Factors Affecting College Completion and Student Ability in the U.S. since 1900*

Kevin Donovan   Christopher Herrington

University of Notre Dame   University of South Alabama

October 2013

PRELIMINARY AND INCOMPLETE - PLEASE DO NOT CITE.
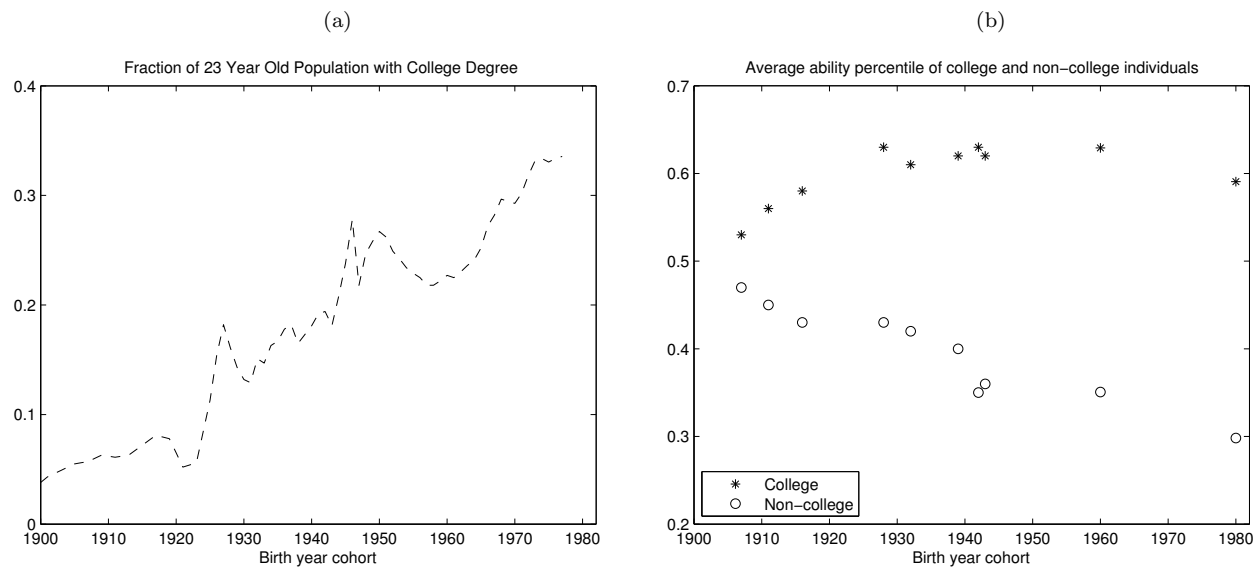
Abstract

We develop a dynamic lifecycle model to study the increases in college completion and average IQ of college students in cohorts born from 1900 to 1972. The model is discplined in part by constructing a historical time series on real college costs from printed government documents covering this time period. The main result is that the model captures nearly all of the increase in attainment from 1900 to 1950, and half is accounted for by a decrease in college costs during this period. The rise in average college student IQ cannot be accounted for without a decrease in the variance of ability signals, which we attribute to the rise in standardized testing.

1 Introduction

The twentieth century saw a dramatic expansion of higher education in the United States. While less than four percent of the 1900 birth cohort held a bachelor's degree by age 23, that share increased to over thirty percent by the 1970 cohort. Panel (a) of Figure 1 plots this rise for all cohorts born from 1900 through 1977.[1] Concurrent with the increase in college attendance, the gap in measured cognitive ability widened substantially between college students and those individuals who ended their formal education after high school graduation ("non-college" individuals). This pattern is seen in Panel (b) of Figure 1, which plots the average IQ percentile (our proxy for "ability") of college and non-college individuals.[2] For example the average college student born in 1907 had an IQ in the fifty-third percentile, very close to the average non-college individual whose IQ was in the forty-seventh percentile. Yet over the next several decades, the average IQ percentile increased among college enrollees and decreased among those with only a high school degree. This trend of increased ability sorting occurred even as the share of students attempting college grew steadily larger.

Figure 1: College Completion and Average Student Ability in the U.S. since 1900

The goal of this paper is to understand the causes of these two empirical trends. This task is complicated by the vast number of changes in both the aggregate economy and education sector over this time period, as well as a lack of reliable data particularly during the first half of the century. We combat this by developing an overlapping generations lifecycle model which allows for a number of potential explanations to coexist within the model. First, each cohort is populated by an exogenous number of male and female high school graduates, since the changing sex composition of graduates may potentially play a role (e.g. Goldin and Katz, 2010). These graduates then immediately choose whether or not to enter college, subject to endowments. Endowments, which are financial assets and ability to complete a college degree, are heterogenous. Moreover, ability is not perfectly known. Consistent with evidence in Cunha et al. (2005), individuals observe only a noisy signal about true ability, which introduces a measure of uncertainty when combined with the limited borrowing opportunities for college. These features of the model allow us to quantitatively compare a number of potential explanations within the framework of the model.

We then calibrate the model and assess its ability to match the facts presented in Figure 1. Naturally, the predictions of the model rely heavily on the costs and benefits associated with attending, and we therefore take care to measure them properly. The most obvious cost is the explicit tuition requirements of attending college, which takes on some added importance in the presence of limited borrowing opportunities. One contribution of this paper is the construction of time series data for real average college costs. We construct this measure by combining data from a series of printed government documents dating back to 1900. This allows us to feed in an accurate time series of college tuition costs, instead of relying on assumptions such as a constant tuition growth rate. The opportunity cost of attending college, along with the benefit of attaining a degree, is computed by estimating wage profiles from the 1940 to 2000 U.S. Censuses and 2006-2010 American Community Surveys. The wage profiles are allowed to depend on sex, age, and education to accurately capture the changing education earnings premia in the United States. These time series are exogenously fed into the model to capture the relevant tradeoffs faced by high school graduates when choosing to enter college.

After incorporating these time series data, we are left to calibrate the time-invariant parameters of the model. One possibility would be to calibrate these parameters to exactly match statistics from the 1900 birth cohort, feed in the relevant time series, and assess the model's ability to match attainment and sorting statistics. The lack of data from this period eliminates this possibility. Instead, we use more recent and reliable data from the National Longitudinal Survey of Youth and High School & Beyond Survey to calibrate the model to match college attainment and ability sorting statistics for the 1962 birth cohort. Taking these structural parameters as fixed, we then simulate the model beginning with the 1900 birth cohort (i.e., the 1918 high school graduation cohort) and feed in the constructed sex-specific high school graduate rates, college costs, and life-cycle earnings profiles to assess the model's ability to replicate enrollment and average ability changes over time.

The main result is that the model predicts almost the entire increase in college completion from the 1900 to 1950 cohort, even though it is calibrated only to match 1962 statistics exactly. Moreover, outside of over-predicting attainment during World War II, the time series tracks the data extremely well during this time period. A simple accounting decomposition of the college completion rate shows that the increase is driven almost exclusively by a combination of two forces: (1) an increase in the fraction of individuals completing high school and (2) an increase in the fraction of individuals enrolling in college *conditional* on high school graduation. The former is exogenous to our model, and has been previously emphasized by Goldin and Katz (2010). We find that the increase in the high school completion rate accounts for approximately fifty percent of the attainment increase. The other half of the increase is accounted for by the increase in the enrollment rate, which is an endogenous object in the model. This endogeneity leads us to a series of counterfactual experiments to understand the underlying forces driving the enrollment increase. Our counterfactual results show that the increase in college enrollment, conditional on high school graduation, is almost exclusively driven by a decrease in real college costs (relative to income) for these cohorts. Combining the decomposition with our counterfactual results implies that the decrease in college costs accounts for almost fifty percent of the increase in college completion rates for cohorts born 1900 to 1950.

When we consider cohorts born after 1950, the model significantly overpredicts college

completion rates. This is a common issue (e.g. Card and Lemieux, 2001), and is due to the large, perfectly forecasted increase in the college earnings premia around 1950. While we over predict college attainment, we still match college enrollment during this time period. We over-predict the fraction of college enrollees who graduate college, suggesting that theories that include time-varying cohort ability may be a promising channel in understanding post-1950 education decisions.

We lastly turn to the ability of the model to capture increased ability sorting over time. Again through a series of counterfactual experiments, we consistently find that changes in economic factors (i.e., earnings premia, college costs, opportunity costs, and asset endowments) have little impact on ability sorting. Instead, the key feature in the model that accounts for this phenomenon is uncertainty about ability. We show that a decrease in the variance of ability signals can generate an increase in ability sorting similar to that in the data. We attribute this change to the increases in standardized testing which improved students knowledge of their own ability relative to other students in their cohort, as discussed in Hoxby (2009).

This paper is primarily related to the literature on college attainment in the United States over time, which includes recent work by Garriga and Keightley (2007) and Hendricks and Leukhina (2011). More closely related to our work, Castro and Coen-Pirani (2012) ask whether educational attainment over time can be explained by changes in the college earnings premia. Their complete markets model underpredicts college attainment for pre-1950 cohorts, while the combination of limited borrowing with our college tuition time series allows us to match it quite well. Hendricks and Schoellman (2012) study early 1900 college attainment and ability sorting, but take college completion and student ability as given in order to understand the changes in the college earnings perium in a complete markets model. By contrast, we seek to understand the economic factors that affected college completion and average student ability. As it relates to post-1950 cohorts, the aforementioned Castro and Coen-Pirani (2012) show that cross-cohort variation in learning ability can alleviate the overprediction of college attainment. Keller (2013), on the other hand, develops a model of college attendence and quality choice, and points to the slowdown of human capital rental rates as the cause of the post-1950 slowdownin attainment. Lochner and Monge-Naranjo

(2011) point to student loan policies with limited commitment as the driving force behind post-World War II student ability sorting. Their focus on the latter half of the twentieth century forces us to exclude the student loan innovations they consider, since Figure 1b shows that a majority of the sorting occurred in cohorts before 1950.

The rest of this paper proceeds as follows. Section 2 begins by decomposing college attainment growth into changes in high school graduation rates and other factors. Section 3 then lays out a life-cycle model of college decisions which we use to discipline our analysis. Section 4 discusses the calibration, including our construction of historical college costs. Section 5 lays out the results of the quantitative analysis, while Section 6 considers the robustness of the main results to different specifications. Finally, Section 7 concludes.

