5-2018

Three Essays on Human Capital and the Distribution of Income

Zhiqi Zhao
Clemson University, zhiqiz@clemson.edu

Three Essays on Human Capital and the Distribution of Income

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Economics

---

by
Zhiqi Zhao
May 2018

Accepted by:
Dr. William R. Dougan, Committee Chair
Dr. Howard Bodenhorn
Dr. Robert K. Fleck
Dr. Chungsang Lam

Abstract

This dissertation exams topics on the human capital and the distribution of income.

The first chapter investigates the insurance value of progressive taxation with heterogeneous risk aversion. The Investment in human capital is lower when the returns to it are subject to uninsurable risk. Progressive income taxation offers a degree of insurance against such risk. Offsetting this effect are the two well-known distortions imposed by progressive taxation: lower expected net-of-tax returns to human-capital acquisition and distortion of the labor-supply decision. The net efficiency effect of progressive income taxation is therefore ambiguous, but there is a presumption that some degree of progressivity can be welfare-improving for risk-averse individuals. To derive the degree of progressivity that may be desirable on efficiency grounds, I construct a general-equilibrium model of an economy with two sectors, calibrated to approximate the U.S. labor market, that differ in terms of the productivity of human capital and the variability of lifetime earnings. Individuals, who differ only in terms of their risk aversion, sort themselves into the two sectors. The simple version of this model, which ignores the labor-leisure choice, suggests that a relatively high degree of income-tax progressivity maximizes aggregate welfare as measured by workers willingness to pay for the insurance being provided. When each workers supply of labor is allowed to vary in response to marginal tax rates, the efficient degree of progressivity is similar to that of the U.S. tax code.

The second chapter exams the worker quality and education premium in the United States. The education premium in the U.S. has been increasing since the 1980s, but the rate of increase has slowed. One explanation for the slowdown is the decrease in the quality of workers who attended some college relative to the quality of workers who obtained a high school diploma. This paper develops a measure of worker quality to estimate the impact of college and high school graduates quality on wages and the education premium. The measure of worker quality uses a weighted average of an occupational skill index. I link occupational skill to the measurement of quality because the

variance of college wages is increasing over time and is directly related to occupational choice. I find that a 1 percent increase in the quality of both high school and college graduates results in a 0.36 percent increase in the education premium, which is an economically meaningful change. One reason for the decline in the quality of college graduates comes from the increasing college enrollment over time. I find that a 1 percent increase in college enrollment leads to a decline in the quality of college graduates by 0.11 percent, and has nearly no effect on the quality of high school graduates.

The third chapter explores the top-coded earnings. The wage and salary earnings, when above some particular value, are censored in the ACS, March CPS and Decennial Census. Others have addressed this issue of top-coding and also used non-public data to develop various alternative multipliers. We improve on this research, which focuses on producing multipliers for the whole population, by developing multipliers that are demographic and regional specific. It is not unreasonable to expect that the earnings distribution varies across racial, gender, and education dimensions, and, even, across geographies. Comparison of the results from our methodology will be made to those proposed by others, with special focus on investigating differences in the education premium over time.

Dedication

To my parents, Jingang Zhao and Chengmei Xia, for their unconditional support.

Acknowledgments

This dissertation would not have been possible without help from my committee members. I am very grateful for my advisor, Dr. William Dougan, who constantly challenged me on many aspects of my research ideas. He spent a tremendous amount of time providing feedback on the first chapter of my dissertation. Without his scope of knowledge and scholarly guidance, it would likely never have come to fruition. I would also like to thank Dr. Chungsang Lam for his continued support throughout my graduate studies as a mentor and a good friend. I am also thankful for the feedback regarding my research and its writing provided by Dr. Howard Bodenhorn and Dr. Robert Fleck.

I further appreciate all of the helpful feedback I received from participants of the Clemson Public Workshop and Labor Workshop. I would like to thank Casey Rothschild, Mike Hoy, and Stephen Shore for valuable comments at the Huebner Doctoral Colloquium. All errors within this dissertation are my own.

Finally, I would like to thank my family and friends for their incredible support throughout my graduate studies. I am most grateful to my parents, Mr. Jingang Zhao and Mrs. Chengmei Mei, for their unwavering support in my pursuits and for all of the sacrifices they have made to help me accomplish them.

Table of Contents

List of Tables

List of Figures

Chapter 1

The Insurance Value of Progressive Taxation with Heterogeneous Risk Aversion

1.1 Introduction

Sectors differ in the mean and variance of earnings. Because risk-averse workers dislike high earnings variance, a high mean earning must compensate for the high variance to attain a given value of expected utility. Workers will sort into sectors based on their preferences over the mean and variance of earnings such that less risk-averse workers will accept a higher variance of earnings in return for a higher mean earning. Progressivity in the average tax rate reduces the expected net-of-tax returns to human-capital acquisition, but it reduces the variance of lifetime earnings, thus providing insurance for risk-averse workers. An extension of the core model allows progressivity in the marginal tax rate to distort the labor supply decision. Once the labor-leisure distortion is taken into account, the efficient degree of progressivity is lower than that implied by the model without a distortion.

My core model considers two distinct channels through which progressivity affects the average tax rate. First, through the mean-consumption channel, progressive taxation reduces the expected net-of-tax returns to human-capital acquisition. Workers must accumulate human capital

in order to gain access to higher average earnings in some sectors. The cost of obtaining human capital is certain in my model, but it does not generate refundable tax credits. Therefore, progressive income taxation reduces the expected net-of-tax returns to human-capital acquisition and consequently, mean consumption. Second, through the variance-of-consumption channel, a progressive tax offers insurance to workers in the sectors with a high volatility of lifetime earnings. Assume there is a distribution of lifetime earnings. When realized lifetime earnings are high (low), workers pay a higher (lower) tax rate; progressive income taxation shrinks the variance of consumption and provides insurance to risk-averse workers.[1] This insurance effect cannot be provided by the firm, which cannot create a wage contract to insure against the variation in lifetime earnings. In addition, individual savings cannot insure against this risk due to the uncertainty of lifetime earnings. Therefore, tax policy is an approach the government can use to fill the missing insurance market. These two effects jointly determine whether there is an efficient degree of progressivity in the tax rates, in the aggregate economy, with workers that are heterogeneous in risk aversion.

Progressive income taxation also distorts the labor supply. A higher marginal income tax rate will give workers less incentive to work. In an extension of my core model, I introduce the labor supply decision, which lowers the efficient degree of progressivity. After calibrating my second model using Panel Study of Income Dynamics (PSID) data for the year 2000, I find that the efficient marginal tax rate in the low income bracket is 21.7%, and the efficient marginal tax rate in the high income bracket is 36.5%.

In this paper, the approach to welfare analysis follows Harberger (1971), who takes the individual willingness to pay as the measure. The idea is to compensate each worker with a certain amount of consumption in order to make the worker indifferent to a flat tax regime or a progressive tax regime (i.e., the expected utility of the worker in the flat tax regime and the progressive tax regime is the same). Since workers are heterogeneous in risk aversion, their compensation amounts will differ. Some workers will gain and some workers will lose under a progressive tax regime, compared to a flat tax regime. Therefore, each worker with a different level of risk aversion has an individual efficient degree of progressivity. In order to capture the aggregate level of welfare, I sum each worker's gain or loss, which follows the criteria of Kaldor-Hicks efficiency. Hence, the efficient

[1]For example, suppose a person wants to be a lawyer. He doesn't know whether he will be a successful lawyer or an unsuccessful lawyer. Before making the decision to become a lawyer, he understands the distribution of lifetime earnings. But he doesn't know whether his draw will be from high lifetime earnings or low lifetime earnings. The level of lifetime earnings will be revealed after he becomes a lawyer.

degree of progressivity is found in the aggregate level.

This paper considers three effects in unison: the insurance effect, the reduction of expected net-of-tax returns to human-capital acquisition, and the labor supply distortion under uncertainty. This project also links heterogeneity in risk aversion among workers with income taxation topics.

The literature on how progressivity in the marginal rate distorts the number of hours worked is directly related to optimal income taxation (Mirrlees 1971; Saez 2001; Sachs et al. 2016). These models consider a world without uncertainty. Optimal marginal tax rates are interpreted in terms of redistribution, insurance, and incentive effects (Low and Maldoomb 2004; Boadway and Sato 2011; Heathcote et al. 2017). They consider the optimal income taxation under uncertainty, but do not consider sectoral choices. Brown and Rosen (1987) discuss how the market allows individuals to substitute the mean level of the wage for its variability across occupations and further predict how lowering the rate of taxation on earnings would impact an individual occupational choice, though optimal income taxation or the efficient degree of progressivity was not discussed in their paper.

Further tax literature explores how progressivity in the average rate reduces the expected net-of-tax returns to human-capital acquisition (Eaton and Rosen 1980; Guvenen et al. 2014). Stantcheva (2017) derives the optimal taxation and human capital policies in a life cycle model with risky human capital. These papers do not address how progressive taxation influences sectoral choice on an individual level. If an individual wants to sort into a high-skill sector, he needs to accumulate a high level of human capital. A progressive tax may make him less likely to choose a sector offering a high return, which therefore reduces the net-of-tax return to human-capital acquisition.

Though progressive taxation reduces the expected net-of-tax returns to human-capital acquisition and distorts the labor supply decision, it also provides insurance against risk of lifetime earnings, which reduces the variance of consumption (Varian 1980; Kniesner and Ziliak 2002). Heathcote et al. (2017) discusses how a progressive tax system can substitute for imperfect private insurance against idiosyncratic earnings risk, though it does not consider sectoral choices and heterogeneity in risk aversion.

One of the assumptions in my model is that workers are homogeneous in ability but heterogeneous in risk aversion when sorting into two sectors with different levels of lifetime earnings uncertainty. Rothschild and Scheuer (2013) consider optimal progressive taxation in a model where individuals can self-select into one of several possible sectors based on heterogeneity in a multidimensional skill vector. Cubas and Silos (2015) discuss progressive taxation and risky career choices.

These general equilibrium models include heterogeneous abilities but do not consider heterogeneous risk aversion.

It is well documented in the labor literature that heterogeneous risk aversion is an important but unobserved factor that influences career choice. Guiso and Paiella (2004) use household survey data to construct a direct measure of absolute risk aversion based on the maximum price a consumer is willing to pay to buy a risky asset. They find that risk-averse consumers are less likely than the risk-prone to be self-employed and to be entrepreneurs and they are more likely to work in the public sector after controlling for the level of income, wealth, personal characteristics, educational attainment, and other attributes. When analyzing sectoral choice, considering heterogeneity in risk aversion is important because of the self-selection problem. If risk aversion is unobservable, estimates of the effect of labor income risk on sectoral choice will be inconsistent because the measure of income risk is correlated with the error term that contains an unobserved preference parameter. Moreover, Hagedorn et al. (2017) argue that observable worker and firm characteristics account for only 30% of the observed variation in wages. In recent studies, career choice and heterogeneity in risk aversion are jointly considered (Cozzi 2011; Barth et al. 2017). Lockwood et al. (2015) discuss heterogeneity in preferences and optimal redistribution in an optimal tax model. Gartner et al. (2017) argue the individual risk preferences and the demand for redistribution.

This paper explores the effect of progressive taxation on the equilibrium allocation of heterogeneously risk averse workers across two sectors. Furthermore, it presents an analysis of relative welfare change with the implementation of a progressive tax, as compared to a flax tax, under uncertainty. In the core model, two effects will be jointly considered in the welfare analysis of progressivity in the average rate: the insurance effect through the channel of reduction in the variance of consumption, and the reduction in expected net-of-tax returns to human-capital acquisition through the channel of mean consumption. All agents in the core model are full-time, full-year workers in the sector of their choice.[2] In the extension of the core model, I include the distortion of the labor supply decision caused by progressivity in the marginal income tax rate. After adding the labor-leisure choice, and calibrating the model using PSID data, I find the efficient degree of progressivity, which is less progressive than in the core model.

The structure of the paper is as follows: In section 1.2, I provide empirical evidence that individuals who are less risk averse will choose sectors with higher unobserved variance of wages.

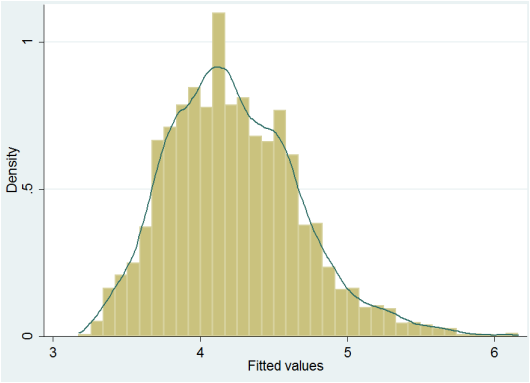
[2]Full-time, full-year workers are defined as working 35-plus hours per week and 40-plus weeks per year.

An income taxation model with a fixed labor supply will be shown in section 1.3. In section 1.4, a model with a variable labor supply will be analyzed. Section 1.5 concludes the paper.

1.2 Empirical Evidence of Heterogeneous Risk Aversion

One of the key features in my model is that workers are heterogeneous in risk aversion. In the 1996 PSID, the questionnaire includes questions related to hypothetical job choices. Based on the individual answers to the questions, Kimball, Sahm, and Shapiro (2009) estimate the average coefficient of relative risk aversion in each of six categories. Following their work, I further impute the individual coefficient of relative risk aversion. The detailed imputation method is given in Appendix A. Figure 1.1 shows that workers are indeed heterogeneous in risk aversion. The model starts from

Figure 1.1: The Distribution of Risk Aversion

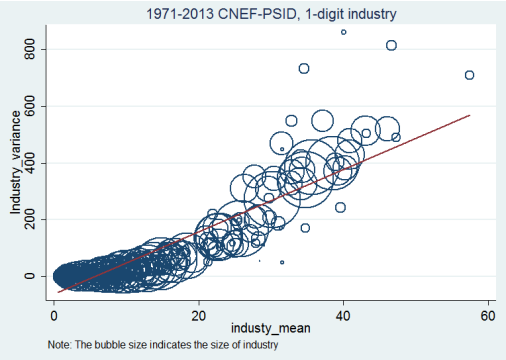


Note: The x-axis indicates the coefficient of relative risk aversion. A higher number means a worker is more risk averse.

the premise that individuals who are less risk averse will sort into sectors with higher unobserved variance of wages. In addition, I assume there is a positive correlation between average wage and unobserved variance of wages across sectors. These are reasonable realistic assumptions based on the empirical evidence. The sample I choose to demonstrate the empirical evidence is full-time workers between 20 and 60 years old in the period 1971-2013. The real wage is equal to annual labor earnings divided by annual work hours, adjusted by the 1999 Consumer Price Index. Following Bonin et al. (2007), I discard observations of full-time employed workers whose wage information is extremely implausible, thus dropping observations of those in the top 1-percentile and bottom 1-percentile of the wage distribution. There are nine categories of industry. I calculate the mean and

the unobserved variance of wages for each industry each year. The measure of unobserved variance is similar to that of Bonin et al. (2008) and Fouargea et al. (2014). The unobserved variance is obtained by the regression of the Mincer equation. I regress hourly mean wage for each industry each year on their education, gender, and experience, and include both a quadratic and cubic term for experience as well as industry fixed effects. The error is clustered at the industry level. Then I calculate the variance of the residual for each industry each year as the unobserved variance of wages. Figure 1.2 shows that there is a positive relationship between average wage and unobserved variance of wages across sectors each year. Figure 1.3 displays that, on average, workers who are more risk averse will sort into industries that have a smaller unobserved variance of wages. The bubble indicates the size of the industry.

Figure 1.2: The Relationship between Average Wage and Unobserved Variance of Wages Across Industries



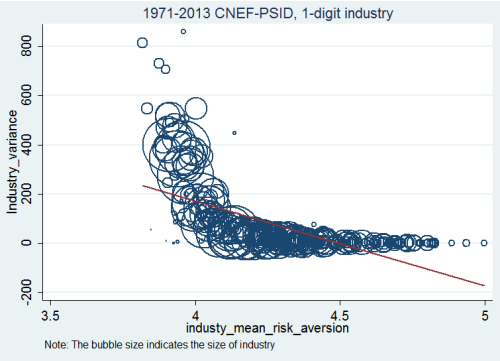
Note: The x-axis measures the mean real wage of workers in each industry each year. The y-axis measures the unobserved variance of wages in each industry each year by using residuals from the Mincer equations. The bubble indicates the size of the industry. The figure shows that a higher average wage compensates for a higher unobserved variance of wages in each industry each year.

1.3 Income Taxation and Welfare with Fixed Labor Supply

1.3.1 Setup

Suppose workers exhibit homogeneity in ability but heterogeneity in risk aversion. The labor market is divided into two sectors, each of which produces a homogeneous goods. The individual wage is endogenous in the general equilibrium framework, which is determined by the marginal product in the sector and by individual-specific labor productivity shocks. Exogenous policy changes (e.g., income taxation reform) can influence the allocation of workers between the two sectors and change

Figure 1.3: The Relationship between Average Risk Aversion of Workers and Unobserved Variance of Wages Across Industries



Note: The x-axis measures the average coefficient of relative risk aversion of full-time workers in each industry each year. A higher number means workers, on average, are more risk averse. The y-axis measures the unobserved variance of wages in each industry each year by using residuals from the Mincer equations.The bubble indicates the size of the industry. The figure shows that, on average, workers who are more risk averse will sort into industries that have a smaller unobserved variance of wages each year.

the equilibrium wage in the labor market.

The production functions in the aggregate level are:

$$\begin{aligned} Y_1 = \alpha_1 L_1^{\beta_1} = \alpha_1 \Big[n(1-h(\theta_m)) \int_0^{+\infty} \eta_{i1} dF(\eta_{i1}) \Big]^{\beta_1}, \\ Y_2 = \alpha_2 L_2^{\beta_2} = \alpha_2 \Big[nh(\theta_m) \int_0^{+\infty} \eta_{i2} dF(\eta_{i2}) \Big]^{\beta_2}, \end{aligned} \tag{1.1}$$

where Y_1 and Y_2 are total output; L_1 and L_2 are total effective labor in sector 1 and sector 2. The total factor productivities are α_1 and α_2 in sector 1 and sector 2 and $\alpha_1 < \alpha_2$. The labor shares are β_1 and β_2. The proportion of labor working in sector 2 is $h(\theta_m)$, and θ_m is the coefficient of risk aversion for the marginal worker. The total population is n. I assume workers who are less risk averse (i.e., $\theta < \theta_m$) will choose to work in sector 2, while those who are more risk averse (i.e., $\theta > \theta_m$) will choose to work in sector 1. Therefore, the marginal worker's coefficient of risk aversion, θ_m, is endogenous in my model. It can be solved numerically and also depends on the tax regimes.

