Essays in the Economics of Education

by

Brian Christopher Clark

Department of Economics
Duke University

Date: ____________________
Approved:

____
Peter Arcidiacono, Supervisor

____
Arnaud Maurel

____
V. Joseph Hotz

____
Seth Sanders

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Economics
in the Graduate School of Duke University
2016

Abstract

Essays in the Economics of Education

by

Brian Christopher Clark

Department of Economics
Duke University

Date: ______________________
Approved:

____
Peter Arcidiacono, Supervisor

____
Arnaud Maurel

____
V. Joseph Hotz

____
Seth Sanders

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics in the Graduate School of Duke University
2016

Abstract

This dissertation is comprised of three essays in the economics of education. In the first essay, I examine how college students' major choice and major switching behavior responds to major-specific labor market shocks. The second essay explores the incidence and persistence of overeducation for workers in the United States. The final essay examines the role that students' cognitive and non-cognitive skills play in their transition from secondary to postsecondary education, and how the effect of these skills are moderated by race, gender, and socioeconomic status.

To my loved ones: for always supporting me

Contents

List of Tables

List of Figures

Acknowledgements

First, I must thank my advisors, Peter Arcidiacono and Arnaud Maurel, for their invaluable guidance throughout my time at Duke, and the opportunity to learn from their example. I would also like to thank the other two members of my dissertation committee, V. Joseph Hotz and Seth Sanders for their crucial help, insightful comments, and encouragement. My work has also benefitted from the frequent help and comments of seminar participants in the Duke Labor Lunch Group, and seminar participants at the Federal Trade Commission and the United States Census Bureau, along with the opportunity to present my work through Duke Gradx and the North Carolina Council of Graduate Schools. This research would also not be possible without access to data provided by the Office of Population Research at Princeton University in Chapter 3, and the North Carolina Education Research Data Center in Chapter 4.

For Chapter 3, my co-authors and I gratefully acknowledge financial support from the W.E. Upjohn Institute for Employment Research through the Early Career Research Grant No. 13-141-02B. We also thank seminar participants at Duke, John Hopkins and UNC - Chapel Hill; as well as the attendees of the 2013 North American Summer Meeting of the Econometric Society (Los Angeles), the workshop "Empirical Research in the Economics of Education" (Rovira i Virgili University, October 2013), the workshop "Skill Mismatch: Microeconomic Evidence and Macroeconomic Relevance" (Mannheim, April 2014), the 2014 Annual Meeting of the Society for

Economic Dynamics (Toronto), the 2014 Annual Congress of the European Economic Association (Toulouse), and the 2015 Econometric Society World Congress (Montreal) for useful comments and discussions.

Lastly, I have to thank my loved ones for their unconditional support. I am grateful to my parents Chris and Judy Clark for nurturing my interests and for the sacrifices they made to enable my academic and career pursuits, and my twin brother Casey Clark for keeping me grounded. I thank my friends Marie Hull, Kate Duch, and Kristoph Kleiner for sharing all of the joys and challenges of graduate school, and for showing me the kind of person I want to be. And finally, to my partner Justin Harris, for helping me become the person I want to be.

1

Introduction

I present three essays on the economics of education in this dissertation, which are unified by a focus on higher education: from how students choose to attend college or pursue a major, to how they utilize their education in the labor market.

In Chapter 2, I examine how labor market shocks influence college students' decisions to switch majors. Over one third of college students in the US switch majors; and since students' beliefs about their expected labor market returns are a determinant of their major choice, shocks to these beliefs likely play a role in their decision to switch majors. For example, sophomore engineering majors at UT Austin were 40% more likely to switch majors during the 2001-02 recession, when the unemployment rate for engineering graduates increased to double that of the average college graduate. However, typical models of major choice are unable to explain this behavior because they assume students' beliefs follow rational expectations. Hence, engineering students would correctly predict that the unemployment rate would recover by their expected graduation in 2004-05. I am the first to explain why students switch majors in response to labor market shocks by more flexibly modeling their beliefs, including how persistent they believe labor market shocks will be. I show that students' choices are consistent with beliefs that the shock to the unemployment rate is highly persistent, and use simulations to illustrate the impact that beliefs have on students' welfare. In particular, giving students rational expectations reduces the probability that they switch out of engineering during the recession by 4.8% and increases the expected earnings of sophomore engineering students by $1.8m.

Chapter 3, which is co-authored by Clément Joubert and Arnaud Maruel, analyzes the career dynamics of workers in the United States who have more schooling than their peers in the same occupation. We use data from the National Longitudinal Survey of Youth, 1979 combined with the Current Population Survey to analyze transitions into and out of overeducated employment, together with the corresponding effects on wages. Overeducation is a fairly persistent phenomenon at the aggregate and individual levels, with 66% of workers remaining overeducated after one year. Overeducation is not only more common, but also more persistent among blacks and low-AFQT individuals. Further, the hazard rate out of overeducation drops by about 60% during the first 5 years spent overeducated. However, the estimation of a mixed proportional hazard model suggests that this is attributable to selection on unobservables rather than true duration dependence. Lastly, overeducation is associated with lower current as well as future wages, which points to the existence of scarring effects.

Lastly, Chapter 4, which is co-authored with Ying Shi, documents the role that cognitive and non-cognitive skills play in determining students' transitions from secondary to postsecondary education. We are able, with an administrative dataset on primary and secondary public school students in North Carolina provided by the North Carolina Education Research Data Center (NCERDC), to examine how these unobserved skills drive students decisions to persist in high school, choose to attend college, choose an initial college major, and how intensively to apply to colleges. Furthermore, we are able to examine the differential distribution and effects of these skills by race, gender, and socioeconomic status, and reveal significant disparities in educational outcomes. In particular, ignoring socioeconomic status masks significant heterogeneity in the probability of graduating from high school or attending college. For example, low SES white males have the second-lowest unconditional probability of graduating from high school, and a 6.7% lower unconditional probability of graduating from high school compared to the next lowest group, low SES white females. We then specify a factor model to examine these patterns and show that, while the model is able to explain the high school attainment gaps for most groups, we are still unable to completely explain the lower rates of graduation for low SES white males, which could point to better outside labor market options for these individuals.

2

Shocked out of Your Major: Do Labor Market Shocks Prompt Major Switching?

College students' labor market earnings vary significantly by their choice of major, and students' beliefs about these earnings play a role in that choice.[1] However, it is not clear whether these beliefs are consistent with rational expectations, and if students choices are *ex ante* optimal. Two papers illustrate this debate: Freeman (1976a) shows students enrolling in engineering fail to anticipate regular long term cycles in labor market returns, while Berger (1988) finds students' major choices are consistent with their long term returns. This distinction is especially important for the over one third of college students in the US that switch majors:[2] if their beliefs are not consistent with rational expectations, labor market shocks could prompt them to sub-optimally switch majors. However, research on students' major choice and switching decisions commonly ignore this distinction and assume students have

[1] Arcidiacono, Hotz, and Kang (2012)

[2] 37.6% of students that entered a Bachelor's degree program in 2003-04 switched by 2009. Source: U.S. Department of Education, National Center for Education Statistics, 2003-04 Beginning Postsecondary Students Longitudinal Study, Second Follow-up (BPS:04/09). Computation by NCES PowerStats on 9/9/2015.

rational expectations. So we do not know whether beliefs are consistent with rational expectations; and if they deviate, how, and what the welfare implications are for students.

I answer these questions by modeling how the major switching behavior of students from The University of Texas at Austin responded to the economic expansion in the 1990's and the 2001-02 recession, which disproportionally affected engineering graduates. Wiswall and Zafar (2015b) suggest students use major-specific labor market shocks like these to update their beliefs about future earnings, but whether students switch majors in response depends in part on how sensitive their beliefs are to shocks. For example, consider sophomore engineering students that enrolled in 2000. They observe the unemployment rate for engineering graduates increase from 2.01% to 5.12% in 2002, before recovering to 2.31% by their graduation in 2004. These sophomores should forecast this recovery if their beliefs about the persistence of the shock are consistent with rational expectations, and hence we would not expect their beliefs about earnings to change significantly nor for them to switch majors at a rate significantly higher than other cohorts. However, these students switch out of engineering at a rate 40% higher than sophomores from the 1998 and 1999 cohorts. I am able to capture this behavior by specifying a sequential model of major choice and the labor market that relaxes rational expectations assumptions. My results show students choices are consistent with the belief that labor market shocks are more persistent than the rational expectations level, which best fits historical trends. Students' choices may also be consistent with the belief that shocks are fully persistent and follow a random walk.

These results suggest, by showing students' switching behavior is not consistent with rational expectations about the persistence of labor market shocks, that students are basing their major choices on *ex ante* inaccurate labor market forecasts. For example, sophomore engineering students in 2002 would predict an unemployment rate of 3.2 p.p at graduation if they had rational expectations, which is near the actual rate of 2.9 percentage points However, students' choices imply they believe shocks are more persistent, and predict that the unemployment rate will be 24% higher, at 4.2 percentage points These predictions then affect students' beliefs about the expected present value of utility from earnings in engineering, which fall -1.33% relative to their initial expectations, a rate over five times larger than the drop with rational expectations. I also simulate students' choices with rational expectations to show that inaccurate forecasts prompt students to make suboptimal major choices: 12.3% of these sophomores switch majors with their actual expectations, which falls to 11.8% with rational expectations. This drop is also relevant from a policy perspective, as it implies giving these 1,130 engineering students accurate labor market expectations is associated with a $1.8m increase in their expected lifetime earnings.

However my results do not wholly reject the rational expectations assumptions that are widely used in studies of college major choice;[3] rather they illustrate criticism that the validity of rational expectations depends in part on correctly specifying students' information sets. Indeed, my model and data share features with other sequential models of major choice that also control for expected labor market returns, like Arcidiacono (2004) and Beffy, Fougere, and Maurel (2012). However these and other papers form a statistical model of expected earnings with longitudinal data on students' actual future earnings,[4] and allow expectations to vary by combinations of observed and unobserved characteristics. However Manski (1993) argues that drawing conclusions about how students' choices respond to labor market returns in these models depends on accurately modeling the contents of the information set

[3] See Altonji, Blom, and Meghir (2012) and Altonji, Arcidiacono, and Maurel (2015) for an overview of the literature.

[4] This is a common approach on papers that study related issues or use alternative models as well. These include Berger (1988), papers that allow for uncertainty over completion of college (Altonji, 1993) and major (Montmarquette, Cannings, and Mahseredjian, 2002), and differential accumulation of human capital across majors (Kinsler and Pavan, 2014).

they use.[5]

This paper is consistent with one of two increasingly common approaches for addressing this problem, which uses students' choices to infer the properties of their information sets. One solution, which was suggested in Manski (2004) and is used in a number of recent studies of college major choice, involves directly eliciting students' subjective expectations. However, since subjective expectations data is relatively rare, this approach often requires conducting a survey,[6] which is prohibitively costly for this and many other research questions.[7] Cunha and Heckman (2007) addresses this issue with an alternative technique, which allows researchers to identify components of students' information sets with more readily available longitudinal data. The concept behind this approach is that students' choices will only be explained by factors that are in their information sets, as in Carneiro, Hansen, and Heckman (2003) which found agents are unable to forecast shocks to their future earnings *ex ante*.

But unlike the approach in Carneiro, Hansen, and Heckman (2003), which also uses longitudinal data to study whether future shocks to earnings enter the information set, I use contemporaneous labor market data to study how known shocks enter the information set. These contemporaneous shocks are particularly relevant for the long term labor market outcomes of students, as Altonji, Kahn, and Speer (2014) demonstrates by documenting significant and long lasting wage penalties for individuals that leave college and enter the labor market during a recession. Therefore,

[5] Manski (1993) illustrates this point by specifying a stationary model of schooling choice. He finds that if a model uses ability to form expected earnings when students do not, it will erroneously show that students' choices do not respond to the returns to education.

[6] Arcidiacono, Hotz, and Kang (2012), Zafar (2011), Wiswall and Zafar (2015b) and Wiswall and Zafar (2015a) all use data on students' expectations from proprietary surveys on undergraduate students from Duke, Northwestern, and NYU.

[7] Obtaining expectations data is particularly burdensome in this case because it requires observing multiple cohorts over a long period of time that contains a labor market shock, and preferably one that disproportionally affects one major.

students may have an incentive to switch majors if they observe a disproportionally large negative shock to the labor market for graduates with their major. However, whether students have this incentive depends crucially on how persistent they perceive the labor market shock to be.

I start with a benchmark rational expectations model of expected labor market returns that incorporates contemporaneous labor market shocks. Students forecast major-specific returns over their labor market career by first generating predicted log wages and unemployment rates that control for returns to experience from their prevailing rates, which I estimate from the Survey of Income and Program Participation (SIPP).[8] Then they use two separate AR(1) processes to forecast corresponding future values, where the mean of each process varies by major,[9] and use them to form the expected present value of wages over their career at graduation. This baseline specification for expected wages is consistent with other rational expectations models, like Arcidiacono (2004), with the exception of using contemporaneous data. Instead of using future earnings, students' use historical data on log wages and unemployment rates to estimate the AR(1) parameters, which have relatively low rates of persistence $0.6 \leqslant \rho \leqslant 0.7$.

The advantage of this specification for the benchmark model is that it allows for several deviations from these rational expectations assumptions. For example, the AR(1) processes on log wages and unemployment rates allow students to predict shocks will be more persistent than the rational expectations level. I can also allow students to have labor market forecasts with varying degrees of myopia, as in Berger

[8] The SIPP provides necessary information on wages, labor force status, college major, labor market experience, and a range of other potential controls that I use to control for major and potential experience.

[9] AR(1) processes allow me to easily vary the persistence of labor market shocks, and are commonly used to model similar shocks. Guvenen (2007) notes that AR(1) processes fit income dynamics well, and Blanchard et al. (1992) uses low-order AR process to model regional deviations in unemployment rates.

(1988), and collapse labor market shocks across majors to test whether students respond to aggregate or major-specific labor market shocks. Lastly, I also allow students' utility over wages in the labor market to take a constant relative risk aversion (CRRA) form that allows for varying levels of risk aversion.[10]

I am able to consider these alternatives to rational expectations because I have data that allows me to observe students' major switching behavior on a semester-by-semester level, however I have to make a number of additions to the framework model to allow for this additional variation. Unlike Blom, Cadena, and Keys (2015), which observes students' final major only, semester-level data is able to show when students switch majors and more precisely control for their information set at the time they switched. However, students are also likely to face different switching costs depending on when they switch, which could delay their graduation, impact their grades, and drive responses to the same labor market shock that varies across cohorts. I account for these complicating factors by flexibly controlling for direct switching costs that vary by the major students switch from and into, what semester students are in when they switch, and a range of other factors.[11] Lastly I allow for switching majors to delay graduation by specifying a framework sequential major choice model that allows for a stochastic graduation process.

However, the features that enable this sequential model of major choice to show whether students' beliefs about the labor market are consistent with rational expectations, and control for the other important factors that mediate their responses to labor market shocks, also make it difficult to estimate. Most importantly, allowing students to make new major choices in every semester and giving them a choice over

[10] Beffy, Fougere, and Maurel (2012) and Buchinsky and Leslie (2010) also use this functional form, which is useful because it adds risk aversion while also nesting the more common log utility used in papers like Arcidiacono (2004).

[11] These include students' cumulative GPAs, whether they are switching for a second or higher time, family socioeconomic status, and whether they are switching back into a previously-held major.

a disaggregated set of majors implies a complex future value term: and one that is computationally-intensive for traditional full information methods.[12] I solve this problem by using conditional choice probability (CCP) techniques first developed by Hotz and Miller (1993), which allows me to collapse the future value term down to a function of the one-period ahead conditional probability of dropping out.

The rest of the paper proceeds as follows: Sections 2.1-2.3 introduce three key components that model the labor market, students' beliefs about labor market outcomes, and complex switching costs that mediate their propensity to switch. Then in Section 2.4 I outline a framework model of major choice, which motivates the features of a basic model and two additional features that enable me to include the effect of labor market shocks. Section 2.5 describes the demands this more complex model imposes on the estimation specification and data, and outlines how the CCP estimator and data satisfy these demands. Lastly, I discuss the estimates of this new model in Section 2.6, along with counterfactual welfare simulations; Section 2.7 concludes.

2.1 Labor market shocks

I begin in this section by specifying the first of the three key components of my sequential major choice model, which captures the labor market shocks. First I discuss what features are necessary in the data in order to document these shocks, the properties of the Survey of Income and Program Participation (SIPP) data I use, and how it meets these requirements. Then I use the SIPP data to document the effects of the 2001-02 recession on wages and unemployment rates by major, and the possible implications these patterns have for how students would respond to the recession. Finally, I use these conclusions to implement how labor market shocks

[12] See Arcidiacono and Ellickson (2011) for a detailed discussion of the computational burden of these models with full information methods.

enter students' information sets.

2.1.1 Data requirements

The literature on college major choice uses many different sources of labor market data, but the most popular in recent research fall into three categories: panel data on students' future labor market returns, survey data on subjective earnings expectations, and aggregated labor market data that is contemporaneous with students' time in college.

Each of these sources of data has benefits. Panel data, like the NLS72 used in Altonji (1993) and Arcidiacono (2004), is able to more accurately control for factors that determine both performance in college and the labor market; such as unobserved heterogeneity that drives selection into engineering programs and high-earning occupations in the labor market. Similarly, data on subjective expectations, like that used in Wiswall and Zafar (2015a) and Arcidiacono, Hotz, and Kang (2012), is able to more accurately control for the labor market components of students' information sets. Lastly, contemporaneous labor market data, like that in Buchinsky and Leslie (2010) and Blom, Cadena, and Keys (2015), only contains information that would be available to students at the time they make their decisions, and can also be linked to many different sources of data on student major choices.

However, contemporaneous labor market data is best suited to studying the effects of labor market shocks on choices in this context. Studies that use subjective expectations data often use proprietary survey instruments that are costly to administer; and it is also not possible to plan *ex ante* for a survey to collect data on student behavior during a recession. Similarly, while longitudinal wage data is useful in a number of settings, it must still be matched to contemporaneous data to capture labor market shocks that enter students' information sets while in college. Furthermore, longitudinal wage data is problematic for studying the effects of labor market shocks because many such datasets, like the NLS72, survey a relatively small number of students across many institutions that make controlling for heterogenous switching costs difficult.

This contemporaneous labor market data must also document shocks to wages and unemployment at a level of aggregation that matches the shock that enters students' information sets. For example, Blom, Cadena, and Keys (2015) studies the effect of aggregate labor market shocks on major concentrations, and finds that students respond to labor market shocks by shifting to majors with higher returns. However, research using surveys of subjective expectations by Wiswall and Zafar (2015b) shows that students update their expectations about the labor market most strongly in response to shocks that affect their major, rather than aggregate shocks. Hence, these data must also allow me to construct labor market shocks that are aggregated both by and across majors. Finally, in order to distinguish whether students respond to major-specific or aggregated labor market shocks, I require a dataset that contains shocks that impact certain majors more than others.

The data I use from the 1990-2008 panels of the SIPP meets all three of these requirements. First, it is a large, nationally-representative sample of individuals that allows me to construct a set of contemporaneous labor market shocks from 1990-2013. Second, each panel records information on each individual's level of educational attainment, major in college, and whether they are a college dropout; which allows me to construct both major-specific and aggregated labor market shocks. Finally, this sample also contains the economic expansion in the 1990s and the 2001-02 recession, which disproportionally affected college graduates with engineering degrees.

2.1.2 Survey of Income and Program Participation

In this section, and in the remainder of the paper, I construct the labor market data from a sample that pools information on individuals from every wave of the 1990-

1993, 1996, 2001, 2004, and 2008 panels of the SIPP. I create a relevant sample for the labor market returns of college graduates and dropouts by keeping only those individuals that: are either employed or unemployed and looking for work, are not enrolled in school, have attempted at least some college, and are 21-59 years old. Then I classify individuals as either dropouts if they did not graduate from college, or graduated with a major classified into one of five categories: engineering, the sciences, business, liberal arts, and fine arts.[13]

I then create a series of log hourly wages and unemployment rates aggregated within and across majors, and dropouts. For wages I take either the wage rate for hourly employees, or the equivalent hourly wage for salaried workers, for respondents' primary occupation in each survey year;[14] and I also generate an indicator for unemployment status in the survey month. I then generate estimates of the wage and unemployment rates for each labor market category by calculating the mean value for each category, calendar year pair.[15]

Finally, in the following results I do not consider the effects of further education or occupational choice on labor market returns. These effects are studied in many other papers, including Altonji, Kahn, and Speer (2014) and Blom, Cadena, and Keys (2015), where they are used to either link educational choices to historical labor market data that does not record major, or to study the role occupation concentration

[13] The SIPP also records the college majors of respondents with agriculture or forestry, home economics, police science or law enforcement, and technical or vocational degrees. However I drop these majors from my sample, since these majors are either not commonly offered by four-year degree granting institutions, or are offered by colleges in associates' or vocational programs.

[14] I calculate equivalent hourly wages for salaried workers by dividing their total earnings by the estimated total hours worked in a given month. Total hours are calculated as the typical number of hours a salaried respondent works in a week, multiplied by the number of weeks in the survey month.

[15] I allow students to choose majors in every semester in the final specification of the model. Since students would then be updating their labor market information each semester, I use a dataset where I linearly interpolate values for the labor market in the first and second halves of the year from the annual data.

and diversification plays in mediating the effects of a recession. While these effects are interesting *per se*, they are not necessary to identify the effects of labor market shocks on major choice. Stratifying the wage and unemployment data by major or dropout broadly controls for the range of occupations, graduate education, and other choices students can make after earning their bachelor's degrees that also affect earnings.[16]

2.1.3 Wage and unemployment trends

These data reveal a number of interesting patterns in wage and unemployment rates that are relevant for this study. I start with Figure 2.1, which plots the unemployment rates for graduates in each of the five major categories, all college graduates, and college dropouts from 1991-2008. This period of time is useful because it allows me to compare the features of the early 1990s recession with the so-called "dot-com" recession in 2001-02.

This graph shows that the 2001-02 recession disproportionally affected graduates with engineering degrees. In particular, the unemployment rate among engineering graduates increased from 1.02 percentage points in the spring of 1999 to a peak of 5.33 percentage points in fall 2002, which remained above 4 percentage points in the following year. Both the level and rate of change in this unemployment rate exceeded that of every other major; including similar majors like the sciences, which peaked at 2.95 percentage points in the fall of 2001, and the average for all college graduates which peaked at 2.85 percentage points in the fall of 2002. Perhaps the most striking result is that the peak in the unemployment rate for engineers in 2002-

[16] This approach requires the implicit assumption that the empirical distribution of graduate education and occupational choice would not significantly change between individuals that are currently in the market and current students. For example, this could be violated if current engineering students are much more likely to go on to obtain an MBA by age 45 than current 45 year old engineering graduates. However, this only becomes a problem if current college students also take this change into account when making their forecasts, which research on the accuracy of student labor market forecasts in Betts (1996) and other papers suggests is unlikely.

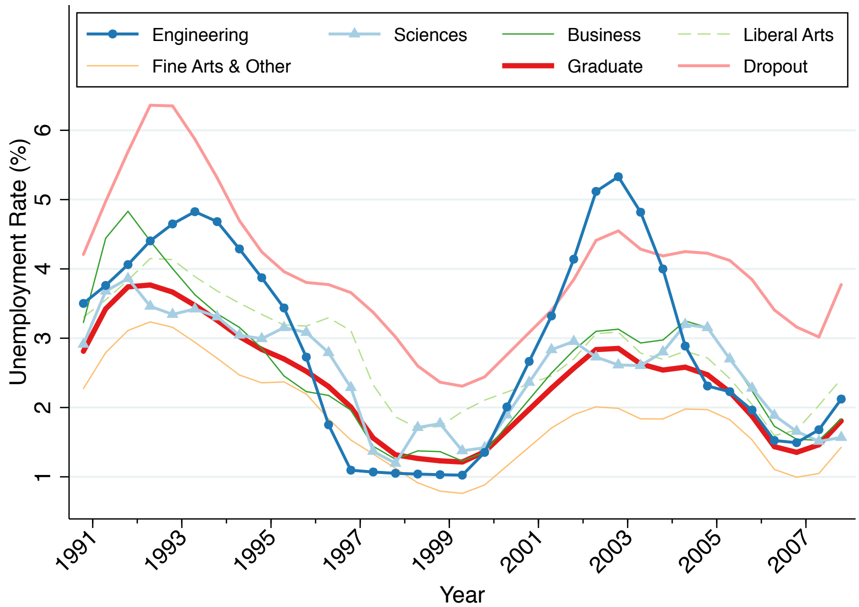

FIGURE 2.1: Unemployment rate by year

2003 rivaled that of college dropouts, which face uniformly higher unemployment rates than college graduates in all other years of the sample.

Furthermore, the disproportionate effect of the 2001-02 recession for engineering majors was unique relative to the previous recession in the early 1990s, and unusual given trends in the overall strength of the engineering labor market. While the unemployment rate for engineers rose to 4.82 percentage points in the spring of 1993 and was higher than that of every other major, it was not significantly larger than many other majors, such as business and liberal arts. In contrast, the unemployment rate for engineers in 2002 was over two percentage points higher than business majors, and over double that of liberal arts majors. Lastly, this increase in the unemployment is especially striking given the relative strength of the engineering labor market from 1995-2000, and in 2004-2007, when engineering majors experienced one of the lowest unemployment rates of all college graduates.

Figure 2.2, which plots the expected log wages by group over the same period,

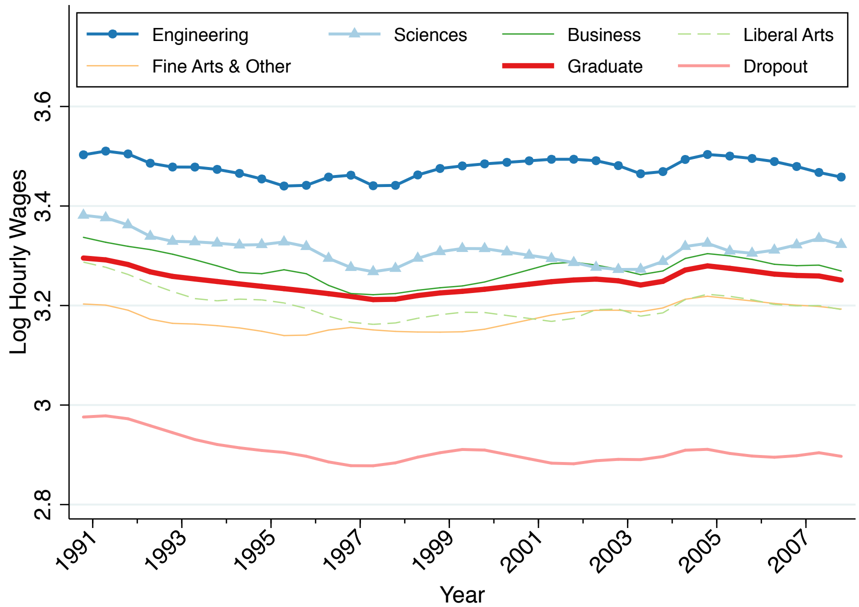

FIGURE 2.2: Log wages by year

also shows how this recession impacted the otherwise high labor market returns for engineering graduates. This graph confirms findings in the literature that show engineering and science majors have the highest expected earnings on average, which incentivizes students to select these majors. It also shows that the disproportionate impact of the 2001-02 recession on engineering graduates operated primarily through an increase in the unemployment rate, rather than a significant decrease in the earnings of employed individuals. Log wages for engineering graduates were relatively stable during the recession, and only fell by 0.03 log points in the spring of 2003 from the peak in the fall of 2001. Furthermore, this drop in log wages is consistent with that of graduates in other fields, unlike the relative increase in the unemployment rate for engineering graduates.

Finally, these labor market patterns also suggest that the effects of this labor market shock varied, depending on an individual's level of potential labor market experience.[17] Although Figure 2.3 shows that log wages for engineering graduates with up to five years of potential experience were similarly stable during the recession; Figure 2.4 shows that the unemployment rate for recent graduates does not increase as dramatically relative to other majors. Instead, the lower wage rate for recent science graduates shown in Figure 2.3 relative to Figure 2.2 suggest that the relative drop in the labor market value of an engineering degree may not be as severe in students' early careers. Hence, whether students account for returns to experience in their labor market forecasts may be important for explaining significant changes in their major choices during the recession.

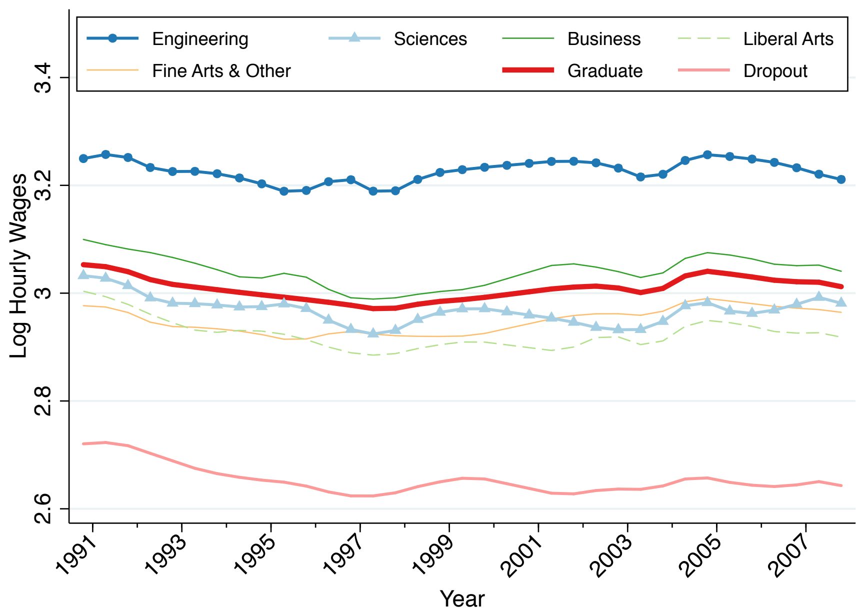

FIGURE 2.3: Log wages by year (0-5 yrs. potential experience)

2.1.4 Implementation: Students' information set

These results suggest students' information sets and the model must have three characteristics with respect to labor market data. First, labor market returns must

[17] I measure potential labor market experience as either the difference between the survey year and the year an individual earned their bachelor's degree, or the difference between the survey year and age 20 for individuals that dropped out of college.

