Three Essays on the Evolution of the Determinants of Educational Attainment and its Consequences

M. Yasin Arafat

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Economics, Science

Nicolaus Tideman, Chair
Wen You
Alec Smith
Eric Bahel

September 19, 2018
Blacksburg, Virginia

Keywords: Inequality of Opportunity, Education, Social Factors, Stepwise Selection, Logistic Regression Model, Machine-Learning Process, Age Distribution, Graduation, Income, Decomposition

Three Essays on the Evolution of the Determinants of Educational Attainment and its Consequences

M. Yasin Arafat

Abstract

The dissertation focuses on the different determinants of education, their effects on the educational outcome, and the overall effect of education on the lifetime consequences.

The first chapter focuses on the inequality of educational opportunity across different demographic factors. This chapter employs a broader set of social factors to provide fresh insights into the inequality situation in the USA relative to those of the extant literature. The chapter employs polynomial trends for the effects of social factors to identify long-term trends in the determinants of the differences in attainment of each of four achievements (high school graduation, some college, college graduation, and post-college work) across different endogenous social groups. Using the Panel Study of Income Dynamics (PSID) data for the years of 1968-2013, we show how inequality of educational opportunity and its determinants have evolved over the years. The chapter utilizes the machine-learning process and logistic regression model to identify inequality of opportunity.

The second chapter examines the age demographic distribution of graduates across cohorts from 1940 until 1990. Using the PSID data, the paper explored the first and second moment of the age of graduating from high school and college across the US. To deal with the data deficiencies, a large part of the chapter dealt with data preparation. The chapter provides a unique method of extracting information on the graduating age of the individuals both from high school and from college. The results show a large dispersion across the full sample. The data truncated to a standard length, however, provides a much smaller dispersion and much smaller moments. The chapter concludes that as the time passes, people tend to attain education at a younger age.

The third chapter investigates the trends of contribution of different factors of income starting from 1910 cohort. Following Mincer (1974), a wave of papers studied how various factors contribute to the earnings of individuals. This paper contributes to that literature in three ways: (i) using the PSID data, it computes the actual working experience of the individuals, (ii) it studies the cohorts who were born in 1910 or afterwards, unlike the existing papers, and (iii) it adds two variables—technological progress and the occupation with which individuals start their careers—to an extended Mincerian equation. The results re-emphasize the importance of education in lifetime earnings. The results also show that while some of the determinants of income have become more important over the years, other factors have not changed much in importance.

JEL Codes: I21, I26, I24, J11, J15, J16, J31

Three Essays on the Evolution of the Determinants of Educational Attainment and its Consequences

M. Yasin Arafat

General Audience Abstract

The reason for choosing the theme 'Evolution of the Determinants of Educational Attainment and its Consequences' was to investigate the different determinants of education, their effects on the educational outcome, and the overall effect of education on the lifetime consequences. Education is considered as one of the tools to eradicate poverty. Yet, countries with high educational coverage keeps suffering from poverty, a reason for which is higher inequality of opportunity.

In the first chapter, entitled 'Inequality in Educational Opportunity in the United States', opportunity inequality in education is illustrated. Much inequality stems from differences in educational attainment. A lack of educational attainment puts an individual behind in the career race, even before the race has started. While individuals are responsible for some of the differences in educational attainment, there are factors outside the control of individuals that play substantial roles. The inequality that arises from these factors is known as inequality of opportunity. This paper focuses on inequality of educational opportunity across socioeconomic background, race, and sex. The factors that are analyzed for their contributions to inequality of educational opportunity are father's education, father's occupation, mother's education, and economic status of the individual's family. The results show that inequality of opportunity has seen a consistent decline for high school completion. The inequality of opportunity (IO) declines for obtaining some college education for the bottom two social groups and remained persistent for the relatively more advantaged group. For college/post-college education, the IO is much lower and, in general, remained persistent across the social strata. Although the females were behind the males – given the equal opportunity – regardless of the race and socioeconomic status during the beginning and the mid twentieth century, the scenario reversed in the late twentieth century. In terms of educational disparity among races, African Americans trail their White counterparts along all the years.

The second chapter 'First and Second Moments of the Age Distributions of Graduates' looks into the age characteristics (mean and variance) in graduating from high school and college across the cohorts from 1940s to 1990s. The idea of the paper largely came from the first chapter of the dissertation as we assumed the lack of opportunity at the earlier age could delay the attainment of education. The paper intends to find out the average age of graduation over the years. In the process, the paper put forward a method to extract the information of age of graduation from the Panel Study of Income Dynamics (PSID) data, as the database does not readily avail the information. The chapter concludes that as the time passes, people tend to attain education at a much younger age.

Titled as 'Factors Affecting Income: Education, Experience, and Beyond', the third chapter investigates the contribution of different factors – education, experience, parental endowments, and labor market conditions – in the returns to education using the PSID data and compare the more recent scenarios with the past. This paper focuses on the trend of the rate of return to different factors of income across the two cohorts – those born between 1910 and 1950, and those born after

1950 – while identifying the changes in the returns for the same education level over time. The paper aims to find out how the contribution of the different factors of earning has changed in the USA over the years. The paper also intends to find out the role of technological progress in reducing the earning gaps across the different social groups. The results re-emphasize the importance of education in lifetime earnings. Experience has become a more important factor of income over the years. The chapter also suggests that income of an individual is a monotonic function of socioeconomic endowments and better endowments resulted in higher returns. Lastly, the chapter finds that the technological investment is progressive in manner.

Dedication

My beloved and lovely wife, Rabita Reshmeen Banee – a Ph.D. student at Virginia Tech, who has been by my side through thick and thin, with all the support and encouragement, to finish the dissertation. My beautiful mother Rafia Haque (a wonder woman herself) who has always been the pillar of my success and achievement.

Acknowledgement

I am ever grateful to Dr. Nicolaus Tideman[1] for providing numerous suggestions and countless number of sessions to converse towards writing the dissertation. I would also take this opportunity to thank my dissertation committee members – Dr. Wen You[2], Dr. Eric Bahel[3], and Dr. Alec Smith[4] – for all the comments and suggestions to improve the dissertation. I also want to thank Dr. John A. Bishop[5] (whom I have met at the Southern Economic Association Conference, 2017) and Dr. Juan Gabriel Rodiguez[6] for their important comments and suggestions on the first chapter of the dissertation. Lastly, I want to show my gratitude to all the friends, family, and well-wishers who supported me throughout.

[1] Professor, Department of Economics, Virginia Polytechnic Institute & State University
[2] Associate Professor, Department of Agricultural and Applied Economics, Virginia Polytechnic Institute & State University
[3] Associate Professor, Department of Economics, Virginia Polytechnic Institute & State University
[4] Assistant Professor, Department of Economics, Virginia Polytechnic Institute & State University
[5] Professor, Department of Economics, East Carolina University
[6] Professor, Department of Economic Analysis I, Universidad Complutense de Madrid

Table of Contents

List of Figures

List of Tables

Chapter 1

Inequality in Educational Opportunity in the USA

Abstract

Much inequality of income stems from differences in educational attainment. While individuals are responsible for some of the differences, there are factors beyond their control that play substantial roles. The inequality that arises from these factors is known as inequality of opportunity. The focus of this paper is on the inequality of educational opportunity across race and gender. The social factors that are analyzed for their contributions to inequality of educational opportunity are father's education, father's occupation, mother's education, and economic status of the individual's family. This paper employs a broader set of social factors to provide fresh insights into the inequality situation in the USA relative to those of the extant literature. The paper employs polynomial trends for the effects of social factors to identify long-term trends in the determinants of the differences in attainment of each of four achievements (high school graduation, some college, college graduation, and post-college work) across different endogenous social groups. Using the Panel Study of Income Dynamics (PSID) data for the years of 1968-2013, we show how inequality of educational opportunity and its determinants have evolved over the years. The paper utilizes the machine-learning process and logistic regression model to identify inequality of opportunity.

1.1 Introduction

Education is an important determinant of financial success. Therefore, inequality of opportunity with respect to education increases wealth and income inequality. With the persistent decrease in the percentage of American children earning more than their parents[7], inequality of opportunity has become one of the burning issues to discuss in the 'Land of Opportunity,' the United States of America. With many entities fighting to reduce the gaps in income and wealth between the rich and the poor, the question arises as to what means for reducing these gaps are available. Education is considered as a key instrument for alleviating poverty and leveling the playing field of life.

[7] Source: Equality of Opportunity Project [Chetty et al. (2013)]

Although universal education is widely acclaimed, not all the people can avail themselves of education. Hence, inequality starts through the limited access to education.

Inequality of opportunity has been formalized. Comparing people from two social groups, A and B, if those from group A have systematically less success than those from group B, then, John Roemer (1993, 1998) argues, those in group A should not be held responsible for this difference. Roemer (1993, 1998) formalized the idea that differences in the rates of success between social groups should be labeled inequality of opportunity. When inequality of opportunity arises with respect to educational success, it leads to increasing inequality of income and wealth.

A lot of work has been done since Roemer formalized the idea of inequality of opportunity. Bourguignon et al. (2007) measured inequality of opportunity in Brazil, drawing a distinction between circumstances and efforts, in order to see which variable has a higher effect on the inequality, while Chechchi and Peragine (2010) did the same in the case of Italy. Pistolesi (2009) measured the inequality of opportunity for income acquisition in the case of United States from the year 1968 to 2001. Inter-country comparisons were done by Lefranc et al. (2009a) showing the relationship between income inequality and opportunity inequality in the case of nine developed western countries. Ferreira and Gignoux (2011) compared the opportunity inequality of six Latin American countries using both a parametric approach following Bourguignon et al. (2007) and a non-parametric approach following Chechchi and Peragine (2010). In the Aaberge et al. (2011) the authors measured the long term ex-ante and ex-post inequality in the case of Norway. There have been a few research conducted on educational opportunity. Foguel and Veloso (2014), using estimated probabilities, measured the opportunity inequality to access the daycare and preschool services in Brazil. Asadullah and Yalonetzky (2012) measured the inequality of educational opportunity in India using four different indices. In addition to measuring inequality within countries, the literature explores other economic aspects of inequality. In Alesina and Angeletos (2005), the authors showed how various perspectives about income inequality can affect the redistributive policies of a society. Marrero and Rodriguez (2010) showed how inequality of opportunity is related to growth, studying 23 states of the United States from 1980 to 1990. Some other empirical works like Rosa-Dias (2009) explored inequality of opportunity in health, and Roemer (2004) examined how inequality of opportunity affects the intergenerational correlation of incomes.

Our paper seeks to measure inequality of educational opportunity in USA. For this analysis we took into account six features of individuals: mother's education, father's occupation, father's education, socioeconomic condition, race, and gender. The aim is to examine the extent of inequality in educational achievement that arise from these factors in the US. This is the first paper to use Roemer's opportunity inequality theory to compare the extent of educational inequality that rises because of gender differences versus the inequality that rises because of class differences. Our focus is on the *social lotteries* as discussed by Rawls[8]. That is, we are interested in the inequality that arises from the socioeconomic background – economic status, parental education, and parental occupation - of the individuals. Graaf and Ganzeboom (1993), Cobalti and Schizzerotto (1993), Buchman, Charles and Sacchi (1993) and many others have showed the effects of father's education and occupation on the educational attainment of sons and daughters for different countries, in the book *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. However, none of these papers took into account the class or gender factor into their analysis. In our paper, we investigate inequality of educational opportunity in USA, using the Roemerian framework. Although Chechchi and Peragine (2010) applied the Roemerian

[8] We explain the social lotteries in Section 1.2.

framework to measure the earnings inequality between men and women, they only took into account the educational attainment of the parents. That is to say, our paper uses a broader set of social variables to analyze inequality of educational opportunity. Moreover, our approach to identifying inequality of opportunity is different from the methods used by most of the existing economics literature. Instead of decomposing the inequality into circumstances and residuals, we implement the probability of obtaining different levels of education to measure inequality.

In addition to measuring inequality of educational opportunity arising from social factors, the paper measures educational inequality across race and gender. The paper also contributes to identifying the social factors that affect educational achievement at different stages of education. Unlike previous papers, this paper implemented a machine learning process to identify the relevant factors. However, this paper takes into account only social factors as determinants of educational attainment. Hence, other factors such as educational infrastructure are not included in this analysis.

Instead of decomposing the inequality into opportunity and effort, like most of the economics papers do, we measure the inequality in terms of differences in probabilities of achieving benchmarks. Decomposition is not feasible for our analysis, since we do not identify the level of effort. The distribution of opportunity is determined, using the *indirect approach*, from the functional relationship between education and the six circumstance variables that we have taken into account. *Type* in our model is endogenously determined by taking the mean log-odds of the fitted values (of the first hurdle) as the values of the opportunity sets. Types remain universal across the hurdles. The paper takes account of the effects of circumstances on effort. While the paper does not decompose inequality into opportunity inequality and effort inequality, it recognizes that residual inequality (inequality that arises from the factors other than circumstances) is a function of circumstance. Hence, the inequality in our model can be decomposed into direct inequality of opportunity and indirect inequality of opportunity. Keeping that in mind, the paper supports the redistribution of income at the very beginning, in line with the *ex-ante compensation principle*. That would eliminate the effect of circumstances on effort.

We use data collected from the PSID (Panel Study of Income Dynamics) to measure inequality of opportunity in education. This paper contributes by comprehensively taking into account the effects of different variables – base variables, polynomials, and interactions – relating to social background in determining attainment in education for different social group. We implement the machine learning process to select from the large number of candidate variables. The social strata in this paper is endogenously determined. This paper measures the probability of attaining different level of education. While Mare (1980) had conducted a study of probability of educational transition, using different socioeconomic factors, with eight education strata, this paper contains only four education strata. Inequality of opportunity is measured for every cohort by taking the difference in probability between two social strata where the highest social group is treated as the reference group. The paper also conducted an analysis from the perspective of race and gender.

The paper is organized as follows. The next section discusses how the concept of inequality of opportunity evolved over time. We present the framework and concept of inequality of opportunity in section 1.3. Section 1.4 deals with the data. In section 1.5, we explain the model and the method. We show the results in section 1.6, and section 1.7 summarizes.

1.2 Emergence of Equality of Opportunity

The literature on equality dates back a long time and was more of a statistical issue prior to Rawls' (1958) philosophical contribution. The Rawlsian philosophy was based on specifying justice beyond maximizing civil liberties. His idea – the 'difference principle' - was to make people equal by having institutions that allocate the worst off individuals the maximum feasible level of 'primary goods'. He viewed primary goods as the goods necessary for the success of any life plan. Other philosophers like Jensen (1969) and Herrenstein (1971), posited in their papers that the existence of inequality is due to differential intelligence of individuals and hence achieving an equal income distribution is not possible. Their views conflicted with the other thinkers like Bowles (1973) and Colnisk (1974) who thought of inequality as a consequence of unequal opportunity. However, it was Dworkin (1981), Arneson (1989) or Cohen (1989) who introduced the personal responsibility into the literature on equality. Dworkin (1981 a,b) argued that a lack of internal resources can be overcome by external wealth and introduced responsibility into achieving equality in a society. Unlike Rawls, Dworkin emphasized preferences instead of primary goods and showed that preferences of individuals matter into the welfare distribution in the society. He argued that if a person has fewer internal resources, then it is the responsibility of the society to fill that gap by distributing the wealth accordingly. Later, Cohen (1989) suggested that a person should not be fully responsible for his preferences; as preferences are sometimes conditional on the circumstances that the individuals face, he criticized Dworkin's work. That is, these philosophers have brought the concept of effort in the form of responsibility into equality theory. It was not until the late 80s and early 90s that economists involved themselves in the discussion of the equality theory. Roemer (1993, 1998) put forward a framework for making policies for achieving equality through an objective known as 'objective equalisandum' named by Roemer and

Betts (2007). Roemer introduced into the discussion of equality the term ‘circumstances’ which are the factors for which individuals cannot be held responsible. Circumstances – as defined by Roemer - are the conditions in which an individual has no hand. That is, from the viewpoint of an individual, a circumstance is a ‘constant’ or given, and effort is a variable. Hence, both efforts and circumstances are responsible for the emergence of inequality. Roemer proposed that an individual should not be held responsible for the ‘constant’ and should be compensated on the basis of a compensation principle as opposed to reward principle, which is awarded for the hard work of an individual under the term ‘effort’. Circumstances include sex, race, place of birth, social background i.e. parent’s education, parent’s occupation, family wealth; while effort is comprised of years of schooling, hours of work etc. The inequality that rises from the circumstances is known as inequality of opportunity and that rises from efforts are known as inequality due to effort.

1.3 An Inequality of Opportunity Framework

In the discussion of Equality of Opportunity (EoP), we have two encompassing views – Meritocratic and Egalitarian. The Meritocratc view says that inequality in outcomes is due to the inequality of the effort and choices that individuals make, without particularly decomposing the reasons of inequality (de Barros, Ferreira, Vega & Chanduvi, 2009). On the other hand, the Egalitarian view – which is due to Roemer (1993, 1998) - points out both the factors for which the individuals are not responsible (circumstances), and the ones for which they are responsible (efforts and choices). Egalitarianism requires that the distribution of outcomes be independent of circumstances. The underlying idea of egalitarianism is to make a level playing field for all the individuals in the society so that all can achieve the skills they want to (de Barros, Ferreira, Vega & Chanduvi, 2009).

Considering the egalitarian view put forward by Roemer, both circumstances and effort can be held responsible for the outcome inequality in an opportunity unequal society. The differences in circumstances can lead individuals to different opportunity sets of ethnicity, gender, family background, place of birth besides mental (talent, patience, diligence etc.) and physical characteristics (eyesight, disabilities etc.) that they inherit at birth or might acquire at some later stages in life. Even the government policies can be considered as a circumstance as they might end up affecting a certain group of people differentially. The opportunity sets –put simply- are the sets of various factors which open the doors of (dis)advantages to the individuals. Let us consider two variables: gender and ethnicity. The opportunity sets, in that case, will be as follows: {male, black}, {male, white}, {female, black}, {female, white}.

Like other approaches the Roemerian approach has its limitations. A few circumstance factors are stated in the literature but there are many more unnoticed factors which lead the society towards unequal outcome. We are not able to take all of them into account as firstly we have data limitations and secondly we are not certain about the causality between inequality and each of those factors individually. Adding to this, we have another difficulty as the variables are interrelated and one might affect another variable through several channels- directly or indirectly. An individual at the bottom quintile of income or wealth might not get the equal access to the basic necessities (i.e. education, health, sanitation etc.) which in turn lead to an unequal opportunity and at the same time the basic necessities can directly be attributable to opportunity inequality in society. Similarly, the social treatment of children might be affected by the location of the household or the motivation can be affected by the neighborhood where children grow up. Hence, finding out every possible variable and potential channels through which one circumstance might affect another is a daunting task. Therefore, we are not trying to be entirely additive in terms of the factors and exhaustive in pointing out every possible channels. However, we are putting the factors into several groups mentioned below.

1. Differences in talent and motivation: This can overlap with luck, as discussed below i.e. exogenous genetic factors. A person can be inheritably more talented or less talented. This is beyond the individual control. A person with a special talent has a probability to have outcome higher than the person who lacks that talent. This introduces us to Rawls' idea of *natural lottery*. It is disputable whether any compensation should be provided in such case. According to Nozick (1977), the individuals own the natural endowments and hence are entitled to enjoy the full benefit from them. Hence, his idea is not to make compensation for people with a bad natural lottery.

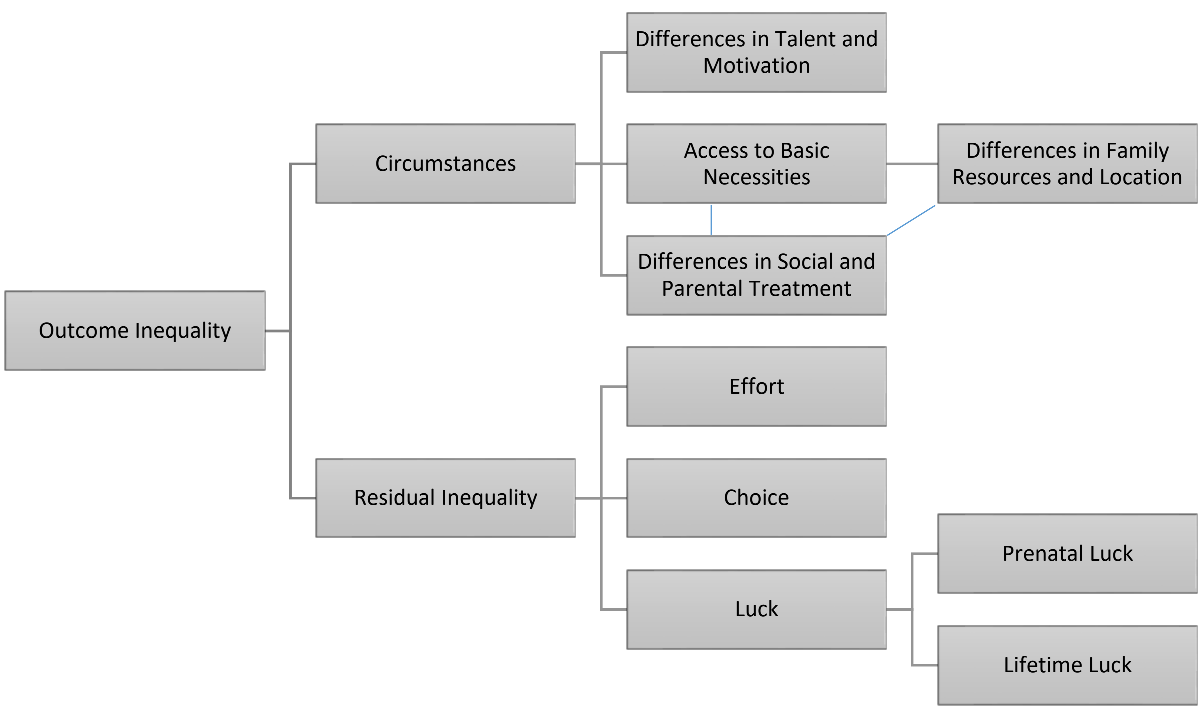

Figure 1.1: Decomposition of Outcome Inequality (Source: (Paes De Barros, Ferreira, Vega & Chanduvi, 2009) (modified))

Also, a person can be less talented due to the lack of nutrition at the early age, from the family impoverishment for which the society ought to make compensation. Motivation also varies with the family background or peers. As stated in Zimmerman (2003), peer effects are highly linked with better SAT scores. These are more intended towards Rawls' idea of *social lottery*.

2. Differences in access/provision of basic opportunities
   a. Differences in social treatment (which indirectly affects the access to opportunities): i.e. differences in parental as well as social treatment of male and female children.
   b. Differences in family resources and location: i.e. parental wealth, birthplace etc.

3. Differences in social treatment (which directly affects the inequality of opportunity)

For the last two factors, the inequality is due to the differences in the family's social status and connections. They fall under the *social lottery* for which it is almost indisputable that the society should provide full compensation.

Other than the factors discussed above (the factors for which the individuals cannot be held responsible), we are left with the factors for which the society is not responsible to compensate (efforts, choice) and the factors for which it is ambiguous whether the society should compensate or not (luck). For effort level and choice making, the individuals are directly responsible whereas luck is a cumbersome factor for which compensation might or might not take place based on the generosity of the entity making the compensation. In order to see the whole picture we need to consider these factors. For better understanding, we group these variables under 'residual inequality'- a term used in Paes De Barros, Ferreira, Vega & Chanduvi (2009). It is crucial to mention that the factors under the residual inequality are conditional on the circumstance factors. The residual inequality factors are described as follows:

1. Effort: The effort that individuals put forward to achieve something is the key to their accomplishments. A person working 40 hours a week, will be better off than a person working 20 hours a week, given they are under the same circumstance group. It is difficult to observe the effort level of the individual. Most common indicators are annual hours of work, years of schooling. However, it is easy to see that effort level depends on many factors starting from social and parental background, talent etc. For example, preference to be educated is one of the key factors in earning education. Keeping in mind the 'sheepskin effect', one might or might not be willing to take education for various reasons, which are not trivial to understand. Whereas a person at a higher circumstance level might be less willing to take education because of the parental wealth or the bequests they are supposed to receive, a person at a lower circumstance level might be less willing to take education, as the opportunity cost might be higher for her and her family. Therefore, the society should compensate for the second case and let the first one go. Furthermore, in the first scenario, the effort level might be affected by the availability of resources and in the second scenario, the lack of resources can be held accountable for the lower effort level. Hence, the circumstance variables can influence the education preference, which has a large effect on the effort to obtain education.

2. Choice/Preference: Another factor for which individuals can be held responsible is the choice they make. If a person chooses not to partake the college education, he will be deprived of the benefits of college education. Similar to effort, the choice individuals make

are dependent on the several circumstance factors as described in Cohen (1989). This corresponds to Dworkin's idea of option luck which entails that 'the risk is taken deliberately, is calculated, isolated, anticipated and avoidable' (Lefranc, Pistolesi and Trannoy, 2009)

3. Luck: Luck is a factor for which a person cannot be held responsible. However, it is dubious whether society should compensate for it or not.

    a. Prenatal luck: Prenatal luck is the treatment that the children receive before their birth which includes, but are not limited to, care and diet. Talent can be considered as a prenatal luck as it can be genetic and might depend on the partner choice of the parents. Also, lack of talent can be due to the malnutrition of mother during pregnancy. The children of poor parents have a higher probability of being malnourished and hence can be less talented (Brown & Pollitt, 1996). Birth associated disabilities (such as oxygen deprivation causing brain damage) can also be put under the prenatal luck. For these factors, egalitarians argue that compensation should be provided.

    b. Lifetime luck: In contrast, the lifetime luck is luck experienced after birth. It can be regarded as luck comprised of lottery outcomes and accidents. An example is early pregnancy in life which can hinder the further progress of both parents or the demise of parents at an age earlier than expected. Now this is debatable whether compensations should be provided in such cases. However, compensation can be put forward for a person who experienced a disability at a later stage of life i.e. experiencing an accident which leads to a permanent loss of eyesight of an archaeologist or claudication of a coal miner.

Based on the discussion above, we can say that inequality depends on factors that can be divided into two groups – circumstance and residual factors.

$$\text{Inequality (I)} = f\ (c, r(c))$$

Where c is the circumstance vector and r stands for the vector of residual causes of inequality.

1.4 Data

The data that are used are gathered from the "Panel Study of Income Dynamics" (PSID), a longitudinal survey conducted by the University of Michigan. We used the data for 1968 through 2013. The dataset contains information on the social background and demographic composition of each household in the survey.

The base explanatory variables used in this research are father's/mother's education (FE/ME), father's occupation (FO), socioeconomic condition (SC), race (RA), and gender (GE). These will be called the 'social factors.' The first step in processing the data is the removal of the data items that were not used in the analysis. All the variables that entered the analysis were subsequently categorized. Parental education was categorized into less than high school education, high school graduates, and at least some college education. The occupation of father was categorized into laborers and farmers, craftsmen and operatives, and lastly clerks, professionals and managers. We separate the income groups into poor, middle class, and rich. All the individuals in the sample are of either black or white racial groups. The limited numbers of individuals in the sample do not let us consider the individuals of other races. The genders considered are either male or female.

Other than the above-mentioned social factors, we also have the birth years of individuals, which enter as polynomials. The birth year of an individual is derived in two steps. First, we calculate a reported birth year for each time an individual is interviewed, by subtracting the reported age in that year from the interview year. In the second step, using the modal value of the reported birth year, we record the standardized birth year for every individual. We analyze the impact of birth year as a polynomial in the standardized value of the birth year, as we suspect that in our model, birth year has a fluctuating influence that can best be captured by a polynomial of a potentially large degree. We use the Chebyshev polynomials to maintain orthogonality among the components of the polynomials.

Table 1.1: List of Explanatory Variables

Variables	Category 1	Category 2	Category 3
Father's Occupation (FO)	Farmers and Laborers	Craftsmen and Operatives	Clerks, Managers, and Professionals
Parental Education (FE/ME)	Less than High School	High School Graduates	Beyond High School
Income group (IN)	Poor	Middle Class	Rich
Race (RA)	White	Black	
Gender (SE)	Male	Female	
Birth Year (BYST)	1920-1988		

The dependent variables are four binary variables representing the educational attainment of the individuals with respect to four hurdles: high school graduation, some college, college graduation, and some post-college work. For each hurdle, we have a separate set of variables selected through a careful examination of the data.

We collapsed the data, based on the unique IDs of the individuals. When data for an individual were inconsistent, we took the first non-missing values for the gender, race, father's occupation, father's education, mother's occupation, and economic status; last non-missing values for the educational attainment; and the modal value of the implied birth year.

Based on the implied birth year, we have data for individuals born as far back as 1881. However, because of a paucity of early observations, we are unable to use the whole dataset. We are using only the data for individuals born in or after 1920. For each of the hurdles, we have set an age constraint for inclusion in the final data set. For high school graduation, we are not using data for the individuals graduating above the age of 25 years, and above 27 years for a first report of some college education. For the college and post college education, we have used a cut-off of 29 years old. After implementing these restrictions, we have 23,749 observations for possible high school graduation, 17,906 observations for high school graduates who might have some college education, 9,058 observations for college students who might graduate from college, and 4,506 observations of college graduates who might have some post-college education.

The figures below show the distributions of our sample over the years. We have fewer observations for the early 20th century across all of the hurdles.