I assume that individual i's productivity shock in sector j follows a log-normal distribution $\eta_{ij} \sim lnN(\frac{-\sigma_j^2}{2}, \sigma_j^2)$, with $\sigma_2^2 > \sigma_1^2$, $E(\eta_{ij}) = 1$ and $Var(\eta_{i2}) > Var(\eta_{i1})$. Therefore, the mean wage for workers in sector 2 is higher than the mean wage for workers in sector 1 in order to compensate for the higher variance in sector 2. Each worker is paid a wage equal to his productivity of labor.

The timing of the model is as follows. First, workers must choose whether or not to obtain human capital before sectoral choices. In order to get into sector 2, which requires high human

capital, workers need to pay the cost. Workers also know their risk aversion. Second, they choose their sector. Third, after that decision has been made, they learn their productivity shock and hence their wage. Under no tax scenario, income is equal to consumption. Appendix B shows that the consumption for an individual who chooses to work in sector 1 or sector 2 is [3]

$$c_{i1} = \eta_{i1}\alpha_1\beta_1(n(1-h(\theta_m)))^{\beta_1-1}, \tag{1.2}$$

$$c_{i2} = \eta_{i2}\alpha_2\beta_2(nh(\theta_m))^{\beta_2-1} - \gamma, \tag{1.3}$$

where γ is the cost of accumulating human capital (i.e., the cost of higher productivity α_2) to work in sector 2.

On the worker side, the utility function exhibits constant relative risk aversion. Thus, the maximization problem is in the context of Von Neumann-Morgenstern expected utility. The utility function is given by

$$U(c) = \frac{c^{1-\theta}-1}{1-\theta}, \tag{1.4}$$

where c is consumption. The constant coefficient of relative risk aversion is $\theta \in [\underline{\theta}, \overline{\theta}]$, which is defined as $\theta = c\frac{-U''}{U'}$. A large value of θ implies that the worker is relatively more risk averse. I assume that θ follows a continuous distribution in the population, which will be calibrated from the data.

1.3.2 Equilibrium Conditions

Since workers care about after-tax earnings, that is, their consumption level, I examine the progressiveness of the average tax rate as a response to the progressiveness of the marginal tax rate in the tax function. Following Guner et al. (2011,2012) and Guner et al. (2014), I use the *log* tax

[3] This is a static model, which represents the steady state. Thus, there are no savings in this model. The after-tax wage is equal to consumption.

functional form[4]. The average tax rate and corresponding consumption is

$$\begin{aligned} t(w) &= a + \rho \cdot log(w), \\ c(w) &= (1-a) \cdot w - \rho \cdot w \cdot log(w) \end{aligned} \tag{1.5}$$

where c is the consumption and w is the wage. In this functional form, a is the scale parameter and ρ captures the curvature of the tax function. If ρ is equal to 0, then it is equivalent to the flat taxation case.

Therefore, the government has access to the following instruments: a progressive labor income tax with two free parameters that it can choose, and a proportional corporate income tax on profits which will be discussed later..

Based on the above functional form, the consumption for worker i in sector 1 or 2 is

$$\begin{aligned} c_{i1} &= (1-a)\big[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,P})))^{\beta_1-1}\big] - \rho\big[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,P})))^{\beta_1-1}\big]log\big[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,P})))^{\beta_1-1}\big], \\ c_{i2} &= (1-a)\big[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,P}))^{\beta_2-1}\big] - \rho\big[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,P}))^{\beta_2-1}\big]log\big[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,P}))^{\beta_2-1}\big] - \gamma, \end{aligned} \tag{1.6}$$

where $h(\theta_{m,P})$ is the proportion of workers choosing sector 2 and P indicates progressive tax regime.

The equilibrium condition in this model (for the marginal worker m) is $E[U_m(c_{m2})] = E[U_m(c_{m1})]$, which is identical to

$$\begin{aligned} &\int_0^{+\infty} \frac{\Big\{(1-a)w_{m1} - \rho w_{m1} log(w_{m1})\Big\}^{1-\theta_{m,P}} - 1}{1-\theta_{m,P}} f(\eta_{m1}) d\eta_{m1} \\ &= \int_0^{+\infty} \frac{\Big\{(1-a)w_{m2} - \rho w_{m2} log(w_{m2}) - \gamma\Big\}^{1-\theta_{m,P}} - 1}{1-\theta_{m,P}} f(\eta_{m2}) d\eta_{m2}, \end{aligned} \tag{1.7}$$

where $\eta_{mj} \sim lnN(\frac{-\sigma_j^2}{2}, \sigma_j^2)$. The above equation is nonlinear, so the solution $\theta_{m,P}$ will be given in the simulation section under certain parameter values.

[4]I have tried two other functional forms. The first corresponds to the function used in Heathcote et al. (2016): $c(w) = aw^{\rho}$. In order to keep tax revenue constant, when I reduce the value of ρ to make the tax system more progressive, the value of a needs to be increased. Then I find that the marginal tax rate, $1 - a\rho w^{\rho-1}$, becomes negative if a is large enough. Thus, this functional form does not fit my model well. The second functional form corresponds to the function used in Guvenen et al. (2013). The function form of total tax paid is $T = aw^{\rho}$. When I increase the value of ρ to make the tax system more progressive, the value of a needs to decrease to hold the tax revenue constant. However, when the values of ρ and a are simultaneously changed, the degree of progressiveness only changes in a very small amount. That is, there is an upper bound for the degree of progressiveness when the values of ρ and a are changed simultaneously to make the tax system more progressive and, at the same time keep the tax revenue constant. Hence, the second tax function form is not ideal.

In the flat tax regime F, the value of ρ is equal to zero. The value of a is equivalent to the flat tax rate τ. Based on the same logic, I find the marginal worker $\theta_{m,F}$, who is indifferent to choosing sector 1 or sector 2. The equilibrium allocation in sector 2 is equal to $h(\theta_{m,F})$.

1.3.3 Welfare Analysis

There are two levels of welfare analysis in this paper. The first is the individual level. Because individuals are heterogeneous in risk aversion, it is important to explore the difference in welfare change in the two tax regimes with respect to the degree of risk aversion. The second is the aggregate level. It can tell us whether there exists an efficient degree of progressivity based on the insurance effect through the variance-of-consumption channel, and on the reduction of expected net-of-tax returns to human-capital acquisition through the mean-consumption channel.

Since workers are heterogeneous in risk aversion, adding their utility together to measure the aggregate welfare is not appropriate. I follow Harberger (1971), who uses the individual's willingness to pay to measure welfare, and mitigates the aggregation problem. The detailed methodology follows Lucas (1987). The idea is to compensate with a certain amount of consumption to make worker i indifferent to either a progressive tax regime (P) or a flat tax regime (F). Individual i might remain in the same sector, or he might switch sectors between the two tax regimes. But the compensating consumption itself will not affect the worker's decision about which sector to enter. It's a hypothetical compensation after the endogenous sectoral choice is made.

The formulas for the welfare change for individual i if he does not change sector to work (i.e., both in sector j), or if he switches from sector j to sector k to work are given by

$$\begin{aligned} E[U_{ij,P}(c_{ij})] &= E[U_{ij,F}(c_{ij} + \Delta c_i)], \quad or \\ E[U_{ik,P}(c_{ik})] &= E[U_{ij,F}(c_{ij} + \Delta c_i)]. \end{aligned} \tag{1.8}$$

If there is a welfare gain for worker i in the progressive tax regime, then $\Delta c_i > 0$. If there is a welfare loss for worker i in the progressive tax regime, then $\Delta c_i < 0$. Workers with different levels of risk aversion, θ_i, will be compensated with different levels of consumption Δc_i. Therefore, after compensating by the amount of consumption Δc_i, the welfare level of worker i will be exactly the same in the progressive tax regime and the flat tax regime.

In the aggregate level, following the criterion of Kaldor-Hicks efficiency, the total welfare

change for all workers involves aggregating the compensating level of consumption and checking how it changes when the tax system is more progressive:

$$\Delta c = \sum_i \Delta c_i. \tag{1.9}$$

In order to fully consider the change in welfare in the economy under the general equilibrium framework, I attribute the residual of the output to the capital owners. They are risk neutral and pay the flat corporate income tax. The change in welfare for the capital owners is equal to the difference of the residual output between a progressive tax regime and a flat tax regime after imposing the constant flat corporate income tax rate τ_c.

$$\Delta\Pi_1 = (1-\tau_c)\alpha_1(n(1-h(\theta_{m,P})))^{\beta_1}[1-\beta_1] - (1-\tau_c)\alpha_1(n(1-h(\theta_{m,F})))^{\beta_1}[1-\beta_1], \tag{1.10}$$

$$\Delta\Pi_2 = (1-\tau_c)\alpha_2(n(h(\theta_{m,P})))^{\beta_2}[1-\beta_2] - (1-\tau_c)\alpha_2(n(h(\theta_{m,F})))^{\beta_2}[1-\beta_2]. \tag{1.11}$$

The total welfare change for capital owners is

$$\Delta\Pi = \Delta\Pi_1 + \Delta\Pi_2. \tag{1.12}$$

The calculation of the welfare change in the aggregate economy is equal to the total welfare change for all workers plus the total welfare change for capital owners. This measurement is based on money instead of utility, and it varies with the degree of progressivity in income taxation.

$$\Delta TW = \Delta c + \Delta\Pi. \tag{1.13}$$

Welfare comparisons are based on the equal tax revenue condition. That is, the expected tax revenue collected from all workers in the flat tax regime and in the progressive tax regime is equal, and the degree of progressivity does not affect expected tax revenue. The tax revenue can be used as government spending, but it will not impact wages and it doesn't impact different individuals differently.[5] Therefore, under the parametric assumption of the tax function, when ρ is changed to

[5]As Eaton and Rosen (1980) argue, "the assumption that the government is only concerned with the expected revenue need not imply that the government is risk neutral. If the shock associated with wage is independent across individuals and if the number of individuals is large enough,then the law of large numbers will guarantee the government a constant total revenue despite uncertainty on the individual level. The government is, in this case, simply a more efficient risk pooler than the individual."

make the tax system more progressive, the value of a needs to be changed too in the tax function to make the expected tax revenue stay the same. The relationship between a and ρ, given the constant tax revenue, is described in Appendix C.

1.3.4 Calibration

In the data set, I focus on working-age individuals, aged 20-60 in the year 2000, whose population weight is nonzero. Each observation in the sample is weighted by its PSID supplied sample weight. I drop those individuals who are not full-time workers, because the core model assumes there is no labor-leisure choice (i.e., no adjustment of hours of work).[6] Wage is equal to the individual's labor earnings divided by the work hours of the individual. I further drop observations with reported hourly wages below the federal minimum wage rate of \$5.15/hour or above \$100/hour. The remaining size of the sample is 2,353 individuals.

Since η_{i1} and η_{i2} follow a lognormal distribution, there is some probability that the wage might be zero, Therefore, I use a shifted lognormal distribution instead to calibrate the parameters in the distribution.

I first calibrate h_m, which is the proportion of workers who attend at least some college. For the labor share β_1 and β_2, I impose values of 2/3.[7] For the cost of human capital γ, it is approximately 20% of the mean hourly wage for workers who attend at least some college. Given h^*, β_1, β_2, and γ, productivity parameters α_1 and α_2 are calibrated to match the target data moments: the pre-tax mean hourly wage in sector 1 and that in sector 2. The pre-tax mean hourly wage in sector 1 is equal to the pre-tax mean hourly wage for all workers who had obtained at most a high school diploma in the year 2000: \$16.41. The pre-tax mean hourly wage in sector 2 is equal to the pre-tax mean wage for all workers who had attended at least some college in the year 2000: \$24.84. I search for the parameters α_1 and α_2 that minimize the squared deviations between the model and data moments. Table 1 shows each parameter in the model.

The remaining unknown variables are σ_1^2 and σ_2^2. Although they are unobserved, they can be solved for in the equilibrium with no tax case given the other calibrated parameters. Suppose $\sigma_2^2 = 3\sigma_1^2$; [8]then I am searching σ_1^2 to solve equation (1.7) if $a = 0$ and $\rho = 0$. Given the following

[6]In the Cross-national Equivalent File version of PSID, if the individual had positive wages and worked at least 1,820 hours last year (35 hours per week on average), then the individual was employed full-time.

[7]I am not aware of any paper that offers a reasonable but different β in different industries in the U.S. Thus, I adopt the labor share in the aggregate economy in the U.S, which is 2/3.

[8]The linear relationship between σ_2^2 and σ_1^2 is assumed in order to have one unknown parameter to be solved in

Table 1.1: Calibrated Parameters In Core Model

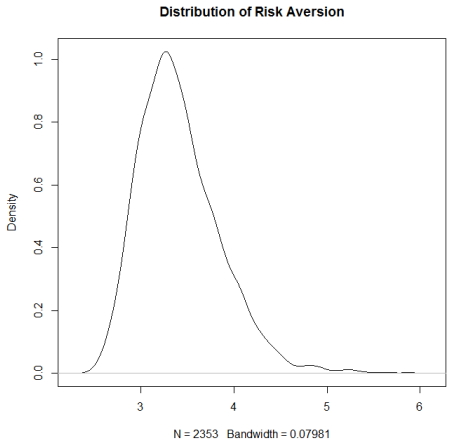
Parameters	Value
h^*	0.521
β_1	0.67
β_2	0.67
α_1	85.4
α_2	100.0
γ	4.97

equation, the ratio of $Var(w_2)$ and $Var(w_1)$ is approximately 2.66. Once σ_2^2 and σ_1^2 are calibrated, they are considered exogenous and will not be changed if the tax regime is changed.

$$\frac{Var(w_2)}{Var(w_1)} = \frac{[(e^{\sigma_2^2}-1)(\alpha_2\beta_2(n(1-h_m))^{\beta_2-1})]^2}{[(e^{\sigma_1^2}-1)(\alpha_1\beta_1(nh_m)^{\beta_1-1})]^2} \tag{1.14}$$

The estimated risk aversion distribution in the sample approximately follows a shifted log-normal distribution. Thus, I draw a three-parameter lognormal distribution (shape=0.30, scale=0.3, threshold=2.0), which approximates the targeted risk aversion distribution for the sample in the year of 2000. It is shown in Figure 1.4.

Figure 1.4: The Distribution of Risk Aversion in the Sample



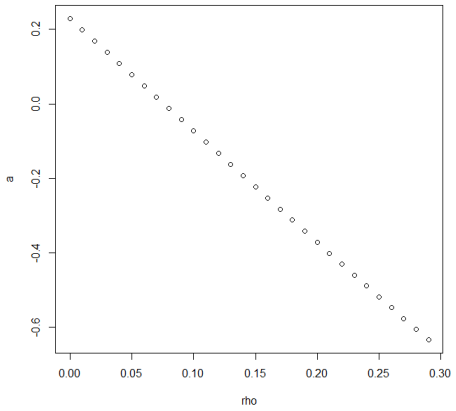
Note: Figure 1.4 shows the distribution of the coefficient of relative risk aversion for full-time workers in the sample in the year of 2000.

one equation. More generally, other linear relationship could also be assumed.

1.3.5 Results

I first simulate the model under the flat tax regime, where $a = 0.23$ and $\rho = 0$ (i.e. $\tau = 0.23$) in the year 2000. Given the values of a, ρ, and the above calibrated parameters shown in Table 1, the equilibrium allocation can be solved using equation (1.7). After the equilibrium allocations, I simulate two individual welfare changes by using equation (1.8), whose coefficient of risk aversion is the minimum and the maximum in the distribution of θ. Then I simulate the change of aggregate welfare in equation (1.13). These simulations will be continuously done as the tax system becomes more progressive, that is, when the value of ρ becomes bigger. All welfare comparisons are done under a constant tax revenue condition, meaning that as ρ becomes bigger a must correspondingly become smaller. Figure 1.5 shows how ρ and a must vary in order to keep tax revenue constant.

Figure 1.5: The Relationship between ρ and a for Constant Tax Revenue

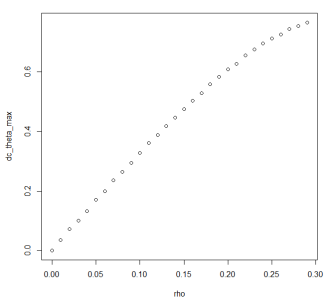


Note: There are two parameters, ρ and a, in the tax functional form I assume. In order to keep the tax revenue constant, when ρ is increased to make the tax system more progressive, a must be adjusted correspondingly. The tax revenue neutral relationship between ρ and a is shown here.

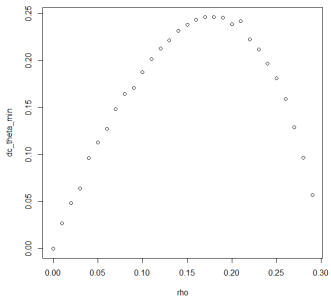
1.3.5.1 Individual Welfare Change with Respect to the Change in ρ

Equation (1.8) provides the formulas to calculate the individual welfare change. Since the worker with the lowest coefficient of risk aversion (i.e., the least risk averse worker) will always stay in sector 2, and the worker with the highest coefficient of risk aversion (i.e., the most risk averse worker) will always stay in sector 1, it is worthwhile to explore their welfare change. Figure 1.6 shows the welfare change for these workers as ρ increases and the tax becomes more progressive. For the least risk averse worker ($\theta = 2.51$), welfare increases initially because of the insurance effect.

Figure 1.6: Individual Welfare Change when Tax is More Progressive



(a) Welfare Change for the Most Risk Averse Worker



(b) Welfare Change for the Least Risk Averse Worker

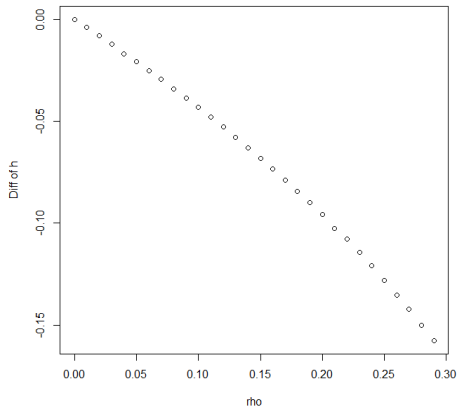
However, at higher levels of progressivity the human capital effect dominates and welfare decreases. For the most risk averse worker ($\theta = 5.88$), a more progressive tax provides more insurance. Unlike workers in sector 2, the most risk averse worker does not need to pay the cost to enter sector 1; thus the mean consumption does not change as much when the tax is more progressive. Based on these facts, welfare strictly increases for the most risk averse worker as the tax becomes more progressive.