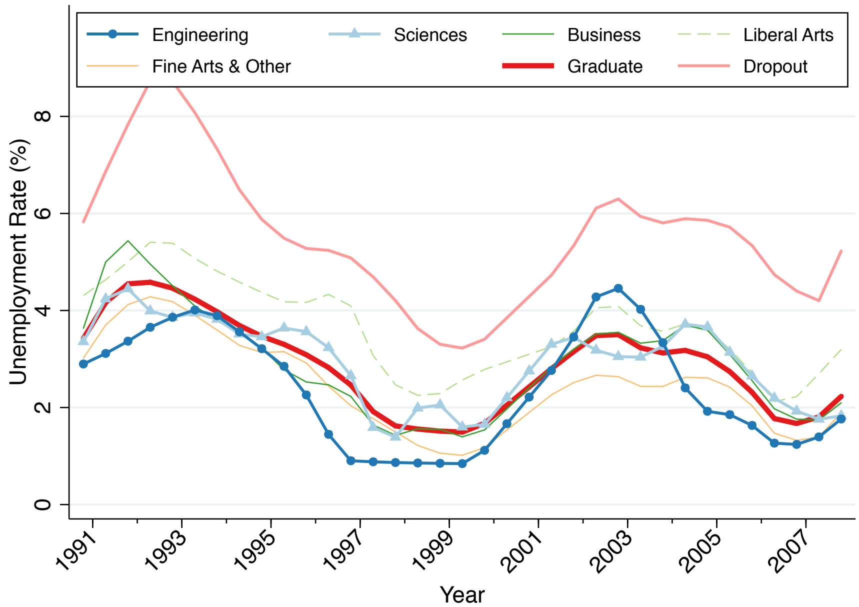

FIGURE 2.4: Unemployment rate by year (0-5 yrs. potential experience)

be allowed to vary by major. Second, returns should also vary by potential labor market experience. Finally, the model should account for labor market shocks that update at the same frequency students' choose majors, in order to accurately capture the labor market shocks in students' informations sets and their responses.

Results from the literature also confirm the need to control for these first two factors. Research on the role of potential experience in the labor market has long established that it is an important determinant of wages, which is documented in surveys that include Katz and Autor (1999), and more recently Acemoglu and Autor (2011). And while research that studies how students update their beliefs is more recent, it also confirms that students update them more strongly in response to major specific data, both for labor market data in Wiswall and Zafar (2015b) and Wiswall and Zafar (2015a).

I control for the contribution of each of these factors to labor market returns by estimating weighted OLS and Logit regressions of log wages and unemployment status that include these factors as regressors.[18] In practice, I allow both series to be a function of a major specific intercept; respondents' demographic and geographic characteristics X, calendar month and fixed effects for overlapping panels X_t,[19] major-specific fixed effects and returns to a quadratic in potential experience X_k, and calendar year fixed effects X_{kt} that also vary by major:

$$\log W_{kt} = \gamma_{1w} + \gamma_{2w}X + \gamma_{3w}X_t + \gamma_{4wk}X_k + \gamma_{5wk}X_{kt} + \epsilon_{wk} \tag{2.1}$$

$$\log\left(\frac{P_{kt}}{1-P_{kt}}\right) = \gamma_{1p} + \gamma_{2p}X + \gamma_{3p}X_t + \gamma_{4pk}X_k + \gamma_{5pk}X_{kt}$$

Once I have these estimates I generate separate series for wages and unemployment rates by major, demographic characteristics,[20] and levels of labor market experience. These series then enter students' information sets, where they are used to forecast the evolution of the labor market.

Finally, accurately capturing the labor market shocks in students' information sets imposes a constraint on the frequency that information on wages and unemployment rates update, which must match the semester-level frequency that students' switch majors. If the model constrains students' information sets to update less frequently, it will smooth out the magnitude of the labor market shocks that students actually use when they switch majors. To see this, suppose the series of unemployment rates generated by Equation 2.1 updates biannually. In this case the sequence of unemployment rates for engineering graduates generated by the model would pool the peak unemployment rate of 5.33 percentage points in fall 2002 with the far lower

[18] I weight each regression by the person frequency weights that are recorded for each individual in the SIPP.

[19] I include fixed effects by SIPP panel in order to control for the any effects that vary across the overlapping 1990-1993 SIPP panels. To do this I choose one reference panel for each calendar year from 1990-1995 and generate indicators for observations from each of the remaining panels. I then remove the average effects of these indicators.

[20] In practice I do not allow returns to vary by demographic characteristics. The labor market data is instead weighted by the empirical distribution of these characteristics.

relative employment rate of 3.32 percentage points in the spring of 2001. Hence the resulting series would significantly dampen the magnitude of the labor market shock generated by the 2001-02 recession, and potentially mask the shock which drives students' major switching behavior.

Similarly, this constraint on the labor market component of the model also requires students to make decisions at least as frequently as the labor market data updates. Suppose students' information sets update with the labor market data, but the model constrains them to choose majors biannually, as in two period models of major choice. This model would then pool the responses of students both before and after the peak of the labor market shock, and hence would dampen the observed major switching response to the recession.

2.2 Student beliefs

The next component of the model establishes how students use the shocks in their information set to form beliefs about the expected utility of the labor market. I focus in particular on how expected utility may deviate from the rational expectations value along three dimensions: the length of students' labor market forecasts, how persistent they believe these shocks to be, and how risk averse they are. Each of these three factors can significantly dampen or amplify the effects of a recession on students' behavior, depending on how they are specified. Furthermore, Betts (1996) and others propose a number of possible explanations for why individuals' forecasts could vary from standard models of the labor market. For example, younger college students may find forecasting especially burdensome, and could be constrained in the level of complexity and sophistication or horizon of their forecasts for the evolution of labor market shocks.

The first of these three dimensions, variance in the length of students' forecasts, is a common potential explanation for why their decisions could deviate from the predictions of a standard model. For example, a negative labor market shock could prompt a student to switch if they only value the next several years of returns, which are disproportionally affected by shocks. These myopic expectations are largely discounted in the literature on college major choice, as in Berger (1988) that found students' college major choices were consistent with the long run expectations for labor market returns by major, rather than myopic expectations. However, this result is based on individuals' future labor market earnings, instead of contemporaneous labor market shocks. Since myopic expectations could drive students to switch majors in response to a shock they observe in college, I include it by modifying the length of students' labor market forecasts.

In the rest of this section I describe how the model allows beliefs and expected utility to vary along the remaining two dimensions: first, how persistent students believe shocks to wage and unemployment rates are; and second, how risk averse students are when they form the expected utility of the labor market. Finally, I describe the implications of modifying beliefs along each of these two dimensions on students' major choices.

2.2.1 Forecasts of labor market values

Students' forecasting process begins when they learn about the value of shocks to the labor market. This happens at the beginning of each decision period, when nature reveals prevailing wage and unemployment rates.[21] Students then combine this new information with historical labor market data, and use the resulting data to estimate separate AR(1) processes for how shocks to wages and unemployment evolve from period to period:[22]

[21] In this section I use subscripts that denote major-specific shocks to and forecasts for wage and unemployment rates. However, these could either be major-specific or aggregate rates for all college graduates. In either case the rates for college graduates and dropouts follow separate processes.

[22] The length of my SIPP sample makes it difficult to accurately estimate these AR(1) parameters for early years of the sample. In practice I use observations from the whole sample to estimate

$$\log W_{kt+1} = c_{wk} + \rho_w \log W_{kt} + \epsilon_{wkt} \tag{2.2}$$

$$\log\left(\frac{P_{kt+1}}{1-P_{kt+1}}\right) = c_{pk} + \rho_p \log\left(\frac{P_{kt}}{1-P_{kt}}\right) + \epsilon_{pkt} \tag{2.3}$$

In each process, ρ is constrained to be the same across all majors and dropouts, hence shocks across majors follow roughly the same rate of decay.[23] However, the mean of each process is likely to vary across majors; for example, the wage rate for engineering majors is 34.2% higher than that for graduates in the liberal arts. I then let the drift parameters c_{wk} and c_{pk} vary by major, in order to allow the mean of each process to converge to these major-specific values.

Once students have these parameter estimates that determine the evolution of shocks, they use the following formula to generate the vectors of wages and unemployment rates for each period in their forecast of the labor market, $n \in \{t+1, \ldots, T_w\}$:

$$\log\left(\frac{P_{kt+n}}{1-P_{kt+n}}\right) = c_{pk} \sum_{t'=1}^{n} \rho_{pkt}^{t'-1} + \rho_{pkt}^{n} \log\left(\frac{P_{kt}}{1-P_{kt}}\right) \tag{2.4}$$

$$\min\left(\rho_p^*, 1 - \frac{c_{pk}}{P_{kt}}\right) \leqslant \rho_{pkt} \leqslant \max\left(\rho_p^*, 1 - \frac{c_{pk}}{P_{kt}}\right)$$

Where ρ_{pkt} denotes the persistence parameter that students actually use, which varies from the rational level ρ_p^* which best fits historical data, and the level that implies the labor market shock will be perfectly persistent.[24] Finally, students take these

these values, and I assume throughout that these wage and unemployment rates will follow similar persistence patterns over time. Hence students do not re-estimate the AR(1) parameters each period.

[23] I test for whether the persistence parameters vary by major for both processes, and am unable to reject the null that $\rho_k = \rho$ for all k. The p-values for the log wage and unemployment F-tests are 0.428 and 0.406, respectively.

[24] In this specification I assume that students will always use the optimal prediction for the drift parameter c_{pk}. However, I have run alternative specifications of the model where I allow students' estimates for c_{pk} vary over time, but these models do not perform as well as the baseline model. These results are consistent with the findings of Arcidiacono, Hotz, and Kang (2012), which find students' beliefs about the expected level of earnings are consistent with actual values.

forecasts of the labor market and form the expected utility of wages from entering the labor market in major k or from dropping out. They do so by first forecasting sequences of wages and unemployment rates after entering the labor market with a given set of parameters governing each AR(1) process. Students then use these sequences to form the expected utility of the labor market for each future period in their forecast, and discount and sum them to form the expected value of the labor market for each major and dropping out.

2.2.2 Predictions with rational expectations

One typical approach in the literature is to assume students have rational expectations over the future of the labor market, that they are forward looking, and exhibit relatively low levels of risk aversion. Hence, in this framework rational individuals form their forecasts of the sequence of wage and unemployment rates by using the optimal AR(1) parameters estimated using historical labor market data.[25] Since students are forward-looking, they generate these forecasts for the foreseeable length of their labor market career. And lastly this benchmark model follows the literature in assuming students have low levels of risk aversion, by using the log of wages to form the utility of wages.[26]

Figure 2.5 plots the expected value of the labor market for this standard rational expectations model, for students in their fourth semester of college. The persistence of each sequence is relatively low, with $\rho \approx 0.7$, and each returns to the mean relatively quickly even by the first year after graduation, when $\rho^3 \approx 0.34$. Therefore even a strong relative labor market shock to the market for engineering graduates, like

[25] In this and later sections I assume students forecast out 30 years from labor market entry. However, these results are consistent with labor market careers that are similar in length to the 12 years used in Berger (1988) or longer.

[26] See Arcidiacono (2004), Beffy, Fougere, and Maurel (2012), and Kinsler and Pavan (2014) for examples of other sequential college major choice models that use log specifications for wages or earnings.

the one driven by the 2001-02 recession, does not significantly decrease the expected labor market value of that degree at graduation for current students; which is why standard rational expectations models predict students' choices should be relatively stable even during the 2001-02 recession.

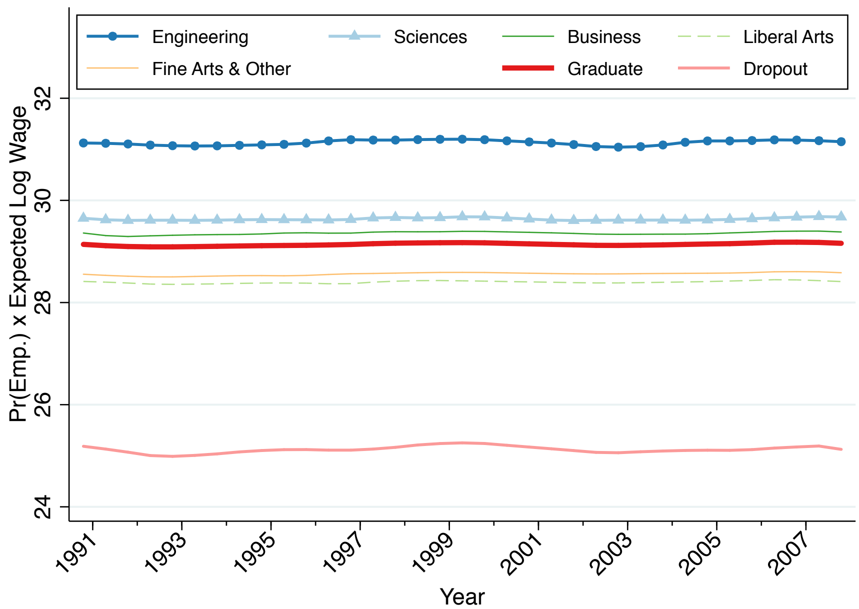

FIGURE 2.5: Expected LM value by year at $t = 4$ (Log, rational ρ)

However, the literature has also shown examples of college students deviating from the predictions of rational expectations models. Freeman (1976a) showed as far back as the 1970s that enrollments and graduation rates for engineers exhibited regular cobweb patterns: students reacted to strong labor market conditions and high wage premiums for engineers by enrolling at high rates, only to graduate four years later into a labor market with low wage premiums, saturated with engineers. Furthermore, research by Betts (1996) suggests that students early in their undergraduate careers have inaccurate labor market forecasts, showing that even for a sample of high ability students from the University of California San Diego, their median absolute error in predicting wages was 20%. While Arcidiacono, Hotz, and Kang (2012) found students' expected wages were more accurate in a sample of students from Duke, a cautious interpretation of this discrepancy suggests students' labor market forecasts may deviate from the rational expectations model. And since these deviations may influence students' choices in a way that drives their major choice and switching decisions, I make several additions to their beliefs and the forecasting process which can capture some of this deviation.

2.2.3 Additions to beliefs and forecasting process

The first addition I make to students' beliefs is to allow the persistence of the labor market shocks to vary from the optimal value ρ^*, since forecasts depend upon how persistent students assume labor market shocks are. I set the bounds on students' level of persistence in Equation 2.4, p_{pkt} in this example, to be between the level that best fits the historical data p_p^* and an upper bound such that the current labor market values for wages and unemployment persist perfectly into the future.[27] Figure 2.6 plots the expected value of the labor market with the same log utility and length of career as Figure 2.5, but with perfectly persistent labor market shocks. The levels of expected utility are directly comparable in these figures, since the only difference between these specifications is the persistence of shocks. The value of the engineering labor market decreases significantly in the aftermath of the 2001-02 recession in the model with perfectly persistent shocks relative to the rational expectations model. Therefore, the same shock could prompt engineering students to seek out alternative majors like the sciences or business if they perceive the shock to be highly persistent, while students with rational expectations would not.

[27] This value is set such that the mean of the AR(1) process is equal to the current labor market value; hence the series of forecast values immediately converges to the current value. I do this by setting the mean of the process μ equal to the current value and then solve the expression $\mu = \frac{c}{1-\rho}$ for ρ. Note that since the bounds of ρ depend on the current labor market value, these bounds must be calculated separately for each labor market value k in time period t. I use the same procedure to update the drift parameters in specifications where I allow the mean of each labor market process to vary by cohort.

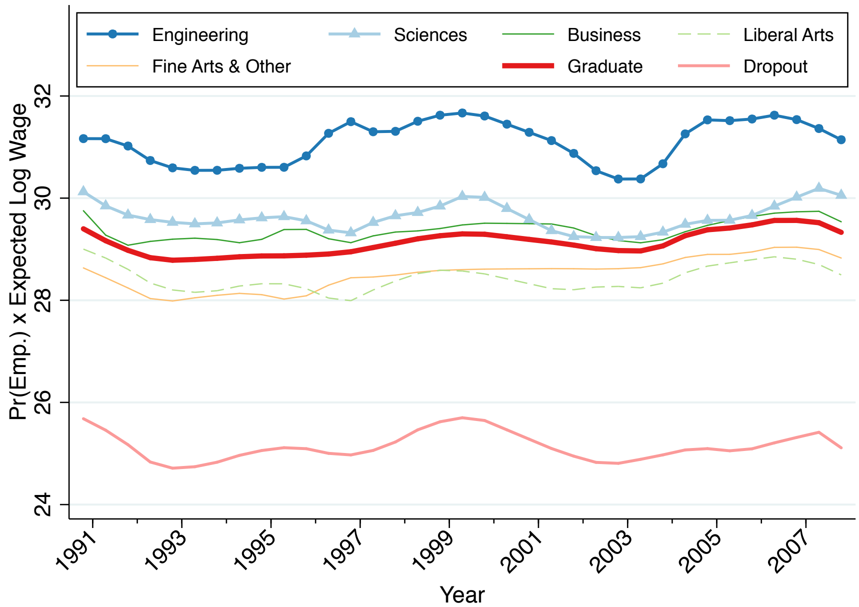

FIGURE 2.6: Expected LM value by year at $t = 4$ (Log, persistent ρ)

In these graphs expected utility simply takes the form of log wages, which I use in my baseline specification. However, since my framework model allows for alternative specifications for the utility from wages in Equation 2.13, I am able to generalize students' utility with a constant relative risk aversion (CRRA) function, where the utility of wages varies with students' level of risk aversion:

$$u_c = \frac{W^{1-r_a} - 1}{1 - r_a} \text{ , if } \quad r_a \neq 1 \tag{2.5}$$

$$u_c = \log W \text{ , if } \quad r_a = 1 \tag{2.6}$$

In one example, Beffy, Fougere, and Maurel (2012) study the effect of expected wages on major choice for a sample of French students that entered the labor market in the early- to mid-1990s; which experienced an economic expansion that disproportionally benefitted science, law, economics, and management graduates. They find low elasticities of major choice with respect to expected earnings, and that their finding is robust to both log and CRRA specifications for the expected utility of earnings. However, in this application they fix the level of risk aversion to a standard value of $r_a = 1.1$, which implies a level moderately higher than that of log utility, where $r_a = 1$. It could still be the case that students' reactions to labor market shocks are consistent with even higher levels of risk aversion than this standard value. For example, Buchinsky and Leslie (2010) also uses a CRRA specification for wages and allows for risk aversion levels up to $r_a = 2$, while Wiswall and Zafar (2015a) estimates risk aversion parameters as high as $r_a = 5$.[28] Figure 2.7 shows the compounding effect that these higher levels of risk aversion have on the value of the labor market during a recession, where unlike log utility specifications, students with this high level of risk aversion would perceive the value of an engineering degree to fall below that of nearly all other degrees during the recession.[29] Controlling for this will allow me to capture the compounding effects high levels of risk aversion may have on students' major choices.

2.2.4 Implications for students' major choices

Varying each of these three components of the expected utility from earnings could have an effect on students' response to labor market shocks. To see this, consider an engineering student that experiences the strong negative shock to wages and unemployment rates for engineering graduates during and after the 2001-02 recession.

First, if this student has myopic expectations they will put more weight on earnings soon after graduation, when the labor market returns of engineering graduates are relatively weaker, and less or no weight on later years after the labor market has

[28] Bliss and Panigirtzoglou (2004) compares the CRRA risk aversion parameters they estimate using data on options contracts with the literature. While some papers like Epstein and Zin (1991) estimate risk aversion parameters near one, their review finds most studies estimate levels that are much higher, such as Aït-Sahalia and Lo (2000) which estimates $r_a = 12.7$.

[29] Since changing the risk aversion parameter modifies students' preferences for earnings, I cannot directly compare the level of utility across Figures 2.5-2.7. However, Figure 2.7 is informative because it shows that students with high levels of risk aversion would perceive the relative value of an engineering degree to fall below that of alternative majors, while individuals with $r_a = 1$ would not.

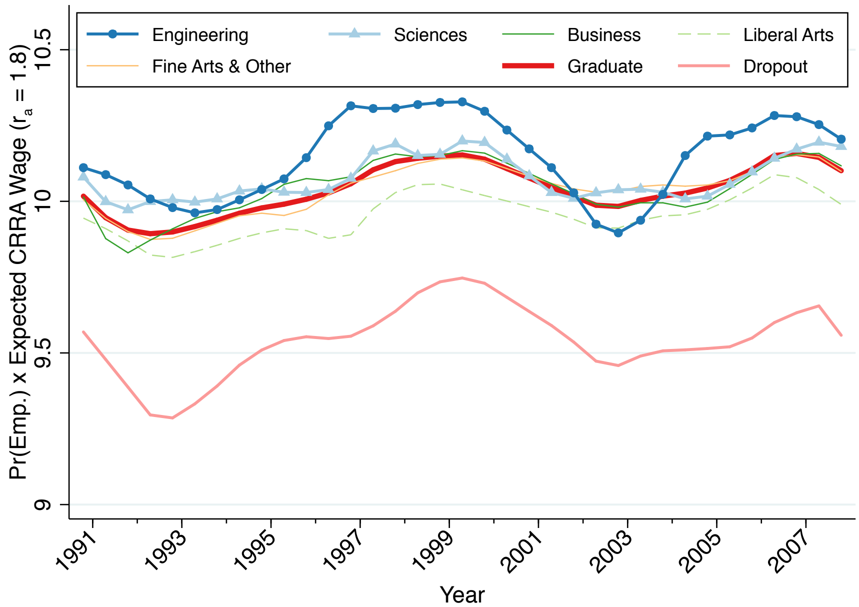

FIGURE 2.7: Expected LM value by year at $t = 4$ (CRRA $r_a = 1.8$, persistent ρ)

recovered. Likewise, a student that believes these labor market shocks will be highly or perfectly persistent will also discount the value of an engineering degree, relative to forecasts that use the "optimal" estimate for persistence. And lastly, high levels of risk aversion may compound the effects of myopic expectations and highly persistent shocks. Given the increase in the unemployment rate for engineering graduates during the recession, students with high levels of risk aversion would earn lower levels of expected utility from the labor market, relative to individuals with lower levels of risk aversion.

In each of these examples, particular modifications to student beliefs modified the utility of labor market earnings for engineering students. Since these earnings play a role in determining the future value of a major choice, each could prompt engineering students to switch our of engineering and into an alternative major in response to the 2001-02 recession and labor market shock.

2.3 Major switching costs

Since the shocks generated by the labor market and belief components of the model present potential incentives for students to switch majors, it is also important to control for the offsetting effort costs that deter students from switching majors. To see this, suppose students observe a labor market shock that prompts them to switch majors, but that they face significant effort costs to switch to this new major. A model that does not control for these effort costs would misinterpret these students' response, and assume that their low elasticity of major choice with respect to earnings is driven by an insensitivity to the shock, rather than burdensome switching costs.

However, these new components also impose additional requirements on my model of major choice, which are necessary to accurately control for these switching costs; and on the data required to estimate these complex switching costs. First, since students may decide to switch majors in any semester as labor market shocks and incentives to switch evolve, I must also explore whether switching costs also vary from period to period, or along other dimensions. Second, determining whether switching costs are heterogenous requires a large administrative dataset, which records students' major choice decisions on at least an annual basis. Finally, these data must also cover a time period before, during and after a significant labor market shock which disproportionally impacts earnings of certain majors, such as the 2001-02 recession. Otherwise, I will not be able to determine whether students' major choices respond to major-specific or aggregate labor market shocks.

In the rest of this section I describe the data from the University of Texas at Austin I use to solve these problems. Second, I document patterns from students' major choices that switching costs are heterogeneous across a number of dimensions, including: the major students switch from and into, the semester students are in when they switch majors, student demographics and performance, and whether a student is switching back to a prior major. Lastly, I use the results of these patterns to characterize switching costs in the model, which impose direct costs in the semester students switch, and indirect costs that operate through penalties in grades and delayed graduation.

2.3.1 Data: University of Texas at Austin

I use administrative data collected by the Texas Higher Education Opportunity Project (THEOP) from the University of Texas at Austin, which records application and linked transcript data on freshman applicants from 1992 to 2003. These data include information on 59,043 students with 360,385 semester-level transcript records, including students' declared major; number of credit hours taken; whether the record was from a fall, spring, or summer term; and the calendar year the term was taken. The application data also records basic measures of student demographics, standardized test scores, high school characteristics and rank, information on AP tests taken, and parent's level of education, among others. Finally, each record includes a measure for whether a student graduated during the sample period.[30]

Table 2.1 presents the summary statistics for individuals included in my sample, which match the characteristics of the student body described in UT Austin's internal reports.[31] Nearly 93% of students are in-state, and a similar proportion attended a public high school. This university is also typical for institutions in the United States for having a slight majority of female students, and students are also likely to come from a household where at least one parent has a bachelor's degree or more: only 17.3% of students fall into a "low-SES" household where parents' highest level

[30] The THEOP data does not include an indicator for whether students dropped out during the sample period. I therefore construct a measure where students are considered dropouts if they are not observed graduating and do not complete a semester in the final three periods I observe in the sample. The remaining students that do not graduate or drop out are considered "censored" observations.

[31] These figures are contained in the January 2015 revision of the University of Texas at Austin Accountability Report, available from the Texas Higher Education Coordinating Board (THECB).

of education is some college or lower. UT Austin is also a highly ranked institution with competitive academic programs across each of the five major categories I define.[32] Students are similarly high-achieving, and while I normalize SAT scores within application cohorts, they have a median composite score of 1,220 on a scale of 1,600, with median subscores of 600 on the verbal and 620 on the quantitative sections.

The final important feature to note about this sample is that it covers the implementation of the Texas "top 10%" law in the late-1990s, where public universities in Texas were required to admit any student in the top 10% of their graduating high school class, if they attended a public high school in Texas. This policy was instituted to increase the representation of underserved populations after universities were prohibited from admitting students according to traditional affirmative action policies. In practice, this policy greatly increased the proportion of students admitted from the top decile of Texas high schools, which is reflected in the high mean high school percentile rank of 87.2 in this sample. However, while UT Austin initially responded to the increase in applicants under the "top 10%" policy by increasing the number of students, they eventually instituted policies to limit the number of these applicants. While this policy does not have direct implications for the effect of labor market shocks on student major choices, I flexibly control for the time-varying effects of UT Austin's "top 10%" policy by adding fixed effects by application cohort.[33]

[32] UT Austin includes University of California-Berkeley, UCLA, UCSD, University of Michigan, UNC Chapel Hill, and University of Wisconsin-Madison in their National Comparison Group (NCG) of peer institutions. See UT Austin's Institutional Reporting, Research, and Information Systems (IRRIS) reporting publications for more information.

[33] I include these fixed effects in the dropout and graduation Logits, as a quadratic time trend in the GPA regressions, but exclude them from the second stage estimator in my preferred specification. However, my results are robust to including these fixed effects in the second stage estimator.

Table 2.1: Summary statistics by student

Variable	Mean	SD	Min	Max
Male	47.5%	(0.4994)	0	1
Black	3.4%	(0.1806)	0	1
Hispanic	13.4%	(0.3411)	0	1
Asian or Pac. Isl.	17.1%	(0.3768)	0	1
HS percentile rank	87.235	(11.8318)	0	99.9
SAT verbal pct. rank	50.445	(28.8674)	1	99
SAT quant. pct. rank	50.541	(28.9653)	1	99
Took any AP test	58.5%	(0.4927)	0	1
High school				
In-state	92.9%	(0.2574)	0	1
Private	8.8%	(0.2827)	0	1
% Disadvantaged	19.7%	(0.1643)	0	100
Parent's education				
High school	6.2%	(0.2414)	0	1
Some college	11.1%	(0.3138)	0	1
Graduate	34.8%	(0.4762)	0	1
Imputed values				
SAT subscores	4.0%	(0.1951)	0	1
HS percentile rank	8.5%	(0.2787)	0	1
HS private	2.0%	(0.1386)	0	1
% Disadvantaged	28.2%	(0.4502)	0	1
Graduate	56.8%	(0.4953)	0	1
Dropout	11.9%	(0.3238)	0	1
Censored	31.3%	(0.4637)	0	1
Observations				59,043

SAT verbal and quantitative percentile ranks for each student are generated within their enrollment cohort. Missing high school percentile ranks are set equal to the median value within an application cohort. Missing SAT subscores are imputed by taking a student's national percentile score on the SAT composite, and calculating the subscores that correspond to the same percentiles within the verbal and quantitative sections. The percentage of disadvantaged students is generated by a regression of high school demographic characteristics on the % disadvantaged within an application cohort.

2.3.2 *Direct costs*

I start documenting the switching cost patterns for this sample of students by examining the major switching behavior of engineering students before, during and after the 2001-02 recession. Figure 2.8 plots the fraction of engineering students that switch to another major at the end of a semester, both by the semester that students switch and the application cohort students belong to. This figure shows that, for students in the pre-recession fall 1998 entry cohort, switching rates out of engineering vary significantly by the semester students are in: rates increase from a low of roughly 2.57 percentage points in the first semester, quickly rise to a peak of nearly 9.85 percentage points in the third semester, and then gradually fall to a low of 3.09 percentage points in the seventh semester as students progress through their undergraduate career.

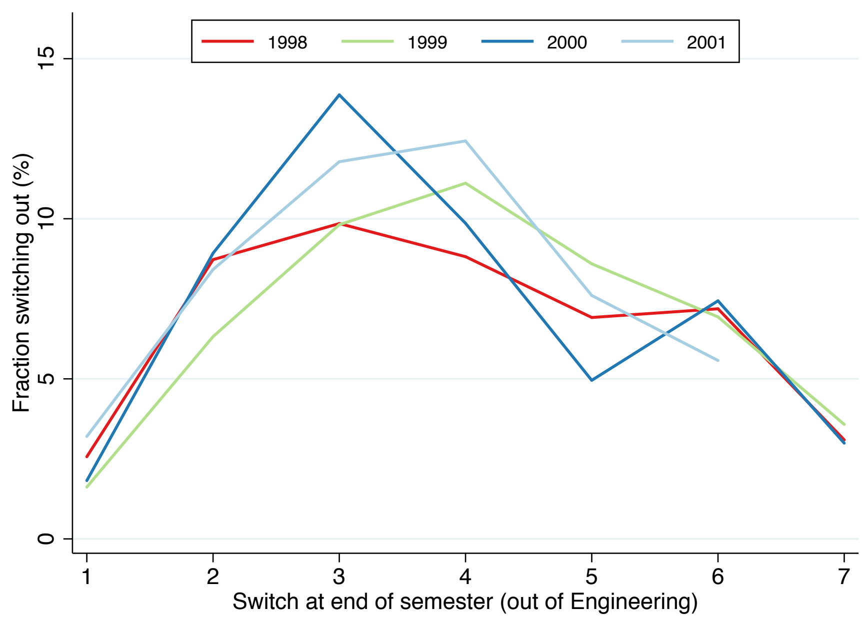

FIGURE 2.8: Switching rate out of Engineering by semester and cohort

Students in the fall 2000 entry cohort, which experienced the peak of the effects of the 2001-02 recession on the labor market for engineering graduates in their third semester, switched out of engineering at a rate that was nearly 40.8% higher than the rate for the 1998 cohort. Students in the fall 1999 entry cohort exhibited similar increases in their switching rate out of engineering when they learned of the shock in their fourth and fifth semesters. However, these students, who learned of the shock one year later into their undergraduate career, had a response that was significantly dampened relative to the fall 2000 cohort. In particular, the switching rate out of engineering increased by a smaller rate of 24.1% for students switching at the end of their fifth semester.

These results suggest both that students responded to the labor market shock generated by the 2001-02 recession by switching out of engineering majors, and that this effect could be dampened by switching costs that vary over time. The response of students in the fall 2000 cohort imply that this labor market shock presented a strong incentive to switch majors. And while the fall 1999 cohort observed the same labor market shock, their response was relatively attenuated; even though they were more likely to directly experience the effects of the shocks by graduating one year earlier, on average. Similarly, when I compare the switching rates of students in the fall 1998 entry cohort to earlier cohorts, I see that the rate of students switching out of engineering falls steadily throughout the 1990s, from a mean of 13.1 percentage points for the 1992-1996 cohorts, to the low of 9.85 percentage points for the 1998 cohort, which coincides with the economic expansion of the 1990s. Hence, students may also respond to positive shocks by choosing to remain in engineering at higher rates.