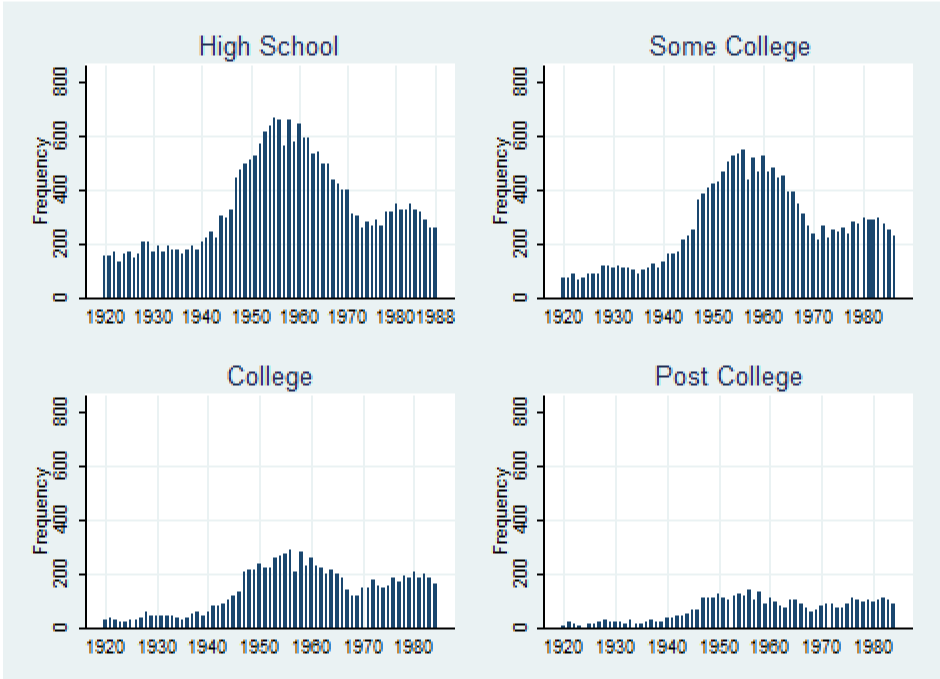

Figure 1.2: Sample Distribution across the Birth Years across the Hurdles (BY 1920-1988) [Source: PSID]

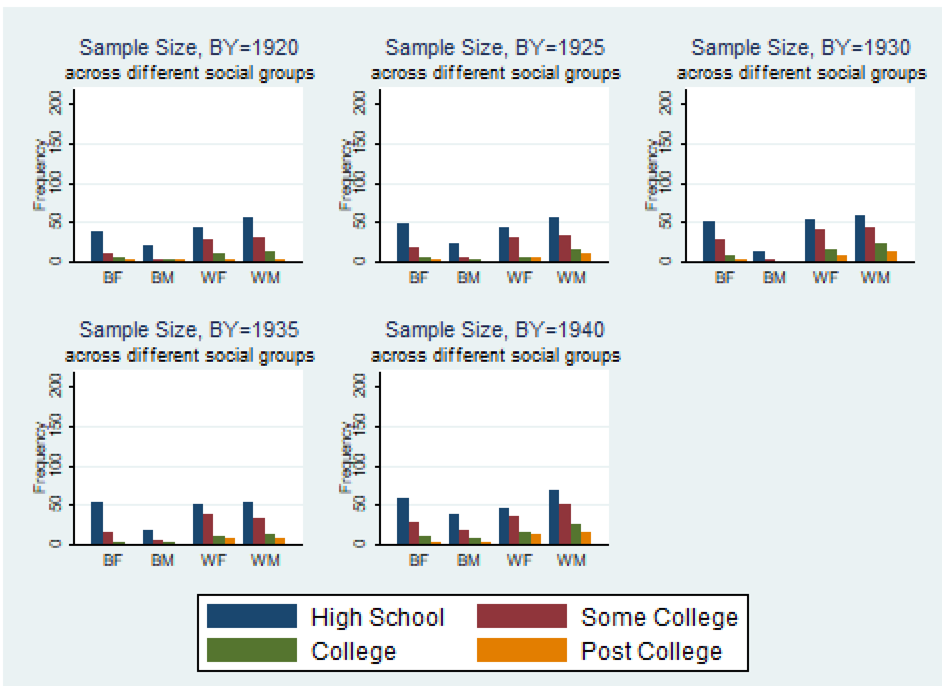

Figure 1.2a: Sample Distribution across Race and Gender (BY 1920-1940)

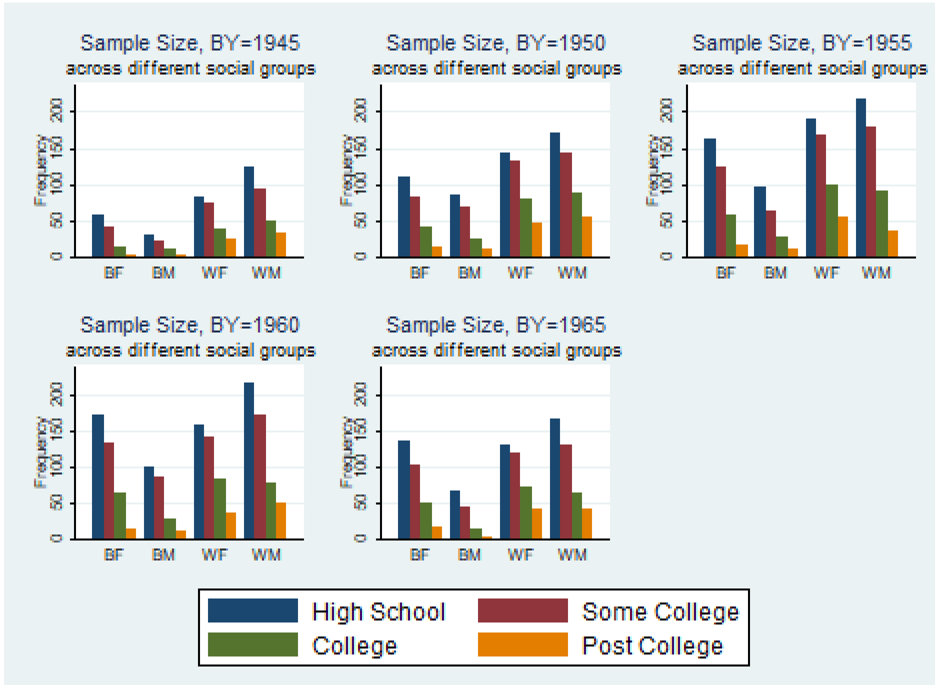

Figure 1.2b: Sample Distribution across Race and Gender (BY 1945-1965)

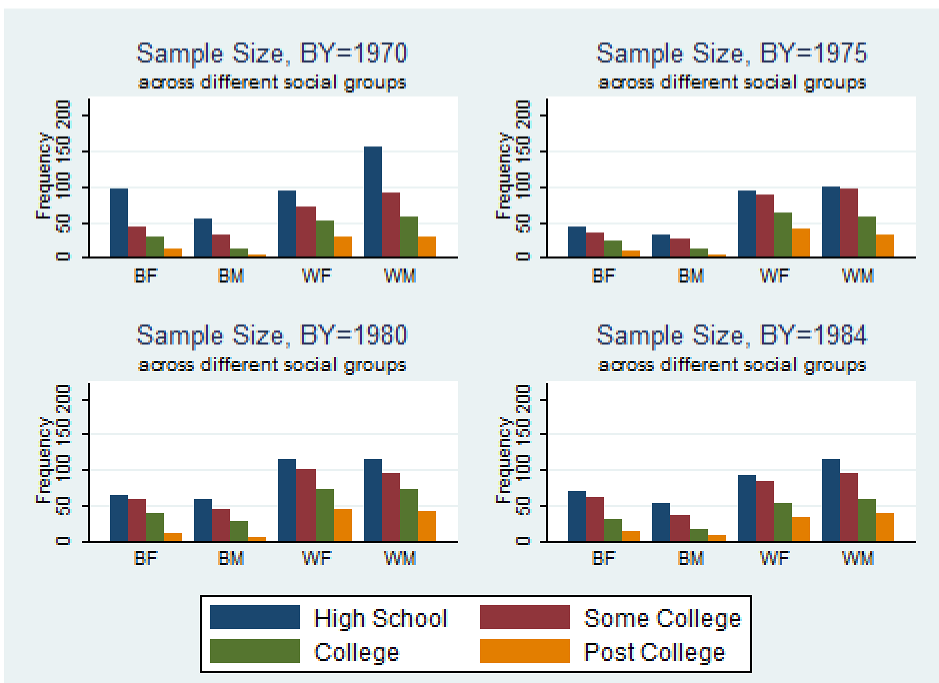

Figure 1.2c: Sample Distribution across Race and Sex (BY 1970-1984)

1.5 Measuring Inequality of Opportunity

The method used in this paper addresses some issues with the existing methodologies. It lets us choose the circumstances variable systematically, rather than choosing haphazardly from the large number of variables in the existing literature. Moreover, it provides the opportunity to build a social advantage factor endogenously. This addresses the problem of heterogeneous circumstances of the individuals; we can identify them according to their overall circumstances, i.e. offsetting lower parental endowments with higher social endowments. As we are looking only into the circumstances variables, and, as discussed earlier, some of the effort variables can depend on circumstances, we experience omitted variable bias. The coefficients are biased upward. We follow Ferreira and Gignoux (2014) in saying that our function *F(X)* is a valid upper bound for the outcome of interest. If any other circumstances are going to be added, it is only going to lower the probability and hence increase the inequality.

We build a model for each of our educational achievements, or 'hurdles,' by following a machine- learning process. The primary reason for following a machine-learning process is the presence of a large number of candidate features, discussed later in this section. Prior to 1997, very few analyses used more than 40 features (variables) to explore any idea. The invention of new techniques changed the practice, and many analyses now use hundreds to tens of thousands of features (Guyon & Elisseeff, 2003). There are arguments for selecting a subset of features that has good predictive power, as opposed to ranking the predictive power of individual features. Following Kohavi and John (1997) and Guyon and Elisseeff (2003), we choose to select a subset of features, since a ranking of the individual features would throw away some of the features that are redundant and yet relevant.

Methods of variable selection are divided into wrapper, filter, and embedded methods. We use the wrapper method, which takes the process of variable selection as a black box (no prior knowledge) to score subsets of variables based on their predictive power. Different search strategies – backward elimination, best-first, branch-and-bound, simulated annealing, genetic algorithms – are available. However, because of its computational advantage and simplicity, we implement the forward selection method for our feature selection.

Our goal is to look into the inequality in educational attainment that arises from the inequality in social factors. The base features that we take into account are birth year, parental education, father's occupation, socioeconomic condition, race, and gender. Some of these features appear in the existing literature—Bourguignon, Ferreira and Menéndez (2005), Peragine and Serlenga (2007), Pistolesi (2008), Marrero and Rodriguez (2010). However, none of the papers has used all of the variables that we are using in this paper. Another difference between this paper and the existing papers is the way the features are chosen. We selected them without any prior knowledge. We want to explore the effects, if there are any, of these features at the different levels of educational achievement. We anticipate having interdependence among the variables (i.e., father's education can affect his occupation, socioeconomic condition can be largely dependent on parental factors), and hence we expand our feature set by constructing conjunctive features, or interaction terms. We also have the presence of polynomials in the birth year. After selecting the base features, we explored the possibility of adding polynomial terms for birth year. In ordering the selection of variables, we follow Guyon and Elisseef (2003), who say, "…it is more reasonable to use a wrapper with linear predictor as a filter and then train a more complex non-linear predictor on the resulting variable." Out of concern for losing too many degrees of freedom, we do not go beyond two-way interactions between pairs of social factors. However, we do use birth year polynomials in the conjunctive features.

For the addition of each of the features, a likelihood-ratio test is conducted, and for the polynomials (including the conjunctive features), local and global tests are conducted. Based on this, our function of interest is,

$$Y = F(X, Z)$$

$$Y = Probability\ of\ Obtaining\ y\ level\ of\ education;$$

$$X = \{\ polynomials\ in\ time\}; Z = \{base\ variables, interaction\ terms\}$$

The separate models for each hurdle (*y*) are as follows:

$$\boldsymbol{Model\ 1:\ \pi^1(high\ school\ graduation) = \alpha + \beta^1 X^1 + \gamma^1 Z^1 +}\ \varepsilon$$

where,

$$X^1 = \{\sum_{i=1}^{8} BY^i\}$$

$$Z^1 = \{\sum_{i=0}^{8} FO * ME * BY^i, \sum_{i=0}^{8} FO * FE * BY^i, \sum_{i=0}^{8} FO * IN * BY^i, \sum_{i=0}^{8} FO * RA * BY^i, \sum_{i=0}^{5} ME * IN * BY^i, \sum_{i=0}^{8} FE * RA * BY^i, \sum_{i=0}^{1} IN * RA * BY^i, \sum_{i=1}^{8} IN * SE * BY^i, \sum_{i=1}^{8} SE * RA * BY^i\}$$

$$\boldsymbol{Model\ 2: \pi^2(some\ college\ education) = \alpha + \beta^2 X^2 + \gamma^2 Z^2 +}\ \varepsilon$$

where,

$$X^2 = \left\{\sum_{i=1}^{8} BY^i\right\}$$

$$Z^2 = \{\sum_{i=0}^{8} FO * ME * BY^i, \sum_{i=0}^{7} FO * FE * BY^i, \sum_{i=0}^{8} FO * IN * BY^i, \sum_{i=0}^{8} FO * RA * BY^i, \sum_{i=0}^{8} FO * SE * BY^i, \sum_{i=0}^{8} IN * RA * BY^i, \sum_{i=1}^{8} IN * SE * BY^i, \sum_{i=1}^{8} SE * RA * BY^i\}$$

$$\boldsymbol{Model\ 3}: \boldsymbol{\pi^3}(\boldsymbol{college\ graduation}) = \boldsymbol{\alpha} + \boldsymbol{\beta^3 X^3} + \boldsymbol{\gamma^3 Z^3} + \varepsilon$$

where,

$$X^3 = \{\sum_{i=1}^{7} BY^i\}$$

$$Z^3 = \{\sum_{i=0}^{2} FO * FE * BY^i, \sum_{i=0}^{5} FO * IN * BY^i, \sum_{i=0}^{8} FO * RA * BY^i, ME, \sum_{i=0}^{8} FE * RA * BY^i, \sum_{i=0}^{5} IN * RA * BY^i\}$$

$$\boldsymbol{Model\ 4}: \boldsymbol{\pi^4}(\boldsymbol{starting\ post\ college\ education}) = \boldsymbol{\alpha} + \boldsymbol{\beta^4 X^4} + \boldsymbol{\gamma^4 Z^4} + \varepsilon$$

where,

$$X^4 = \{\sum_{i=1}^{4} BY^i\}$$

$$Z^4 = \{\sum_{i=0}^{1} FO * RA * BY^i\}$$

These equations are the results of machine-learning processes. We run logit regressions based on these models for each educational achievement. We present the estimates of the log odds of success with the hurdles based on the following logit model

$$ln[\frac{P(Y = y|X, Z)}{1 - P(Y = y|X, Z)}]$$

We find the mean log-odds (MLO) for each combination of paternal endowments (FE, FO), each maternal endowment (ME), and each family endowment (IN). Based on these values of mean log odds, we build a social-advantage factor (SAF) for each individual and use it to define four quartiles: most advantaged, relatively more advantaged, relatively less advantaged, and least advantaged.

$$SAF = Q_i^{FOFE} + Q_j^{ME} + Q_k^{IN}$$

where $i, j, k \in z$ and Q_z^X represents the mean log-odds for group m in factor X.

Each group contains one quarter of the population. We transform the log-odds into a probability *(F(X))* that follows a logistic function, to measure opportunity inequality. The probability of individual *i* in social group *m* completing hurdle *y* is modeled as,

$$F_{i,m}^{y}(X, Z) = \widehat{p_{\iota,m}} = \frac{\exp(x_{i,m}\hat{\beta} + z_{i.m}\hat{\gamma})}{1 + \exp(x_{i,m}\hat{\beta} + z_{i,m}\hat{\gamma})}$$

by taking the difference between the reference social strata and the relevant social strata.

$$Opportunity\ Inequality, IO = G_j^y(X, Z);$$

$$G_j^y = F_a^y(.) - F_j^y(.)\ ; a, j \in m$$

$$where\ a\ is\ the\ reference\ group\ (most\ advantaged) and\ j\ is\ any\ other\ group$$

For race(r) and gender (s) analysis,

$$G_{j,r,g}^{y} = F_{a,r,s}^{y}(.) - F_{j,r,s}^{y}(.)\ ; a, j \in m$$

1.6 Results

We have separate results for each hurdle. For each hurdle, we have shown the probability of graduating/attending the hurdle and the related inequality of opportunity (IO). We further anatomize the results for different race and gender groups.

1.6.1 High School

Figure 1.3 depicts the inequality of opportunity (IO) in the United States of America (USA). The left panel shows the probability of graduating from high school while the right panel shows the inequality of opportunity, which was calculated by taking the difference of the reference category (most advantaged) and the current category. IO was the maximum for the individuals born in the early twentieth century and has ebbed continuously since then. The results are also presented in Table 1.2.

On average, individuals who are the least advantaged and born in the 1920s, had a staggering 71–percentage point (pp) difference in the probability of graduating from high school with the individuals of the same cohort and from the most advantaged group and born in the same year. The difference came down to 32- pp for individuals born in 1950 and 8- pp for those born in 1988 – an ever-decreasing trend. The same is true for relatively less advantaged group. The IO was 43-pp for the 1920 cohort but only around 7-pp for the 1988 cohort. For the more advantaged groups, the probability dipped in 1930 and 1940. In other years, the probability of graduating for these groups were almost perfect.

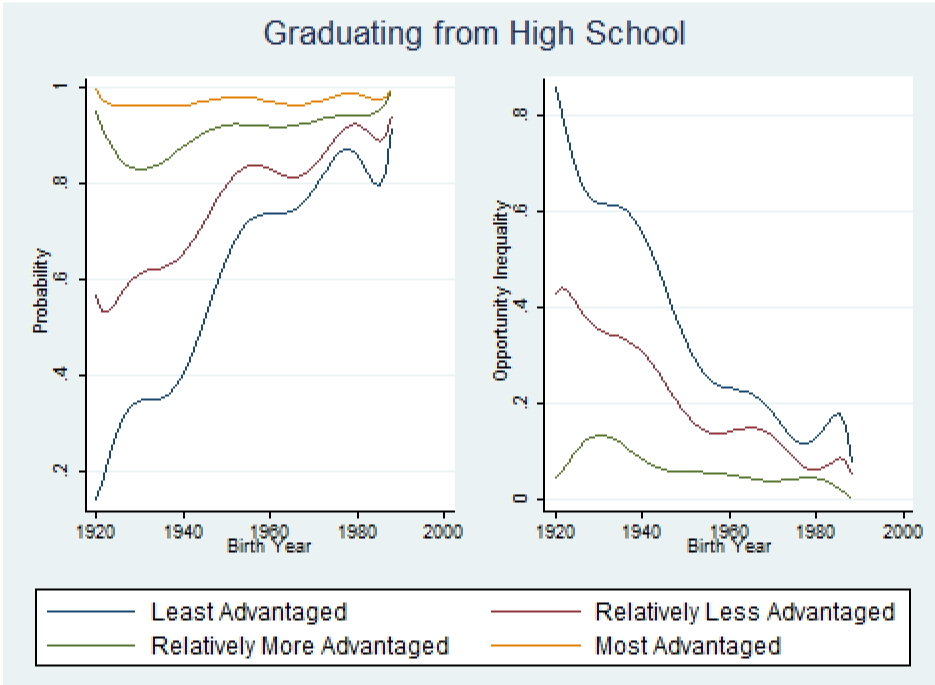

Figure 1.3: Probability/Opportunity Inequality of High School Graduation

The paper also investigated the differences in probability across the race and sex by controlling the social advantage factor (SAF). Figure 1.3a and 1.3b depict the results. The most advantaged group led the line with the highest probability across all the race and gender groups throughout. The figure is consistent with what we have seen in Figure 1.3 – the most advantaged groups leading the three groups with the probability being monotonic among the social groups – the more advantaged groups have higher probability of graduation.

Table 1.2: Probability (Average) of Graduating from High School

Birth Year	Least Advantaged	Relatively Less Advantaged	Relatively More Advantaged	Most Advantaged
1920-1929	0.26	0.56	0.88	0.97
1930-1939	0.36	0.63	0.85	0.96
1940-1949	0.51	0.72	0.90	0.97
1950-1959	0.71	0.83	0.92	0.98
1960-1969	0.75	0.82	0.92	0.97
1970-1979	0.84	0.89	0.94	0.98
1980-1988	0.83	0.91	0.94	0.98

The IO was much bigger between the blacks and the whites in the early 20th century with the whites, in general, leading the line[9]. A white male born in 1920 and is in the least advantaged group leads his black counterpart by 18- pp. The difference is just 5-pp in 1988. For the most advantaged group, the white males born in 1988 leads the black males only by 2-pp. The advantaged groups born in 1988 had nearly the full chance of graduating from high school.

The females – regardless of the race and social advantage groups – dominate the males in the probability of graduating from high school in most of the years reported in Table 1.A.1.1. Although the females led the line, the male were never far behind. For 1988 cohort, however, the males led the females across all the groups. The least advantaged white male born in 1930 had 15-pp difference with least advantaged white female. The difference was 12-pp for the black population. In 1970, the difference was 10-pp and 5-pp respectively. For the most advantaged group the difference was 3-pp for the whites and 4-pp for the blacks in 1930, while in 1980 there is no difference between the whites and only 3-pp for the blacks.

[9] Results are presented in Table 1.A.1.1 in the appendix. Figure 1.A.2.5 – 1.A.2.8 in the appendix show all the graphs (Figure 1.3 – Figure 1.6) with 95% confidence interval.

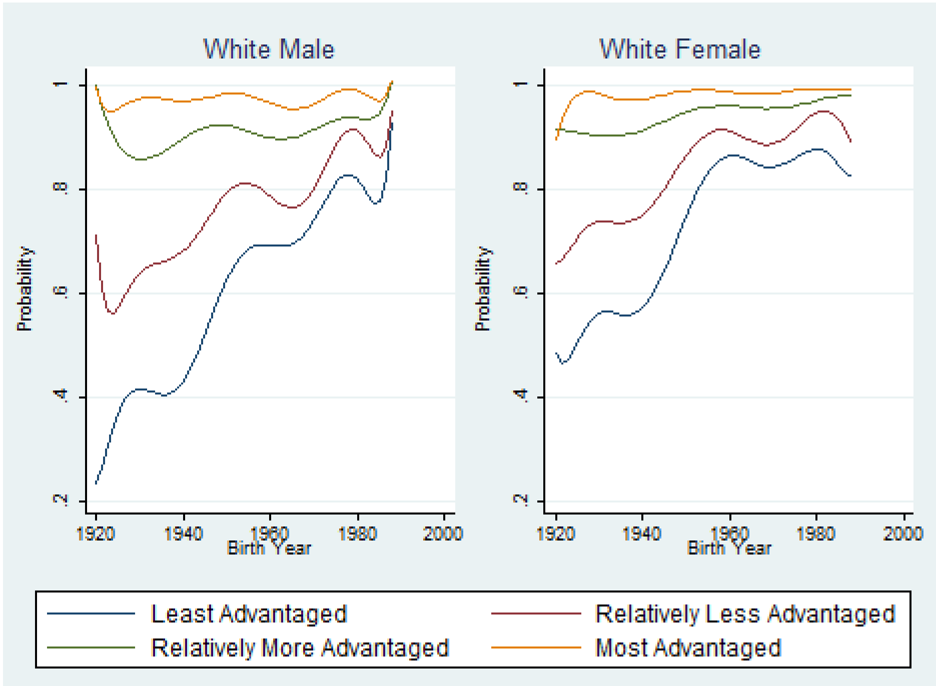

Figure 1.3a: Probability of High School Graduation across Race and Gender (White)

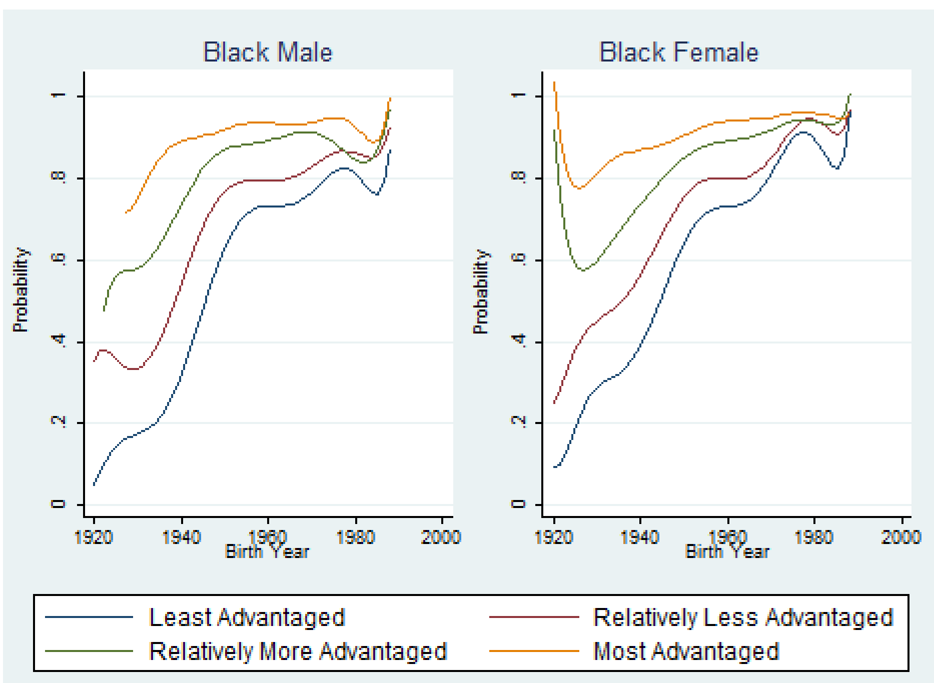

Figure 1.3b: Probability of High School Graduation across Race and Gender (Black)

1.6.2 Some College

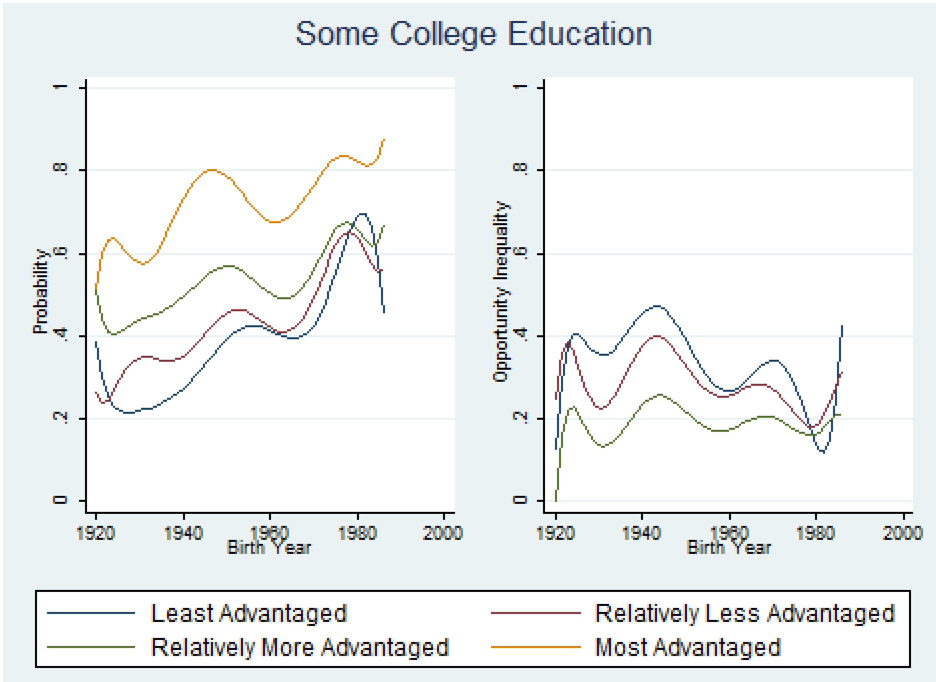

Figure 1.4: Probability/Opportunity Inequality of Attending Some College

The individuals who has graduated from high school are only considered for this analysis. Probability of attending some college has an upward trend across the cohorts, and is much flatter than the probability of high school graduation. The figure shows a slight downward trend in IO over the years, as compared to the steep downward trend for the high school graduation. The higher up the groups are, the higher is the probability of obtaining some college education. Born in 1920s, the least advantaged group are 34-pp behind than the most advantaged group. The difference came down to only 20-pp for the 1980s cohorts. The least advantaged groups are the most unequal followed by the relatively less advantaged and relatively more advantaged groups respectively. The most advantaged groups have the highest probability of attending some college. For the relatively more advantaged group, the IO remained persistent over the cohorts. The fluctuation is not very frequent for the other two social groups too.

The order of the social groups remains as a monotonic function of the social groups for the race and sex analysis. The white males of the most advantaged group of 1970s cohorts had 35-pp more chance of attending some college than the males in the least advantaged group and 20-pp higher probability than the relatively most advantaged males. The black males from the top social stratum had 60 percent probability of attending some college while the black males from the bottom stratum had only 44 percent probability. Among the females, the whites from the top stratum had 87 percent probability and the blacks from the same stratum had 78 percent probability – a difference of 9-pp. In general, the whites, regardless of the gender, led their black counterparts of the same gender in attaining some college education. The difference is much lower between the lowest social stratum of the two races and grows bigger as we go up to the social ladder. That means as the blacks climb up the social ladder, the attainment of some college education is relatively lower than their white counterparts from the same social strata. For the 1960s cohorts,

Table 1.3: Probability (Average) of Attending Some College

Birth Year	Least Advantaged	Relatively Less Advantaged	Relatively More Advantaged	Most Advantaged
1920-1929	0.26	0.29	0.43	0.60
1930-1939	0.24	0.34	0.46	0.62
1940-1949	0.33	0.40	0.53	0.78
1950-1959	0.41	0.45	0.54	0.73
1960-1969	0.40	0.43	0.51	0.70
1970-1979	0.54	0.59	0.63	0.81
1980-1984	0.63	0.59	0.64	0.83

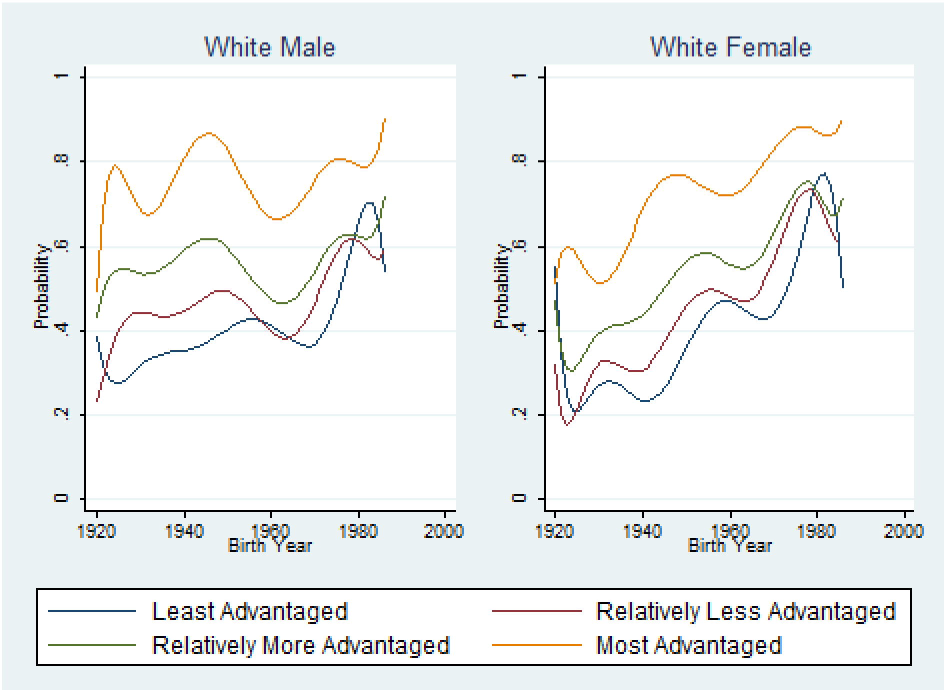

Figure 1.4a: Probability of Attending Some College across Race and Gender (White)

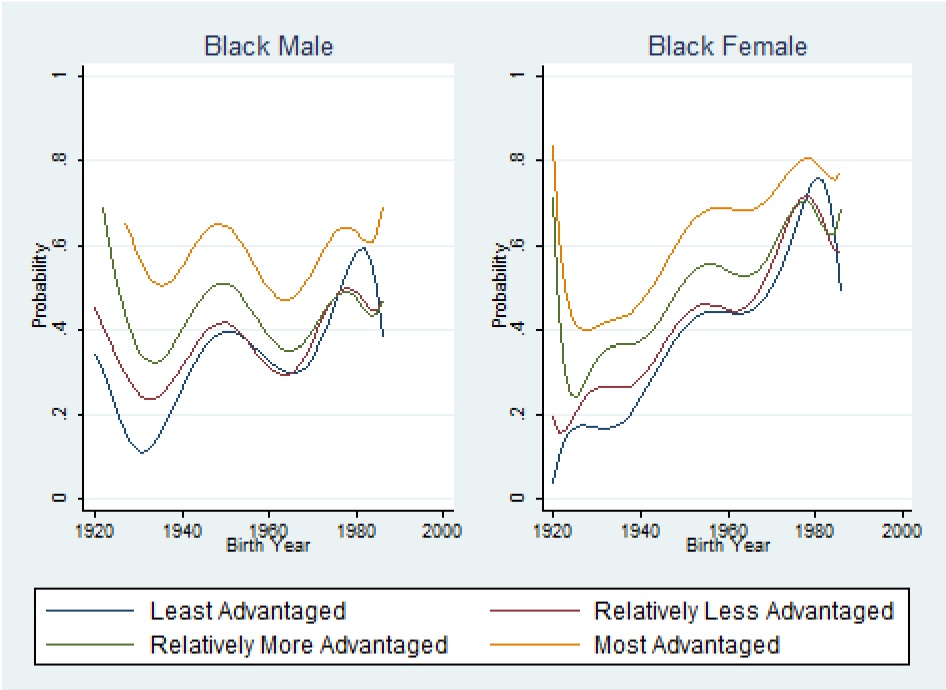

Figure 1.4b: Probability of Attending Some College across Race and Gender (Black)

the difference between the least advantaged groups of the two races is 7-pp, while it is 19-pp for the most advantaged group. The white females led the black females across all the strata and cohorts, except the lowest social strata where the black females – in few cohorts – led their white counterparts. Regardless of the race, the females – on average – led the males in attending some college education from 1950 onwards. The probability of attaining some college has increased over the years nonetheless. The inequality gap between the white and black males widened while it narrowed down for the females of the lower social strata. The results are presented in Table 1.A.1.2.