2   Accounting for College Attainment

The model developed in Section 3 includes cohorts of high school (HS) graduates who are fed into the model exogenously. Our focus is on post-HS graduation factors such as college costs and earnings premia, which affect college attendance decisions. We motivate our interest in these post-HS graduation factors with an accounting exercise to decompose the factors driving the increase in college attainment over our sample. Our measure of college attainment for cohort t, the share of twenty-three year olds with a college degree, can be decomposed as three separate factors

$$\frac{P_t^{grad}}{P_t^{23}} = \left(\frac{P_t^{HS}}{P_t^{23}} \right) \left(\frac{P_t^{enroll}}{P_t^{HS}} \right) \left(\frac{P_t^{grad}}{P_t^{enroll}} \right). \tag{2.1}$$

P^{HS}, P^{enroll}, and P^{grad} are the number of individuals born at year t that complete high school, enroll in college, and graduate college. The first ratio is the fraction of all twenty-three year olds who have graduated high school, which we term the graduation rate. This ratio will be exogenous when we turn to the model, and therefore will not be influenced by the cost factors focused on in this paper. The second ratio is the fraction of high school graduates who enroll in college, the enrollment rate. Enrollment is endogenous to the model. The last is the fraction of college attendees who graduate college, or the graduation rate. Again, this ratio is endogenous to the model. Using equation (2.1), the growth rate of college

Table 1: Growth rate decomposition

	γ^{col}	Levels		Percent of γ^{col}	
		γ^{hs}	γ^{endog}	γ^{hs}	γ^{endog}
1900 - 1972	2.043	1.604	0.439	0.785	0.215
1900 - 1920	0.190	1.140	-0.950	5.997	-4.997
1920 - 1972	1.853	0.465	1.388	0.251	0.749
1950 - 1972	0.236	-0.014	0.250	-0.058	1.058

attainment between cohorts born at t and t' is approximated by

$$\gamma_{t',t}^{col} \equiv \log\left(\frac{P_{t'}^{grad}}{P_{t'}^{23}}\right) - \log\left(\frac{P_t^{grad}}{P_t^{23}}\right) = \gamma_{t',t}^{hs} + \gamma_{t',t}^{enroll} + \gamma_{t',t}^{grad}. \tag{2.2}$$

where γ^{hs}, γ^{enroll}, and γ^{grad} are the growth rates for the three ratios on the right hand side of (2.1). Unfortunately we do not have sufficient data back to 1900 to separately compute γ^{enroll} and γ^{grad}. However, we can compute the two jointly. That is, with data on college attainment and high school graduation rates, $\gamma^{enroll} + \gamma^{grad}$ can be computed as the residual of (2.2). We therefore rewrite equation (2.2) as

$$\gamma_{t',t}^{col} = \gamma_{t',t}^{hs} + \gamma_{t,t'}^{endog} \tag{2.3}$$

where γ^{endog} is the growth rate of factors endogenous to our model, and therefore potentially affected by changes in college costs and earnings premia. Table 1 decomposes γ^{col} for the entire sample (birth cohorts 1900 to 1972) and three subperiods.

If one considers the entire time period, college attainment increases by over two hundred percent. From an accounting perspective, over two-thirds of this can be accounted for by the 157% increase in high school completion, γ^{hs}. This result, however, is skewed by large differences across sub-samples of the entire time period. For 1900 to 1920 birth cohorts, the college attainment increases a relatively small nineteen percent. This masks the underlying heterogeneity of a huge increase in high school graduation (114%) and a huge decrease in endogenous factors (95%). From 1920 onwards, however, when the bulk of the college attainment increase occurs, over three-fourths of the increase in attainment is accounted for

by endogenous factors. This suggests that increases in high school graduation rates cannot be relied upon as the sole factor driving subsequent increases in college completion. Rather, it is vital to understand the factors affecting college enrollment and completion decisions for a full understanding of college attainment trends during the 20th century. The remainder of this paper will develop and calibrate a model to quantitatively assess the main contributors to this growth.

3 Model

We develop overlapping generations model to investigate the causes of increased college completion and increased ability sorting.[3] The relevant features include borrowing limits, uncertain ability, and risky completion of college education. The notation introduced in this section is summarized in Table 3 of the Appendix.

Demographics and Preferences Time in the model is discrete, and a model period is one year. Each period, N_{mt} males and N_{ft} females are born, each of whom lives for a total of T periods. Let $a = 1, 2, \ldots, T$ denote age. Each individual maximizes expected lifetime consumption, given by

$$\mathbb{E}_0 \sum_{a=1}^{T} \beta^{a-1} \left(\frac{c_a^{1-\sigma} - 1}{1 - \sigma} \right).$$

Endowments and Signals Individuals are ex-ante heterogeneous along three dimensions: their sex, m or f, initial asset endowment k_0, and ability to complete college, denoted α. The probability that any individual completes his or her current year of college is given by $\pi(\alpha)$, where $\pi' > 0$. Ability is drawn from a normal distribution $F(\alpha)$, and initial assets are drawn from a Pareto distribution $G(k)$. The distributions $F(\alpha)$ and $G(k)$ are not independent, so endowments for ability and assets may be correlated. Section 4.4 discusses further the justification and parameterization of these distributions.

While sex and asset endowments are perfectly observable, ability α is not. Instead, each individual receives a signal $\theta = \alpha + \varepsilon$ at the beginning of life. The error term is $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Because assets and ability are jointly distributed, individuals actually receive two pieces of

---

[3]The counterpart to ability in the data is IQ.

information about ability – the signal θ and asset endowment k_0. Let $\nu = (k_0, \theta)$ be the information an individual has about his true ability. After the initial college enrollment decision, ability α becomes publicly observable.

Education Decisions   The population we are considering consists of high school graduates, so that birth in this model translates to a high school graduation in the real world. At birth, every individual decides whether or not to enroll in college, given sex, asset endowment k_0, and ability signal θ. This is the only time this decision can be made. Once enrolled in college, individuals can only exit college by either graduating or by failing out with annual probability $\pi(\alpha)$. After failure, individuals enter the labor force and may not re-enroll, consistent with the finality of dropout decisions discussed in Card and Lemieux (2001). Graduating college requires C years of full-time education at a cost of λ_t per year. If an individual decides to not enter college, he or she immediately enters the labor market and begins to work.

Labor Market   We adopt the common assumption that individuals of different ages, a, sex s, and education, e, are different inputs into a constant returns to scale production function that requires only labor. Therefore, wages depend on age, sex, education level, and the year. We write wages as $w_{a,t}(e, s)$ for $s \in \{f, m\}$ and $e \in \{0, 1, \ldots, C\}$. While ability α has no direct effect on realized wages, it does affect expected wages because higher ability students are more likely to graduate college and earn higher wages.

Savings   Each individual can borrow and save at an exogenous interest rate r_t. Individuals cannot die in debt, so $k_{T+1} = 0$. During life, borrowing is constrained to be a fraction $\gamma \in [0, 1]$ of expected discounted future earnings. Therefore, individuals must keep assets k_t each period above some threshold $\bar{k}$, where

$$\bar{k} = -\gamma \cdot \mathbb{E} \sum_{n=a}^{n=T} \frac{w_{n,t}}{1 + r_t}$$

Note that both the expectations operator and wage can depend on a number of factors, including ability α, age a, year t, education e, and sex s. Therefore, the borrowing constraint will be written as the function $\bar{k}(\alpha, a, t, e, s)$. In a slight abuse of notation, we will write

$\bar{k}(a, t, e, s)$ when the borrowing constraint does not depend on ability α, as is the case once an individual finishes college.

3.1 Timing and Recursive Problem

At the beginning of year t, N_{mt} men and N_{ft} women are born at age $a = 1$. Again, each individual is initially endowed with assets k_0, sex s, ability α, and a signal θ of true ability. Immediately, each individual decides whether or not to enroll in college. If he or she enrolls in college, true ability is immediately realized, and the individual proceeds through college. In the case of leaving college due to failure or graduation, he or she proceeds to the labor market and works for the remainder of his or her life. Individuals who do not enroll in college proceed directly to the labor market, where they receive the wage associated with age a, education $e = 0$, and sex s.

Recursive Problem for Worker   For individuals currently not enrolled in college, their ability is irrelevant for their decision problem. Therefore, the value of entering year t at age a with assets k, years of college education e, and sex $s \in \{f, m\}$ is:

$$
\begin{aligned}
V_{a,t}^w(k, e, s) &= u(c) + \beta V_{a+1,t+1}^w(k', e, s) \\
s.t. \quad c + k' &= (1+r)k + w_{a,t}(e, s) \\
k' &\geq \bar{k}(a, t, e, s) \\
k_{T+1} &= 0
\end{aligned}
$$

Recursive Problem for College Student   If instead an individual is currently enrolled in college, he has already completed e years of his education and must pay λ_t in college costs for the current year. The probability that he passes and remains enrolled the next year, however, depends on his ability α. Recall that α is known with certainty as soon as the education decision is made, so there is no uncertainty about ability.

   The value of being enrolled in college at year t at age a, with assets k, ability α, e years

of education completed, and sex $s \in \{f, m\}$ is:

$$V_{a,t}^c(k, \alpha, e, s) = u(c) + \beta \Big[\pi(\alpha) V_{a+1,t+1}^c(k', \alpha, e+1, s) + (1 - \pi(\alpha)) V_{a+1,t+1}^w(k', \alpha, e, s) \Big]$$
$$s.t. \quad c + k' - \lambda_t = (1+r)k$$
$$k' \geq \bar{k}(\alpha, a, t, e, s)$$

If an individual is in his last year of college at age $a = C$, he moves directly into the workforce next period, so the recursive formulation becomes

$$V_{C,t}^c(k, \alpha, C, s) = u(c) + \beta V_{a+1,t+1}^w(k', \alpha, C, s)$$
$$s.t. \quad c + k' - \lambda_t = (1+r)k$$
$$k' \geq \bar{k}(a, t, C, s)$$

The College Enrollment Decision Given the value of being enrolled in college and working, it is possible then to define the educational decision rule at the beginning of life. Recall that at this point, α is unknown, but each individual receives information $\nu = (k_0, \theta)$. Each individual then constructs beliefs over possible ability levels by using Bayes' Rule.

Let $F(\alpha; k_0, \theta)$ be the cumulative distribution function of beliefs (as defined by Bayes' Rule) over ability levels. Given all this, an individual born in year t of sex s with assets k_0 and signal θ enters college if and only if the expected value of entering college is higher than the (certain) value of entering the workforce. This is given by the inequality

$$\int_\alpha V_{1,t}^c(k_0, \alpha, 1, s) F(d\alpha; k_0, \theta) \geq V_{1,t}^w(k_0, 0, s). \tag{3.1}$$

Let $\phi(k_0, \theta) = 1$ if the individual decides to enroll, and $\phi(k_0, \theta) = 0$ otherwise.