Many of these engineering students switched to alternative majors in the sciences, and are likely to have done so for two reasons: first, engineering and science majors are close substitutes, with students commonly switching between each; and second, the sciences featured a labor market that was relatively unaffected by the 2001-02 recession. For example, 60.9% of engineering students that switch majors at the end of their third semester switch into the sciences. While just 11.5% of students switching out of the sciences at the end of their fourth semester switch into engineering, this is likely due to the more stringent requirements for switching into engineering. Indeed, the rate of students switching into engineering from other majors is much lower, like fine arts, where 6.0% of students switching out at the end of their fourth semester switch into engineering. Figure 2.9 shows the same switching rate graph for students in these science majors, and also shows complementary patterns. For students in the fall 1998 entry cohort, switching rates out of the sciences increase from approximately 5.22 percentage points in the first semester to a peak of nearly 11.36 percentage points in the fourth semester, before falling sharply in the fifth and later semesters. The fall 2000 cohort, by contrast, switched out of the sciences in this peak switching semester at a rate of 7.20 percentage points, which was 36.6% lower than a typical cohort.

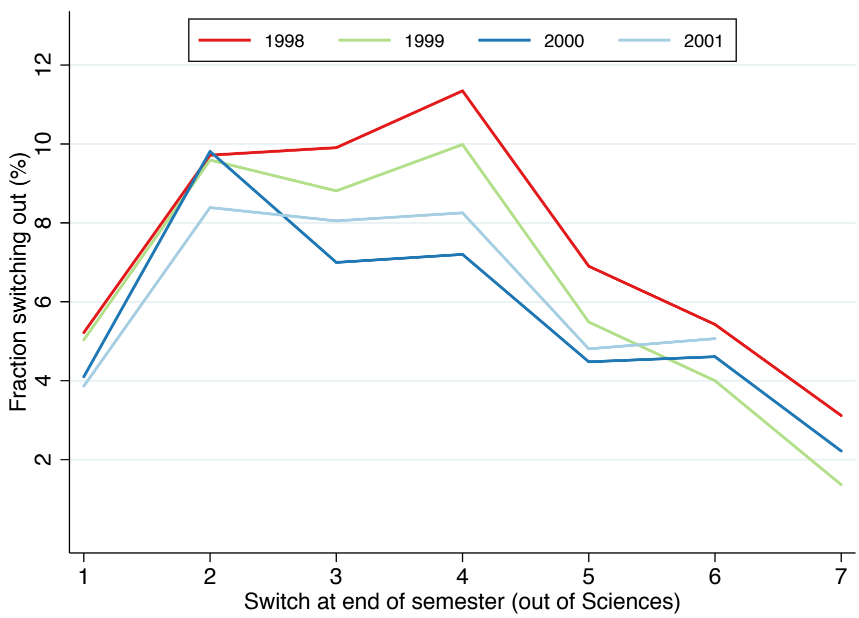

FIGURE 2.9: Switching rate out of the Sciences by semester and cohort

These results for science majors offer additional suggestive evidence that students respond to major specific labor market shocks, and that the pattern of switching costs varies both over time and by the major students switch from and into. First, these results are consistent with students responding to major specific, rather than aggregate shocks. If students were simply responding to the same aggregate shock, we would not expect to see engineering students increase their rate of switching into other majors, while science majors decrease their switching rate. And second, if switching costs were constant over time and equal across the majors that students switch from and into, the response to the labor market shock would not be attenuated for older cohorts, and we would not expect to see the switching rates from engineering to the sciences vary significantly from the rate of sciences students switching to engineering.

I first control for these patterns in switching costs by implementing a direct effort cost for switching majors:

$$
\begin{aligned}
c_{kt} = {} & \gamma_{c1jk} S_{jkt} + \gamma_{c2jk} S_{jkt} \mathbb{1}(t=3) + \gamma_{c3jk} S_{jkt} \mathbb{1}(t=4) + \gamma_{c4jk} S_{jkt} \mathbb{1}(t=5) \\
& + \gamma_{c5jk} S_{jkt} \mathbb{1}(t \geqslant 6)\, t + \gamma_{c4k} S_{mk} + \gamma_{c6k} S_{bk} + \gamma_{c7k} S_k * \mathbb{1}(\text{LowSES})
\end{aligned}
\tag{2.7}
$$

In this specification, the effort cost of switching majors is a function of the major that students switch from and into S_{jkt}, which is also allowed to vary by the semester the switch occurs; an indicator for whether a student has already switched majors S_{mk} or if they are switching back to a previously-held major S_{bk}; and if the student comes from a low SES household.

2.3.3 Indirect costs: Grades and probability of graduating

Switching majors may also impose an indirect cost on students by both impacting their academic performance in later semesters, or by reducing the probability that they graduate on time. To see these two effects, consider the case of an engineering student that switches to a business major in response to the 2001-02 recession.

Although it may be the case that engineering is more rigorous than business, the coursework in business may be different enough that switchers may need to satisfy a number of course requirements in order to graduate within the same timeframe as their previous major. As a result, these students may be forced to take a heavier course load, reduce their available time to study for each class, and therefore earn a lower semester GPA that they would have if they started in business instead of engineering. Similarly, this same student may cope with this increased number of course requirements by taking the same course load and spreading these courses over additional semesters, which would mechanically decrease their probability of graduating on time.

I flexibly control for these indirect effects of switching majors on performance and the probability of graduating by allowing the grade production and stochastic graduation processes to depend on whether students switched majors in prior periods. I present regression results for an OLS regression of semester GPA and a Logit for the probability of graduating at the end of the semester in Tables 2.2 and 2.3.[34] These results are consistent with indirect effects that switching costs may have through grades and the probability of graduation. In particular, I find significant negative effects of switching majors on students' semester GPA, and a positive effect for switching back to a previously-held engineering major. Likewise, students that only recently switched into their current major, as measured by the total fraction of credits taken in their current major, are significantly less likely to graduate at the end of a given period.

[34] I use these same regression results in the sequential model of major choice that I specify in Section 5.

Table 2.2: Semester GPA regressions

	Semester GPA (by major)				
	Engineering	Sciences	Business	Liberal Arts	Fine Arts
Constant	2.687*** (0.0951)	3.012*** (0.0984)	2.906*** (0.0928)	3.039*** (0.0798)	3.446*** (0.0981)
Male	−0.048*** (0.0069)	−0.036*** (0.0051)	−0.032*** (0.0040)	−0.039*** (0.0038)	−0.106*** (0.0050)
Black	−0.092*** (0.0156)	−0.083*** (0.0134)	−0.117*** (0.0112)	−0.079*** (0.0095)	−0.107*** (0.0110)
Hispanic	−0.066 (0.0096)	−0.068 (0.0086)	−0.058*** (0.0070)	−0.042*** (0.0058)	−0.048*** (0.0073)
Asian or Pac. Isl.	−0.034*** (0.0065)	−0.020*** (0.0058)	−0.034*** (0.0049)	−0.014** (0.0063)	−0.033*** (0.0068)
HS percentile rank	0.003*** (0.0003)	0.003*** (0.0003)	0.003*** (0.0002)	0.002*** (0.0002)	0.003*** (0.0002)
SAT verbal pct. rank	0.000* (0.0001)	0.000*** (0.0001)	0.000*** (0.0001)	0.000*** (0.0001)	0.000*** (0.0001)
SAT quant. pct. rank	0.000*** (0.0001)	0.001*** (0.0001)	0.000 (0.0001)	0.000*** (0.0001)	0.000*** (0.0001)
Took any AP test	0.009 (0.0061)	0.023*** (0.0055)	0.013*** (0.0042)	0.009** (0.0038)	0.012** (0.0048)
High school					
In-state	−0.026** (0.0116)	−0.025** (0.0114)	−0.028*** (0.0086)	−0.039*** (0.0079)	−0.017* (0.0094)
Private	0.048*** (0.0124)	0.044*** (0.0111)	0.000 (0.0081)	0.014** (0.0071)	0.008 (0.0088)
% Disadvantaged	−0.001*** (0.0002)	−0.001*** (0.0002)	−0.002*** (0.0002)	−0.002*** (0.0001)	−0.001*** (0.0002)
Lagged GPA					
Semester	0.293*** (0.0269)	0.155*** (0.0285)	0.315*** (0.0251)	0.245*** (0.0214)	0.280*** (0.0289)
Semester^2	−0.002 (0.0051)	0.029*** (0.0052)	−0.005 (0.0044)	0.010*** (0.0039)	0.009* (0.0051)
Cumulative	−1.398*** (0.0691)	−1.206*** (0.0718)	−1.163*** (0.0679)	−1.164*** (0.0591)	−1.546*** (0.0737)
Cumulative^2	0.314*** (0.0123)	0.258*** (0.0125)	0.252*** (0.0117)	0.250*** (0.0103)	0.314*** (0.0126)
Switched majors					
Switch in current semester	−0.022 (0.0200)	−0.044*** (0.0117)	0.061*** (0.0093)	0.061*** (0.0084)	0.069*** (0.0095)
Multiple majors	−0.035 (0.1337)	−0.177* (0.0973)	−0.040 (0.0911)	−0.231*** (0.0722)	0.241*** (0.0937)
Lagged GPA interactions					
Semester	−0.120 (0.0788)	0.177*** (0.0552)	0.059 (0.0564)	0.164*** (0.0449)	−0.065 (0.0563)
Semester^2	0.032** (0.0144)	−0.026*** (0.0098)	−0.005 (0.0095)	−0.024*** (0.0079)	0.014 (0.0096)
Cumulative	0.171*** (0.0168)	0.209*** (0.0157)	0.197*** (0.0143)	0.227*** (0.0127)	0.331*** (0.0149)
Cumulative^2	−0.027*** (0.0028)	−0.030*** (0.0026)	−0.029*** (0.0024)	−0.033*** (0.0022)	−0.053*** (0.0025)
Dispersion	0.5911	0.5345	0.4852	0.5138	0.4902
Observations	51,186	49,735	69,639	90,688	55,873

Standard errors in parentheses. Model also includes a quadratic in application cohort, missing public or private HS, and imputation flags for SAT subscores, high school rank, and high school % disadvantaged. SAT percentile ranks are calculated relative to a students' freshman class and are reported on a scale of $[0, 99.9]$. Lagged GPA is interacted with an indicator for whether students have ever switched majors in the current or prior periods. Sample includes only individuals with nonzero semester GPAs and lagged cumulative GPAs.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.3: Logit model of graduation

	Graduate at the end of semester (by major)									
	Engineering		Sciences		Business		Liberal Arts		Fine Arts	
Constant	−42.459***	(1.2750)	−33.059***	(1.0580)	−45.766***	(1.0156)	−38.846***	(0.7941)	−33.929***	(1.1436)
Male	−0.037	(0.0606)	0.040	(0.0450)	−0.364***	(0.0401)	−0.356***	(0.0331)	−0.387***	(0.0489)
Black	−0.189	(0.1478)	0.105	(0.1198)	−0.382***	(0.1130)	−0.371***	(0.0844)	−0.178*	(0.1077)
Hispanic	−0.507***	(0.0886)	−0.061	(0.0774)	−0.273***	(0.0685)	−0.247***	(0.0506)	−0.213***	(0.0707)
Asian or Pac. Isl.	−0.020	(0.0568)	−0.046	(0.0510)	−0.239***	(0.0481)	−0.424***	(0.0543)	−0.148**	(0.0653)
Low SES	0.002	(0.0689)	−0.148**	(0.0645)	−0.058	(0.0576)	−0.118***	(0.0442)	−0.065	(0.0563)
HS percentile rank	0.009***	(0.0025)	0.005*	(0.0026)	0.009***	(0.0020)	0.005***	(0.0014)	−0.003	(0.0021)
SAT verbal pct. rank	−0.003***	(0.0009)	0.000	(0.0009)	−0.003***	(0.0008)	−0.005***	(0.0007)	−0.005***	(0.0009)
SAT quant. pct. rank	0.004***	(0.0012)	0.003***	(0.0010)	0.003***	(0.0009)	−0.001	(0.0007)	−0.004***	(0.0010)
Took any AP test	0.070	(0.0541)	0.111**	(0.0492)	0.143***	(0.0422)	0.101***	(0.0336)	−0.004	(0.0461)
Semester GPA	2.503***	(0.2828)	2.171***	(0.2376)	2.106***	(0.2226)	1.998***	(0.1633)	0.485**	(0.2391)
Semester GPA^2	−0.372***	(0.0482)	−0.335***	(0.0406)	−0.308***	(0.0378)	−0.305***	(0.0281)	−0.051	(0.0392)
High school										
In-state	0.100	(0.1028)	−0.088	(0.1014)	−0.154*	(0.0874)	−0.224***	(0.0725)	0.039	(0.0957)
Private	0.317***	(0.1145)	0.205**	(0.1024)	−0.072	(0.0847)	0.159**	(0.0646)	0.134	(0.0899)
% Disadvantaged	0.000	(0.0020)	−0.002	(0.0018)	−0.002	(0.0016)	−0.004***	(0.0013)	−0.007***	(0.0017)
Semester										
Fall	−0.926***	(0.0477)	−1.206***	(0.0447)	−1.606***	(0.0384)	−1.353***	(0.0313)	−1.217***	(0.0424)
Semester number	2.842***	(0.2085)	3.721***	(0.1819)	4.740***	(0.1901)	3.857***	(0.1421)	3.793***	(0.1973)
Semester number^2	−0.111***	(0.0117)	−0.176***	(0.0103)	−0.221***	(0.0117)	−0.182***	(0.0086)	−0.170***	(0.0109)
Lagged values										
Cumulative GPA	4.765***	(0.5956)	0.661	(0.5217)	4.130***	(0.4612)	3.253***	(0.3689)	2.522***	(0.5283)
Cumulative GPA^2	−0.593***	(0.0962)	0.035	(0.0843)	−0.542***	(0.0752)	−0.381***	(0.0607)	−0.224***	(0.0848)
Total credits	0.207***	(0.0130)	0.149***	(0.0115)	0.221***	(0.0135)	0.213***	(0.0104)	0.149***	(0.0107)
Total credits^2	−0.001***	(0.0001)	−0.001***	(0.0000)	−0.001***	(0.0001)	−0.001***	(0.0001)	−0.001***	(0.0000)
Switched majors										
HS percentile rank	−0.014**	(0.0057)	−0.004	(0.0048)	−0.008**	(0.0035)	−0.009***	(0.0026)	0.000	(0.0037)
% credits in prior major	−2.114***	(0.3090)	−1.020***	(0.2213)	−2.097***	(0.1960)	−1.364***	(0.1314)	−0.256	(0.1614)
LL		-60,166.79		-52,010.54		-67,721.53		-83,824.49		-67,261.68
Observations		19,480		20,459		29,734		39,107		24,258

Table reports point estimates, with standard errors in parentheses. Model also includes indicators for application cohort, missing public or private HS, an indicator for multiple majors, interactions with the major switch indicator and a quadratic in lagged total credits, and imputation flags for SAT subscores, high school rank, and high school % disadvantaged. SAT percentile ranks are calculated relative to a students' freshman class and are reported on a scale of $[0, 99.9]$. Sample includes only individuals with nonzero semester GPAs and lagged cumulative GPAs, and students that are defined as "able to graduate," which includes: students in their 6th or higher semester, and students with transcripts that are not censored by the end of data collection in 2004.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.4 Framework model of major choice

In this section I specify a sequential model of college major choice that meets two requirements. First, it establishes a standard framework that nests other sequential models of major choice used in the literature. This allows me to document the effect of labor market shocks on major choice, while still controlling for other factors that research has shown to be significant determinants of major choice. Second, I have shown above that allowing major choice to depend on labor market shocks requires a number of key modifications to the model, which include increasing the frequency of student major choice decisions and allowing individuals to graduate at different times; along with more complex labor market, belief and switching cost components. This framework is also flexible enough to allow for these additional features, which allows me to capture the effects of labor market shocks on major choice.

In this framework model, students make a sequence of choices $d_{kt} \equiv (d_{0t}, \ldots, d_{Kt})$ during their time in college $1 \leqslant t \leqslant T$ in order to maximize the expected present value of utility in Equation 2.8, which includes the flow utility in period t, $u_{kt}(z_t)$ from either attending college in major $k \in \{1, \ldots, K\}$ or dropping out $k = 0$.[35] Students must also consider the effect of their choices on the expected utility of earnings from entering the labor market by graduating in a particular major at the end of the period $d_{kt} = 1, d_t^g = 1$ or by dropping out $d_{0t} = 1$. Finally, entering the labor market is an absorbing state, so students are not allowed to continue in college once they have graduated, or reenter if they choose to drop out.[36][37]

[35] In the rest of this section I drop the i subscript denoting individuals for simplicity.

[36] This implies $d_{kt'} = 0 \quad \forall \quad k \in \{0, \ldots, K\}, t' > t$ if $d_t^g = 1$ or $d_{0t} = 1$.

[37] While other papers like Arcidiacono et al. (2014) allow students to re-enter college after choosing to drop out, adding this feature to my model would increase the complexity of the conditional choice probability estimation I use to make the model tractable. Although this may be a worthwhile tradeoff in some settings, only 506 individuals in my sample of nearly 60,000 have an attendance gap of 12 months or more, which would indicate that they dropped out for a meaningful length of time and returned.

$$E\left\{\sum_{t=1}^{T}\sum_{k=0}^{K}\beta^{t-1}\left[d_{kt}\left(1-d_{0,t-1}\right)\left(1-d_{t-1}^{g}\right)\left[u_{kt}\left(z_{t}\right)+\epsilon_{kt}\right]\right.\right. \tag{2.8}$$

$$\left.\left.+d_{k,t-1}\left(d_{0,t-1}+d_{t-1}^{g}\right)u_{wk}\right]\right\}$$

This model deviates from many sequential models of major choice by allowing both for a variable number of choice periods T, and for a stochastic process that determines if and when a student is allowed to graduate $\{d_1^g, \ldots, d_T^g\}$.[38] In other papers, such as Arcidiacono (2004), college consists of two choice periods split roughly between a student's first two and later years in college, which is followed by the labor market. As a result, students are allowed to switch majors or drop out only once in their career; and if students choose to remain in college, they are constrained to graduate and enter the labor market at the end of the second period.

Relaxing these constraints allows this model to capture additional variation in students' major choice behavior. Increasing the number of choice periods allows me to separately identify the incentives and costs that students face when choosing to switch majors at different times in their freshman and sophomore years, or in their junior and later years. Also, allowing graduation to follow a stochastic process will let me capture heterogeneity in whether and when students are allowed to graduate. For example, this would allow students in engineering programs to take longer to graduate than students in liberal arts, or for switching majors to delay graduation by a semester as students make up required courses, as I discuss in Section 2.3.3.

Despite these changes and increased flexibility, this framework also nests these

[38] Beffy, Fougere, and Maurel (2012) also endogenizes students' length of time in college in a study of the effects of expected earnings on the major choices of French students. However they do this with respect to the level of degree students attain: e.g. associate's, bachelors, or master's degree. In this paper by contrast, I endogenize how long it takes students to complete their bachelor's degree.

standard two period models of major choice with deterministic graduation,[39] and students' decisions follow a similar sequence of events:

1. Students enter a period and face a choice set where they can either remain in college in one of K majors, or drop out.

2. If students choose to remain in college, nature reveals information on shocks and whether a student is able to graduate at the end of the period.

3. In the next period, students that did not drop out or graduate update their beliefs and make another major choice or dropout decision. Students that drop out or graduate enter the labor market and earn the corresponding flow utility from the labor market.[40]

4. At the end of the final period T, students that choose to remain in college graduate with certainty $d_T^g = 1$ and enter the labor market.

Once students choose a major k, they receive a flow utility of pursuing the major, which is a linear function of demographic measures and high school characteristics X_D, ability X_A, measures that vary with experience in college and time trends X_t, grade point average in that period G_{kt}, and a time-varying effort cost c_{kt}.

$$u_{kt} = \psi_{1k} + \psi_{2k} X_D + \psi_{3k} X_A + \psi_{4k} X_t + \psi_{5k} G_{kt} + c_{kt} \tag{2.9}$$

Previous studies have shown that measures of demographic characteristics are important determinants of relative major preferences.[41] Fixed effects also play an

[39] Collapsing this framework to the standard model requires two changes. First, set the number of choice periods to $T = 2$. Second, constrain the graduation process such that the probability of graduating is zero for all but the final period $\{d_t^g\}_{t \leqslant 1} = 0$, when students that remain in college graduate with certainty $d_2^g = 1$.

[40] Graduates earn the labor market returns that correspond to the final major they selected in college. This specification allows, but does not require returns to be major-specific.

[41] Gemici and Wiswall (2014) and Zafar (2013), for example, find that differences in preferences for majors across genders play a significant role in explaining gender gaps in major choice.

important role in studies that use data from multiple institutions or cohorts. For example, including effects for each cohort from a longitudinal sample of students can capture changes in the relative preferences for majors over time, peer effects from increasing admission standards, and administrative changes that influence the quality of instruction in a major, among others.

It is also likely that the varying direct effort and financial costs of studying in a particular program influence major choice.[42] However, the financial structure of higher education, particularly in the United States, makes studying the effects of direct financial costs difficult. Since students in the US may apply, enroll and pay for attending an institution before or without choosing a major, it is likely that the the within-period, within-institution differences in attendance costs between majors are small relative to this total cost of attendance.

As a result, this and other papers that study the major choice decisions of students in the US focus primarily on per-period effort costs that are allowed to vary by major. In particular, studies find that the flow utility and relative appeal of majors is much lower for students that switch in from a different major in the previous period, relative to a similar student that did not switch in from a different major.[43] These deterrent effects of switching majors are then interpreted as the flow "switching cost" of changing majors.

Lastly, student ability and academic performance also have significant and interacting effects on major choice. Stinebrickner and Stinebrickner (2011), Arcidiacono (2004) and others have shown that relative verbal and quantitative ability drive sorting into majors; where students with relatively high quantitative ability prefer majors

[42] Hoxby and Turner (2013) and Bettinger et al. (2012) suggest that students' perceptions about the direct cost of college play a role in their application decisions, and Stinebrickner and Stinebrickner (2008) finds that credit constraints play a role in students' decision to drop out of college. Hence, students may also consider direct financial costs when choosing a major if, for example, a particular choice of major would require them to finance an extra year of college, or pursue a graduate degree.

[43] Arcidiacono (2004)

like engineering and the sciences, while students with relatively high verbal ability prefer liberal and fine arts. Students' academic performance, typically measured by GPA, also influences the flow utility of each major. In some models performance impacts choices by manipulating labor market returns; but better grades could also directly influence preferences by exposing students to better peers, earn them more attention from professors, or improve their course selection with preferential registration times and access to enriched classes. However, since ability is also likely to play a role in determining a student's academic performance, I need an exclusion restriction in the production function for grades in order to separately identify the effects of ability and grades on major choice.

$$\begin{aligned} G_{kt} &= \psi_{1gk} + \psi_{2gk} X_{Gt} + \psi_{3gk} G_{t-1} + \psi_{4gk} \bar{G}_{t-1} + \epsilon_{gkt} \\ X_{Gt} &= [X_{Dg}, X_A, X_t] \end{aligned} \tag{2.10}$$

In this specification, per-period grades are a linear function of a set of covariates X_{Gt} that include demographic characteristics that determine grades X_{Dg},[44] ability and measures that vary with experience in college. Since these terms are all included in the flow utility, I also include lagged measures of a student's previous period GPA G_{t-1} and cumulative GPA $\bar{G}_{t-1}$.[45] These lagged terms allow me to separately identify the contribution of ability and performance to the flow utility of a major, since the utility depends only on student performance in that semester.

Students then weigh the flow utility of pursuing a major k, determined by these functions for flow utility and grades, against the implications of that choice on the value of their best option in the next period. They then choose the major that

[44] In my later specification, these include all of the characteristics in the flow utility except a student's socioeconomic status. I assume that SES will only impact grades through the properties of a student's high school, as measured by the proportion of disadvantaged students.

[45] Note that these measures of performance are not major specific. I will include terms that capture varying returns to GPA for students that switch from other majors in the switching costs section.

maximizes this weighted flow and future utility, which is measured by the conditional value function:

$$v_{kt}\left(d^g_{t-1}, z_t\right) = u_{kt}\left(z_t\right) + \beta \int_{z_{t+1}} V_{t+1}\left(d^g_t, z_{t+1}\right) f_{kt}\left(z_{t+1}|z_t\right) \tag{2.11}$$

This conditional value function for choosing major k in period t is equal to the flow utility of that major net of the preference shock ϵ_{kt}, plus the expectation of the discounted one period ahead value function taken over every possible future state z_{t+1}.[46] Then the future value term is simply the expected value of the discounted stream of flow utilities, from college and the labor market, that students receive from the optimal sequence of choices $\{d^*_{kt'}\}_{t+1 \leqslant t' \leqslant T}$.

$$\begin{aligned} V_{t+1}\left(d^g_t, z_{t+1}\right) = \max_{d^*_{kt'}} E_t \Bigg\{ \sum_{t'=t+1}^{T} \sum_{k=0}^{K} \beta^{t'-t-1} \big[& d^*_{kt'}\left(u_{kt'}\left(z_{t'}\right) + \epsilon_{kt'}\right) \\ & + d_{k,t'-1}\left(d_{0,t'-1} + d^g_{t'-1}\right) u_{wkt'} \big] \Bigg| z_{t+1} \Bigg\} \end{aligned} \tag{2.12}$$

The value of the labor market enters the objective function as the expected discounted value of earnings u_{wkt} of working in a career that starts at time t. These returns can take many forms, but in this paper the labor market is characterized by wages W_{kt} and the probability of being unemployed in each period P_{kt}.[47]

$$u_{wkt}\left(z_t\right) = \alpha_w E_t \left[\sum_{t'=t}^{T_w} \beta^{t'-t}\left(1 - P_{kt'}\right) u_c\left(W_{kt'}\right) \Bigg| z_t \right] \tag{2.13}$$

[46] A student's choices impact the future value term through the state variable z_t, which includes the vector of choices from the previous period $\{d_{k,t-1}\}$.

[47] I normalize the flow utility of being unemployed in period t to be equal to zero to simplify the flow utility from labor market earnings.

Students forecast the evolution of wages and unemployment conditional on their information set at time t, and then use those series to form the expected utility from wages $u_c(W_{kt'})$ for a career that lasts for $T_w - t$ periods. This specification allows for several alternative forms for expected utility. While log wages are one common form, I also allow for CRRA utility.

2.5 Estimation

While this new sequential model of college major choice is able to capture both the incentives and obstacles to switching majors that are prompted by labor market shocks, the new framework and components it requires to do so makes it difficult to implement. In this section I discuss the constraints imposed by the model, and how I use techniques from the literature on conditional choice probabilities (CCPs) to solve these problems and develop a feasible model.

2.5.1 Model constraints

The properties of the labor market, student beliefs, and major switching costs impose several constraints on the model that complicate the estimation procedure: a relatively large choice set, multiple periods, stochastic graduation, and a complex set of switching costs. The cumulative effect of each of these constraints is that the future value term becomes very complex. In particular, since students are allowed to choose between dropping out and five majors in each period, and I observe students taking up to 13 semesters at UT Austin, the future value term must account for all of the possible flow utilities in up to 12 subsequent periods. Furthermore, it must also account for how those flow utilities change depending on the choices students make in each subsequent period, and how those choices affect their probabilities of graduating at the end of each period.

This sort of complicated future value term is both very cumbersome to specify and computationally intensive to estimate using traditional full information maximum likelihood (FIML) methods, which are best suited to future value terms with small numbers of future periods. In fact, one of the reasons why sequential models of college major choice with few periods are appealing is because the future value term in these models is relatively simple, with only $t-1$ periods. However, since students are able to make a terminal choice in each period by dropping out, this future value term can be greatly simplified by using conditional choice probability methods, developed by Hotz and Miller (1993).

2.5.2 Conditional choice probabilities

In the rest of this section I implement this solution by deriving the CCP estimator for this model, where I follow Arcidiacono and Miller (2011). I start by considering students' objective function for their decisions in college, which I motivated in the framework model of major choice in Equation 2.8. This expected discounted sum of flow utilities from college and the labor market is maximized by a Markov decision rule which determines the optimal choice for a student in period t, given the state $\left\{d_{t-1}^{g}, z_t\right\}$ and choice-specific preference shock $\epsilon_t \equiv (\epsilon_{0t}, \ldots, \epsilon_{Kt})$; I then define the optimal decision rule as $d_t^o\left(d_{t-1}^{g}, z_t, \epsilon_t\right) \equiv \left(d_{0t}^o\left(d_{t-1}^{g}, z_t, \epsilon_t\right), \ldots, d_{Kt}^o\left(d_{t-1}^{g}, z_t, \epsilon_t\right)\right)$.

Given this optimal decision rule, I can obtain the vector of conditional choice probabilities $p_t\left(d_{t-1}^{g}, z_t\right)$ of a student in state $\left\{d_{t-1}^{g}, z_t\right\}$ making every choice $k_t \in \{0, \ldots, K\}$ by integrating out the preference shocks for each component of the optimal decision rule vector:

$$p_{kt}\left(d_{t-1}^{g}, z_t\right) \equiv \int d_{kt}^o\left(d_{t-1}^{g}, z_t, \epsilon_t\right) g\left(\epsilon_t\right) d\epsilon_t \tag{2.14}$$

I can then redefine the *ex-ante* value function in state $\left\{d_{t-1}^{g}, z_t\right\}$ as the expected value at period t of the discounted optimal stream of flow utilities after the graduation indicator is revealed to the individual, but before the preference shock ϵ_{kt} is revealed:[48]

$$V_t\left(d^g_{t-1}, z_t\right) = E_t\left\{\sum_{t'=t}^{T}\sum_{k=0}^{K}\beta^{t'-t}\left[d_{kt'}\left(1-d_{0,t'-1}\right)\left(1-d^g_{t'-1}\right)\left[u_{kt'}\left(z_{t'}\right)+\epsilon_{kt'}\right]\right.\right. \tag{2.15}$$

$$\left.\left. + d_{k,t'-1}\left(d_{0,t'-1}+d^g_{t'-1}\right)u_{wk}\right]\right| z_t\Bigg\}$$

This new specification for the future value function allows me to then modify the conditional value function, by modeling the future value term explicitly as an expectation over whether students are allowed to graduate at the end of the current period or not. The conditional value function[49] of making choice $k_t \in \{0, \ldots, K\}$ is still equal to the flow utility of college net of the preference shock ϵ_{kt}, plus the expectation of the future value term. However, I simply take the expectation of this future value term across the possible future states $\{d^g_t, z_{t+1}\}$ of graduating at the end of period $t+1$ or not:

$$v_{kt}\left(0, z_t\right) = u_{kt}\left(z_t\right) + \beta\sum_{z_{t+1}}\left[Pr\left(d^g_t = 0|z_{t+1}\right)V_{t+1}\left(0, z_{t+1}\right) + \right. \tag{2.16}$$

$$\left.\left(1-d_{0t}\right)Pr\left(d^g_t = 1|z_{t+1}\right)V_{t+1}\left(1, z_{t+1}\right)\right]f_{kt}\left(z_{t+1}|z_t\right)$$

where the value function after graduation reduces to $V_{t+1}\left(1, z_{t+1}\right) = \sum_{k=1}^{K} d_{kt}u_{wk}$.[50]

While $V_{t+1}\left(0, z_{t+1}\right)$ is a complex expression in the conditional value functions for decisions to remain in college $k_t \in \{1, \ldots, K\}$, it reduces to the labor market

[48] The timing works as follows: the student makes choice d_{kt}, the preference shock ϵ_{kt} is revealed, and then the graduation indicator d^g_t is revealed.