1.6.3 College

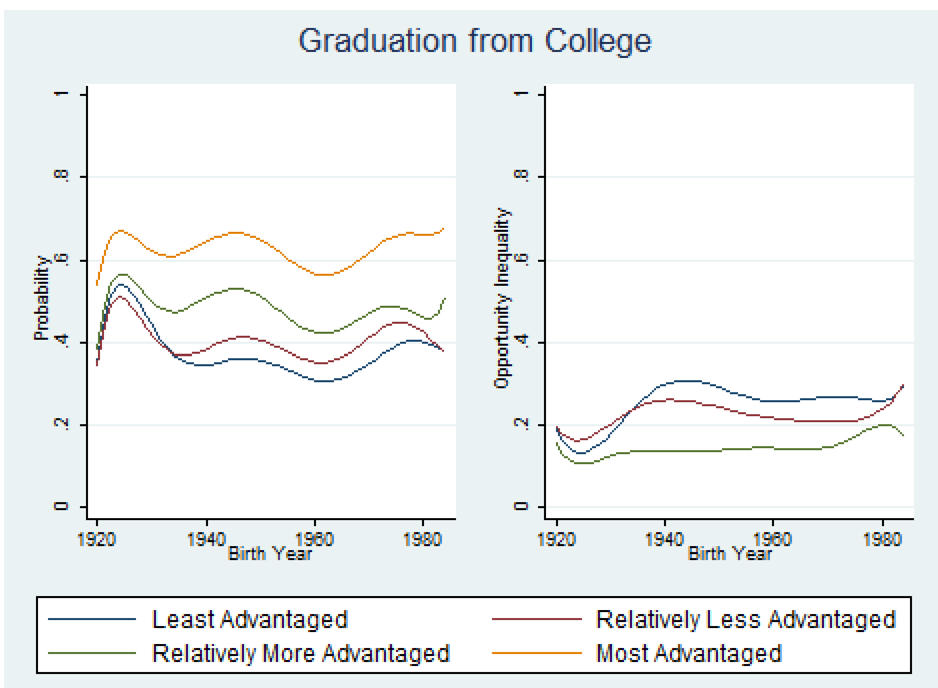

Figure 1.5: Probability/Opportunity Inequality of Graduating from College

Figure 1.5 presents the probability of graduating from college and the inequality of opportunity across the social advantage groups. Except for the early cohorts, the probability of graduating from college had almost persisted across the cohorts. The inequality of opportunity graph has experienced an upward turn until the 1940 cohorts for the two social strata from the bottom while, for the most advantaged group the inequality persisted. After 1940, the least advantaged group has the highest IO across the cohorts and it consistently went up until 1940.

As presented in Table 1.4, the 1920s cohorts had a 16-pp difference between the least and the most advantaged group. For the same groups, the 1950s and 1980s cohorts had 27-pp difference. The 1920s relatively less advantaged cohort had only 16-pp difference, which increased to 27-pp for the 1980s cohort of the same tier. The difference between the top two tiers remained consistent across the cohorts.

Table 1.4: Probability (Average) of Graduating from College

Birth Year	Least Advantaged	Relatively Less Advantaged	Relatively More Advantaged	Most Advantaged
1920-1929	0.47	0.47	0.53	0.63
1930-1939	0.39	0.38	0.48	0.62
1940-1949	0.35	0.40	0.53	0.66
1950-1959	0.34	0.38	0.46	0.61
1960-1969	0.32	0.37	0.44	0.57
1970-1979	0.38	0.44	0.48	0.65
1980-1984	0.39	0.39	0.48	0.66

Figure 1.5a and 1.5b shows us that the probability of graduating from college is a monotonic function of the social strata. For college graduation there is a virtual equality among the whites and the blacks meaning the gender difference does not have a major impact on earning a college degree[10]. If we ignore some of the anomalies (that arises because lack of observations) at the earlier cohorts, the maximum difference found between the black males and females were 8-pp, and 6-pp

[10] Figure 1.A.2.3a, Figure 1.A.2.3b, and Table 1.A.1.3

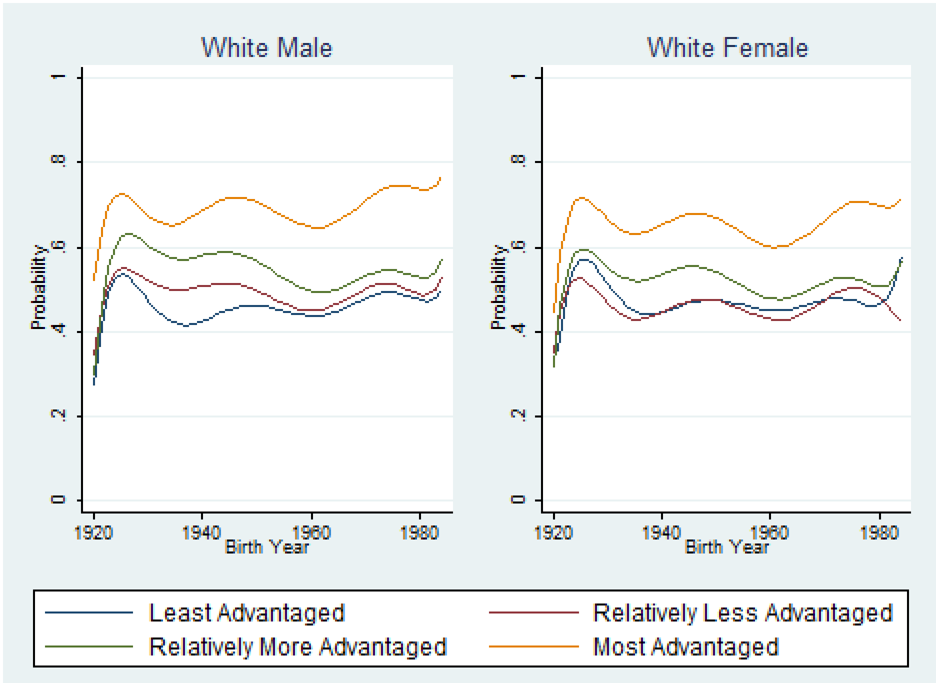

Figure 1.5a: Probability of College Graduation across Race and Gender (White)

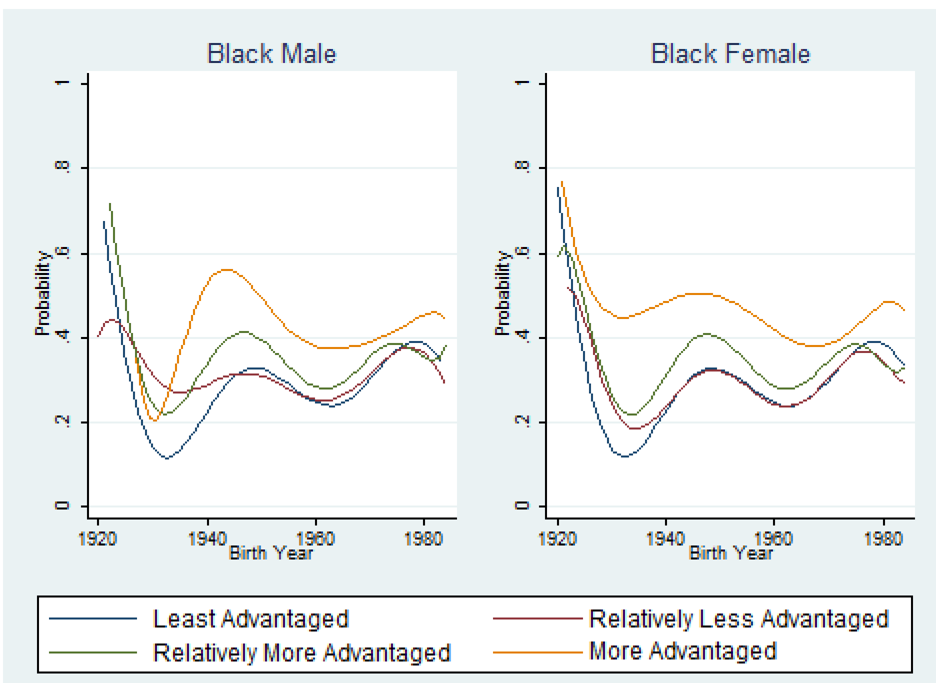

Figure 1.5b: Probability of College Graduation across Race and Gender (Black)

for the white males and females, for the relatively less advantaged cohort of 1930s where the males were leading. Within the same gender, the whites had the higher probability of having college degree. There is a 30-pp difference between the white and black males and 26-pp difference between the white and the black females of the most advantaged 1970s cohorts.

1.6.4 Post College

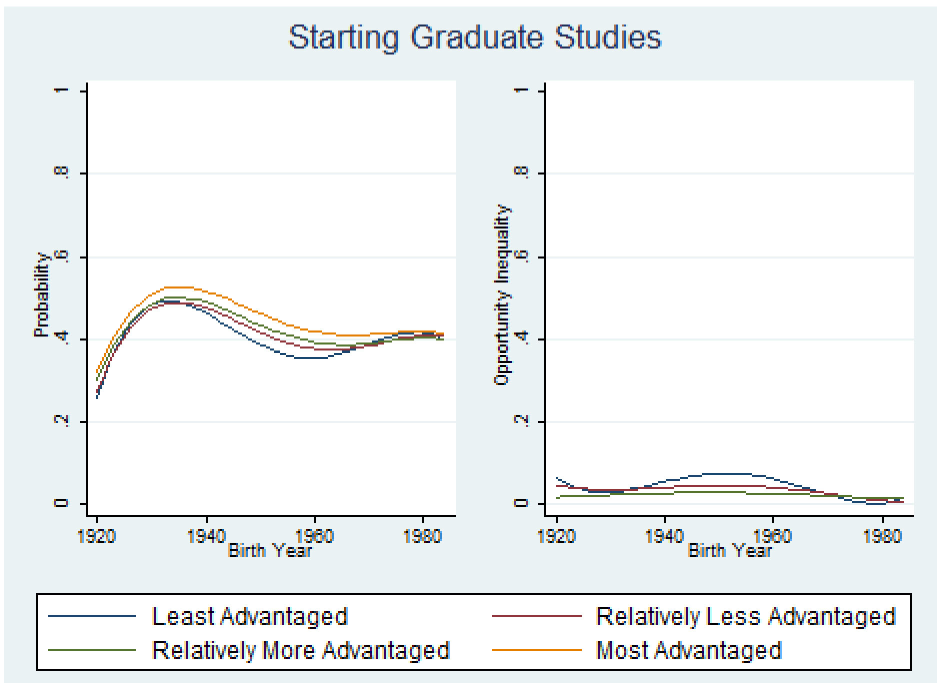

Figure 1.6: Probability/Opportunity Inequality of Starting Post College Education

The most advantaged group has led in starting the post college studies. The probability of starting the post-graduate studies increased until 1940s cohorts, continued to decline for all the social strata but the bottom two which experienced an increase in probability afterwards to converge with the other social groups. The IO shows the shape of a normal distribution for the least advantaged group, and the persistent but vanishing IO for the other two groups. For 1920s cohorts, the IO was

3-pp and 5-pp for the bottom two social strata respectively. The difference was 7-pp and 5-pp, for 1950s cohorts, between the least advantaged and most advantaged, and the relatively less advantaged and the most advantaged group respectively. For the 1970s cohorts, the IO was only 0-pp, 1-pp and 2-pp for the bottom 3 tiers respectively. Although IO graph shows varying trends across the social strata, the inequality remained much smaller. This indicates the decreasing role of social factors in postgraduate education.

Table 1.5: Probability (Average) of Starting Post College Education

Birth Year	Least Advantaged	Relatively Less Advantaged	Relatively More Advantaged	Most Advantaged
1920-1929	0.40	0.38	0.41	0.43
1930-1939	0.49	0.49	0.49	0.52
1940-1949	0.40	0.45	0.47	0.49
1950-1959	0.37	0.39	0.41	0.44
1960-1969	0.38	0.38	0.39	0.41
1970-1979	0.39	0.40	0.40	0.42
1980-1984	0.42	0.41	0.40	0.42

For starting post college education, there is no difference between the two genders. Among the white population, there is no apparent differences between the bottom 2 tiers (Figure 1.6a and Figure 1.6b). For the same tiers, the difference is very small for the black population. The whites led the blacks until 1960s cohorts, after which the blacks, in general, has been leading the whites. While the blacks have been experiencing an upsurge in probability of entering graduate school, the whites have been experiencing the opposite. The results are deducted from the table Table 1.A.1.4, and Figures 1.A.2.4 (a, b) below.

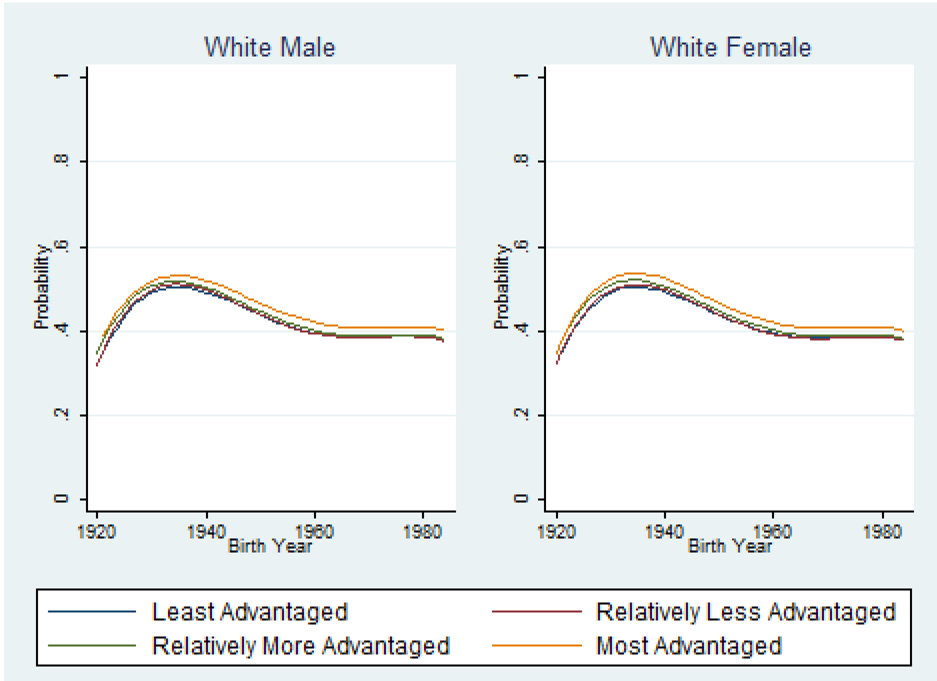

Figure 1.6a: Probability of Starting Post College across Race and Gender (White)

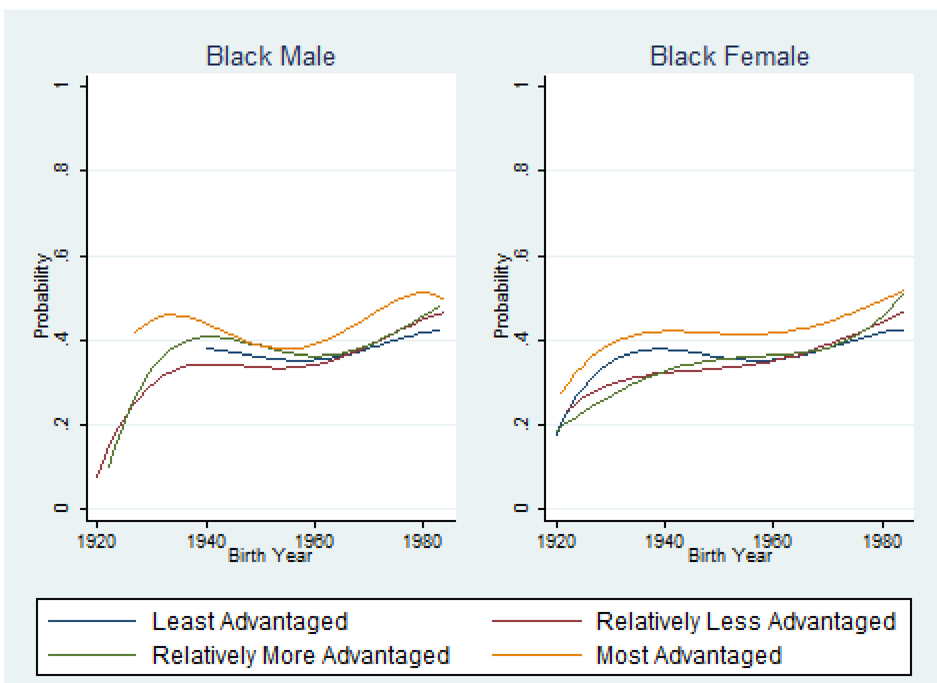

Figure 1.6b: Probability of Starting Post College across Race and Gender (Black)

1.7 Conclusion

The purpose of this paper is to examine the inequality of opportunity in education that arises because of the differences in social factors. We have examined the differences in educational attainment arising from social factors across the social groups (using our own index), races, and genders. We have found some interesting results. In USA, the inequality of opportunity has seen a consistent decline for high school completion. The inequality of opportunity (IO) declines for obtaining some college education for the bottom two social groups and remained persistent for the relatively more advantaged group. For college/post-college education, the IO is much lower and, in general, remained persistent across the social strata.

Individuals from the top-most social strata have nearly the full chance of graduating from high school, regardless of the race and gender. Blacks were far behind Whites in graduating from high school in the earlier cohorts, but the gap eventually narrowed down. Females are however leading males. For some college education, black males are trailing their white counterparts while white females are leading the black females, except for the least advantaged group. The difference in probability between the blacks and whites has seen a consistent upsurge. Males are leading their female counterparts, regardless of the races, from 1950s cohorts onwards. For college education, the difference in probability of graduation is much higher between the races than it is between the genders. Interestingly, between the two races, the disparity rose as we moved up to the social strata. For both college and post-college, the probability of two genders were converging while the two races experienced a persistent gap between them. Inequality of opportunity gets lower in magnitude as we move up to the education ladder which means the social factors become less and less relevant. This implies that, although it is much harder for someone with lower social attributes to obtain education at the beginning of the educational process; the difficulty fades if s/he persists.

References:

Aaberge, R., Mogstad, M. & Peragine, V.: Measuring long-term inequality of opportunity, Journal of Public Economics 95, 193-204 (2011)

Alesina, A, & Angeletos, G.M.: Fairness and Redistribution. The American Economic Review, 95(4): 960-980 (2005)

Arneson, R. J.: Equality and equal opportunity for welfare. Philosophical studies, 56(1), 77-93 (1989)

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

Bailey, M. J., & Dynarski, S. M.: Gains and gaps: Changing inequality in US college entry and completion (No. w17633). National Bureau of Economic Research (2011)

Betts, J. and Roemer, J. E.: "Equalizing opportunity for racial and socioeconomic groups in the United States through educational finance reform", in Schools and the Equal Opportunity Problem, P. Peterson (ed.), Cambridge, M.A.: The MIT Press (2007)

Bossert, W.: Redistribution mechanisms based on individual characteristics. Mathematical Social Sciences 29, 1–17 (1995)

Bound, J., & Turner, S.: Going to war and going to college: Did World War II and the GI Bill increase educational attainment for returning veterans? Journal of labor economics, 20(4), 784-815 (2002)

Bourguignon, F., Ferreira, F.H.G. & Menendez, M.: Inequality of opportunity in Brazil. Review of Income and Wealth 53, 585-618 (2007)

Bowles, S.: Understanding unequal economic opportunity. The American Economic Review, 63(2), 346-356 (1973)

Brown, J. L., & Pollitt, E.: Malnutrition, poverty and intellectual development. Scientific American, 274(2), 38-43 (1996)

Buchmann, M., Charles, M., & Sacchi, S.: The lifelong shadow: Social origins and educational opportunity in Switzerland. Persistent inequality: Changing educational attainment in, 13, 177-192 (1993)

Checchi, D. & Peragine, V.: Inequality of opportunity in Italy. Journal of Economic Inequality 8, 429-450 (2010)

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N.: The equality of opportunity project. Retrieved November, 26, 2017 (2013)

Cobalti, A., & Schizzerotto, A.: Inequality of educational opportunity in Italy. Persistent inequality: Changing educational attainment in, 13, 155-176 (1993)

Cohen, G. A.: On the currency of egalitarian justice. Ethics, 99(4), 906-944 (1989)

Conlisk, J.: Can equalization of opportunity reduce social mobility?. American Economic Review 64, 80-90 (1974)

Graaf, P. M., & Ganzeboom, H. B.: Family background and educational attainment in the Netherlands of 1891-1960 birth cohorts. Persistent inequality: Changing educational attainment in, 13, 75-100 (1993)

de Barros, R. P., Ferreira, F. H., Vega, J. R. M., & Chanduvi, J. S. (2009). Measuring Inequality of Opportunities in Latin America and the Caribbean. World Bank Publications (2009)

Dardanoni, V., Fields, G. S., Roemer, J. E., & Puerta, M. L. S. How demanding should equality of opportunity be, and how much have we achieved?. Mobility and inequality: Frontiers of research from sociology and economics, 59-82 (2006)

DiPrete, T. A., & Buchmann, C.: The rise of women: The growing gender gap in education and what it means for American schools. Russell Sage Foundation (2013)

Dworkin, R.: What is equality? Part 1: Equality of welfare. Philosophy & Public Affairs, 185-246 (1981)

Dworkin, R.: What is equality? Part 2: Equality of resources. Philosophy & Public Affairs, 283-345 (1981)

Guyon, I., & Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182 (2003)

Ferreira, F.H.G. and Gignoux J.: The measurement of inequality of opportunity: Theory and an application to Latin America. Review of Income and Wealth, 622-657 (2011)

Ferreira, F.H.G. and Peragine, V.: Equality of Opportunity: Theory and Evidence (2015)

Fleurbaey, M.: Equal opportunity or equal social outcome. Economics and Philosophy, 11, 25-56 (1995)

Fleurbaey, M.: Fairness, responsibility, and welfare, Oxford University (2008)

Fleurbaey, M. and Schokkaert E.: Unfair inequalities in health and health care. Journal of Health Economics 28, 73-90 (2009)

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

Foguel, M. N., & Veloso, F. A.: Inequality of opportunity in daycare and preschool services in Brazil. The Journal of Economic Inequality, 12(2), 191-220 (2014)

Gaer, D.: Equality of opportunity and investment in human capital (Doctoral dissertation, Leuven: Katholieke Universiteit Leuven) (1993)

Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182 (2003)

Herrnstein, R. (1971). IQ. Atlantic Monthly, September, 43-64.

Jensen, A.: How much can we boost IQ and scholastic achievement?. Harvard educational review, 39(1), 1-123 (1969)

Kohavi, R., & John, G. H.: Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324 (1997)

Kornrich, S., & Furstenberg, F.: Investing in children: Changes in parental spending on children, 1972–2007. Demography, 50(1), 1-23 (2013)

Lefranc, A., Pistolesi, N., & Trannoy, A.: Inequality of opportunities vs. inequality of outcomes: Are western societies all alike?. Review Income and Wealth 54(4), 513-546 (2008)

Lefranc, A., Pistolesi, N., & Trannoy, A.: Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. Journal of Public Economics, 93(11), 1189-1207 (2009)

Mare, R. D.: Change and stability in educational stratification. American sociological review, 72- 87 (1981)

Marrero, G. A., & Rodríguez, J. G.: Inequality of opportunity and growth. Journal of Development Economics, 104, 107-122 (2013)

Nozick, R.: Anarchy, State and Utopia. Basic Books (1977)

Peragine, V.: Opportunity egalitarianism and income inequality. Mathematical Social Sciences, 44, 45-60 (2002)

Pistolesi, N.: Inequality of opportunity in the land of opportunities, 1968-2001. Journal of Economic Inequality 7, 411-433 (2009)

Ramos, X., & Van de Gaer, D.: Empirical approaches to inequality of opportunity: Principles, measures, and evidence (2012)

Rawls, J.: Justice as fairness. The philosophical review, 67(2), 164-194 (1958)

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

Rawls, J.: Some reasons for the maximin criterion. The American Economic Review, 64(2), 141-146 (1974)

Roemer, J. E.: A pragmatic theory of responsibility for the egalitarian planner. Philosophy & Public Affairs, 146-166 (1993)

Roemer, J. E.: Theories of distributive justice. Harvard University Press (1998)

Roemer, J. E.: Equal opportunity and intergenerational mobility: going beyond intergenerational income transition matrices (pp. 48-57). Cambridge University Press, Cambridge, England (2004)

Roemer, J. E.: On several approaches to equality of opportunity. Economics and Philosophy, 28(02), 165-200 (2012)

Roemer, J. E., & Trannoy, A.: Equality of opportunity (2013)

Rosa-Dias P. (2009). Inequality of Opportunity in Health: evidence from the UK cohort Study. Health Economics 18, 1057-1074 (2009)

Sen, A. K.: Equality of What? In S. McMurrin (ed.) Tanner Lectures on Human Values, Cambridge: Cambridge University Press (1980)

Zimmerman, D. J.: Peer effects in academic outcomes: Evidence from a natural experiment. Review of Economics and statistics, 85(1), 9-23 (2003)

Appendix

1.A.1: Additional Tables

Table 1.A.1.1: Probability (Average) of High School Graduation (Across Race, Sex, and Social Groups)

	White Male				White Female				Black Male				Black Female			
Birth Year	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA
1920-1929	0.34	0.60	0.91	0.96	0.50	0.69	0.91	0.96	0.12	0.35	0.54	0.69	0.18	0.36	0.66	0.85
1930-1939	0.41	0.66	0.87	0.97	0.56	0.74	0.90	0.98	0.22	0.42	0.66	0.84	0.33	0.49	0.65	0.83
1940-1949	0.52	0.73	0.91	0.97	0.64	0.80	0.93	0.98	0.47	0.65	0.80	0.91	0.50	0.65	0.79	0.88
1950-1959	0.67	0.81	0.91	0.98	0.82	0.90	0.95	0.99	0.70	0.80	0.88	0.93	0.70	0.79	0.87	0.93
1960-1969	0.70	0.77	0.90	0.95	0.85	0.90	0.96	0.98	0.74	0.80	0.90	0.93	0.74	0.81	0.90	0.94
1970-1979	0.80	0.87	0.93	0.98	0.86	0.91	0.96	0.99	0.80	0.86	0.89	0.94	0.88	0.90	0.93	0.96
1980-1988	0.81	0.89	0.95	0.98	0.85	0.93	0.98	0.99	0.79	0.87	0.87	0.92	0.87	0.93	0.95	0.95

Table 1.A.1.2: Probability (Average) of Attending Some College (Across Race, Sex, and Social Groups)

	White Male				White Female				Black Male				Black Female			
Birth Year	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA
1920-1929	0.30	0.37	0.52	0.71	0.28	0.24	0.35	0.56	0.23	0.35	0.52	0.66	0.15	0.20	0.36	0.51
1930-1939	0.34	0.44	0.56	0.72	0.26	0.32	0.41	0.58	0.16	0.26	0.34	0.50	0.18	0.27	0.34	0.42
1940-1949	0.37	0.47	0.60	0.84	0.27	0.36	0.50	0.73	0.34	0.37	0.47	0.65	0.32	0.35	0.43	0.53
1950-1959	0.42	0.46	0.54	0.76	0.43	0.49	0.57	0.74	0.37	0.38	0.45	0.57	0.43	0.45	0.54	0.68
1960-1969	0.38	0.40	0.48	0.68	0.44	0.49	0.57	0.76	0.31	0.31	0.37	0.49	0.45	0.47	0.54	0.68
1970-1979	0.45	0.56	0.60	0.80	0.56	0.67	0.71	0.87	0.44	0.45	0.46	0.60	0.62	0.66	0.66	0.78
1980-1986	0.65	0.59	0.64	0.82	0.69	0.66	0.70	0.87	0.53	0.46	0.45	0.63	0.68	0.63	0.66	0.77

Table 1.A.1.3: Probability (Average) of College Graduation (Across Race, Sex, and Social Groups)

	White Male				White Female				Black Male				Black Female			
Birth Year	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA
1920-1929	0.48	0.50	0.55	0.68	0.53	0.48	0.53	0.65	0.59	0.41	0.43	0.36	0.40	0.43	0.51	0.58
1930-1939	0.43	0.50	0.59	0.66	0.46	0.44	0.52	0.65	0.20	0.26	0.30	0.20	0.16	0.18	0.22	0.49
1940-1949	0.44	0.51	0.58	0.71	0.46	0.47	0.55	0.67	0.32	0.31	0.39	0.50	0.29	0.30	0.38	0.49
1950-1959	0.45	0.47	0.53	0.67	0.46	0.45	0.51	0.64	0.27	0.27	0.31	0.43	0.29	0.28	0.34	0.47
1960-1969	0.45	0.47	0.50	0.66	0.46	0.44	0.49	0.62	0.27	0.29	0.32	0.35	0.25	0.25	0.30	0.38
1970-1979	0.48	0.50	0.54	0.74	0.47	0.49	0.52	0.69	0.37	0.36	0.37	0.44	0.36	0.35	0.37	0.44
1980-1984	0.48	0.50	0.54	0.75	0.53	0.46	0.53	0.69	0.35	0.33	0.36	0.52	0.36	0.31	0.33	0.47

Table 1.A.1.4: Probability (Average) of Starting Post College Education (Across Race, Sex, and Social Groups)

	White Male				White Female				Black Male				Black Female			
Birth Year	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA	LA	RLA	RMA	MA
1920-1929	0.43	0.42	0.44	0.46	0.42	0.42	0.45	0.45	0.38	0.27	0.35	-	0.26	0.27	0.20	0.31
1930-1939	0.50	0.51	0.51	0.53	0.50	0.51	0.52	0.53	0.36	0.33	0.40	0.41	0.38	0.33	0.34	0.43
1940-1949	0.47	0.47	0.48	0.50	0.47	0.47	0.48	0.50	0.35	0.34	0.35	0.38	0.36	0.33	0.35	0.41
1950-1959	0.41	0.42	0.42	0.44	0.42	0.41	0.42	0.44	0.37	0.37	0.39	0.43	0.36	0.35	0.35	0.42
1960-1969	0.39	0.39	0.39	0.41	0.39	0.38	0.39	0.41	0.41	0.43	0.44	0.50	0.38	0.39	0.39	0.43
1970-1979	0.39	0.39	0.39	0.41	0.39	0.38	0.39	0.41	0.42	0.48	0.48	0.48	0.42	0.45	0.46	0.47
1980-1984	0.39	0.38	0.39	0.41	0.38	0.38	0.39	0.40	-	-	-	-	-	-	-	0.45

