on either household income (Columns (1) and (2)) or GPA (Columns (3) and (4)).

Table 5.11: Graduates' Whereabouts (Nested Logit Model)

	(1)	(2)	(3)	(4)
	Postgraduate	Domestic	Postgraduate	Domestic
Female	1.032	0.805	1.034	0.806
Minority	1.257	0.521**	1.259	0.520**
Single	0.791	1.790	0.783	1.780
Home Province	1.325***	0.835	1.327***	0.836
No. Family	0.807***	2.710***	0.806***	2.698***
Missing No. Family	2.445	0.272	2.427	0.271
Urban Hukou	1.592***	0.397***	1.597***	0.396***
GPA	1.254***	1.013		
Log HHincome			1.267***	0.425***
Domestic postgraduate				
Log HHincome	0.134***	0.297***		
GPA			0.756***	0.974
Work				
Log HHincome	0.217***	0.530***		
GPA			0.663***	1.038**
Observations	6801	6801	6801	6801
AIC	3708.0	3976.2	3673.1	3991.5
Loglikelihood	-1836.0	-1970.1	-1818.6	-1977.8

[1] Exponentiated coefficients.
[2] * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
[3] Year Dummies are used as independent variables in each regression.

Comparing AIC and log likelihood for each model, we find that models with nests between work and study perform significantly better than models with nests between staying in China and going overseas. This might indicate that graduates are more likely to choose between study and work first and consider domestic and oversea study after their first decision.

Regarding explanatory variables, household income is an important factor in all model specifications. Higher household income is especially important for those who choose going overseas for postgraduate study. Another crucial factor is urban Hukou which positively affects the choice of study. These findings are consistent with the results from the nested logit model and the multinomial regression model.

The results from GPA are different for the nested logit model compared to the multinomial regression model. GPA is not significant if graduates make decisions between domestic and overseas first. This implies that GPA or personal academic performances have little impact on the decision to go overseas. Instead, it is family backgrounds and economic wellbeing that determine the decision to go overseas.

Generally, the nested logit model show results similar to the multinomial regression model, which increases the robustness of our main findings. In addition, the nested logit model indicates that graduates are more likely to make a decision between work and study than between domestic and overseas postgraduate study.

5.5.4 The Effect of Hukou Status and Parents' Occupation on Graduates Choices: An Alternative Approach

From the regression analysis, we found that Hukou status and parents' occupation are two determinants of graduates' choices. In this subsection, we identified the effect of Hukou status and parents' occupation on graduates choices using the stochastic dominance approach (Lefranc et al., 2008). Assume that graduates' choices follow the criteria of first-order stochastic dominance (FSD) and second-order stochastic dominance (SSD). The definitions of stochastic dominance are as follows:

Definition 5.5.1 *For two different vectors of circumstances* $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ *where* $\mathbf{c} \neq \mathbf{c}'$, $\mathbf{c}$ *first-order stochastic dominance (FSD)* $\mathbf{c}'$ *for choice* j*, i.e.* $\mathbf{c} \succsim_{FSD_j} \mathbf{c}'$ *iff:*

$$F(p_j|\mathbf{c}) \leq F(p_j|\mathbf{c}'), \forall p_j \in \mathbb{R}_+ \tag{5.12}$$

where $F(p_j|\mathbf{c})$ is the cumulative distribution of probabilities p_j of choosing option j. $\mathbf{c}, \mathbf{c}'$ are vectors of circumstances belonging to the set of all possible vectors of circumstances $\mathcal{C}$. In other words, $\mathbf{c} \succsim_{FSD_j} \mathbf{c}'$, when graduates with circumstances $\mathbf{c}$ have higher probabilities in choosing j than graduates with circumstances $\mathbf{c}'$.

However, the first-order stochastic dominance does not consider individuals' risk-averse preference. Given two circumstances, individuals might prefer circumstances which provide more predictable probabilities in choosing to choose j than a less predictable probabilities. To take the risk-averse preference into account, we used the second-order stochastic dominance (SSD) (Lefranc et al., 2008)

Table 5.12: Lorenz Dominance Test (Domestic Postgraduate)

Father	Farmer/Worker	Adminstrator	Teacher
Farmer/Worker	-	$<$	$<$
Adminstrator	-	-	$<$
Teacher	-	-	-
Hukou	Rural	Urban	
Rural	-	$<$	
Urban	-	-	
Mother	Farmer/Worker	Adminstrator	Teacher
Farmer/Worker	-	$<$	$<$
Adminstrator	-	-	$=$
Teacher	-	-	-
Hukou	Rural	Urban	
Rural	-	$<$	
Urban	-	-	

[1] > The row dominates the column.
[2] < The row is dominated by the column.
[3] = The row neither dominates nor is dominated by the column.

Definition 5.5.2 *For two different vectors of circumstances* $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ *where* $\mathbf{c} \neq \mathbf{c}'$, $\mathbf{c}$ *second-order stochastic dominance (SSD)* $\mathbf{c}'$ *for choice* j*, i.e.* $\mathbf{c} \succsim_{SSD_j} \mathbf{c}'$ *iff:*

$$\int_0^1 F(p_j|\mathbf{c})dx \leq \int_0^1 F(p_j|\mathbf{c}')dx, \forall p_j|\mathbf{c} \in \mathbb{R}_+ \tag{5.13}$$

The SSD is equivalent to generalized Lorenz dominance (GLD)(Lefranc et al., 2008). Formally:

$$\forall p_j|\mathbf{c} \in \mathbb{R}_+, \mathbf{c} \succ_{SSD} \mathbf{c}' \Leftrightarrow \forall \pi \in [0,1] GL_{F(\cdot|\mathbf{c})}(\pi) \geq GL_{F(\cdot|\mathbf{c}')}(\pi) \tag{5.14}$$

where $GL_{F(\cdot|\mathbf{c})}(\pi)$ is the value of the generalized Lorenz curve at π for the distribution of $F(\cdot|\mathbf{c})$.