[49] The conditional value function at time t is deterministic for $d^g_{t-1} = 1$ since graduating is an absorbing state and no choices are made in subsequent periods $t' \geqslant t$. Hence, for the rest of this analysis I am only concerned with the value function in state $\left\{d^g_{t-1} = 0, z_t\right\}$.

[50] The above equations imply $d_{k,t-1} \in z_t$. Since the decision to drop out in period t is an absorbing state, $d_{0t} = 1 \Rightarrow d^g_t = 0$ and $V_{t+1}\left(1, z_{t+1}\right)$ is undefined. Hence, in the conditional value function I must multiply $V_{t+1}\left(1, z_{t+1}\right)$ by $\left(1-d_{0t}\right)$.

return u_{w0} if a student decides to drop out. Hence the conditional value function for dropping out in period t is:

$$v_{0t}(0, z_t) = u_{0t}(z_t) + \beta u_{w0} \tag{2.17}$$

Because this terminal choice greatly simplifies the conditional value function, I will only be considering the conditional value functions for choices $k \in \{1, \ldots, K\}$ in the derivation that follows. However, this simplification of terminal choices is the key feature that the CCP estimation method uses to simplify the future value term. Next, I derive the correspondence that allows me to substitute this expression in for the future value term of other non-terminal choices.

Hotz and Miller (1993) show that a mapping $\psi_k(p_t(0, z_t))$, exists between the CCPs I derived in equation 2.14 and the difference between the future value and conditional value functions:[51]

$$\psi_k[p_t(0, z_t)] \equiv V_t(0, z_t) - v_{kt}(0, z_t) \tag{2.18}$$

Then I can get an expression for the conditional value function in terms of the one period ahead conditional value function and the CCP mapping for an arbitrary choice $j \in \{0, \ldots, K\}$. I do so by solving this equivalence for $V_t(0, z_t)$, shifting it forward one period and substituting the resulting expression into the conditional value function at period t:

$$\begin{aligned} v_{kt}(0, z_t) = u_{kt}(z_t) + \beta \sum_{z_{t+1}} [& Pr(d_t^g = 0 | z_{t+1}) [v_{j,t+1}(0, z_{t+1}) + \psi_j [p_{t+1}(0, z_{t+1})]] \\ & + (1 - d_{0t}) Pr(d_t^g = 1 | z_{t+1}) u_{wk}] f_{kt}(z_{t+1} | z_t) \end{aligned} \tag{2.19}$$

[51] Since dropping out at the start or graduating at the end of period t is an absorbing state, individuals make no further choices and $p_{kt'}\left(d_{t'-1}^g, z_{t'}\right) = 0 \; \forall \; k \in \{0, \ldots, K\}, t' > t$. Hence I will only consider the non-trivial CCPs $p_{kt'}(0, z_{t'})$.

However, since the future choice j I selected to create this expression is arbitrary, I can set it to the terminal choice of dropping out in the next period: $j = 0$. Furthermore, since the conditional value function for dropping out is a simple, one period function of the flow utility and the labor market returns from dropping out in equation 2.17, I can use it to greatly simplify the expression for every conditional value function $v_{kt}(z_t)$:

$$v_{kt}(0, z_t) = u_{kt}(z_t) + \beta \sum_{z_{t+1}} \left[Pr\left(d_t^g = 0 | z_{t+1}\right) \left[u_{0,t+1}(z_{t+1}) + \beta u_{w0} + \psi_0 \left[p_{t+1}(0, z_{t+1}) \right] \right] \right.$$
$$\left. + \left(1 - d_{0t}\right) Pr\left(d_t^g = 1 | z_{t+1}\right) u_{wk} \right] f_{kt}(z_{t+1} | z_t) \tag{2.20}$$

Finally, by assuming the preference shocks follow a GEV distribution with schooling options $k \in \{1, \ldots, K\}$ in one nest and entering the labor market as a college dropout $k = 0$ in the other, I get a closed form expression for the mapping $\psi_0(p_t(0, z_t))$:

$$\psi_0(p_t(0, z_t)) = \gamma - \ln(p_{0t}(0, z_t)) \tag{2.21}$$

where γ is the Euler-Mascheroni constant.

It is important to note that the above derivation of the conditional value function assumes a state $\left\{d_{t-1}^g, z_t\right\}$ that is discrete. In this setting, z_t includes a student's major in the prior period d_{t-1},[52] and the vector of labor market shocks to wages and unemployment rates. While the major in the prior period is discrete, the labor market shocks are continuous random variables. However, since I specify a process for how students forecast the evolution, and hence their expectation, of the future values of these terms, I do not need to explicitly control for their evolution in $f_{kt}(z_{t+1}|z_t)$.

[52] This term is undefined for the first period.

With this specification, the estimation reduces to a simple multinomial Logit model, where the probability of choosing major k during period t is given by:

$$Pr\left(d_{kt}=1|d_{t-1}^{g}=0,z_t\right)=\frac{\exp\left(\frac{v'_{kt}(0,z_t)}{\mu_t}\right)\left[\sum_{k\neq 0}\exp\left(\frac{v'_{tk}(0,z_t)}{\mu_t}\right)\right]^{\mu_t-1}}{1+\left[\sum_{k\neq 0}\exp\left(\frac{v'_{tk}(0,z_t)}{\mu_t}\right)\right]^{\mu_t}} \tag{2.22}$$

where the differenced conditional value function $v'_{kt}\left(0,z_t\right)$ is given by:

$$\begin{aligned}
v'_{kt}\left(0,z_t\right)=&v_{kt}\left(0,z_t\right)-v_{0t}\left(0,z_t\right) \\
=&u_{kt}\left(z_t\right)-u_{0t}\left(z_t\right)-\beta u_{w0} \\
&+\beta\sum_{z_{t+1}}\left[Pr\left(d_t^g=0|z_{t+1}\right)\left[u_{0t}\left(z_t\right)+\beta u_{w0}+\psi_0\left[p_{t+1}\left(0,z_{t+1}\right)\right]\right]\right. \\
&\left.+\left(1-d_{0t}\right)Pr\left(d_t^g=1|z_{t+1}\right)u_{wk}\right]f_{kt}\left(z_{t+1}|z_t\right)
\end{aligned} \tag{2.23}$$

2.6 Results

Figures 2.8 and 2.9, which I discussed in Section 4, suggest that engineering students responded to the significant increase in the unemployment rate for engineering graduates, shown in Figure 2.1, by switching to alternative majors that were less hard hit by the 2001-02 recession. I complement these figures by estimating my structural model with a benchmark rational expectations specification: students forecast the value of a labor market career that lasts for 30 years, with “optimal” AR(1) parameters for log wages and unemployment that best fit the historical data, and a log specification for the utility of wages.[53]

In the following section I present results from this baseline model, and demonstrate the relevance of the labor market in students’ major choices and switching

[53] This corresponds to a CRRA utility specification with $r_a = 1$.

behavior, and the role switching costs play in mediating their response to labor market shocks and other incentives to switch majors. I then turn to the role deviations from rational expectations play in explaining students' behavior by reestimating the model with alternative specifications for students' labor market beliefs and preferences for risk. I conclude with simulation exercises designed to illustrate the welfare implications of having students base their major choice decisions on beliefs about the labor market that deviate from rational expectations.

2.6.1 Relevance of labor market values

I allow the value of the labor market to operate on students' major choices though two channels. First, the expected present value of utility from labor market earnings can enter students' objective functions directly at the end of a given semester for students that are able to graduate. Second, the labor market enters indirectly at the end of future semesters, which I transform into the conditional choice probability of dropping out in the next semester. The intuition for this specification is that, if I observe the probability of a student dropping out in their current major increase relative to alternative majors, it must reflect a lower value of that major in the future. Hence, that student will be more likely to switch majors.

The structural model estimates in Table 2.4 show the direct effect of the labor market on students' choices. I first constrain the value of the expected utility from the labor market to be equal across all majors, because otherwise students would value monetary returns at different rates, despite constructing these returns in the same way across each major.[54] The positive parameter estimate on expected wages suggests that students are more likely to choose majors with higher expected wages, all else equal.

[54] To see this, assume engineering and the sciences both have the same value of expected utility from labor market wages. If I allow the value of wages to vary across majors, students could place different values on identical expected utility from wages.

Table 2.4: Structural utility parameter estimates

	Major				
	Engineering	Sciences	Business	Liberal Arts	Fine Arts
Constant	−4.756*** (0.4237)	−3.846*** (0.4590)	−6.292*** (0.5659)	−5.860*** (0.4582)	−1.691*** (0.5060)
Fall semester	0.554** (0.2702)	−2.031*** (0.2527)	−0.863** (0.3358)	−0.812*** (0.2360)	−0.412* (0.2489)
Male	0.657*** (0.0384)	0.115*** (0.0360)	0.330*** (0.0382)	0.044 (0.0338)	−0.173*** (0.0362)
Black	0.477*** (0.0887)	0.558*** (0.0895)	0.397*** (0.0952)	0.142* (0.0815)	0.361*** (0.0845)
Hispanic	0.192*** (0.0559)	0.203*** (0.0548)	0.170*** (0.0599)	0.060 (0.0508)	0.106** (0.0539)
Asian or Pac. Isl.	0.258*** (0.0464)	0.279*** (0.0457)	0.451*** (0.0505)	−0.201*** (0.0487)	−0.072 (0.0495)
Low SES	−0.087 (0.0527)	−0.006 (0.0543)	0.004 (0.0665)	−0.009 (0.0544)	−0.002 (0.0559)
SAT verbal pct. rank	−0.006*** (0.0007)	−0.002*** (0.0007)	−0.006*** (0.0007)	0.005*** (0.0007)	−0.004*** (0.0007)
SAT quant. pct. rank	0.005*** (0.0008)	0.002*** (0.0008)	0.001* (0.0008)	−0.008*** (0.0007)	−0.005*** (0.0008)
HS percentile rank	−0.005*** (0.0015)	0.001 (0.0015)	−0.006*** (0.0016)	−0.009*** (0.0013)	−0.007*** (0.0014)
Took any AP test	0.063* (0.0365)	0.043 (0.0364)	−0.043 (0.0388)	−0.012 (0.0341)	−0.062* (0.0363)
High school					
In-state	0.156** (0.0690)	0.152** (0.0701)	0.069 (0.0730)	0.109* (0.0657)	0.132** (0.0672)
Private	−0.194*** (0.0686)	−0.197*** (0.0675)	0.055 (0.0678)	0.114* (0.0599)	−0.054 (0.0633)
% Disadvantaged	−0.001 (0.0013)	0.001 (0.0013)	−0.003** (0.0014)	0.002* (0.0012)	0.003** (0.0013)
Predicted performance					
Semester GPA	4.111*** (0.2913)	3.365*** (0.3089)	5.204*** (0.3775)	5.286*** (0.3085)	2.167*** (0.3357)
Semester GPA^2	−0.616*** (0.0527)	−0.522*** (0.0544)	−0.713*** (0.0657)	−0.774*** (0.0540)	−0.200*** (0.0571)
Switching cost	Major				
Prior major	Engineering	Sciences	Business	Liberal Arts	Fine Arts
Engineering		−5.140*** (0.2280)	−5.702*** (0.3165)	−3.126*** (0.2302)	−4.325*** (0.2732)
Sciences	−3.742*** (0.3111)		−5.575*** (0.3685)	−1.939*** (0.1964)	−1.831*** (0.2097)
Business	−6.995*** (0.5325)	−7.197*** (0.6755)		−3.249*** (0.2681)	−6.504*** (0.3624)
Liberal Arts	−6.672*** (0.6842)	−6.142*** (0.4066)	−6.725*** (0.3360)		−3.766*** (0.2459)
Fine Arts	−4.601*** (0.4058)	−5.299*** (0.3825)	−11.836*** (0.4005)	−4.524*** (0.2547)	
Interactions					
Multimajor	−0.217 (0.1929)	−0.022 (0.1619)	−0.299** (0.1185)	−0.141 (0.1094)	−0.135 (0.1316)
Switch back	1.419*** (0.0983)	0.726*** (0.0789)	1.851*** (0.0833)	0.708*** (0.0685)	0.931*** (0.0704)
Low SES	0.154 (0.1720)	−0.108 (0.1165)	−0.284*** (0.0918)	−0.110 (0.0676)	0.078 (0.0922)
Labor market					All Majors
Expected wages					1.374*** (0.0046)
LL					-95,000.34
Observations					324,742

Table reports point estimates, with standard errors in parentheses. Model also includes indicators for application cohort, imputed SAT subscores and imputed high school rank. Sample includes all individuals that were excluded from the estimation of the first-stage semester GPA, graduation, and dropout regressions.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The first column of Table 2.6 reports the parameter estimates for the Logit I use to predict the CCP of dropping out in the next semester, conditional on students' choices in the current semester. Since only approximately 7,000 student-semester level records are associated with a dropout in the following semester, I pool the dropout regression across all majors and semesters. The parameter estimate on the change in students' expected utility from wages suggests that students are less likely to drop out if their forecast of the expected earnings for a major increase relative to their prediction in their first semester. Hence if students are currently enrolled in that major, they will be less likely to switch to an alternative major, holding the returns to other majors constant. These results are also consistent with the behavior of engineering students during the 2001-02 recession, which switch out of engineering as the expected utility from wages in engineering falls more than alternative majors.

2.6.2 Role of switching costs

These structural estimates suggest that the value of major-specific labor market returns play a role in students' major choice, and that shocks to these values incentivize switching behavior; however these results also suggest that major switching costs play a significant role in students' choices by offsetting incentives to switch. I model these offsetting effects as operating through three channels: first, switching costs can directly impose an effort cost on students that make a switch in a given period; second, they can operate indirectly by influencing students' predicted performance; and finally, they can also operate indirectly by influencing the probability that students graduate at the end of a given period.

I present a subsample of the direct switching costs in Table 2.4, which include costs that vary by the major that students switch from and into, whether they are switching for a second or later time or back into a previously-held major, and

Table 2.5: Structural utility switching cost parameter estimates

By prior major	Engineering	Sciences	Business	Liberal Arts	Fine Arts
			Major		
Engineering		−5.140*** (0.2280)	−5.702*** (0.3165)	−3.126*** (0.2302)	−4.325*** (0.2732)
Sciences	−3.742*** (0.3111)		−5.575*** (0.3685)	−1.939*** (0.1964)	−1.831*** (0.2097)
Business	−6.995*** (0.5325)	−7.197*** (0.6755)		−3.249*** (0.2681)	−6.504*** (0.3624)
Liberal Arts	−6.672*** (0.6842)	−6.142*** (0.4066)	−6.725*** (0.3360)		−3.766*** (0.2459)
Fine Arts	−4.601*** (0.4058)	−5.299*** (0.3825)	−11.836*** (0.4005)	−4.524*** (0.2547)	
Cumulative GPA					
Engineering		0.394*** (0.0660)	−0.380*** (0.0807)	−0.724*** (0.0717)	−0.155* (0.0800)
Sciences	−0.215** (0.0916)		−0.401*** (0.0963)	−0.772*** (0.0612)	−0.712*** (0.0645)
Business	0.634*** (0.1467)	0.704*** (0.1931)		−0.375*** (0.0810)	0.662*** (0.1006)
Liberal Arts	0.120 (0.1941)	0.375*** (0.1169)	0.434*** (0.0976)		−0.163** (0.0736)
Fine Arts	−0.100 (0.1175)	0.319*** (0.1106)	1.781*** (0.1076)	−0.013 (0.0744)	
Other					
Multimajor	−0.217 (0.1929)	−0.022 (0.1619)	−0.299** (0.1185)	−0.141 (0.1094)	−0.135 (0.1316)
Switch back	1.419*** (0.0983)	0.726*** (0.0789)	1.851*** (0.0833)	0.708*** (0.0685)	0.931*** (0.0704)
Low SES	0.154 (0.1720)	−0.108 (0.1165)	−0.284*** (0.0918)	−0.110 (0.0676)	0.078 (0.0922)
Semester interactions					
Third semester	Engineering	Sciences	Business	Liberal Arts	Fine Arts
Engineering		0.984*** (0.1560)	2.682*** (0.2507)	1.157*** (0.1824)	0.938*** (0.1943)
Sciences	0.523*** (0.1859)		2.381*** (0.2682)	0.546*** (0.1552)	0.372** (0.1658)
Business	0.943*** (0.2807)	−0.174 (0.3591)		0.678*** (0.1976)	1.618*** (0.2075)
Liberal Arts	0.987*** (0.3915)	0.399 (0.2489)	1.152** (0.1980)		0.640*** (0.1707)
Fine Arts	0.075 (0.2535)	0.158 (0.2108)	2.085*** (0.2107)	0.579*** (0.1744)	
Fourth semester					
Engineering		1.542*** (0.1690)	1.780*** (0.2741)	1.268*** (0.1929)	0.931*** (0.2073)
Sciences	0.478** (0.2134)		1.324*** (0.3086)	0.851*** (0.1760)	0.722*** (0.1860)
Business	1.197*** (0.2910)	−0.285 (0.3666)		0.788*** (0.2142)	0.443* (0.2475)
Liberal Arts	0.920** (0.4109)	0.403 (0.2553)	0.316 (0.2298)		0.447** (0.1903)
Fine Arts	−0.304 (0.2974)	−0.537** (0.2464)	1.610*** (0.2322)	0.800*** (0.1872)	
Fifth semester					
Engineering		1.088*** (0.1678)	2.699*** (0.2595)	1.516*** (0.1889)	1.575*** (0.1967)
Sciences	0.304 (0.2074)		2.134*** (0.2812)	1.188*** (0.1621)	0.627*** (0.1787)
Business	1.043*** (0.2913)	0.195 (0.3362)		1.194*** (0.2010)	0.886*** (0.2320)
Liberal Arts	0.477 (0.4311)	0.344 (0.2507)	0.613*** (0.2122)		0.714*** (0.1776)
Fine Arts	0.160 (0.2557)	−0.448* (0.2341)	1.976*** (0.2191)	0.840*** (0.1762)	
Sixth+ semester x t					
Engineering		0.060*** (0.0209)	0.317*** (0.0333)	0.202*** (0.0226)	0.131*** (0.0239)
Sciences	−0.132*** (0.0270)		0.195*** (0.0362)	0.061*** (0.0189)	−0.022 (0.0201)
Business	−0.079** (0.0401)	−0.074* (0.0417)		0.090*** (0.0235)	0.035 (0.0269)
Liberal Arts	−0.034 (0.0503)	−0.021 (0.0286)	−0.073*** (0.0272)		−0.066*** (0.0211)
Fine Arts	0.029 (0.0271)	−0.111*** (0.0264)	0.112*** (0.0281)	0.029 (0.0212)	

Semester switching cost interactions are broken out by the semester a student is in when they first start their new major. For example, if a student switches from engineering in their third semester to the sciences in their fourth, they will incur a major switching cost of $-5.140 + 1.542 = -3.598$, in addition to the contribution from other switching costs. Table reports point estimates, with standard errors in parentheses. Model also includes indicators for application cohort, imputed SAT subscores and imputed high school rank. Sample includes all individuals that were excluded from the estimation of the first-stage semester GPA, graduation, and dropout regressions.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

their family socioeconomic status.[55] These parameter estimates are consistent with my expectations. Namely, all of the switching costs that vary by students' prior and new major are negative, which implies that, holding all else equal, the relative appeal of choosing a particular major will be lower for students that would have to switch

[55] I proxy for socioeconomic status by controlling for parents' level of education, where a student is defined as low SES if neither of their parents have a college degree. See Table 2.5 for the full set of switching costs that vary by semester and students' cumulative GPA at the time they switch.

Table 2.6: Logit model of dropout

	Rational		Best Fit		Full Persistence	
Constant	85.181	(54.7302)	29.003***	(7.8573)	19.093***	(4.1490)
Male	−0.052*	(0.0310)	−0.052*	(0.0310)	−0.052*	(0.0310)
Black	−0.275***	(0.0707)	−0.275***	(0.0707)	−0.276***	(0.0707)
Hispanic	−0.143***	(0.0456)	−0.143***	(0.0456)	−0.143***	(0.0456)
Asian or Pac. Isl.	−0.112***	(0.0404)	−0.112***	(0.0404)	−0.111***	(0.0404)
Low SES	−0.123***	(0.0395)	−0.124***	(0.0395)	−0.124***	(0.0395)
HS percentile rank	0.008***	(0.0012)	0.008***	(0.0012)	0.008***	(0.0012)
SAT verbal pct. rank	0.005***	(0.0006)	0.005***	(0.0006)	0.005***	(0.0006)
SAT quant. pct. rank	0.000	(0.0007)	0.000	(0.0007)	0.000	(0.0007)
Took any AP test	−0.174***	(0.0305)	−0.173***	(0.0305)	−0.173***	(0.0305)
High school						
In-state	−0.407***	(0.0615)	−0.407***	(0.0615)	−0.408***	(0.0615)
Private	−0.078	(0.0612)	−0.078	(0.0612)	−0.078	(0.0612)
% Disadvantaged	0.003***	(0.0011)	0.003***	(0.0011)	0.003***	(0.0011)
Semester						
Semester	−0.702***	(0.0258)	−0.697***	(0.0258)	−0.696***	(0.0259)
Semester^2	0.064***	(0.0022)	0.064***	(0.0022)	0.064***	(0.0022)
Fall	−0.575***	(0.0304)	−0.574***	(0.0304)	−0.574***	(0.0304)
Switched majors						
Current semester	0.152**	(0.0639)	0.154**	(0.0639)	0.154**	(0.0639)
Multiple majors	0.509	(0.3140)	0.504	(0.3139)	0.503	(0.3138)
Semester switched	−0.050**	(0.0251)	−0.050**	(0.0251)	−0.050**	(0.0251)
HS percentile rank	−0.006	(0.0036)	−0.006	(0.0036)	−0.006	(0.0036)
GPA						
Semester GPA	−2.597***	(0.0657)	−2.595***	(0.0657)	−2.595***	(0.0657)
Semester GPA^2	0.323***	(0.0152)	0.322***	(0.0152)	0.322***	(0.0152)
Lagged cumulative GPA	−3.572***	(0.1315)	−3.569***	(0.1315)	−3.569***	(0.1315)
Lagged cumulative GPA^2	0.502***	(0.0261)	0.502***	(0.0261)	0.502***	(0.0261)
Major						
Engineering	8.887	(6.1565)	2.571***	(0.8857)	1.450***	(0.4665)
Sciences	2.890	(1.8740)	0.984***	(0.2762)	0.646***	(0.1521)
Business	3.220	(2.4493)	0.730**	(0.3637)	0.288	(0.2002)
Fine Arts	1.587*	(0.8328)	0.733***	(0.1288)	0.580***	(0.0770)
Expected wages						
Major (Change rel. $t = 1$)	−1.464***	(0.3962)	−0.302**	(0.1192)	−0.147**	(0.0675)
Major ($t = 1$)	−2.750	(1.9382)	−0.764***	(0.2793)	−0.413***	(0.1480)
LL		-1,241,727.68		-1,241,680.30		-1,241,674.97
Observations		276,999		276,999		276,999

Table reports point estimates, with standard errors in parentheses. Model also includes indicators for application cohort, imputed SAT subscores and imputed high school rank. Sample includes only individuals with nonzero semester GPAs and lagged cumulative GPAs, and students that are defined as “able to drop out,” which includes: students that do not graduate at the end of the period, and students with transcripts that are not censored by the end of data collection in 2004.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

into that major. Furthermore, the relative magnitude of these switching costs are also consistent with my priors about the relative costs of switching into a major by a students' prior major. For example, students that wish to switch into business from more rigorous majors like engineering and the sciences incur a lower switching cost, on average, than students that switch into business from a liberal arts or fine arts program.[56]

These switching estimates also suggest that costs vary significantly depending on the students' characteristics and the characteristics of the major they wish to switch into. For example, students that choose to switch back into a major that they have previously held also experience lower costs, which is consistent with students either more easily meeting transfer requirements or by having higher levels of major-specific human capital. Lastly, there is suggestive evidence that switching costs are higher for students from low SES households, which is consistent with the potentially large opportunity costs of switching majors that may be even more burdensome for low SES students.

Switching costs can also influence students' choices through their expected semester grades. I capture this channel in the model by allowing the value of a major to depend on a student's predicted performance, and for that predicted performance to depend, in turn, upon their major switching choices. However, these results depend on whether students' major choices respond to their expected performance. The parameter estimates for a quadratic in predicted semester GPA in Table 2.4 support this, and suggest that students are more likely to choose a major if they predict a higher level of performance in that major, which is consistent across each alternative major. Similarly, students are more likely to drop out in a given period if they predict lower GPAs across all majors.

[56] These relative switching cost results also hold broadly for switching costs that vary by semester in Table 2.5, as is the case for students switching into business in their fifth semester.

This indirect effect of switching costs through grades also depends on whether students' semester GPAs vary by their switching behavior, which I show evidence of in my first stage regression of semester GPAs. I estimate each of these semester GPA regressions separately by major and pool across semesters. The general results from this regression are consistent with theory across all majors, for example: higher high school percentile rank, SAT scores, and lagged semester GPA are all associated with higher semester GPAs.[57] I implement switching costs as a pair of indicators for whether students have switched into a new major in the current semester and whether they switched majors in any prior semesters, and an interaction of the maximum of both indicators with quadratics in students' lagged semester and cumulative GPAs. The parameter estimates for these indicators suggest that switching into a new major is associated with higher semester GPAs in all majors but the sciences. Estimates for the lagged GPA interactions also suggest students that switch majors have a weaker relationship between their lagged cumulative GPA and their current semester GPA. This makes sense, given that students' lagged semester GPA will be more highly correlated with their current GPA than their cumulative GPA if they switch majors.

The third and final channel through which switching costs can impact students' choices is delayed graduation. Since students' only obtain the utility from labor market earnings after graduating, any factors that significantly advance or delay graduation for a major, such as a late major switch that requires a student to take a large number of extra courses, will change the relative appeal of that major. I estimate the probability of graduating at the end of every semester separately by major, but pool semesters within each major. I also constrain the probability of graduating to be equal to zero for all semesters before a students' junior year, and

[57] While I find large negative estimates for the linear parameter on lagged cumulative GPA, the positive estimate on the quadratic term still allows for higher semester GPAs to be associated with higher cumulative GPAs in the previous semester, provided students' cumulative GPAs are higher than 2.00 on a four point scale.

equal to one for students that remain in the sample up to their thirteenth semester.

I find general results for the dropout regression that are expected: students are more likely to graduate at the end of a given period if they are at the end of an academic year, have completed more semesters and credit hours, and have higher semester and cumulative GPAs. This last result is consistent with a UT Austin requirement that all students must have a cumulative GPA of 2.00 or higher in order to be eligible to graduate. Switching costs enter this regression primarily though an interaction between an indicator for ever switching majors and the percentage of a students' cumulative credits that were taken in a prior major. This factor, which is highly correlated with when students switched into their final major, suggests that switching into a major late in one's academic career is associated with a lower probability of graduating at the end of any given semester.

2.6.3 Do students deviate from rational expectations?

This baseline specification with rational expectations over future labor market outcomes fits broad patterns of students' major choice and switching behavior well, but does not fit how these patterns vary across cohorts as well. This baseline model performs well in matching students choices when I pool cohorts. Table 2.7 shows the actual rate of students switching out of a given major that vary both by a students' prior major and the semester when they switch, and compares it to the predicted switching rates generated by the model on the same sample. These predictions closely match the actual switching rates, even for prior major and semester pairs where switching is relatively rare, and deviate by less than 0.01 percentage points in each cell. However, the model doesn't fit these rates as well when I look at them separately by application cohort. For example, while 9.85% of engineering students from the fall 1998 application cohort switch out of engineering at the end of their third semester, the model predicts that 11.55% of these students will switch. Likewise, the model under-predicts the same rate for students in the fall 2000 application cohort by 1.59 percentage points

While these results all suggest that labor market shocks play a significant role in students' decision to switch majors, I have so far ignored the effect of beliefs that deviate from the "optimal" values I use in the benchmark model. So I next explore whether other sets of labor market beliefs besides my baseline rational expectations model do a better job of explaining students' switching behavior across cohorts as they observe shocks to the labor market. I do this by relaxing three assumptions from my rational expectations specification: how persistent they believe shocks to log wages and unemployment will be, whether they have a myopic view of the labor market, and whether they have preferences that are more risk averse over wages. For these results, I start by running a compound grid search over the persistence of the shock to wages and the persistence of the shock to unemployment rates for the log utility model.

Table 2.7: Model fit: Fraction switching out (by prior major)

	Engineering		Sciences		Business		Liberal Arts		Fine Arts	
Semester	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
1	2.82%	2.82%	5.22%	5.22%	1.77%	1.77%	1.80%	1.80%	3.18%	3.18%
2	8.78%	8.78%	9.71%	9.71%	4.82%	4.82%	3.76%	3.76%	7.92%	7.92%
3	12.34%	12.34%	8.24%	8.24%	3.03%	3.03%	2.64%	2.64%	6.53%	6.53%
4	11.86%	11.86%	10.07%	10.07%	3.74%	3.74%	3.18%	3.18%	9.37%	9.37%
5	6.28%	6.28%	5.04%	5.04%	1.70%	1.70%	1.10%	1.10%	4.51%	4.51%
6	6.45%	6.45%	5.05%	5.05%	1.78%	1.78%	1.18%	1.18%	4.13%	4.13%
7	3.55%	3.55%	3.04%	3.04%	1.45%	1.45%	1.18%	1.18%	3.01%	3.01%
8	3.72%	3.72%	3.43%	3.43%	2.27%	2.27%	1.49%	1.49%	1.99%	1.99%

Table reports fraction of students in major k in semester t that switch to another major in $t+1$.

I first report results on the role that the perceived persistence of labor market shocks plays in students' behavior. In this implementation of the grid search, each AR(1) process varies between the "optimal" persistence used in the baseline model, and "perfect" persistence: where students believe that labor market shocks follow a random walk. I then evaluate the fit of each model by calculating and comparing the AIC criterion for each, which I report in Table 2.8.[58] The panel corresponding to a 30 year forecast of the labor market contains the AIC for my baseline rational expectations model. And while the AIC values for models with perfect persistence in the log wage shocks do not fit as well as the baseline model, these results suggest both models with higher persistence in the labor market shock better fit the data. In particular, when I compare the best fit model to the rational expectations model, its AIC value suggests that the baseline model is 0.001 times as probable to as the best fit model to minimize information loss.[59] These results also suggest that the model where students assume shocks to unemployment are perfectly persistent also better fit the data than the baseline model.[60]

This result is also consistent across other alternative specifications for the expected utility of the labor market. For example, either the model with high or perfect persistence of the unemployment shock has a fit that dominates that of the rational expectations persistence parameters for calibrations where students take expectations over either the first fifteen years of their careers, or just the first year. However, I am unable to distinguish between whether these alternative models fit the data better than the best fit model with a 30 year labor market forecast. There-

[58] Since these alternative models simply change the form of the expected value of utility from wages, I neither add nor subtract parameters to the model between these calibrations. Hence, I am unable to perform likelihood ratio or similar tests of relative model fit.