3.2 Measures of College Attainment and Average Ability

The two outcomes we are interested in comparing between model and data are college attainment and average student ability (for both college and non-college individuals). The

former is measured by the share of twenty-three year olds in a birth cohort with a college degree. The latter is measured by computing in the model the average ability percentile of individuals who choose to forego college to begin working immediately after high school (the "non-college" group), and the average ability percentile of individuals who choose to enroll in college, regardless of whether or not they graduate (the "college" group).

4   Calibration

The goal of this paper is to assess the roles played by a number of features of the economy in understanding ability sorting and college enrollment over time. We therefore take a multi-faceted approach to parameterizing the model. First, we construct historical data series for N_{mt}, N_{ft}, and λ_t, which are incorporated directly into the model. Second, we estimate life-cycle wage profiles $w_{a,t}(e, s)$, which are taken as given by model individuals solving their dynamic problem. Third, we set values for T, C, r_t, β, ρ_t, $\mu_{\alpha,t}$, $\mu_{k,t}$, $\sigma_{\alpha,t}$, $\sigma_{k,t}$, and $\pi(\alpha)$ for consistency with relevant data. Finally, we calibrate σ_ϵ and γ in order to match the two main outcomes of interest - college attainment and average student ability - in the last year of our sample, 1972. Each of these are discussed in more detail below.

4.1   Historical Time Series Data

As previously mentioned, N_{mt} males and N_{ft} females are "born" into the model each year, meaning they graduate high school and enter the model eligible to make college enrollment decisions. We take high school completion, and thus the population of potential college en-rollees, as exogenous. The series for N_{mt} and N_{ft} are taken directly from the U.S. Statistical Abstract Historical Statistics, and we use linear interpolation to supply missing values.

Annual college costs per student, λ_t, are calculated as the average tuition and fee expenses paid out-of-pocket by students each year.[4] Note that because we measure average *out-of-pocket* costs in the data, λ_t accounts for changes over time in the average amount of financial aid received by students in the form of public and private scholarships and grants. Full

---

[4]Additional student expenses, such as room and board, could also be included, and in fact we do consider these costs as a robustness exercise in Section 6. We choose to leave these out of the benchmark specification because such costs are usually more accurately classified as consumption rather than education expenses, and must be paid regardless of college enrollment status.

details of the data construction are relegated to Appendix A. Briefly, however, we compute λ_t each period as the total revenues from student tuition and fees received by all institutions of higher education divided by the total number of students enrolled in those institutions. The complete time series is constructed by splicing together data from historical print sources including the *Biennial Surveys of Education* (1900 to 1958) and the *Digests of Education Statistics* (since 1962).

4.2 Life-Cycle Wage Profiles

Life-cycle wage profiles $w_{a,t}(e, s)$ are estimated using decennial U.S. Census data from 1940 through 2000, along with American Community Survey (ACS) data from 2006-2010. Each ACS data set is a 1% sample of the U.S. population, so that when combined they constitute a 5% of the U.S. population, similar to a decennial census. The data are collected from the Integrated Public-Use Microdata Series (IPUMS) (Ruggles et al., 2010), and include wage and salary income, educational attainment, age, and sex. From age and education data we compute potential labor market experience, x, as age minus years of education minus six. We assume that wages can be drawn from one of three education categories - high school, some college, or college. These correspond to $e = 0$, $e \in [1, C-1]$ and $e = C$ in the model. For each education category, we estimate wage profiles for the non-institutionalized population between ages 17 and 65 who report being in the labor force using the following regression:

$$\log(w_{i,t}) = \delta_{i,t}^b + \sum_{j=1}^{4} \beta_j^s x_{i,t}^j \tag{4.1}$$

where i denotes individuals, b is birth-year cohort, s is sex, and x is potential labor market experience. In words, we regress log wages on a full set of birth year dummies plus sex specific quartics in experience.

4.3 Exogenous Parameters

Parameters set exogenously prior to solving the model are: T, C, r_t, β, and $\pi(\alpha)$. We set the length of working life at $T = 48$, implying that individuals born into the model at age 18 would retire at age 65. The number of periods required to complete college is $C = 4$, so

that all individuals in the model have post-secondary education $e \in \{0, 1, 2, 3, 4\}$.[5] The real interest rate is set to $r_t = 0.04$ in all periods, and the discount rate is $\beta = 0.96$, a standard value in models with annual periods.

Finally, we need to set the annual probability of passing college, $\pi(\alpha)$. Note that $\pi(\alpha)$ is a reduced form way to capture college non-completion for any reason, including failure and voluntary drop-out. We employ the simple assumption that an individual's *cumulative* probability of completing college equals her percentile rank in the ability distribution. For example, an individual whose ability is higher than 75% of the peers in her birth-year cohort will complete college with probability 0.75, conditional on enrollment. With the length of college set to $C = 4$, there are 3 independent opportunities for failure - after the first, second, and third years of school. Thus, the *annual* probability $\pi(\alpha)$ is simply the cumulative probability raised to the power one-third.

4.4   Joint Distribution of Assets and Ability

Our construction of the joint ability-asset distribution offers significant flexibility in matching the shapes of these distributions. Recall from Section 3 that α only affects an individual's probability of passing college. Furthermore, our interest in "ability" is limited to understanding changes over time in the average ability of college versus non-college students *within cohorts*. In other words, we only care here about the relative ability of students within the same birth year, as in the data from Figure 1b, not across birth years. As this is our objective, we do not have to worry about trends in average student ability (such as the so-called "Flynn effect") and can normalize the ability distribution for each birth cohort.

To construct this distribution, we need information on student ability (IQ) and initial parental transfers for education. To our knowledge, no single data source contains both, so we merge information from the National Longitundal Survey of Youth in 1979 (NLSY79) and the National Center for Education Statistics' program called High School & Beyond (HSB). The basic procedure is as follows, and details are in Appendix B. In the NLSY79 data we observe AFQT percentile scores for each student, and we compute their high school grades from transcript data. We assume the underlying AFQT raw scores are distributed $N(0, 1)$,

---

[5]We are not presently concerned with educational attainment beyond the bachelor's degree level, so we do not model post-graduate education in this paper.

which allows us to translate percentiles into raw grades. The unconditional distribution of grades is distributed $logN(\mu_{gpa}, \sigma_{gpa})$. To construct our link from GPA to $AFQT$, we run the regression

$$AFQT_i = \alpha_0 + \alpha_1 \log(GPA)_i + \varepsilon_i \tag{4.2}$$

Note that due to the distributional assumptions on grades, the underlying AFQT scores can be written as

$$AFQT \sim N(\alpha_0 + \alpha_1 \mu_{gpa}, \ \alpha_1^2 \sigma_{gpa}^2 + \sigma_\varepsilon^2) \tag{4.3}$$

The distribution in (4.3) gives us the relationship between IQ scores (here AFQT) and GPA, which is the missing link in the HSB dataset. Since the HSB dataset does contain grade information, (4.3) will allow us to construct an estimated distribution of IQ in the HSB data. Using the HSB dataset, we construct the empirical unconditional distribution of high school grades. We then assume HSB grades are distributed $logN(\mu_{HSB}, \sigma_{HSB})$ and estimate these two parameters to match the empirical grade distribution. We show in the Appendix that the estimated distribution matches the empirical grade distribution in HSB extremely well. With the distribution of grades, we then utilize the link between GPA and and AFQT derived from the NLSY. That is, $\alpha_0, \alpha_1, \sigma_\varepsilon$ imply

$$IQ_{HSB} \sim N(\alpha_0 + \alpha_1 \mu_{HSB}, \ \alpha_1^2 \sigma_{HSB}^2 + \sigma_\varepsilon^2).$$

Since we have the marginal distribution of IQ_{HSB}, we now need to construct the marginal distribution of transfers T. The empirical distribution of transfers has a large mass near zero. Therefore, we assume the marginal distribution of transfers follows a generalized Pareto distribution. The location parameter is set to zero, since zero is the minimum transfer value. The shape and scale parameters are set to minimize the squared errors between the empirical cdf and the generalized Pareto cdf. Again, in the Appendix we show that the estimated distribution matches the unconditional cdf derived from HSB quite well.

Our unconditional distributions for ability (IQ) and assets are normal and Pareto, which allow us to accurately describe the unconditional marginal distributions. A potential downside of this is that there is no simple way to combine the two marginals into a joint distribution. To overcome this, we utilize a Frank copula to combine the (continuous) marginal

distributions into a joint distribution. The Frank copula takes the form

$$C(u,v) = \frac{-1}{\rho} \log \left(1 + \frac{(\exp(-\rho u) - 1)(\exp(-\rho v) - 1)}{\exp(-\rho) - 1} \right)$$

where ρ governs the dependence of draws. Our joint distribution of α and k can therefore be written as

$$H(\alpha, k) = C\left[F(\alpha), G(k)\right]$$

where F is the normal cdf of ability, and G is the Pareto cdf of assets. The best fit to the data for the parameter ρ is $\rho = 4.46$. Intuitively, $\rho > 0$ implies positive dependence between the two draws. The Kendall rank coefficient is 0.42, implying that high school graduates with higher IQ on average have higher initial asset holdings.

The procedure just described provides the mean and variance of the asset distribution, as well as the correlation with ability, for the cohort born in 1962. We assume that the correlation of ability and assets is constant over time, but we do want to allow for changes over time in the mean and variance of the initial asset distribution. Recall that we interpret k_0 as a reduced-form way of capturing all of the personal financial resources available to a new high school graduate, including but not limited to parental gifts and bequests, and the individual's own prior income and savings. With this in mind, we require that the mean and standard deviation of initial assets in the model to track the mean and standard deviation of income in U.S. data. To this end, we start with $\mu_{k,t}$ equal to the annual mean real income per person, as in Piketty and Saez (2006) so that the average real asset endowment in the model equals the actual real mean income in the U.S. each year.

Piketty and Saez (2006) also provide historical data on the share of income received by the top ten percent of individuals, as well as the cut-off income level for the 90th percentile. Assuming that the U.S. income distribution is log-normal as predicted by Gibrat's law, we can use these data to back out the implied standard deviation of the U.S. income distribution each year. The procedure is as follows. Let real income in year t, denoted Y_t, be a random variable with realization y_t such that $Y_t \sim \ln \mathcal{N}(\mu_t, \sigma_t^2)$ and the associated cumulative distribution function is $F_Y(y_t; \mu_t, \sigma_t^2)$. Observed data are the real mean income in the U.S. in year t, denoted $\bar{y}_t$, and the 90th percentile of real income in year t, denoted $y_{90,t}$. A standard

property of the log-normal distribution is that $\mathbb{E}[Y_t] = \exp(\mu_t + \frac{\sigma_t^2}{2})$. Since $\mathbb{E}[Y_t] = \overline{y}_t$ is observed, we can guess a value $\tilde{\sigma}_t^2$ and solve for the associated mean of the distribution:

$$\tilde{\mu} = \ln(\overline{y}_t) - \frac{\tilde{\sigma}_t^2}{2}$$

Next, we compute $1 - F_Y(y_{90,t}; \tilde{\mu}, \tilde{\sigma}_t^2)$, which would be the fraction of total income received by those with income above the threshold value $y_{90,t}$ if the mean and variance of the income distribution were actually $\tilde{\mu}$ and $\tilde{\sigma}_t^2$. This process continues iteratively until we find a value σ_t^2, and associated μ_t such that the fraction of income received by the top ten percent equals that observed in the data. We then set $\sigma_{k,t} = \sigma_t$.