1.3.5.2 Aggregate Welfare Change with respect to the Change of ρ

I next consider whether there is a progressive tax rate that maximizes aggregate welfare. Figure 1.7 shows that when the tax is more progressive, some workers move from sector 2 to sector 1. The economic intuition is that although a higher degree of progressiveness provides more insurance to workers in sector 2, the average tax rate in the progressive tax regime in sector 2 is higher than the average tax rate in sector 1, which decreases the mean after-tax wage in sector 2 more. When the cost of the lower mean after-tax wage is greater than the benefit of lower variance in sector 2, workers start to move from sector 2 to sector 1.

Giving the changing allocation of workers across sectors as the tax rate becomes more progressive, it is easy to understand the welfare change for capital owners shown in Figure 1.8. When the tax is more progressive, fewer workers are employed in sector 2. Output is reduced in sector 2, and the capital owner in sector 2 must pay higher wages because the remaining workers have higher marginal productivity. Thus, the capital owner in sector 2 is worse off (Figure 1.8(b)). Similar analysis is applied in the opposite direction for the capital owner in sector 1 (Figure 1.8(a)). Figure 1.8(c) shows the net change of welfare for both capital owners in the economy. The net effect

Figure 1.7: Difference in Proportion of Workers in Sector 2: Progressive vs. Flat Tax Rate



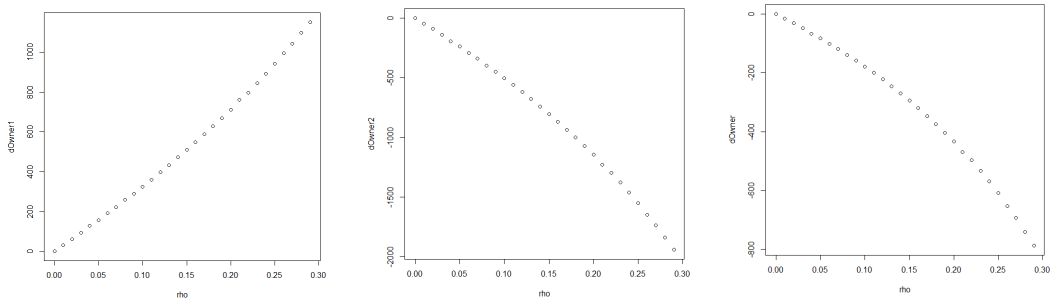
Note: When the tax rate is more progressive (i.e., the parameter in the tax function, ρ, is larger), workers are moving from sector 2 to sector 1.

shows the welfare of capital owners strictly decreases as the tax system becomes more progressive because total output is reduced.

Figure 1.9 presents the hump relationship between the change in welfare for the whole economy and ρ. In the aggregate level, when the income tax rate becomes more progressive, total welfare first increases and then decreases. The aggregate effect is composed of the insurance effect through the variance-of-consumption channel, the reduction of expected net-of-tax returns to human-capital acquisition through the mean-consumption channel, and the reduction of output. At low levels of progressivity, the insurance effect dominates the reduction in expected net-of-tax returns to human-capital acquisition and the output effect, which makes total welfare increase. However, when the level of progressivity is very high, the reduction in expected net-of-tax returns to human-capital acquisition is large, which dominates the insurance effect. More workers choose to work in the low productivity sector (i.e., sector 1) and total output is reduced. Thus, total welfare eventually decreases.

The maximal welfare gain (i.e., dTW) under progressive income tax system, as compared with flat income tax system, is the point where $\rho = 0.200$ and $a = -0.372$. I transfer the values of those two parameters into average tax rates, because the insurance effect and the reduction in return to human capital are based on the progressivity of average tax rates. In addition, the progressive average tax rates lead to an output reduction. The corresponding average tax rates ($ATRs$) are shown below. The marginal tax rate is the average tax rate plus $\rho = 0.200$.

Figure 1.8: Welfare Change for the Capital Owners When the Tax is More Progressive



(a) Welfare Change for Owner in Sector 1

(b) Welfare Change for Owner in Sector 2

(c) Welfare Change for Both Owners

Note: Figure 1.8 shows the welfare change for capital owners when the tax system is more progressive, that is, the parameter in the tax function, ρ, is larger.

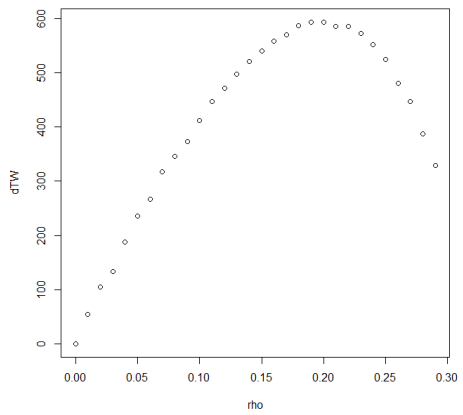
$$ATR_1 = \begin{cases} 8.7\% & w_{1,min} = 9.91 \\ 18.8\% & w_1 = 16.41 \\ 44.5\% & w_{1,max} = 59.49. \end{cases}$$

$$ATR_2 = \begin{cases} 23.9\% & w_{2,min} = 21.21 \\ 27.1\% & w_2 = 24.85 \\ 53.2\% & w_{2,max} = 91.92. \end{cases}$$

1.4 Income Taxation and Welfare With Variable Labor Supply

The additional distortion of progressiveness is the labor supply decision. The following model will deal with the trade-off among the insurance effect, the reduction of expected net-of-tax returns to human-capital acquisition, and the labor supply distortion imposed by progressive income taxation. In order to make the model more tractable, I use a discrete distribution of realized wage

Figure 1.9: Welfare Change for Workers and Owners in the Aggregate Level



Note: Figure 1.9 shows shows the hump relationship between the change in welfare for the whole economy and the degree of progressiveness. When the parameter in the tax function, ρ, is larger, the tax system is more progressive. Thus, there is an efficient degree of progressivity.

and a piecewise linear tax function.[9]

1.4.1 Setup

There are two sectors in the model. In sector 1, wage is exogenously given as w_1; thus no uncertainty about earnings exists in sector 1. In sector 2, there are three possible exogenous wages: w_{2L} with probability P_L, w_{2M} with probability P_M, and w_{2H} with probability P_H. That is, uncertainty about earnings exists in sector 2. According to compensating differential theory, a higher variance of wage needs to be compensated by a higher mean wage. Therefore, I assume $w_{2H} > w_{2M} > w_{2L} = w_1$.

Workers are heterogeneous in risk aversion. Those who are more risk averse will choose sector 1, and those who are less risk averse will choose sector 2. Once workers sort into sectors based on their coefficient of relative risk aversion, they will decide how many hours to work based on the income taxation structure. I assume workers cannot move between sectors when income taxation becomes more progressive or less progressive. Following Eichenbaum et al. (1988) and French (2005), the utility function is given by

[9]With no variation in labor supply in this model, the results are consistent with the results in the core model, which has a continuous distribution of realized earnings and a continuous tax function.

$$U(c,l) = \frac{(c^{\lambda} l^{1-\lambda})^{1-\theta}}{1-\theta} \tag{1.15}$$

where c is consumption, l is leisure, and θ is the coefficient of relative risk aversion.

1.4.2 Leisure Choice

1.4.2.1 Progressive-Tax Case

In the progressive tax regime, there is a piecewise linear tax code with three rates: t_L, t_M, and t_H. Workers in sector 1 will be taxed at t_L. Workers in sector 2 will be taxed at t_L, t_M, or t_H based on their realized earnings.

The consumption and the total tax paid by worker i in sector 1 is

$$c_{i1} = (1-t_L)w_1(H-l_1) + B_p, \tag{1.16}$$

where H is the time endowment and B_p is the lump-sum rebate from the tax revenue[10]. The optimal level of leisure that worker i will choose is given by the first order condition: $\frac{\partial u}{\partial l_1} = 0$. After some simple algebra, the amount of leisure that worker i will choose and the corresponding utility level are

$$\begin{aligned} l_1(t_L) &= (1-\lambda)\Big[H + \frac{B_p}{(1-t_L)w_1}\Big], \\ U_{i1,P} &= \frac{\left\{\Big[(1-t_L)w_1\lambda H + \lambda B_p\Big]^{\lambda}\Big[(1-\lambda)(H + \frac{B_p}{(1-t_L)w_1})\Big]^{1-\lambda}\right\}^{1-\theta_i}}{1-\theta_i}. \end{aligned} \tag{1.17}$$

For worker i in sector 2, the consumption function is

$$c_{i2} = \begin{cases} (1-t_L)w_{2L}(H-l_{2L}) - \gamma + B_p & \text{with } P_L \\ (1-t_L)Y_L + (1-t_M)[w_{2M}(H-l_{2M}) - Y_L] - \gamma + B_p & \text{with } P_M \\ (1-t_L)Y_L + (1-t_M)(Y_M - Y_L) + (1-t_H)[w_{2H}(H-l_{2H}) - Y_M] - \gamma + B_p & \text{with } P_H, \end{cases}$$

[10]In this model, I rebate the tax revenue to every worker in an equal amount. Therefore, on average, no income effect imposed by the progressive tax will exist on the labor supply decision. That is, the substitution effect shows the decrease in labor supply if the income tax is more progressive.

where $Y_L = w_{2L}(H - l_{2L})$ is the cutoff in the first tax bracket and $Y_M = w_{2M}(H - l_{2M})$ is the cutoff in the second tax bracket.

Taking first-order conditions, the optimal level of leisure that worker i will choose and the corresponding indirect utility levels are

$$\begin{aligned} l_{2L}(t_L) &= \frac{1-\lambda}{(1-t_L)w_{2L}}\Big[(1-t_L)w_{2L}H - \gamma + B_p\Big], \\ l_{2M}(t_M) &= \frac{1-\lambda}{(1-t_M)w_{2M}}\Big[(1-t_M)w_{2M}H + (t_M - t_L)Y_L - \gamma + B_p\Big], \\ l_{2H}(t_H) &= \frac{1-\lambda}{(1-t_H)w_{2H}}\Big[(1-t_H)w_{2H}H + (t_M - t_L)Y_L + (t_H - t_M)Y_M - \gamma + B_p\Big]; \end{aligned} \tag{1.18}$$

$$\begin{aligned} U_{i2,L} &= \frac{\left\{\Big[(1-t_L)w_{2L}\lambda H - \lambda\gamma + \lambda B_p\Big]^{\lambda}(l_{2L}(t_L))^{1-\lambda}\right\}^{1-\theta_i}}{1-\theta_i}, \\ U_{i2,M} &= \frac{\left\{\Big[(1-t_L)Y_L + (1-t_M)[w_{2M}(H-l_{2M}) - Y_L] - \gamma + B_p\Big]^{\lambda}(l_{2M}(t_M))^{1-\lambda}\right\}^{1-\theta_i}}{1-\theta_i}, \\ U_{i2,H} &= \frac{\left\{\Big[(1-t_L)Y_L + (1-t_M)(Y_M - Y_L) + (1-t_H)[w_{2H}(H-l_{2H}) - Y_M] - \gamma + B_p\Big]^{\lambda}(l_{2H}(t_H))^{1-\lambda}\right\}^{1-\theta_i}}{1-\theta_i}. \end{aligned} \tag{1.19}$$

Therefore, the expected utility under the progressive tax regime in sector 2 can be expressed as

$$E[U_{i2,P}] = U_{i2,L} * P_L + U_{i2,M} * P_M + U_{i2,H} * P_H.$$

1.4.2.2 Flat-Tax Case

In the flat tax regime, $t_L = t_M = t_H = \tau$. Consumption for worker i in sector 1 is then

$$c_{i1} = (1-\tau)w_1(H - l_1) + B_\tau, \tag{1.20}$$

where B_τ is the lump-sum rebate from the total tax revenue.

The optimal level of leisure that worker i will choose and the corresponding utility are

$$\begin{aligned} l_1(\tau) &= (1-\lambda)\Big[H + \frac{B_\tau}{(1-\tau)w_1}\Big], \\ U_{i1,F} &= \frac{\left\{\Big[(1-\tau)w_1\lambda H + \lambda B_\tau\Big]^{\lambda}\Big[(1-\lambda)(H + \frac{B_\tau}{(1-\tau)w_1})\Big]^{1-\lambda}\right\}^{1-\theta_i}}{1-\theta_i}. \end{aligned} \tag{1.21}$$

For worker i in sector 2, the consumption function is

$$c_{i2} = \begin{cases} (1-\tau)w_{2L}(H - l_{2L}) - \gamma + B_\tau & \text{with } P_L \\ (1-\tau)w_{2M}(H - l_{2M}) - \gamma + B_\tau & \text{with } P_M \\ (1-\tau)w_{2H}(H - l_{2H}) - \gamma + B_\tau & \text{with } P_H. \end{cases}$$

Taking first-order conditions, the optimal level of leisure that worker i will choose and the corresponding utility levels are

$$\begin{aligned} l_{2k}(\tau) &= (1-\lambda)\Big[H - \frac{\gamma}{(1-\tau)w_{2k}} + \frac{B_\tau}{(1-\tau)w_{2k}}\Big], \\ U_{i2,k}(\tau) &= \frac{\left\{\Big[(1-\tau)w_{2k}\lambda H - \lambda\gamma + \lambda B_\tau\Big]^{\lambda}\Big[(1-\lambda)(H - \frac{\gamma}{(1-\tau)w_{2k}} + \frac{B_\tau}{(1-\tau)w_{2k}})\Big]^{1-\lambda}\right\}^{1-\theta_i}}{1-\theta_i}, \end{aligned} \tag{1.22}$$

where $k \in \{L, M, H\}$. Thus, the expected utility can be expressed as

$$E[U_{i2,F}] = U_{i2,L} * P_L + U_{i2,M} * P_M + U_{i2,H} * P_H. \tag{1.23}$$

1.4.3 Defining the Tax Rebates

Under the flat tax regime, the total tax revenue collected is

$$E[R_\tau] = n(1-h_m)\tau w_1(H - l_1(\tau)) + nh_m P_k \sum_{k=L}^{H} \tau w_{2k}(H - l_{2k}(\tau)), \tag{1.24}$$

where $k \in \{L, M, H\}$.

Under the progressive tax regime, the total tax revenue collected is

$$\begin{aligned} E[R_p] =& n(1-h_m)t_L w_1(H - l_1(t_L)) \\ &+ nh_m\Big\{t_L w_{2L}(H - l_{2L}(t_L))P_L + \big[t_L Y_L + t_M[w_{2M}(H - l_{2M}(t_M)) - Y_L]\big]P_M \\ &+ \big[t_L Y_L + t_M(Y_M - Y_L) + t_H[w_{2H}(H - l_{2H}(t_H)) - Y_M]\big]P_H\Big\}. \end{aligned} \tag{1.25}$$

where h_m is the proportion of workers in sector 2. This value will be calibrated from the current tax system. The expected tax revenue should be equal under the progressive tax regime and flat tax regime (i.e., $E[R_\tau] = E[R_p]$). The tax rebates are defined as $B_\tau = \frac{E[R_\tau]}{n}$ for the flat tax regime, and $B_p = \frac{E[R_p]}{n}$ for the progressive tax regime. Under the contant tax revenue condition, $B_\tau = B_p$. All the tax revenue is rebated to the workers and every worker gets the same amount of lump-sum rebate.

The above total amount of tax revenue is a function of optimal level of leisure. The optimal level of leisure is a function of the tax rebate. I use iteration to solve B_τ and B_p.

1.4.4 Welfare Analysis

The measure of welfare is determined by using the same approach as under the core model. The idea is to compensate with a certain amount of consumption to make worker i indifferent to a progressive tax regime or a flat tax regime. For workers in sector 1, the amount of consumption Δc_i needs to meet

$$U_{i1,P}(c_{i1}, l_{1,P}) = U_{i1,F}(c_{i1} + \Delta c_i, l_{1,F}). \tag{1.26}$$

For workers in sector 2, the amount of consumption Δc_i needs to meet

$$E[U_{i2,P}(c_{i2}, l_{2,P})] = E[U_{i2,F}(c_{i2} + \Delta c_i, l_{2,F})]. \tag{1.27}$$

In the aggregate level, there are no capital owners, and the total welfare measure is determined by aggregating the compensating consumptions for all workers and checking how it changes

when the tax system is more progressive:

$$\Delta c = \sum_i \Delta c_i. \tag{1.28}$$

1.4.5 Calibration

I focus on working-age individuals, aged 20-60 in the year 2000, whose population weight is nonzero. Each observation in the sample is weighted by its PSID supplied sample weight. For the value of the time endowment, $H = 16(hours/day) * 5(days/week) * 52(weeks) = 4,160$. I drop individuals whose annual hours of work are above the time endowment. According to the CNEF-PSID codebook, "If the individual had positive wages in the previous year and worked at least 52 hours, then the individual was employed. Otherwise, the individual was not employed." Thus, I drop individuals whose annual hours of work are below 52 hours.

An individual's hourly wage is equal to the individual's labor earnings divided by work hours of the individual. I further drop observations of those whose hourly wage is below the federal minimum wage rate of \$5.15/hour. The remaining size of the sample is 3,236 individuals.

In the sample, there are two groups: workers who had obtained at most a high school diploma in the year 2000, and workers who had attended at least some college in the year 2000. Wage is exogenous in my model. The pre-tax mean hourly wage in sector 1 is equal to the pre-tax mean hourly wage for all workers who had obtained at most a high school diploma in the year 2000: \$16.66 (i.e., $w_1 = 16.66$). I assume that $w_1 = w_{2L} = 16.66$. The pre-tax mean hourly wage in sector 2 is equal to the pre-tax mean wage for all workers who had attended at least some college in the year 2000: \$27.08 (i.e., $w_{2M} = 27.08$). Because the wage follows a lognormal distribution, thus $ln(w_{2M}) - ln(w_{2L}) = ln(w_{2H}) - ln(w_{2M})$. I can then calculate $w_{2H} = 43.99$. Given the values of w_{2L} and w_{2M}, I observe these two points at the 37.8 percentile and 67.3 percentile of the wage distribution. Hence, I set $P_L = 0.378$, $P_H = 0.327$, and $P_M = 1 - P_L - P_H = 0.295$.

I use the mean of annual hours of work divided by the time endowment to approximate the value of λ in my sample. I calculate a value of 0.488, which is between 1/3 in Kydland and Precott (1982) and 0.5 in French (2005).

For the values of t_H, I choose the statutory marginal tax rates 31%. The value of t_M is 23%, which is the average federal income tax rate in the year 2000. I assume $\tau = t_M = 23\%$. Based on the constant tax revenue condition under the progressive tax regime and flat tax regime, I can

Table 1.2: Calibrated Parameters In Extended Model

Parameters	Value
h_m	0.533
w_1	\$16.66
w_{2L}	\$16.66
w_{2M}	\$27.08
w_{2H}	\$43.99
P_L	0.378
P_M	0.295
P_H	0.327
λ	0.488
γ	\$10,739

further impute t_L, which is around 20%.