[59] I calculate this value by constructing $\exp\left(\frac{(AIC_{pref} - AIC_{base})}{2}\right)$

[60] However, I am unable to distinguish between which of these two models best fits the data, as the same calculation implies that the perfectly persistent model is 0.247 times as likely to minimize information loss.

Table 2.8: Model AIC by labor market specification

	Log utility					
	30 yr. forecast		15 yr.		1 yr.	
	ρ_w^*	$\bar{\rho}_w$	ρ_w^*	$\bar{\rho}_w$	ρ_w^*	$\bar{\rho}_w$
ρ_p^*	190,542.67	190,616.37	190,539.01	190,612.83	190,620.32	190,674.88
$\frac{1}{2}\left(\rho_p^* + \bar{\rho}_p\right)$	190,527.73	190,588.57	190,523.66	190,584.77	190,561.04	190,622.34
$\bar{\rho}_p$	190,530.53	190,571.27	190,525.54	190,567.56	190,522.63	190,580.51
	CRRA utility ($r_a = 1.8$)					
	ρ_w^*	$\bar{\rho}_w$	ρ_w^*	$\bar{\rho}_w$	ρ_w^*	$\bar{\rho}_w$
ρ_p^*	190,616.59	190,628.05	190,616.89	190,626.86	190,835.64	190,863.38
$\frac{1}{2}\left(\rho_p^* + \bar{\rho}_p\right)$	190,610.92	190,624.34	190,610.64	190,624.34	190,713.74	190,741.77
$\bar{\rho}_p$	190,647.13	190,658.47	190,639.74	190,650.34	190,626.07	190,651.27

Table reports the AIC calculated for a model with 271 parameters with different calibrations of students' expected utility from wages. ρ_w^* and ρ_p^* denote the optimal AR(1) parameters for the process governing the evolution of shocks to log wages and the unemployment rate, which I use in the rational expectations model. $\bar{\rho}_w$ and $\bar{\rho}_p$ denote the value of ρ that corresponds to students believing that labor market shocks are perfectly persistent, and hence follow a random walk.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

fore, my results offer suggestive evidence that is consistent with Berger (1988), that students do take a long-term rather than myopic view of the labor market. Finally my results from the same exercise, run using CRRA utility over wages with a high level of risk aversion, suggest that risk aversion does not play a significant role in explaining students' major switching responses to labor market shocks. This implies that students may respond to policies that seek to improve their outcomes by improving their labor market forecasts. Otherwise, if students' behavior was best explained by a rational expectations model with high risk aversion, their choices would be due to preferences, rather than poor forecasts.

2.6.4 Simulations and welfare implications

One of the benefits of estimating this structural model is that I am able to explore the welfare implications of students' labor market beliefs by running several counterfactual simulations. In particular, I can compare students' major choices in the best fit model with their response to alternative labor market forecasts, which include the "optimal" labor market forecast from the rational expectations model and forecasts from counterfactual labor market shocks.

In the first simulation exercise I use the parameter estimates from the best fit model to predict students' choices if they used the optimal AR(1) persistence parameters to form their labor market forecasts. I then plot the cumulative fraction of engineering students that switch out of their major by the end of their fourth semester in Figure 2.10 and by the end of their sixth semester in Figure 2.11. These figures suggest that giving students more accurate labor market forecasts would change their behavior during both the economic boom of the late 1990s and the 2001-02 recession: 0.6 percentage points more engineering students would switch out from the fall 1998 entry cohort during the expansion by the end of their fourth semester, and 1.9 percentage points fewer students from the fall 2000 cohort would switch out during the peak of the recession. These results are similar for the fraction switching out by the end of their sixth semester, where I predict that 2.2 percentage points fewer students from the 2000 cohort would switch out of engineering.

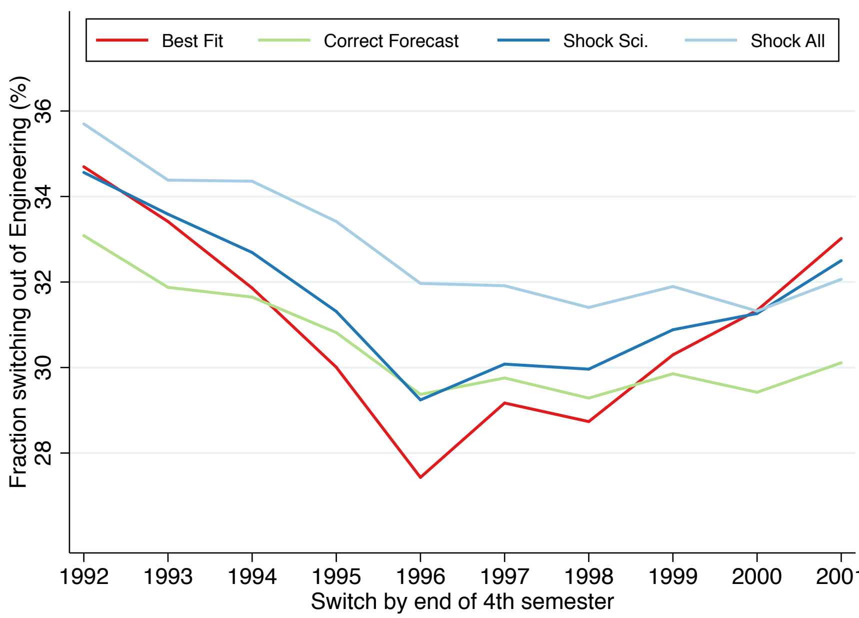

FIGURE 2.10: Fraction switching out of Engineering by $t = 4$ (by LM type)

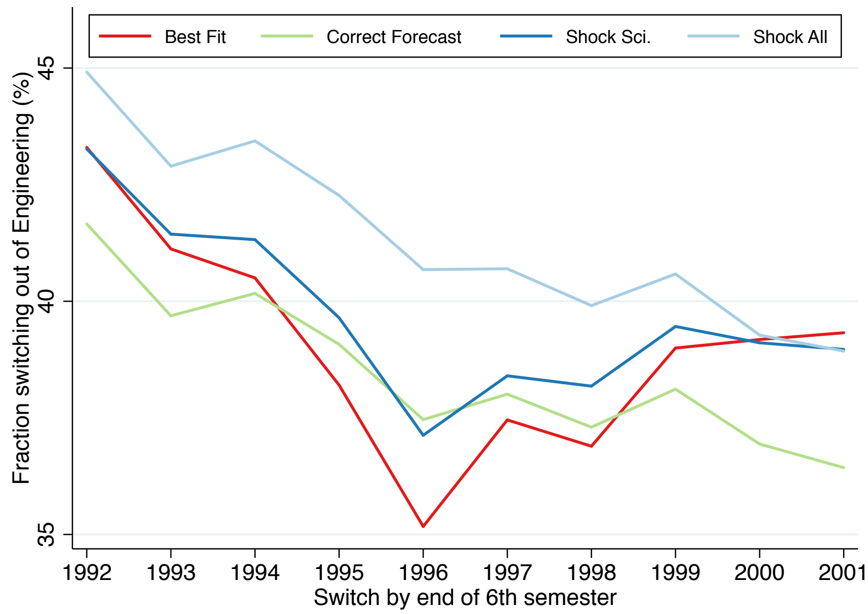

FIGURE 2.11: Fraction switching out of Engineering by $t = 6$ (by LM type)

The properties of the labor market shock generated by the 1990s expansion and 2001-02 recession also influenced students' choices, beyond just its magnitude. Figure 2.10 and 2.11 also show predicted cumulative switching rates for counterfactual labor market shocks, and suggest that students respond to major specific labor market shocks, instead of just aggregate shocks. For these simulations I use students' beliefs from the best fit model and generate two alternative labor markets shocks to the unemployment rate: one where I flip the shocks to engineering and the sciences, and another where I restrict all shocks to match that of engineering students. These results suggest that engineering students would be more likely to switch out of engineering during the expansion if the sciences experienced a similarly low unemployment rate, which similarly holds if all other majors experienced the same unemployment rate. However, these alternative shocks do not significantly change the predicted cumulative switching rates of students during the peak of the recession. This may suggest, as in Zafar (2011), that students are more responsive to positive labor market shocks. Hence the observed difference in switching rates out of engineering between the 1998 and 2000 cohorts was driven more by the 1998 cohort being overly optimistic about the labor market prospects of an engineering degree.

Lastly, these results also suggest a number of potential welfare and policy implications for how students and colleges should respond to recessions, particularly for the accuracy of students' labor market forecasts. I also calculate the expected earnings for students from the fall 2000 cohort when they choose whether to switch majors at the end of their third semester under their beliefs from the best fit model with and without the rational labor market forecast. I find that, for these 1,130 students, giving them the rational expectations labor market forecast is associated with major choices that increase their total expected present value of earnings by $1.8m, which is driven by a greater percentage of students choosing to remain in engineering.[61] These results are also relevant beyond just students prospective earnings, and could play a role in policies designed to increase persistence in and graduation rates from engineering programs. Colleges or government agencies could also give students far more accurate forecasts of how they expect labor market shocks to evolve, which could further improve the *ex ante* optimality of their forecasts. However, my results suggest that even improving the quality of students' forecasts with the simple AR(1) processes used in this model would have significant effects on students' choices.

2.7 Conclusion

Students' expectations about their future labor market outcomes are an important determinant of their major choice. However, studies of college major choice often assume that students have rational expectations about these outcomes, and do not

[61] I generate these results by calculating students' expected earnings from working in a particular major with the rational expectations forecast of the labor market. I then discount these with $\delta = \frac{1-0.95}{0.95} = 5.26\%$ and multiply them by the corresponding predicted probability of students choosing that major.

take into account how students' beliefs could deviate from rational expectations assumptions. I explore the role that deviations from rational expectations play in explaining students major switching behavior in response to labor market shocks, and find evidence that students' beliefs do deviate from rational expectations that influence their major switching behavior.

Understanding why college students switch majors is important for all of the same reasons that have been documented in studies of static major choice, because it is simply another major choice. However, models that allow students to switch majors also have to account for additional features that influence students' incentives to switch majors, such as how their beliefs about the value of possible majors evolve over time, and how the costs of switching majors impede their ability to switch in response to incentives. The literature has already established that students' evolving beliefs about their academic ability play a significant role in their decision to switch majors. Furthermore, being able to switch majors in response to these shocks has been shown to improve students' welfare, since they are better able to match their abilities to the requirements of their major.

The literature and patterns in the data have also found suggestive evidence that shocks to labor market returns are also a determinant of students' switching behavior; however, unlike their response to ability shocks, students' behavior is not consistent with welfare-maximizing rational expectations. I show that the 2001-02 recession, and its disproportionate impact on engineering graduates, prompted engineering students at the University of Texas at Austin to switch out of engineering majors at a rate 40.8% higher than in typical years. Furthermore, this behavior presents a puzzle: students switched out of these majors with the highest earnings and historically favorable labor market conditions, and did so in response to a shock that rational expectations forecasts of the labor market indicated would largely dissipate by the time they graduated and entered the labor market.

I explain this behavior by introducing a new sequential model of major choice that can capture the effect that students' beliefs about labor market shocks have on their major choices and decision to switch majors. I start by specifying a basic model that captures the relevant determinants of major choice that have been established in the literature, and modify it to capture the effect of labor market shocks by allowing students to choose majors every semester and for graduation to follow a stochastic process. I then add new components that capture the labor-specific data in students' information sets, a model of their labor market beliefs that allows me to control for several channels that would drive their response to the 2001-02 recession, and complex switching costs that mediate their propensity to switch.

Unfortunately, all of these new features make estimating the model infeasible, without significantly decreasing the complexity of the future value term and matching it with a dataset that can identify all of the additional parameters. I overcome these problems by first specifying a conditional choice probability estimator that significantly reduces the complexity of the future value term to a simple function of the one-period ahead CCP of dropping out. Finally, I pair this simplified estimator with data from UT Austin, which is able to identify the additional parameters of the model, including the complex major switching costs.

The estimates from this model show that students' beliefs about labor market shocks and their impact on future labor market outcomes are able to explain their response to the 2001-02 recession. In particular, students respond to major-specific labor market shocks, and their choices are consistent with a set of beliefs that assume shocks to the unemployment rate will be highly persistent over their labor market careers. Finally, major switching costs are a significant obstacle to their response to these shocks, and vary significantly by the major that students switch from and into, the semester they are in when they switch, and a range of other factors.

Finally, these results also have significant implications for students' welfare, which

I demonstrate with counterfactual simulations. Students' beliefs about the labor market prompt them to switch out engineering during the recession at a rate 7.86% higher than they would with rational expectations. This not only has important implications for policies designed to increase the number of STEM graduates, but would lead to a decrease in the present discounted value of expected future earnings of $1.8m for students switching out of engineering.

3

The Career Prospects of Overeducated Americans

As American students accumulate college loan debt (1.08 trillion dollars as of December 31, 2013, Source: NY Fed), there is a growing concern that expensive skills acquired in college may be underutilized in low-paying jobs. Existing studies estimate that around a third of American workers are "overeducated" - i.e. have more schooling than is necessary for their job.[1] These would include, for example, a college graduate working as a cashier in a store. Estimated wage returns to this surplus schooling average 4.3%, or about half of the returns to required schooling. Thus, schooling mismatch appears to be an important source of the *ex post* heterogeneity in returns to schooling documented in the literature (see, e.g., Carneiro, Hansen, and Heckman, 2003). At the aggregate level, overeducation could reflect skill mismatch and an inefficient allocation of workers to jobs.

What cross-sectional data misses is the possibility that overeducated workers may only be temporarily underemployed before switching to a job that requires their level of schooling. Further, low unobserved ability, compensating non-pecuniary job

[1] Leuven and Oosterbeek (2011) obtain these estimates of the incidence of overeducation and returns to overschooling by averaging estimates from 151 studies.

characteristics and career mobility considerations could rationalize apparent overeducation without the implication of a suboptimal schooling choice.[2] In order to understand how much of a problem overeducation really is, it is crucial to go beyond the cross-sectional stylized facts and investigate longitudinal patterns.

This paper provides the first analysis of the career dynamics of overeducated U.S. workers. Specifically, we use the National Longitudinal Survey of Youth 1979 (NLSY79), combined with the pooled 1989-1991 waves of the Current Population Survey (CPS), to examine first the determinants of overeducation, second how overeducation incidence changes along the career, and third how wages differ across overeducated, matched or undereducated employment, and as a function of previous overeducated employment spells. Most of our analysis will focus on the causes and consequences of overeducated employment among two and four-year college graduates, who make up the bulk of overeducated workers.[3]

While the literature has paid comparatively little attention to the longitudinal dimension, analyzing transitions into and out of overeducated employment, together with their effects on wages, is key to disentangling the role played by labor market frictions versus other theories of overeducation.[4] For example, if overeducation was due only to search frictions, one would expect this type of mismatch to be transitory and concentrated early in the career. Conversely, selection on ability, compensating wage differentials or career mobility motives would generate persistence in the

[2] The career mobility factor was initially investigated by Sicherman and Galor (1990). The general idea is that high-skilled workers may face higher promotion probabilities in low-skilled jobs. It follows that forward-looking individuals may choose to become overeducated.

[3] Throughout the paper, we use "matched" as a shorthand for being neither overeducated nor undereducated for one's job. While our analysis primarily focuses on overeducation, it is worth noting that undereducation is also relatively common, in particular among individuals who have completed 14 years of education, and, as such, has the potential to account for some of the wage dispersion within this schooling category.

[4] In practice we restrict our analysis to individuals who have completed their highest level of education over the sample period (1982-1994), so that all changes in overeducation status result from a change in employment rather than from a change in schooling attainment.

overeducation patterns.[5] The individual persistence and duration dependence of overeducated employment, together with the wage penalties associated with it, are also important for the design of unemployment insurance and training programs. For instance, encouraging early exit from unemployment may push more workers into overeducated work with potentially negative long-term effects on earnings.

The question of overeducation was first brought to the attention of economists and policy-makers by Freeman (1976b), who argued that excess supply of college graduates was causing the decline in the college wage premium observed in the U.S. during the 1970's. While the cross-sectional properties of overeducation are well-studied (see, e.g., Verdugo and Verdugo, 1989; Alba-Ramirez, 1993; Kiker, Santos, and de Oliveira, 1997; Hartog, 2000), still little is known about the evolution of overeducation over the life cycle, although, as argued above, dynamics are of clear interest in this context. U.S. evidence is particularly scarce.[6] A notable exception is Rubb (2003) who provides evidence from the CPS that overeducation displays a substantial degree of persistence, with around 30% of the individuals overeducated in year t switching to a job which matches their level of education in year $t+1$. While duration dependence and dynamic selection effects imply that these transition rates are likely to decrease over the length of the spell, the CPS panels are too short to address this question adequately.

It is worth pointing out that while we borrow the wording "required level of education" from the existing literature, defining and measuring that concept is not

[5] See Gottschalk and Hansen (2003) who show that, in a model with two sectors, two skills (college and non-college) and heterogeneous preferences for each sector, some college workers may choose to work in the non-college sector. Uncertainty in returns to schooling is another channel which could generate persistence in overeducation, since some individuals may find it *ex post* optimal to be overeducated (see Lee, Lee, and Shin, 2014).

[6] Several studies have used German, British, Canadian or Australian data to estimate panel wage regressions (Bauer, 2002; Frenette, 2004), dynamic random effect models of overeducation exit (Mavromaras et al., 2013) or simply document overeducation status transitions (Dolton and Vignoles, 2000).

a trivial task. In this paper we use a statistical measure, in a similar spirit as, e.g., Verdugo and Verdugo (1989) and Kiker, Santos, and de Oliveira (1997). Namely, we compute the mode of the distribution of schooling in the 1989-1991 CPS for each occupation, which we define using the 1980 3-digit Census Occupation classification. We also restrict the CPS sample used to compute the mode to individuals in the same birth cohorts as the NLSY79 respondents. The required levels of education are then defined as those within 15 percentage points of the schooling mode.[7] The typical overeducated worker in our sample has two or four years of college education, but is working as a secretary or a cashier, say, among a majority of high school graduates. Relative to alternative approaches in the literature, and in particular those measuring the required levels of education with the General Educational Development (GED) scale provided by the Dictionary of Occupational Titles (see, e.g., Hartog, 1980, and Rumberger, 1987), this approach is arguably more transparent and has the benefit of directly generating requirements in terms of years of education. As such, while our required schooling variable should be a pretty good indicator of the amount of schooling needed to perform well in an occupation, it is most cleanly interpreted as the schooling attainment of a typical worker in one's occupation.

We document longitudinal patterns of overeducation for the NLSY79 cohort up to 12 years after labor market entry. Overeducation incidence within the cohort decreases as workers progress through their careers, but remains sizeable 12 years after the first job. This suggests that, while frictions are likely to play a role, we need to appeal to other economic mechanisms to explain this long-term persistence. Overeducation is also a fairly persistent phenomenon at the individual level, with around 66% of overeducated workers remaining in overeducated employment after

[7] Using this type of conservative measure of mismatch allows us to mitigate the risk of misclassification, which could arise if two or more occupations with different required levels of education were aggregated up at the 3-digit level. Depending on the aggregation level used for the educational attainment, our measure of overeducation yields incidence levels of between 18% and 25% among all workers, and as much as about 40% among college graduates.

one year. We find that blacks and low cognitive ability workers (as measured by their AFQT scores) are not only more likely to be overeducated, but also less likely to switch into matched jobs. That is, the longitudinal dimension magnifies the cross-sectional black-white and cognitive ability gaps.

The hazard rate out of overeducated work is also strongly decreasing in overeducation duration, and drops by about 60% after five years. We estimate a mixed proportional hazard model (Elbers and Ridder, 1982) of overeducated employment duration to investigate whether this decreasing hazard rate reflects selection on unobservables or true duration dependence. While composition effects based on observable characteristics explain away some of the duration dependence, further controlling for unobserved heterogeneity largely wipes it out. In other words, the duration of the overeducated employment spell does not have a significant impact on the probability to exit overeducation. Instead, we identify large unobservable differences in the hazard rate: while overeducation is found to be very persistent for 30% of the sample, the rest is much more likely to exit quickly, in keeping with a frictional view of overeducation. The latter pattern provides clear evidence that there is generally more to overeducation than selection on unobserved ability and preferences.

We then revisit the classical augmented wage regression used in the overeducation literature.[8] In particular, we account for productivity differences across the two heterogeneity types identified in the duration model. Importantly, we find that, controlling for current overeducation status, past overeducation entails a sizeable wage penalty of between 2.6% and 4.2%, which persists over four years. This provides a likely candidate mechanism behind the negative wage effects of graduating dur-

[8] Augmented wage regressions, pioneered by Duncan and Hoffman (1981) replace the usual years-of-education regressor with three terms: years of required education in the current occupation, years of education in excess of that required level and years of education below that required level. The corresponding coefficients are interpreted as returns to required education, returns to overeducation and returns to undereducation. In a similar spirit, a number of recent papers stress the importance of relaxing the homogeneous and linear returns to schooling assumption in the classical Mincer regression (see, e.g., Belzil and Hansen, 2002, and Heckman, Lochner, and Todd, 2006)

ing a recession recently discussed in the literature (Kahn, 2010, Liu, Salvanes, and Sorensen, 2012, Oreopoulos, von Wachter, and Heisz, 2012, and Altonji, Kahn, and Speer, 2014), since overeducation is likely to be more frequent during recessions, consistent with the cyclical upgrading literature (Bils and McLaughlin, 2001).

The remainder of the paper is organized as follows. Section 3.1 describes the data used in the analysis and the construction of the required schooling measure. Section 3.2 discusses the determinants of overeducation in our sample. Section 3.3 documents the longitudinal patterns in the incidence of overeducated employment along the career. Section 3.4 estimates a mixed proportional hazard model of overeducated employment duration, which allows us to separate true duration dependence from dynamic selection on observed and unobserved worker attributes. Section 3.5 presents results pertaining to the effect of overeducation on wages, and Section 3.6 concludes.

3.1 Data

Our main data source is the NLSY79, which is a nationally representative sample of 12,686 young men and women who were 14-22 years old when they were first surveyed in 1979.[9] We pool the observations for the 6,111 individuals that comprise the core civilian cross section of the NLSY79, from the 1982 to the 1994 rounds, which results in 79,443 person-year observations.[10] Then we cut 5,947 person-year observations

[9] Of the 12,686 individuals interviewed in the initial 1979 wave, these data provide information for respondents on a yearly basis from 1979 to 1994 and biyearly afterwards. The initial wave is comprised of a core civilian cross-section of 6,111 and an oversample of 5,295 black, Hispanic, and economically disadvantaged individuals born between January 1, 1957 and Dec. 31 1964. This is further supplemented by a military sample of 1,280 individuals born in the same period. We only keep the core cross-sectional sample of the NLSY79 in order to maintain a consistent sample between the NLSY79 and the CPS.

[10] We restrict our sample to these years because it is the largest contiguous period where the NLSY79 reports the 1980 Census Occupation codes on an annual basis. We use the 1980 codes because they better reflect the set of occupations available over the period of interest than the 1970 codes.

with a level of education that was unreported or less than 12 years; 22,272 where the individual had not entered the labor market permanently; and 6,228 that were non-interviews. We define the date of (permanent) entry into the labor market as the first survey year where (i) the individual is employed in the civilian labor force, (ii) works more than 26 weeks out of the year, (iii) is not enrolled in school as of May 1st of the survey year, and (iv) has reached their highest level of education over the sample period 1982-1994. After making these cuts, we are left with a total of 44,996 observations corresponding to 4,895 distinct individuals.

The main variables of interest are the highest level of completed education, the occupation (measured using the 1980 3-digit Census code) and the hourly wage at the time of each interview.[11] Besides these, the variables used in our analysis include age, minority status, gender and place of birth, cognitive and non-cognitive skill measures, geographical location and the corresponding local unemployment rate, family characteristics, a measure of hazards associated with the current occupation and employment history (see Table 3.1).

3.1.1 Measuring required schooling

The NLSY79 does not have direct measures of required schooling in the job occupied by the respondent. In this paper, we use a statistical measure for the required level of education, in a similar spirit as, e.g., Verdugo and Verdugo (1989) and Kiker, Santos, and de Oliveira (1997). Namely, for each given occupation in the 1980 3-digit Census Occupation classification, we compute the required level of education from the pooled monthly samples of the 1989-1991 waves of the CPS. Those years were chosen with two considerations in mind. First, they sit in the middle of the

[11] In practice, we use the occupation and hourly wage corresponding to the current or most recent job at the time of the interview. When individuals hold multiple jobs at the same time, we use the occupation and wage corresponding to the job in which the respondent worked the most hours. We adjust for inflation by reporting all wages in constant dollars and then drop the top and bottom 2.5% of the reported person-year wages for every survey wave.

date range that we are analyzing. This minimizes the extent to which technological change might have altered the schooling requirements in some occupations. Second, the average unemployment rate (5.9% between 1989 and 1991, Source: Bureau of Labor Statistics) was low during these years. This reduces the likelihood that a bad labor market would push so many highly educated individuals into low schooling requirement occupations that the modal worker in those occupations would have more schooling than is necessary for their job. In order to obtain required schooling levels that are pertinent to the NLSY79 sample, we restrict the age range within each year of the CPS to that of the NLSY79 cohorts at that time (see Appendix A for additional details on the CPS sample used in the analysis). Required schooling in a given occupation code is then defined as the mode of the distribution of the levels of education among the individuals working in that occupation.[12]

3-digit occupational codes correspond to a high level of disaggregation, and constitute the finest description of occupations available for a long representative panel of U.S. workers. Census occupational codes were defined, among other things, according to the skills involved in performing the job. For example, 3-digit codes distinguish among sales occupations between jobs that involve increasing degrees of knowledge and task complexity: from street vendors, to cashiers, sales workers (subdivided into eight different industry groups), sales representatives (6 industry groups), sales engineers and sales supervisors. However, there still might be unobserved heterogeneity in schooling requirements within some 3-digit occupations. One concern would be that a 3-digit occupation contains several occupations with different schooling requirements. With this in mind, for occupations such that the frequencies of two or more schooling levels are within 15 percentage points of each

[12] Finite sample variability should not be a major concern here given the large number of observations (on average above 1,000) which are used to estimate the mode of each occupation. See, e.g., Dutta and Goswami (2010), who show that, for a Bernoulli distribution with sample size larger than 100, the mode of the empirical distribution matches the population mode with a probability close to 0.9.

other, we choose to use a more conservative definition of over (and under)-education. Specifically, workers whose schooling attainments fall within the range defined by these schooling levels are classified as matched, while those with a higher (lower) level of education are defined as overeducated (undereducated). It is important to note that our results are robust to the choice of other cutoffs.[13]

In order to mitigate concerns with classification error on attained schooling, we collapse our years of education variable into four categories: 12-13 years, 14-15, 16-17, and over 18 years of completed education. This classification is natural since each category, simply referred to as 12, 14, 16 and 18 years of schooling in the rest of the paper, corresponds to high school graduates, two-year and four-year college graduates, and graduate school.[14]

A clear advantage of this method is that it generates requirements directly in terms of years of education. This is in contrast to a common approach in the literature which maps occupations into skills first, and then skills into years of education (using, for example, the GED scale). In practice the latter approach is problematic, in particular since there is no clear consensus on the mapping between the skill content of occupations, as measured by the GED scale, and years of schooling (Leuven and Oosterbeek, 2011). On the other hand, one limitation of our measure of required schooling is that it is based on the distribution of schooling attainment among workers employed in a given occupation, which is an equilibrium outcome of labor supply and demand decisions. Ultimately, one can think about the workers we identify as overeducated as those who have more schooling than their peers, in the sense of having more schooling than the modal worker in their current occupation.

[13] Specifically, we considered four alternative definitions of over (and under)-education, using (i) the mode only (i.e. no cutoff), (ii) a 5% cutoff, (iii) a 10% cutoff and (iv) a 20% cutoff. Our results were overall very similar to those obtained using our baseline definition.

[14] It is also possible that errors in the occupation codes recorded in each interview of the NLSY79 could generate artificial transitions between overeducation statuses. In that case, our estimates of overeducation persistence could still be interpreted as lower bounds.

3.1.2 Summary statistics

Table 3.1 presents summary statistics for the NLSY79 variables used in our analysis. Of the 44,996 observations in our final sample, 84.7% are employed, 10.6% are out of the labor force, and 4.8% are unemployed. 66.1% are high school graduates, 12.9% have two years and 14.8% have four years of college education, and 6.3% have some graduate school experience.

Figure 3.1 shows the distribution of years of overeducation, generated by subtracting an individual's observed highest level of completed education, from the level of education required by their occupation. More than half of all observations have a perfect match of observed and required education, while progressively smaller fractions exhibit one or more years of education level mismatch. Comparing the two panels shows that collapsing schooling attainment into four categories preserves the shape of the distribution. It also mitigates the concern that small errors in the measure of years of schooling attainment will generate overeducation status misclassification.

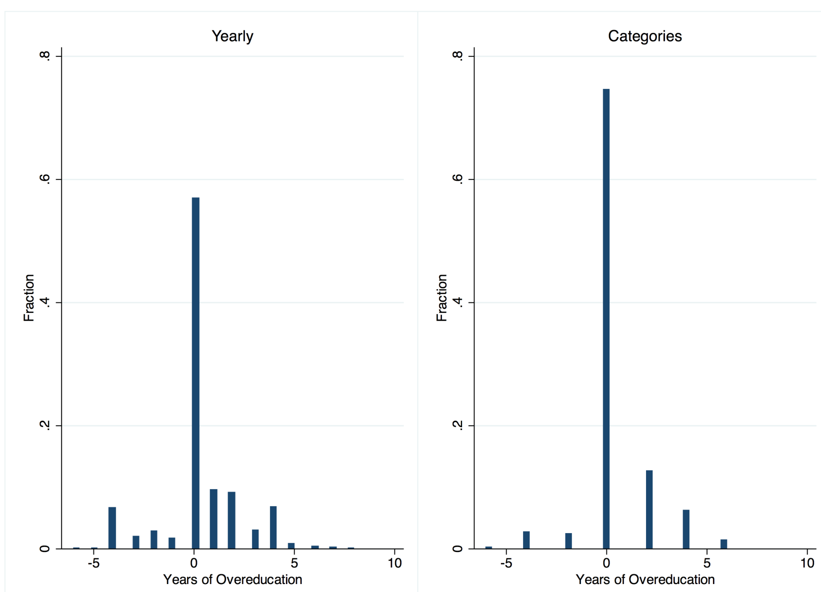

FIGURE 3.1: Distribution of overeducation

Table 3.1: Summary statistics for pooled cross section (1982-1994)

Variable	Mean	(Std. Dev.)	N[1]
Required education (yrs.)	12.871	(1.751)	34,535
Overeducation (yrs.)	0.365	(1.635)	37,087
Age	27.503	(3.904)	44,996
Black	0.112	(0.315)	44,996
Hispanic	0.058	(0.233)	44,996
Female	0.514	(0.5)	44,996
Born in The United States	0.963	(0.189)	44,996
AFQT	0.503	(0.274)	43,052
Rotter Scale in 1979	8.44	(2.371)	44,669
Sociability in 1985	2.894	(0.669)	43,897
12 Years of education	0.661	(0.474)	44,996
14 Years of education	0.129	(0.335)	44,996
16 Years of education	0.148	(0.355)	44,996
18 Years of education	0.063	(0.242)	44,996
GED (No HS diploma)	0.09	(0.286)	44,996
Northeast	0.194	(0.395)	44,748
South	0.331	(0.471)	44,748
West	0.171	(0.377)	44,748
Urban	0.785	(0.411)	43,828
Unemployment rate (in %)	6.814	(1.374)	44,748
HH in SMSA	0.767	(0.423)	42,708
Mother's edu. (yrs., 1979)	11.965	(3.387)	41,395
Father's edu. (yrs., 1979)	11.714	(2.557)	43,102
Lived with both parents	0.673	(0.469)	44,996
Employed	0.847	(0.36)	44,996
Out of labor force	0.106	(0.307)	44,996
Unemployed	0.048	(0.213)	44,996
Tenure (1k wks.)	0.17	(0.173)	41,782
Number of jobs	1.407	(0.867)	44,996
Work experience (1k hrs.)	17.222	(8.800)	29,743
Weeks unemployed	2.291	(7.991)	44,275
Hourly wage	12.717	(6.158)	38,717
Occupational hazards	−0.138	(0.952)	37,889

Unemployment rate (by region) corresponds to the annual unemployment rate for each Census region (BLS Local Area Unemployment Statistics data). The measure of occupational hazards is drawn from a sample of the 1970 Census which included occupational characteristics from the Dictionary of Occupational Titles (DOT). An indicator for a DOT occupation having hazards was averaged within each 1980 Census occupation code; we then convert this percentage measure to a Z-score. Tenure is defined as the total number of weeks an individual has worked in a given occupation across all employers. The total number of jobs is measured since the last interview. Wages are measured in constant dollars.