4.5 Calibrated Parameters

Finally, we choose the variance of the noise on the ability signal, σ_ϵ, to match the average ability of college relative to non-college individuals in 1962, and we choose the borrowing constraint, γ, to match the college completion rate, also in 1962. The variance on the noise of the ability signal is $\sigma_\epsilon = 0.2$. The borrowing constraint is $\gamma = 0.025$, which means that in any given period an individual can borrow up to 2.5% of his expected lifetime income. After an individual completes college this amount is known with certainty because the wage profiles are given, but during college the expected lifetime income is conditional on the probability of passing college.

5 Results

[NOTE: These results were computed with a previous calibration. They have not yet been updated, but will be soon. Our apologies for any confusion.]
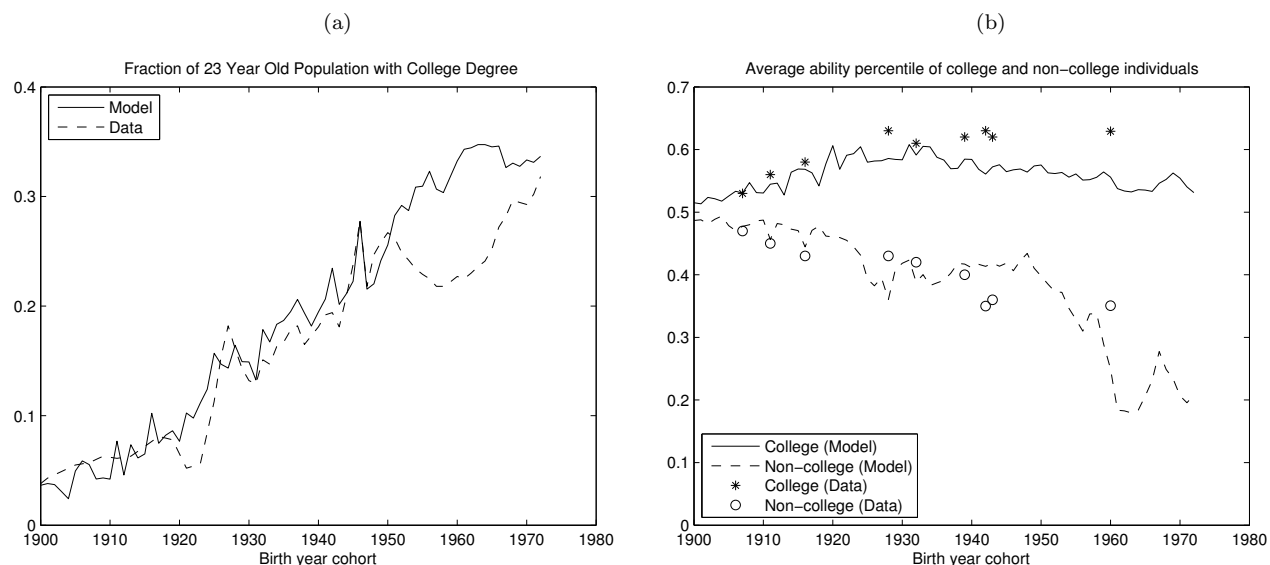
Our main computational exercise consists of first simulating the model for U.S. birth cohorts from 1900 through 1972 (i.e., students who graduated high school from 1918 through 1990), verifying that the model replicates important features of the historical data, and then running counterfactual simulations to quantify the impact of changes in direct college costs, education earnings premia, and opportunity costs of college (foregone wages) on college completion and average student ability. Having discussed the benchmark model

parameterization, we now examine how well the simulated model matches U.S. data.

5.1 Benchmark Model Fit

Figure 2 depicts the model predictions along with historical U.S. data for college completion and average student ability. The measure of college completion that we choose to match is the share of 23-year-olds with a college degree. While educational attainment is often measured later in life to capture those who complete college at older ages, we prefer this series for a couple of reasons. First, to our knowledge it is the only measure of college completion with consistent time series data for birth cohorts back to 1900. Second, our model is not constructed to evaluate college enrollment decisions of older students who: (i) are generally less financially-dependent upon parents when paying for education; (ii) face different opportunity costs of school after having been in the workforce for some time; and (iii) may anticipate different return on investment in education due to later-life completion.

Figure 2: Benchmark Model Results

(a)   (b)

Panel (a) of Figure 2 shows that, overall, the model replicates well the trends in U.S. college completion over much of the 20th century, with one notable exception. The model does not capture the initial decline and subsequent increase in college completion for cohorts born in the 1950s and 1960s. This deviation is due primarily to the modeling assumption

Table 2: Measures of Fit for Various Model Specifications

Model \ Cohorts	Fraction of 23-year-olds with College Degree				Average Ability Difference
	1900-1972	1900-1925	1926-1950	1951-1972	1900-1972
Benchmark	0.158	0.015	0.009	0.133	0.034
Imperfect foresight	0.055	0.020	0.013	0.022	0.028
Constant costs rel. to income	0.134	0.024	0.011	0.099	0.034
$\text{Corr}(\alpha, k_0) = 0.30$	0.183	0.023	0.017	0.143	0.038
Include room and board	0.159	0.019	0.017	0.123	0.035

that individuals know their lifetime wage profile with certainty, implying that they can perfectly forecast changes in the education earnings premium. Later we consider alternative assumptions, and find that the model can generate more accurate predictions over this time period.

Panel (b) of Figure 2 plots the average ability percentile of students who attempt college (even if they do not complete), and those who have only high school education. While we only have a few reliable data points to match, those we do have show a clear pattern of increased sorting by ability over time. For cohorts born at the beginning of the 20th century, college and non-college students had similar ability on average, but the ability gap widened throughout the century. This general pattern is also predicted by the model.

In order to facilitate quantitative comparison with alternative specifications, we also provide measures of model fit over various time periods in Table 2. The measure of fit we report is the sum of squared deviations between model and data. The columns labeled "Fraction of 23-year-olds with college degree" refer to the series in Panel (a) of Figure 2. For this series, we compute the fit over all cohorts 1900-1972, and three subsamples: 1900-1925, 1926-1950, and 1951-1972. As seen in the "Benchmark" model specification in Panel (a) of Figure 2, the model matches the data very closely for cohorts born pre-1950, but does less well for cohorts born after 1950. The column labeled "average ability difference" measures how well the model matches the difference between the average ability percentile of college and non-college individuals. We only report the full sample for this statistic because there are so few data points to match within the sub-sample periods.
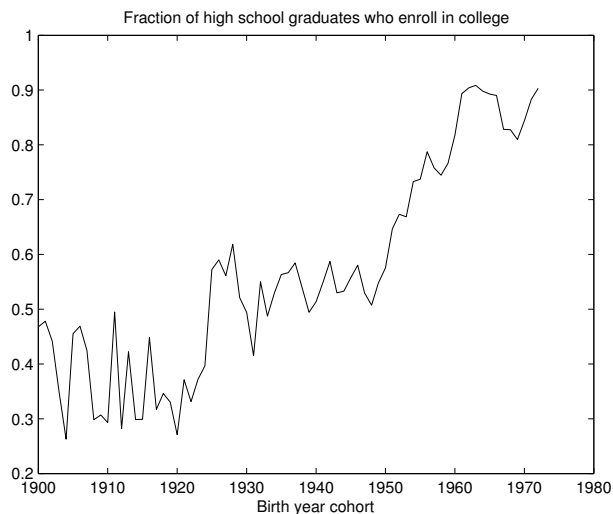
5.2 Decomposition of Benchmark Results

Recall from Section 2 that our measure of college completion – the fraction of twenty-three year olds with a college degree – can be decomposed as:

$$\frac{P_t^{grad}}{P_t^{23}} = \left(\frac{P_t^{HS}}{P_t^{23}}\right)\left(\frac{P_t^{enroll}}{P_t^{HS}}\right)\left(\frac{P_t^{grad}}{P_t^{enroll}}\right). \tag{5.1}$$
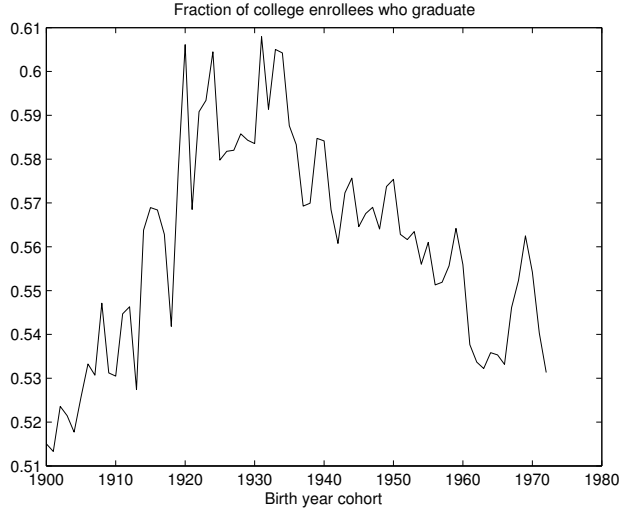
where P^{HS}, P^{enroll}, and P^{grad} are the number of people that complete high school, enroll in college, and graduate college. While the first is exogenous, the second and third terms are endogenous to the model. Since any change in completion can be driven by a change in one of the three ratios on the right hand side, we begin with a decomposition of college completion into these three ratios. This will also help guide our counterfactual experiments. Figures 3 and 4 plot the time series for the two endogenous ratios.

Figure 3: College Enrollment Conditional on High School Graduation

First, Figure 3 plots the share of high school graduates that enroll in college, as predicted by the model. In the language of equation (5.1), this is P_t^{enroll}/P_t^{HS}. Figure 3 shows that for cohorts born between 1900 and 1920, college enrollment rates conditional on high school graduation were between 30 and 50 percent, albeit with a lot of noise. This rate increased for cohorts born in the 1920s and generally remained between 50 and 60 percent for cohorts through 1950, after which the rate again increased substantially.