Based on the definition of stochastic dominance and the results from multinomial regression, we conducted Lorenz dominance tests. The results are shown in Table 5.12 and Table 5.13. Table 5.12 shows Lorenz dominance tests on the probabilities of choosing domestic postgraduates and Table 5.13 shows Lorenz dominance tests on the probabilities of choosing overseas postgraduates. We use the estimators in columns (3) and (4) of Table 5.9 and three categories of parents' occupations: farmers and manual workers; government officers, administrators and civil servants; and teacher. We find that students with parents who are farmers and manual workers are Lorenz dominated by the other two occupations in choosing both domestic and overseas postgraduates. Teachers dominate the other two occupations in domestic postgraduates but are dominated by administrators in overseas postgraduates.

We also compare two Hukou statuses: rural and urban. We find that rural Hukou is

Table 5.13: Lorenz Dominance Test (Oversea Postgraduate)

Father	Farmer/Worker	Adminstrator	Teacher
Farmer/Worker	-	<	<
Adminstrator	-	-	>
Teacher	-	-	-
Hukou	Rural	Urban	
Rural	-	<	
Urban	-	-	
Mother	Farmer/Worker	Adminstrator	Teacher
Farmer/Worker	-	<	<
Adminstrator	-	-	>
Teacher	-	-	-
Hukou	Rural	Urban	
Rural	-	<	
Urban	-	-	

[1] > The row dominates the column.
[2] < The row is dominated by the column.
[3] = The row neither dominates nor is dominated by the column.

dominated by urban in both domestic and overseas postgraduates. These results suggest that graduates' choices are substantially affected by circumstances such as Hukou status and parents' occupations.

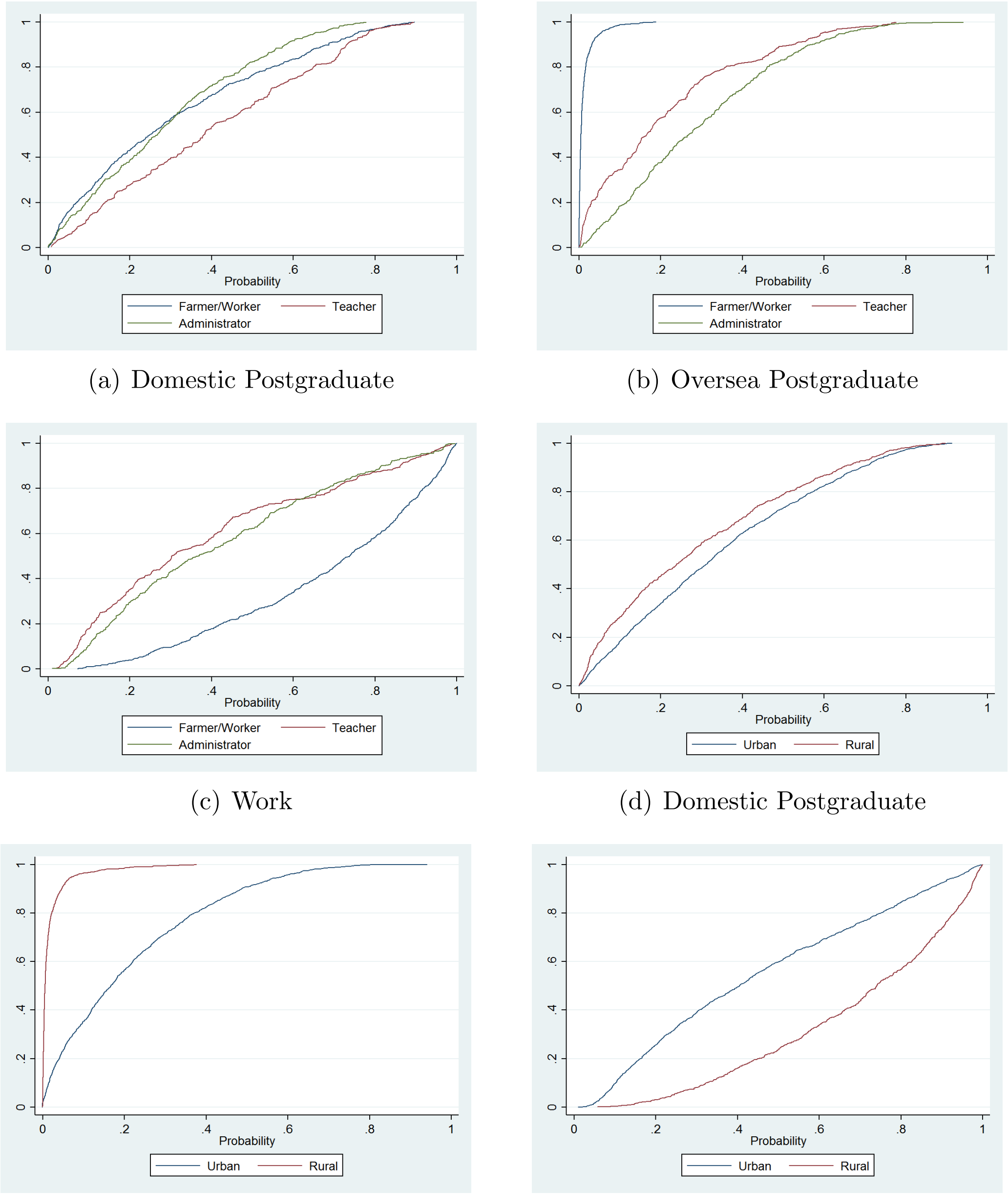
Figure 5.7 shows the cumulative distribution of probabilities on father's occupation and Hukou status. The figures in the first row show the cumulative distributions of three different occupations for each choice. Students with a father who is a farmer or a worker are less likely to choose domestic or overseas postgraduates. A father who is an administrator is more likely to send his children overseas.