[1] Person-year observations

Table 3.2: Overeducation status proportions by education level

	Yearly	Categories				
		All	12	14	16	18
Undereducated	15.9%	14.1%	4.7%	16.3%	2%	
Matched	59.5%	68%	95.3%	18.1%	60.5%	28.7%
Overeducated	24.6%	17.9%		65.7%	37.4%	71.3%

Sample excludes individuals that have completed less than 12 years of education.

In Table 3.2, we further break down overeducation status by our categorical measure of attained schooling. It is apparent that overeducation is mechanically absent from the lowest schooling level (12 years of education) while undereducation is absent from the highest schooling level (18 years of education). Consistent with the existence of a relatively small number of jobs requiring 14 or 18 years of education, a very large fraction of the observations corresponding to those schooling levels exhibit education level mismatch. Although a larger share of jobs require 16 years of education, it is interesting to note that 37.4% of college graduates are overeducated, typically working in a job requiring 12 years of education. Table 3.3 lists the ten occupations accounting for the most number of observations in overeducated employment. Secretaries and Sales workers account for the largest numbers of overeducated workers. For most of the occupations in the table the modal worker has 12 years of schooling, and the typical overeducated worker has two or four years of college education. One notable exception is teachers, which typically have 16 years of schooling in the CPS.

Table 3.3: Ten most frequent occupations among overeducated workers

1980 Census		Required Edu.		Observations	
Code	Occupation	Min.	Max.	OE	Total
313	Secretaries	12	12	420	1,525
274	Sales workers, other commodities	12	12	394	960
243	Supervisors and proprietors, sales occupations	12	12	323	772
337	Bookkeepers, accounting, and auditing clerks	12	12	243	649
19	Managers and administrators, n.e.c	12	16	223	2,290
156	Teachers, elementary school	16	16	197	318
389	Administrative support occupations, n.e.c.	12	12	163	415
276	Cashiers	12	12	121	675
433	Supervisors, food preparation and service occupations	12	12	117	254
376	Investigators and adjusters, except insurance	12	12	113	243

The columns for required education correspond to the lowest (Min.) and highest (Max.) levels of required education that we calculate from the CPS. Note that these values will differ if the mode and second largest observed levels of education are separated by fewer than 15 percentage points, as in occupation 19, Managers and administrators.

3.2 The determinants of overeducation

Before moving on to the analysis of the career dynamics of overeducated workers, we start by documenting the cross-sectional determinants of overeducation in our sample. We report in Table 3.4 the estimation results from a Probit model, which allows the probability of overeducation to depend on a set of socio-demographic characteristics, ability measures, family characteristics, a measure of hazards associated with the current occupation and employment history.[15] We stratify the regression by schooling level, consistent with our focus throughout the paper on overeducation as a labor market, rather than educational, phenomenon.

AFQT scores exhibit a negative and significant relationship with overeducation at all schooling attainments. This negative relationship between cognitive ability and the likelihood of overeducation, which is in line with prior findings in the literature (see, e.g., Allen and van der Velden, 2001 and Chevalier and Lindley, 2009), is consistent with ability and schooling attainment being labor market substitutes.

Also in accordance with existing studies, females are about 5 to 13 percentage points more likely to be overeducated than males. Women may place more value on non-pecuniary characteristics associated with low-requirement jobs, such as flexibility in hours worked or their proximity to home, making it easier to combine work and home production activities. Alternatively, this could reflect discrimination on the part of employers. At any rate, given the existence of a substantial and persistent wage penalty of being overeducated (discussed in Section 3.5), this result implies that overeducation is an important aspect of the gender wage gap, which is absent from most of the literature on this question.[16]

[15] In practice we use a partial maximum likelihood estimator, clustering standard errors at the individual level. The resulting inference is robust to serial correlation in the unobserved determinants of overeducation.

[16] One notable exception is the early analysis by Frank (1978).

Table 3.4: Probit model of overeducation status

	14 Years		16 Years		18 Years	
Black	0.159***	(0.0555)	0.035	(0.0673)	0.107	(0.1045)
Hispanic	0.122*	(0.0628)	−0.081	(0.0697)	−0.023	(0.0866)
Female	0.055	(0.0392)	0.125***	(0.0290)	0.105***	(0.0374)
Born in The United States	0.206**	(0.0838)	−0.016	(0.0624)	0.021	(0.0926)
AFQT†	−0.132***	(0.0511)	−0.148**	(0.0622)	−0.187*	(0.0995)
Rotter Scale in 1979	0.010	(0.0076)	0.006	(0.0065)	−0.003	(0.0077)
Sociability in 1985	−0.036	(0.0256)	−0.028	(0.0218)	0.018	(0.0285)
GED (No HS diploma)	−0.051	(0.0829)	−0.282	(0.2360)	−0.095	(0.2206)
Northeast	−0.027	(0.0522)	0.002	(0.0355)	−0.005	(0.0529)
South	−0.011	(0.0416)	0.007	(0.0370)	−0.008	(0.0452)
West	−0.003	(0.0520)	0.006	(0.0426)	−0.021	(0.0492)
Unemployment rate	0.016*	(0.0093)	0.003	(0.0078)	−0.019*	(0.0099)
HH in SMSA	−0.077	(0.0541)	−0.006	(0.0428)	0.032	(0.0504)
Tenure (1k wks.)	0.078	(0.0976)	−0.056	(0.1017)	0.396***	(0.1176)
Work experience (1k hrs.)	−0.005	(0.0062)	−0.018***	(0.0056)	0.014*	(0.0073)
Work experience^2 (1k hrs.)	0.003	(0.0022)	0.003	(0.0019)	−0.006**	(0.0028)
Number of jobs	0.031*	(0.0161)	0.053***	(0.0140)	−0.031*	(0.0175)
Total unemp. (wks.)	0.002***	(0.0006)	0.000	(0.0008)	0.002	(0.0014)
Occupational hazards	0.130***	(0.0328)	0.279***	(0.0414)	0.223**	(0.0897)
χ^2		201.92		377.69		160.54
LL		-1,737.63		-2,235.03		-816.98
Observations		3,125		4,064		1,763

Model also includes controls for industry, age, mother and father's education level in 1979, a dummy for living with both parents in 1979, and a dummy for living in an urban area at the time of interview. Unemployment rate (by region) corresponds to the annual unemployment rate for each Census region (BLS Local Area Unemployment Statistics data). Entries correspond to marginal effects (standard errors in parentheses).
† Values converted to Z-scores
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note that there is also a strong positive correlation between minority status and overeducation at the 14 years of schooling level. Blacks and Hispanics are 15.9 percentage points and 12.2 percentage points, respectively, likelier than whites to be overeducated among that group. Among college graduates and above, the relationship becomes insignificant, which is possibly a result of a strong selection into college based on unobservable skills that are negatively correlated with overeducation.

Lastly, the evidence regarding the theory that overeducated workers accept lower wages in exchange for better non-pecuniary job characteristics is mixed. Workers in hazardous jobs are actually more likely to be overeducated at all three levels of schooling. On the other hand, overeducated workers are more likely to hold several jobs (except for those with 18 or more years of schooling), which possibly reflects a higher flexibility for these overeducated jobs. These results complement previous studies that have interpreted a negative correlation between overeducation and job satisfaction as evidence against the compensating wage differential model of overeducation (see, e.g., Hersch, 1991, and Korpi and Tåhlin, 2009).

3.3 The dynamics of overeducation

Figure 3.2 displays the aggregate incidence of overeducated employment over the first 12 years of work for the NLSY79 cohort, for workers with at least some college education. Overall, the incidence of overeducation decreases by about 12 percentage points, from 62.3% to 50.4%, over the first 12 years of respondents' careers.[17] While the decline in aggregate overeducation rates as the career progresses does suggest that overeducation is in part frictional, the most striking feature of this graph is that the incidence of overeducation remains very high more than ten years after labor market entry. Overall, this is a clear indication that overeducation is a persistent phenomenon.

In Figure 3.3, we further disaggregate this graph along several observable characteristics. Blacks do not exhibit the same reduction in overeducation as whites (Panels 1 and 2). Similarly, Panels 3 and 4 show that overeducation among females decreases much less than among males. Lastly, individuals at higher AFQT quartiles see a larger decline in overeducation incidence than those at lower quartiles (Panels 5 through 8). Overall, these results suggest that overeducated black, female and low-AFQT workers are less likely to receive and/or accept offers from matched jobs. These dynamic patterns therefore accentuate their already higher propensity to be overeducated.

Individuals' patterns point to a similar story. Table 3.5 presents the fractions

[17] Using a self-reported measure of overeducation, Dolton and Vignoles (2000) also find that its incidence among a 1980 cohort of U.K. university graduates decreased over time, from 38% to 30% 6 years after graduating.

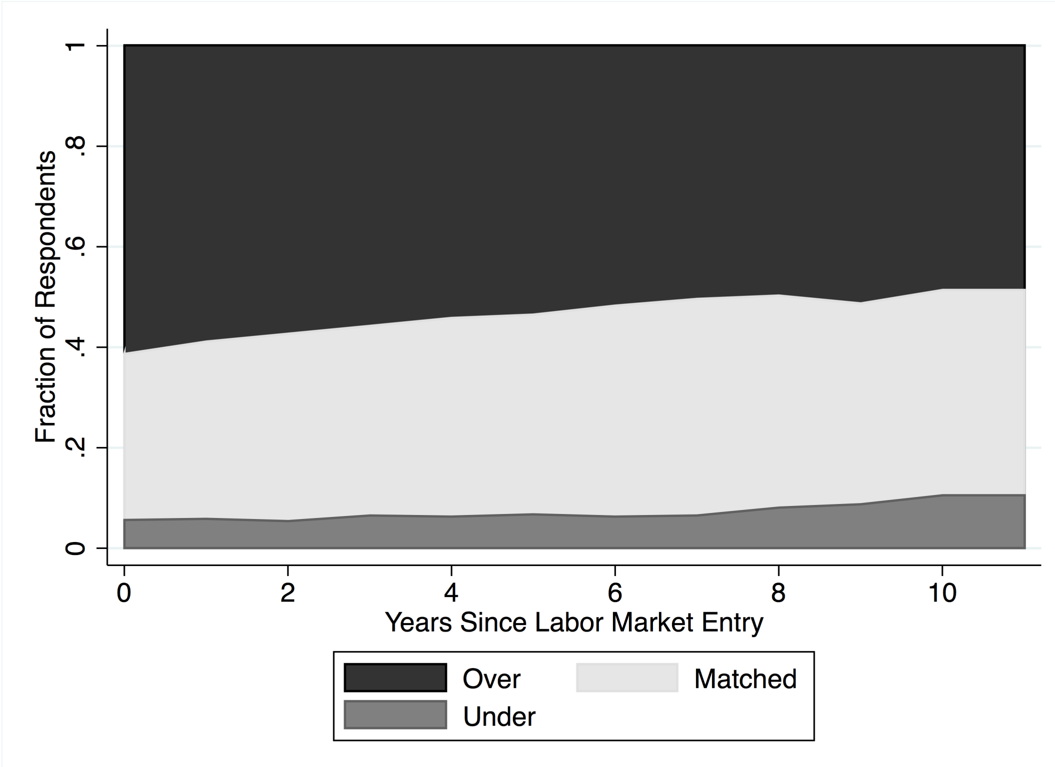

FIGURE 3.2: Composition of overeducation (fraction of respondents)

of individuals who are non-employed (defined as unemployed or out of the labor force), undereducated, overeducated or matched at interview time, conditional on the status reported during the interview one year before, for workers with at least some postsecondary education. Overeducation is also persistent at the individual level, with 65.9% of workers remaining overeducated after one year. This fraction is 20 points higher than the unconditional overeducation rate (46.3%). By comparison, non-employed individuals have a smaller 49.0% chance of remaining in that state.[18]

We break down these transition rates further by gender and race in Tables 3.6 and 3.7. Men are slightly more likely than women to remain in an overeducated job but also more likely to transition into matched jobs. Differences by race are sizeable: first, overeducation is more persistent among blacks relative to whites, second

[18] With a different measure of overeducation and using CPS data, Rubb (2003) obtains a level of persistence for overeducated individuals of 73%.

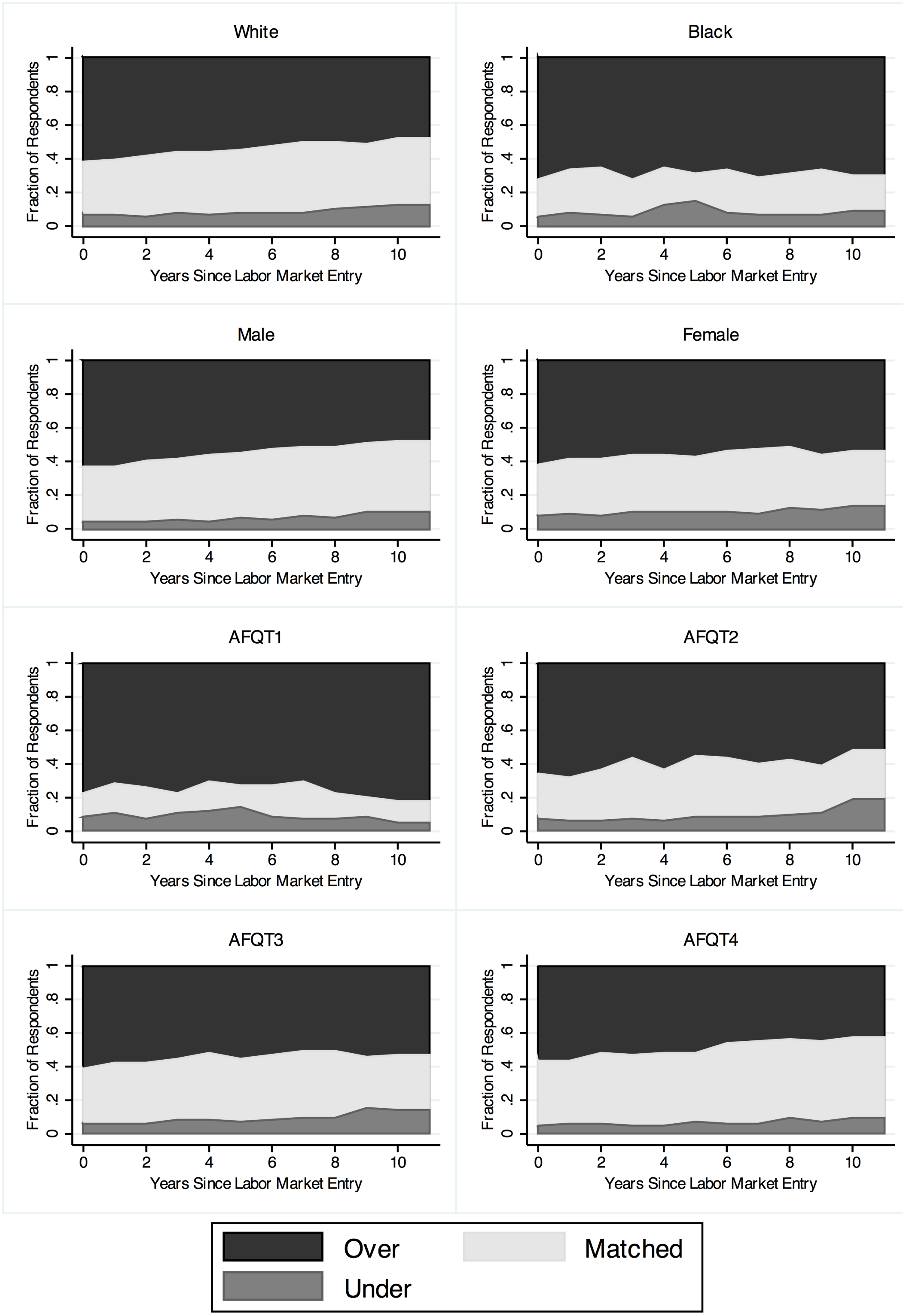

Figure 3.3: Composition of overeducation (14 years of education or more)

Table 3.5: Transition matrix

$t-1$	t OE	UE	Matched	Non-emp.
Overeducated	65.9%	4.2%	25.1%	4.8%
Undereducated	29.0%	44.9%	20.8%	5.4%
Matched	28.5%	3.9%	64.1%	3.6%
Non-employed	30.8%	4.7%	15.5%	49.0%
Total	46.3%	7.5%	37.6%	8.6%

Sample has entered the labor market and has a schooling attainment of 14 years or more.

Table 3.6: Transition matrix by gender

(Male) $t-1$	t OE	UE	Matched	Non-emp.
Overeducated	66.3%	3.7%	26.5%	3.5%
Undereducated	28.9%	49.0%	18.8%	3.4%
Matched	29.5%	3.0%	65.2%	2.2%
Non-employed	49.2%	7.0%	18.8%	25.0%
Total	48.7%	6.7%	40.4%	4.2%
(Female) $t-1$	t OE	UE	Matched	Non-emp.
Overeducated	65.5%	4.7%	23.8%	6.0%
Undereducated	29.0%	42.3%	22.0%	6.6%
Matched	27.3%	4.7%	63.0%	5.0%
Non-employed	23.5%	3.7%	14.2%	58.5%
Total	44.1%	8.2%	35.0%	12.6%

Sample has entered the labor market and has a schooling attainment of 14 years or more.

overeducated blacks are much less likely than whites to transition to a matched job, and third these matched spells are less persistent for blacks than for whites. Finally, as illustrated in Table 3.8, the persistence of overeducation decreases monotonically with AFQT scores. All of these patterns in individual transitions confirm that the aggregate persistence described earlier does not result from cancelling flows into and out of overeducation.

Table 3.5 also shows that transitions into overeducation are equally likely among

Table 3.7: Transition matrix by race

(White) $t-1$	OE	UE	t Matched	Non-emp.
Overeducated	65.4%	3.9%	26.0%	4.6%
Undereducated	27.4%	44.4%	22.6%	5.6%
Matched	27.9%	3.7%	64.8%	3.6%
Non-employed	29.4%	4.6%	16.8%	49.2%
Total	45.5%	7.1%	39.0%	8.5%
(Black) $t-1$	OE	UE	t Matched	Non-emp.
Overeducated	70.1%	6.8%	17.2%	5.9%
Undereducated	40.0%	48.0%	8.0%	4.0%
Matched	35.8%	6.4%	54.1%	3.7%
Non-employed	40.4%	5.3%	7.0%	47.4%
Total	54.2%	11.2%	24.0%	10.5%

Sample has entered the labor market and has a schooling attainment of 14 years or more.

workers who were undereducated, matched or non-employed in the previous year.[19] Among males, however, non-employed workers are much likelier to transition into overeducation than matched or undereducated workers (Table 3.6). Across all categories of workers, but especially for black and low-AFQT workers, the non-employed are more likely to transition into overeducation than into matched jobs (Table 3.7 and 3.8). In addition, transitions into matched jobs are more common among the overeducated than the non-employed. Taken together, these patterns suggest that for some workers overeducation is a pathway from non-employment into matched employment.

Finally, the NLSY79 data allows us to go beyond annual transitions, and report the hazard rates out of overeducated employment as a function of the duration of overeducation (see Figure 3.4).[20] After 3 years being overeducated, the probability of

[19] Note that some of the transitions into overeducation from undereducated or matched employment are likely to mask non-employment spells occurring between two consecutive interviews.

[20] For any given year t, these hazard rates are computed as the number of individuals leaving

Table 3.8: Transition matrix by AFQT quartile

(Lowest) $t-1$	OE	UE	t Matched	Non-emp.
Overeducated	72.8%	6.2%	15.4%	5.6%
Undereducated	42.1%	42.1%	5.3%	10.5%
Matched	37.1%	9.7%	50.0%	3.2%
Non-employed	44.7%	2.6%	0.0%	52.6%
Total	58.0%	11.0%	19.3%	11.7%

(2nd) $t-1$	OE	UE	t Matched	Non-emp.
Overeducated	71.5%	6.3%	16.9%	5.3%
Undereducated	46.0%	34.9%	12.7%	6.3%
Matched	31.7%	1.8%	58.7%	7.8%
Non-employed	30.8%	0.0%	15.4%	53.8%
Total	52.9%	6.7%	28.5%	12.0%

(3rd) $t-1$	OE	UE	t Matched	Non-emp.
Overeducated	66.3%	4.1%	24.7%	5.0%
Undereducated	26.7%	45.9%	20.7%	6.7%
Matched	30.5%	4.4%	61.4%	3.6%
Non-employed	32.4%	9.4%	17.3%	41.0%
Total	46.9%	8.7%	36.3%	8.2%

(Highest) $t-1$	OE	UE	t Matched	Non-emp.
Overeducated	62.4%	2.9%	30.5%	4.2%
Undereducated	21.8%	46.5%	28.9%	2.8%
Matched	25.6%	3.8%	68.0%	2.5%
Non-employed	23.8%	4.2%	19.0%	53.0%
Total	41.7%	6.4%	44.5%	7.4%

Sample has entered the labor market and has a schooling attainment of 14 years or more.

exiting that state, defined as starting a new matched or undereducated employment spell, drops from 39% to only 20%. This number further drops to 15% and 10% after 5 and 10 years, respectively. Overall, while this pattern is consistent with a negative duration dependence associated with overeducated employment, it could also result from compositional effects (permanent heterogeneity correlated with the hazard rate out of overeducated employment). We attempt to tell these two effects apart in the next section.

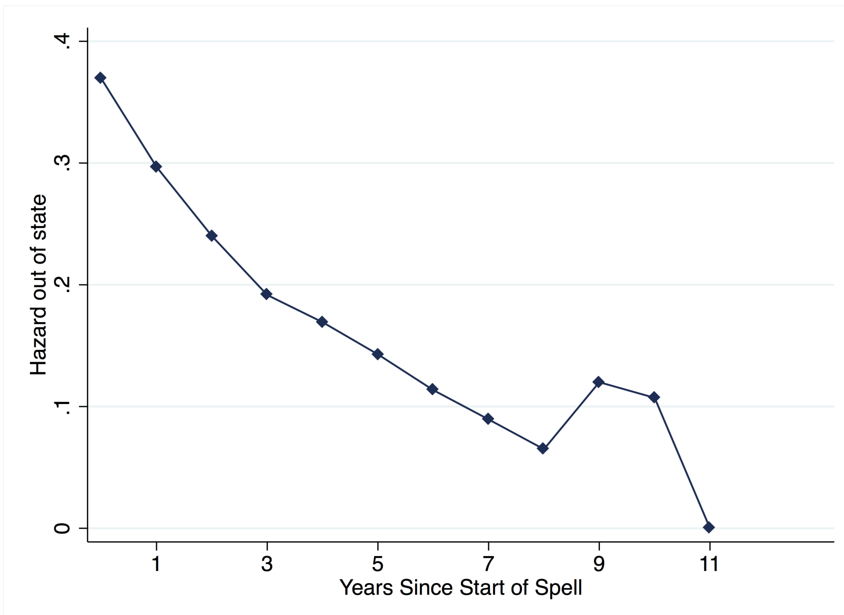

FIGURE 3.4: Hazard out of overeducation

3.4 Duration dependence versus dynamic selection

The results discussed so far provide some suggestive evidence of duration dependence in overeducation status, with a strongly decreasing hazard rate out of overeducation. However, in order to establish the role played by true, rather than spurious duration dependence, we need to control for dynamic selection on worker attributes.

overeducated employment in year t, divided by the number of individuals who are still overeducated at the beginning of year t.

Specifically, we assume that the duration of the first spell of overeducated employment is determined by a mixed proportional hazard model, where the baseline duration follows a Weibull distribution.[21] The probability distribution function (pdf.) and cumulative distribution function (cdf.) of the duration of the overeducation spell, conditional on the set of observed individual characteristics x_i and the unobserved heterogeneity ν_i, are respectively given by:[22]

$$f\left(t|\mathbf{x}_i,\nu_i,\alpha,\theta\right)=\exp\left(\mathbf{x}_i\theta\right)\alpha t^{\alpha-1}\nu_i\exp\left[-\exp\left(\mathbf{x}_i\theta\right)t^{\alpha}\nu_i\right], \tag{3.1}$$

$$F\left(t|\mathbf{x}_i,\nu_i,\alpha,\theta\right)=1-\exp\left[-\exp\left(\mathbf{x}_i\theta\right)t^{\alpha}\nu_i\right]. \tag{3.2}$$

Following Heckman and Singer (1984), we assume that the unobserved heterogeneity follows a discrete distribution with R points of support. The parameters (α,θ) and the unobserved heterogeneity distribution are then estimated by maximizing the log-likelihood of the data, which is given by:

$$\ell=\sum_{i=1}^{N}\left[d_i\log f\left(t_i|\mathbf{x}_i,\alpha,\theta\right)+\left(1-d_i\right)\log\left[1-F\left(T_i|\mathbf{x}_i,\alpha,\theta\right)\right]\right]$$

where N is the number of individuals in the sample with at least one overeducated spell, $d_i=1$ if individual i leaves the overeducated state before the end of the survey (0 otherwise), t_i is the duration of the first overeducation spell (observed if $d_i=1$) and T_i is the length of time to the end of the survey. The pdf. of the overeducation spell duration is given by $f\left(t|\mathbf{x}_i,\alpha,\theta\right)=E_{\nu}\left(f\left(t|\mathbf{x}_i,\nu_i,\alpha,\theta\right)\right)$ and the cdf. is given by $F\left(t|\mathbf{x}_i,\alpha,\theta\right)=E_{\nu}\left(F\left(t|\mathbf{x}_i,\nu_i,\alpha,\theta\right)\right)$, where $E_{\nu}(.)$ denotes the expectation operator

[21] By definition, this model is estimated on individuals who have at least one overeducated spell, which mechanically excludes individuals with only 12 years of schooling.

[22] While using a parametric specification allows us to get more precise estimates, it is important to note that the mixed proportional hazard model is identified nonparametrically from single-spell data (see Elbers and Ridder, 1982).

with respect to the distribution of ν. Throughout our analysis, we consider that an overeducation spell ends when the individual starts a new employment spell in an occupation which does not require less than their level of schooling.[23]

It is interesting to compare the estimates for $R = 1$ (i.e., without unobserved heterogeneity) and $R = 2$ (i.e., with two unobserved heterogeneity types).[24] Columns 2,4 and 6 of Table 3.9 report the estimation results corresponding to the case without unobserved heterogeneity with alternative sets of individual controls. The estimated α is well below one (between 0.77 and 0.84 depending on the specification), which means that the hazard out of overeducation is strongly decreasing in the duration of overeducation even after controlling for *observed* heterogeneity. In other words, while part of this negative duration dependence is attributable to selection on observables, as shown by the increase in the estimated α parameter going from specification (3) to (1), the exit rate is still declining with the duration of the overeducation spell after controlling for an extensive set of observed characteristics.

[23] The NLSY79 allows us to compute the required level of education only for the current or most recent job at the time of each interview (the "CPS job"). To keep the exposition simple, the duration model we present in the text ignores the existence of job spells that might have occurred in between two interviews. We also estimated a model that explicitly takes into account these between-interview employment spells, treating the corresponding overeducation status as missing, which resulted in negligible differences in the estimation results.

[24] We estimated the model with more than two types, but including these additional types did not significantly improve the fit of the model. It resulted in higher BIC criteria than the model with two types while attributing similar values for the key parameters.

Table 3.9: Duration models (first overeducation spell)

	(1)				(2)				(3)			
	MPH		PH		MPH		PH		MPH		PH	
α	1.031	(0.0459)	0.836***	(0.0262)	0.989	(0.0347)	0.822***	(0.0261)	1.115***	(0.0404)	0.768***	(0.0261)
$P(V_1)$	0.289***	(0.0803)			0.218***	(0.0565)			0.472***	(0.0440)		
$P(V_2)$	0.711***	(0.0803)			0.782***	(0.0565)			0.528***	(0.0440)		
V_1	0.153***	(0.0615)			0.094***	(0.0557)			0.142***	(0.0243)		
V_2	1.344***	(0.1167)			1.253***	(0.0740)			1.768***	(0.1202)		
Constant	−3.432***	(0.3097)	−3.092***	(0.2324)	−3.024***	(0.3293)	−2.723***	(0.2041)	−4.086***	(0.1095)	−3.339***	(0.0880)
Black	−0.313*	(0.1860)	−0.265**	(0.1344)	−0.288*	(0.1653)	−0.287**	(0.1265)				
Hispanic	0.136	(0.2694)	0.015	(0.1665)	0.261	(0.3226)	0.015	(0.1610)				
Female	0.002	(0.0977)	−0.176***	(0.0664)	0.102	(0.0908)	−0.078	(0.0633)				
Born in The United States	−0.294	(0.2541)	−0.150	(0.1884)	−0.318	(0.3206)	−0.198	(0.1869)				
AFQT†	0.149***	(0.0547)	0.101***	(0.0390)	0.150***	(0.0522)	0.104***	(0.0370)				
Rotter Scale in 1979†	0.026	(0.0463)	0.027	(0.0324)	0.016	(0.0447)	0.022	(0.0319)				
Sociability in 1985†	−0.026	(0.0526)	0.010	(0.0319)	−0.012	(0.0513)	0.019	(0.0321)				
14 Years of edu.	−0.733***	(0.1071)	−0.587***	(0.0769)	−0.829***	(0.1053)	−0.692***	(0.0739)				
18 Years of edu.	−1.706***	(0.1297)	−1.388***	(0.0969)	−1.560***	(0.1269)	−1.290***	(0.0951)				
GED (No HS diploma)	0.257	(0.3226)	0.153	(0.2489)	0.211	(0.3194)	0.102	(0.2463)				
Northeast	0.035	(0.1185)	−0.002	(0.0943)								
South	0.346**	(0.1453)	0.223***	(0.0822)								
West	0.035	(0.1407)	0.004	(0.1012)								
Unemployment rate†	−0.073	(0.0510)	−0.066*	(0.0368)								
HH in SMSA	0.088	(0.1465)	0.101	(0.1087)								
Tenure (1k wks.)	−0.230	(0.4995)	−0.010	(0.3710)								
Work experience (1k hrs.)†	−0.050	(0.0675)	−0.031	(0.0519)								
Work experience^2 (1k hrs.)†	−0.036	(0.0527)	−0.017	(0.0305)								
Number of jobs†	−0.025	(0.0488)	0.016	(0.0333)								
Total unemp. (wks.)†	−0.166***	(0.0579)	−0.133***	(0.0467)								
Occupational hazards	−0.207***	(0.0728)	−0.191***	(0.0465)								
LL		5,321.9		5,348.6		5,344.9		5,374.8		5,442.7		5,501.3
Observations		1,648		1,648		1,648		1,648		1,648		1,648

Model also includes controls for age, mother and father's education level in 1979, a dummy for living with both parents in 1979, and a dummy for living in an urban area at the time of interview. Models estimated on the first overeducation spells of individuals with at least one overeducation spell. Unemployment rate (by region) corresponds to the annual unemployment rate for each Census region (BLS Local Area Unemployment Statistics data). Standard errors in parentheses.