1.A.2: Additional Graphs

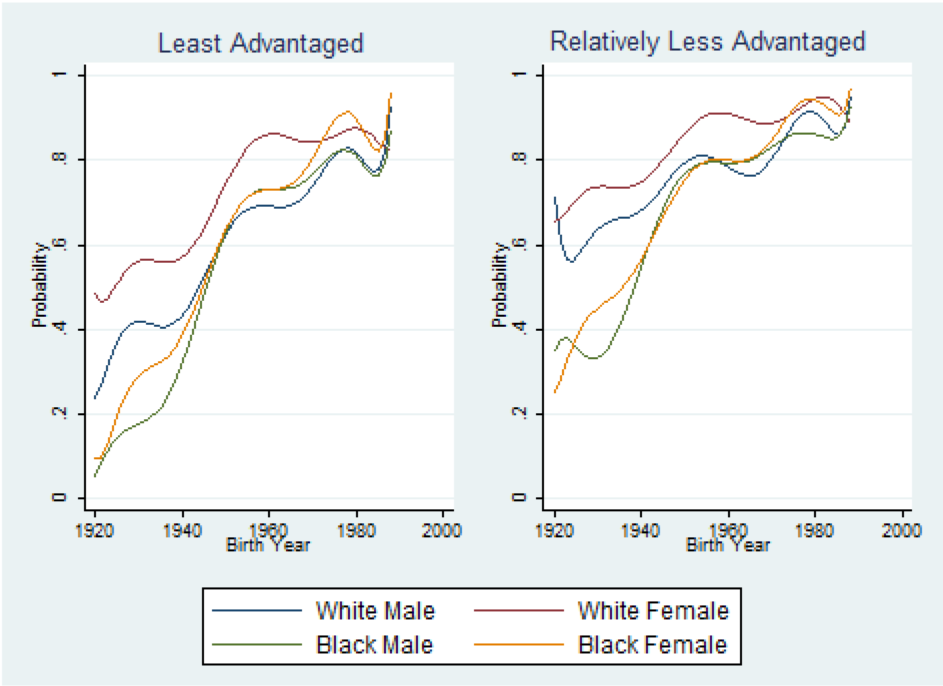

Figure 1.A.2.1a: Probability of High School Graduation across Social Groups (Less Advantaged)

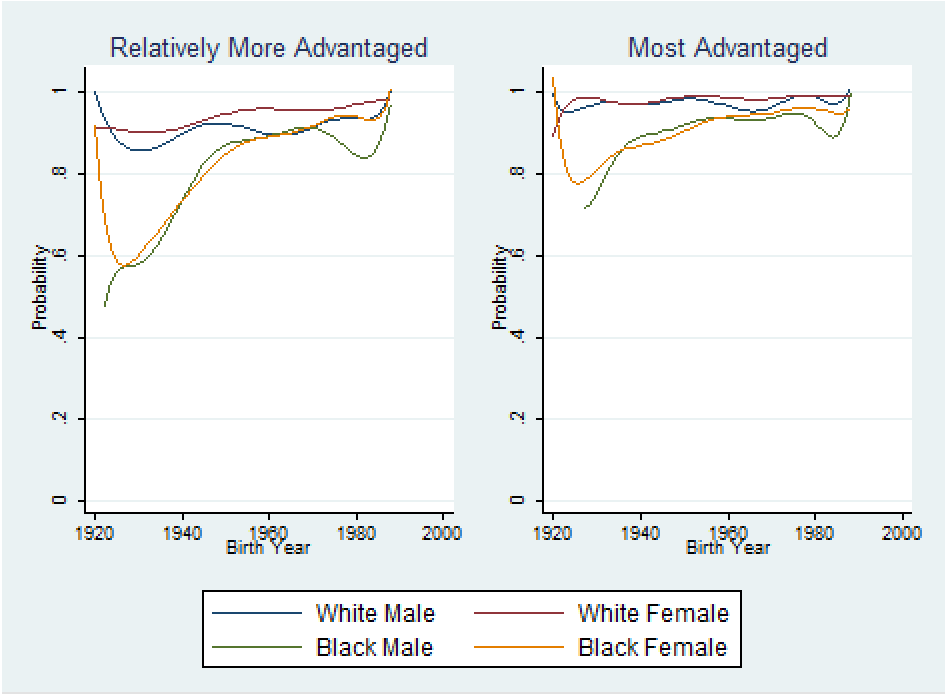

Figure 1.A.2.1b: Probability of High School Graduation across Social Groups (More Advantaged)

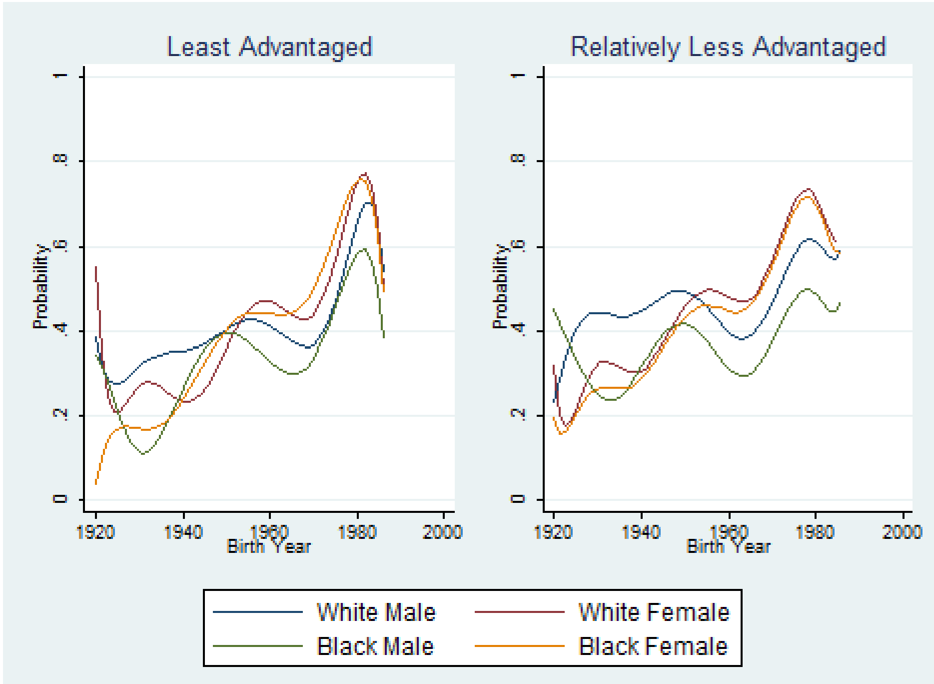

Figure 1.A.2.2a: Probability of Some College Education across Social Groups (Less Advantaged)

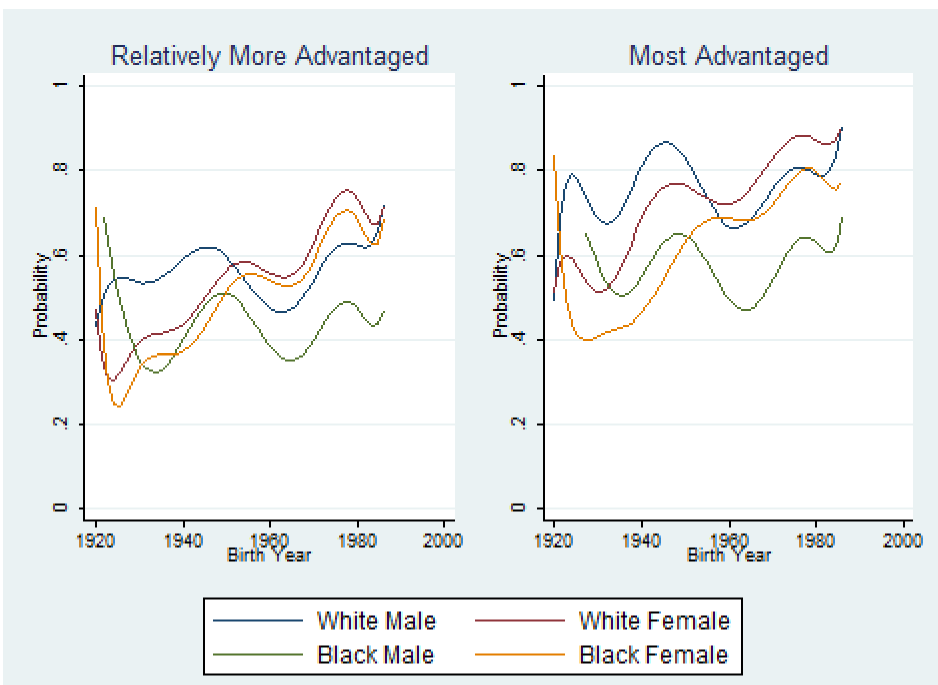

Figure 1.A.2.2b: Probability of Some College Education across Social Groups (More Advantaged)

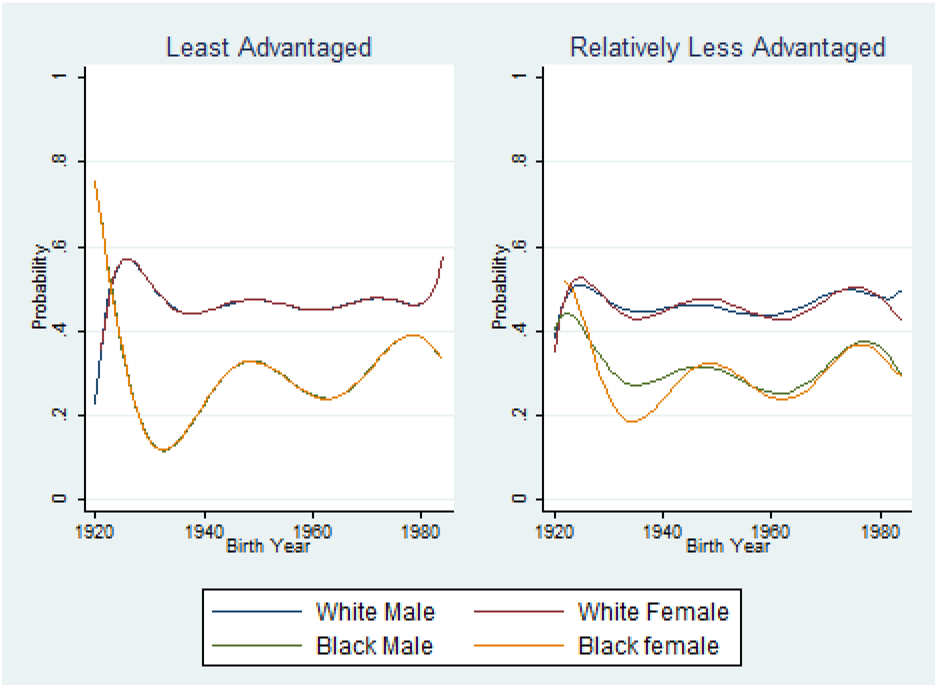

Figure 1.A.2.3a: Probability of College Graduation across Social Groups (Less Advantaged)

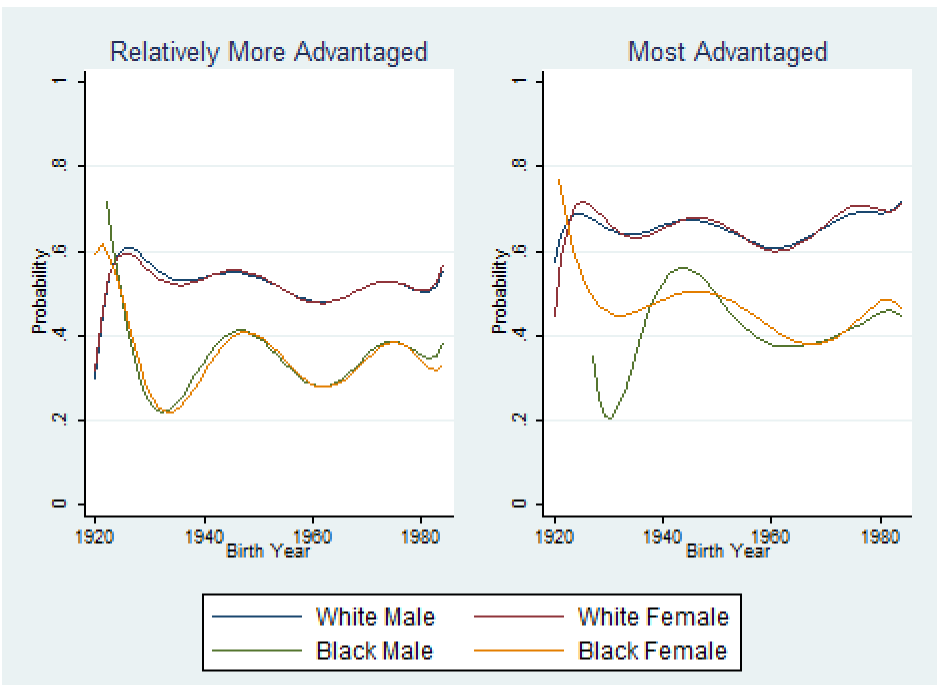

Figure 1.A.2.3b: Probability of College Graduation across Social Groups (More Advantaged)

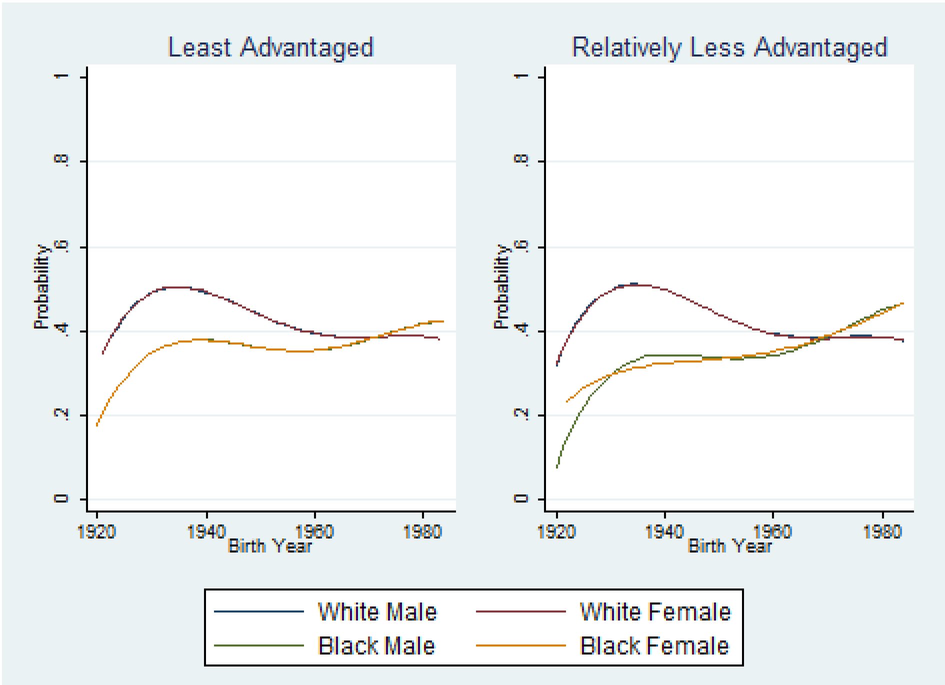

Figure 1.A.2.4a: Probability of Starting Post College across Social Groups (Less Advantaged)

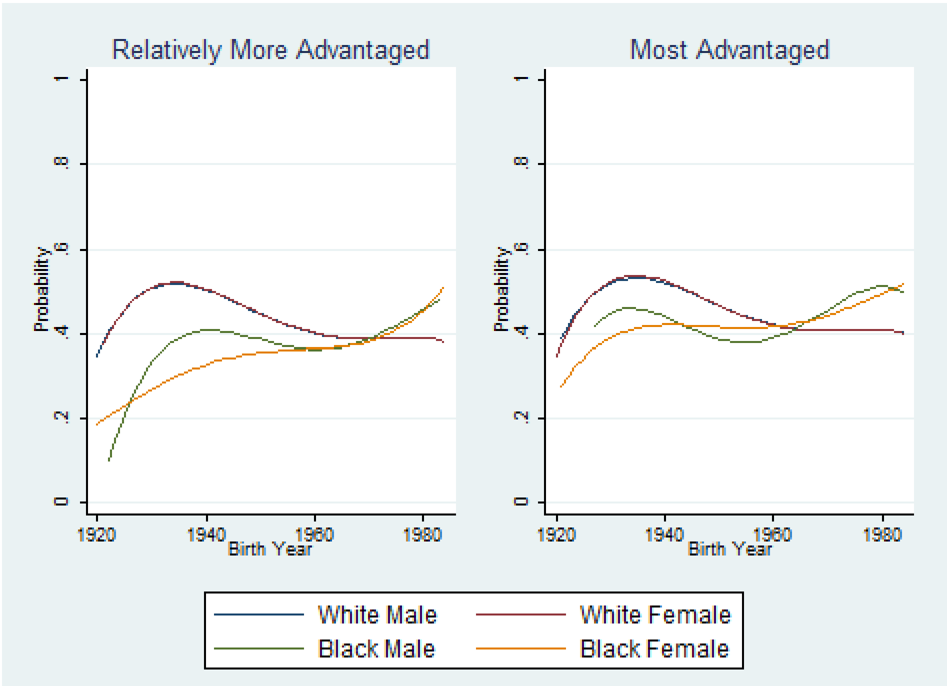

Figure 1.A.2.4b: Probability of Starting Post College across Social Groups (More Advantaged)

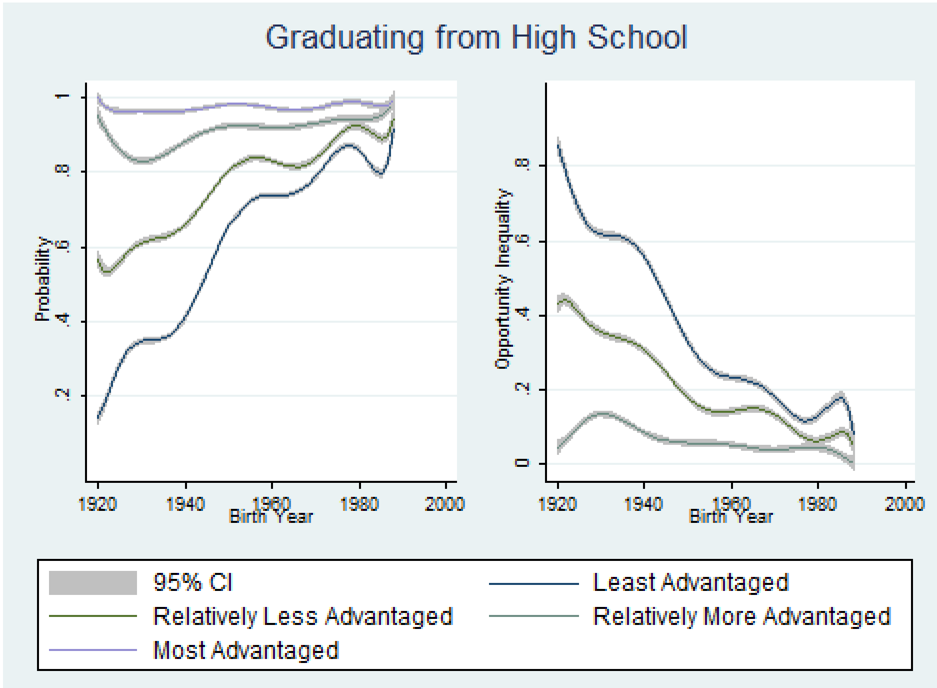

Figure 1.A.2.5a: Probability/Opportunity Inequality of Graduating from High School

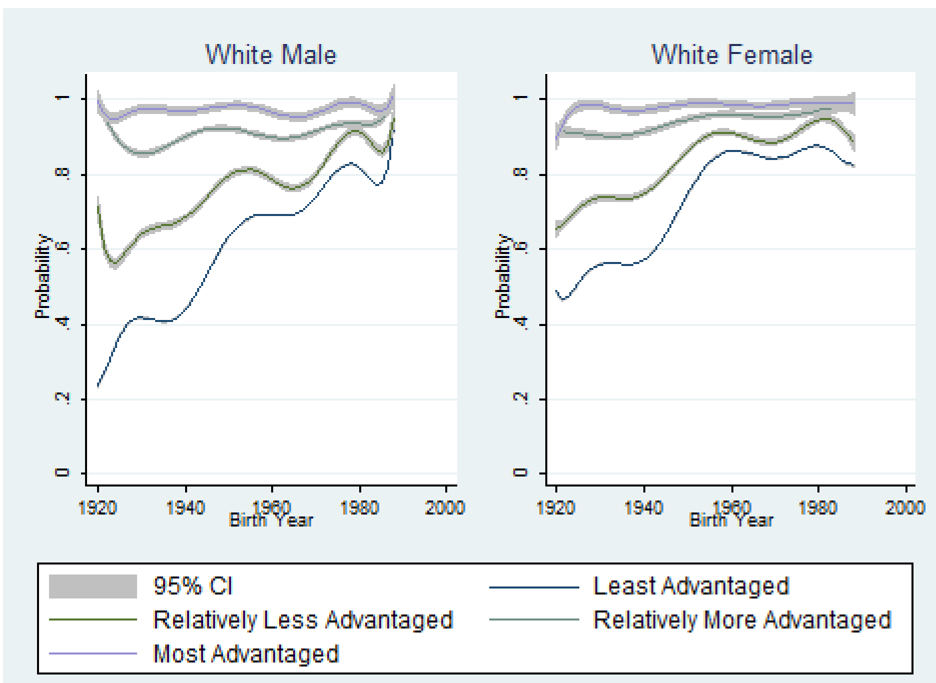

Figure 1.A.2.5b: Probability of Graduating from High School across Race and Gender (White)

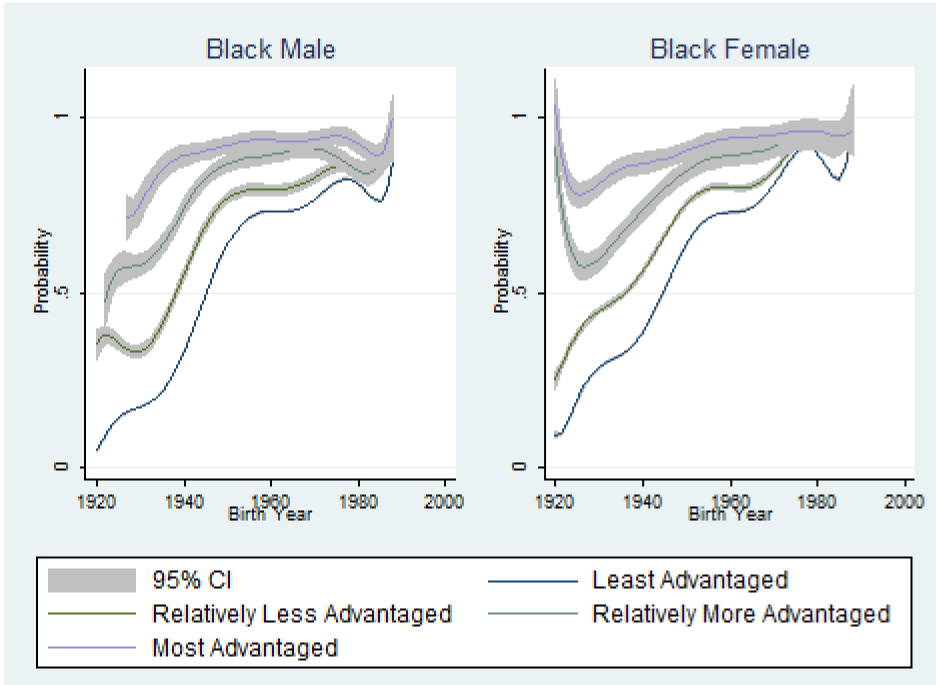

Figure 1.A.2.5c: Probability of Graduating from High School across Race and Gender (Black)

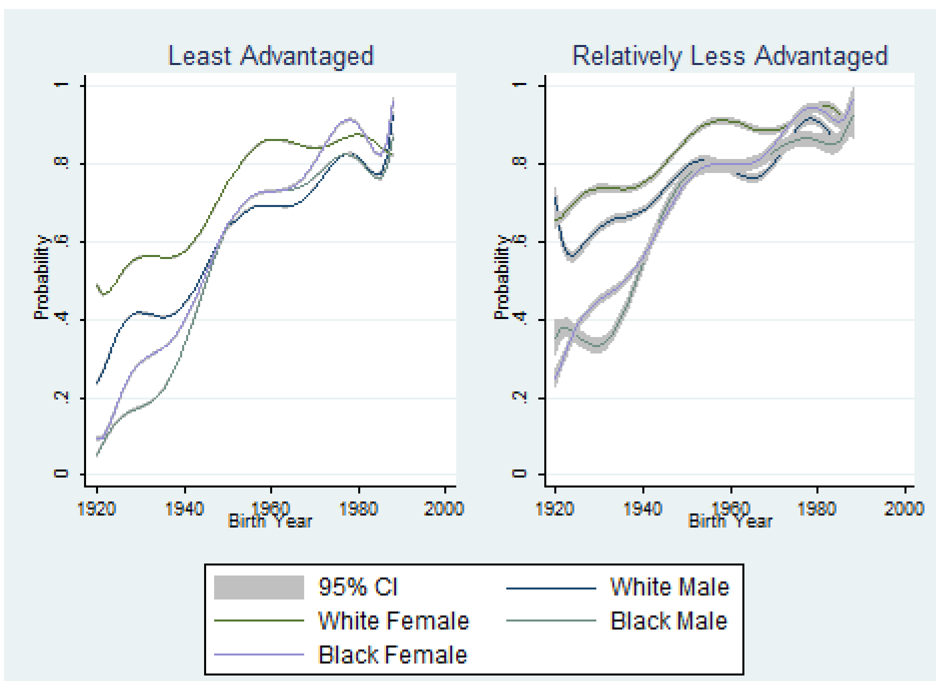

Figure 1.A.2.5d: Probability of Graduating from High School across Social Groups (Less Advantaged)

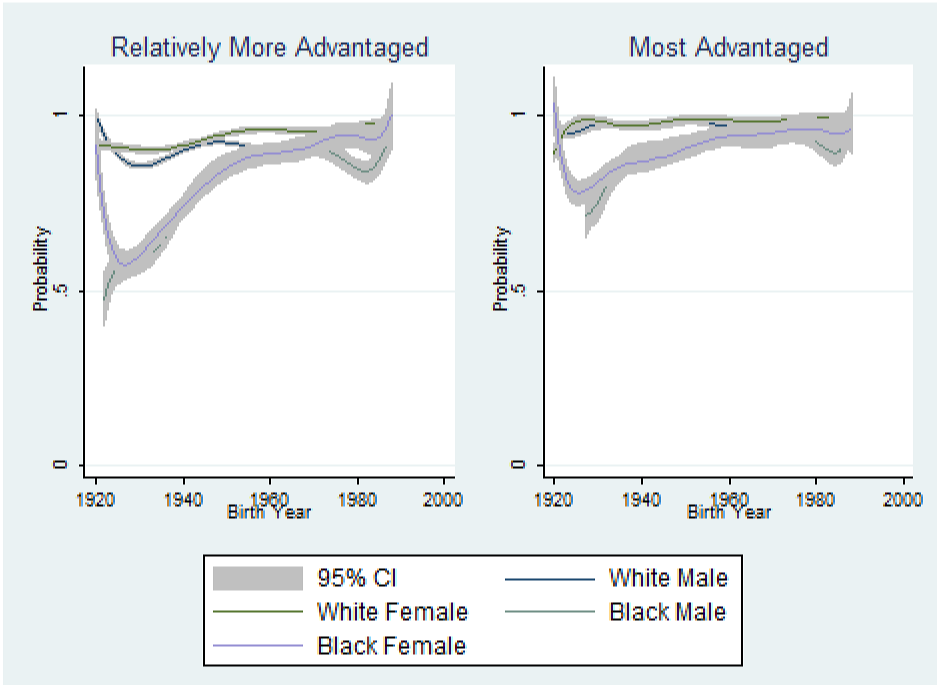

Figure 1.A.2.5e: Probability of Graduating from High School across Social Groups (More Advantaged)

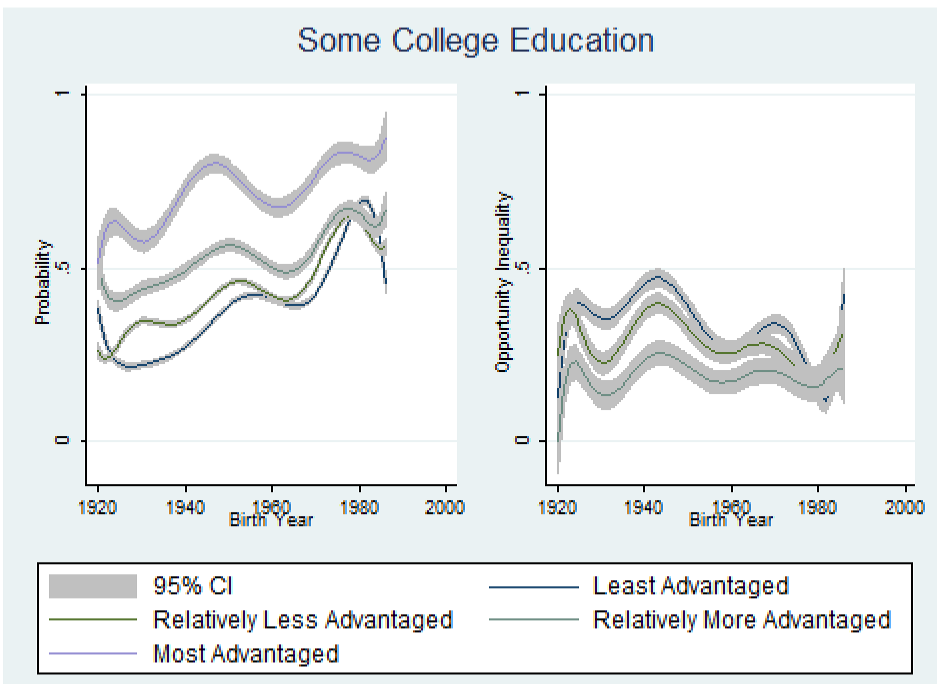

Figure 1.A.2.6a: Probability/Opportunity Inequality of Attending Some College

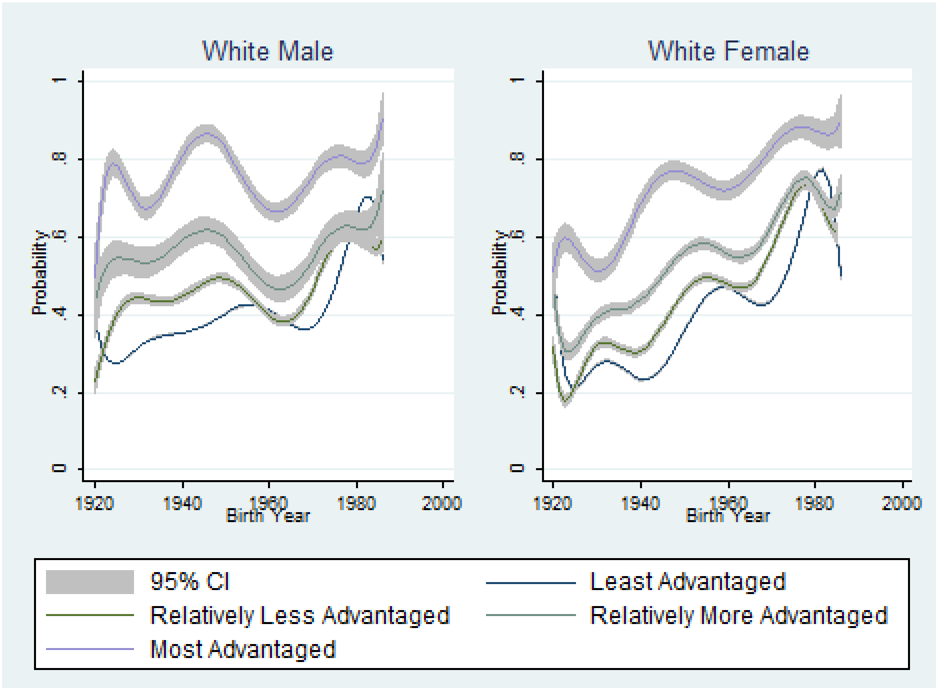

Figure 1.A.2.6b: Probability of Some College Education across Race and Gender (White)