Figure 4: College Pass Rate

Fraction of college enrollees who graduate

The third term in equation (5.1) is the share of college enrollees that graduate by age twenty-three. This is given by the ratio P_t^{grad}/P_t^{enroll} and is plotted in Figure 4. While Figure 4 shows that the college pass rate has a fair amount of year-to-year noise, the hump-shaped trend is still evident. From the 1900 through 1930 birth cohorts, the college pass rate increased from about 51% to nearly 61%. After the 1930 cohort, however, this trend reverses, and the pass rate steadily declines back down to around 53%. This result is consistent with evidence from Bound et al. (2010), who compare the high school class of 1972 (roughly birth cohort 1954) to that of 1992 (birth cohort 1974) and find a significant decrease in college completion conditional on enrollment. In our model, this pattern is due entirely to the ability composition of college students. Recall from Panel (b) of Figure 2 that the average ability of college enrollees was generally increasing through the 1930 cohort, then decreasing in the following cohorts. Unfortunately, we have found no reliable historical data to compare with the model's predicted pass rates. However, the National Center for Education Statistics (NCES) does provide more recent data we can use for a rough comparison. For the cohort beginning college in 1996 (assuming they are around 18 years old on average, this would be approximately the 1976 birth cohort), the share completing college within five years was 50.2%.[6] Our last birth cohort in the model is 1972, so the comparison is not perfect, but the model pass rate of 53.1% for that cohort is quite close.

---

[6]See Table 341 in the 2010 *Digest of Education Statistics*.

To isolate the relative importance of each ratio in equation (5.1), we replot college attainment in turn holding college enrollment and pass rates constant at their 1900 cohort level. We then use equation (5.1) to predict the hypothetical college completion rates. This is plotted in Figure 5, along with the benchmark prediction for college completion.

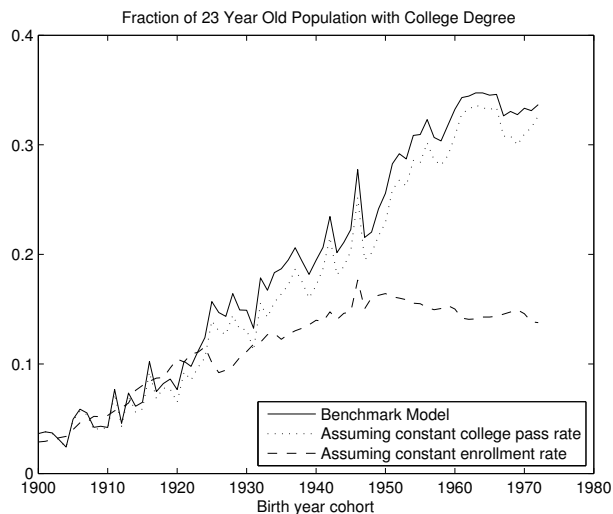Figure 5: College Completion if Enrollment Rates and Pass Rates were Constant

Figure 5 shows that if the college enrollment rate had remained constant instead of rising after the 1920 cohort, the model would have under-predicted college completion rates by more than half by the end of the time series.[7] Similarly, if the college pass rate had instead remained constant at the 1900 value of 51.5%, then college completion would have been lower, but not significantly different from the benchmark. This suggests that the two key factors affecting college completion in the model are the high school graduation rate and the college enrollment rate. Since only the latter is endogenous, we now tunr to a serious of counterfactual experiments to understand the underlying mechanics of the changing enrollment rate in the model.

---

[7]In Figure 5, we assume that the the college enrollment rate conditional on high school graduation is constant at 36.9%, which is the average enrollment rate for cohorts 1900 through 1920.

5.3 Counterfactual Experiments

5.3.1 Variation over time in precision of ability signals.

In the benchmark specification, we chose $\sigma_\epsilon = 0.2$ to match the difference in average ability of college and non-college students from the 1972 birth cohort. Holding σ_ϵ constant, however, generated no sorting in average ability, indicating that changes in the economic factors such as real college costs and earnings premia were not also responsible for the increased sorting by ability. We now allow σ_ϵ to vary over time and ask what else may have accounted for the increased sorting seen in Panel (b) of Figure 1.

Unfortunately, we do not have direct evidence on the precision with which individuals in a given cohort know their own ability relative to their peers. At a qualitative level, it is likely that this precision has increased – i.e., $\sigma_{\epsilon,t}$ has likely decreased – over time. In the early part of the 20th century, no standardized exams existed to compare students within cohorts across schools. Those college admissions exams that did exist were generally school-specific, so there was little scope for comparison of students across schools. During World War I, the U.S. military began testing recruits using the Army Alpha and Army Beta aptitude tests. By World War II, these tests were replaced by the Army General Classification Test (AGCT), a precursor to the Armed Forces Qualification Test (AFQT). On the civilian side, the introduction of the Scholastic Aptitude Test (SAT) in 1926 started a trend toward more widespread use of standardized exams as a college admissions criteria. As standardized testing became more common, students obtained more and more precise signals of their own ability relative to peers. In the modern era, virtually every student contemplating college takes either (or both) of the SAT or the ACT (American College Testing) exams. Even those who do not take these college admissions exams still have quite precise information about their relative ability because other standardized exams are mandated at public schools.

With this historical background in mind, we make the following assumptions on the time series structure of $\sigma_{\epsilon,t}$. For cohorts making college decisions prior to World War II, i.e., those born 1900 through 1923 and graduating high school from 1918 through 1941, we assume that $\sigma_{\epsilon,t}$ decreases linearly from $\sigma_{\epsilon,1900} = 2$ to $\sigma_{\epsilon,1923} = 0.2$. For cohorts born after 1923, $\sigma_{\epsilon,t}$ remains constant at 0.2. This is an admittedly ad hoc construction, but in a simple

way it captures the trend of each subsequent cohort getting slightly better information than the previous cohort as aptitude and ability tests became more common in the time between the world wars. By the completion of World War II, such tests were in widespread use and students likely had quite precise signals about their own ability relative to peers.

[ADD RESULTS LATER].

5.3.2 What if real college costs increased proportional to real disposable incomes?

College enrollment is obviously tied closely to tuition costs. We therefore begin by asking how college completion rates and average student ability would have differed over the time period in question if real college costs were constant with respect to real average income. Figure 6 depicts the actual time series data for real college costs that we use in the benchmark model (solid line), along with a hypothetical series for college costs which are a constant fraction of annual real average income (dashed line). From 1920 to around 1940, the actual series exceeds the hypothetical series due the the fact that per student tuition and fees spiked relative to income during the Great Depression. Then from the early 1940s until about 1990, the hypothetical series is above the actual series. Holding all else constant, we would expect that individuals in the counterfactual model facing the hypothetical college costs should attend college in greater numbers for the cohorts born from about 1900 to 1920 (those in school from around 1920 to 1940), and fewer of those born after 1920 would attend college.
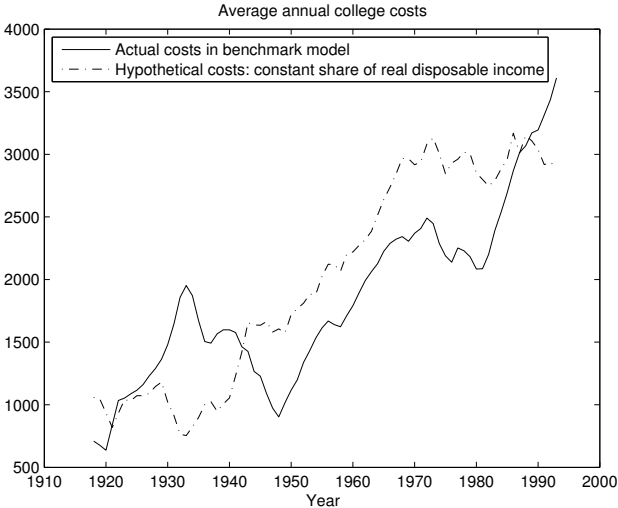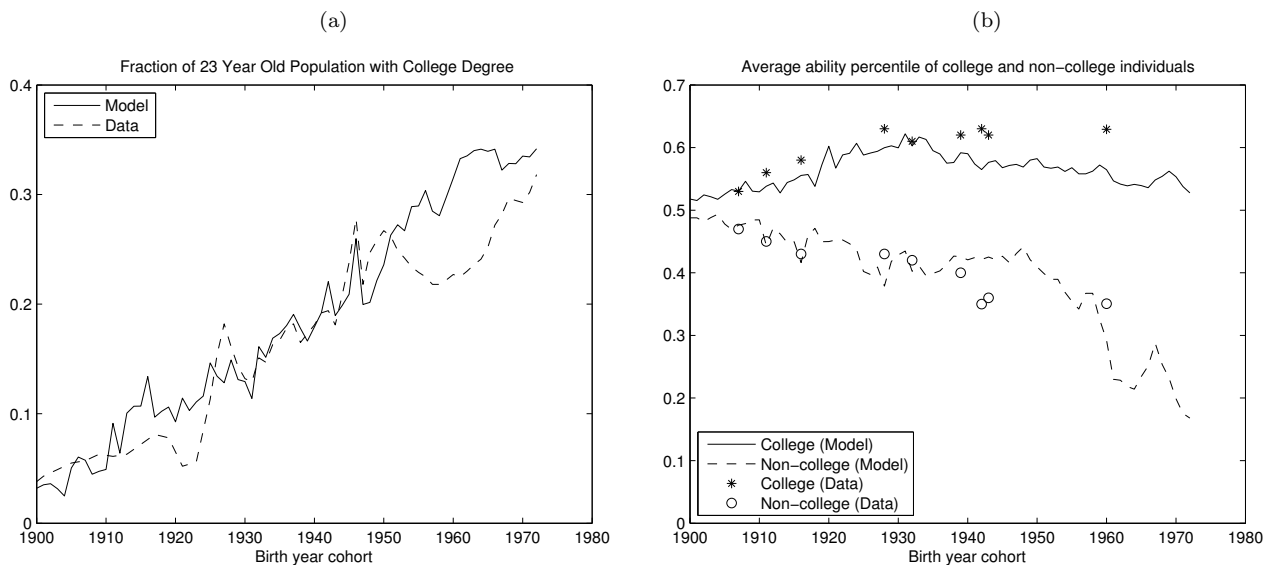
Figure 6: College Costs

Figure 7 largely confirms these predictions. Relative to the data, the model predicts too many people attending college for those cohorts born between about 1910 and 1925. For the cohorts from 1925 through 1950, the model does predict slightly fewer college graduates, but still matches the data quite closely. And finally, for the cohorts born after 1950, the model still predicts more college graduates than in the data. However, as can be seen in Table 2, the model fit improves over this period since the sum of squared deviations fall from 0.133 to 0.099, a decrease of more than 25%. Turning to Panel (b) of Figure 7, there are hardly any discernible differences in average ability of college and non-college students relative to the benchmark model. This can also be confirmed by noting that sum of squared deviations for the average ability difference in Table 2 is unchanged from the benchmark value of 0.034. We conclude that the fluctuations in real college costs relative to real income are not a major factor in accounting for the increased ability sorting over time.