† Values converted to Z-scores

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

By contrast, once we allow for unobserved heterogeneity, we obtain values for α that are very close to one, signifying the absence of true duration dependence (Table 3.9, columns 1, 3, and 5). In other words, the duration of the first overeducated employment spell does not seem to have a significant impact on the probability to exit overeducation. The estimation results point to the existence of two groups of individuals with markedly different dynamics. These two groups are identified from the variation in the duration of the first overeducation spell, conditional on observed characteristics. The first group has a low hazard (type 1, 28.9% of the ever-overeducated in the sample) while the second group is much more likely to exit overeducation quickly (type 2, 71.1%). As type 2s exit the pool of overeducated individuals, the probability that a random individual exits overeducation declines, as they are more and more likely to be a low-hazard, type 1 individual.

The ratio between the two unobserved heterogeneity parameters is $\frac{V_1}{V_2} = 0.11$, which implies that type 2s exit overeducation almost 10 times as fast as type 1s. Interestingly, the stark difference in hazard rates between the two unobserved groups suggests that overeducation follows different mechanisms in each case. The high exit rate in group 2 is consistent with a frictional view of overeducation. For the remaining, low-hazard third of the sample, it could be that their aptitude (after controlling for AFQT) is not sufficient for jobs that match their level of formal schooling. Alternatively, they may have preferences for non-pecuniary, unobserved job characteristics found in some of the jobs that require less schooling, thus translating into highly persistent overeducation. In the next section we explore these ideas further as we examine whether the two types differ in the wage patterns they exhibit.

The coefficients on observable characteristics allow us to complement and refine the analysis of the raw transition dynamics presented in Section 3.3. The negative coefficient on blacks in all specifications of the duration model implies that the lower exit rates out of overeducation for that group are robust to controlling for observable and unobservable heterogeneity. Similarly, the positive effect of AFQT scores on overeducation exit also carries over from the raw transitions into the mixed proportional hazard model. The coefficient for women is also significantly negative when the full set of observable characteristics is included in the proportional hazard model. However once unobserved heterogeneity is taken into account, being female no longer has a direct effect on the probability of exiting overeducation. This runs against the notion that discrimination would keep women in jobs where they are overeducated.

Lastly, an interesting finding is that time spent unemployed in the past has a negative effect on overeducation exit. This result complements a previous finding in the literature that longer time spent unemployed reduces the probability of finding a job (see e.g. Kroft, Lange, and Notowidigdo, 2013). It also suggests that higher overeducation persistence is one of the mechanisms through which past unemployment affects future wages (see e.g. Schmieder, von Wachter, and Bender, 2013).

3.5 Effects of overeducation on wages

3.5.1 Wage dynamics

While the regressions found in the literature focus on the cross-sectional correlation between current overeducation and wages, we examine whether initial overeducation is also associated with lower wages later in the career. Figure 3.5 describes the median hourly wage among workers with 14, 16 and 18 years of education as they progress through their career, conditional on their overeducation status. The striking result here is that the penalties from overeducation at the start of the career appear very persistent over time for the first two schooling categories (14 and 16 years). In the following sections, we provide additional evidence of these effects by showing that the negative association between past overeducated employment and wages still holds after controlling for observed heterogeneity (including AFQT scores and measures of non-cognitive skills), and for the unobserved heterogeneity types identified in the duration analysis.

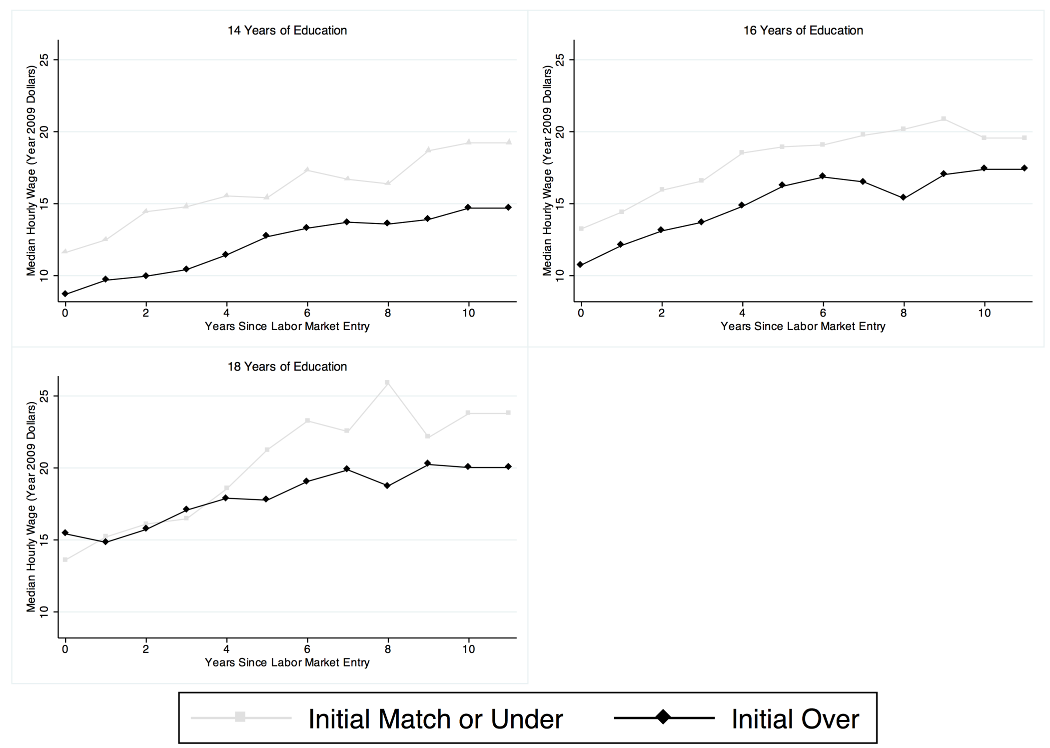

FIGURE 3.5: Path of median hourly wage

3.5.2 Augmented wage regressions

The impact of overeducation on wages has been measured in the literature by applying OLS to the following log-wage equation introduced by Duncan and Hoffman (1981):

$$\log w_{it} = \alpha^r S_{it}^r + \alpha^o S_{it}^o + \alpha^u S_{it}^u + X_{it}'\beta + \varepsilon_{it}$$

where, for any given individual i in year t, S_{it}^r, S_{it}^o and S_{it}^u denote respectively the number of years of required schooling, years of overeducation (years of schooling above the required level) and years of undereducation (years of schooling below the required level), X_{it} a vector of controls (including ability measures, socio-demographic background characteristics, labor market experience and experience squared) and ε_{it} an idiosyncratic productivity shock. This model, which nests the standard Mincerian wage regression ($\alpha^r = \alpha^o = -\alpha^u$), allows for the estimation of separate wage returns to the (i) required years of schooling, (ii) years of overeducation, and (iii) years of undereducation.

Table 3.10, Panel 1, presents the pooled OLS parameter estimates for this model using our data.[25] Unlike for the rest of our analysis, individuals with 12 years of schooling are part of the estimation sample. While none of these individuals are overeducated, including them helps identify the returns to the required years of schooling, years of overeducation and undereducation. The results from our sample are broadly consistent with prior studies. Namely, we see a return of 9.6% for each additional year of education, less than half the rate of return (3.8%) for additional years of education over the required level, and substantial wage penalties for years of undereducation (-6.9% for each additional year of undereducation).

These estimates also reveal a number of expected results. In particular, we see a gender wage gap of 15.5%, and a significant compensating wage differential for hazardous occupations. We also find significant wage premia relative to the North Central region of 9.7% and 6.9% for Northeast and West, respectively, consistent with higher average costs of living in these regions. Lastly, measures for labor market experience all have the expected sign. Those with greater occupation-specific tenure, and total labor market experience have higher wages. Conversely, those with frequent and/or long unemployment spells (as measured by total unemployment experience in weeks), and those frequently switching jobs during the year have lower wage rates, all else equal.

[25] For the occupations such that the discrepancy between the frequency of the mode and the second most frequent schooling level is less than 15 percentage points, individuals whose schooling level falls with the range defined by the two most frequent schooling levels are assumed to be matched, and years of required schooling is set equal to their actual level of education. Individuals whose schooling level is higher (lower) than the upper (lower) bound of the aforementioned range are assumed to be overeducated (undereducated), and years of required schooling is set equal to that upper (lower) bound.

Table 3.10: Augmented log-wage regressions

	(1)		(2)	
	$\ln(HourlyWage)$		$\ln(HourlyWage)$	
Required education (yrs.)	0.096***	(0.0046)	0.108***	(0.0051)
Schooling above required (yrs.)	0.038***	(0.0055)	0.066***	(0.0065)
Schooling below required (yrs.)	−0.069***	(0.0065)	−0.080***	(0.0069)
Black	0.007	(0.0253)	0.003	(0.0256)
Hispanic	0.041	(0.0339)	0.046	(0.0338)
Female	−0.155***	(0.0148)	−0.148***	(0.0147)
Born in The United States	−0.049	(0.0390)	−0.049	(0.0389)
AFQT†	0.110***	(0.0126)	0.111***	(0.0128)
Rotter Scale in 1979	−0.004	(0.0029)	−0.004	(0.0029)
Sociability in 1985	0.006	(0.0104)	0.005	(0.0104)
GED (No HS diploma)	0.001	(0.0256)	−0.010	(0.0254)
Northeast	0.097***	(0.0191)	0.101***	(0.0190)
South	−0.027	(0.0167)	−0.022	(0.0167)
West	0.069***	(0.0213)	0.076***	(0.0213)
Unemployment rate	0.001	(0.0032)	0.000	(0.0032)
HH in SMSA	0.129***	(0.0185)	0.131***	(0.0186)
Tenure (1k wks.)	0.250***	(0.0356)	0.264***	(0.0353)
Work experience (1k hrs.)	0.035***	(0.0030)	0.036***	(0.0030)
Work experience^2 (1k hrs.)	−0.008***	(0.0010)	−0.008***	(0.0010)
Number of jobs	−0.060***	(0.0069)	−0.059***	(0.0069)
Total unemp. (wks.)	−0.001***	(0.0002)		
Occupational hazards	0.045***	(0.0066)	0.044***	(0.0066)
Constant	1.052***	(0.1028)	0.962***	(0.1068)
Unemployed (t-1)			−0.076***	(0.0239)
Unemployed (t-2)			−0.028	(0.0182)
Unemployed (t-3)			−0.003	(0.0158)
Unemployed (t-4)			−0.038*	(0.0219)
Overeducated (t-1)			−0.039***	(0.0129)
Overeducated (t-2)			−0.042***	(0.0105)
Overeducated (t-3)			−0.039***	(0.0114)
Overeducated (t-4)			−0.026**	(0.0127)
Overeducated (t-1), spell ongoing			0.013	(0.0182)
Overeducated (t-2), spell ongoing			0.021	(0.0279)
Overeducated (t-3), spell ongoing			0.007	(0.0386)
Overeducated (t-4), spell ongoing			−0.007	(0.0589)
Adjusted R^2		0.372		0.375
F		112.6		79.9
Observations		12,775		12,775

Model also includes controls for age, mother and father's education level in 1979, a dummy for living with both parents in 1979, and a dummy for living in an urban area at the time of interview. Model estimated on individuals with 12 years of schooling or higher, at least four years after labor market entry. Unemployment rate (by region) corresponds to the annual unemployment rate for each Census region (BLS Local Area Unemployment Statistics data). Standard errors are clustered at the individual level.
† Values converted to Z-scores
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Following-up on the dynamic patterns discussed in Section 6.1, we report in Panel 2 of Table 3.10 the estimation results from a log-wage regression which is further augmented with four lags in overeducation status. We also include interactions between lagged overeducation status and an indicator for the fact that the overeducation spell is ongoing. This disentangles the effect of past completed overeducation spells on wages when an individual is currently matched or undereducated, from the effect of the duration of an ongoing overeducated spell on current wages if an individual is still overeducated. Notably, we find wage penalties of 2.6-4.2% per year associated with up to four lags of overeducation. Those effects, which show that having been overeducated in the past is associated with lower wages in current jobs irrespective of whether one is overeducated or not in them, are both statistically and economically significant.[26] It is interesting to compare the magnitude of these "scarring" effects with those generated by past unemployment. While the penalty associated with the first lag of unemployment is found to be sizeably stronger than that associated with overeducation (7.6% vs. 3.9%), the penalty associated with past unemployment is also less persistent as further lags do not significantly affect wages.

These results show that the overeducation scarring effects suggested by the raw wage data remain after including a rich set of controls, including measures of cognitive and non-cognitive ability. To the extent that, consistent with the cyclical upgrading literature (see, e.g., Bils and McLaughlin, 2001), overeducated employment is likely to be more frequent during recessions, it follows that overeducation is a candidate mechanism behind the negative and persistent wage effects of graduating during a recession recently uncovered in the literature (Kahn, 2010, Liu, Salvanes, and Sorensen, Oreopoulos, von Wachter, and Heisz, 2012, and Altonji, Kahn, and Speer, 2014).

[26] Older lags, on the other hand, do not generate significant additional penalties.

3.5.3 Accounting for heterogeneity types

Next, we extend our wage regression to account for unobserved heterogeneity in productivity. To do so, we add a type-specific intercept and estimate the model by weighted least squares, using as weights the posterior type probabilities obtained from the mixed proportional hazard model that we estimated in the previous section. For each individual in the sample, we use Bayes' rule to compute the posterior probabilities of being each type from the estimated likelihood of the duration of the first overeducation spell and that individual's characteristics.

The duration model estimates imply a strong correlation between overeducation status and the type probabilities. It follows that the inclusion of the type-specific intercept will mitigate the endogeneity biases that may affect the estimated returns to schooling and the penalties for past overeducation, to the extent that heterogeneity in the exit rate out of overeducated employment overlaps with confounding productivity heterogeneity.[27] Note that this model can only be estimated on the sample used for the mixed proportional hazard model, which is comprised of individuals with at least one overeducation spell.

Estimation results are reported in Table 3.11. The first two panels correspond to the model without lagged overeducation regressors, while Panels 3 and 4 incorporate four lags in overeducation status. In both cases we contrast our type-weighted least squares model (Panels 1 and 3) with simple OLS (Panels 2 and 4). A robust finding across specifications is that returns to overeducation are much lower than returns to required schooling, resulting in large wage penalties associated with overeducated

[27] Alternative approaches to correcting for the correlation between unobserved productivity and overeducation status present serious challenges in this context. The instrumental variables approach requires a valid instrument for required schooling in addition to the usual instrument for schooling. In the fixed effect approach, the returns to schooling are identified off of individuals changing jobs with different required schooling levels, and, in particular, switching across matched, undereducated and overeducated jobs. These transitions are likely to be correlated with changes in unobserved wage determinants. Besides, the fixed effect estimates may not be generalizable to those who never switch overeducation status.

Table 3.11: Augmented log-wage regressions with heterogeneity types

	(1)		(2)		(3)		(4)	
	$\ln(HourlyWage)$		$\ln(HourlyWage)$		$\ln(HourlyWage)$		$\ln(HourlyWage)$	
Required education (yrs.)	0.074***	(0.0090)	0.076***	(0.0086)	0.075***	(0.0087)	0.076***	(0.0086)
Schooling above required (yrs.)	0.019*	(0.0099)	0.015	(0.0092)	0.029***	(0.0096)	0.028***	(0.0098)
Schooling below required (yrs.)	−0.062***	(0.0179)	−0.060***	(0.0180)	−0.064***	(0.0189)	−0.063***	(0.0175)
Black	0.047	(0.0417)	0.050	(0.0450)	0.040	(0.0435)	0.041	(0.0444)
Hispanic	0.034	(0.0600)	0.030	(0.0609)	0.037	(0.0607)	0.036	(0.0578)
Female	−0.099***	(0.0215)	−0.100***	(0.0225)	−0.097***	(0.0235)	−0.097***	(0.0236)
Born in The United States	−0.115**	(0.0544)	−0.114*	(0.0595)	−0.117**	(0.0548)	−0.116**	(0.0536)
AFQT†	0.188***	(0.0368)	0.189***	(0.0410)	0.181***	(0.0364)	0.180***	(0.0373)
Rotter Scale in 1979	−0.006	(0.0049)	−0.006	(0.0053)	−0.006	(0.0052)	−0.006	(0.0050)
Sociability in 1985	0.004	(0.0180)	0.005	(0.0190)	0.001	(0.0191)	0.001	(0.0176)
Has GED (No HS diploma)	0.018	(0.1106)	0.026	(0.1048)	0.015	(0.1010)	0.017	(0.1115)
Northeast	0.131***	(0.0313)	0.128***	(0.0325)	0.134***	(0.0329)	0.133***	(0.0300)
South	0.010	(0.0278)	0.004	(0.0304)	0.009	(0.0300)	0.008	(0.0292)
West	0.052	(0.0353)	0.052	(0.0343)	0.059	(0.0378)	0.059*	(0.0322)
Unemployment rate	0.006	(0.0059)	0.005	(0.0057)	0.007	(0.0064)	0.007	(0.0057)
HH in SMSA	0.105***	(0.0343)	0.110***	(0.0331)	0.112***	(0.0349)	0.114***	(0.0343)
Tenure (1k wks.)	0.152**	(0.0620)	0.147**	(0.0663)	0.166**	(0.0665)	0.167***	(0.0610)
Work experience (1k hrs.)	0.039***	(0.0069)	0.041***	(0.0069)	0.043***	(0.0077)	0.044***	(0.0070)
Work experience^2 (1k hrs.)	−0.012***	(0.0022)	−0.012***	(0.0023)	−0.013***	(0.0025)	−0.013***	(0.0023)
Number of jobs	−0.061***	(0.0121)	−0.062***	(0.0134)	−0.060***	(0.0124)	−0.060***	(0.0116)
Total unemp. (wks.)	−0.002***	(0.0006)	−0.001***	(0.0006)				
Occupational hazards	0.054***	(0.0146)	0.051***	(0.0160)	0.049***	(0.0159)	0.049***	(0.0151)
Type 1	−0.120**	(0.0571)			−0.033	(0.0552)		
Constant	1.161***	(0.2016)	1.129***	(0.2226)	1.321***	(0.2172)	1.321***	(0.2016)
Unemployed (t-1)					−0.112*	(0.0635)	−0.114**	(0.0574)
Unemployed (t-2)					−0.051	(0.0385)	−0.051	(0.0407)
Unemployed (t-3)					−0.013	(0.0402)	−0.013	(0.0379)
Unemployed (t-4)					−0.033	(0.0479)	−0.033	(0.0490)
Overeducated (t-1)					−0.060***	(0.0131)	−0.062***	(0.0131)
Overeducated (t-2)					−0.048***	(0.0118)	−0.050***	(0.0118)
Overeducated (t-3)					−0.041***	(0.0117)	−0.044***	(0.0120)
Overeducated (t-4)					−0.030**	(0.0136)	−0.032**	(0.0133)
Overeducated (t-1), spell ongoing					0.031	(0.0200)	0.030	(0.0204)
Overeducated (t-2), spell ongoing					0.012	(0.0306)	0.010	(0.0286)
Overeducated (t-3), spell ongoing					0.027	(0.0388)	0.026	(0.0354)
Overeducated (t-4), spell ongoing					0.010	(0.0636)	0.006	(0.0547)
Observations		3,906		3,906		3,906		3,906

Model includes controls for age, mother and father's education level in 1979, a dummy for living with both parents in 1979, and a dummy for living in an urban area at the time of interview. Estimates not reported.
Model estimated on individuals with at least one overeducated spell, at least four years after labor market entry. Specifications (1) and (3) are estimated by Weighted Least Squares, using as weights the posterior type probabilities from the MPH model, and Specifications (2) and (4) are estimated by pooled OLS.
Unemployment rate (by region) corresponds to the annual unemployment rate for each Census region (BLS Local Area Unemployment Statistics data).
Standard errors are clustered at the individual level.
† Values converted to Z-scores
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(as opposed to matched) employment. We find that returns to overeducation are between 4.6 and 6.1 percentage points lower than returns to required schooling. For individuals who graduate with a four year college degree, this corresponds to wage penalties ranging from 18.4% (Panel 3) to 24.4% (Panel 2) if they work in a high school job instead of a college job.

Individuals with a low propensity to exit overeducation (Type 1) tend to have significantly lower wages than their high propensity counterparts. In the first specification, the wage differential is 12%. However, it substantially decreases in magnitude, by more than 70%, once we control for lagged overeducation (Panel 3). This implies that Type 1 individuals have lower wages mostly as a consequence of having been overeducated in the past, through a form of scarring effects. While not statistically significant at standard levels, the remaining penalty is economically non-negligible at 3.3%. This provides some suggestive evidence that Type 1 individuals may also be inherently less productive than Type 2s - for whom overeducation is primarily frictional - over and above any adverse effects from past overeducation spells.

It is also interesting to note that, for all four specifications, both the returns to required schooling and years of overeducation and the penalties to years of undereducation are lower than in the wage regressions discussed in the previous section. This finding is consistent with overeducation being negatively correlated with ability, since the sample used for these wage regressions is composed of individuals with at least one reported overeducation spell.

Lastly, controlling for the heterogeneity type does not significantly affect the returns to required schooling, years of overeducation or penalties to years of undereducation. Importantly, the penalties for having been overeducated in the past (Panel 3) also remain significant at the 1% level, sizeable and quantitatively similar to the estimates obtained without controlling for unobserved heterogeneity. In other words, they are not likely to be driven by unobserved productivity differences between those who have and have not been overeducated in the previous years. Instead, these results support the existence of causal scarring effects from past overeducation, reminiscent of what has been identified in the case of long-term unemployment (Arulampalam, Gregg, and Gregory, 2001).

3.6 Conclusion

Although economists and policy-makers have long been concerned about the determinants and wage effects of overeducation, little is still known about its dynamics along the career. This paper combines data from the NLSY79 and CPS to provide the first analysis of the career dynamics of overeducated U.S. workers. Overall, we find overeducation to be a persistent phenomenon, particularly for blacks and low-ability workers, with sizeable and lasting negative effects on wages. These wage penalties bear similarities with the scarring effects that have been found to accompany prolonged unemployment spells.

Controlling for dynamic selection on unobservable characteristics is key in this context. While the exit rate from overeducation decreases quickly over the duration of an overeducation spell, the estimation of a mixed proportional hazard model suggests that, after controlling for both observed and unobserved heterogeneity, true duration dependence is not strong. Rather, the propensity to exit overeducated employment appears to be very heterogeneous among workers.

We find that past unemployment increases the duration of future overeducation spells, thus indicating that overeducation is likely to be one the mechanisms through which the scarring effects on earnings associated with unemployment spells operate. The scarring effects associated with overeducation could also account for some of the negative wage effects of graduating during a recession which have been recently uncovered in the literature.

From a policy standpoint, since both overeducation and unemployment (see, e.g., Saporta-Eksten, 2014) are associated with negative and persistent wage shocks, a relevant question becomes whether a marginal increase in unemployment duration is more or less harmful, in terms of lifetime earnings, than entering an overeducation spell. The answer has bearing on the design of unemployment insurance - such as whether early exit should be encouraged at the cost of more mismatch - and the appropriate evaluation of the performance of employment agencies.

In sum, our results show that overeducation is a complex phenomenon that involves a number of the classical ingredients of labor economics: human capital, search frictions, ability differences and, perhaps, compensating wage differentials. In order to quantify the importance of each mechanism and explore the effects of unemployment insurance or schooling subsidy programs, a promising avenue would be to estimate a dynamic model of schooling and occupational choice that would nest these different channels, while allowing for correlated unobserved heterogeneity in job mobility and productivity.

4

A Factor Analysis of College Major: Explaining Gaps by Race, Gender, and Socioeconomic Status

Disparities in educational attainment and postsecondary outcomes by gender, race, or family background are of eminent concern to policymakers and occur at several stages, including high school completion, college attendance, and major choice. First, successfully completing high school is a well-established and important marker of future economic well-being. Second, students that then choose to continue on to, and graduate from, college earn significantly more than high school graduates (Grogger and Eide, 1995; Card, 1999). And third, labor market outcomes can further diverge if socio-demographic groups differ systemically in their choice of postsecondary institutions and field of study, given the sizable returns to college quality and select majors. The wage gains associated with certain majors rival even that of the college wage premium (Altonji, Blom, and Meghir, 2012). As such, these educational decisions carry significant implications for future labor market outcomes.

This paper examines a set of outcomes governing the crucial transition from high school to college, focusing on educational attainment at the end of secondary schooling, college application behavior, and major intentions. It documents gaps by gender, ethnicity, and socioeconomic status, expanding the literature to explore the interactions between these attributes. In this way we are able to fully explore gaps in educational outcomes across the socioeconomic spectrum for a given ethnicity and/or gender group, for example distinguishing between the plight of disadvantaged white males from disadvantaged African American males.

We examine three possible explanations for cross-group disparities: high school characteristics, cognitive skills, and non-cognitive skills. Existing research offers mixed evidence on the role of school quality in explaining equity gaps (Cook and Evans, 2000; Fryer and Levitt, 2004; Autor et al., 2016).[1] Proxies for cognitive skills, on the other hand, have been established as relevant determinants for both educational attainment and major choice (Paglin and Rufolo, 1990; Weinberger, 1999; Heckman and Krueger, 2005). In recent years there has been a shift from an unidimensional focus on cognitive abilities toward the contributions of both cognitive and non-cognitive personality factors in ameliorating disadvantage (Heckman, Stixrud, and Urzúa, 2006; Cunha, Heckman, and Schennach, 2010).

Operationalizing this type of analysis is challenging because it requires information on students' schooling attainment and postsecondary intentions, as well as data from high school and earlier that can be used to identify school characteristics, cognitive skills, and non-cognitive skills. We also require a sufficiently large sample that enables the exploration of outcomes for more granular classifications of students

[1] The majority of evidence in the literature focuses on racial gaps in achievement. The contribution of school quality to the narrowing black-white performance gap is less than 12% for math in one study (Cook and Evans, 2000). While it plays a more prominent role in explaining racial gaps for students entering kindergarten and first grade, the role of high school quality diminishes in a re-evaluation of the same students several years later (Fryer and Levitt, 2004, 2006). Evidence on the role of school quality in determining gender gaps is more limited, although what exists suggests that men and women respond differentially to school inputs. Autor et al. (2016) finds varying gradients between school quality and educational and behavioral outcomes by gender. Boys benefit more from high quality middle schools, such that the usual female advantage in outcomes is declining in school quality.

by socioeconomic status (SES). We overcome these challenges by using administrative data from the North Carolina Education Research Data Center (NCERDC). The sociodemographic information contained in the NCERDC allows us to segment students by gender, ethnicity, and SES.[2] Rich data on student standardized test scores, course-level grades, disciplinary infractions, homework effort, and high school attended enable the construction of covariates corresponding to each of our three possible explanations. Data on school exit, graduation, and dropouts, alongside College Board information on college applications and major intentions, complete the dataset. While the NCERDC forms the core sample, we supplement our analyses using the nationally representative National Longitudinal Survey of Youth 1997 (NLSY97) to ensure that the educational gaps we document are valid outside of North Carolina.

One of the challenges in analyzing the contribution of cognitive and non-cognitive factors to academic outcomes is that observed measures (e.g. test scores) are noisy proxies for underlying skills. Furthermore, they are themselves the products of a dynamic process in which familial background and school inputs, such as teacher quality, exert influences on the student. To address measurement error and endogeneity concerns, we specify a sequential model of educational attainment, college application intensity, and college major choice that is consistent with the factor analysis literature, in particular Cameron and Heckman (2001) and Aucejo (2015). We identify two latent skills that are mutually independent, and fixed by the time students make decisions about their transition to higher levels of secondary education and college.[3] These latent cognitive and non-cognitive skills are allowed to be correlated with observed demographic characteristics, which enables us to analyze the

[2] We focus on eight student subgroups, representing the interactions between dichotomous variables for gender, white or African American, and free and reduced lunch eligibility.

[3] Restricting each factor to be constant after students enter school is a common assumption in the literature, as documented in Aucejo (2015).

effects of each factor by gender, race, and SES.[4] The measurement system we use also allows observed cognitive and non-cognitive proxies to vary by other influences such as schooling inputs, which allows us to address measurement error. We are also able to control for the dynamic selection process that drives students' sequential educational choices. Doing so enables us to separately identify both factors, their distributions across demographic groups, and the importance of each factor at each stage of the model.

We find significant observed differences in student outcomes by race, gender, and SES. Disadvantaged white females are less likely to finish high school than their disadvantaged black female peers, in spite of performing better on 8th grade standardized tests in both math and reading. In contrast, the high school attainment of high SES black females are comparable to their high SES white female peers, despite a deficit in earlier academic performance. Demarcations of major choice along gender lines are consistent across ethnicity and throughout the SES distribution. Females are more likely to choose health-related majors over STEM, while the opposite holds for men. We specify the factor model to identify the relative contributions of cognitive and non-cognitive skills to each outcome. In doing so, we find that controlling for these latent factors allows us to explain gaps in outcomes across socio-demographic groups that cannot be accounted for with simple reduced form methods. The factor

[4] The literature finds that non-cognitive factors such as social skills and effort are important determinants of graduation from high school and college, and a student's ultimate level of educational attainment after high school. Heckman, Stixrud, and Urzúa (2006) show that the effect of non-cognitive skills on the probability of graduating from high school and continuing to college are comparable in magnitude to that of cognitive skills. Factor models have also demonstrated the role that differences in non-cognitive abilities play in educational attainment across gender and race. Urzúa (2008) uses The National Longitudinal Survey of Youth, 1979 (NLSY79) to show that unobserved non-cognitive ability is a significant determinant of black-white gaps in educational attainment. Aucejo (2015) extends this analysis using data from the NLSY97 to show both factors play a significant role in explaining the gender gap in educational attainment both within and across races. In particular, Aucejo finds that while cognitive factors are a significant determinant of transitions between education levels, non-cognitive factors primarily explain the gender gap in educational attainment, particularly for African Americans.

model explains high school attainment disparities between advantaged white females and all groups with the exception of low SES white males. The combination of cognitive and non-cognitive factors also close the gap on college application intensity for everyone except for high SES black students. One area where the model comes up short is in college major intentions, where gaps remain robust to the explanatory factors considered.