I draw a three-parameter lognormal distribution (shape=0.3, scale=0.3, threshold=2.0), which approximates the targeted risk aversion distribution in the sample.

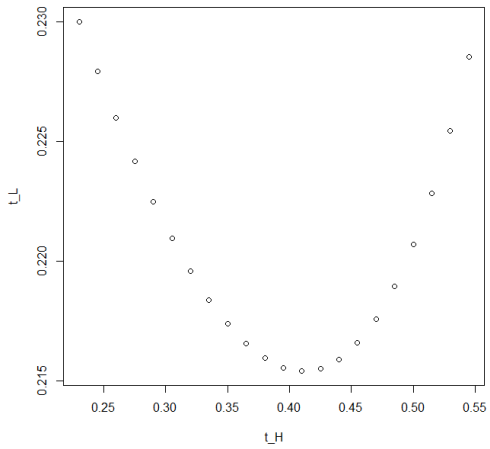
The allocation of workers in sector 2 (h_m) is the proportion of workers who had attended at least some college in the year 2000, which is 0.533. Since I know the distribution of θ and also the cumulative distribution function evaluated at θ_m, $h(\theta_m) = 0.533$, I can find the marginal worker's θ_m who is indifferent to choose sector 1 or sector 2 in the year 2000, $U_{m1} = E[U_{m2}]$. Using this equation, I can calibrate the only remaining unknown parameter,γ, the cost of accumulating higher human capital, which is $\$10,739$. The unit of γ is the annual earnings. Table 1.2 summarizes the calibration values for each parameter.

1.4.6 Results

I start at the point where marginal tax rates, $t_L = t_M = t_H = 23.0\%$. Then I make the tax system more progressive: increasing t_H in a increment of 1.5%. To keep the tax revenue constant, I search t_L to match each t_H. The relationship between t_L and t_H is shown in Figure 10. When $t_H > 41\%$, the labor supply distortion is large. The tax revenue collected from the high income bracket decreases in a large amount. Thus, the marginal tax rate t_L eventually increases to keep the tax revenue constant, although it is still smaller than τ to keep the insurance effect working. Figure 11 presents how the more progressive tax system distorts the labor supply decision: leisure $l_{2H}(t_H)$ increases when t_H increases.

By using equations (1.26) and (1.27), individual i's welfare difference between the progressive tax and flat tax regimes can be captured by Δc_i. By using equation (1.28), the aggregate welfare

Figure 1.10: The Relationship between Marginal Tax Rate t_L and Marginal Tax Rate t_H



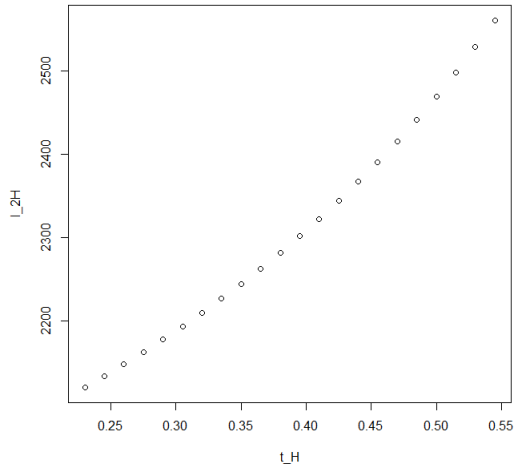
Note: Under the more progressive tax system, when t_H increases, t_L needs to be adjusted in order to keep the tax revenue constant. In addition, the flat tax regime indicates $t_H = t_L$, which is shown at the initial point in Figure 1.10.

difference between the progressive tax and flat tax regimes can be captured by Δc.

Since I assume workers are heterogeneous in risk aversion, even though they pay the same tax rate in sector 2, the insurance effect will be different for workers with different degree of risk aversion. Figures 1.12(a) and 1.12(b) display the hump shapes between Δc and t_H for the marginal worker (i.e., $\theta_{marg} = 3.38$) and the least risk averse worker (i.e., $\theta_{min} = 2.48$). Since the worker with $\theta_{marg} = 3.38$ is more risk averse than the worker with $\theta_{min} = 2.48$, his insurance effect will be larger than that for the least risk averse worker. Therefore, consumption needs to compensate more for him in the flat tax regime in order to match the expected utility level in the progressive tax regime. That is, Δc_{marg} is larger than Δc_{min} before the inflection point. In addition, as Figures 1.12(a) and 1.12(b) show, the efficient degree of progressivity for $\theta_{marg} = 3.38$ is $t_H = 0.335$ and $t_L = 0.218$; the efficient degree of progressivity for $\theta_{min} = 2.48$ is $t_H = 0.290$ and $t_L = 0.222$. That is, the efficient degree of progressivity is more progressive for the marginal worker because of larger insurance effect. When the tax system is very progressive, the distortion of labor supply and the reduction of expected net-of-tax returns to human-capital acquisition dominate the insurance effect; thus the marginal worker and least risk-averse worker are both worse off in the progressive tax regime, compared to flat tax regime.

Figure 1.12(c) shows the hump shape between Δc_{max} and t_H for the most risk-averse worker (i.e., $\theta_{max} = 5.83$). He is always better off in the progressive tax regime, compared to flat tax regime

Figure 1.11: The Relationship between Leisure l_{2H} and Marginal Tax Rate t_H



Note: When the tax system is more progressive (i.e., the marginal tax rate t_H increases), the labor supply distortion is larger (i.e., yearly hours of leisure, $l_{2H}(t_H)$, increases).

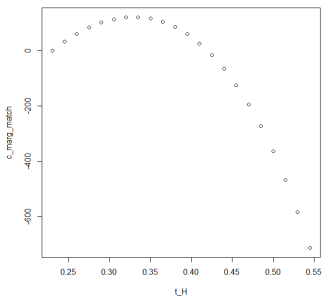
(i.e., $\Delta c_{max} > 0$), because the most risk-averse worker pays t_L, which is always smaller than τ. When keeping tax revenue constant, t_L first decreases than increases; thus his welfare gain first increases than decreases.

Figure 1.13 shows a hump shape between aggregate welfare change Δc and t_H and the efficient degree of progressivity is at $t_L = 21.7\%$ and $t_H = 36.5\%$, given $t_M = 23.0\%$. When the tax is more progressive compared to the flat tax, the insurance effect initially dominates the labor supply distortion and the reduction of expected net-of-tax returns to human-capital acquisition. After the tax becomes more progressive than the efficient point, the labor supply distortion and the reduction of expected net-of-tax returns to human-capital acquisition dominate the insurance effect.

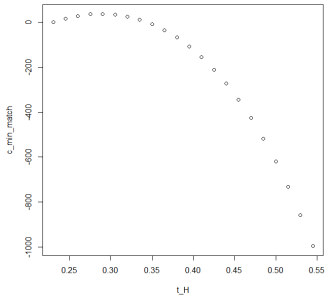
1.5 Conclusion

The welfare analysis in this paper is based on efficiency instead of redistribution. Three effects are considered in unison: the insurance effect, the reduction of expected net-of-tax returns to human-capital acquisition, and the labor supply distortion. I use each worker's willingness to pay for the insurance being provided as a welfare measure on the individual level, which depends on each worker's degree of risk aversion. Based on the Kaldor-Hicks efficiency criterion, I further aggregate each individual's willingness to pay to find the efficient degree of progressivity. Therefore, the

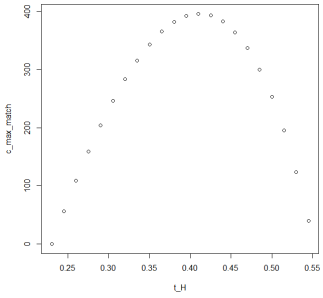
Figure 1.12: Individual Welfare Change with Respect to the Individual's Degree of Risk Aversion



(a) Welfare Change for θ_{marg} in Sector 2



(b) Welfare Change for θ_{min} in Sector 2



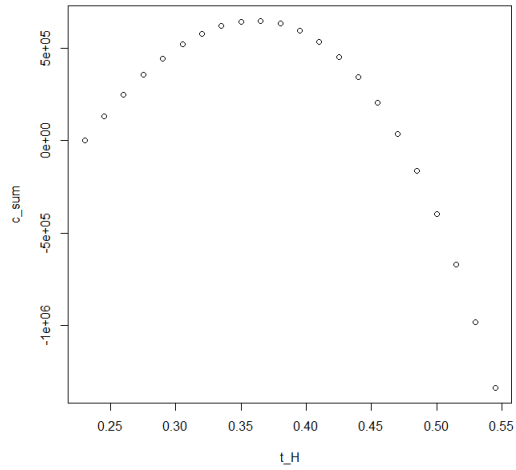
(c) Welfare Change for θ_{max} in Sector 1

willingness to pay approach provides a fruitful lens for examining the efficient degree of progressivity for income taxation.

To characterize the efficient degree of progressivity, I construct models to examine how progressivity for income taxation influences workers with heterogeneous risk aversion to sort into two sectors with different levels of lifetime earnings uncertainty. In the core model, as in Thaler and Rosen (1976), there is a "hedonic wage locus", which indicates how the market allows individuals to substitute the mean level of the wage for its variability across sectors. However, in my general equilibrium framework, the variability-return locus varies with respect to changes in the tax rate. Furthermore, the shape of the indifference curve between the mean and variability of consumption is different for workers with different levels of risk aversion. In the extension of the core model, I further account for the labor supply response with respect to the degree of progressivity for income taxation. After calibrating the model, I find the efficient degree of progressivity is similar to that of the U.S. tax code.

The sorting mechanism in my paper is based on heterogeneous risk aversion. In future work, I will add an additional dimension of sorting: heterogeneous ability. The two dimensions will jointly determine how workers sort into two sectors: a sector with high uncertainty of lifetime earnings and a sector with low uncertainty of lifetime earnings. Moreover, I will further research how the joint dimensions of heterogeneous ability and heterogeneous risk aversion determine the efficient degree of progressivity.

Figure 1.13: The Relationship between Aggregate Welfare Change and Marginal Tax Rate t_H



Note: As the tax system becomes more progressive, the aggregate welfare first increases and then decreases after an inflection point. The efficient degree of progressivity is at the marginal tax rates $t_L = 21.7\%$ and $t_H = 36.5\%$, given $t_M = 23.0\%$.

Chapter 2

Worker Quality, Wages and the Education Premium in the United States, 1980-2005

2.1 Introduction

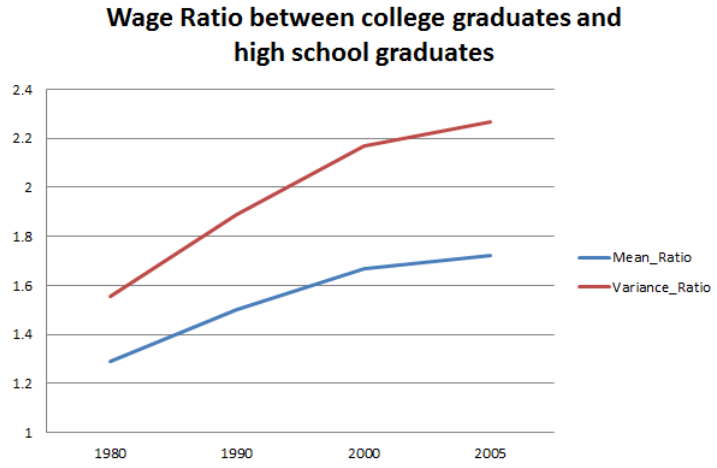
The education premium, measured as the relative wages of college graduates to high school graduates, has increased dramatically since the 1980s. Acemoglu (2002) argues that skill-biased technology has caused the growth in wage inequality in the U.S during the past two decades. Autor, Katz, and Kierney (2008) show that education premium growth was especially large in the early 1980s, and continued at a slower pace through the mid-1990s.[1] An important question is the extent to which the slower growth is explained by the relative change of quality of college graduates?

This paper explores how the quality of college and high school graduates affects wages and the education premium. Figure 2.1 shows trends of the mean wage ratio and variance wage ratio for college graduates and high school graduates for full time-full year white male workers aged 26-60 in

[1] Pierce (2001) argues that inequality growth in broader measures of compensation (fringe benefits) slightly exceeds wage inequality growth over the 1981 to 1997 period. Since currently available data often lack information on benefit costs and multiple benefits categories, he used Employment Cost Index (ECI) data. Unfortunately, no worker demographic or human capital information is available in the ECI, which is the focus of my paper. Therefore, I do not consider the inequality or the education premium in nonwage compensation, which includes health and life insurance, several forms of leave, pension and savings plans, bonuses and unemployment insurance.

U.S.[2] The relative variance increases more than the relative mean wage over time. Thus, attending at least some college is a risky decision because some people after obtaining a college degree may be on the left tail of the returns to college and receive a very low wage. The distribution of the returns to college is related to the graduate's occupational choice. Hence, I link my new measurement of the quality of college graduates and quality of high school graduates to an occupation skill index based on a worker's place in the wage-skill distribution.

Figure 2.1: Trends of Mean Wage Ratio and Variance Wage Ratio



This new measurement of quality is a weighted occupational skill index in the state of birth. The reason why I want to link this new measurement to occupations is that the variance for the return to college is increasing over time, and the return to college is directly related to the occupation choice. The result shows that worker quality influences the wage for college graduates more than the wage for high school graduates.

There are some classic measurements of human capital accumulation in a birth state in the literature. For example, the average SAT scores reflects the human capital accumulation before college in a birth state. The average ranking of colleges and universities shows the quality of colleges as well as the human capital accumulation in colleges. The new measurement in this paper includes all human capital accumulation before people start to work. Although the human capital accumulation in the new measurement has a longer period than the former two human capital accumulations, there may be some correlation among them. Hence, I will give evidence about the validity of the new measurement.

[2]The data comes from 1980, 1990, 2000 Census and 2005 American Community Survey.

Carneiro and Lee (2011) argue that increased college enrollment led to a decline in the average quality of college graduates between 1960 and 2000, resulting in a decrease of 6 percentage points in the college premium. It is the first paper to decompose the college premium into price (quantity) effects and composition (quality) effects. They argued that college enrollment is an important factor influencing the quality of college graduates. They also regressed the wage for the high school group on college enrollment. However, college enrollment is not a good measurement for the average quality of high school graduates. Furthermore, college enrollment only affects the quality of college education, and it does not fully capture the quality of workers with a college degree, which is the quality of college graduates. Because quality of college education measures only human capital accumulation within college, but the quality of college graduates measures all human capital accumulation until individuals get a college degree. Therefore, I use a new measure of quality of college graduates and test how it influences the wage.

The structure of this paper is as follows: Section 2.2 describes the empirical strategy I use to measure worker quality and the validity of the new measurement. The empirical model is presented in section 2.3. The data and the empirical results are provided in sections 2.4 and 2.5, respectively. Section 2.6 shows sensitivity analysis. Section 2.7 concludes.

2.2 Methodology

I measure the quality of college and high school graduates by the average skill level in their state of birth using two steps. First, I revise the method of Autor and Dorn (2013) to calculate the skill percentile for each occupation in the whole nation level in every year. Second, I calculate average skill level in each state of birth for each group.

2.2.1 Skill percentile for occupations

First, in order to get consistent categories of occupations, I refer to the system developed by Dorn (2009). There are a new occupation system with 330 "occ1990dd" codes that provides a balanced panel of occupations covering the 1980, 1990, and 2000 Censuses and the 2005 ACS. Next, in each year, according to the mean real weekly wage for each occupation [3], I rank the skill level for

[3]Skill index is likely to be contaminated by price level effects. However, Glaeser and Resseger (2009) argue that bigger cities can attract more skilled workers and agglomeration makes workers more skilled too. In addition, more productive workers can live in cities because they can afford a higher cost of living. Thus, the contamination of skill

Table 2.1: 1980 Skill Percentile for Occupations

Occupation Category	Mean Real Weekly Wage	Employment Share	CER	Skill Percentile
Food preparation and service workers	89.10	0.00064	0.00064	0.064
Shoemaking machine operators	94.00	0.00024	0.00088	0.088
Hotel clerks	96.06	0.00015	0.00103	0.103
Child care workers	102.40	0.00025	0.00128	0.128
Clothing pressing machine operators	102.44	0.00015	0.00143	0.143
...	...	...	...	...
...	...	...	...	...
Optometrists	302.37	0.00024	0.99227	99.227
Health and therapy occupations	304.02	0.00014	0.99241	99.241
Podiatrists	359.22	0.00006	0.99247	99.247
Dentists	390.19	0.00124	0.99371	99.371
Physicians	395.38	0.00629	1.00000	100.00

Notes: Mean weekly wage is adjusted by CPI1999. The skill percentile is equal to the cumulative employment share multiplied by 100. CER means Cumulative Employment Share.

each occupation by using cumulative share of employment. The methodology for developing skill percentile for occupation j in year t is described in the following function:

$$SkillPercent_{jt} = \sum_{k=1}^{j} EmployShare_{kt} * 100, \tag{2.1}$$

where occupation index j ranges from 1 to 330 and mean wage $\overline{w_k} < \overline{w_j}$ for $\forall$ k<j in year t.

Table 2.1 gives an example to illustrate how to calculate the skill percentile for occupations in 1980. The bottom 5 and top 5 occupations are showed according to their mean weekly wage. The skill percentile is equal to the cumulative employment share multiplied by 100. Similarly, the skill percentile for occupations is calculated in the same way for the year 1990, 2000 and 2005. So each year has its own occupation skill in the national level.

2.2.2 Average skill level in states of birth

The birthplace of workers is predetermined, though they can choose to work in any occupation in any state. Consistent with the previous literature, I assume workers are born and educated in the same state.[4]

index should not be an important issue. But in the section of robustness checks, I correct real wage by running wages on a vector of city dummies and use residual wages to construct the skill percentiles.

[4]Card and Krueger (1992) show the evidence that there are relatively low mobility rates of preschool and school-age children. They assumed that an individual attends public elementary and secondary schools in his state of birth. However, the assumption that people attend college in the state of birth needs to be tested. Unfortunately, there is no cross section individual data available for the place of education in the state level. Following Winters (2012), one noisy and not ideal way to infer the college location of recent graduates (i.e. those who are ages 23-27 in the census and likely in school five years prior) is to use five-year migration data in the decennial censuses in 2000 and prior. There is one main five-year migration data in the state level in IPUMS: state or country of residence 5 years ago (MIGPLAC5). By using MIGPLAC5 and the variable of state of birth, the result shows that there is nearly 66 percent of college graduates studying in the same state of birth in U.S in the year 2000. Therefore, the assumption

Then I calculate average skill level in each state of birth for particular age group and education group in particular year. This index is a new measurement for the quality of workers which is directly related to where they were born. The quality here does not only capture the education resources in particular state of birth, it also reflect the average abilities and average family background of workers. These factors influence the wage received in their occupational state (not necessarily the same as the state of birth).