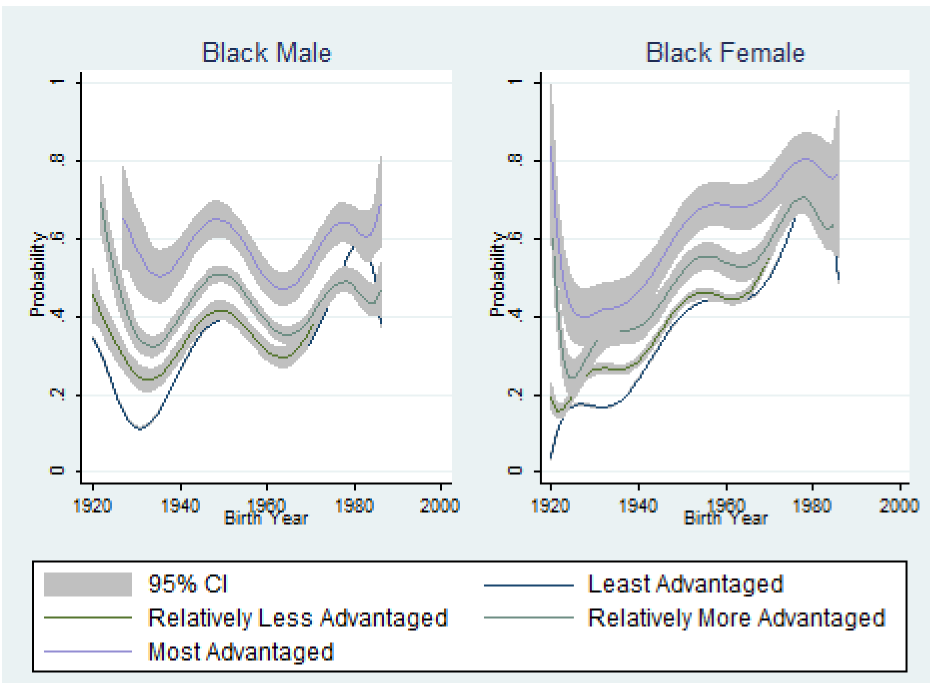

Figure 1.A.2.6c: Probability of Some College Education across Race and Gender (Black)

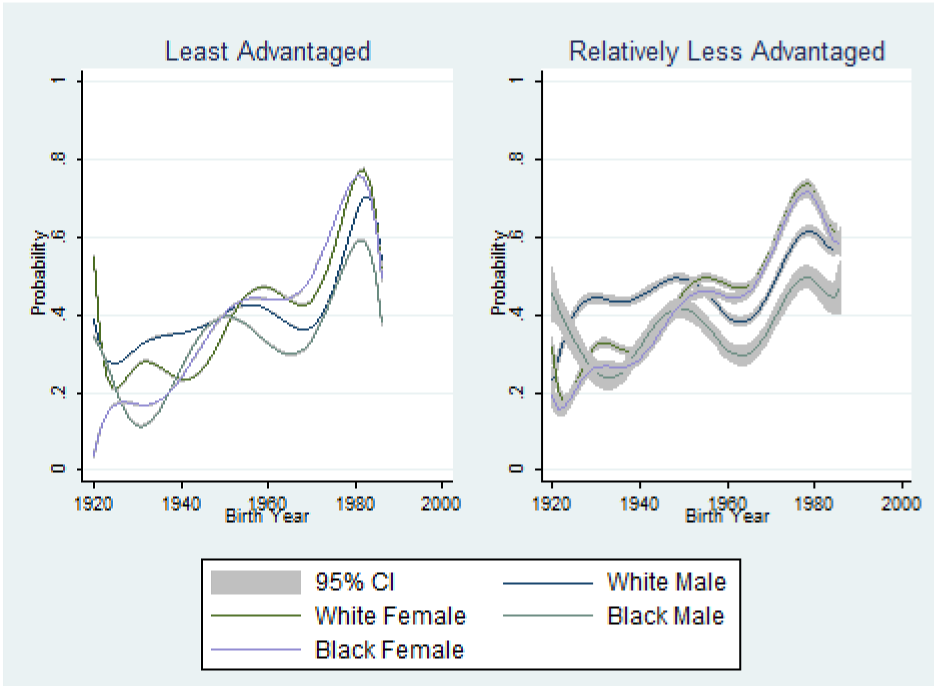

Figure 1.A.2.6d: Probability of Some College Education across Social Groups (Less Advantaged)

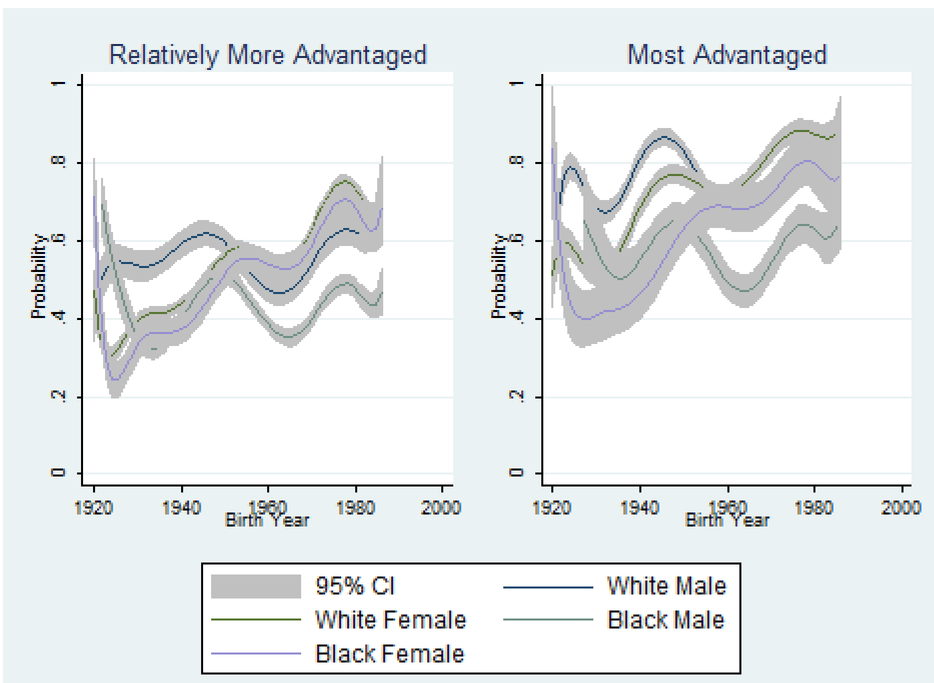

Figure 1.A.2.6e: Probability of Some College Education across Social Groups (More Advantaged)

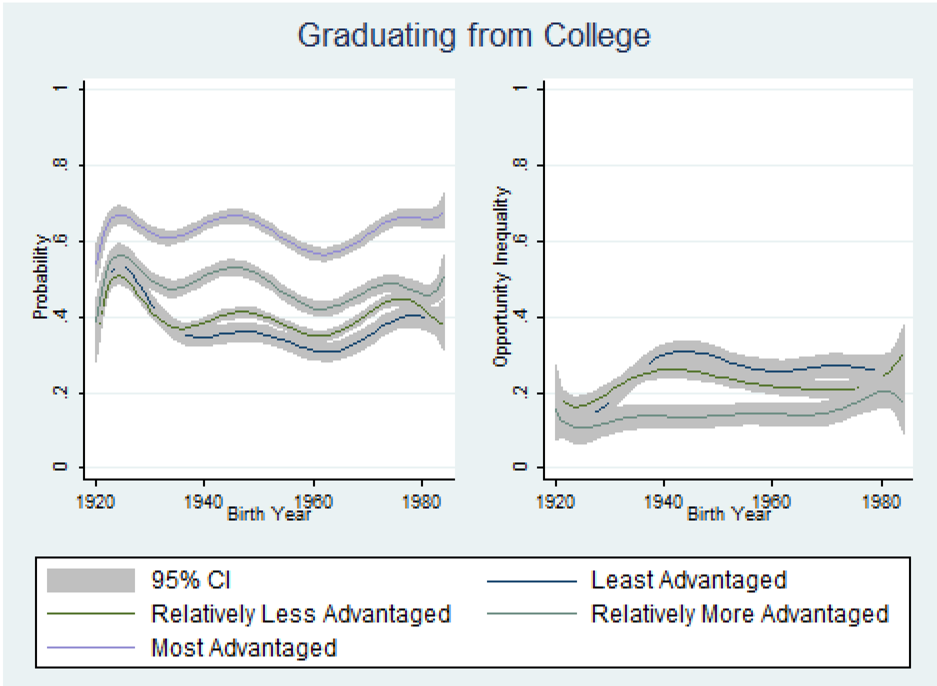

Figure 1.A.2.7a: Probability/Opportunity Inequality of Graduating from College

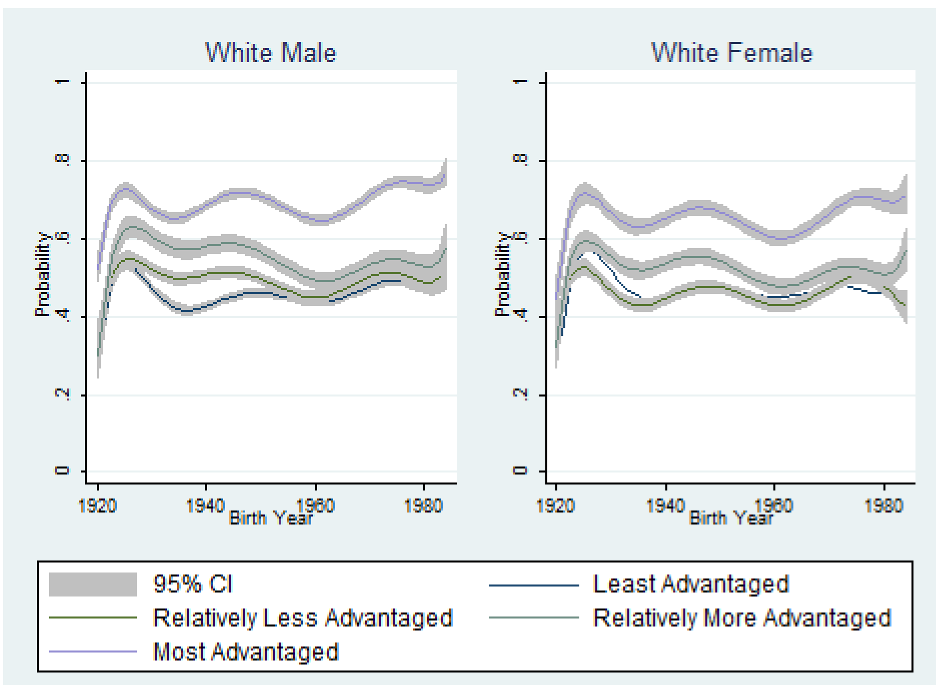

Figure 1.A.2.7b: Probability of Graduating from College across Race and Gender (White)

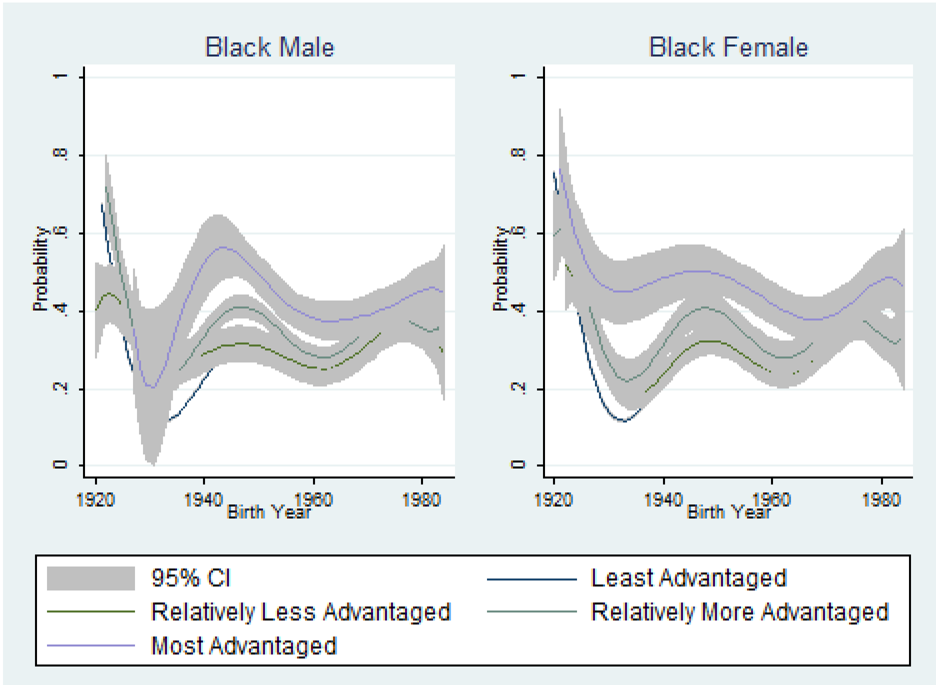

Figure 1.A.2.7c: Probability of Graduating from College across Race and Gender (Black)

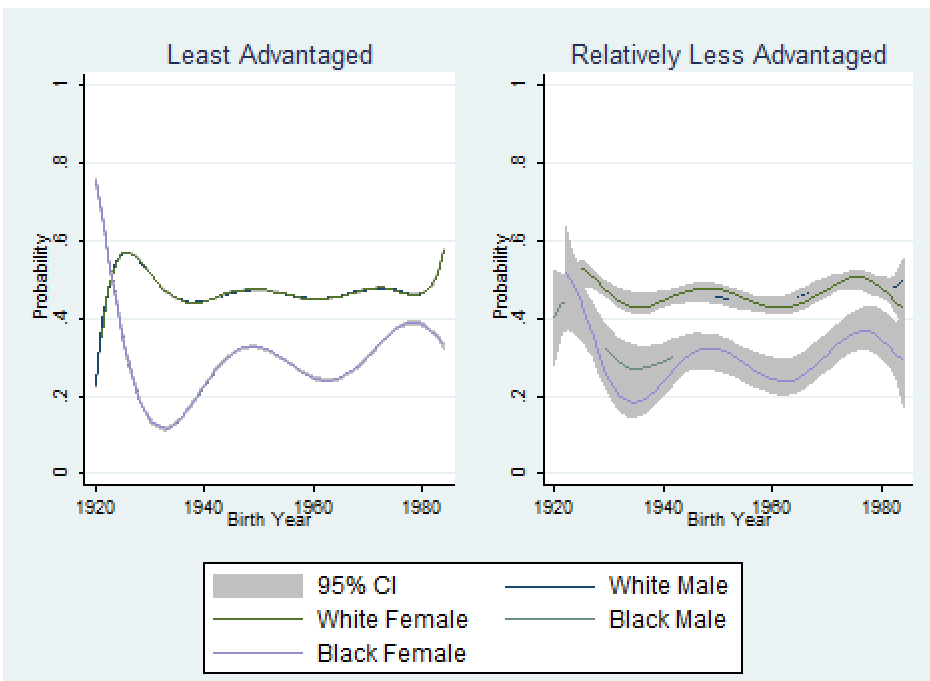

Figure 1.A.2.7d: Probability of Graduating from College across Social Groups (Less Advantaged)

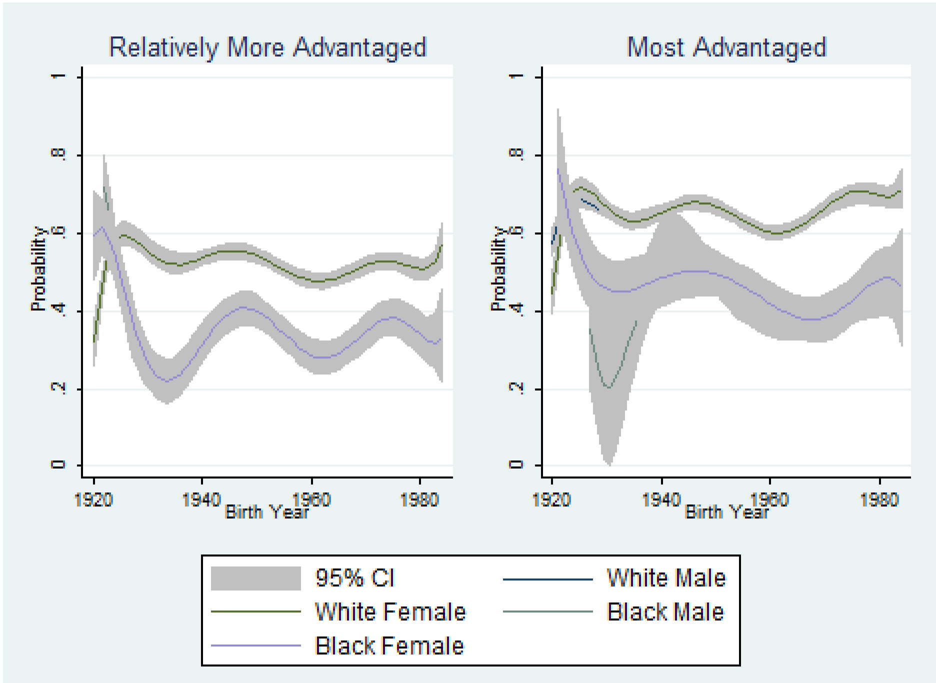

Figure 1.A.2.7e: Probability of Graduating from College across Social Groups (More Advantaged)

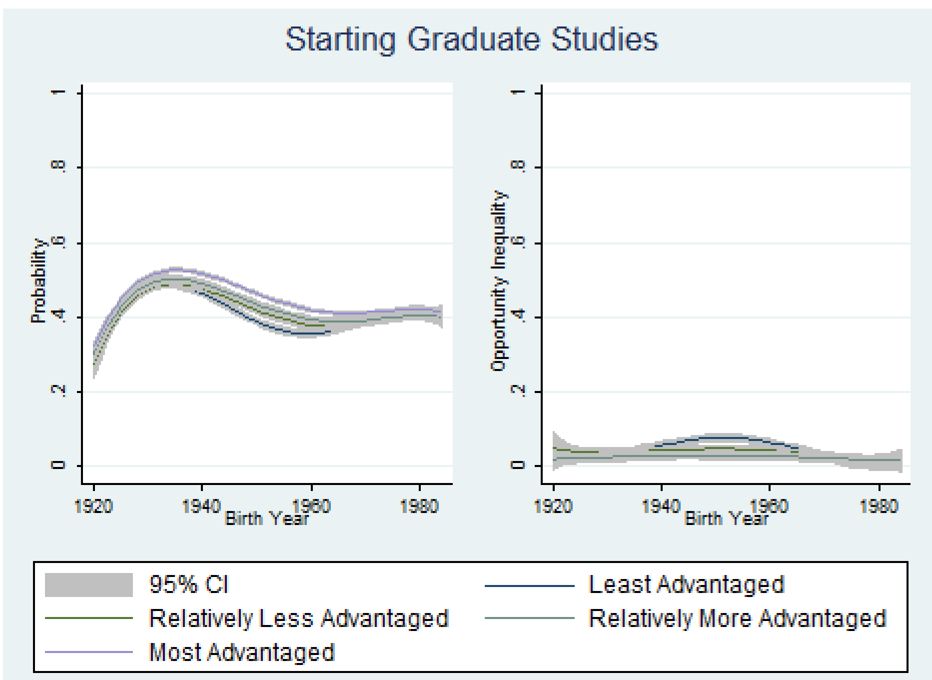

Figure 1.A.2.8a: Probability/Opportunity Inequality of Starting Post College Education

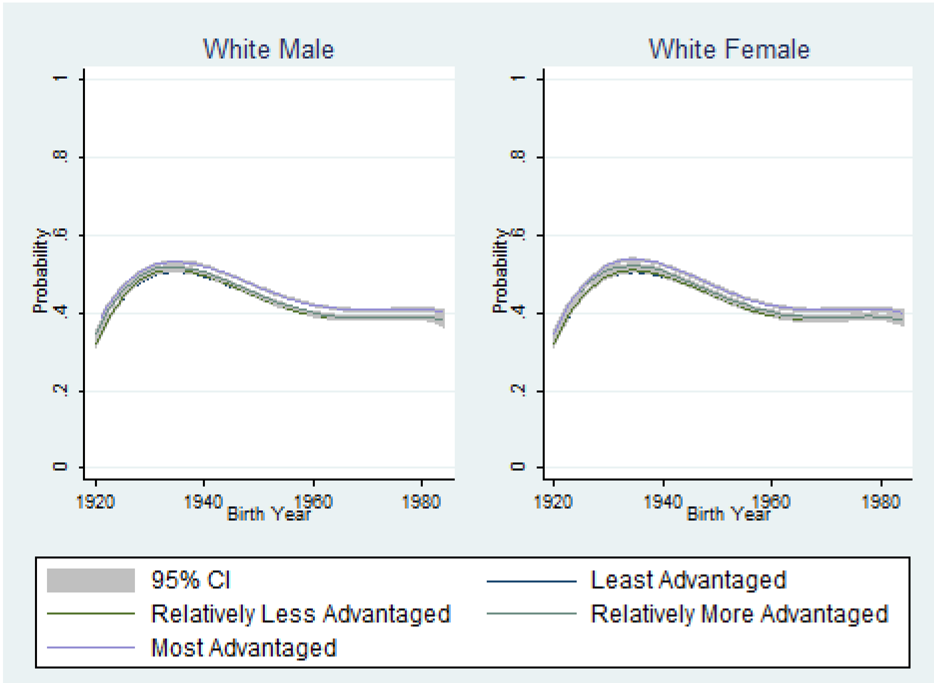

Figure 1.A.2.8b: Probability of Starting Post College Education across Race and Gender (White)

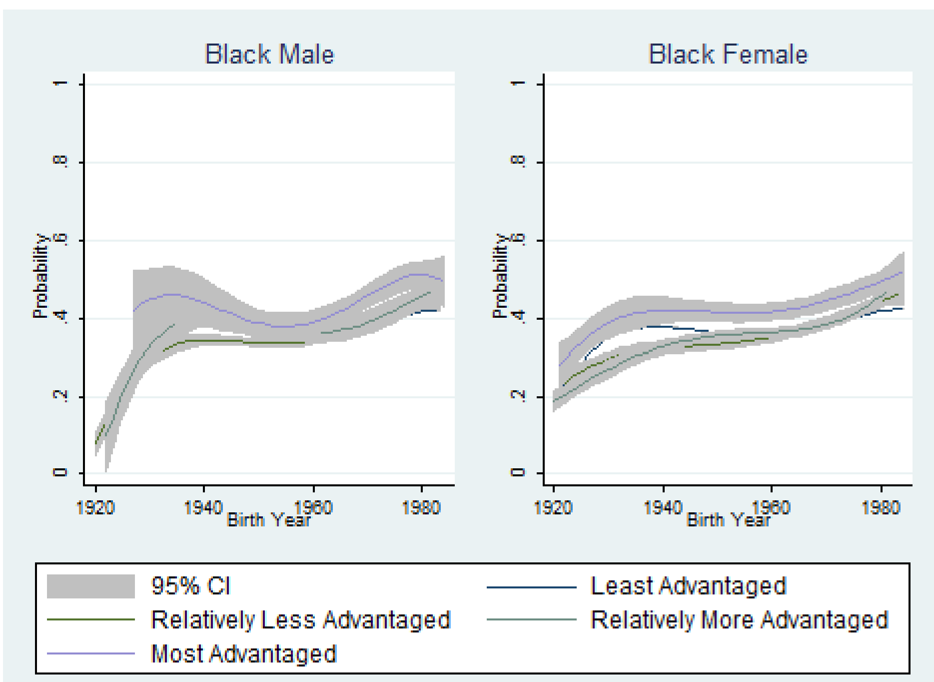

Figure 1.A.2.8c: Probability of Starting Post College Education across Race and Gender (Black)

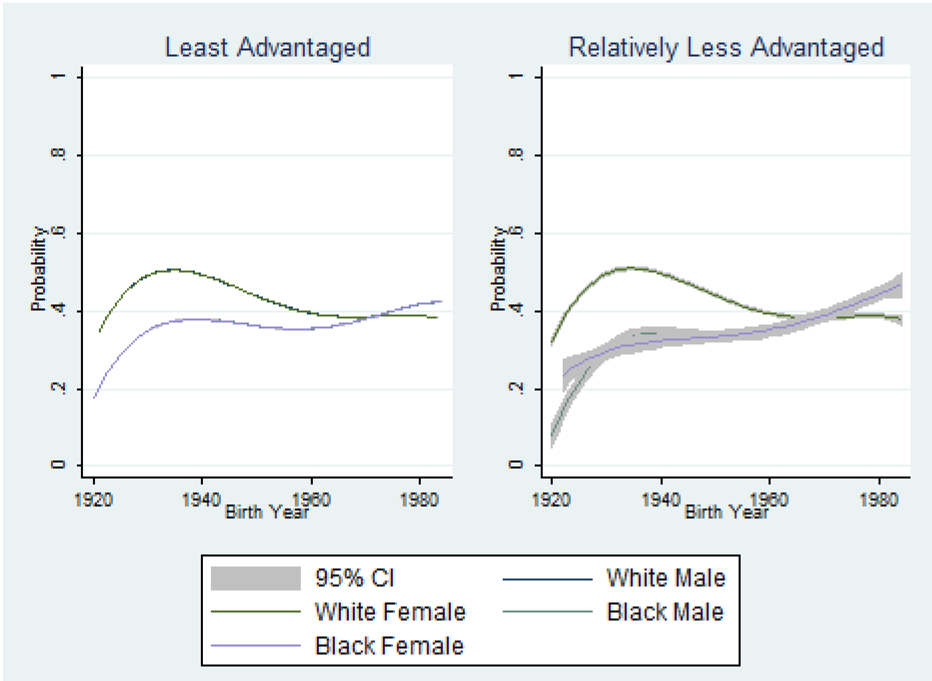

Figure 1.A.2.8d: Probability of Starting Post College Education across Social Groups (Less Advantaged)

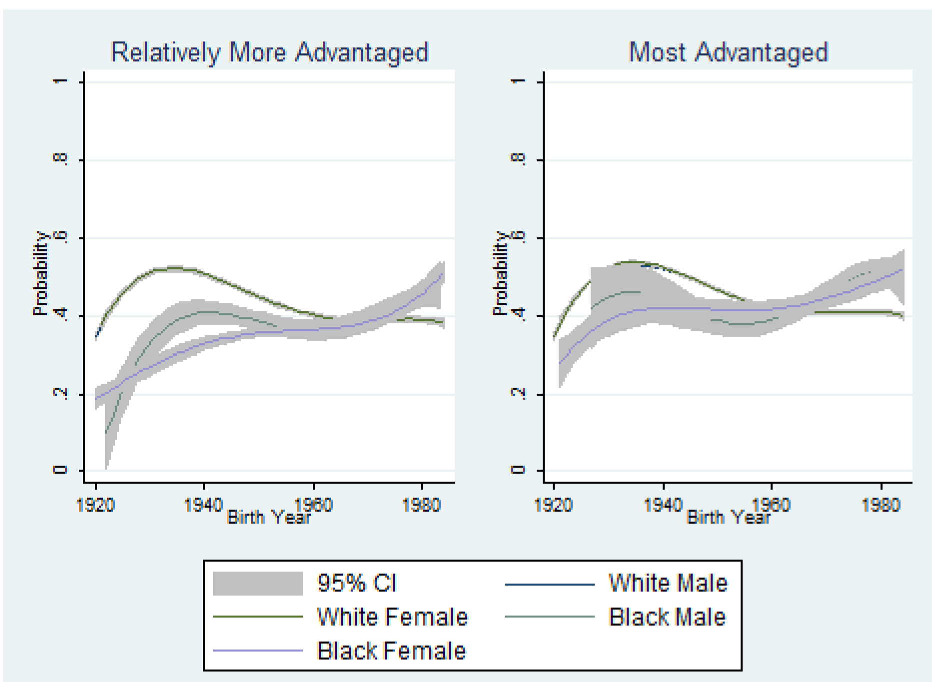

Figure 1.A.2.8e: Probability of Starting Post College Education across Social Groups (More Advantaged)

Chapter 2

First and Second Moments of the Age Distributions of Graduates

Abstract

The paper examines the age and demographic distribution of graduates across cohorts from 1940 until 1990. According to a report by Organisation for Economic Co-operation and Development, there is an increasing trend in the number of non-traditional students in Europe and in the US. Using the Panel Study of Income Dynamics data, the paper explores the first and second moment of the age of graduating from high school and from college across the US. To deal with data deficiencies, a large part of the paper deals with data preparation. The paper provides a unique method of extracting information on the graduating age of the individuals both from high school and from college. The results show a large dispersion across the full sample. The data truncated to a standard length, however, provides a much smaller dispersion and much smaller moments. The paper concludes that as the time passes, people tend to attain education achievements at a younger age.

2.1 Introduction

Much credit goes to Mincer (1962, 1974) and Becker (1962) for formalizing the concept of the returns to education. However, prior to their formalization, the importance of education in having a comfortable life was unequivocal. Belief that educational attainment plays a crucial part in a country's human capital formation is a quintessential component in both micro- and macro-economic thinking. The breadth of the existing literature on returns to education is huge, exploring different aspects of educational returns, starting with modelling the returns to education and moving to the contributions of different determinants of education across demographic groups. While it is important to look into the returns to education across demographic groups – race, gender, socioeconomic condition – for policy prescriptions, the age at which individuals attain educational credentials is also important for understanding the returns to education, because that affects the impact of education on life-long earnings.

According to a report by the Organization for Economic Co-operation and Development (OECD) in 2006, there is an increasing trend in the number of non-traditional students pursuing college education, increasing the diversity of entering college students in the US. A traditional student is an individual entering college right after graduating from high-school and is 19 years old (Gilardi & Guglielmetti, 2011). The definition of a non-traditional student is not that straightforward; there are at least three definitions of non-traditional students (Kim, 2002). The first definition classifies students who are older than 23 or 25 when starting college as non-traditional (Metzner & Bean, 1987). Rendon, Jalamo, and Nora (2000) put forward another definition, under which students are classified as non-traditional based on their race, ethnicity, and socioeconomic background. The third definition focuses on risk factors for dropping out, following the US Department of Education: delayed enrollment, part-time enrollment, full-time employment, financial independence, dependents etc. (Gilardi & Guglielmetti, 2011). Although these definitions were designed for college students, high school education can also have non-traditional students. There are individuals finishing high school early and starting college at a much younger age. Also, individuals can have enrollment breaks while in high school or can finish high school later. There can also be impacts of policies such as G.I. Bill, which provided an opportunity for veterans returning from war to have further education.

An interesting issue raised by the increasing proportion of non-traditional students is the consequent change in the distribution of the age at which students graduate from high school and college. This paper looks into the evolution of these distributions. Analyzing the Panel Study of Income Dynamics (PSID) data, the paper finds a decreasing trend in the first moment of the age distribution. A large part of this trend can be attributed to a more dispersed age range group at the beginning of the observation year. In order to deal with the large dispersion, we further analyze the truncated sample for the both educational attainment. The results show much smaller and much more uniform moments.

The rest of the paper goes as follows. We present the method of data preparation and frequency distribution in section 2.2. Section 2.3 of the paper contains the results and section 2.4 concludes.

2.2 Data and Distributions

2.2.1 Data Preparation

The paper traces changes in the distribution of high school and college graduation age over recent decades, for the US population. The data, collected from PSID, include education level for each individual who was interviewed, each time he or she was interviewed, from 1968 to 2013. Each individual has a unique ID. Since the aim of the paper is to trace changes in the distribution of graduation age, much effort was put into determining the correct education levels of the individuals at each point in time, and therefore when they graduated. We cannot claim, however, to have found the graduation age for every individual with 100% accuracy. The graduation ages reported in this paper are informed estimates for many individuals, as the data do not yield indubitable truth.

The data we have used were sifted in several steps. We started with 75,252 individuals, with 2,859,576 observations in total. For many individuals, the education level was unreported or was a nonsensical value throughout the time span. We excluded those individuals, as they provided no information of interest for this research. For individuals who had unreported education (or if reported, an education value > 18) in few years, the education level was first replaced by its last non-missing value and, if it still remained unreported, then with the next non-missing education level. We got rid of the individuals who did not have any credible reported education level. After these steps, we were left with 52,051 individuals.