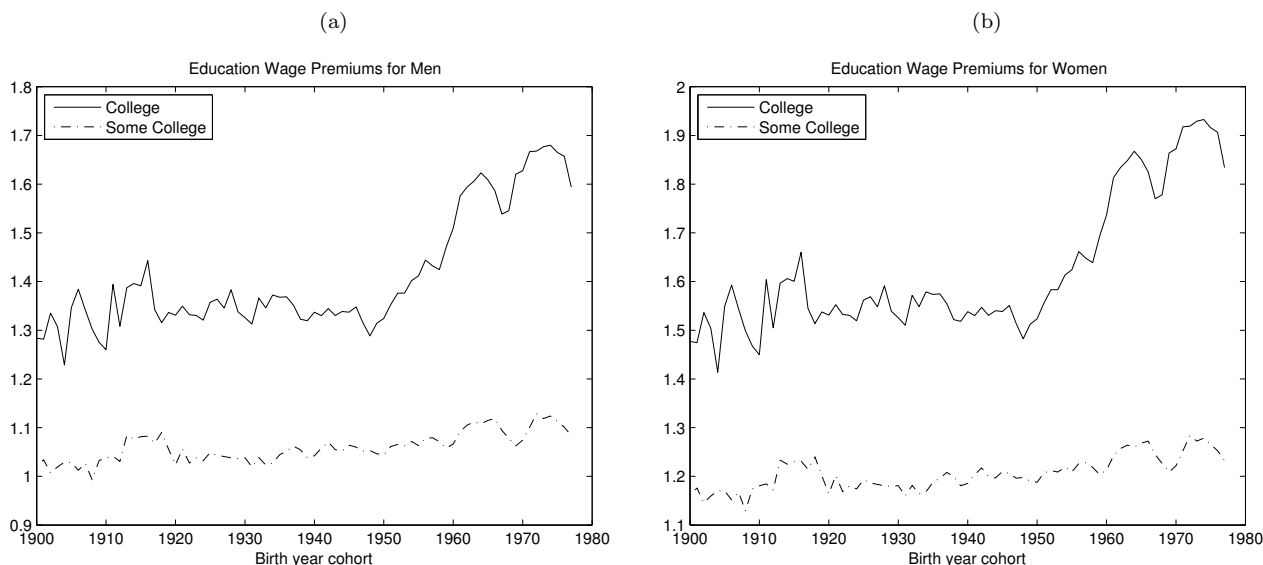Figure 7: Results with Alternative College Costs

(a)

(b)

5.3.3 What if individuals do not have perfect foresight of education earnings premia?

Figure 8 shows that for cohorts born in the U.S. prior to 1950, the education premia implied by our estimated life-cycle wage profiles exhibit some year to year variation, but essentially no trend. Beginning around the 1950 cohort, however, the college earnings premia began

increasing steadily. We now examine how the model predictions for college completion and average student ability would differ if, instead of predicting changes in the education premium exactly, model individuals expected an historical average education earnings premia to prevail in the future as well.
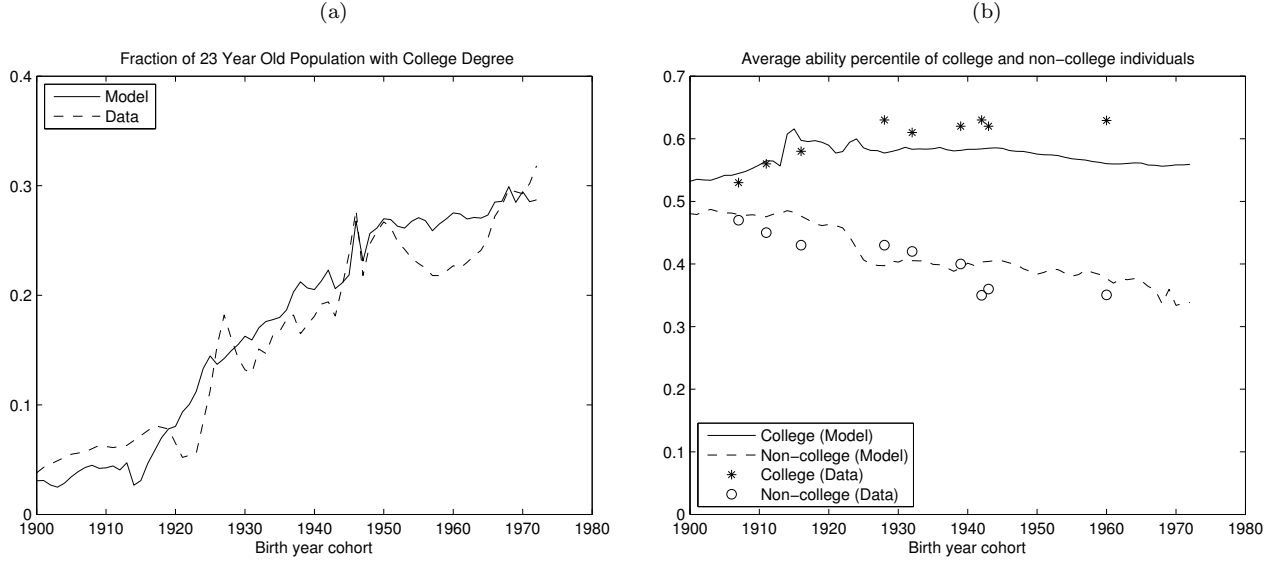
Figure 8: Education Premia Implied by Estimated Life-Cycle Wage Profiles

(a)   (b)

For this exercise, we assume that the high school wage for each cohort is observable, but the earnings premia for individuals who complete college or some college are not observable. Rather, individuals observe a moving average of the earnings premia earned by previous cohorts and assume their own cohort's earnings premia will be the same. Thus, as the true college earnings premium begins rising, newly born cohorts will predict the increase imperfectly and with several years lag.

Figure 9 shows the model predictions under this counterfactual experiment, assuming a 25-year moving average. Relative to the benchmark model results, notice that the model now comes much closer to the actual college completion rate in the data for cohorts born after 1950. The model still does not capture all of the decline for the cohorts in the 1950s, but as Table 2 clearly shows, this specification fits the data much better than the benchmark assumption that individuals perfectly forecast changes in the education premia. Over the entire time period, the sum of squared deviations declines by almost two-thirds from the

Figure 9: Results with Imperfect Foresight of Education Earnings Premia

(a)   (b)

benchmark value of 0.158 to 0.055. All of this gain is due to the 1951-1972 cohorts, where the sum of squared deviations changes from 0.133 to 0.022, a decrease of more than 83%. Additionally, the model's ability to match changes in average ability of college and non-college students also improves under this specification. According to the last column of Table 2, the sum of squared deviations declines from 0.034 to 0.028. These improvements strongly suggest that perfect foresight of education earnings premia is a problematic assumption. Accurately modeling students' expectations about the returns to education is crucial for understanding college enrollment decisions, particularly during periods of time when education premia are changing rapidly.
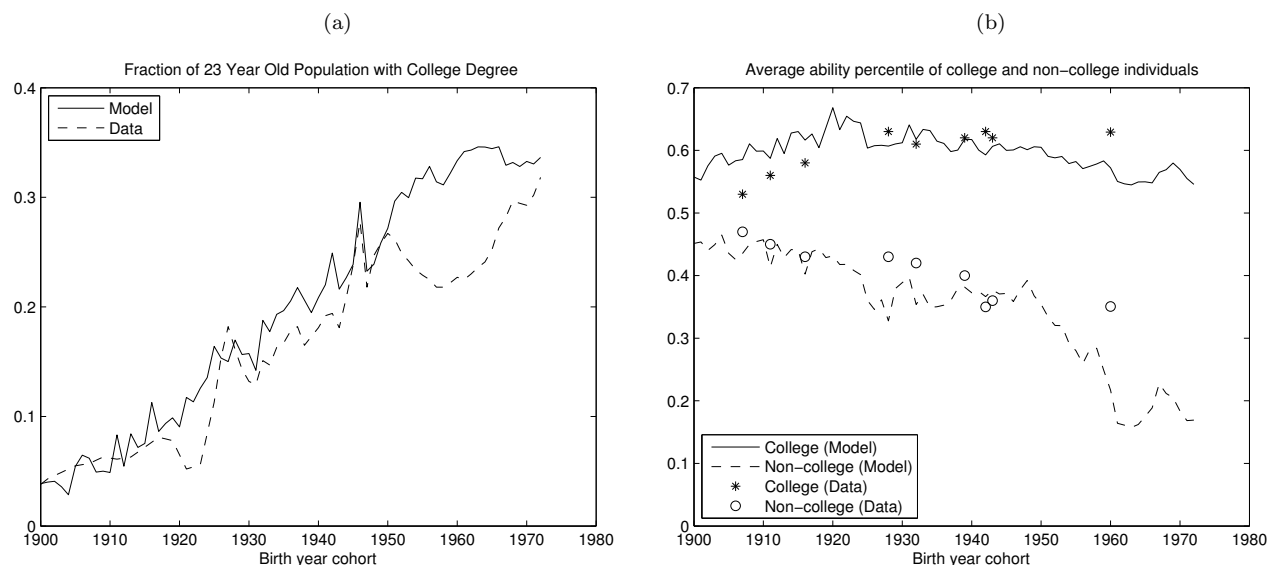
6   Robustness

Having discussed the benchmark model results and counterfactual experiments, we now make a few remarks about the robustness of some modeling assumptions. In particular, we made the strong assumption that ability and initial assets were uncorrelated. We also assumed that room and board were excluded from college costs. We now relax these assumptions and see how they affect the results.

6.1 Correlation of Ability and Initial Assets

In the benchmark specification, we assumed that the random endowments for ability and assets were uncorrelated. However, there is evidence to suggest that these may be positively correlated, and we want to understand how this affects the results. We maintain the assumption that α and $\log(k_0)$ share a bivariate normal distribution, only now we set $\rho = 0.3$. All other parameters are maintained as in the benchmark specification. Figure 10 shows the model predictions for college completion and ability sorting between college and non-college individuals.