Since I want to focus on the education or skill premium[5], I separately calculate the average skill level for the high school graduates and college graduates at each state of birth in each year. Also, this quality index is different across the age groups. The following equation describes the calculation of the average skill level for the age group a, education group k, in state of birth b in year t:

$$AvegSkillPercent_{kabt} = \sum_{i=1}^{N_{kabt}} SkillPercent_{ikabt}/N_{kabt}, \tag{2.2}$$

where individual i born in state of birth b in age group a, education group k and work anywhere in a particular occupation corresponding with the skill percentile in year t. N_{kabt} is the total number of people in the sample born in state b belonging to the age group a, education group k in year t.

For example, how to calculate the quality of college graduates in the age group 25-30 born in New York State in year 1980? I average skill percentile for all people who are born and get college degree in New York State in the age group 25-30 in 1980. Each person in that group has occupation in a state (*not necessarily New York State*) and the occupation has skill level.

Table 2.2 illustrates how to calculate the quality index for college graduates in the age group 25-30 born in New York State in year 1980:

$$AverageSkillLevel = \frac{0.448 + 0.412 + 3.2 + 0.286 + 6.666 + ...595.362 + 297.723 + 595.582 + 4471.695 + 25700}{7 + 4 + 25 + 2 + 33 + ... + 6 + 3 + 6 + 45 + 257} = 58.744. \tag{2.3}$$

The key point of this methodology is that the average skill level captures the quality of the supply side. However, it is possible that mobility is so low across state of birth and state of work, and at the same time, the demand for some occupations is very high within a state of birth[6]. Thus, the demand shock for some occupations within the state may contaminate the measurement

that people attend college in the state of birth is reasonable based on Census and ACS data. Besides this strategy, I also refer to a small sample from National Center for Education Statistics and it shows that more than 90 percent of first-time first-year students go to college in the same state of residence in 2000.

[5]In this paper, I do not distinguish the education premium and skill premium.

[6]Thanks for this comment from Steven Durlauf and Christopher Taber in 2015 Summer School on Socioeconomic Inequality in UChicago.

Table 2.2: In 1980, the average quality of college graduates, age 25-30, who were born in New York State

Occupation Category	Skill Percentile (1)	Number of observations (2)	Product(1)*(2)
Food preparation and service workers (444)	0.064	7	0.448
Hotel clerks (317)	0.103	4	0.412
Child care workers (468)	0.128	25	3.2
Clothing pressing machine operators (747)	0.143	2	0.286
Waiters and waitresses (435)	0.202	33	6.666
...	...	...	...
...	...	...	...
Optometrists (87)	99.227	6	595.362
Health and therapy occupations (89)	99.241	3	297.723
Podiatrists (88)	99.247	6	595.582
Dentists (85)	99.371	45	4471.695
Physicians (84)	100.00	257	25700

Notes: The skill percentile is calculated in the first step by the nation level. The number of observations means how many observations are for the occupations i in this group.

Table 2.3: The average mobility rate across state of birth and state of residence for FTFY white male workers,aged 25-60

Year	State Level	State−High School Graduates	State−College Graduates
1980	0.373	0.304	0.454
1990	0.394	0.296	0.459
2000	0.398	0.309	0.465
2005	0.399	0.312	0.460

Notes: Column 1 includes all the samples. Column 2 only includes workers who attended at least some colleges. Column 3 only includes workers who got at most high school diploma.

of quality. However, Table 2.3 gives evidence that average mobility rate across states is quite high, especially for the workers who attended at least some college−college graduates group. Therefore, the contamination from the demand shock should not be a big concern for the new measurement of quality.

2.2.3 The validity of the new measurement

This new measurement of quality captures the human capital accumulation in the state where workers are born and educated. How is it related to some traditional measurements of human capital accumulation in a birth state, like average SAT score across states? I will run the regression on the new measurement to SAT score for the college group to see whether they have any correlation.

The average SAT scores across states come from National Center for Education Statistics. Available data for SAT scores is discontinuous across years. Since the SAT test was re-centered in 1995, I converted the recentered scale in 1995 to the original scale before 1995 in order to compare

the SAT scores across years [7].

Assume people take their SAT test around 16 years old. I will only focus on workers who are 26 years old and attended at least some college in order to match the years in my data with the years of the SAT data. In addition, 26 years old workers' SAT scores are more related to my measurement of quality because they have just worked for a few years rather than older workers whose wage may no longer reflect their SAT scores. Thus, I match 26 years old worker's quality in 1990 to SAT score in 1980; worker's quality in 2000 to SAT score in 1990; worker's quality in 2005 to SAT score in 1995.

Table 2.4 column 1 shows the regression result, and they have a negative correlation. However, there is one important explanation why this result is strange: perhaps people in different states have a different preference for SAT and ACT. It is possible that the average SAT score is very high in Wisconsin because only top students want to take the SAT in order to apply to universities in the northeast. However, the average quality of college graduates there may be very low. Thus, in column 2, I add one important independent variable: percent of graduates taking the SAT in 1993 [8].

Table 2.4 column 2 displays that there is nearly no correlation between average SAT scores and the new proposed measurement of quality. This was not expected. At least one important reason can explain why there is no correlation: SAT scores are actually not good to measure the quality of college graduates because the new measurement of quality includes all the human capital accumulation until people graduate from college, and SAT just measures human capital accumulation before people get into college. That means human capital accumulated in college or university is much more important than human capital accumulated before going to college. Thus, there is nearly no correlation between SAT score and my new measurement. I will further explore this point of view.

[7] http://research.collegeboard.org/programs/sat/data/equivalence/sat-mean

[8] This is the earliest data I can get for the percent of high school graduates taking the SAT across states. Thus, I assume it will not change much from 1980 to 1993, and I will use it for all years 1980, 1990, 1995.

Table 2.4: Correlation of average SAT scores and the new measurement of quality across states

	AvgSkill Level(1)	AvgSkill Level(2)
SAT Verbal	-0.025**	0.023
	(0.011)	(0.023)
Percent of graduates taking SAT		0.070**
		(0.029)
Observations	150	150
SAT Math	-0.022**	0.013
	(0.010)	(0.018)
Percent of graduates taking SAT		0.060**
		(0.026)
Observations	150	150

Notes: I exclude District of Columbia since there is no average SAT data in 1980.

2.3 The Empirical Model

2.3.1 The wage structure

I utilize the framework of wage structure created by Carneiro and Lee (2011). Suppose that the wage of each individual i, of age a, at time t, who is born and goes to school in state b and works in state r (which may or may not be equal to b) can be written as:

$$W^k_{iatrb} = \Pi^k_{atr} * U^k_{i,t-a,b}, \tag{2.4}$$

where W^k_{iatrb} is the wage, Π^k_{atr} is the price of k-type skill for those with schooling level k [9], in age group a in year t working in state r, and $U^k_{i,t-a,b}$ is the individual specific endowment of k-type skill for those in cohort $t-a$ and in state b. After averaging across individuals in each group, it follows from (2.4) that

$$W^k_{atrb} = \Pi^k_{atr} * U^k_{atrb}, \tag{2.5}$$

Thus, the average wage of workers in each group (k, a, t, r, b) is the sum of the price effect (or quantity) part and composition effect (or quality) part. The price effect, Π^k_{atr} is the interaction of the demand for skill with the quantity of skill supplied in the local labor market. The composition effect, U^k_{atrb} is determined by the quality of college and high school graduates across state of births b in a particular year.

[9] k equals two: high school graduates group and college graduates group.

Taking logs from equation (2.5):

$$w^k_{atrb} = \pi^k_{atr} + v^k_{atrb}. \tag{2.6}$$

This equation is the basis of the empirical models.

2.3.2 The composition/quality part

The composition part v^k_{atrb} varies over groups because of changes in the average quality of college and high school graduates. If a worker is born in the state with higher average quality, then he will get higher wage in the state where he works after controlling the price or quantity effect. Thus, the main variable of interest in the composition part is average quality of workers for the age group a in particular state of birth b in particular year t: P_{tba}.

Migration is one concern in order to specify P_{tba} in the composition part v^k_{atrb}. Since migration is not random (see, e.g., Heckman, Layne-Farrar, and Todd 1996, and Dahl 2002), it will lead to self selection.[10] Thus, I control for selective migration by using time-varying interactions (state of birth)×(state of residence): γ_{ktrb}.

In addition, Carneiro and Lee (2001) hypothesize that the proportion of college-goers in cohort $t-a$ born in region b is one of the important factor determining differences in the average quality of college graduates across regions of birth. They calculate an average proportion of individuals who attend at least some college for each cohort and region of birth (common across years), by averaging this number across all years. However, in my new measure of average quality of both college and high school graduates, I use the average skill level for each group (t, a, b). Moreover, comparing skill levels changes over time is difficult because they are a function of mean wages in each year. The skill percentile for particular occupation, for example CEO, is higher in 1990 than in 1980. This could be the result of higher quality of CEOs, higher demand for their work, or lower supply in 1990, which all influence the mean wage of CEOs. Thus, in the composition/quality part, I add a year fixed effect, γ_t.

School resources may very across states. Well endowed states could have higher quality schools, which simultaneously lead to higher worker quality and affect worker's wage in future. I

[10]The intuition is based on the cost-benefit analysis. Persons migrating long distances will tend to have higher earnings in all destination regions than those persons making short moves or no mores at all because their migration costs will be higher too.

account for this by using state of birth fixed effects, interacted with age: γ_{kba}

Therefore, the composition variable v^k_{atrb} is formulated as follows:

$$v^k_{atrb} = \gamma_{ktrb} + \gamma_{kba} + \gamma_t + \phi_k(P_{tba}), \tag{2.7}$$

where γ_{ktrb} is a state-of-birth by state-of-residence fixed effect that is interacted with year dummies (capturing the selective migration). The variables γ_t and γ_{kba} are year fixed effect and state of birth fixed effect. The form $\phi_k(P_{tba})$ is a function of average skill level for the college and high school graduates in state of birth b for each particular age group in particular year (capturing the quality effect).

2.3.3 The reduced form model

After inserting equation (2.7) into (2.6), I get the first main reduced form model :

$$w^k_{atrb} = \gamma_{katr} + \gamma_{ktrb} + \gamma_{kba} + \gamma_t + \beta_k * \phi_k(P_{tba}) + \epsilon_{atrb}, \tag{2.8}$$

where γ_{katr} is full interactions of age-time-state fixed effects (capturing the price effects). The first function form I use for $\phi_k(P_{tba})$ is P^k_{tba}.

Once I control for the price effect, γ_{katr}, on the wage, I can determine the composition effects β_k. The price effect is the interaction of the demand for skill with the quantity of skill supplied in the local labor market. The composition is determined by the quality of college and high school graduates in the state of birth b for particular age group in particular year.

2.3.4 Identification Problem

In the empirical model, I use state as the geographical unit, but Carneiro and Lee (2011) argue that the reason why they do not use the state as the geographical unit is that the resulting cell sizes would be too small for estimates to be reliable. It is reasonable because there are 2858 groups (51 states*7 age groups*2 education groups*4 years), which leads to no variation within some groups. Thus, the empirical model is hard to identify if I use state level instead of region level. However, it is not likely that quality of college graduates and quality of high school graduates is the same within the census region. Within a state, the quality is more likely to be the same. The

identification problem and the quality similarity are the trade off in this scenario. I will omit some interaction of fixed effects in order to identify the model in the state level compared to the model in Carneiro and Lee (2011). Thus, the reduced model I want to estimate is

$$w^k_{atrb} = \gamma_{ra} + \gamma_b + \gamma_t + \gamma_m + \phi_k(P_{tba}) + \epsilon^k_{atrb}, \tag{2.9}$$

where γ_m is a dummy indicates whether on average there is migration from state of birth b to state of residence r in the group (a,t,k).

2.3.5 Instrumental Variable

The skill percentile is calculated for the information of mean wages and employment shares of occupations, so the average skill percentile P_{tba} may include some component of the mean wage as the dependent variable. To correct this endogeneity problem, I use an instrumental variable approach. The IV for P_{tba}, $P_{tb(a-1)(a+1)}$, is the weighted average skill level of the adjacent age groups, in the same year, same state of birth, and same education group. For example, for the quality of college graduates in the age group 31-35 in 1980, who were born in New York State, I use weighted average quality of college graduates of age group 25-30 and age group 36-40 in 1980, who were born in New York State, as the IV.

There are two main reasons for creation of this instrumental variable. First, the weighted average quality of college graduates of adjacent age groups is similar as the average quality of college graduates in this particular age group. Second, I can avoid the situation that the dependent variable and independent variable is in the same wage cell because now the IV's age group and dependent variable's age group are different.

However, empirically weak instruments can produce biased IV estimators. Stock and Yogo (2005) tabulate critical values that enable using F-statistic form of the Cragg-Donald (1993) statistic to test whether given instruments are weak. The Cragg-Donald Wald F statistic passes the 10 percent critical value. Thus, rejection of their null hypothesis represents the absence of a weak-instruments problem here.

2.4 Data

2.4.1 Data Description

I use cross-sectional data from the 1980, 1990, 2000 US Censuses and 2005 ACS. Compared to Careiro and Lee (2011), the data come from different census years, but the construction of the data are the same. I focus on white males aged from 25 to 60, and aggregate them into seven age groups: 25-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60. I consider 51 states of birth and 51 states of residence eliminating from the sample those individuals who are foreign-born. For the education groups, the individuals are divided into two categories: high school graduate and college graduate. Thus, individuals are grouped into cells defined by five variables: year, state of residence, state of birth, schooling (high school or college) and age group.

For each cell I compute the relevant log of average weekly wages for full-time/full-year workers. Weekly wages for high school graduates are obtained by taking only males with exactly 12 years of schooling and dividing annual income by annual weeks worked. Weekly wages for college graduates are obtained in an analogous way for individuals with exactly 16 years of schooling.[11]

2.4.2 Summary Statistics

Descriptive statistics are summarized in Table 2.5. In Panel A, each column reports the sample mean and standard deviation of the dependent variable: real weekly wage for cell (a,r,b) in different years for college groups. In both panels, the sample mean and the standard deviation of the wage increases over time. In addition, Table 2.5 also reports the ratio of the mean wage for college graduates to the mean wage for high school graduates. It also increases over time, which suggests the education premium has increased. However, it increases at slower pace over time.

2.5 Empirical Results

Table 2.6 reports estimation results that are obtained by implementing the econometric framework described in Section 3. Robust standard errors are reported in the table, clustered by

[11] Carneiro and Lee (2011) argues that in order to get measures of wages for clearly defined and relatively homogeneous groups of individuals, they ignore high school dropouts, those with some college, those with post-graduate studies.

Table 2.5: Summary Statistics

	1980	1990	2000	2005
Panel A. College				
Dependent Variable: Weekly Wage	233.19	684.23	1310.57	1800.20
	(91.25)	(346.60)	(765.08)	(1158.23)
Observations (groups)	11140	12731	13664	9000
Panel B. High School				
Dependent Variable: Weekly Wage	169.51	422.37	765.04	1018.23
	(54.77)	(170.92)	(368.47)	(552.25)
Observations (groups)	12717	12758	14258	9072
Ratio W^C/W^H	1.38	1.62	1.71	1.77

Notes:(1) The dependent variable is real weekly wage (adjusted to 1999 dollars). Weekly wages for high school graduates are obtained by taking only males with exactly 12 years of schooling. Weekly wages for college graduates are obtained in an analogous way for individuals with exactly 16 years of schooling. (2) Ideally, the number of observations should be 18207 (51 birth states*51 residence state* 7 age groups) in each education category each year. Since there are so many groups, some groups may be quite small, and there will be no observations. For example, there is no observation in age group 36-40, born in Delaware, working in Washington in 1980 for the FTFY white male workers.

year and state of birth.

In column 1 of Table 2.6, I only use the year fixed effect in the regression. Since we cannot compare the skill percentile for occupation across years as mentioned early, the year fixed effect is a very important control variable. Both estimates are significant for the two education groups.

Column 2 of Table 2.6 shows the estimation results for the reduced model (9). If the quantity part is fixed, both college graduates and high school wages respond substantially to changes in the quality of workers. If the quality of high school graduates increases by one percentile, on average their wages will increase by 0.67 percent. Further, if the quality of college graduates increases by one percentile, on average their wages will increase by 0.98 percent.

In column 3 of Table 2.6, I use the IV for the average skill level for the particular age group in each state of birth in each year. Column 3 only includes the year fixed effect. The magnitude of estimates is smaller than those in column 1 of Table 2.6. However, in column 4 of Table 2.6, when I use the IV and estimate the reduced form model (9) with all fixed effects included, the estimates are both statistically and economically significant and they are larger in magnitude than the estimates in column 2 of Table 2.6. If the quality of high school graduates increases by one percentile, on average the high school graduates wages will increase by 0.81 percent.

Since the college premium at time t is defined as $E({w_t}^C) - E({w_t}^H)$, this requires subtracting the college and high school wage equations and averaging across all ages, state of birth and state of residence. Given the wage equation is linear in all variables, I compute the effect of quality on

Table 2.6: Regression of Wages on Quality of College and High School Gradues

	Reduced-form model (1)	Reduced-form model (2)	2SLS (3)	2SLS (4)
Panel A. College				
Quality of college graduates	4.075***	0.984***	3.792***	1.165***
	(0.131)	(0.119)	(0.194)	(0.254)
R^2	0.742	0.786	0.742	0.786
Observations	46535	46535	46535	46535
Panel B. High School				
Quality of high school graduates	3.023 ***	0.670***	2.867***	0.806**
	(0.114)	(0.111)	(0.157)	(0.387)
R^2	0.776	0.806	0.776	0.806
Observations	48805	48805	48805	48805
Included Year Fixed Effect	Yes	Yes	Yes	Yes
Included Other Fixed Effects	No	Yes	No	Yes

Notes: The dependent variable is log real weekly wage in each cell. The independent variable "Quality of college graduates" is measured by the average skill percentile in each age group, each region of birth and each year, which ranges from 0 to 1.

Significant codes: 0.01 '***' 0.05 '**' 0.1 '*'.

the college premium as the difference between the coefficients on P_{tab} in college and high school equations. In column 4 of Table 2.6, the difference is 0.36, implying that the college premium increases by 0.36 percent if the quality of both high school and college graduates increases by one percentile.

2.6 Robustness Checks

In section 2, I argued that the skill index may be contaminated by price level effects. In order to test this conjecture, I first regress the weekly real wage on the city dummies, then use the residual wage to construct the skill percentile. The average skill level is calculated by the adjusted skill percentile according to section 2. The estimation results for the reduced form model (9) are showed in columns (1)-(4) of Table 2.7 . In column 4 of Table 2.7, the difference of the two estimates is 0.36, implying that the college premium increases by 0.36 percent if the quality of both high school and college graduates increases by one percentile. The magnitude of this result is similar as the magnitude calculated Table 2.6, which indicates the result is robust[12].