The figures in the second row show the cumulative distributions on Hukou status. Rural students are more likely to choose work but less likely to choose domestic and overseas postgraduates than their urban counterparts. These results are consistent with what we find in the multinomial regression analysis and suggest the influence of circumstances on graduates' choices.

5.6 Conclusion

This essay uses graduates' administrative data from one university in China during 2008 to 2015. Through the analysis of this dataset, we find a widening gap in admission to tertiary education and its outcome between rural and urban residents. In particular, the proportion of rural graduates decreased from 50% in 2008 to less than 25% in 2015; the rate of the decrease is much faster than the rate of increase in urbanization rate — from less than 50% to more than 50%. This finding is consistent with other research on educational inequality in China such as Connelly and Zheng (2003), Wang et al. (2013)

Figure 5.7: The Cumulative Distributions of Probabilities on Father's Occupation and Hukou status



(a) Domestic Postgraduate

(b) Oversea Postgraduate

(c) Work

(d) Domestic Postgraduate

(e) Oversea Postgraduate

(f) Work

and Zhang et al. (2015).

However, there are differences between this study and other related research. For example, the longitudinal data in this thesis highlights the increase of educational inequality for a longer period when focused on tertiary education. This thesis shows a decrease in the share of rural students in highly-ranked colleges. In contrast, the study by Wang (2014) reports an increase in the share of rural students in lower-quality colleges. Both studies show a widening rural-urban gap in college enrolment though in different perspectives.

The rural-urban gap in China could be due to unequal allocations of educational resources (Wang, 2014) and the restriction on the education of rural migrants' children (Zhang et al., 2015). This thesis detected another factor — the economic conditions of households. Having a higher household income could be an advantage for students when it comes to enrollment in a highly-ranked university even when they did not perform well in the college entrance examination. We find that, for those who have already been admitted to this university, the entrance score is negatively related to household income. In contrast, when estimating a similar regression on GPA, the relationship between household income and academic performance disappears. These results provide evidence that students with a higher household income might have more options in choosing universities than their lower-household-income counterparts. Due to the lack of data on those who failed admission to universities, this thesis cannot examine the impact of household income on university admission.

Household income is also a main determinant of graduate outcomes. Graduates with a high household income are more likely to enroll for a postgraduate degree. The household income of graduates choosing an overseas postgraduate program is significantly higher than those who choose to do a domestic postgraduate program or to work. This finding is supported by research on educational expenditure in China. Qian and Smyth (2011) find that household income has significant effects on expenditures on domestic and overseas education. Students with a higher household income are more likely to go overseas for education with financial support.

With lower household income, rural graduates have fewer opportunities in their whereabouts after graduation. Graduates with a rural Hukou are much less likely to choose to study for a postgraduate degree after graduation than their urban counterparts, let alone for an overseas postgraduate. However, the lack of opportunities for rural graduates is not because of their lower academic performance in college. Instead, rural graduates have similar GPAs to their urban counterparts. In addition, the academic performance in college is uncorrelated with graduates' Hukou status and household income. This means that GPA can represent individuals' effort on academic performance which is independent of circumstances. Although GPA affects graduates' whereabouts, it fails to explain why rural graduates exclude the choice of studying overseas.

Based on these findings, we conclude that inequality of opportunity in tertiary education in China has significantly escalated in the last decade. Students who entered highly-ranked universities relied more on their circumstances such as family backgrounds, parents' socioeconomic status and Hukou status. They also faced unequal opportunity when they graduated. Having similar academic performances, graduates with urban Hukou and wealthy parents had more opportunities in their whereabouts.

Since our data is from a highly-ranked university, this study only focuses on the admission and graduates' outcome in highly-ranked universities. The admission and graduates' outcome in a low-quality university might be a different story. Nevertheless, using this dataset, this paper sheds light on reasons and highlights the trend of inequality of opportunity in tertiary education and addresses the issues of rural-urban inequality in China.

Chapter 6

Conclusion

6.1 Summary and Findings

The main objective of this thesis is to improve the empirical methodology in measuring inequality of opportunity, and to apply the new methodologies to gain better understanding of inequality of opportunity in China. Specifically, we study inequality of opportunity in terms of income and education in China. The measurement of inequality of opportunity in income shows that income is not fairly distributed as it is influenced by circumstances. To study the educational inequality of opportunity, we examine whether students' opportunities to access a highly-ranked university and their opportunities after graduation are fair and not determined by circumstances. Ideally, equal opportunity holds when the rate of enrolment and the graduate outcomes are identical for those with different circumstances.