Figure 10: Results with Positive Correlation between Ability and Initial Assets

(a)   (b)

Relative to the benchmark model results, two things are notable. The positive correlation between ability and assets increases college completion minimally throughout the time period, and it increases the difference in ability between college and non-college students during earliest birth cohorts. Both of these effects reduce the model fit slightly, as seen in Table 2. The increase in completion is simply due to the fact that higher ability students are now more likely to have greater financial resources as well, thus making them more likely to attend college. The effect on average ability is also quite intuitive. Recall that individuals receive information $\nu = (k_0, \theta)$, where $\theta = \alpha + \varepsilon$ is the noisy signal of true ability α. As ρ increases k_0 becomes more informative about α, so individuals with high initial assets

will infer that they have higher ability, and thus be more likely to enroll in college. This increases the average ability of individuals who attempt college, while simultaneously decreasing the average ability of non-college individuals. The effect is largest for earlier birth cohorts because later birth cohorts received more accurate signals about their true ability.

6.2 College Costs Including Room and Board

College costs in the benchmark model were restricted only to tuition and fees. Now, we take a broader view of college costs and examine whether or not the results are sensitive to the inclusion of room and board expenses. Like the earlier time series data on college tuition and fees, we construct this data from printed historical government documents. The details are found in appendix A. For this experiment, all calibrated values are maintained just as in the benchmark economy, with the exception of the borrowing constraint, γ. We need to adjust γ because students now face additional college expenses, so college completion rates would be too low if we held γ constant at the benchmark value. The new borrowing constraint which allows us to match the time series of college completion is $\gamma = 0.04$.

Figure 11: Results for College Costs including Tuition, Fees, Room, and Board
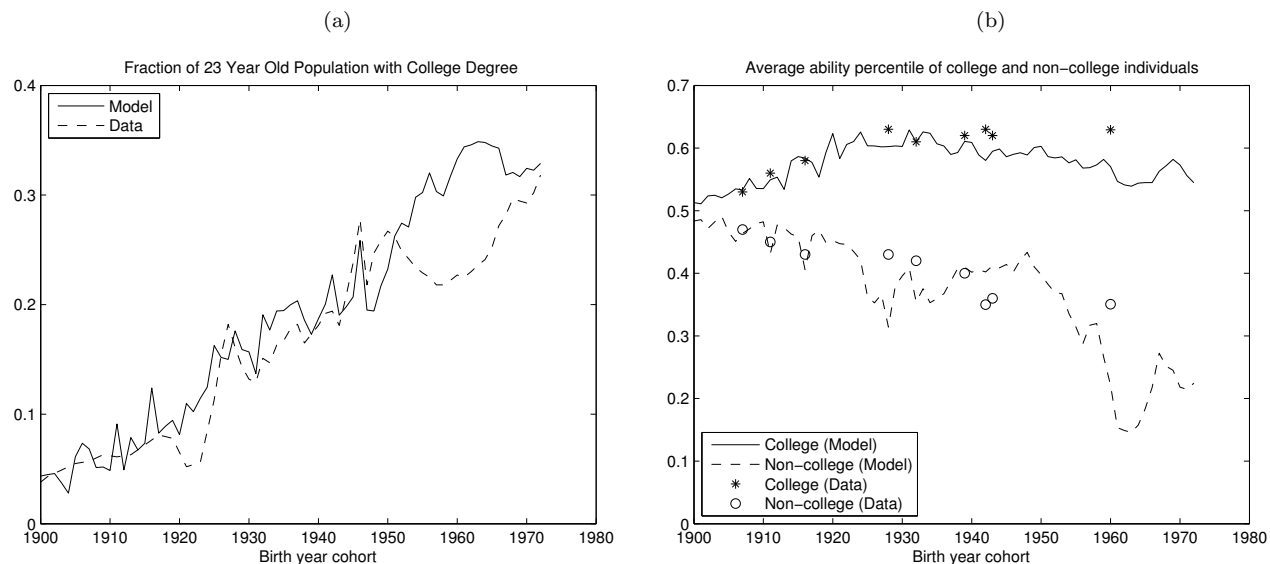
(a)

(b)

Figure 11 shows the model predictions for college completion and average student ability when room and board costs are included. Relative to the benchmark results in Figure 2, very

little has changed. The model still predicts college completion rates in line with the data up until the 1950s and 1960s cohorts, when model and data diverge. Additionally, average ability of college and non-college students diverges over time just as in the benchmark model. Referring to Table 2, it is clear that while the model fits college completion slightly worse than the benchmark model pre-1950, it does slightly better post-1950. On the whole, this model fits almost exactly as well as the benchmark model for both college completion and average ability difference.

7   Conclusion

We develop an overlapping generations model with unobservable ability and borrowing constraints to investigate post-secondary completion and ability sorting in the birth cohorts of 1900–1972. To discipline our model, we digitize and utilize historical data series including statistics on college costs and high school graduation rates. We find that the share of high school graduates enrolling in college and the subsequent college pass rate are both key for understanding increased college graduation rates. However, we find no evidence that economic factors – including real college costs, opportunity costs, education wage premia, or asset endowments – have a major impact on increasing ability sorting over time. We do find, however, that a decrease in the variance of ability signals can properly match this fact, a trend which we attribute to increases over time in standardized testing.

An important deviation between the benchmark model and historical data is that the model does not properly match college completion after the 1950 birth cohort. We show that this could be due to individuals having imperfect foresight about the college earnings premium. If individuals observe a moving average of the earnings premia from previous cohorts and use this to estimate the future earnings premium, then changes in the earnings premium are taken into account only with a lag. We build this into the model and find that it significantly improves the model's fit. We therefore view this as evidence of backward looking wage estimation when making college enrollment decisions.

An interesting use of this framework would be an extension to multiple countries. Evidence suggests that ability is strongly related to growth (Hanushek and Kimko, 2000), but the

causality from formal schooling to economic growth is somewhat tenuous (Bils and Klenow, 2000). If developing countries have very little ability sorting between education levels, as was the case in the early U.S., there may be a weak correlation between education level and labor efficiency. In a cross-country context, this could arise due to tighter borrowing constraints or less precise signals about true ability. We will explore this link in future research.

References

M. Bils and P. J. Klenow. Does schooling cause growth? *American Economic Review*, 90 (5):1160–1183, 2000.

J. Bound, M. F. Lovenheim, and S. Turner. Why have college completion rates declined? an analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics*, 2(3):129, 2010.

D. Card and T. Lemieux. Dropout and enrollment trends in the postwar period: What went wrong in the 1970s? In *Risky behavior among youths: An economic analysis*, pages 439–482. University of Chicago Press, 2001.

R. Castro and D. Coen-Pirani. Do returns to skill explain the evolution of educational attainment in the u.s.? Mimeo. Université de Montréal., 2012.

F. Cunha, J. Heckman, and S. Navarro. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers*, 57(2):191–261, 2005.

C. Garriga and M. P. Keightley. A general equilibrium theory of college with education subsidies, in-school labor supply, and borrowing constraints. Federal Reserve Bank of St. Louis Working Paper, 2007.

C. Goldin and L. F. Katz. *The Race between Education and Technology*. Harvard University Press, 2010.

E. A. Hanushek and D. D. Kimko. Schooling, labor-force quality, and the growth of nations. *American Economic Review*, 90(5):1184–1208, 2000.

L. Hendricks and O. Leukhina. The return to college: Selection bias and dropout. Mimeo. University of North Carolina., 2011.

L. Hendricks and T. Schoellman. Student abilities during the expansion of u.s. education, 1950-2000. Mimeo. University of North Carolina., 2012.

C. M. Hoxby. The changing selectivity of american colleges. *Journal of Economic Perspectives*, 23(4):95–118, 2009.

E. Keller. The slowdown in american educational attainment. Mimeo. University of Iowa., 2013.

L. J. Lochner and A. Monge-Naranjo. The nature of credit constraints and human capital. *American Economic Review*, 101(6):2487–2529, October 2011.

T. Piketty and E. Saez. The evolution of top incomes: A historical and international perspective. *American Economic Review*, 96(2):200–205, 2006.

S. Ruggles, J. T. Alexander, K. Genadek, R. Geoken, M. B. Schroeder, and M. Sobek. *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]*. Minneapolis: University of Minnesota, May 2010. URL http://www.ipums.org/.

T. D. Snyder, editor. *120 Years of American Education: A Statistical Portrait*. United States Department of Education, 1993.

P. Taubman and T. Wales. *Mental Ability and Higher Educational Attainment in the 20th Century*. National Bureau of Economic Research, 1972.

Appendices

A Data

We take several historical data series as exogenous to the model, and this section details the construction of those series. Data are taken from several sources in order to construct a consistent series since 1900. From 1900 to 1958, most data were collected every two years and published in the Biennial Survey of Education (BSE). Since 1962, the Digest of Education Statistics (DES) has been published annually. Other publications including the annual U.S. Statistical Abstract, the Bicentennial Edition "Historical Statistics of the United States: Colonial Times to 1970", and "120 Years of American Education: A Statistical Portrait" help in bridging breaks between series, as well as verifying continuity of series that may have changed names from year to year. Also, many data were revised in later publications, so we take the most recent published estimates where available.

First, let c_t be the total annual cost of college per student. We assume that the total cost for educating *all* students in the U.S. in a given year equals the total revenues received in the current period by all institutions of higher education. Dividing this by the total enrollment each year yields the total annual cost per student. Alternatively, one could use the total current expenditures rather than revenues as the measure of total cost, but this makes little difference quantitatively because revenues and expenditures track each other quite closely. In addition, the revenue data is preferable because it allows us to determine how much of costs are paid out-of-pocket by students for tuition and fees, and how much comes from other sources such as state, local, and federal governments, private gifts, endowment earnings, auxiliary enterprises (athletics, dormitories, meal plans, etc.), and other sources. The numerator for c_t is constructed as follows:

- 1997-2000: total current revenue must be computed as the sum of current-fund revenue for public and private institutions, from the DES.

- 1976-1996: total current revenue equals "current-fund revenue of institutions of higher education" from the DES.

- 1932-1975: total current revenue equals "current-fund revenue of institutions of higher

education" in "120 Years of American Education: A Statistical Portrait".

- 1908-1930: total current revenue equals "total receipts exclusive of additions to endowment" for colleges, universities, and professional schools, from the BSE.

- 1900-1908: total current revenue equals "total receipts exclusive of additions to endowment" for colleges, universities, and professional schools, and is computed as (income per student)*(total students, excluding duplicates) from the BSE. Continuity with later years can be verified using the "income per student" series, which was published from 1890-1920.

The denominator for c_t is constructed as follows:

- 1946-2000: total fall enrollment for institutions of higher education, from the DES.