As an alternative specification, I use the functional form

$$\phi_k(P_{tba}) = P_{tab}/(1 - P_{tab}) \tag{2.10}$$

[12]Besides this method to check the price level effects, the skill percentile can be calculated by using real weekly wage. Here this real weekly wage is adjusted by cities' CPI index, not national CPI. It is considered for future research.

Table 2.7: Regression of Wages on Quality of College and High School Gradues

	Model (1)	Model (2)	2SLS (3)	2SLS (4)	Model (5)	2SLS (6)
Panel A. College						
Quality of college graduates	4.127***	0.999***	3.962***	1.356***	0.125***	0.131***
	(0.117)	(0.121)	(0.179)	(0.283)	(0.017)	(0.327)
R^2	0.744	0.786	0.744	0.786	0.786	0.786
Observations	46535	46535	46535	2268	46535	46535
Panel B. High School						
Quality of high school graduates	3.367***	0.703***	3.403***	0.998**	0.207***	0.173
	(0.119)	(0.122)	(0.166)	(0.480)	(0.041)	(0.129)
R^2	0.777	0.806	0.777	0.806	0.806	0.806
Observations	48805	48805	48805	48805	48805	48805
Included Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Included Other Fixed Effects	No	Yes	No	Yes	Yes	Yes
Using residual wage to construct skill percentile	Yes	Yes	Yes	Yes	No	No
New Function Form	No	No	No	No	Yes	Yes

Notes: The dependent variable is log real weekly wage in each cell. The independent variable "Quality of college graduates"' is measure by the average skill level in each age group, each region of birth and each year, which ranges from 0 to 1.

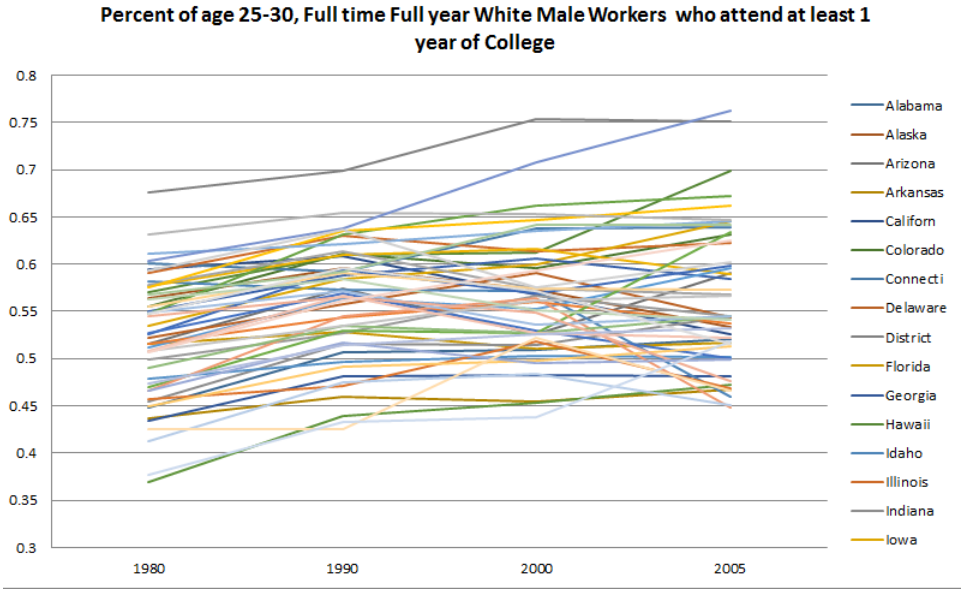
Significant codes: 0.01 '***' 0.05 '**' 0.1 '*'.

for $\phi_k(P_{tba})$. As Carneiro and Lee (2011) argue, this function is a strictly increasing function of P_{tab} and can take any nonnegative value. The estimation results are shown in columns (5)-(6) of Table 2.7. Column 5 shows that for a quality increase of high school graduates from the 50th percentile to the 60th percentile (i.e. $\phi_k(P_{tba})$ increases from 1 to 1.5), the average high school graduates' wage increases by 0.10 percent. Column 6 shows that the education premium will decrease by 0.02 percent if the quality of both high school graduates and college graduates increases from the 50th percentile to the 60th percentile.

2.7 One Dimension of Quality of Workers: Evidence from College Enrollment

There are several important factors influencing the quality of workers, and one of them is the college enrollment. Figure 2.7 shows that the college enrollment is increasing over time, and it varies across states of birth in my sample. Two mechanisms in the literature explain why the increase of college enrollment will decrease the quality of college graduates. Since the college enrollment increases, the marginal student, whose quality is lower than the quality of average students in college, is drawn into college thus it results in a decline in average quality. The other mechanism focuses on the averages. The rise in college attendance means that resources at each college have to be spread more thinly across students, resulting in a lower quality education for each enrolled student. As Carneiro and Lee (2011) argues, it is hard to distinguish those two mechanism empirically.

Figure 2.2: Trends of College Enrollment across States



It is certainly a challenge and very interesting question. However, I will not try to distinguish those two mechanism here because in the scenario of this paper, quality is actually a weighted occupation index. It is interesting to ask how the increase of college attendance affects the occupation distribution of college graduates. The mechanism is that higher college enrollment pushes college graduates entering less skilled occupations, which indicates their quality decreases on average. Thus, college enrollment is causally effecting the quality of college graduates in the state of birth.

However, it is unlikely that college enrollment is the only important factor determining differences in the average quality of workers across the states of birth. For example, school resources may vary across states. Well endowed states (e.g. New York State) could have higher quality schools, which simultaneously lead to higher worker quality and higher college enrollment. Thus, in order to catch the causal effect, I need to account the state of birth fixed effects. I interact the state of birth with age: γ_{ba}, which indicates state of birth has different effect on different cohort. The empirical model is as following:

$$P^k_{tba} = \gamma_{ba} + \gamma_t + \beta_k E_{tba} + \epsilon_{atrb}, \tag{2.11}$$

where P^k_{tba} is the average quality of workers in age group a, state of birth b, education group k in year t; E_{tba} is the college enrollment for the same group[13]; γ_t is year fixed effect.

[13]The college enrollment is defined as the proportion of individuals who attend at least some college in the sample.

Table 2.8: Regression of Quality of College and High School Graduate on College Enrollment

	Reduced-form model
Panel A. College	
Proportion in college	-0.109***
	(0.012)
Observations	1428
Panel B. High School	
Proportion in college	-0.014
	(0.011)
Observations	1428

Notes: The dependent variable "Quality of college graduates" is measured by the average skill level in each age group, each state of birth and each year, which ranges from 0 to 1. The independent variable is the the proportion of individuals who attend at least some college in the sample, which ranges from 0 to 1.

Significant codes: 0.01 '***' 0.05 '**' 0.1 '*'.

Table 2.8 reports estimation results. Robust standard errors are reported, clustered by state of birth and year. The results show that one percent increase in college enrollment leads to 0.11 percent decrease in average quality of college graduates. But one percent increase in college enrollment only causes 0.01 percent decrease in average quality of high school graduates. Therefore, the college enrollment has larger impact on the quality of college graduates than the quality of high school graduates.

It is clear that increase of college enrollment will decease the quality of college graduates. But why the estimate is negative even for the high school group, though it is not significant? The intuition is that the original marginal "good" high school graduates can now go to college because the college enrollment increases. This further pushes down the average quality of high school graduates.

2.8 Conclusion

In this paper, I develop a new measurement of quality of high school and college graduates, the average skill level in each state of birth. This measurement links the worker's quality to skill level of occupations. I use that measure to estimate the effect of relevant worker quality on the observed education premium, finding that increasing both the quality of college and high school graduates by one percent will cause 0.36 percent of increase in the education premium. In addition, one of the key factor influencing quality is the college enrollment in each state of birth. I find that increases in college enrollment will lead to the decline of quality of college graduates more than the quality of high school graduates by nearly 0.1 percent.

It is still puzzling why there is nearly no correlation between average SAT scores and the new proposed measurement of quality. Though those two measure different time periods of human capital accumulation, the details remain obscure. Thus, I will further explore that correlation.

Chapter 3

Top-coded Earnings

3.1 Introduction

The top-coding issue is raised because the wage and salary earnings, when above some particular value, are censored in the American Community Survey (ACS), the March Current Population Survey (CPS) and Decennial Census. Researchers who rely on the top-codes to represent real earnings value of the individuals who earn more than the predetermined top-codes value could potentially lead to misleading results. For example, when the top-code values are lower than the actual earnings and the affected groups are in the high income group, the measurement of income inequality in the United States may be distorted. We aim to explore this issue by using the RDC internal data. We focus on wage and earnings variables in the ACS, March CPS and Decennial Census.

In order to deal with the top codes issues, researchers have adopted various methods to mitigate the problem. For example, adjustment multipliers, like 1.4 or 1.5, are used for the censored earnings in the literature. However, as Armour, Burkhauser, and Larrimore (2014) argued, such an approach may misstate top earnings if the wrong multiple is used or if the appropriate multiple changes over time. Since there is no agreement and method to evaluate the common adjustment factors, we aim to develop a new method while evaluating existing methods in this project. By using our method and multipliers, people can get more information from the public data for the top earnings in different demographic groups without jeopardizing confidentiality.

Other researchers have addressed this issue of top-coding and also used non-public data to

develop various alternative multipliers. We improve the method by carefully relating the relationship between the top code in the public data and the highest income individuals in the internal data. We further improve the multiplier approach by developing multipliers that are demographic and region specific. It is reasonable to expect that the earnings distribution varies across racial, gender, and education dimensions, and, even, across geographies. Comparison of the results from our methodology will be made to those proposed by others.

In existing literature, researchers are picking different methods to mitigate the top-coded issue and even when they are using a multiplier, they pick different values. This project is most valuable when we can objectively use the data to support our method against other methods. This requires a way to evaluate which method or which multiplier value is better. One of the best ways to assess the methods is to quantify on average, how good the estimated earnings generated by the method is different from the actual internal earning data. This is the most important part of this project and the internal Census Bureau data is essential to make it work. Therefore, in section 3.3.7, we will use public data to simulate the situation in the internal data and compare different methods before we have access to the internal data.

3.2 Literature Review

The Census Bureau only provides the censored earnings in the public data in order to protect confidentiality. Several multipliers are used to handle the top code issues. One is by David H. Autor, Lawrence F. Katz, and Melissa S. Kearney (AKK) (2008) and Autor and Dorn (2013) , and the other is by Lemieux (2006).

AKK (2008) used the factor 1.5 to adjust the top codes for both March CPS-Annual Social and Economic Supplement (CPS-ASEC) and CPS-Outgoing Rotation Groups (CPS-ORG). The same multiplier is also found in Katz and Murphy (1992) and the data they used is March CPS. In addition, Autor and Dorn (2013) adopted that top-coded yearly wages are multiplied by a factor of 1.5 for Decennial Census and American Community Survey (ACS). Lemieux (2006) adjusted the top-coded wages by a factor of 1.4 in CPS-Outgoing Rotation Groups (CPS-ORG). The same approach is also found in Beaudry, Green, and Sand (2016). Those papers do not give detailed reasons why they use those particular factors but our guess is that those factors are estimated by using Pareto distribution. That is, the Pareto distribution method assumes that the top level earnings follow

Pareto distribution and use the estimated parameter in the Pareto distribution to get the factor. The details will be discussed in the section 3. However, since there is no method to judge which factor is better or more reasonable, we want to develop a new method in our project.

Some authors use multipliers that are higher, time-varying, and differ by group. Hirsch and Macpherson (2015) show separate male and female Pareto estimates by year for 1973-2014 using public CPS-Merged Outgoing Rotation Groups (CPS-MORG) files. Values rise over time and are higher for men than women (for 2014, they are 2.06 for men and 1.81 for women). Certainly, it is interesting to explore the difference of multiplier between males and females. In this project, besides sex, we want to further estimate the multipliers for different race and regions. Moreover, we want to focus on March CPS, Decennial Census and ACS instead of CPS-MORG.

Recently, Armour, Burkhauser, and Larrimore (ABL) (2014)[1] use a continuous Maximum Likelihood estimator along with internal March CPS data and produce a series of estimates of top earnings in the CPS data. They found that previous approaches for imputing top-coded earnings systematically understate top earnings. Their estimates start with actual top earnings from the internal CPS data combined with a Pareto estimate using these data for internally censored observations.

However, our methodology and focus are different from theirs. First, the estimators they got are used for mean earning/wage for the whole population each year. We focus on the estimators for different demographical groups, for example, white and nonwhite. Since top code issues are closely related to white and non-white, male and female wage and earning inequality, the distinction of estimators between different demographical groups will be more useful for research. For example, if researchers are interested in the earnings inequality between male and female, then the difference of top codes adjustments for earnings in the male group and the female group will be very important.

In addition, for each demographic group, we will assume a different distribution according to the real data, not necessarily the Pareto distribution. Though ABL (2014) realized that the earnings distribution may not be Pareto far enough below the public-use top-code threshold to obtain reasonable estimates of the scale parameter, they used Pareto imputation for internally censored observations. In our methodology, we are the first to release the assumption that top earnings can be described by the Pareto distribution and simply use the mean square error method to get the estimator and evaluate how well these adjustment factors do. We will compare our

[1] A common limitation is their failure to account for non-responses and imputed earnings values. There are certainly some non-responses/missing of earnings in the public data, which may lead to inaccurate measure of wage inequality. But in the project, we will not deal with this issue.

methodology with others in section 3.3.

In the public March CPS, Decennial Census and ACS, mean earnings among top-coded workers are reported in some years. But we want to emphasize that mean value may be a good moment for some particular distributions (e.g. symmetric distribution) above the public top code. For different demographic groups and region groups, the distribution varies hugely. Therefore, the estimated adjustment factors multiplying the top codes varies across groups and is not necessarily mean earnings among top-coded individuals in most cases.

3.3 Methodology

3.3.1 Internal March CPS, Decennial Census and ACS

As Larrimore et al. (2008) mentioned, even the internal March CPS data, which are not subject to top coding, have been censored to various degrees over time, albeit well above those in the public files. In this project, we assume that the internal top-coded earnings are true earnings.[2] However, if RDC internal files provide Census information on means above the top codes, then we will treat those means as the true earnings. In addition, the internal files may use swapped values as top codes in recent years. In those cases, we will treat those swapped values as the true earnings. In general, we want use as much Census information as we can to deal with the internal top-coded earnings in the RDC internal files.

3.3.2 Calculating the Adjustment Factor

There is no consensus on the correct multiplier factor for the top codes. Possible values range from 1.4 to 1.6 according to the literature. Hence, a formal method to evaluate how well these adjustment factors do would be useful.[3]

Our methodology to calculate the adjustment factor is based on the mean square error. Suppose the true wage and salary income for each individual i in group g in year t is defined as Y_{igt}. The wage and salary income, which we can observe in the public data (e.g. IPUMS), is defined as I_{igt}. T_{igt} is a dummy variable and if $T_{igt} = 1$, then the wage and salary income for this individual in

[2]We would like to thank comments from ACRDC Review Board and Melissa Banzhaf.

[3]Our approach is not necessary the best approach. That is, the top code multipliers we estimate are not necessary the best multipliers. But we want to develop a new methodology, compare different multipliers based on our method and use the internal data. Further, we want check whether our new multipliers can tell a different story about the education premium in recent decades.

group g is censored. There are three groups (without interaction) we want to use to get estimator each year: Male and Female, White and Nonwhite, and Nine Census Regions.

The Y_{igt} and I_{igt} in the equation (3.1) below corresponds to WSAL-VAL, ERN-VAL, WS-VAL variables available in the March CPS data set, and QINCWG variable available in the Decennial Census and WAGP in the ACS. The statistical equation we want to estimate is as follows:

$$Min_{a_{gt}} \sum_{i} (Y_{igt} - \hat{y_{igt}})^2 \tag{3.1}$$

where $y_{igt} = (1 - T_{igt}) * I_{igt} + T_{igt} * a_{gt} I_{igt}$. In addition, the predicted value for annula earnings $\hat{y_{igt}} = I_{igt}$ if $T_{igt} = 0$, otherwise $\hat{y_{igt}} = \hat{a_{gt}} I_{igt}$ if $T_{igt} = 1$. Thus, we can get the estimator $\hat{a_{gt}}$. Comparing with the multiplier 1.4 or 1.5 or the adjusted top earnings in Armour, Burkhauser, and Larrimore (2014), we can check which adjustment is better by comparing the above mean square error. These estimated adjustment factors, and, especially if they differ across surveys, could be made available by Census to other researchers as an additional product accompanying the public use version of the data.

3.3.3 Using the Adjustment Factor to Assess Earnings and Wage Inequality

Consider the education premium for full-time, full-year male workers. Before we calculate the wage, the annual earnings variables (WSAL-VAL, ERN-VAL, WS-VAL; QINCWG; WAGP) need first to be adjusted since they are censored in the public data. Using our method, adjusted earnings can be obtained by multiplying the censored earnings with the adjusted factor for male each year we estimate. Then the adjusted weekly wage is calculated by using adjusted earnings divided by the weeks worked in the whole year.

Since we got the adjusted weekly wage for each individual, the mean wages of college group (i.e. individuals attended some college and those college graduates) and high school graduates group (i.e. individuals in this group have had at most high school diploma) can be calculated. Hence, relative mean wages of college graduates to high school graduates is defined as the education premium. Trend is defined as the growth of education premium.

Autor, Katz and Kearney (2008) shows that the education premium (wage inequality) increased rapidly in the 1980s, and continued to rise in the 1990s, but at a slower pace. They simply

used 1.5 to adjust the top codes for annual earnings in March CPS and CPS-ORG. However, if we use our estimated multiplier factor when replicating their paper, would the education premium still increase rapidly after 1980s? Or would the education premium still increase at a slower pace during 1990s? Since the overwhelming majority of top-coded observations are college educated individuals, the education premium will be sensitive to treatment of top-code multipliers. Hubbard (2011) identified and examined the biasing effect of top-codes on the college wage premium by using CPS public use files. He concluded that there has been essentially no gender difference in the college wage premium for at least a decade after identifying and correcting a bias in estimates of college wage premiums.

The advantage of our project is that we can further explore those questions by using internal ACS, March CPS and Decennial Census files. Though one can use public-use CPS files and easily examine the sensitivity of rate of return estimates with respect to changes in the multiplier, we do not know which multiplier fits the true rate of return best. Only access to the internal files can we get the true or accurate rate of return.