We address several empirical issues in three main essays. One issue is the independence between circumstances and effort assumed by most of the literature. This assumption simplifies the approach to measure inequality of opportunity, but at the cost of potentially underestimating the extent of inequality. In Essays I and II, we identify two different ways in which circumstances could be affected by effort. First, different types, which are groups with individuals sharing the same circumstances, could have different effort distributions. Second, circumstances can have different effects on income for different levels of effort.

In Essay I, we capture the effect of circumstances on effort through differences in variances across types. Variances in income distributions imply that there are differences in individual effort among individuals with the same circumstances. We use Maximum-Likelihood Estimation to parameterize both mean and variance for each income distribution conditional on the circumstances. This method allows us to estimate not only the overall effect of circumstances on income but also the effect of each circumstance variable such as gender, ethnicity and parents' socioeconomic status on income. We believe that measures showing the effect of each circumstance variable are more useful for policy consideration than a composite measure. Circumstances contributing more to income inequality should be given greater attention and be prioritized in policy design.

In Essay II, we address the effect of circumstances on effort by examining the heterogeneous effects of circumstances on income between different levels of effort. Most literature assumes a homogeneous effect of circumstances on income and an unobserved effort due to data limitation. Inequality of opportunity still can be measured (Ferreira and Gignoux, 2008) without knowing effort. However, without information on effort, the measure could be biased if the effects of circumstances on income differ across levels of effort. To identify this heterogeneous effect, we use a latent-class model in which each level of effort is represented by a class so that levels of effort can be estimated even though they cannot be observed. To allow for heterogeneity, the effects of circumstances on income are parameterized separately for each level of effort.

Using two different approaches to identifying the correlation between circumstances and effort, we find that the heterogeneous effect between types accounts for about 20% of income inequality and the heterogeneous effect of circumstances on income between levels of effort accounts for about 29%. These figures imply that neglecting the impact of circumstances on effort could lead to a significantly underestimated measure of inequality of opportunity.

In terms of the measure of inequality of opportunity, we use both an ex-post measure — a measure requiring the same outcome for those who exert the same degree of effort, and an ex-ante measure — a measure requiring the same average outcome for those with different circumstances. The former measure emphasises the respect for individual effort while the latter measure pays more attention to compensating the disadvantaged. These two measures correspond to two basic principles with regard to equal opportunity: the reward principle and the compensation principle (Fleurbaey, 2008). However, these two principles can lead to controversial results (Fleurbaey and Peragine, 2013). A measure consistent with the reward principle might suggest that a larger proportion of income inequality should be respected while a measure consistent with the compensation principle might indicate that a larger proportion of income inequality should be compensated. In Essay II, we find some evidences of the incompatibility between the ex-ante and ex-post measure. Results show that inequality of opportunity is about 41% when using an ex-post measure compared to 48% when using an ex-ante measure. These results suggest that more of income inequality should be respected (100% − 41% = 59%) if using an ex-post measure and more of income inequality should be compensated (55%) if using an ex-ante measure. Considering that there are unobserved circumstances and effort, the real gap between the ex-ante and ex-post measures could be higher. This inconsistent result between the ex-ante and ex-post approach needs more investigation in any future study.

In Essay III, we develop an approach to measure inequality of opportunity in the access to a highly-ranked university and its graduate outcomes. This study is different from most empirical literature in measuring inequality of opportunity in that graduate outcome is a categorical variable. In the empirical literature, the outcome of interest is either a continuous variable such as income or a binary variable such as having tertiary education or not. A categorical variable is different from a continuous variable in that it sometimes cannot be directly ranked. High income or having tertiary education clearly is an advantage over low income or not having tertiary education. In contrast, a categorical variable could have no order but options related to individual choices and preferences. In Essay III, there are three whereabouts for graduates: work, postgraduate study in China, or oversea postgraduate. Inequality of opportunity exists if these graduates' whereabouts are affected by circumstances. To evaluate inequality of opportunity given graduate outcomes, we use stochastic dominance (Lefranc et al., 2008) to examine the effect of circumstances on the categorical outcome. There is inequality of opportunity if the distributions of graduates outcomes, given one circumstance, is second-order stochastically dominated by the distributions of graduates outcomes given other circumstances.

6.2 Policy Implication

By addressing these issues and improving the methodology, we closely examine inequality of opportunity in China in both income and education on the basis of the conventional framework developed by Roemer (1998). In Essay I, unfair income inequality accounts for about 50% of total income inequality if the impact of circumstances on effort is considered. In Essay II, unfair income inequality accounts for about 48% to 56% of total income inequality. These figures suggest a large proportion of income inequality is unfair and due to circumstances such as gender, ethnicity, the Hukou Status and parents' socioeconomic backgrounds.

We do not limit our study to providing a rough and general figure for indicating inequality of opportunity in China. Instead, we identify types associated with lack of opportunity — the opportunity-deprivation profile purposed by Ferreira et al. (2011). For example, Ferreira et al. (2011) found that ethnic minority is the most important reason for the deprivation of opportunities in Brazil. In China, our findings suggest that geographic factors such as the Hukou status and residential provinces are the most important factors for deprived opportunities. These findings indicate substantial rural-urban inequality and regional disparity in China. A rural household earns about half of the income an urban household does, and rich provinces have an average income over four times that of poor provinces.