- 1938-1946: resident college enrollments, from the BSE. Continuity with the later series can be verified in that year 1946 data matches in both.

- 1900-1938: total students, excluding duplicates, in colleges, universities, and professional schools, from the BSE. Continuity with the later series can be verified in that year 1938 data matches in both.

Second, we construct two time series which estimate the share of annual college costs paid out-of-pocket by students. One measure, λ_t, includes only tuition and fees paid by students, and the other measure, ϕ_t includes tuition, fees, room, and board. In each year λ_t equals total tuition and fees paid by all students divided by total current revenue received by institutions of higher education. Similarly, ϕ_t equals total tuition, fees, room, and board aid by all students divided by total current revenue received by institutions of higher education. In each case, the measure of total current revenue is the same time series as was used above in constructing c_t. The time series for λ_t is constructed as follows:

- 1997-2000: current fund revenues from tuition and fees for all institutions of higher education is computed as the sum of the series for public and private institutions, from the DES.

- 1976-1996: current fund revenues from student tuition and fees, from the DES.

- 1930-1975: current fund revenues from student tuition and fees, from "120 Years of American Education: A Statistical Portrait".

- 1918-1930: receipts of universities, colleges, and professional schools for student tuition and fees, from BSE.

- 1900-1918: we are unable to obtain proper data for these years.

The time series for ϕ_t is constructed as follows:

- 1976-2000: Average tuition, fees, room, and board paid by full-time equivalent (FTE) students is obtained from the DES. We multiply this by enrollment of FTE students, also from the DES, and divide by the current fund revenues to compute ϕ_t.

- 1960-1976: we are unable to obtain proper data for these years.

- 1932-1958: Data available biennially on 1total revenues from student tuition and fees, as well as revenue from auxiliary enterprises and activities (room and board), in the BSE. ϕ_t computed as the sum of these, divided by total current revenue.

- 1900-1930: ϕ_t computed as total revenue from student fees (included tuition, fees, room, and board) divided by total current revenue.

B   Procedure Connecting NLSY79 and High School and Beyond

Our goal is to get a joint distribution of parental transfers and IQ. High School and Beyond (HSB) has data on transfers and grades, but no information on IQ. The NLSY has data on grades and IQ, but no information on transfers. This note details how these two data sets are combined to construct a joint distribution of transfers and IQ.

B.1   Distribution of IQ in HSB using NLSY

We first construct high school grades in NLSY by utilizing the grades for individual classes. Since NLSY GPA is originally calculated out of 4.0, we multiply all grades by 25 to reach a scale of 100. This variable is GPA. We assume GPA is lognormally distributed. Since the NLSY only has percentiles of AFQT scores and not the raw scores, we assume that

the underlying AFQT scores are distributed by a standard normal. Using the percentiles of AFQT, we back out the implied underlying AFQT score, which we denote with variable name $AFQT$.

The first step is to construct the link between IQ and grades in the NLSY. To do so, we run the regression

$$AFQT_i = \alpha_0 + \alpha_1 log(GPA)_i + \varepsilon_i \tag{B.1}$$

Note that GPA rescale keeps the standard deviation of the regression error term the same as if we were calculating out of 4.0. It follows from the assumptions on the distribution of grades that the distribution of underlying AFQT scores can be written as

$$AFQT = \alpha_0 + \alpha_1 \log(GPA) + \varepsilon \sim N(\alpha_0 + \alpha_1\mu_{GPA}, \ \alpha_1^2\sigma_{GPA}^2 + \sigma_\varepsilon^2)$$

since the regression assumes $\varepsilon \sim N(0, \sigma_\varepsilon)$. The regression gives coefficients $\alpha_0 = -8.00$ and $\alpha_1 = 1.91$, and predicts an error term standard deviation of $\sigma_\varepsilon = 0.88$.

We now turn to the High School and Beyond (HSB) dataset to construct the link between IQ and transfers. Since HSB does not include actual IQ data, we have to use the information from the NLYS construction. HSB asks students which grade bin they fall into. The bins are (1) 90-100, (2) 85-89, (3) 80 - 84, (4) 75-79, (5) 70-74, (6) 65-69, (7) 60-64, (8) lower than 60. Let $\bar{g}_j$ and $\underline{g}_j$ be the maximum and minimum grades in any grade bin j. There are no observations in bin (8). If grades are distributed according to some distribution with cdf F, each bin j includes mass

$$\widetilde{M_j} = F(\bar{g}_j) - F(\underline{g}_j).$$

Let M_j be the empirical mass in each grade bin calculated in the HSB data. We assume the underlying grade distribution is lognormally distributed $logN(\mu, \sigma)$, and choose μ and σ to minimize the sum of the squared errors

$$(\mu_{HSB}, \sigma_{HSB}) \in argmin \sum_j (\widetilde{M_j} - M_j)^2.$$

We now have the underlying grade distribution in HSB, which is distributed $logN(\mu_{HSB}, \sigma_{HSB})$. We assume the link between IQ and grades are identical in HSB and the NLSY, so that we
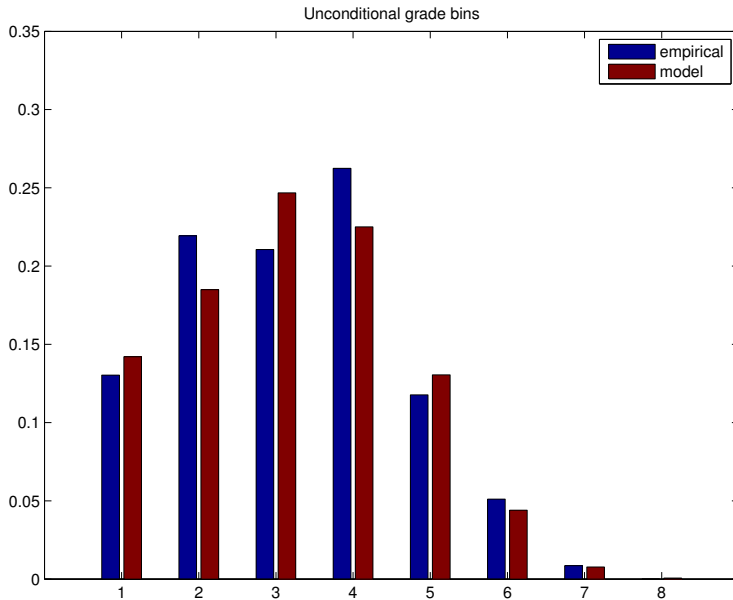
Figure 12: Data and Predicted Grade Bins in HSB

can use the link constructed in (B.1) to generate the distribution of IQ in HSB. That is, $\alpha_0, \alpha_1, \sigma_\varepsilon$ imply

$$IQ_{HSB} \sim N(\alpha_0 + \alpha_1\mu_{HSB}, \ \alpha_1^2\sigma_{HSB}^2 + \sigma_\varepsilon^2).$$

The construction of IQ percentiles from the underlying IQ distribution is then straightforward.

B.2 Marginal Distribution of Transfers

Since we have the marginal distribution of IQ_{HSB}, we now need to construct the marginal distribution of transfers T. To construct the initial asset holdings, we calculate the total transfers to high school graduates in the 4 years after high school graduation and construct the present value of these transfers utilizing an interest rate of 4%. Since the empirical distribution of transfers has a large mass at zero, we assume that the distribution follows a generalized Pareto distribution. The location parameter is set to zero, since zero is the minimum transfer value. The shape and scale parameters are set to minimize the squared errors between the empirical cdf and the generalized Pareto cdf.
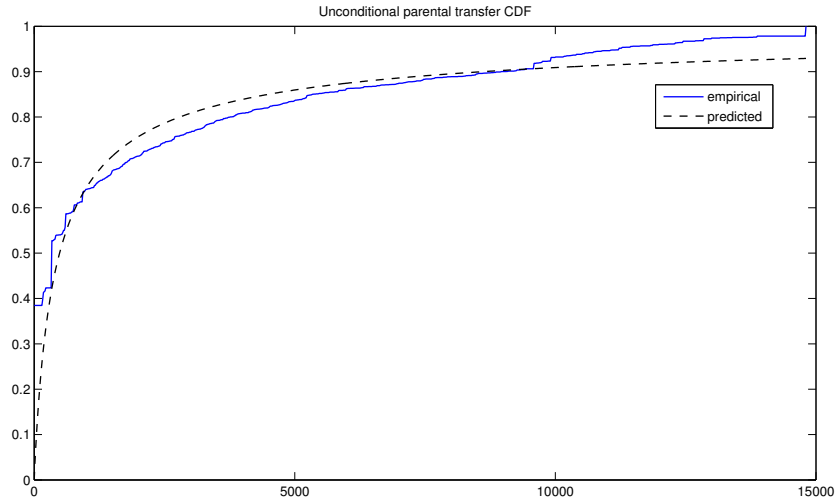
Figure 13: Data and Predicted Transfer CDF in HSB

B.3 Constructing the Joint Distribution

Since our marginal distributions are normal and Pareto, there is no simple way to combine these into joint distribution. We use a Frank copula, which allows for arbitrary distributions to be combined into joint distributions, as long as the marginal cdfs are continuous. The parameter ρ governs the dependence of draws on each other. Roughly, $\rho > 0$ implies positive linear correlation and $\rho < 0$ implies negative linear correlation between draws from the two distributions.

Table 3: Summary of Notation

Symbol	Definition
N_{mt}	Number of males born into model (i.e., graduating high school) each year
N_{ft}	Number of females born into model (i.e., graduating high school) each year
a	Age of individual, where $a = 1, 2, ..., T$
s	Sex of individual, where $s \in \{f, m\}$
k_0	Initial asset endowment
α	Ability endowment
$\pi(\alpha)$	Annual probability of passing college, given ability α
ρ_t	Correlation between initial asset and ability endowments
$\mu_{\alpha,t}$	Mean of ability distribution
$\mu_{k,t}$	Mean of initial asset distribution
$\sigma_{\alpha,t}$	Standard deviation of ability distribution
$\sigma_{k,t}$	Standard deviation of initial asset distribution
θ	Signal of true ability, where $\theta = \alpha + \varepsilon$
ε	Error term on signal of true ability
σ_ε	Standard deviation on distribution for ε
ν	Vector of variables that are informative about true ability, where $\nu = (k_0, \theta)$
C	Number of years required to graduate from college
e	Years of education completed by individual, where $e \in \{0, 1, ..., C\}$
λ_t	Annual cost of college in year t
$w_{a,t}(e, s)$	Wage in year t for individual of age a, sex s, and education e
γ	Individuals may not borrow more than a fraction γ of expected discounted future earnings
$\bar{k}$	Minimum asset level for individual, given age, sex, education, and γ