3.3.4 Methodology Comparison

In the literature, there are two main methods to estimate multipliers. We want to compare their methodologies to our methodology and use simulated results from public data to argue that our methodology is better in certain criteria.

The first method (Method 1) is based on the assumption that the distribution of top earnings can be described by the Pareto distribution in the public data. The common approach is to assume that the distribution is Pareto above some lower cutoff point (X_c) in the earning distribution and choose a second cutoff point above that pointtypically the topcode threshold itself (X_T) in the public data. In the literature, the lower cutoff point (X_c) is usually assumed at the 80th percentile of the earning distribution in the public data

The Pareto distribution is defined by the following cumulative distribution function:

$$P(X < x) = 1 - (\frac{X_c}{X})^{\alpha} \tag{3.2}$$

where x is a given value of earnings (weekly earning) larger than X_c, which is the scale or cutoff parameter given in the distribution, and α is the shape parameter of the distribution. Then the

mean above any threshold y is given as:

$$M(y) = (\frac{\alpha}{\alpha - 1})y \tag{3.3}$$

The Pareto shape parameter is then:

$$\hat{\alpha} = \frac{ln(\frac{C}{T})}{ln(\frac{X_T}{X_C})} \tag{3.4}$$

where C represents the number of individuals with earnings above the lower cutoff point (X_c) and T represents the number of individuals with earnings above the topcode threshold X_T, which is equal to y here.

The second method (Method 2) is explored by ABL (2014). They used the internal data and used Maximum Likelihood Hill estimator. Under this approach, the continuous, closed-form solution for estimating the Pareto parameter is:

$$\hat{\alpha} = \frac{M}{Tln(X_T) + \sum\limits_{X_m < X_i < X_T} ln(X_i) - (M+T)ln(X_m)} \tag{3.5}$$

where M is the number of individuals with earnings between the lower cutoff (X_m) and censoring point (X_T), T is the number of individuals with earnings at or over the topcode or censoring point, and X_i is the earnings of an individual. Using this formula allows individuals between the cutoff and censoring points to contribute to the CDF with their actual earnings, while those at the censoring point contribute to the CDF with the information that they have earnings at least as high as the censoring point.

According to those two methods, the adjustment factor or the multiplier is $\frac{\hat{\alpha}}{\hat{\alpha}-1}$.

3.3.5 Data

The Census Data we use are: March Current Population Survey (CPS) 1980-2015; Decennial Census 1970, 1980, 1990, 2000; American Community Survey (ACS) 2005-2014.

The data we want to use is wage and salary income (WSAL-VAL), earnings from longest job (ERN-VAL), earnings from other work including wage and salary earnings (WS-VAL) in the March CPS from 1980-2015. In addition, we will also use the wage and salary income (WAGP) in ACS from 2005-2014 and (QINCWG) in Census from 1970-2000. The same analysis will be performed

Table 3.1: Changes of Top codes in March CPS

	WSAL-VAL		ERN-VAL		WS-VAL	
Year/Eanings Variable	TCTV	VARW	TCTV	VARW	TCTV	VARW
Year 1980-1981	$50,000	$50,000				
Year 1982-1984	$75,000	$75,000				
Year 1985-1987	$99,999	$99,999				
Year 1988-1995			$99,997	$99,997	$99,997	$99,997
Year 1996-2002			$150,000	mean	$25,000	mean
Year 2003-2010			$200,000	mean	$35,000	mean
Year 2011			$250,000	swapped	$47,000	swapped
Year 2012-2015			$250,000	swapped	$50,000	swapped

Notes: *TCTV* indicates "Top-Coded Threshold Value". *VARW* indicates "values above replaced with". *means* indicates "mean earning among all top-coded workers (various among different demographic groups)" and *swapped* indicates "all earning values greater than or equal to the swap value were systematically swapped with other topcoded cases".

Table 3.2: Changes of Top codes in Decennial Census

	QINCWG	
Year/ Earnings Variable	Top-Coded Threshold Value	Values above top code replaced with
Year 1970	$50,000	$50,000
Year 1980	$75,000	$75,000
Year 1990	$140,000	Higher amounts are coded as the state means of values above the listed Top Code value
Year 2000	$175,000	Higher amounts are coded as the state means of values above the listed Top Code value

on both data sets, as they represent different surveys. The Census Bureau can learn from parallel analyses about differences that arise from different surveys.

Internal data is necessary for this project. Although there are different multiplier factors for top codes in the literature, without the internal data above the publicly top-coded values, we cannot evaluate those adjustment factors. Since we havent had access to the internal data yet, we want to treat the public data set as internal data set, then we compare the adjustment factors by using the three methodologies.

There are four main changes for the earnings top codes during 1980-2015 in March CPS, two main changes during 1970-2000 in Decennial Census and two main changes during 2005-2014 in ACS, which are described in the Tables 3.1, 3.2 and 3.3.

Table 3.3: Changes of Top codes in ACS

	WAGP	
Year/ Earnings Variable	Top-Coded Threshold Value	Values above top code replaced with
Year 2005-2014	99.5th Percentile in State	Higher amounts are coded as the state means of values above Top Code value

3.3.6 Project Output and Disclosure Risk

There are three outputs we want to obtain:

Estimated multiplier factors for (male, female), (white, nonwhite) and (nine census regions) each year by using our method. In addition, standard errors from estimation will be included in results. Those estimated multipliers will be available to other researchers. In addition, in order to compare different methods, we will report the estimated multiplier factors for the whole population by using method 1 and method 2.

Weighted sum of the square of residual error[4] each year (see equation (3.1)). For example, in the group male and female in 1990 March CPS, we will get two estimated adjustment factors: $\alpha_M 1990$, $\alpha_F 1990$ and two sum of the square of residual errors: $\sum_{i=1}^{p} \epsilon_{iM1990}^2$ and $\sum_{i=1}^{q} \epsilon_{iF1990}^2$, where p is the number of observations of male and q is the number of observations of female in the sample. Thus, the statistic, weighted sum of the square of residual error, we want to report is

$$\sum_{i=1}^{p} \epsilon_{iM1990}^2 + \sum_{i=1}^{q} \epsilon_{iF1990}^2 \tag{3.6}$$

The other weighted sum of the square of residual error we want to report is calculated by the common multiplier factor in the literature (e.g. 1.5 for all demographic groups) in equation (3.1):

$$\sum_{i=1}^{p+q} \mu_{i1990}^2 \tag{3.7}$$

According to our methodology, if (6) is smaller than (7), then our adjustment factors are better.

We are sensitive to the possibility that certain demographic groups may not be very heavily represented in the highest earnings levels, which means releasing a multiplier for that demographic group might be viewed as actually releasing the earnings for that very small group. The following Table 3.4, Table 3.5 and Table 3.6 show sample size with top-coded values for WSAL-VAL, ERN-

[4]Weight here means Census personal weight.

Table 3.4: Cell Size in Decennial Census

Demographic Group or Regional Group	Estimated Sample Size with top-coded values for QINCWG in public version of the Census
1970 Male; 1970 Female	1313; 68
1980 Male; 1980 Female	19978; 1290
1990 Male; 1990 Female	32687; 2840
2000 Male; 2000 Female	56793; 10282
1970 White; 1970 Nonwhite	1357; 24
1980 White; 1980 Nonwhite	20248; 1020
1990 White; 1990 Nonwhite	33383; 2144
2000 White; 2000 Nonwhite	59584; 7491
1970 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	110;330;290;83;176;50;108;43;191
1980 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	1296;3822;3879;1323;3085;888;2241;910;3825
1990 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	2705;7636;5099;1603;5916;1289;3020;1364;6895
2000 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	5076;12947;9215;2985;12081;2672;5880;3463;12755

Notes: Nine Census Regions are New England; Middle Atlantic; East North Central; West North Central; South Atlantic; East South Central; West South Central; Mountain; Pacific.

Table 3.5: Cell Size in ACS

Demographic Group or Regional Group	Estimated Sample Size with top-coded values for WAGP in public version of the Census
2005 Male; 2005 Female	11503; 1756
2006 Male; 2006 Female	11756; 1826
2007 Male; 2007 Female	12009;2007
... ...	... ...
2005 White; 2005 Nonwhite	12315; 217
2006 White; 2006 Nonwhite	12516; 235
2007 White; 2007 Nonwhite	12783; 291
... ...	... ...
2005 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	642;1746;1891;756;2803;849;1589;918;2065
2006 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	682;1812;1977;744;2771;860;1475;1015;2246
2007 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	676;1914;2133;779;2928;921;1501;972;2192
... ...	... ...

Notes: Nine Census Regions are New England; Middle Atlantic; East North Central; West North Central; South Atlantic; East South Central; West South Central; Mountain; Pacific.

VAL, WS-VAL; QINCWG; WAGP in public version of the March CPS, Census, and ACS. These tables indicate that, for most groups, cell size wont be a substantial concern. However, we will work with the CES Disclosure Officer to ensure that all released statistics will be based on sufficiently large cell sizes to ensure confidentiality and will be modified if any disclosure concerns are raised.

It may be a concern that at high levels of earnings, the sample sizes for groups in March CPS are too thin to provide reliable estimates. Such cuts of the data are more feasible in the decennial Censuses and ACS. In order to deal with those concerns, if the sample sizes are too small in the March CPS, we will merge those groups to a bigger group and then estimate the multiplier. For example, the sample size for nonwhite in 1980 is only 26. Therefore, we will merge the nonwhite and white in 1980 to a bigger group to make the estimate more reliable.

Table 3.6: Cell Size in March CPS

Demographic Group or Regional Group	Estimated Sample Size with top-coded values for WSAL-VAL in public version of the Census
1980 Male; 1980 Female	738; 34
1981 Male; 1981 Female	962; 41
... ...	... ...
1980 White; 1980 Nonwhite	746; 26
1981 White; 1981 Nonwhite	958; 45
... ...	... ...
1980 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	73; 102; 105; 64; 106; 27; 64; 80; 151
1981 NE; MA; ENC; WNC; SA; ESC; WSX; MO; PA	79; 129; 150; 86; 151; 24; 68; 110; 206
... ...	... ...

Notes: Nine Census Regions are New England; Middle Atlantic; East North Central; West North Central; South Atlantic; East South Central; West South Central; Mountain; Pacific.

Table 3.7: Multipliers for Male/Female in 1980 March CPS

1980 March CPS	Method 1	WS of SRE 1	Method 2	WS of SRE 2	Our Method	WS of SRE
Male	1.50	3.63E+14	1.39	1.83E+14	1.16	3.19E+13
Female	1.57	1.29 E+14	1.43	9.10E+13	1.28	7.59E+13
Male+Female	1.60	5.16E+14	1.45	2.45E+14		1.08E+14

Notes: $WSofSRE$ means "weighted sum of the square of residual error".

3.3.7 Simulation Results

Since the second method and our method are based on the internal data and we have not had access to the internal data yet, we want to treat the public data set as internal data set, which we call it hypothetical internal data set, and get the estimators according to the three methods. Then we can compare those different multipliers by comparing the mean square error in equation (3.1).

The simulation idea is showed in the Figure 3.1. The line represents the earning distribution in a particular demographic group in a particular year.

Based on the above simulation idea, we can get multipliers based on three different methods. We only use 1980 March CPS Male/Female and White/Black as an example.

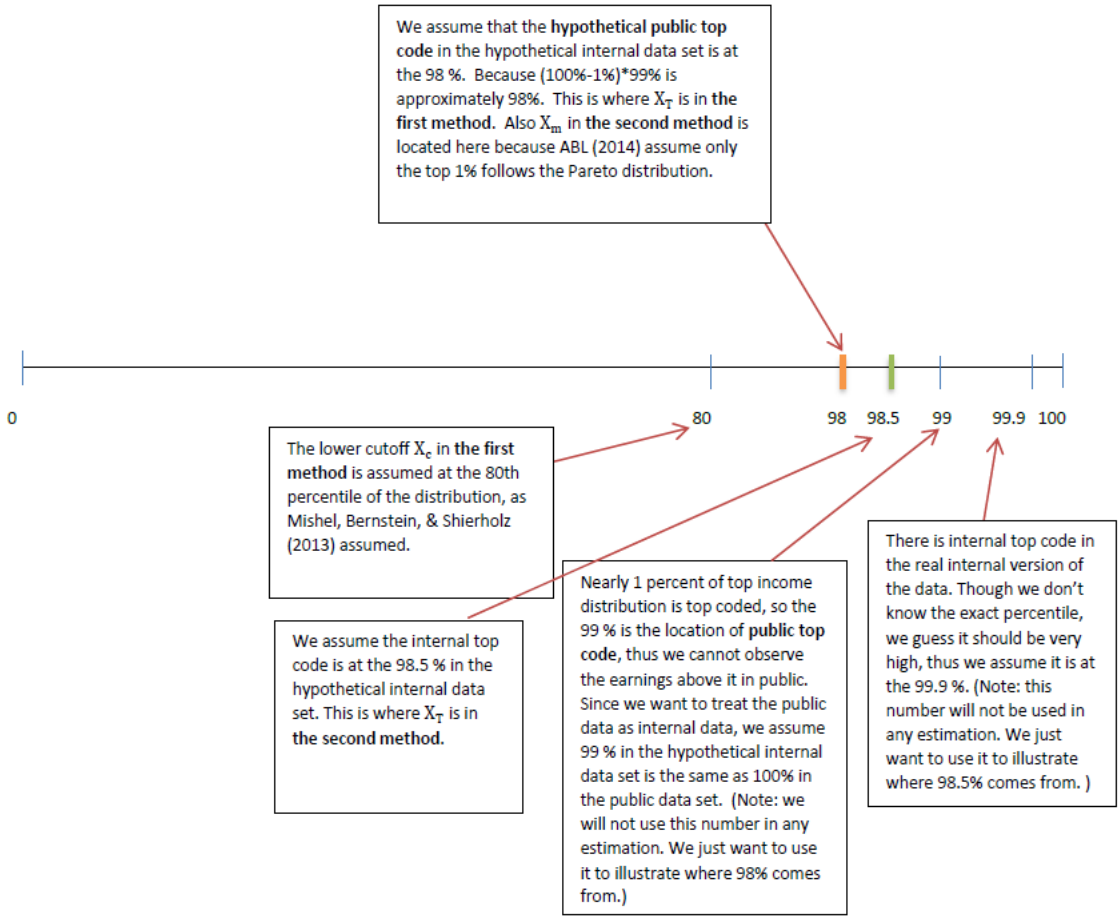
Table 3.7 and Table 3.8 show that using our method, not only we can get different multipliers

Table 3.8: Multipliers for White/Black in 1980 March CPS

1980 March CPS	Method 1	WS of SRE 1	Method 2	WS of SRE 2	Our Method	WS of SRE
White	1.61	5.58E+14	1.38	1.69E+14	1.25	1.13E+13
Black	1.57	3.82E+14	1.24	1.55E+13	1.23	1.55E+13
White+Black	1.61	5.37E+14	1.41	2.02E+14		1.29E+14

Notes: $WSofSRE$ means "weighted sum of the square of residual error".

Figure 3.1: Simulation Idea



for different demographic groups without assuming Pareto distribution, but our weighted sum of the square of residual errors are also the smallest. More importantly, the multipliers obtained from our method are smaller than multipliers obtained from methods in literature. If our multipliers are used to adjust topcodes, the income inequality within various demographic groups is not as severe as the literature indicates.

3.4 Conclusion

The primary purpose of this project is to evaluate the concepts and practices underlying the dissemination of top-coded earnings variables in a variety of Census products, such as the American Community Survey, the March Current Population Survey and the Decennial Census. The project

illustrates the implications of this practice of top-coding for understanding the differences in wage distributions across demographic and social characteristics of the top-coded population.

Adjustment multipliers, like 1.4 or 1.5, are used for the censored earnings in the literature. We improve on this research, which focuses on producing multipliers for the whole population, by developing multipliers that are demographic and region specific. It is not unreasonable to expect that the earnings distribution varies across racial, gender, and education dimensions, and even across geographies. In future, we hope to have access to the internal data to estimate the multipliers without creating a simulating surrounding.

Appendices

Appendix A Imputation of Risk Aversion

In the theoretical model, one of the key variables is the individual specific coefficient of risk aversion θ. Kimball, Sahm, and Shapiro (2008) have developed direct survey measures of risk tolerance based on hypothetical choices and appropriate econometric techniques for dealing with the inevitable measurement error in questionnaires. Kimball, Sahm, and Shapiro (2009) also present the risk tolerance imputations for the survey responses in the PSID. I follow their imputation methods based on the gambling questions in the 1996 PSID. The questions are as follows:

Suppose you had an occupation that guaranteed you income for life equal to your current total income. Also suppose that occupation was your/your family's only source of income. Then you are given the opportunity to take a new, and equally good, occupation with a 50-50 chance that it will double your income and spending power. But there is a 50-50 chance that it will cut your income and spending power by a third. Would you take the new occupation?

Individuals who answered that they would take this risky occupation were then asked about a riskier occupation:

Now, suppose the chances were 50-50 that the new occupation would double your/your family's income and 50-50 that it would cut it in half. Would you still take the new occupation?

In contrast, individuals who would not take the initial risky occupation were asked about a less risky occupation:

Now, suppose the chances were 50-50 that the new occupation would double your/your family's income and 50-50 that it would cut it by 20%. Then, would you take the new occupation?

Conditional on their first two responses, individuals were asked to consider a risky occupation with either a 75% downside risk or a 10% downside risk. These responses allow us to order individuals into six categories. Kimball, Sahm, and Shapiro (KSS) (2009) used maximum likelihood estimation and then impute the conditional expectation of risk aversion for each category c: θ_c.

Table 9: Imputation of Risk Preference in KSS (2009)

Response Category	Risk Tolerance	Risk Aversion
1	0.27	6.7
2	0.40	4.2
3	0.49	3.5
4	0.60	2.8
5	0.79	2.2
6	1.22	1.4

NOTE: This table comes from part of Table 1-Risk Tolerance in the PSID in Kimball, Sahm, and Shapiro (2009).

Thus, only one risk aversion in each category is available in their paper, which is shown in Table 3. However, I assume there is a continuous distribution of risk aversion in the whole population in the theoretical model, and I will make a further imputation based on their results. Guiso and Paiella (2005) show that aversion to losses is less pronounced among people with higher levels of education. Thomas, Armin, Huffman, and Uwe (2010) propose that exogenous personal characteristics, like age and gender, determine an individual's risk aversion. Thus, I use an individual's number of years of education (e), age (a), $age^2/100$ ($a^2/100$), male (m), children (c), and race (r) to predict their risk aversion based on the available six levels of risk aversion. The whole sample I use for my prediction is composed of employed people between 20 and 60 years old, whose population weights are nonzero. The imputation model is as follows:

$$\theta_c = \beta_0 + \beta_1 e + \beta_2 a + \beta_3 a^2/100 + \beta_4 m + \beta_5 c + \beta_6 m * c + \vec{\beta_7} r + \epsilon,$$

where $e \in \{1, 2, ..., 17\}, a \in \{20, 21, ..., 69\}, m \in \{1 = Male, 0 = Female\}, c \in \{1 = HaveChildren, 0 = NoChildren\}$, r is a vector of dummies for race and $\vec{\beta_7}$ is a vector of coefficients.