More importantly, our findings suggest that geographic factors play an increasingly important role in education. In the third essay, we find that the proportion of students from rural areas attending a highly ranked university decreased from more than 50% in

2008 to about 20% in 2015. Even when rural students have entered and completed a bachelor degree in a highly-ranked university, they are still less likely to study towards a postgraduate degree. These findings are in line with researches on rural-urban inequality in educational opportunity (e.g. Wang et al. 2013, and Zhang et al. 2015). The difference is that our study emphasizes inequality of opportunity in graduate outcomes for graduates studying at the same highly-ranked university. Distribution unfairness not only exists in the labour market and the access of education; it also affects the decisions of individuals to invest in their own human capital. This rising inequality in educational opportunity might lower the intergenerational mobility in the next generation in China.

6.3 Future Research

The focus of this thesis is mainly on improving the empirical methods and measuring inequality of opportunity in China in terms of income and education. However, the improvement is limited by the data we used. For example, we used the China Family Panel study for Essay I and II, but this dataset only contained two waves at the time we worked on these two essays. Future improvement on the empirical methods can be made by using a panel model based on a dataset with a longer period. In Essay III, our data is the administrative data from one highly ranked university. The result will be more concrete if data from other universities become available.

In addition, this thesis contributes less to the theoretical aspect of inequality of opportunity. For example, our study points out that circumstances could affect effort in different ways and the biases due to omitting these effects are identified empirically. On the contrary, we do not claim that these effects must be compensated. Future research could clarify whether these different types of effects of circumstances on effort should be considered unfair or not.

Bibliography

Afridi, F., Li, S. X., and Ren, Y. (2015). Social identity and inequality: The impact of China's hukou system. *Journal of Public Economics*, 123:17 – 29.

Alain, T., Sandy, T., Florence, J., and Marion, D. (2010). Inequality of opportunities in health in France: a first pass. *Health Economics*, 19(8):921–938.

Almas, I., Cappelen, A. W., Lind, J. T., Sø rensen, E. O., and Tungodden, B. (2011). Measuring unfair (in)equality. *Journal of Public Economics*, 95:488–499.

Arneson, R. (1989). Equality and equal opportunity for welfare. *Philosophical Studies*, 56(1):77–93.

Ashraf, M. A., Liu, S., Ismat, H. I., and Tsegay, S. M. (2017). Choice of higher education institutions: Perspectives of students from different provinces in China. *Frontiers of Education in China*, 12(3):414–435.

Awaworyi, S. and Mishra, V. (2014). Returns to education in China: A meta-analysis. Monash Economics Working Papers 41-14, Monash University, Department of Economics.

Balczar, C. F. (2015). Lower bounds on inequality of opportunity and measurement error. *Economics Letters*, 137:102 – 105.

Barry, B. (2005). *Why Social Justice Matters*. Themes for the 21st century. Wiley.

Becker, G. S. and Lewis, H. G. (1974). Interaction between quantity and quality of children. In *Economics of the family: Marriage, children, and human capital*, pages 81–90. UMI.

Becker, G. S. and Tomes, N. (1986). Human capital and the rise and fall of families. *Journal of Labor Economics*, 4(3):S1–39.

Bertaut, C. and Starr-McCluer, M. (2000). Household portfolios in the United States. Finance and Economics Discussion Series 2000-26, Board of Governors of the Federal Reserve System (U.S.).

Biao, X. (2007). How far are the left-behind left behind? a preliminary study in rural China. *Population, Space and Place*, 13(3):179–191.

Björklund, A., Jäntti, M., and Roemer, J. E. (2012). Equality of opportunity and the distribution of long-run income in Sweden. *Social Choice and Welfare*, 39(2):675–696.

Bossert, W. (1995). Redistribution mechanisms based on individual characteristics. *Mathematical Social Sciences*, 29(1):1–17.

Bourguignon, F., Ferreira, F. H. G., and Menndez, M. (2007). Inequality of opportunity in Brazil. *Review of Income and Wealth*, 53(4):585–618.

Brunori, P., Ferreira, F. H. G. F. H. G., and Peragine, V. (2013). Inequality of opportunity, income inequality and economic mobility: Some international comparisons. In *IZA Discussion Paper No.7155*, number 7155.

Brunori, P., Peragine, V., and Serlenga, L. (2012). Fairness in education: The italian university before and after the reform. *Economics of Education Review*, 31(5):764 – 777.

Cameron, A. and Trivedi, P. (1990). The information matrix test and its implied alternative hypotheses. Papers 372, California Davis - Institute of Governmental Affairs.

Carroll, C. D. (1998). Why do the rich save so much? NBER Working Papers 6549, National Bureau of Economic Research, Inc.

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

Checchi, D., Peragine, V., and Serlenga, L. (2010). Fair and unfair income inequalities in europe. IZA Discussion Papers 5025, Institute for the Study of Labor (IZA).

Chen, C.-N., Tsaur, T.-W., and Rhai, T.-S. (1982). The Gini coefficient and negative income. *Oxford Economic Papers*, 34(3):473–478.

Chen, Y. and Cowell, F. A. (2015). Mobility in China. *Review of Income and Wealth*, 63(2):203–218.

Clarke, G. R. (1995). More evidence on income distribution and growth. *Journal of Development Economics*, 47(2):403 – 427.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, 99(4):906–944.

Connelly, R. and Zheng, Z. (2003). Determinants of school enrollment and completion of 10 to 18 year olds in China. *Economics of Education Review*, 22(4):379–388.

Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, 27:79–102.

Cowell, F. A. (1985). Measures of distributional change: An axiomatic approach. *The Review of Economic Studies*, pages 135–151.

Cowell, F. A. (2000). Measurement of inequality. *Handbook of Income Distribution*, 1:87–166.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844.