In the next step, to ensure a coherent data set, we set some conditions. The conditions are as follows:

$$(C1)\ E_{y_2} - E_{y_1} \leq\ y_2 - y_1$$

$$(C2) E_{y_2} - E_{y_1} \geq 0$$

where,

$$y_1 = interview\ at\ year\ (t-1)$$

$$y_2 = interview\ at\ year\ t$$

$$E_{y_1} = reported\ education\ at\ y_1$$

$$E_{y_2} = reported\ education\ at\ y_2$$

$$t \in [1968{,}2013]$$

C1 tells us whether people attain more than one year of education in a year and *C2* tells us whether the educational attainment decreases, which is not believable, instead of increasing in a year.

These conditions provide hints as to whether any of the individuals/data collectors have misreported the education level of an individual in any of the years. An individual with the normal progression of education would see the above two conditions met. For 3,236 individuals (which we named as *'inconsistent observations'*) at least one of the above two conditions was not met. We have consequently sought to make them consistent.

The inconsistent observations were dealt in three steps. In the first step, we used the *mode* function to determine the number of inconsistencies. For individuals who had one inconsistency over the years, the inconsistent observation was replaced with the modal value of that individual's reported education. For instance, if we have $x > 1$ observations of k and 1 observation of j, where $j, k \in Z^+ \in [1, 17]$, for one individual, we have replaced j with k. For the next round of inconsistency, we have built an inconsistency score index. The index is as follows:

$$Inconsistency\ score\ =\ 0\ if\ education_L \geq\ 0$$

$$Inconsistency\ score\ =\ 0\ if\ education_L \leq year_L + 1$$

$$Inconsistency\ score\ =\ 0.1\ if\ education_L \leq year_L + 2$$

$$Inconsisntecy\ score = 1\ if\ education_L == year_L + 3$$

$$Inconsistency\ score = 2\ if\ education_L > year_L + 3$$

$$Inconsistency\ score = 2\ if\ education_L < 0$$

where,

$$education_L = E_{y_2} - E_{y_1}$$

$$year_L = y_2 - y_1$$

The individuals who had a sum of inconsistency scores of six or above were dropped from the sample. For the individuals with the sum of inconsistency scores greater than or equal to four but less than six, their inconsistencies were resolved manually, based on the most plausible interpretation of the data. For individuals who had only two numbers for the reported values of education, which were incoherent or violated at least one of the two conditions, maximum and minimum functions were employed to sort out the actual educational attainment. If the difference between the *maximum* and *minimum* value was one, the maximum educational attainment was deemed the actual educational attainment. Figure 2.1 and Figure 2.2 shows the distribution of data across the different steps of data preparation.

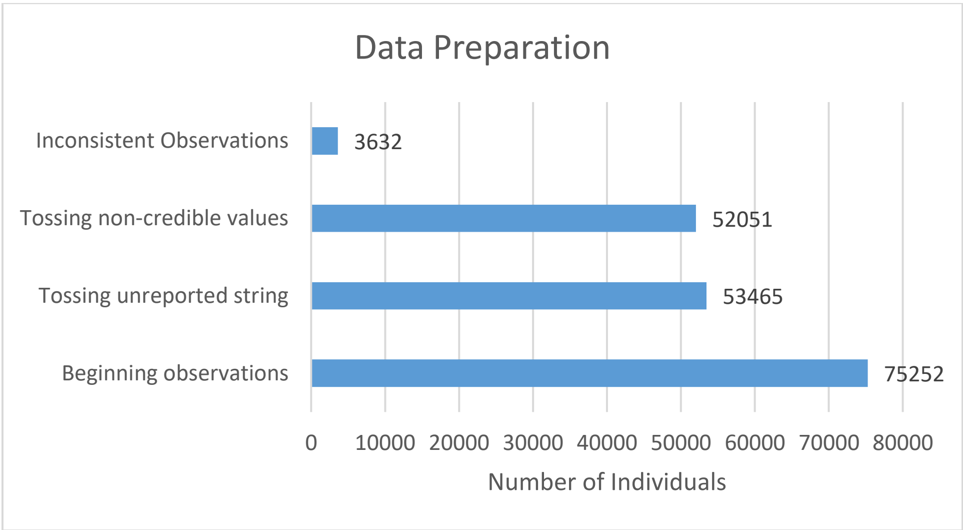

Figure 2.1: Sample Size at the Different Stages of Data Preparation

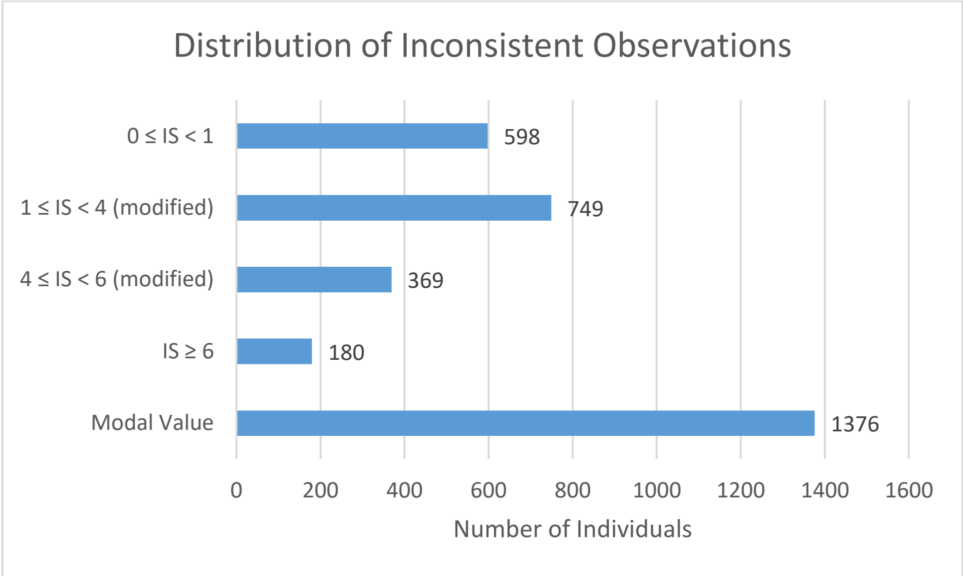

Figure 2.2: Distribution of Inconsistent Observations

At the end, after all the modifications, we were left with 1,875,819 observations of 51,511 individuals for our analysis.

2.2.2 Frequency Distribution

For this analysis, we have two achievements – high school graduation, and college graduation. In this paper, we aim to identify the frequency distribution for each achievement, for individuals born in a particular cohort. The individuals for this analysis were born between 1940 and 2002. The individuals were interviewed in years between 1968 and 2013. The observations were separated into three groups: non-graduates, old graduates, and new graduates. For this analysis, we took into account only those who were reported to be newly graduated in any particular year. We did not account for the old graduates because we do not know the year in which they achieved an educational attainment. Based on this information, we calculate a frequency distribution for each of the six decade cohorts from the 1940s to the 1990s. We collected the mean and standard deviation from the reported samples for each of the cohorts.

2.3 Results

2.3.1 High School Graduation

Table 2.1: Moments of High School Graduation

Decade	Mean	Standard Deviation	Standard Error
1940-1949	34.656	12.221	0.388
1950-1959	28.377	10.424	0.233
1960-1969	22.426	5.572	0.092
1970-1979	20.185	4.055	0.071
1980-1989	19.397	2.538	0.052
1990-1996	18.366	1.119	0.034

Table 2.1 presents the mean and standard deviation of the age distribution for graduating from high school. As reported, the mean is much higher for the earliest (1940-1949) cohort of our analysis and there is a decreasing trend in mean age across the cohorts. As we move to the more

recent cohorts, the average age of graduating from high school declines and averages about 18 for the most recent cohort. The story is more comprehensible if we examine the standard deviation, in column (3) of Table 2.1. It says that we have a more dispersed age range at the earlier cohorts, and dispersion shrinks as we move to the more recent cohorts. It would be plausible to presume that as the time passes, the average age of high school graduation would increase for the recent cohorts.

Table 2.2, 2.3, and 2.4 show the moments of high school graduation truncated at the age of 23, 33, and 43. The tables suggest that the number of non-traditional students shrink as we truncate the age of individuals and the mean age of graduation declines as we move to the more recent decades.

Table 2.2: Moments of High School Graduation Truncated at Age 23

Decade	Mean	Standard Deviation	Standard Error
1940-1949	20.97	1.37	0.09
1950-1959	19.25	1.83	0.06
1960-1969	19.15	1.43	0.03
1970-1979	18.89	1.25	0.02
1980-1989	18.77	1.44	0.03

Table 2.3: Moments of High School Graduation Truncated at Age 33

Decade	Mean	Standard Deviation	Standard Error
1940-1949	24.52	3.88	0.17
1950-1959	22.98	5.01	0.13
1960-1969	20.93	3.98	0.07
1970-1979	19.73	3.01	0.05

Table 2.4: Moments of High School Graduation Truncated at Age 43

Decade	Mean	Standard Deviation	Standard Error
1940-1949	28.87	7.46	0.27
1950-1959	25.93	7.50	0.18
1960-1969	21.83	5.64	0.09

Figure 2.3 through 2.6 shows us the graphical representation of Table 2.1 to 2.4, respectively, at the 5% significance level. Figures 2.7 through 2.12 provide the frequency distribution of each of the cohorts. The density is much higher for all the cohorts around the ages of 18 and 19, except for the 1940s cohort, which experiences a fluctuation in density across ages.

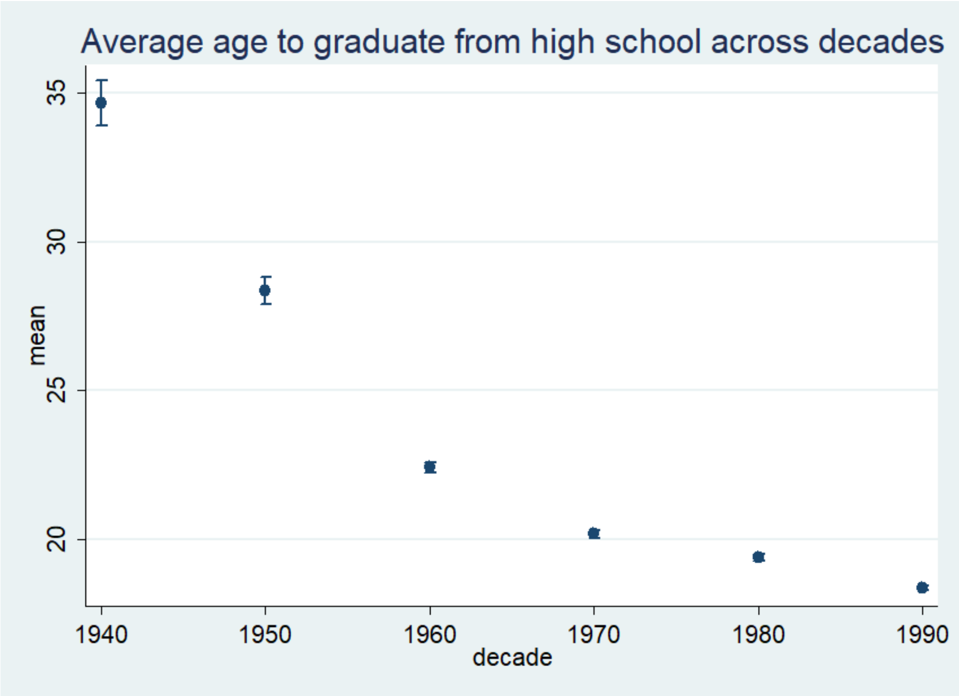

Figure 2.3: Moments of High School Graduation

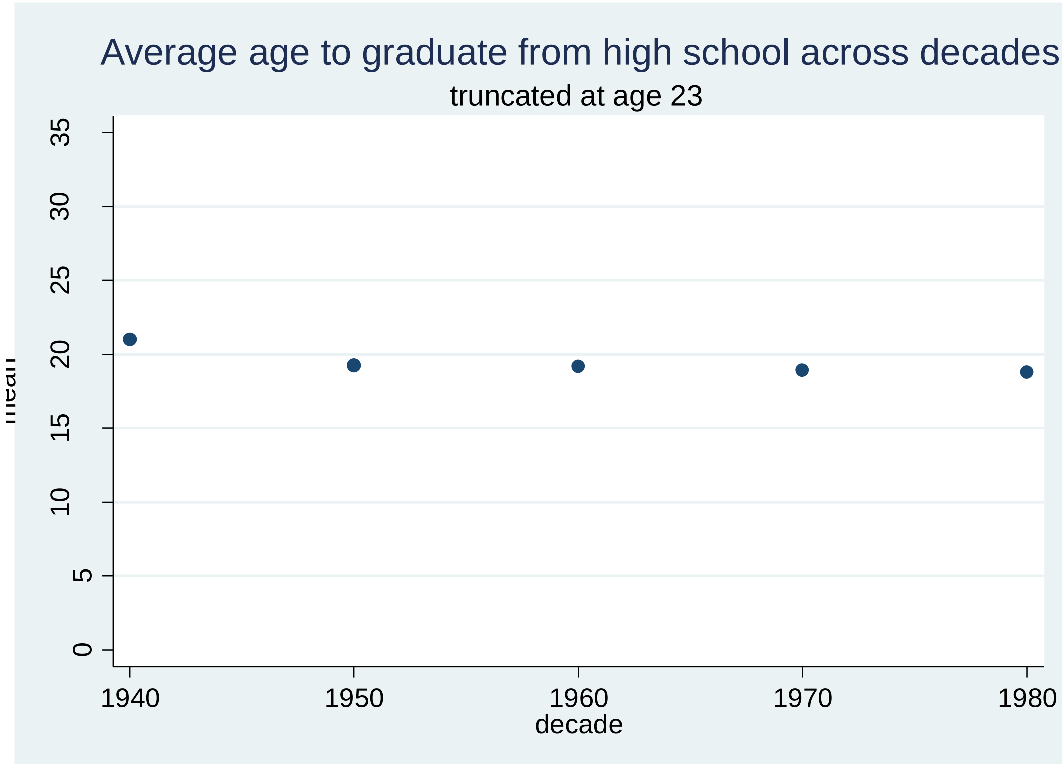

Figure 2.4: Moments of High School Graduation (truncated at age 23)

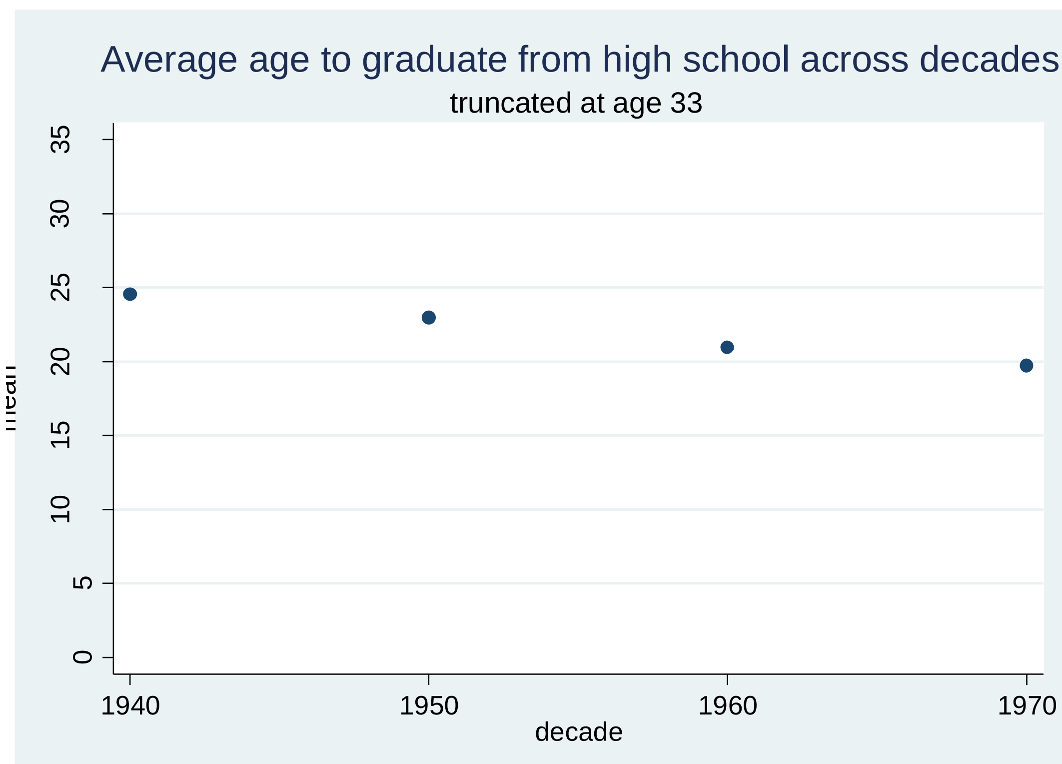

Figure 2.5: Moments of High School Graduation (truncated at age 33)

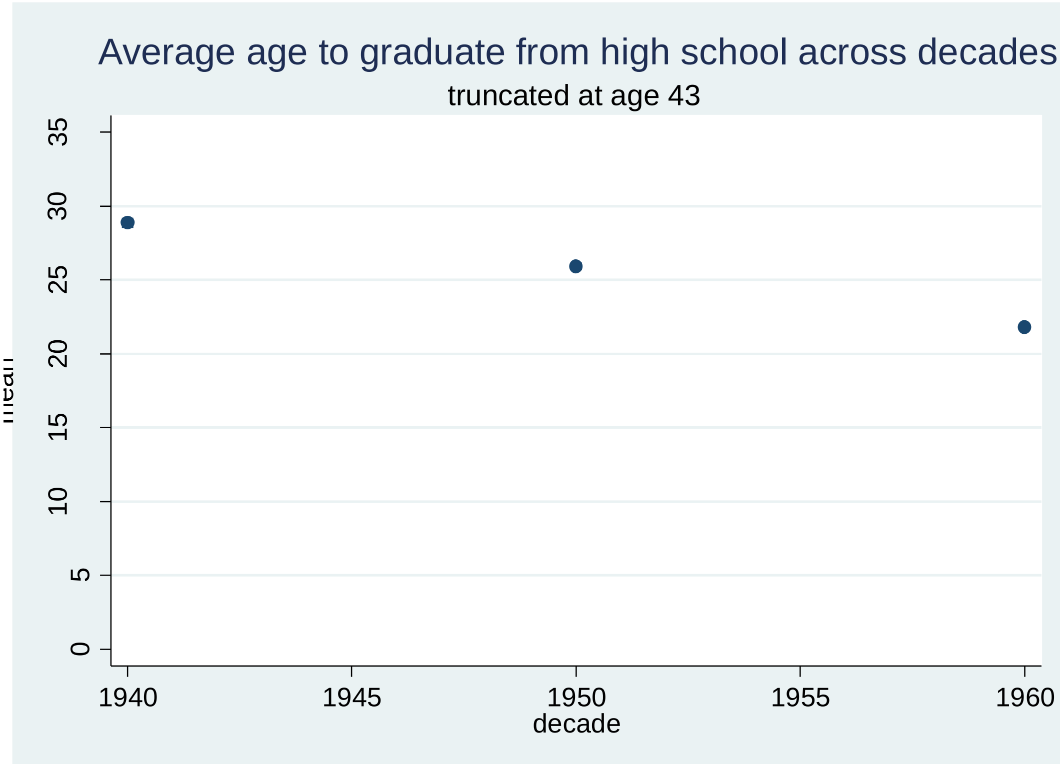

Figure 2.6: Moments of High School Graduation (truncated at age 43)

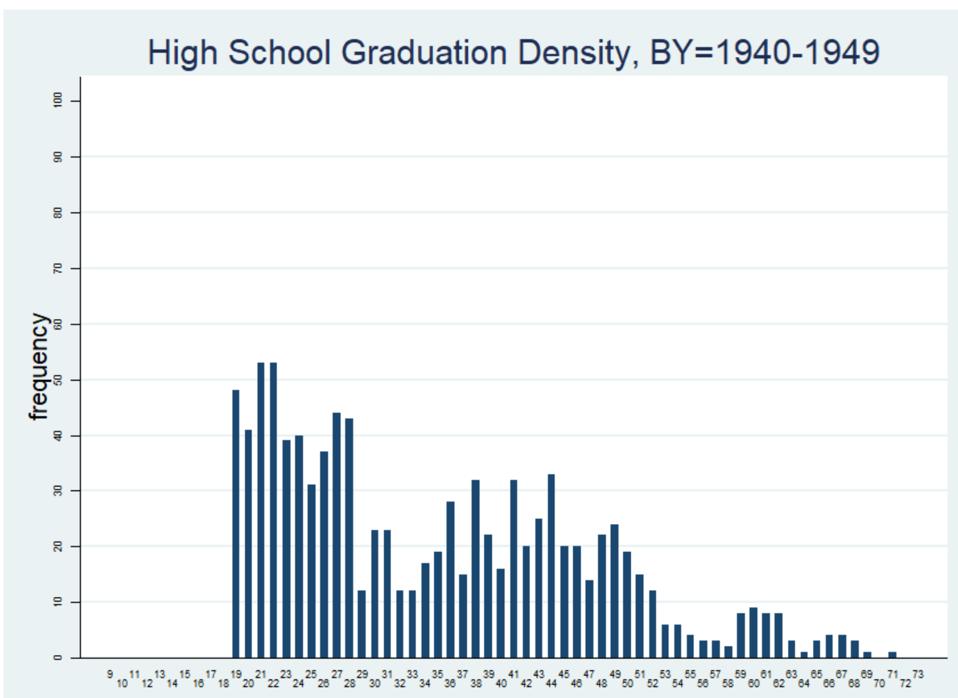

Figure 2.7: High School Graduation Density (Birth Year: 1940-1949)

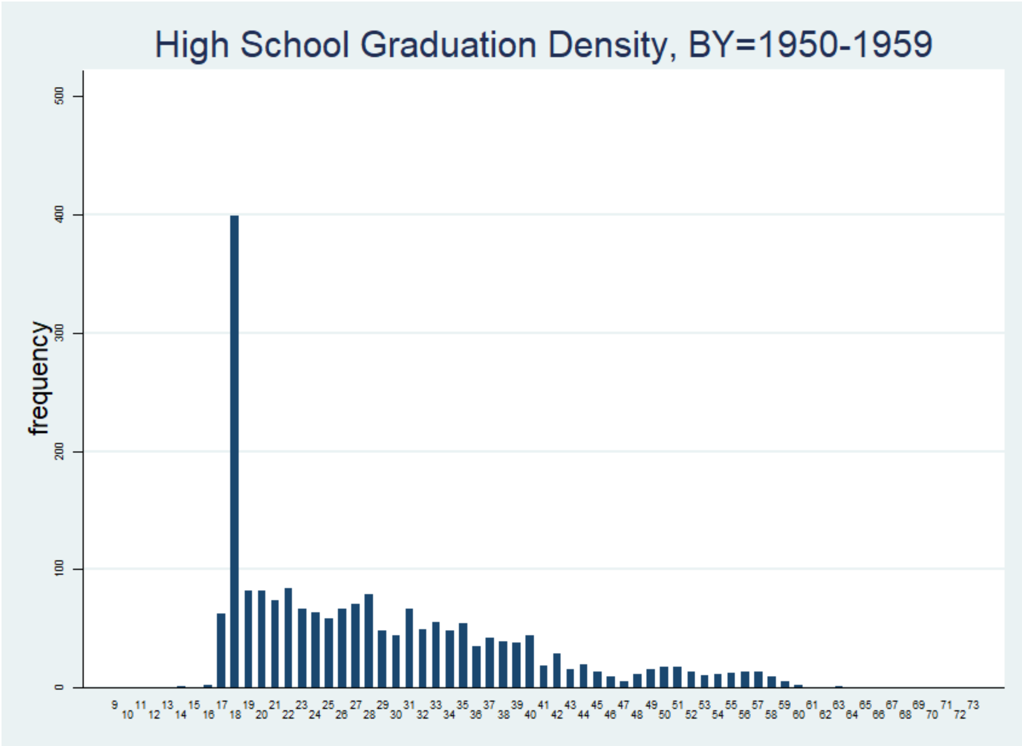

Figure 2.8: High School Graduation Density (Birth Year: 1950-1959)

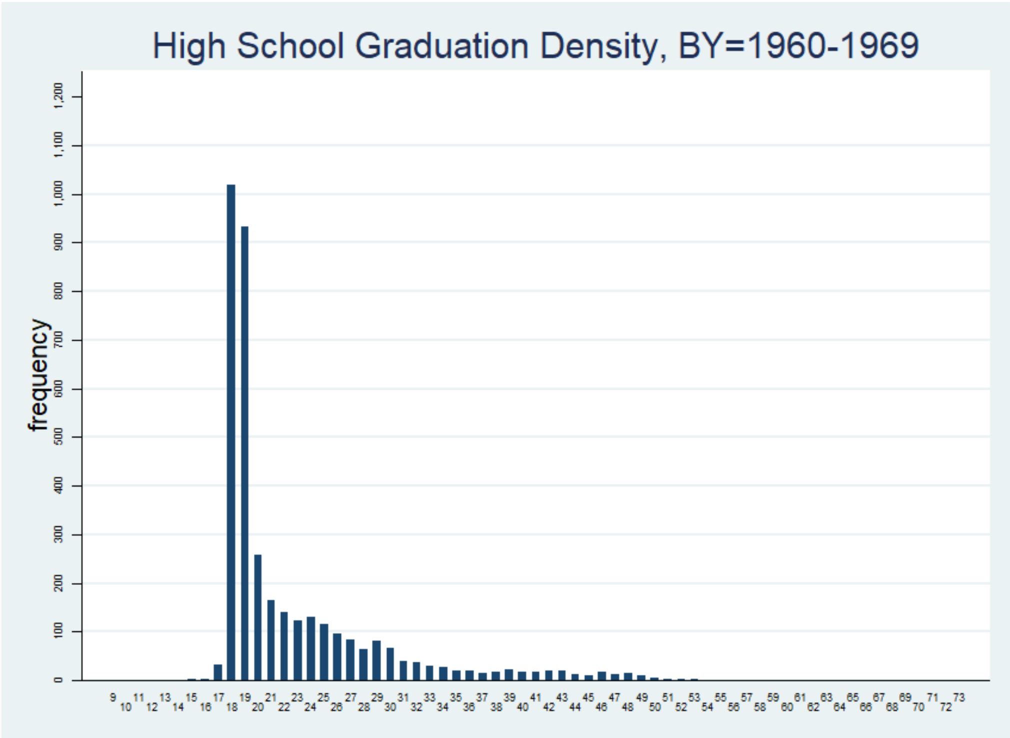

Figure 2.9: High School Graduation Density (Birth Year: 1960-1969)

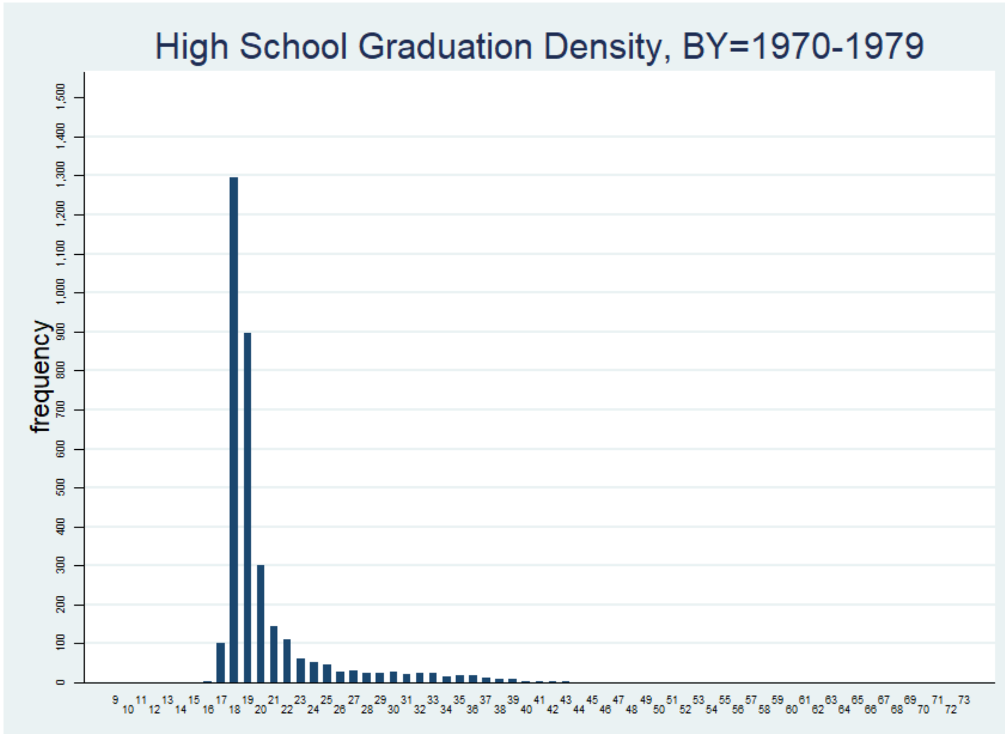

Figure 2.10: High School Graduation Density (Birth Year: 1970-1979)

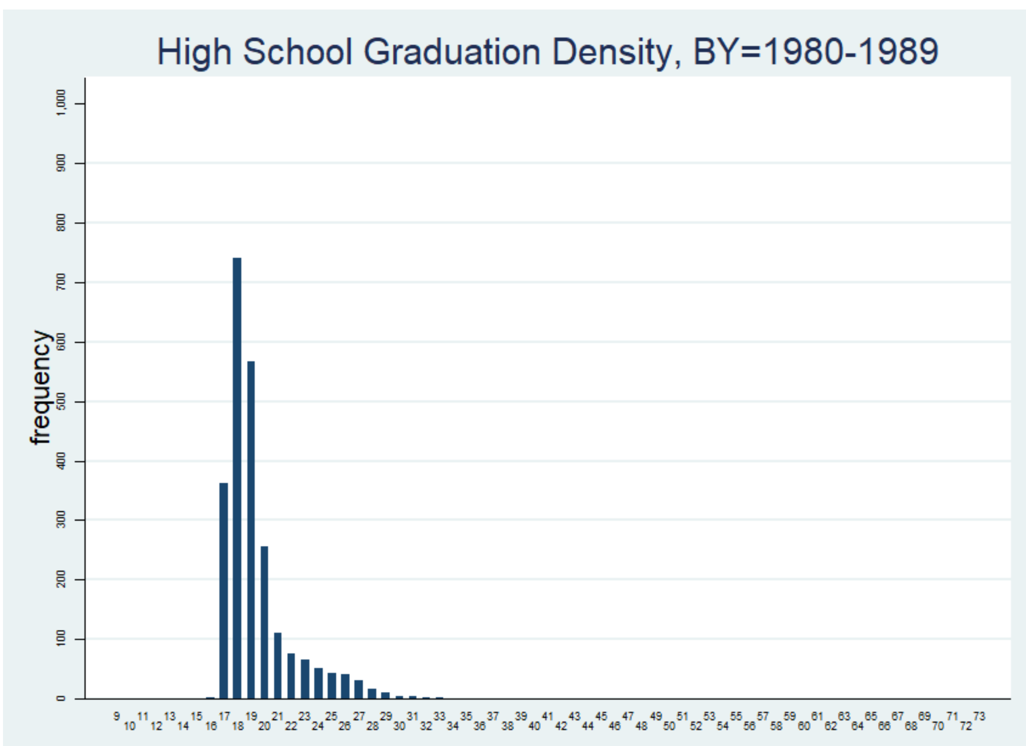

Figure 2.11: High School Graduation Density (Birth Year: 1980-1989)

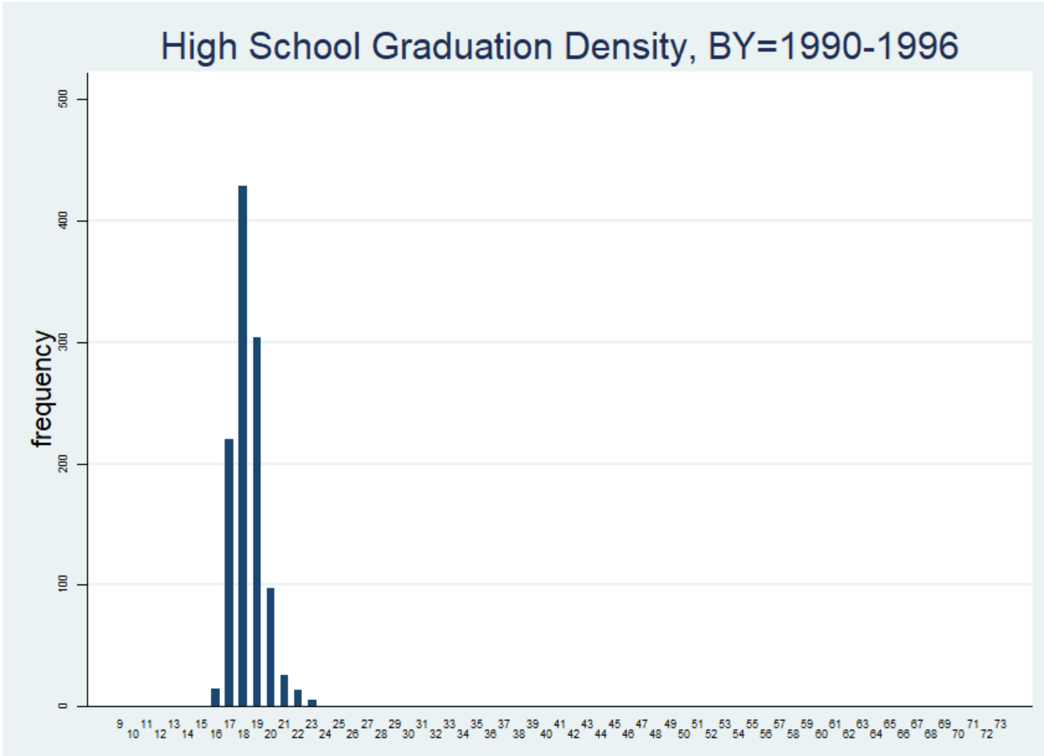

Figure 2.12: High School Graduation Density (Birth Year: 1990-1996)

2.3.2 College Graduation

Table 2.5: Moments of College Graduation

Decade	Mean	Standard Deviation	Standard Error
1940-1949	33.526	14.036	0.718
1950-1959	28.066	11.499	0.445
1960-1969	27.235	6.980	0.313
1970-1979	24.651	6.936	0.334
1980-1989	22.28	4.305	0.255

Table 2.5 provides the moments for the college graduation. Like the average age of high school graduation, the average age of college graduation is declining as we move towards the more recent cohorts. Figure 2.13 provides a graphical representation of Table 2.5. Looking at the figures (Figures 2.16 to 2.20), we can see that the age range is much higher for the 1940s cohort, producing the higher average. Table 2.6 and Table 2.7 present the moments of college graduation of the

truncated samples. The moments of truncated samples are much lower and have small dispersion across the decades.