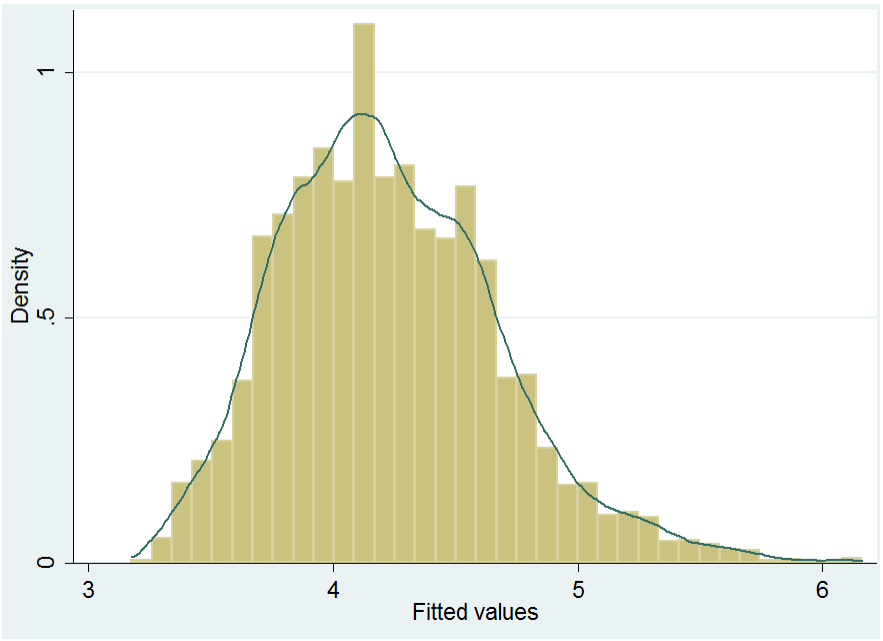
Based on the above model, I run regressions and get the predicted value of θ_c on each category (e, a, m, c, $m * c$, r). Therefore, there will be a distribution of θ_c among the population. Table 4 reports the regression results for the whole sample and for subsamples of single and married individuals. Figure 14 shows that the shifted lognormal distribution of risk aversion among populations fits the data pretty well.

Table 10: Regression of θ_c on Personal Characteristics

	All	Single	Married
Years of education	-0.092***	-0.060**	-0.093***
	(0.012)	(0.029)	(0.015)
Age	-0.059***	-0.039	-0.034
	(0.022)	(0.053)	(0.031)
$Age^2/100$	0.112***	0.087	0.082**
	(0.027)	(0.072)	(0.038)
Male	-0.318***	-0.413***	-0.172*
	(0.072)	(0.134)	(0.103)
Children	0.040	0.015	0.039
	(0.077)	(0.206)	(0.105)
Male*children	0.025	0.638**	-0.110
	(0.103)	(0.319)	(0.134)
Black	0.414***	0.280*	0.325**
	(0.083)	(0.157)	(0.134)
American Indian, Aleut, Eskimo	0.317*	0.494	0.441*
	(0.186)	(0.466)	(0.249)
Asian, Pacific Islander	-0.137	-1.003	-0.083
	(0.476)	(0.876)	(0.575)
Hispanic	-0.185	0.010	0.126
	(0.414)	(0.898)	(0.534)
Other race	0.291	0.200	0.424
	(0.210)	(0.468)	(0.275)
Constant	5.932***	5.004***	5.394***
	(0.422)	(0.965)	(0.611)
Observations	4747	949	2793

NOTE: A larger value of θ implies more risk aversion. Significant codes: 0.01 '***' 0.05 '**' 0.1 '*'.

Figure 2: The Distribution of Risk Aversion



Note: The x-axis indicates the coefficient of relative risk aversion. A higher number means a worker is more risk averse.

Appendix B Individual Consumption

The pre-tax wage for one unit of effective labor in sector 1 or sector 2 is

$$w_1 = \alpha_1\beta_1 L_1^{\beta_1-1} = \alpha_1\beta_1\Big[n(1-h(\theta_m))\int_0^{+\infty}\eta_{i1}dF(\eta_{i1})\Big]^{\beta_1-1},$$
$$w_2 = \alpha_2\beta_2 L_2^{\beta_2-1} = \alpha_2\beta_2\Big[nh(\theta_m)\int_0^{+\infty}\eta_{i2}dF(\eta_{i2})\Big]^{\beta_2-1}.$$

The above equations are identical to

$$w_1 = \alpha_1\beta_1 L_1^{\beta_1-1} = \alpha_1\beta_1\Big[n(1-h(\theta_m))\Big]^{\beta_1-1},$$
$$w_2 = \alpha_2\beta_2 L_2^{\beta_2-1} = \alpha_2\beta_2\Big[nh(\theta_m)\Big]^{\beta_2-1}.$$

If each worker is paid a wage equal to her productivity, then the wage for individual i in sector 1 or sector 2 is

$$w_{i1} = \eta_{i1}\alpha_1\beta_1\Big[n(1-h(\theta_m))\Big]^{\beta_1-1},$$
$$w_{i2} = \eta_{i2}\alpha_2\beta_2\Big[nh(\theta_m)\Big]^{\beta_2-1}.$$

Therefore, the consumption for an individual who chooses to work in sector 1 or sector 2 is

$$c_{i1} = \eta_{i1}\alpha_1\beta_1(n(1-h(\theta_m)))^{\beta_1-1},$$
$$c_{i2} = \eta_{i2}\alpha_2\beta_2(nh(\theta_m))^{\beta_2-1} - \gamma,$$

where γ is the cost of accumulating human capital (i.e., the cost of higher productivity α_2).

Appendix C Constant Expected Tax Revenue Condition

C.1 Expected Tax Revenue in the Flat Tax Regime

Under a flat tax regime, the expected tax revenue collected from workers is

$$E_w[R_\tau] = \underbrace{\tau(n(1-h(\theta_{m,F}))) \int_0^{+\infty} \left[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,F})))^{\beta_1-1}\right] \frac{1}{\eta_{i1}\sigma_1\sqrt{2\pi}} e^{\frac{-((log(\eta_{i1})-\mu_1)^2)}{2\sigma_1^2}} d\eta_{i1}}_{\text{total tax revenue collected in sector 1}}$$

$$+ \underbrace{\tau(nh(\theta_{m,F})) \int_0^{+\infty} \left[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,F}))^{\beta_2-1}\right] \frac{1}{\eta_{i2}\sigma_2\sqrt{2\pi}} e^{\frac{-((log(\eta_{i2})-\mu_2)^2)}{2\sigma_2^2}} d\eta_{i2}}_{\text{total tax revenue collected in sector 2}}$$

where $\mu_1 = \frac{-\sigma_1^2}{2}$ and $\mu_2 = \frac{-\sigma_2^2}{2}$.

C.2 Expected Tax Revenue in the Progressive Tax Regime

Under a progressive tax regime, the total expected tax revenue collected from workers is

$$E_w[R_\rho] = \underbrace{n(1-h(\theta_{m,P})) \int_0^{+\infty} a\left[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,P})))^{\beta_1-1}\right] + \rho\left[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,P})))^{\beta_1-1}\right] log\left[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,P})))^{\beta_1-1}\right] f(\eta_{i1}) d\eta_{i1}}_{\text{total tax revenue collected in sector 1}}$$

$$+ \underbrace{nh(\theta_{m,P}) \int_0^{+\infty} a\left[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,P}))^{\beta_2-1}\right] + \rho\left[\eta_{i2}\alpha_2\beta_2(n(1-h(\theta_{m,P})))^{\beta_2-1}\right] log\left[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,P}))^{\beta_2-1}\right] f(\eta_{i2}) d\eta_{i2}}_{\text{total tax revenue collected in sector 2}}.$$

C.3 Relationship

The tax revenue neutral condition implies $E_w[R_\tau] = E_w[R_\rho]$. Therefore, the relationship between τ and $(\rho, a)$is

$$\tau^* = \frac{E_w[R_\rho]}{n(1-h(\theta_{m,F})) \int_0^{+\infty} \left[\eta_{i1}\alpha_1\beta_1(n(1-h(\theta_{m,F})))^{\beta_1-1}\right] f(\eta_{i1}) d\eta_{i1} + nh(\theta_{m,F}) \int_0^{+\infty} \left[\eta_{i2}\alpha_2\beta_2(nh(\theta_{m,F}))^{\beta_2-1}\right] f(\eta_{i2}) d\eta_{i2}}.$$

where $f(\eta_{i1}) = \frac{1}{\eta_{i1}\sigma_1\sqrt{2\pi}} e^{\frac{-((log(\eta_{i1})-\mu_1)^2)}{2\sigma_1^2}}$, $f(\eta_{i2}) = \frac{1}{\eta_{i2}\sigma_2\sqrt{2\pi}} e^{\frac{-((log(\eta_{i2})-\mu_2)^2)}{2\sigma_2^2}}$, $\mu_1 = \frac{-\sigma_1^2}{2}$, and $\mu_2 = \frac{-\sigma_2^2}{2}$.

In addition, when the value of ρ is changed, the value of a needs to be changed to match $E_w[R_\rho]$ given the initial value of ρ and a.

Appendix D Solution Algorithm

The computational procedure used to solve the equilibrium allocations and further welfare calculations in different progressive tax regimes can be represented by the following algorithm:

1) Given all α_1, α_2, β_1, β_2, σ_1^2, σ_2^2, a_1 and ρ_1 (i.e., $a_1 = 0.23$ and $\rho_1 = 0$), the equilibrium allocation of workers $h(\theta_{m,P1})$ in the progressive tax regime can be solved through equation (6).

2) Once $h(\theta_{m,P1})$ is solved, the related tax revenue, $E_w[R_{\rho 1}]$, can be calculated.

3) Then I increase the value from ρ_1 to ρ_2 to make the tax system more progressive. I define an interval of $a_2 \in [\underline{a_2}, \overline{a_2}]$. For every possible value of a_2, I find the new equilibrium, $h(\theta_{m,P2})$, and then $E_w[R_{\rho 2}]$. I search the value of a_2 in the interval to minimize the distance between $E_w[R_{\rho 2}]$ and $E_w[R_{\rho 1}]$. Therefore, I could find the matched a_2 and unchanged ρ_2.

4) The compensating consumption for individuals can be calculated based on the above values, and total welfare change can be further analyzed.

5) I further increase ρ_2 to ρ_3 and do the procedures 3) and 4) again.

Bibliography

[1] Daron Acemoglu. Technical change, inequality, and the labor market. *Journal of economic literature*, 40(1):7–72, 2002.

[2] Philip Armour, Richard V Burkhauser, and Jeff Larrimore. Using the pareto distribution to improve estimates of topcoded earnings. *Economic Inquiry*, 54(2):1263–1273, 2016.

[3] David H Autor, Lawrence F Katz, and Melissa S Kearney. Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323, 2008.

[4] Daniel Barth, Stephen H Shore, and Shane T Jensen. Identifying idiosyncratic career taste and skill with income risk. *Quantitative Economics*, 8(2):553–587, 2017.

[5] Paul Beaudry, David A Green, and Benjamin M Sand. The great reversal in the demand for skill and cognitive tasks. *Journal of Labor Economics*, 34(S1):S199–S247, 2016.

[6] Robin Boadway and Motohiro Sato. Optimal income taxation with uncertain earnings: A synthesis. 2011.

[7] Holger Bonin, Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Economics*, 14(6):926–937, 2007.

[8] James N Brown and Harvey S Rosen. Taxation, wage variation, and job choice. *Journal of Labor Economics*, 5(4, Part 1):430–451, 1987.

[9] David Card and Alan B Krueger. Does school quality matter? returns to education and the characteristics of public schools in the united states. *Journal of political Economy*, 100(1):1–40, 1992.

[10] Pedro Carneiro and Sokbae Lee. Trends in quality-adjusted skill premia in the united states, 1960-2000. *American Economic Review*, 101(6):2309–49, 2011.

[11] Marco Cozzi. Risk aversion heterogeneity, risky jobs and wealth inequality. Technical report, Queen's Economics Department Working Paper, 2011.

[12] German Cubas and Pedro Silos. Progressive taxation and risky career choices. 2015.

[13] H David and David Dorn. The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–97, 2013.

[14] Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3):1238–60, 2010.

[15] David Dorn. *Essays on inequalitiy, spatial interaction, and the demand for skills*. na, 2009.

[16] Jonathan Eaton and Harvey S Rosen. Taxation, human capital, and uncertainty. *The American Economic Review*, 70(4):705–715, 1980.

[17] Martin S Eichenbaum, Lars Peter Hansen, and Kenneth J Singleton. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. *The Quarterly Journal of Economics*, 103(1):51–78, 1988.

[18] Didier Fouarge, Ben Kriechel, and Thomas Dohmen. Occupational sorting of school graduates: The role of economic preferences. *Journal of Economic Behavior & Organization*, 106:335–351, 2014.

[19] Eric French. The effects of health, wealth, and wages on labour supply and retirement behaviour. *The Review of Economic Studies*, 72(2):395–427, 2005.

[20] Manja Gärtner, Johanna Mollerstrom, and David Seim. Individual risk preferences and the demand for redistribution. *Journal of Public Economics*, 153:49–55, 2017.

[21] Edward L Glaeser and Matthew G Resseger. The complementarity between cities and skills. *Journal of Regional Science*, 50(1):221–244, 2010.

[22] Luigi Guiso and Monica Paiella. The role of risk aversion in predicting individual behaviors. 2004.

[23] Nezih Guner, Remzi Kaygusuz, and Gustavo Ventura. Taxation and household labour supply. *The Review of economic studies*, 79(3):1113–1149, 2011.

[24] Nezih Guner, Remzi Kaygusuz, and Gustavo Ventura. Taxing women: A macroeconomic analysis. *Journal of Monetary Economics*, 59(1):111–128, 2012.

[25] Nezih Guner, Remzi Kaygusuz, and Gustavo Ventura. Income taxation of us households: Facts and parametric estimates. *Review of Economic Dynamics*, 17(4):559–581, 2014.

[26] Fatih Guvenen, Burhanettin Kuruscu, and Serdar Ozkan. Taxation of human capital and wage inequality: A cross-country analysis. *Review of Economic Studies*, 81(2):818–850, 2013.

[27] Marcus Hagedorn, Tzuo Hann Law, and Iourii Manovskii. Identifying equilibrium models of labor market sorting. *Econometrica*, 85(1):29–65, 2017.

[28] Arnold C Harberger. Three basic postulates for applied welfare economics: An interpretive essay. *Journal of Economic literature*, 9(3):785–797, 1971.

[29] Jonathan Heathcote, Kjetil Storesletten, and Giovanni L Violante. Optimal tax progressivity: An analytical framework. *The Quarterly Journal of Economics*, 132(4):1693–1754, 2017.

[30] James Heckman, Anne Layne-Farrar, and Petra Todd. Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *The Review of Economics and Statistics*, pages 562–610, 1996.

[31] Barry T Hirsch and David A Macpherson. *Union Membership and Earnings Data Book...: Compilations from the Current Population Survey*. Bureau of National Affairs, 2015.

[32] William HJ Hubbard. The phantom gender difference in the college wage premium. *Journal of Human Resources*, 46(3):568–586, 2011.

[33] Lawrence F Katz and Kevin M Murphy. Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics*, 107(1):35–78, 1992.

[34] Miles S Kimball, Claudia R Sahm, and Matthew D Shapiro. Imputing risk tolerance from survey responses. *Journal of the American statistical Association*, 103(483):1028–1038, 2008.

[35] Miles S Kimball, Claudia R Sahm, and Matthew D Shapiro. Risk preferences in the psid: individual imputations and family covariation. *American Economic Review*, 99(2):363–68, 2009.

[36] Thomas J Kniesner and James P Ziliak. Tax reform and automatic stabilization. *American Economic Review*, 92(3):590–612, 2002.

[37] Finn E Kydland and Edward C Prescott. Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, pages 1345–1370, 1982.

[38] Jeff Larrimore, Richard V Burkhauser, Shuaizhang Feng, and Laura Zayatz. Consistent cell means for topcoded incomes in the public use march cps (1976–2007). *Journal of Economic and Social Measurement*, 33(2, 3):89–128, 2008.

[39] Thomas Lemieux. Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review*, 96(3):461–498, 2006.

[40] Benjamin B Lockwood and Matthew Weinzierl. De gustibus non est taxandum: Heterogeneity in preferences and optimal redistribution. *Journal of Public Economics*, 124:74–80, 2015.

[41] Hamish Low and Daniel Maldoom. Optimal taxation, prudence and risk-sharing. *Journal of Public Economics*, 88(3-4):443–464, 2004.

[42] Robert E Lucas and Lucas. *Models of business cycles*, volume 26. Basil Blackwell Oxford, 1987.

[43] James A Mirrlees. An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2):175–208, 1971.

[44] Brooks Pierce. Compensation inequality. *The Quarterly Journal of Economics*, 116(4):1493–1525, 2001.

[45] Casey Rothschild and Florian Scheuer. Redistributive taxation in the roy model. *The Quarterly Journal of Economics*, 128(2):623–668, 2013.

[46] Dominik Sachs, Aleh Tsyvinski, and Nicolas Werquin. Nonlinear tax incidence and optimal taxation in general equilibrium. Technical report, National Bureau of Economic Research, 2016.

[47] Emmanuel Saez. Using elasticities to derive optimal income tax rates. *The review of economic studies*, 68(1):205–229, 2001.

[48] Stefanie Stantcheva. Optimal taxation and human capital policies over the life cycle. *Journal of Political Economy*, 125(6):1931–1990, 2017.

[49] James H Stock and Motohiro Yogo. Vtesting for weak instruments in linear iv regressionv. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,(New York: Cambridge University Press, 2005)*, 2005.

[50] Coen N Teulings. The wage distribution in a model of the assignment of skills to jobs. *Journal of political Economy*, 103(2):280–315, 1995.

[51] Richard Thaler and Sherwin Rosen. The value of saving a life: evidence from the labor market. In *Household production and consumption*, pages 265–302. NBER, 1976.

[52] Hal R Varian. Redistributive taxation as social insurance. *Journal of public Economics*, 14(1):49–68, 1980.

[53] John V Winters. Differences in employment outcomes for college town stayers and leavers. *IZA Journal of Migration*, 1(1):11, 2012.