Dawkins, R. (2006). *The Selfish Gene: 30th Anniversary Edition.* ISSR library. OUP Oxford.

Deb, P. (2008). Finite mixture models. Summer North American Stata Users' Group Meetings 2008 7, Stata Users Group.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336.

Deb, P. and Trivedi, P. K. (2013). Finite mixture for panels with fixed effects. *Journal of Econometric Methods*, 2(1):35–51.

Deutsch, J., Pi Alperin, M. N., and Silber, J. (2018). Using the shapley decomposition to disentangle the impact of circumstances and efforts on health inequality. *Social Indicators Research*, 138(2):523–543.

Deutsch, J. and Silber, J. (2007). *Decomposing Income Inequality by Population Subgroups: A Generalization*, pages 237–253.

Deutsch, J. and Silber, J. (2008). Earnings functions and the measurement of the determinants of wage dispersion: extending the blinderoaxaca approach. In *Work, Earnings and Other Aspects of the Employment Relation*, volume 28, pages 371–427. Emerald Group Publishing Limited.

Devooght, K. (2008). To each the same and to each his own: A proposal to measure responsibility-sensitive income inequality. *Economica*, 75(298):280–295.

Dias, P. R. (2009). Inequality of opportunity in health: evidence from a UK cohort study. *Health Economics*, 18(9):1057–1074.

Donni, P. L., Rodrguez, J., and Dias, P. R. (2015). Empirical definition of social types in the analysis of inequality of opportunity: a latent classes approach. *Social Choice and Welfare*, 44(3):673–701.

Du, Z., LI, R., He, Q., and ZHANG, L. (2014). Decomposing the rich dad effect on income inequality using instrumental variable quantile regression. *China Economic Review*, 31:379 – 391.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610.

Dworkin, R. (1981a). What is equality ? part 2 : Equality of resources. *Philosophy and Public Affairs*, 10(4):283–345.

Dworkin, R. (1981b). What is equality? part 1: Equality of welfare. *Philosophy and Public Affairs*, 10(3):185–246.

Eriksson, T., Pan, J., and Qin, X. (2014). The intergenerational inequality of health in China. *China Economic Review*, 31:392 – 409.

Eriksson, T. and Zhang, Y. (2012). The Role of Family Background for Earnings in Rural China. *Frontiers of Economics in China*, 7(3):465–477.

Ferreira, F. H. G. and Gignoux, J. (2008). The measurement of inequality of opportunity: theory and an application to Latin America. *World Bank Policy Research Working Paper 4659.*

Ferreira, F. H. G. and Gignoux, J. (2014). The measurement of educational inequality: Achievements and opportunity. *The World Bank Economic Review*, 28(2):210.

Ferreira, F. H. G., Gignoux, J., and Aran, M. (2011). Measuring inequality of opportunity with imperfect data: the case of Turkey. *The Journal of Economic Inequality*, 9(4):651–680.

Ferreira, F. H. G., Lakner, C., Lugo, M. A., and Ozler, B. (2014). Inequality of opportunity and economic growth : a cross-country analysis. Policy Research Working Paper Series 6915, The World Bank.

Ferreira, F. H. G., Lanjouw, P., and Neri, M. (2003). A robust poverty profile for Brazil using multiple data sources. *Revista Brasileira de Economia*, 57:59 – 92.

Fleurbaey, M. (2008). *Fairness, Responsibility, and Welfare.* OUP Oxford.

Fleurbaey, M. and Bossert, W. (1996). Redistribution and compensation. *Social Choice and Welfare*, 13(3):343–355.

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

Fleurbaey, M. and Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics*, 28(1):73 – 90.

Foguel, M. and Veloso, F. (2014). Inequality of opportunity in daycare and preschool services in Brazil. *The Journal of Economic Inequality*, 12(2):191–220.

Foley, D. K. (1967). Resource allocation and the public sector. *Yale Economic Studies*, 7(1):45–98.

Foster, J. E. and Shneyerov, A. A. (2000). Path independent inequality measures. *Journal of Economic Theory*, 91(2):199–222.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gamboa, L. F. and Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006-2009. *Economics of Education Review*, 31(5):694 – 708.

González de San Román, A. and De La Rica, S. (2012). Gender gaps in PISA test scores: The impact of social norms and the mother's transmission of role attitudes.

Hambrick, D. and Tucker-Drob, E. (2014). The genetics of music accomplishment: Evidence for geneenvironment correlation and interaction. *Psychonomic Bulletin & Review*, pages 1–9.

Hannum, E. and Wang, M. (2006). Geography and educational inequality in China. *China Economic Review*, 17(3):253 – 265. Symposium on Inequality, Market Development, and Sources of Growth in China under Accelerating ReformTechnology, Human Capital, and Economic Development.

Hardoon, D. (2017). An economy for the 99%: Its time to build a human economy that benefits everyone, not just the privileged few.

Harsanyi, J. (1976). Can the maximin principle serve as a basis for morality? a critique of John Rawlss theory. 12:37–63.

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61:434.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63:309.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica: Journal of the Econometric Society*, pages 1219–1240.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Hufe, P. and Peichl, A. (2015). Lower Bounds and the Linearity Assumption in Parametric Estimations of Inequality of Opportunity. *IZA Discussion Papers*, (9605).

Israeli, O. (2007). A shapley-based decomposition of the r-square of a linear regression. *The Journal of Economic Inequality*, 5(2):199–212.