Table 2.6: Moments of College Graduation Truncated at Age 33

Decade	Mean	Standard Deviation	Standard Error
1940-1949	24.05	3.28	0.21
1950-1959	21.76	5.42	0.25
1960-1969	23.19	5.22	0.27
1970-1979	22.48	5.01	0.26

Table 2.7: Moments of College Graduation Truncated at Age 43

Decade	Mean	Standard Deviation	Standard Error
1940-1949	25.96	5.93	0.36
1950-1959	25.11	8.34	0.34
1960-1969	25.48	7.28	0.34

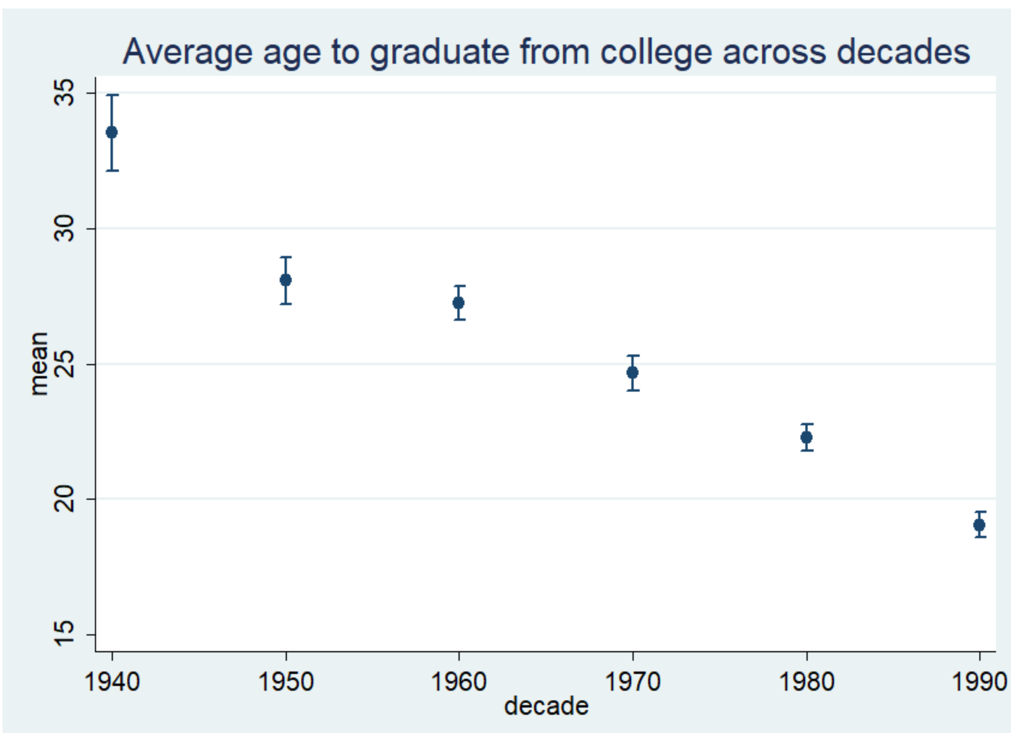

Figure 2.13: Moments of College Graduation

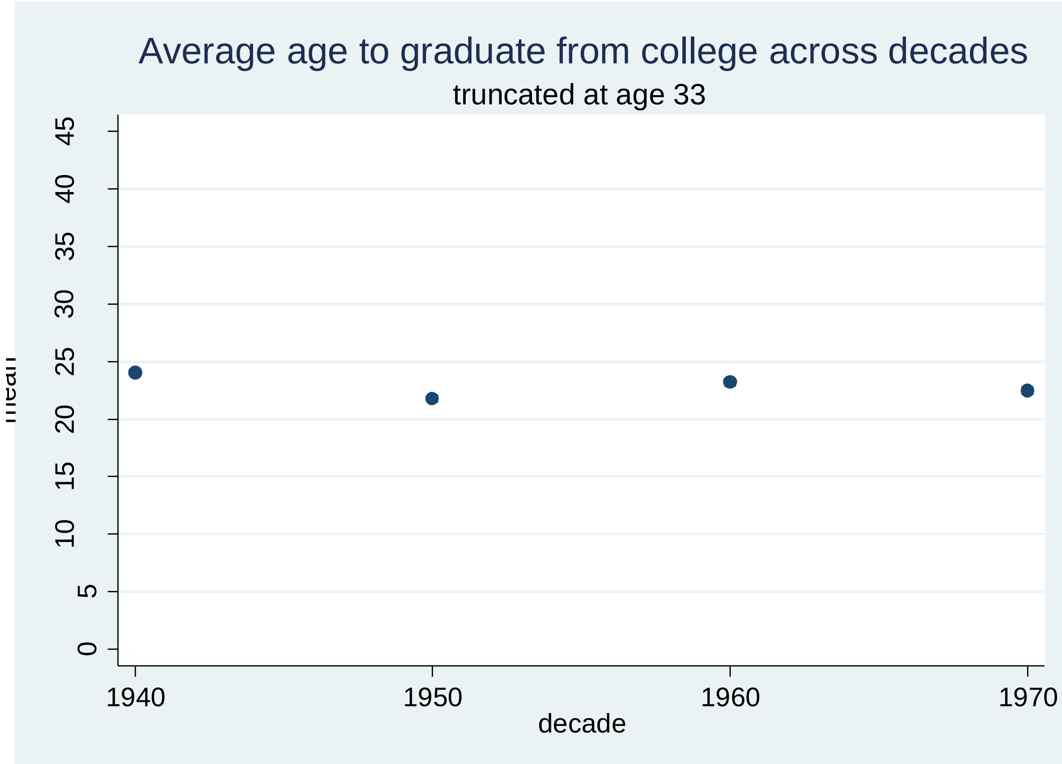

Figure 2.14: Moments of College Graduation (truncated at age 33)

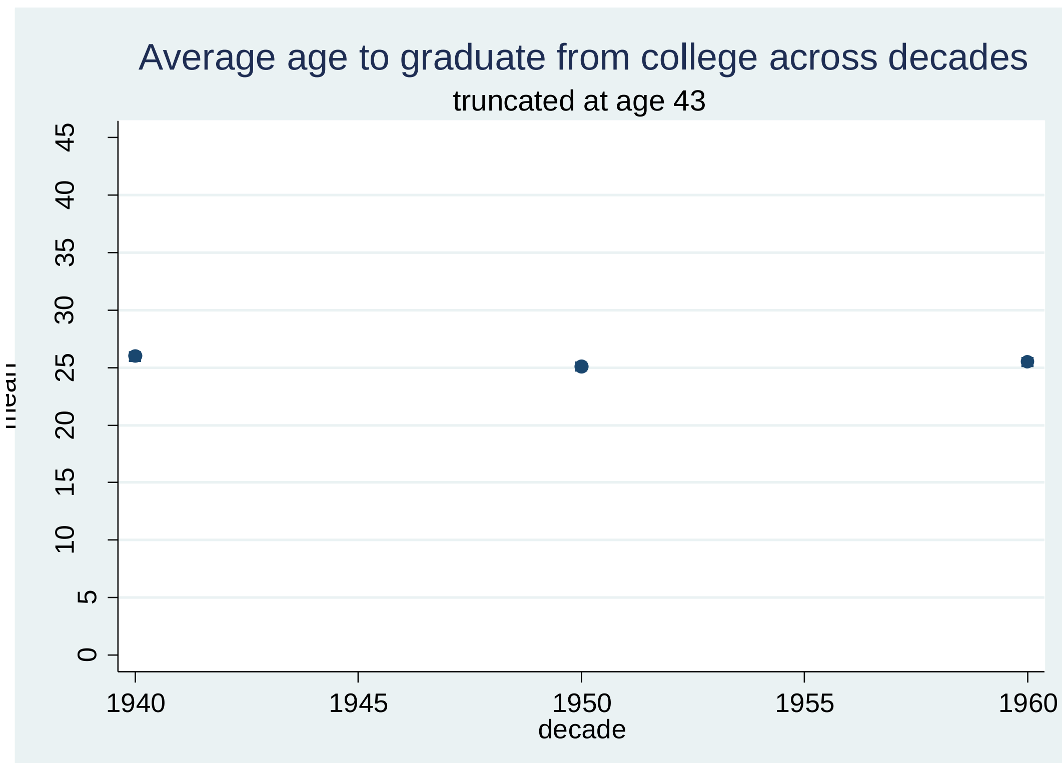

Figure 2.15: Moments of College Graduation (truncated at age 43)

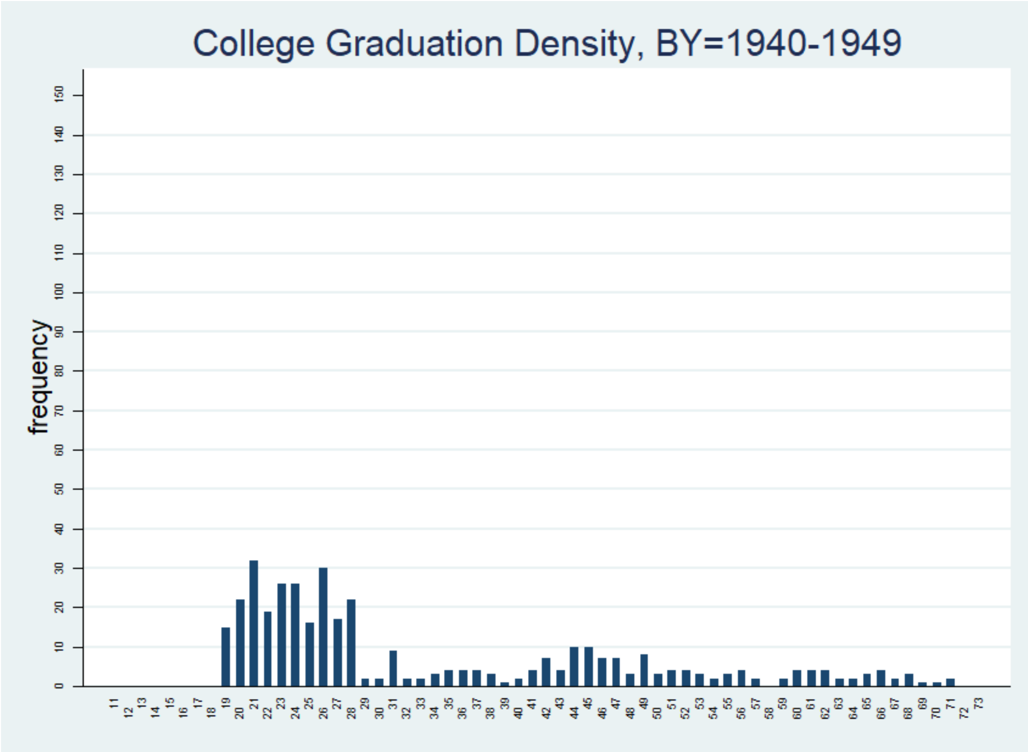

Figure 2.16: College Graduation Density (Birth Year: 1940-1949)

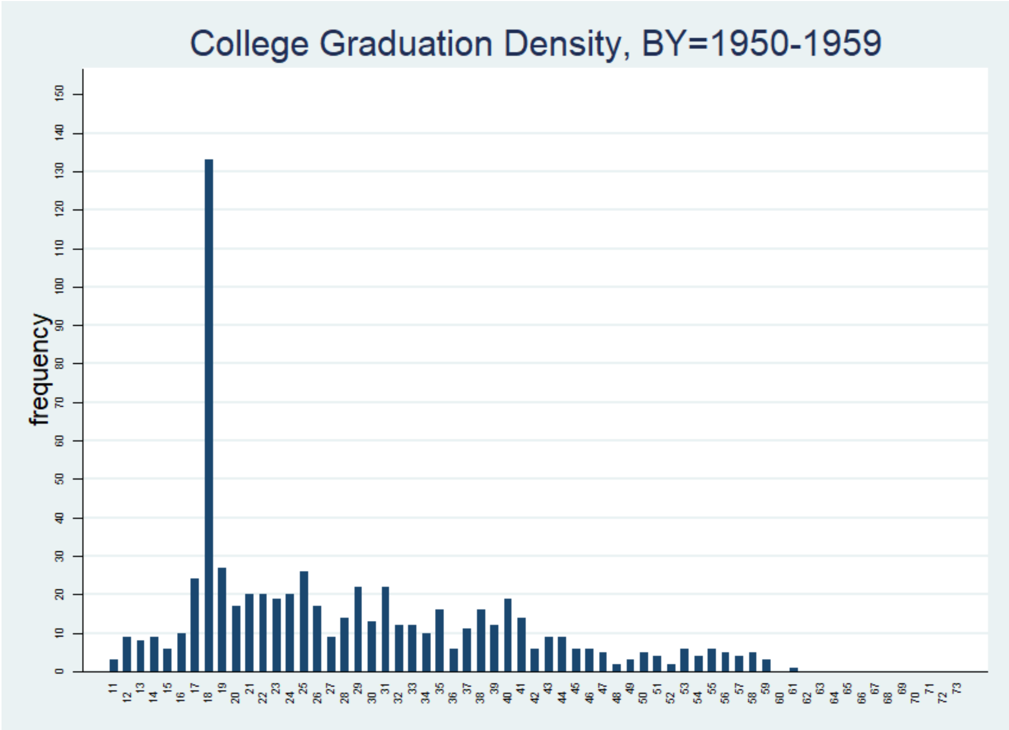

Figure 2.17: College Graduation Density (Birth Year: 1950-1959)

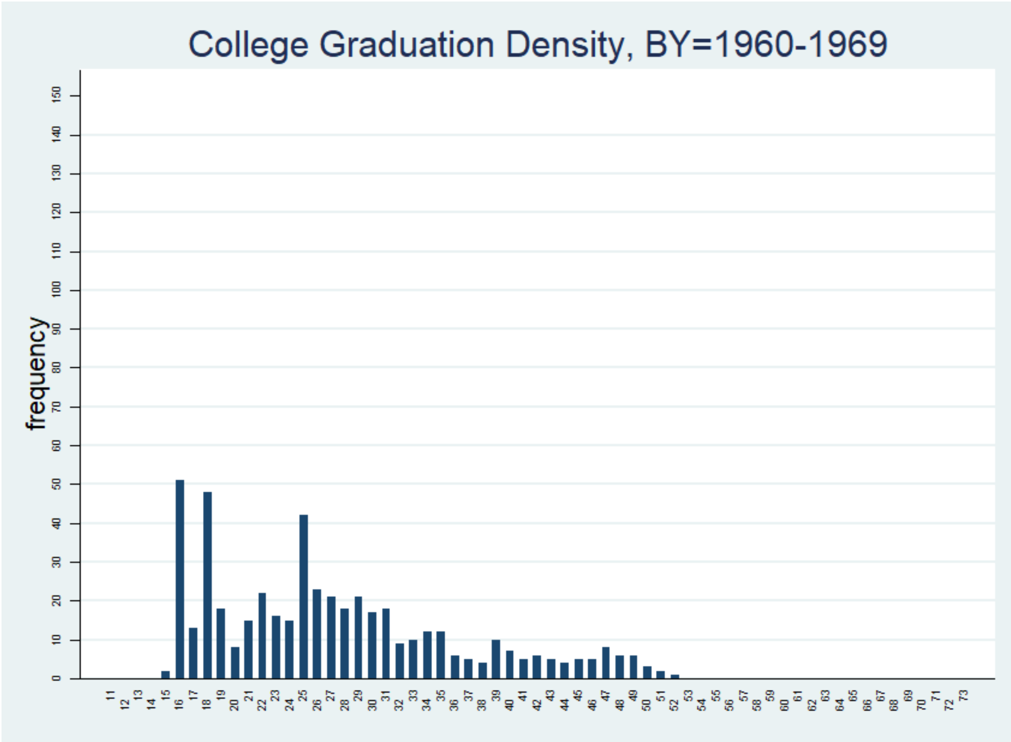

Figure 2.18: College Graduation Density (Birth Year: 1960-1969)

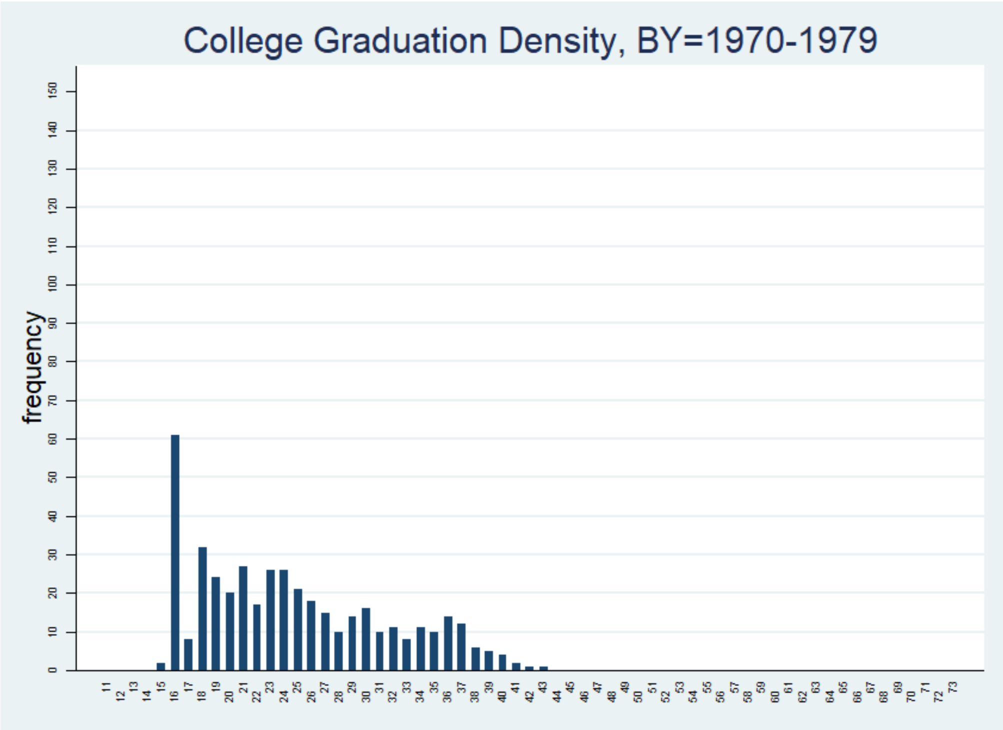

Figure 2.19: College Graduation Density (Birth Year: 1970-1979)

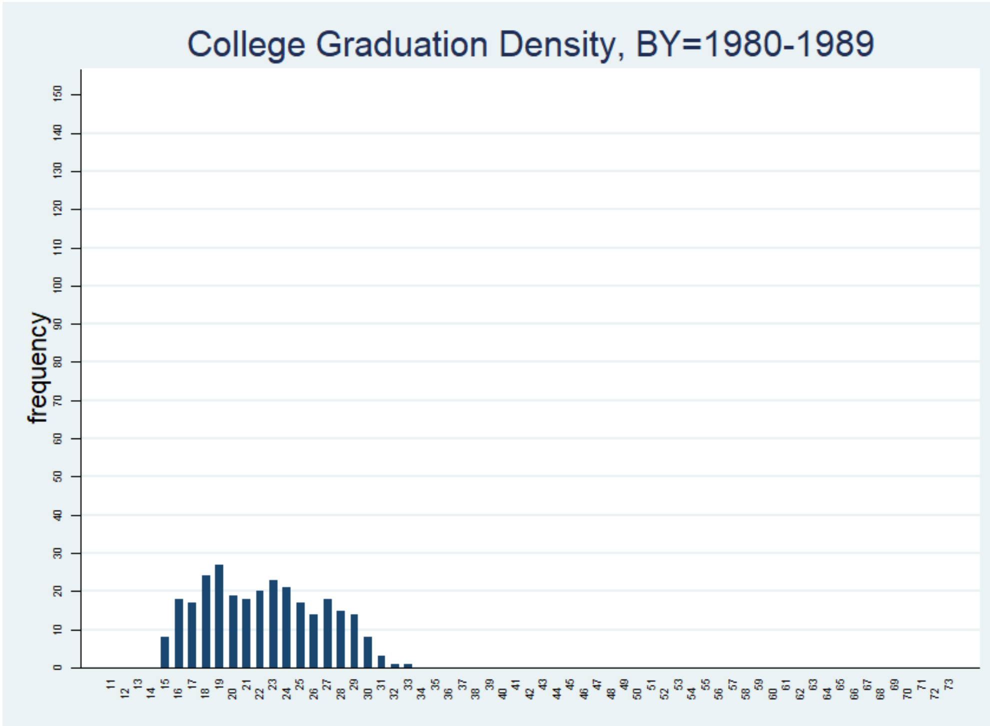

Figure 2.20: College Graduation Density (Birth Year: 1980-1989)

2.4 Conclusion

We investigated a simple yet interesting question in this paper – the changes in graduation age across the years in the US. The idea of this paper came from our rigorous study of different demographic factors and their contribution towards inequality of opportunity. The challenge of this paper was the data limitations of the PSID. PSID data do not directly report the age of graduation for the individuals. Hence, we had to compute the age of graduation (high school or college) from the available data. This paper reports the mean and standard deviation of ages of educational attainment. We found that the average age of graduating from high school and college are declining. While there might be several reasons behind this decline, a good reason would be more older people from earlier cohorts are going back to finish their educations. To further investigate this, we truncated the sample at lower thresholds. The results show a much smaller dispersion. The truncated results also suggests, over time people tend to attain education comparably at a younger age. While it would be interesting to look into the way that the age distribution of educational attainments varies with other demographic factors, that was not possible in this analysis due to data limitations.

Reference:

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, *70*(5, Part 2), 9-49.

Gilardi, S., & Guglielmetti, C. (2011). University life of non-traditional students: Engagement styles and impact on attrition. *The Journal of Higher Education*, *82*(1), 33-53.

Kim, K. A. (2002). ERIC review: Exploring the meaning of" nontraditional" at the community college. *Community College Review*, *30*(1), 74-89.

Metzner, B. S., & Bean, J. P. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in higher education*, *27*(1), 15-38.

Mincer, J. (1962). On-the-job training: Costs, returns, and some implications. *Journal of political Economy*, *70*(5, Part 2), 50-79

Mincer, J. (1974). Schooling, Experience and Earnings. New York: National Bureau of Economic Research.

Organisation for Economic Co-Operation and Development (OECD). (2006). Education at a glance. OECD Publishing. Retrieved August 25, 2018, from http://www.oecd.org/document/

Rendón, L. I., Jalomo, R. E., & Nora, A. (2000). Theoretical considerations in the study of minority student retention in higher education. *Reworking the student departure puzzle*, *1*, 127-156.

Chapter 3
Factors Affecting Income: Education, Experience, and Beyond

Abstract

While there is a huge literature studying the changes across cohorts in returns to different income determinants, the studies lack information concerning what happened prior to 1950. Following Mincer (1974), a wave of papers studied how various factors contribute to the earnings of individuals. This paper contributes to that literature in three ways: (i) using Panel Study of Income Dynamics (PSID) data, it computes the actual working experience of the individuals, (ii) it studies the cohorts who were born in 1910 or afterwards, unlike the existing papers, and (iii) it adds two variables—technological progress and the occupation with which individuals start their careers—to an extended Mincerian equation. The results re-emphasize the importance of education in lifetime earnings. The results also show that while some of the determinants of income have become more important over the years, other factors have not changed much in importance.

3.1 Introduction

The rate of return is one of the primary concern in education economics as well as in labor economics. There are numerous papers analyzing the returns to education using the earnings function introduced by Mincer (1962). Mincer (1962) and Becker (1962) are considered the pioneers of introduction of rates of return in labor economics. They foresaw the change in the U.S. labor market that has taken place starting in the 70s, and Mincer (1974) put forward a model that is considered a major advance in the field. The primary change in the labor market is the great increase in the returns to college education beginning in the 1980s, which slowed down later in the 1990s (Katz and Murphy, 1992; Katz and Autor, 1999; Card and Lemieux, 2001, Carneiro and Lee, 2001). All of these papers looked into the trend in the labor market starting in the 1950s.

Earnings or income is commonly seen as a measure of the return to education. Although education can play a big role in determining the earnings of an individual, earnings do not depend only on education. It has been a great challenge for economists to model income – primarily due to data challenges and the huge presence of unobservable characteristics. It was Becker's (1964) seminal work that showed the path of modeling the internal rate of return to education. The regression coefficient for schooling as a rate of return was justified by Becker and Chiswick (1966). Mincer (1974) put forward a model – famously known as the 'Mincer Equation' - that was later used to estimate the rate of return to schooling. Mincer provided a simple model of income dependent only on education and experience. He provided a simple model due to data limitations. There is a huge existing literature built upon the earlier work done by Mincer. Although the 'Mincer Equation' was consistent with the data available in 1960s, the newly available, richer data made possible an enhanced analysis. Heckman, Lochner and Todd (2006) pointed to the limitations of the Mincer Equation. They argued that, with the availability of new data and tools, the Mincer model cannot provide an adequate estimate of the rate of return to education. Availability of data made it possible to take into account other factors in estimating the rate of return to education. A more recent paper of Ashworth, Hotz, Maurel and Ransom (henceforth AHMR) (2017) provided a model with a much broader set of factors, including family characteristics, individual characteristics, and labor market conditions, along with education and experience, which all play important roles in determining lifetime income.

This paper focuses on the trend of the rate of return across cohorts. Borrowing from AHMR (2017), we want to explore trends in the determinants of income across gender, race, and ethnicity, over time. Unlike the AHMR paper, we include technological progress and the first occupation of the individuals in the model, while we exclude the cognitive abilities of the individuals, due to the data limitations. Moreover, we compute actual experience, instead of using potential experience. We use several specifications in this paper, to obtain rates of return for different factors. We intend to see if technological progress has made any impact on the income of individuals. In addition, our inquiry includes the impact that first occupation can have on the life-long earnings of individuals. This paper contributes to the existing literature by analyzing the earlier cohorts. The earlier papers analyzed the cohorts starting from 1950. Collecting the data from Panel Study of Income Dynamics (PSID), we decompose our datasets into two cohorts – those born between 1910 and 1949 and those born in 1950 or afterwards. We name these two cohorts PSID10 and PSID50, respectively. Finally, we decompose the across-cohort changes in the rates of return. The results show that education, mother's education, and technological investment all became more important over the years, whereas the importance of the father's endowments and social status remained the same.

We use the PSID data for our analysis in order to undertake an analysis on the entire population rather than on particular age cohorts. We use longitudinal analysis in order to measure early-career schooling and work more precisely. From the PSID data, we construct measures of schooling and experience, along with measures of personal and family background characteristics. Using this information, we construct multidimensional measures of human capital investment.

The rest of the paper goes as follows. Section 3.2 presents how the data is constructed; Section 3.3 details the construction of our model and estimation of various specifications of the model; Section 3.4 presents the results of the various specifications; and Section 3.5 discusses the decomposition methods and results. We conclude in Section 3.6.

3.2 Data

We gathered the data from the "Panel Study of Income Dynamics" (PSID), a longitudinal survey conducted by the University of Michigan. We use the data for years 1968-2015. The dataset contains information on the social background and demographic composition of each household in the survey.

Earlier research has showed that the data on work experience is quintessential information in labor economics. The Census and the Current Population Survey (CPS), two of the most representative and largest databases in the USA, do not collect the information on work experience. The National Longitudinal Survey (NLS) and the Panel Study of Income Dynamics (PSID) are the two most highly regarded sources of data in labor economics. Both of these collect work history information.

There are several papers that use data collected from NLS. In fact, AHMR (2017), the paper that inspired this analysis, employed NLS data. This paper differs from AHMR primarily in two ways – the data source and the way that experience is computed.

Why this paper employed the PSID data

As mentioned in Blau and Kahn (2008), the NLS does not represent the entire USA population. Rather, it focuses on particular age cohorts, unlike the PSID data, which represents the entire population with a sample of 5000 households and hence covers the entire working-age range. In addition, PSID data relies on respondents' memories of their work histories, based on annual interviews in which the activities of the past year are catalogued. In this analysis, we use that information to compute the actual work experience of the individuals. Blau and Kahn (2008) found that such retrospective experience data provides information that matches up with the data based on other annual surveys.

For this analysis, we have taken into account variables that can be classified as individual endowments, social endowments, demographic endowments, and local labor market conditions.

Individual Endowments:

Individual endowments include experience, education, and first occupation. We used both the potential experience and actual experience to trace the differences in results, if any.

Typically, in labor economics experience is calculated as:

$$experience_{potential} = age - education - 6$$

Actual experience is imputed using total number of years of experience, number of years of experience with the present employer, and start and end with the previous employers. Due to wide array of years of collecting data, the data frequently varies inconsistently across the years. Hence, special care must be taken to extract the greatest amount of reliable information from the data. In this work we calculated the experience from present and past employment separately. Adding the two would give us the total experience.

$$experience_{total} = experience_{present} + experience_{previous}$$

We select the maximum years of experience for each individual, when the data are inconsistent, and then match it with the maximum income (explained later) of the individual. For data that do not match, we replace the missing values with the maximum years of experience. In order to remove any discrepancy, we get help from the *year of interview* data. We adjust all the data points that do not match with the *year of interview*. As stated earlier, the maximum years of experience is matched with the maximum income of an individual, and we generate a separate variable to represent the difference in the years that the information on these two variables is reported. We then add this difference to the total experience to get the actual experience, only if the total experience data is missing and the difference is less than zero. In addition, if any value of the total experience is non-positive we replace the value with the total experience.