ISSP (2009). Issp 2009 ”social inequality iv” - za no. 5400. http://www.gesis.org/issp/modules/issp-modules-by-topic/social-inequality/2009/.

ISSS (2014). The development of people’s well being 2014. Technical report.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jann, B. et al. (2008). The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4):453–479.

Jusot, F., Tubeuf, S., and Trannoy, A. (2013). Circumstances and efforts: How important is their correlation for the measurement of inequality of opportunity in health? *Health Economics*, 22(12):1470–1495.

Kanbur, R. and Wagstaff, A. (2014). How useful is inequality of opportunity as a policy construct? Working Papers 338, ECINEQ, Society for the Study of Economic Inequality.

Kanbur, R. and Zhang, X. (1999). Which regional inequality? the evolution of rural–urban and inland–coastal inequality in China from 1983 to 1995. *Journal of Comparative Economics*, 27(4):686–701.

Knight, J. and Song, L. (1993). The spatial contribution to income inequality in rural China. *Cambridge Journal of Economics*, 17(2):195–213.

Kolm, S.-C. (1973). super-équité. *Kyklos*, 26(4):841–843.

Koppelman, F. S. and Bhat, C. (2006). *A self instructing course in mode choice modeling: multinomial and nested logit models*. FTA.

Kranich, L. (1996). Equitable opportunities: An axiomatic approach. *Journal of Economic Theory*, 71(1):131–147.

Kuznets, S. (1955). Economic growth and income inequality. *The American Economic Review*, 45(1):1–28.

Lara Ibarra, G. and Martinez Cruz, A. L. (2015). Exploring the sources of downward bias in measuring inequality of opportunity. Policy Research Working Paper Series 7458, The World Bank.

Larsen, C. A. (2016). How three narratives of modernity justify economic inequality. *Acta Sociologica*, 59(2):93–111.

Lazar, A. (2013). *¡italic¿Ex-ante¡/italic¿ and ¡italic¿Ex-post¡/italic¿ Measurement of Inequality of Opportunity in Health: Evidence from Israel*, chapter 14, pages 371–395.

Lefranc, A., Pistolesi, N., and Trannoy, A. (2008). Inequality of opportunities vs. inequality of outcomes: Are Western societies all alike? *Review of Income and Wealth*, 54:513–546.

Lefranc, A., Pistolesi, N., and Trannoy, A. (2009). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics*, 93(11-12):1189–1207.

Li, H., Loyalka, P., Rozelle, S., Wu, B., and Xie, J. (2015). Unequal access to college in China: How far have poor, rural students been left behind? *The China Quarterly*, 221:185–207.

Li, H., Zhang, J., and Zhu, Y. (2008). The Quantity-Quality Trade-Off of Children In a Developing Country: Identification Using Chinese Twins. *Demography*, 45(1):223–243.

Li, H. and Zou, H.-f. (1998). Income inequality is not harmful for growth: theory and evidence. *Review of development economics*, 2(3):318–334.

Li, S., Sato, H., and Sicular, T. (2013). *Inequality in Focus vol. 2(no. 2)*. World Bank.

Li, S., Whalley, J., and Xing, C. (2014). China's higher education expansion and unemployment of college graduates. *China Economic Review*, 30:567–582.

Liu, Z. (2005). Institution and inequality: the hukou system in China. *Journal of Comparative Economics*, 33(1):133–157.

Manna, R. and Regoli, A. (2012). Regression-based approaches for the decomposition of income inequality in Italy, 1998-2008. *Rivista di statistica ufficiale*, 14(1):5–18.

Marrero, G. A. and Rodriguez, J. G. (2013). Inequality of opportunity and growth. *Journal of Development Economics*, 104(0):107 – 122.

McFadden, D. et al. (1973). *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., and Ulln, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science*.

Nardi, M. D. (2002). Wealth inequality and intergenerational links. Staff Report 314, Federal Reserve Bank of Minneapolis.

NBS (2013). *China Statistical Yearbook 1996-2012*. National Bureau of Statistics of China.

NBS (2016). *China Statistical Yearbook 2016*. National Bureau of Statistics of China.

Niehues, J. and Peichl, A. (2014). Upper bounds of inequality of opportunity: theory and evidence for Germany and the US. *Social Choice and Welfare*, 43(1):73–99.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, pages 693–709.

Paes de Barros, R., Ferreira, F. H., Molinas Vega, J. R., and Saavedra Chanduvi, J. (2009). *Measuring inequality of opportunities in Latin America and the Caribbean*. Washington, DC: World Bank.

Park, A. and Wang, D. (2010). Migration and urban poverty and inequality in China. *China Economic Journal*, 3(1):49–67.

Pazner, E. A. and Schmeidler, D. (1978). Egalitarian equivalent allocations: A new concept of economic equity. *The Quarterly Journal of Economics*, pages 671–687.

Peragine, V. (2004a). Measuring and implementing equality of opportunity for income. *Social Choice and Welfare*, 22(1):187–210.

Peragine, V. (2004b). Ranking income distributions according to equality of opportunity. *The Journal of Economic Inequality*, 2(1):11–30.

Peragine, V. and Ferreira, F. (2015). Equality of opportunity: Theory and evidence. *World Bank Policy Research Paper*, (7217).

Pistolesi, N. (2009). Inequality of opportunity in the land of opportunities, 1968-2001. *Journal of Economic Inequality*, 7(4):411–433.