One of the ways we tried to extract the information on *experience* was to examine the maximum years of experience reported. If someone reports wrong information, one way learn this is to crosscheck this information with the reported birth year[11] of the individual. We drop those individuals whose reported years of experience are inconsistent with their birth year information. This finally gives us the actual experience.

$$experience_{actual} = experience_{total}$$

Education level is classified as less than high school, high school, or college education. The first occupation of the individual is categorized as blue-collar jobs and white-collar jobs.

Social Endowments:

The social endowment variables used in this research are father's/mother's education (FE/ME), father's occupation (FO), and socioeconomic condition/income groups (IN). Parental education is categorized into less than high school education, and at least high school education. The occupation of father is categorized into blue-collar jobs, and white-collar jobs. We separate the income groups into poor, middle class, and rich.

Demographic Endowments:

For demography, we collected information on race and gender. All the individuals in the sample are either black, white, or Hispanic. The limited numbers of individuals in the sample do not let us consider individuals of other races. The genders considered are male and female.

Local Labor Market Conditions:

The variables considered for the local labor market are the unemployment rate and technological investment. The unemployment data is collected from the Bureau of Labor Statistics (BLS) for each of the states for 1976-1992. The geographical location of the individuals is matched with the

[11] The birth year of an individual is derived in two steps. First, we calculate a reported birth year for each time an individual is interviewed, by subtracting the reported age in that year from the interview year. In the second step, using the modal value of the reported birth year, we record the standardized birth year for every individual.

location of the current employment of the individuals. Moreover, the year the unemployment rate is matched with the year the income is reported.

The investment in technology has been computed after collecting the data from the Industry Statistics Portal of the U.S. Census Bureau. The data is collected for total R&D investment (flow), current cost of net stock (stock), and current cost of depreciation for 1959-2004. Using this information, we compute the stock of computer and electronics using the following formula:

$$Total\ Investment = Stock + Flow - Depreciation$$

We match the information on technological investment with the interview year reported in the dataset.

Dependent Variable:

The dependent variable in our analysis is the logarithm of labor income of the individuals. We have collected the data for the labor income and taxable income of the individuals. To extract the information on labor income from the taxable income data, we matched the taxable income with the type of income. This has been done in order to ensure the accuracy of the information for most of the individuals. In order to maintain purchasing power parity, we further transformed the nominal income into real income in 2016 dollars. Lastly, we take the log value of the real income.

3.3 Model

We intend to develop a model to estimate the returns to the several factors we have talked in the last section. Our goal is to build a model that takes into account the endogeneity of schooling and experience in estimating the returns across the two birth cohorts.

We desire to estimate the effects of different input factors on income. In particular, we are interested in total years of schooling, years of work experiences. Our model is capable of estimating the effect of different educational attainment, particularly high school and college graduation, on these outcomes. In the following, we will refer to these work experiences, occupation schooling activities and graduation outcomes collectively as *individual endowments*. The vector of individual endowments is given by:

$$(1) x_i^r \equiv (x_{1i}, x_{2i}, x_{3i}, x_{4i}^r, I_i(R_i = 1), I_i(R_i = 2)\ I_i(R_i = 3))'$$

where the variables are: x_{1i}, the number of years of total schooling years; x_{2i}, first occupation of the individual; x_{3i}, the number of years of experience; x_{4i}^r, the number of years of experience with certain educational attainment; $I_i(R_i = 1)$, an indicator equal to 1 if an individual has less then high school education; $I_i(R_i = 2)$, an indicator equal to 1 if an individual has completed at least high school education; and $I_i(R_i = 3)$ an indicator equal to 1 if an individual has completed at least college education. The individual endowments variables are collected from the starting age of 16[12].

Let W_{ij} be the potential income of individual *i* if he is in the activity set *j*=1,2,3. We assume the W_{ij} be determined by the human capital H_i, the occupation-specific skill price of the individual across the local labor market M_i:

$$(2)\ W_{ij} = H_i M_{ij},$$

The log of income, denoted by w_{ij}, is as below:

$$(3)\ w_{ij} = h_i + m_{ij}$$

where $h_i \equiv \ln H_i$ and $m_{ij} \equiv \ln M_{ij}$. We assume that m_{ij} is the function of the labor market in which the individual *i* resides, l_i; and the technological investment in the economy, t_i:

[12] This part is taken care of in the data section

$$(4)\ m_{ij} = \beta_{0j} + \beta_l l_i + \beta_t t_i$$

We assume that the individual's stock of human capital, h_i, is determined by some of the personal characteristics of the individuals, e.g., birth year, race, gender etc., denoted by the vector z_i; the family characteristics, f_i; years of schooling, degree completion, first occupation, and work experience, x_i^r:

$$(5) h_i = \beta_z z_i + \beta_x g(x_i^r) + \beta_f f_i$$

Following (4) and (5), equation (3) can be written as:

$$(6)\ w_{ij} = h_i + m_{ij} = \beta_{0j} + \beta_l l_i + \beta_t t_i + \beta_z z_i + \beta_f f_i + \beta_x g(x_i^r)$$

where $g(.)$ includes: (*i*) a cubic polynomial in experience (*ii*) interactions between school attainment and experience , and (*iii*) indicators of having a high school degree and a college degree (Heckman, Lochner and Todd, 2006).

Lastly, to measure the effects of family background, we ran the full model once and then controlled for the local labor market conditions and the occupation of the individual.

3.4 Results

3.4.1 Returns to Education

Table 3.A.1.1 exhibits the estimates of returns to year of schooling for various specifications. Panel *(a)* presents the returns to an additional year of schooling while panel *(b)* and *(c)* estimates the 'sheepskin effect' of graduating from high school and college, respectively. We reported the results for both the cohorts – PSID10 (1910-1949) and PSID50 (1950 and above). There are eight different specifications. We start with the raw premia in row (i) and end with the full model in row (viii).

We start in Panel *(a)* with the Mincerian and Heckman, Lochner and Todd (HLT) specifications. For the whole sample, the returns to an additional year of schooling are 6.30% and 7.95% for these two specifications respectively. For the PSID10 cohort, the Mincerian equation shows a return of 5.8% while the return is 6.9% for the PSID50 cohort. The HLT specification shows an opposite result, where the older cohort shows a higher return than the younger cohort does by 2.8 percentage points. In both of these specifications, we use the potential experience as explained in the earlier section. If we replace the potential experience with the actual experience(row (*iv*)) in HLT, the returns to schooling falls to only 5.90%, which is lower than for the previous two specifications. The returns to schooling for PSID10 cohort with actual reported experience is 6.7% and that of PSID50 cohort is 5.35%. Adding the socioeconomic endowments to the equation (row (*v*)) reduces the returns to an additional year of schooling further to 5.6%. For PSID10 cohort it is reduced to 6% while for the PSID50 cohort it goes up to 5.56%. In row (*vi*), we add the local labor market condition (unemployment rate). Adding unemployment does not change the estimate significantly. In row (*vii*) and (*viii*) we add technological investment and first occupation of the individual respectively. Adding these two reduce the differences between the two cohorts.

Panel *(b)* shows the returns to graduation from high school. The returns to graduate from high school varies from 6% to 16% depending on the specification. The return is highest for the Mincerian specification which shows that those graduated from high school has 16% more return than those who have less than high school education. For high school returns, PSID50 leads the PSID10 cohort across all the specifications. The returns for the PSID10 cohort is in the bracket of 4.2% to 13.6% whereas for the PSID50 the bracket is 9.5% to 19.2%. When adding the social endowments in the specification (row (*v*)), PSID10 cohort shows 11.1% higher return while the

PSID50 cohort shows a 12.5% higher return. If we ignore the Mincer specification, row (vi) gives us the highest returns for both the cohorts.

Panel *(c)* represents the returns to graduate from college. The return is much higher for graduating from college compared to graduating from high school. Those who graduate from college has 19% to 39% more return than those who graduated from high school but did not graduate from college. For both the PSID10 and PSID50 cohorts the return is the highest when we introduce the actual experience to the specification – 26.1% and 22.90% respectively. As can be noted, in most of the cases, the Mincerian specification overestimates and HLT specification underestimates the returns compared to our specification.

3.4.2 Returns to Experience

Table 3.A.1.2 represents the returns to experience for different specifications and for different educational achievement. Column 1 of Panel *(a)* shows the returns to experience for the whole sample across all the education level. Columns (3) and (5) of panel (a) presents the estimates for the PSID10 and PSID 50 cohort respectively. The specifications chosen for this analysis are limited as it makes the analysis more precise and to the point. We chose Mincerian and HLT specification to find how these specifications differ from the specification in which we introduce the actual experience that we computed. The other specification that we chose is the full model where we controlled for the social endowments, local labor market, technological investment, and first occupation of the individual. Panel *(b)*, *(c)*, and *(d)* presents the estimated returns to experience for different education hurdle – less than high school, at least high school, and at least college.

For the whole sample, the return is in the range of 2 percent to 4.8 percent. The return is the highest for the HLT specification. For Mincer specification, the return is 2.3% which is higher than what we found after substituting actual experience – 2.2 percent – in the specification. All the specifications, except for row (*iii*), exhibit that the return is higher for the PSID50 cohort. When we introduce the actual experience, the return is -1% for PSID50 cohort.

Returns to experience varies across the educational achievement. Panel *(b)* and *(c)* show that the return is higher for those who have at least high school education. For those who could not cross the hurdle of high school, the Mincer specification shows the return is 1.5% whereas the HLT specification shows the result to be 2.8%. If we substitute the actual experience for the potential experience the estimation is 1.1% - much lower than row *(i)* and *(ii)*. For the full model the estimation does not vary much from row *(iii)*. However, with high school education the return is much higher. Row (*iii*) of panel (c) shows the return is 2.4%, double that of panel (b), when we use the actual experience in the specification. The return is 2.1% when we use the full model specification. Mincer specification shows the highest return – 2.6% - for those who at least graduated from high school. If we compare column (3) of both panel *(b)* and *(c)*, we can see the return is higher for those who have at least graduated from high school for row *(i)*, *(iii)*, and *(iv)*. We get the same result for column (5) for the panels.

Panel (*d*) of the table shows the returns to experience for those who has at least graduated from college. Column (1) of the panel shows the estimation are negative except for the Mincer specification. Column (3) of the panel shows the return is negative for the PSID10 cohort. However, for the PSID50 cohort the return is positive (except for the HLT specification) and higher than that of the PSID10 cohort. The return is 12 percent when we use the actual experience and is the highest across all the specifications.

3.4.3 Returns to Social Endowments

Our model also accounts for the social endowments and its impacts on the income of an individual. As can be noted, social endowments can play a big role in earning through education. In this paper we chose to pick father's occupation, parental education, socioeconomic group, race, and gender in the social endowment set. Table 3.A.1.3 represents the estimations of the social endowments. Panel (*a*) exhibits the specification without controlling for local labor markets, technology, and the first occupation of the individuals whereas panel (*b*) controls for these variables. To put in another way, panel (*a*) is row (*v*) of Table 1 and panel (*b*) is the row (viii) of Table 1. Column (1) of Table 2 presents the estimation for the whole sample while columns (3) and (5) presents the estimations for PSID10 and PSID50 cohorts respectively.

We start with the father's occupation in panel (*a*). The reference group for father's occupation is the blue-collar occupation. Column (1) shows that the children of the white-collar professionals are earning more than the children of the blue-collar professionals for the whole sample. For the PSID50 cohort, if the father is a white-collar professional their children earn 8% more than those of blue-collar professionals. The earning is much higher for the PSID50 cohort. For the full model specification in panel (*b*), the results are very similar. For the full sample, the children of white-collar professional fathers are earning 3.8% more than the reference group whereas the estimation is 1.2% and 5.3% for the PSID10 and PSID50 cohort respectively.

Next, we consider the parental education. Panel (*a*) shows that children of the parents with higher education have higher earning ability. Children of the fathers with at least high school education earn 6% more than the children of the fathers with no high school education. Children in the PSID10 cohort earn 10% more if their fathers are at least high school graduates. The children whose mothers are at least high school graduates earn 6% more than if their mothers are not high school graduates. The returns are 6.7% and 4.1% more for the PSID10 and PSID50 cohorts. The results are slightly different if we choose for the full model specification. The return is 2.8% more if fathers are high school graduates than if not. For the PSID10 cohort, the return is 5.6% more. For mother's education, the return is 7% - a little higher than what we have seen in panel (*a*). For the PSID10 and PSID50 cohort the returns are 8% and 5.5% more respectively if mothers are high school graduates.

The socioeconomic groups are a monotonic function of income for the full sample where the middle-class children earn more than their poor counterparts and the rich children earn more than both the other groups. For the full model specification, the middle-class children earn 1.6% and the rich earn 2.2% more than the poor children. Similarly, the whites lead the blacks and the Hispanics in earning. Panel (*b*) shows that the blacks and the Hispanics earn 11.7% and 22.6% less respectively than their white counterparts. The statement is true and consistent across both the cohorts. Lastly, the males lead the females in earning for the whole sample and across the cohorts. In panel (*b*), the females earn 40% less than their male counterparts. The difference is 50% and 30% for the two cohorts respectively.

3.4.4 Other Important Observations

There are some more important observations in our analysis. Once we include the technological progress in our analysis, the result shows that technological investment helps the poor and middle class children to earn more. We also included the first occupation of the individuals in the full model specification. The result shows that if the first occupation is a white-collar job, the earning is 3.4% higher than if it were a blue-collar job. If we separate the two cohorts, the difference is only 1.6% for the PSID10 cohort while the difference is 3.7% for the PSID50 cohort.

The results also found that males are leading females across all the birth years. For the racial analysis, we find that the whites are leading followed by the blacks. The pay gap between black and whites have always been declining but at a very flat rate. The results are reported in Table 3.A.1.3.

3.5 Decomposition

3.5.1 Decomposition Methods:

Decomposition methods are widely used to explain the differences in outcome i.e. wage, health, to find out the trend in outcome differences, to investigate the contribution of the factors in creating these differences. In labor economics, following the seminal papers of Oaxaca (1973) and Blinder (1973), decomposition methods have been widely used to look into the pay gap across different social groups starting from gender to race to socioeconomic groups. In this paper, the Oaxaca-Blinder decomposition method explains the difference in the means of wage between the two groups (e.g., born before and after 1950). The method further analyze the contribution of different factors into formulating that gap.

We have two interest groups: those who were born before 1950 and those who were born in or after 1950. For notational purpose we name the group a for those born in or after 1950 and b for the other. The wage equation can be written as follows:

$$(7)\ w_i = u_i \beta^g + v_i^g$$

Where, u is the vector of determinants

$a, b \in g$

$E\left(v_i^g\right) = 0$

At the mean,

$$(8)\ \bar{w}^a - \bar{w}^b = \bar{u}^a\hat{\beta}^a - \bar{u}^b\hat{\beta}^b$$

Let,

$$(9)\ \Delta\beta = \hat{\beta}^a - \hat{\beta}^b$$

$$(10)\ \Delta u = \bar{u}^a - \bar{u}^b$$

Using (3) and (4),

$$(11)\ \bar{w}^a - \bar{w}^b = (\Delta u + \bar{u}^b)\hat{\beta}^a - \bar{u}^b(\hat{\beta}^a - \Delta\beta) = \Delta u\hat{\beta}^a + \bar{u}^b\Delta\beta$$

$$(12)\ \bar{w}^a - \bar{w}^b = (\Delta u + \bar{u}^b)\hat{\beta}^a - \bar{u}^b(\hat{\beta}^a - \Delta\beta) = \Delta u\hat{\beta}^b + \bar{u}^a\Delta\beta$$

Equation (5) and (6) can be written as,

$$(13)\ \bar{w}^a - \bar{w}^b = \bar{u}^a\hat{\beta}^a - \bar{u}^b\hat{\beta}^b = \Delta u\hat{\beta}^b + \bar{u}^b\Delta\beta + \Delta u\Delta\beta = E + C + CE$$

The difference in the mean wages is comprised of differences in endowments (E), differences in coefficients (C), and differences in the interaction of the two (CE).

Oaxaca's decomposition can also be written as a special case of the following decomposition:

$$(14)\ \bar{w}^a - \bar{w}^b = \Delta u\left[D\hat{\beta}^a + (I - D)\hat{\beta}^b\right] + [u^a(I - D) + \bar{x}^b D]\Delta\beta$$

Where *I* is the identity matrix and *D* is a matrix of weights. If D=1 we retrieve equation (5) and if D=0 we retrieve equation (6).

Besides Oaxaca-Blinder decomposition, in this paper, we have used some other formulations. Cotton (1988) proposed to weight the differences in the *u*'s using the average of the coefficient vectors.

$$(15)\ diag\ (D)\ =\ 1/2$$

where diag (D) is the diagonal of D. Reimers (1983) weighted the coefficient vectors based on the sample fraction of the two groups. If f_a is the proportion of group a, we have,

$$(16)\ diag\ (D) = f_a$$

Finally we used the Neumark (1988) decomposition method which use the coefficients of the pooled data regression, β^P:

$$17 \quad w^a - w^b = \Delta u\beta^a + [u^a(\beta^a - \beta^P) + u^b(\beta^P - \beta^b)$$

3.5.2 Decomposition Results[13]:

Table 3.A.1.4 exhibits the average of income for the two groups – those born before 1950 (high group) and born after 1950 (low group) – and the difference between the groups. The table also presents the contribution in the difference by endowments (E), coefficients (C), and the interaction (CE). As can be seen, the highest contribution in the gap comes from the unexplained portions (C) of the outcome.

The second block of output (Table 3.A.1.5) reports how the different methods of decomposition change the explained (E) and unexplained (C) portions of the gap. The first two columns correspond to the Oaxaca decomposition whereas the third and fourth columns relates to the Cotton and Reimers (for the third column D=0.5 and for the fourth column D=0.506) respectively. Lastly, the last column "*" shows the Neumark's decomposition. For all the methods, the mean values of β's contributes more to the wage gap between the two cohorts.

[13] All the above mentioned model can be can be computed using Ben Jann's Stata routine *decompose*.

Table 3.A.1.6 exhibits each of the individual *u*'s contribution to the overall explained gap. All the decomposition results are consistent in the direction of the gap. Education, Mother's education, and technological investment – all favors more the individuals born on or after 1950. This in turn shows that over the years these variables became more important in earning wage. For example, education plays a more important role in the wage function from 1950 and onwards than it did before 1950. There is virtually no difference between the two groups for some of the *u*'s – father's endowments, socioeconomic status. The two demographic variables – race and sex – favor those born before 1950. For all the decomposition methods, experience accounts for the majority of the explained gap. There is also a heavy presence of race in determining the earning of the individuals. Social background does not directly put any impact over income.

Table 3.A.1.7 provides the estimates, means, and predictions for each *u* for both the two cohorts.

3.6 Conclusion

This goal of this paper was to find out the rates of return to different factors of income. Specifically, we wanted to look into the rate of returns for schooling, experience, and social endowments. We investigated the rate of returns across the two birth cohorts – PSID10 and PSID50. We ran this analysis across several specifications. The results show that the returns to an additional year of schooling is significant – both mathematically and statistically. The return is higher for the more recent cohort. The results also show that the return is more for higher educational attainment. The return for experience, in general, shows that the return is higher for the PSID50 cohort. Our results show that the returns reported in the HLT paper overestimated the actual return. Experience with at least high school education provides much higher return than without high school degree. The returns with a college degree is also higher for the later cohort. We also looked into the returns to the socioeconomic backgrounds of the individuals. We found that income of an individual is a monotonic function of socioeconomic endowments and better endowments resulted in higher returns. The analysis also found that if the first occupation of an individual is a white-collar job the return is higher than if it were a blue-collar job. The paper can loosely conclude that the technological investment is progressive in manner. Lastly, the paper found that the racial and gender gap in earning still exists.

Another goal of the paper was to find out the contribution of different factors of income. We found that the experience has the highest contribution to the explained gap followed by education and race. Education, mother's education, and technological investment all favors the later cohort while other social endowments do not favor or disfavor any of the cohorts.

To conclude, this paper contributed to the existing literature in three ways: (1) using PSID data to compute actual experience, (2) analyzing the cohorts before 1950, and (3) using technological progress and individual's occupation in the extended mincer equation. A rigorous analysis will be required to see the real impact of the technological progress on the income inequality, which can be one significantly important research avenue.

Reference:

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, *70*(5, Part 2), 9-49.

Mincer, J. (1962). On-the-job training: Costs, returns, and some implications. *Journal of political Economy*, *70*(5, Part 2), 50-79.

Becker, G. S. (1964). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. New York: National Bureau of Economic Research, distributed by Columbia University Press.

Becker, G. S. and B. R. Chiswick (1966, March). Education and the distribution of earnings. The American Economic Review 56(1/2), 3586369.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 436-455.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 693-709.

Mincer, J. (1974). Schooling, Experience and Earnings. New York: National Bureau of Economic Research.

Katz, L. F., & Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics*, *107*(1), 35-78.

Katz, L., Autor, D., Ashenfelter, O., & Card, D. (1999). Handbook of labor economics. *Handbook of Labor Economics*, *3*.

Card, D., & Lemieux, T. (2001). Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *The Quarterly Journal of Economics*, *116*(2), 705-746.

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. Handbook of the Economics of Education, 1, 307-458.

Blau, F. D., & Kahn, L. M. (2008). Women's work and wages. *The new Palgrave dictionary of economics*,, *8*, 762-772.

Carneiro, P., & Lee, S. (2011). Trends in quality-adjusted skill premia in the United States, 1960-2000. *American Economic Review*, *101*(6), 2309-49.

Ashworth, J., Hotz, V. J., Maurel, A., & Ransom, T. (2017). Changes across Cohorts in Wage Returns to Schooling and Early Work Experiences (No. w24160). National Bureau of Economic Research.

Appendix

3.A.1 Tables

Table 3.A.1.1: Measures of wage returns to schooling across specifications

	Full		1910-1949 Cohort		>1950 Cohort	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Panel *(a)*: Return to Years of Schooling						
(i) Raw	0.0539667***	0.002908				
(ii) Mincer	0.0630648***	0.002617	0.0584147***	0.003402	0.0693929***	0.004292
(iii) HLT	0.0795534***	0.005654	0.0819193***	0.009527	0.0539976***	0.00945
(iv) Actual Experience	0.0589547***	0.004288	0.0672287***	0.006238	0.0534608***	0.007182
(v) Background	0.0557786***	0.004198	0.0601054***	0.005991	0.0556416***	0.00711
*(vi)*Unemployment	0.0568525***	0.004349	0.0609107***	0.006545	0.0587314***	0.007116
(vii) Technology	0.0556779***	0.004375	0.0592606***	0.006556	0.0575259***	0.007141
(viii) First Occupation	0.0547704***	0.004397	0.0586917***	0.006631	0.0569908***	0.007151
Panel *(b)*: Return to Graduation from HS						
(i) Raw						
(ii) Mincer	0.1604922***	0.017489	0.1364664***	0.022612	0.192364***	0.027822
(iii) HLT	0.0800969***	0.022456	0.0417725	0.030858	0.0947085***	0.035469
(iv) Actual Experience	0.0885028***	0.020993	0.0905785***	0.028699	0.0972812***	0.032589
(v) Background	0.0636514***	0.021278	0.1110568***	0.026868	0.1253272***	0.031409
*(vi)*Unemployment	0.0994159***	0.022762	0.1236749***	0.028947	0.1449103***	0.03163
(vii) Technology	0.1040746***	0.022737	0.1217785***	0.028979	0.1448869***	0.03165
(viii) First Occupation	0.1048626***	0.022774	0.1211153***	0.029092	0.1438989***	0.031679
Panel *(c)*: Return to Graduation from College						
(i) Raw						
(ii) Mincer	0.3966497***	0.019634	0.4101675***	0.028044	0.3422123***	0.026709
(iii) HLT	0.1913802***	0.028704	0.1227445***	0.046332	0.1363255***	0.038467
(iv) Actual Experience	0.2605771***	0.027685	0.2611534***	0.042887	0.2293378***	0.038639
(v) Background	0.2345345***	0.028093	0.2144272***	0.039499	0.2029146***	0.037601
*(vi)*Unemployment	0.2241602***	0.029391	0.205875***	0.043407	0.2074301***	0.037267
(vii) Technology	0.2266281***	0.029427	0.1961794***	0.043554	0.2012725***	0.037321
(viii) First Occupation	0.2267411***	0.029488	0.1982018***	0.043796	0.2004428***	0.037345

Table 3.A.1.2: Measures of wage returns to experience across specifications

	Full		1910-1949 Cohort		>1950 Cohort	
	(1)	(2)	(3)	(4)	(5)	(6)
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Panel *(a)*: Return to Experience						
(i) Mincer	0.0236914***	0.001917	0.0020973	0.003252	.0317195***	0.005262
(ii) HLT	0.0480968***	0.005105	-.0071585	.0240793	.0176602	0.013973
(iii) Actual Experience	0.0223816***	0.00527	0.0045262	0.00822	-.0113365	0.016096
(iv) Full Model	0.0203314***	0.005013	0.0039033	0.007739	.0101221	0.015246
Panel *(b)*: Less than High School Education						
(i) Mincer	.0155998***	.0043236	-.0026247	.0073281	.0049005	.0144681
(ii) HLT	.0276431**	.0128095	-0.007372	0.009796	-.0995301**	.0465019
(iii) Actual Experience	.0114758	.0103171	.0043542	.0139548	-.0667529*	.0404196
(iv) Full Model	0.0112908**	0.004887	.0084435	.0131875	-.0519558	.0377373
Panel *(c)*: At Least High School Education						
(i) Mincer	.0268559***	.0026325	.0069001	.0048961	.0265451***	.0074198
(ii) HLT	.0201425	.0127516	-.015214	.0220819	-.0190043	.0325248
(iii) Actual Experience	.0237464*	.0135586	.0217766	.0196978	.0020774	.0327774
(iv) Full Model	.0213808	.013057	.0139686	.0181529	.0148288	.0304831
Panel *(d)*: At Least College Education						
(i) Mincer	.0339579***	.0057912	.0016443	.0086046	.0495304***	.0156705
(ii) HLT	-.1135526*	.0632343	-.1317578	.0891181	-.2480299	.1609243
(iii) Actual Experience	-.1038373	.0688605	-.162269*	.0896109	.126733	.1931916
(iv) Full Model	-.0848464	.0734265	-.1363959	.1022117	.0458438	.1840067

Table 3.A.1.3: Measures of wage returns to social background across specifications

	Full		1910-1949 Cohort		>1950 Cohort	
	(1)	(2)	(3)	(4)	(5)	(6)
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Panel *(a)*: Without Controlling for Local Labor Markets, Technology, and First Occupation						
Father's Occupation (Ref: Blue Collar)						
White Collar	0.0182425	0.019704	-.042295	.0268339	.0799594***	.0281965
Father's Education (Ref: Less than High School)						
>= High School	.0589942***	.0189829	.1056241***	.0255545	-.0170668	.0278435
Mother's Education (Ref: Less than High School)						
>= High School	.0590774***	.0152976	.0675421***	.0210579	.0417574*	.0215237
Socioeconomic Group (Ref: Poor)						
Middle Class	.0107751	.0154467	-.0123827	.0207369	.0367523	.0227169
Rich	.0173932	.0206817	.0211805	.0286813	-.0004897	.0288429
Race (Ref: White)						
Black	-.093143***	.0160656	-.074370***	.0215583	-.101842***	.0235479
Hispanic	-.198613***	.020034	-.141725***	.0330959	-.187059***	.0266807
Gender (Ref: Male)						
Female	-0.40218***	0.0135605	-.476813***	.0183817	-.296868***	.0194794
Panel *(b)* Full Model: Controlling for Local Labor Markets, Technology, and First Occupation						
Father's Occupation (Ref: Blue Collar)						
White Collar	.038402*	.0201943	.0120119	.0286988	.0532629*	.0280842
Father's Education (Ref: Less than High School)						
>= High School	.0280149	.0194195	.0566778**	.0275488	-.0091408	.0271213

Mother's Education (Ref: Less than High School)						
>= High School	.0690861***	.0159306	.0790401***	.0230339	.0544331**	.0217411
Socioeconomic Group (Ref: Poor)						
Middle Class	.0161554	.0161293	-.0008971	.0226842	.02332	.0226946
Rich	.0224218	.0210532	.0259053	.0309441	.0034342	.0282638
Race (Ref: White)						
Black	-.117975***	.0174996	-.118239***	.0248102	-.109851***	.0243516
Hispanic	-.226151***	.021074	-.208432***	.0343972	-.208941***	.0270891
Gender (Ref: Male)						
Female	-.399168***	.0139212	-.490329***	.0197598	-.304565***	.0193803

Table 3.A.1.4: Mean Income and Contribution of Differential Elements

Mean prediction high (H):	10.881
Mean prediction low (L):	10.73
Raw differential (R) {H-L}:	0.151
- due to endowments (E):	-0.096
- due to coefficients (C):	0.146
- due to interaction (CE):	0.101

Table 3.A.1.5: Contribution of Explained and Unexplained Portion across Different Methods

D:	0	1	0.5	0.506	*
Unexplained (U){C+(1-D)CE}:	0.247	0.146	0.197	0.196	0.075
Explained (V) {E+D*CE}:	-0.096	0.005	-0.046	-0.045	0.075
% unexplained {U/R}:	163.7	97	130.3	129.9	50
% explained (V/R):	-63.7	3	-30.3	-29.9	50

Table 3.A.1.6: Factor Contribution to the Explained Portion

	explained: D =						
	E(D=0)	C	CE	1	0.5	0.506	*
Education	-0.043	0.015	-0.001	-0.044	-0.044	-0.044	-0.041
Experience	0.108	-0.053	-0.074	0.034	0.071	0.071	0.224
Father's Occupation	0	-0.028	0	0	0	0	0
Father's Education	0	0.079	0.002	0.002	0.001	0.001	0.001
Mother's Education	-0.002	0.032	-0.001	-0.003	-0.002	-0.002	-0.002
Economic Status	0	0.004	0	0	0	0	0
Race	0.027	-0.003	0	0.027	0.027	0.027	0.029
Sex	0.003	-0.265	0.002	0.005	0.004	0.004	0.004
Unemployment	0.001	-0.023	0	0.001	0.001	0.001	0.001
Technology	-0.012	0.045	-0.005	-0.017	-0.014	-0.014	-0.009
First Occupation	0.001	-0.016	-0.001	0.001	0.001	0.001	0.001
Constant	0	0.381	0	0	0	0	0
	-0.096	0.146	0.101	0.005	-0.046	-0.045	0.075

Table 3.A.1.7: Estimates, Means and Predictions of the Factors

	PSID10			PSID50			Pooled
	Coef.	Mean	Pred.	Coef.	Mean	Pred.	Coef.
Education	0.058	12.038	0.703	0.057	12.798	0.732	0.054
Experience	0.003	17.958	0.058	0.01	7.448	0.077	0.021
Father's Occupation	0.007	1.351	0.009	0.028	1.343	0.037	0.019
Father's Education	0.058	1.211	0.071	-0.009	1.175	-0.01	0.028
Mother's Education	0.079	1.368	0.108	0.056	1.4	0.079	0.069
Economic Status	0.005	2.207	0.012	0.003	2.158	0.007	0.006
Race	-0.109	1.536	-0.167	-0.107	1.786	-0.191	-0.115
Sex	-0.492	1.411	-0.694	-0.305	1.421	-0.434	-0.4
Unemployment	0.006	7.03	0.039	0.009	6.94	0.061	0.007
Technology	0	5.10E+04	0.146	0	5.70E+04	0.118	0
First Occupation	0.009	1.549	0.014	0.02	1.477	0.029	0.018
Constant	10.58	1	10.58	10.199	1	10.199	10.383
Total			10.879			10.704	