Qian, J. X. and Smyth, R. (2011). Educational expenditure in urban China: income effects, family characteristics and the demand for domestic and overseas education. *Applied Economics*, 43(24):3379–3394.

Qian, X. and Smyth, R. (2008). Measuring regional inequality of education in China: widening coast–inland gap or widening rural–urban gap? *Journal of International Development*, 20(2):132–144.

Qin, X., Wang, T., and Zhuang, C. C. (2016). Intergenerational transfer of human capital and its impact on income mobility: Evidence from China. *China Economic Review*, 38:306 – 321.

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

Rawls, J. (1971). *A Theory Of Justice (Orig Edn)*. Harvard paperback. Harvard University Press.

Rawls, J. (2009). *A Theory of Justice*. Harvard university press.

Rodrguez, J. G. (2008). Partial equality-of-opportunity orderings. *Social Choice and Welfare*, 31(3):435–456.

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

Roemer, J. E. (1993). A pragmatic theory of responsibility for the egalitarian planner. *Philosophy and Public Affairs*, 22(2):146–166.

Roemer, J. E. and Trannoy, A. (2015). Chapter 4 - equality of opportunity. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2 of *Handbook of Income Distribution*, pages 217 – 300. Elsevier.

Salehi-Isfahani, D., Hassine, N. B., and Assaad, R. (2014). Equality of opportunity in educational achievement in the Middle East and North Africa. *Journal of Economic Inequality*, 12(4):489.

Sen, A. (1979). *Equality of What?* Stanford University.

Sen, A. (1995). *Inequality Reexamined*. A Russell Sage Foundation Book. Russell Sage Foundation.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

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.

Sicular, T., Ximing, Y., Gustafsson, B., and Shi, L. (2007). The urban–rural income gap and inequality in China. *Review of Income and Wealth*, 53(1):93–126.

Small, K. A. and Hsiao, C. (1985). Multinomial logit specification tests. *International economic review*, pages 619–627.

Starmans, C., Sheskin, M., and Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, 1:0082.

Swift, A. (2005). Justice, luck, and the family : the intergenerational transmission of economic advantage from a normative perspective. In Bowles, S., Gintis, H., and Groves, M. O., editors, *Unequal chances : family background and economic success*, pages 256–276. Princeton University Press, New York : Princeton, N.J. ; Oxford : Russell Sage Foundation.

The World Bank (2016a). Gini index. http://data.worldbank.org/indicator/SI.POV.Gini.

The World Bank (2016b). Poverty and equity in China. http://povertydata.worldbank.org/poverty/country/CHN.

Theil, H. (1967). *Economics and information theory*. Studies in mathematical and managerial economics. North-Holland Pub. Co.

Thomson, W. (1994). Notions of equal, or equivalent, opportunities. *Social Choice and Welfare*, 11(2):137–156.

Van De Gaer, D. (1993). *Equality of Opportunity and Investment in Human Capital*. Faculteit der Economische en Toegepaste Economische Wetenschappen, Katholieke Universiteit Leuven. Kath. Univ.

Waltenberg, F. D. and Vandenberghe, V. (2007). What does it take to achieve equality of opportunity in education?: An empirical investigation based on Brazilian data. *Economics of Education Review*, 26(6):709 – 723. Economics of Education: Major Contributions and Future Directions - The Dijon Papers.

Wan, G., Lu, M., and Chen, Z. (2006). Globalization and regional income inequality: Empirical evidence from within China. Working Paper Series RP2006/139, World Institute for Development Economic Research (UNU-WIDER).

Wan, G. and Zhou, Z. (2005). Income inequality in rural China: Regression-based decomposition using household data. *Review of Development Economics*, 9(1):107–120.

Wang, Q. (2014). Rural students are being left behind in China. *Nature*, 510(7506).

Wang, X., Liu, C., Zhang, L., Shi, Y., and Rozelle, S. (2013). College is a rich, han, urban, male club: Research notes from a census survey of four tier one colleges in China. *The China Quarterly*, 214:456–470.

Whalley, J. and Zhang, S. (2007). A numerical simulation analysis of (hukou) labour mobility restrictions in China. *Journal of Development Economics*, 83(2):392 – 410. Papers from a Symposium: The Social Dimensions of Microeconomic Behaviour in Low-Income Communities.

Whalley, J. and Zhao, X. (2013). The contribution of human capital to China's economic growth. *China Economic Policy Review*, 2(01):1350001.

Whyte, M. K. (2010). *One country, two societies: rural-urban inequality in contemporary China*, volume 16. Harvard University Press.

Wu, D. and Rao, P. (2017). Urbanization and income inequality in China: An empirical investigation at provincial level. *Social Indicators Research*, 131(1):189–214.

Wu, X. and Treiman, D. J. (2004). The household registration system and social stratification in China: 1955–1996. *Demography*, 41(2):363–384.

Xie, Y. and Zhou, X. (2014). Income inequality in todays China. *Proceedings of the National Academy of Sciences*, 111(19):6928–6933.

Zeng, J., Pang, X., Zhang, L., Medina, A., and Rozelle, S. (2014). Gender inequality in education in China: a meta-regression analysis. *Contemporary Economic Policy*, 32(2):474–491.

Zhang, D., Li, X., and Xue, J. (2015). Education inequality between rural and urban areas of the people's republic of China, migrants children education, and some implications. *Asian Development Review*.

Zhang, Y. and Eriksson, T. (2010). Inequality of opportunity and income inequality in nine Chinese provinces, 1989-2006. *China Economic Review*, 21:607–616.