Our paper makes several contributions to research on disparities in educational outcomes and the transition from high school to college. First, the categorization of students along the three dimensions of gender, ethnicity, and SES permits a more detailed understanding of students' academic success and challenges than previously available US-based research. While studies on racial or gender gaps are prevalent, socioeconomic background is usually relegated to a control variable. Studies on the black-white achievement gap finds that socioeconomic factors account for anywhere between a quarter up to 85% of disparities at various levels of educational experience (Fryer and Levitt, 2004, 2006; Murnane et al., 2006; Reardon and Robinson, 2008; Clotfelter, Ladd, and Vigdor, 2009).[5] A small selection of papers that explore the full interaction of race, gender, and SES acknowledges the importance of distinguishing between ethnicity and SES (Strand, 2014). Ethnic disparities can vary by level of disadvantage, and outcomes can diverge markedly for an ethnic group across the socioeconomic spectrum. For example, using race to confer disadvantage obscures the narrative on deprived white students and better-off African American students.

[5] Fryer and Levitt (2004) is relatively unique in the literature in finding that the vast majority of the black-white achievement gap can be accounted for using a set of socioeconomic controls. While they observe this result for kindergarteners, the explanatory power of SES diminishes significantly by the time students reach third grade (Fryer and Levitt, 2006). Murnane et al. (2006) documents that one-third of the math achievement gap during kindergarten and third grade is eliminated after accounting for family covariates such as SES background. Clotfelter, Ladd, and Vigdor (2009) find that a vector of socioeconomic factors including free and reduced-price lunch eligibility, parental education, gender, age, and district and region type explain approximately one-third of black-white gaps in math and reading from grades 3-8.

Our results point to a more complex set of gaps in student achievement and postsecondary intentions. High SES black students lag behind their advantaged white peers, with the gap more pronounced among males than females, raising questions on the roles of institutional barriers and social expectations that are especially salient for African American students. In a finding that echoes Strand (2014), widespread gaps underscore the scale of challenges faced by poorer white students of both genders.

Second, our paper characterizes how cognitive and non-cognitive factors, which have been shown to be important determinants of educational attainment, contribute to college application behavior and major intentions. Specifically, we examine the number of college applications and students' preferred college major at the end of high school. Our structural model is able to explain more of the gaps between high SES white females and other socio-demographic groups relative to reduced form specifications. Using an outcome of four or more applications and accounting for the two latent factors, we find that high SES blacks are more likely to reach this application threshold than their advantaged white female peers. As such, this paper is the first to dynamically model the link between latent cognitive and non-cognitive skills and college application behavior.

Recently, a growing body of literature has estimated the dynamic process of college major choice. Existing research pays particular attention to the role of cognitive skills and preferences. Measures of cognitive ability have been shown to vary significantly across majors and contribute meaningfully to major choice.[6] Meanwhile,

[6] Paglin and Rufolo (1990) proxy for ability using the average math test scores of recent graduates within major, and are able to explain 79-82% of the variance in observed wages. Weinberger (1999) critiqued Paglin and Rufolo's conclusion that the relative math content of majors chosen by men and women could explain all of the gender wage gap, and use the 1985 sample of the Survey of Recent College Graduates to show a persistent gender wage gap of 9%. Likewise, Arcidiacono (2004) documents significant heterogeneity in SAT verbal and quantitative scores across majors in The National Longitudinal Study of the High School Class of 1972 (NLS72), and uses a dynamic model of college and major choice to show that sorting by, and learning over cognitive ability are important factors.

several studies emphasize the role of preferences over cognitive skills, particularly for explaining gender gaps in major choice. In Turner and Bowen (1999), differences in skills account for less than half of the gender gap in major choice, and are dominated by gender-specific major preferences.[7] Gemici and Wiswall (2014) estimate a dynamic model of human capital investment in years of schooling, field of schooling, and years of experience in the labor market. They likewise find preferences are the primary driver of the gender gap in major choice, even after controlling for these additional factors. One possible explanation for the regular finding that preferences largely account for students' major choices is that studies have yet to rigorously account for the effects of both cognitive and non-cognitive factors. We address this shortcoming by considering the contributions of both cognitive and non-cognitive factors in our model. Finally, the structural model permits us to separately distinguish the effects of each factor at every educational stage. This allows us to better understand the relative importance of high cognitive vs. non-cognitive ability for a number of schooling outcomes.

The remainder of the paper proceeds as follows. Section 4.1 outlines the sample construction using NCERDC data, as well as our use of nationally representative longitudinal surveys. Section 4.2 summarizes the extent of disparities across student subgroups and presents a series of linear probability models to provide suggestive evidence on the contribution of each of our three potential explanations for the observed gaps. We shift from OLS evidence to a description of the factor model in Section 4.3. Section 4.4 presents parameter estimates from the structural model for comparison with previous results. We conclude with a discussion of the implications of our findings in Section 4.5.

[7] Zafar (2013) reaches largely the same conclusion in a study of Northwestern students, that the gender gap is explained primarily by preferences, and that subjective earnings were not a significant determinant of major choice.

4.1 Data

We use administrative data from the state of North Carolina as the basis of our estimation sample. The resulting student population is significantly larger than samples constructed from longitudinal surveys such as the National Educational Longitudinal Study (NELS) (Jacob, 2002) and the NLSY97 (Aucejo, 2015). To ensure the external validity of documented academic gaps, we construct the same eight granular student subgroups along race, gender, and SES using the nationally representative NSLY97.

4.1.1 NCERDC

The primary data source is the NCERDC. This ongoing partnership between the Duke University Center for Child and Family Policy and the North Carolina Department of Public Instruction provides administrative data on the state's public school students going back to the mid-1990s. Data elements include students' standardized test scores, attendance and delinquency records, course-taking histories in high school, graduation, dropout, SAT test scores, and college application behavior. Our main dataset tracks 9th grade students in 2006 and 2007 for four years to the presumed end of their high school careers.[8]

We begin with the roster of 9th grade students over these two cohorts and merge in covariates and outcome measures in succession. Covariates broadly fall into three categories: student socio-demographic information, high school characteristics, and measures of cognitive and non-cognitive skills. We discuss each of these covariates and outcomes in turn.

First, individual characteristics are primarily taken from comprehensive student

[8] High school continuation measures are conditional on students achieving completion outcomes within this four-year window.

roster files on 9th graders.[9] Individuals' socio-demographic variables include ethnicity, sex, and free and reduced-price lunch status. White and African American students comprise nearly 90% of 9th grade public school students in North Carolina during the period of our study. We focus on the two largest racial groups in this study, although the conceptual framework can be readily extended to examine educational outcomes for Hispanic, American Indian, and Asian student populations. Similar to ethnicity and gender, we also use a dichotomous definition of socioeconomic status by identifying students who are eligible for the free and reduced lunch program as "low SES." Children qualify for free lunch under the federally subsidized National School Lunch program if they reside in a household with incomes at or below 130% of the federal poverty level, and they qualify for a reduced-price lunch if household income is between 130% and 185% of the poverty line.[10]

Then we interact the three binary categories of female, African American, and low SES to produce eight mutually exclusive student subgroups that reflect nuanced relationships between ethnicity and family background. Economically disadvantaged African Americans may face systemically different sets of challenges and constraints than their white peers from equally under-privileged households. Our analyses examines the extent to which these differences can manifest in varying educational attainment and postsecondary choices.

The second set of variables comprises high school attributes that come from the Public School Universe. These files provide one record per school on its type, locale, counts of free or reduced price lunch eligible students and student counts by race, ethnicity, and grade. We include an indicator for magnet schools. Locale

[9] The "masterbuild" files, which comprise the basis of the sample, contain one record per student, socio-demographic variables, and corresponding End-of-Grade and End-of-Course scores. An encrypted student ID links all other longitudinal NCERDC files to the base sample.

[10] In North Carolina, children from households that are eligible for receiving Supplemental Nutrition Assistance Program (SNAP) or Cash Assistance (CA) are automatically eligible for free meals.

coding is based on the school building's mailing address and differentiates between city, suburban, and rural. We create indicators for schools located in cities with a population of at least 250,000 or a rural region defined as locations with fewer than 2,500 people or those in ZIP codes designated as rural by the Census Bureau. Lastly, we construct the shares of 9th grade students in the school who are white or African African using student counts by race and grade.

Third, we compile cognitive and non-cognitive measures from multiple source files. We measure cognitive skills using students' End-of-Grade (EOG) math and reading scores from 8th grade, the last year for which statewide standardized testing occurs at the grade level, and End-of-Course (EOC) reading grades from 9th grade.[11] We normalize scores using the population of all test-taking students in a given grade and year to ensure comparability across schools and time. A particular advantage of these statewide data relative to longitudinal surveys such as NLSY79 and NLSY97 is that cognitive measures are taken for the same cohort to ensure comparability, instead of across schooling levels and ages.

Non-cognitive measures include GPA, absences, tardies, and suspensions taken in 9th grade, and time spent on homework during 8th grade. We construct grade-specific weighted GPAs using credits earned and course-level grades from student transcripts. Honors and AP courses, for instance, are allocated more credits than regular coursework. As a result the 9th grade weighted GPA reflects both course levels and performance. In addition, we use a rich set of discipline measures taken at the end of 9th grade.[12] They distinguish between excused and unexcused absences, as well as in- and out-of-school suspensions.[13] Lastly, a data item from 8th grade

[11] We only use 9th grade reading scores because the same English standardized test is taken by all 9th grade students, while there is no corresponding prescribed course for 9th grade math.

[12] The choice of 9th grade discipline data minimizes students' previous interactions with school administrators to moderate the influence of disciplinary histories.

[13] Research shows consistent differences in disciplinary sanctions by race and gender, although

EOG exams provides students' average weekly homework load. We create indicators for no homework ever assigned, less than 1 hour of homework per week, between 1 and 3 hours of homework/week, more than 3 hours of homework per week, and homework assigned but not actually done.[14]

Finally, the outcomes we use in this paper cover several dimensions of students' transition from secondary into higher education, including high school attainment, attending college, number of college applications, and intended major choice. We derive the first dependent variable on attainment from three sets of files: graduation, drop out, and school exit.[15] We merge 9th grade students to graduation files at the end of four years, as well as dropout and exit information from the four years spanning grades 9-12. Specifically, we match a 9th grader in 2006 to 2009 graduation files and 2006-2009 dropout and school exit data to ensure a comprehensive view of their manner of exit.[16] We construct a variable on the student's last observed grade during the four-year observation window. The final sample excludes observations that are missing this grade variable as well as those who exit for reasons that are exempt from graduation rate calculations. These primarily comprise students who exit from the public school system and therefore can no longer observed via graduation files.[17]

evidence on the direct causal influence of bias is limited (Gregory, Skiba, and Noguera, 2010; Kinsler, 2011).

[14] Our choice of non-cognitive measures are consistent with variables used in previous studies on college enrollment. Jacob (2002) proxies for non-cognitive skills with middle school grades, hours spent on homework during 9th grade, a composite measure of disciplinary incidents, and an indicator for elementary school retention. Aucejo (2015) uses 8th grade GPA, retentions and suspensions between grades 1 and 8, involvement in fights, and precocious sex.

[15] School exit files provide the rationale for leaving, which includes graduation and dropping out, along with other reasons such as transfers to home schooling and private schools.

[16] We cross-check graduation and dropout outcomes with the rationale given in school exit data. For instance, we code students as having graduated or dropped out if they are marked as such in either the graduation/dropout or exit files.

[17] The NCERDC exempts students who exit for the following reasons: leaving for another school in the system, leaving to another system, leaving for a different state, death, visiting student status, transfers to within-state private schools and home schools, unconfirmed transfers including detention centers, and students who do not belong in the cohort. All dropped observations comprise

Major choice information is recorded in SAT testing data given in the NCERDC.[18] In addition to students' performance on SAT math and verbal tests, the accompanying SAT Questionnaire asks students about their first choice college major. We codify student responses into five broad categories: STEM, Arts/Humanities/Education/Other, Health, Business/Law/Social Sciences, and Undecided.[19] Since answers are recorded during the latest administration of the SAT a student takes, and the vast majority of students do so during their junior or senior year, the variable in effect describes major intentions during the second half of high school.[20]

SAT files in the NCERDC also enable us to approximate students' college application portfolios. Beginning in 2010, these files include the full list of institutions and scholarships that ever received an official score report from each student. We match these institutional codes to the Integrated Postsecondary Education Data System (IPEDS), providing rich college attributes such as classification and selectivity. By adding up the number of colleges and universities receiving official SAT score reports, we can arrive at a close approximation of how many institutions the student applied to, and the characteristics associated with each.

The final sample spans both the 2006 and 2007 9th grade cohorts. We keep

slightly less than 20% of the sample.

[18] College Board, the non-profit organization developing and administrating the SAT test, shared student-level data with the North Carolina Department of Public Instruction annually from 2009-2014.

[19] The STEM category includes agriculture, biological and biomedical sciences, computer and information sciences, engineering and engineering technologies, mathematics, statistics, and physical sciences. Arts/Humanities/Education/Other includes English literature, cultural and gender studies, foreign languages, history, liberal arts, library science, philosophy, religious studies, visual and performing arts, education, as well as interdisciplinary or vocational fields such as construction, mechanic and repair technologies, natural resources and conservation, leisure and fitness, law enforcement, firefighting, and theology. The health categories captures all health-related professions. Business/Law/Social Sciences includes business management, marketing, communication, journalism, legal professions, psychology, public administration, and other social sciences.

[20] The outcome is conditional on students taking the SAT. The college preferences of students who do not intend to enroll in a 4-year institution or took the ACT in place of the SAT are therefore not reflected in this sample. Using the 2009 cohort, 47% of high school seniors in the NCERDC database took the SAT at least once.

student records with non-missing high school characteristics, 8th grade EOG scores, 9th grade EOC reading scores, hours spent on homework, 9th grade absences, tardies, suspensions, and GPA. Lastly, while the final dataset retains all ethnic groups, we restrict our analyses to a subset of white and African American students.

4.1.2 NLSY97

We supplement our NCERDC analyses with data from the NLSY97 to ensure that the academic disparities we document are not specific to North Carolina or the period we cover. 8,984 individuals between the ages of 12 to 18 were interviewed for the NLSY97 in 1997. The survey tracks them through the 2013-2014 survey wave, when respondents were 28 to 34 years old.

We construct student groups using socio-demographic information on race, gender, and socioeconomic status. Although no data element exists on free and reduced-price lunch status, we can approximate eligibility using constructed variables on the ratio of household income relative to the federal poverty line. All respondents with family incomes at or below 185% of the federal poverty guideline are denoted as low SES. The dependent variable derives from a constructed measure on the highest level of schooling attained. Students are marked as graduating from high school if they indicate schooling completion through at least 12th grade. The final sample of 4,936 respondents comprises all students who completed at least middle school.

4.2 Reduced form evidence

We begin by presenting reduced form evidence on the observed gaps in educational outcomes across race, gender, and socioeconomic groups. Then we corroborate evidence on the magnitude of disparities in the North Carolina with a nationally representative sample from the NLSY97. We then shift from unconditional means to a series of linear probability models that regress student outcomes on high school characteristics, cognitive skills, and non-cognitive skills.

4.2.1 Educational disparities by subgroup

Table 4.1 describes outcomes and covariates for the eight subgroups defined along gender, racial, and socioeconomic dimensions using NCERDC data. The first four columns show summary statistics for high SES students while the final four show the same for more disadvantaged populations. Our reference category in Column 1 is high SES white females.

We juxtapose three sets of outcomes. The first is educational attainment as measured by whether the student completes 10th and 11th grades, and graduates in 12th grade.[21] The second proxies for the number of college applications using the cumulative number of SAT score reports. We use an indicator for whether a student applied to at least four institutions in subsequent reduced form models. The behavioral implications of using four applications as a cutoff derive from the pricing structure of SAT score reports. Students receive the first four reports free of charge and must pay a unit-fee for each subsequent report. As the marginal report is costless up to a point, one would expect students with college intentions to send no fewer than four applications. In addition to evaluating the quantity of applications, we also approximate for quality using the distribution of SAT scores among undergraduate populations at receiving institutions. We define very selective institutions as those with 25 percentile SAT math scores at 600 or higher. Only Duke, Davidson, and UNC Chapel Hill qualify in the state of North Carolina. Lastly, we characterize preferred major choice among those taking the SAT by classifying students' top major into five categories: STEM, Arts/Humanities/Education/Other, Health, Business/Law/Social Sciences, and Undeclared.

[21] These attainment outcomes are conditional on the student being observed in a North Carolina public high school in 9th grade and observed as graduating, dropping out of, or remaining in the public system within four years.

We observe sizable attainment and postsecondary intentions disparities along gender, racial, and socioeconomic dimensions. 94% of advantaged white females graduate from high school compared to 73% of disadvantaged white males and 72% of disadvantaged black males.[22] Two white subgroups document the highest and lowest outcomes in college applications: advantaged white females send 3.6 applications on average to postsecondary institutions, compared to 2.5 among disadvantaged white males. The patterns governing preferred college majors fall largely along gender lines. Between 23-28% of males, regardless of race or SES background, choose STEM majors compared to 7-10% of females. In contrast, health-related majors are strikingly popular with females relative to their male peers.

Covariates span high school attributes, cognitive measures in EOG and EOC standardized scores, and non-cognitive measures including 9th grade GPA, disciplinary history, and homework effort. Interestingly, African American students were more likely on average to enroll in a magnet school. A greater share of African American students attend high schools in urban areas with 250,000 or more residents and institutions with a higher share of black students in 9th grade. African American students also demonstrate significantly lower achievement scores in both math and reading. Relative to high SES white students, who average 0.5 standard deviations above the statewide average in 8th grade math, high SES black students score 0.6 standard deviations lower than their white peers, while disadvantaged black students score 1 standard deviation lower.[23] Generally, all white student subgroups perform above the statewide average in math and science, while the opposite holds for African American students across gender and SES lines.[24]. On the non-cognitive

[22] The 4-year graduation rate across the sample is 86%, higher than the national graduation rate for 2012-2013 of 81%. One possible reason for this discrepancy is the exclusion of students who have missing covariates.

[23] The magnitude of these black-white achievement gaps is consistent with the 0.8 standard deviation gap found by Clotfelter, Ladd, and Vigdor (2009) among eighth graders.

[24] An exception is advantaged black females, who score slightly below average in math and above

side, African American students exhibit a similar GPA disparity and are more likely to accumulate tardies and suspensions.[25]

Several observations emerge from these unconditional averages. The first is the interaction between race and socioeconomic background. Substituting minority status for economic disadvantage can obscure significant disparities across the economic spectrum within each racial or gender group. Relatively well-off black females, for example, have 12th grade graduation rates rivaling their white female counterparts, while under-privileged black females fare significantly worse. Both groups attend 9th grade classrooms of similar racial compositions, but 23% of advantaged black female students enroll in an urban high school, compared to 15% of disadvantaged black female students. The division of race and SES thus enables us to consider a fuller spectrum of family circumstances and the means through which they may influence students' behavior and achievements.

Second, females tend to exhibit higher educational attainment and perform better in academic dimensions that require sustained effort than males in the same race and SES group. Females are 2 to 11 percentage points more likely to graduate high school in 12th grade relative to comparable male peers. They also send more score reports on average to prospective colleges. These advantages are consistent with patterns of achievements that have already emerged by 8th and 9th grades. Female students are ahead across all reading assessments and log similar scores in EOG math. One possible input into these results is hard work, as female students systematically report spending more time on homework. In keeping with higher effort, female students also achieve 9th grade GPAs that are approximately 0.3 points higher than those of their male peers. Meanwhile, stark differences are apparent along gender lines when

average in reading.

[25] Limited information on the nature of disciplinary infraction renders it difficult to determine whether racial disparities in accumulated sanctions are due to severity of actual behavior or some form of discrimination or bias.

Table 4.1: NCERDC: Summary statistics

	High SES				Low SES			
	Female		Male		Female		Male	
	White	Black	White	Black	White	Black	White	Black
Completed 10th grade	98.6%	98.2%	97.6%	95.6%	92.6%	94.4%	89.3%	89.5%
Completed 11th grade	96.3%	95.6%	93.9%	90.6%	84.3%	87.3%	78.4%	77.8%
Graduated 12th grade	94.2%	92.7%	90.7%	86.1%	79.2%	83.1%	72.5%	72.4%
College applications[a]								
Sent any SAT scores	84.9%	83.2%	82.8%	77.2%	73.2%	77.1%	68.1%	71.5%
Sent to 4+ institutions	55.1%	53.0%	53.1%	52.9%	40.6%	51.7%	36.8%	46.5%
Number of institutions	3.637	3.655	3.528	3.603	2.856	3.420	2.543	3.153
No. of very selective institutions	0.565	0.498	0.595	0.475	0.317	0.261	0.331	0.278
College major intentions[b]								
Reported major intention	62.6%	64.0%	53.4%	52.7%	27.8%	40.2%	20.3%	31.0%
STEM	10.2%	10.3%	28.2%	27.0%	8.5%	6.9%	25.5%	23.4%
Arts/Humanities/Edu./Other	26.8%	17.6%	22.0%	19.5%	28.3%	15.9%	25.8%	20.8%
Health	28.6%	32.6%	10.9%	10.8%	30.0%	35.2%	12.1%	9.5%
Business/Law/Social Sciences	16.7%	27.2%	16.2%	26.4%	16.0%	26.0%	13.5%	24.3%
Undeclared major	17.6%	12.3%	22.8%	16.3%	17.4%	15.9%	23.1%	22.0%
Skill proxies: Cognitive								
8th grade math EOG z-scores	0.539	−0.038	0.546	−0.154	0.001	−0.419	0.022	−0.501
8th grade reading EOG z-scores	0.573	0.064	0.452	−0.118	0.087	−0.376	0.004	−0.506
9th grade reading EOC z-scores	0.685	0.200	0.453	−0.071	0.162	−0.247	−0.077	−0.475
Skill proxies: Non-cognitive								
GPA	3.157	2.825	2.888	2.476	2.609	2.453	2.390	2.183
Excused absences	2.519	1.610	2.322	1.527	4.326	2.682	4.082	2.690
Unexcused absences	2.613	2.537	2.689	2.603	5.070	4.668	5.192	4.728
Tardies	1.053	1.497	0.904	1.366	1.323	1.841	1.145	1.840
In-school suspensions	0.073	0.153	0.141	0.232	0.260	0.324	0.420	0.511
Out-of-school suspensions	0.110	0.329	0.228	0.548	0.395	0.834	0.712	1.185
No homework ever assigned	0.5%	0.8%	1.2%	1.6%	1.3%	1.7%	2.6%	2.8%
Less than 1 hour of /week	20.7%	23.2%	27.3%	28.1%	28.6%	31.2%	34.5%	36.4%
Between 1 and 3 hours of /week	46.8%	52.7%	44.8%	48.9%	47.9%	49.9%	42.8%	45.6%
More than 3 hours of /week	31.6%	23.1%	25.3%	20.1%	21.4%	16.6%	16.9%	13.6%
High school characteristics								
Magnet high school	4.3%	11.3%	4.2%	10.9%	1.1%	8.2%	1.3%	7.4%
City with 250K+ residents	14.6%	23.1%	14.4%	22.3%	5.1%	14.6%	5.8%	14.7%
Rural area	53.9%	36.2%	53.7%	37.2%	62.1%	40.7%	61.6%	41.4%
Share of black students	22.2%	46.3%	22.3%	45.1%	20.2%	48.5%	20.9%	48.2%
Share of white students	67.6%	41.8%	67.6%	42.7%	70.2%	39.5%	69.5%	39.9%
Observations	33,251	5,462	33,035	5,125	10,763	15,510	10,351	13,269

Note: Proxies for non-cognitive skills are taken from a student's records in 9th grade, while proxies for their cognitive skills are taken from 8th grade records. The share of students by race are taken from the characteristics of the student's school in 9th grade.

[a] Data on students' SAT score submissions to colleges are available only for the 2007 cohort, hence these data are reported relative to the 2007 cohort only.

[b] Data on students' college major intentions are available only for students that report an intended major on the SAT. Hence, the proportion of students in each major are defined relative to students that report a college major intention.

it comes to choice of majors. Female students repeatedly choose health, arts and humanities, and education fields while male students prefer STEM, business, law, and social sciences.

The final notable pattern is the under-performance of poorer white students across several outcomes. Poorer white males rival black males in low high school graduation rates. Disadvantaged white students also apply to the fewest postsecondary institutions. This is particularly striking given that disadvantaged white females score 0.4 standard deviations higher than disadvantaged black females in both 8th grade reading and math, while the corresponding advantage is closer to 0.5 standard deviations among low-SES male students. Under-privileged white students also perform better in 9th grade GPA and accumulate fewer tardies and suspensions relative to African American students in the same socioeconomic bracket. Yet these advantages in academic and disciplinary track records do not translate to superior outcomes by the end of high school.

To ensure that these patterns are not specific to the North Carolina context, we turn to the national sample available in NLSY97. Table 4.2 distinguishes between the same eight student groups using gender, race, and predicted National School Lunch Program eligibility constructed from household income relative to the federal poverty line. The graduation rate, defined by whether the student ever completed 12th grade, is 82% for the full sample. Subgroups maintain consistent achievement patterns relative to each other as in North Carolina, with the exception of poor black males. Only 56% graduate from high school in the national sample, compared to 72% in North Carolina. The higher persistence of the latter may be partially explained by the positively selected sample containing non-missing data. A notable observation is that poorer white females exhibit lower graduation rates than poorer black females, despite their relative economic advantage, which is consistent with patterns in our NCERDC sample.

Table 4.2: NLSY97: Summary statistics

	High SES				**Low SES**			
	Female		Male		Female		Male	
	White	Black	White	Black	White	Black	White	Black
Complete 12th grade	90.1%	87.3%	86.0%	82.0%	73.3%	75.0%	66.8%	56.4%
Average percent of poverty level	444.5%	355.8%	443.2%	359.4%	108.0%	82.6%	111.4%	79.6%
Observations	444	356	443	359	108	83	111	80

Note: Base sample comprises 4,936 respondents across the eight categories who completed at least 8th grade. A respondent has a high school degree if he or she reports completing the 12th grade. Outcomes are weights-adjusted.

4.2.2 Regressions

In this section we examine the extent to which high school characteristics and proxies for both cognitive and non-cognitive skills can explain cross-subgroup disparities in educational attainment, college application behavior, and intended major choice. We begin with graduation in 12th grade, before switching to preferred major and the dichotomous outcome of whether the individual submits at least four college applications. Collectively, these outcomes reflect multiple dimensions of the transition from secondary into higher education. Segmenting the student population by gender, race, and socioeconomic status informs our understanding of how these specific groups fare during a stage when many students are at risk of exiting the educational system.

Tables 4.3 through 4.5 present results associated with each outcome using a set of linear probability models. We increasingly augment the base model with additional covariates. The leftmost column in Table 4.3 regresses 12th grade graduation on the eight subgroups and a cohort indicator, omitting the reference group of high SES white females. Disadvantaged black males are shown to have the largest graduation deficit at 21.8 percentage points, while under-privileged white males follow closely behind with 21.7 percentage points. The addition of high school characteristics spanning magnet school status, urbanicity, and racial composition of classrooms in the second column has little impact across coefficients.

The third specification incorporates measures for non-cognitive skills, including 9th grade GPA, homework effort exerted in 8th grade, absences, tardies, and suspensions. Their inclusion erases the gap between high SES white females and all black subgroups with the exception of low SES males, for whom approximately three-quarters of the disparity is accounted for. Similar yet smaller effects are observed when cognitive measures are included in lieu of socio-emotional proxies. Among white students, both skill sets exhibit lower explanatory power, owing to the relatively higher performance of students on standardized test scores (4.1).

When both cognitive and non-cognitive measures are included in Column 5, black females are shown to have higher rates of high school graduation, well-off black males are on equal footing with privileged white female peers, and only 15% of the original deficit for under-privileged black males remains. Coefficients across white student groups are reduced by 60%. Thus emerges an educational attainment narrative that bifurcates by race. If efforts to raise the cognitive and non-cognitive measures of black students on par with their advantaged white female peers are successful, we would expect the majority to have at least the same levels of completion. However, deficits in yet unobserved measures are setting back the progress of disadvantaged whites in particular.

Next we turn to students' preferred major choice as reported in College Board's SAT Questionnaire data. Table 4.4 shows a consistent preference for STEM majors among males of all socioeconomic backgrounds. STEM intention among high SES white males is 18.0 percentage points higher than high SES white females. The magnitude of this disparity becomes apparent when considering women's baseline STEM intention of less than 10%. As such, men are more than twice as likely to express STEM major intentions than women. These differences remain robust and are sometimes even exacerbated when controlling for cognitive and non-cognitive skills. Measures such as EOG math and 9th grade GPA are positively correlated with

Table 4.3: Linear probability model: Complete 12th grade

	(1)	(2)	(3)	(4)	(5)
High SES					
Black female	−0.014***	−0.013**	0.014***	0.039***	0.024***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
White male	−0.035***	−0.035***	−0.012***	−0.023***	−0.014***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Black male	−0.081***	−0.080***	−0.010	0.002	0.004
	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
Low SES					
White female	−0.150***	−0.152***	−0.067***	−0.099***	−0.060***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Black female	−0.111***	−0.111***	−0.005	−0.001	0.021***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.004)
White male	−0.217***	−0.218***	−0.098***	−0.146***	−0.088***
	(0.007)	(0.007)	(0.005)	(0.006)	(0.005)
Black male	−0.218***	−0.218***	−0.067***	−0.079***	−0.034***
	(0.009)	(0.008)	(0.006)	(0.006)	(0.005)
Additional controls					
HS characteristics[a]		Y	Y	Y	Y
Skill proxies: Non-cognitive[b]			Y		Y
Skill proxies: Cognitive [c]				Y	Y
Observations	126,766	126,766	126,766	126,766	126,766
R^2	0.054	0.055	0.218	0.146	0.243

Note: Sample includes all students that completed at least 9th grade. Indicators for race/gender/SES categories are defined relative to high-SES white females.

[a] High school characteristics are taken from a student's 9th grade institution, and include indicators for a magnet high school and being located in a city with 250K+ residents or a rural area, and continuous measures for the share of 9th grade students that are black and white.

[b] Non-cognitive skill proxies include a number of measures from a student's 9th grade records, including: cumulative GPA, number of excused and unexcused absences, number of in school and out of school suspensions, number of tardies, and a categorical variable for a student's amount of time spent on homework during 8th grade.

[c] Cognitive skill proxies include a student's 8th grade End-of-Grade scores in math and reading, normalized relative to statewide test scores for the year a student took a given 8th grade EOG test, along with an indicator for whether a student was identified as academically and intellectually gifted in 9th grade.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

STEM participation. Female students have higher GPAs and sometimes higher math standardized scores. Thus elevating men's performance to females' levels results in a gap of greater magnitude.

In contrast to gender gaps, within-gender racial and SES differences are largely erased when considering our full set of covariates. Low SES white and black females were 1.7 and 3.3 percentage points less likely to pursue STEM, respectively. These disparities are no longer significant when either cognitive or non-cognitive measures are included in the model. In conclusion, our models are sufficient for explaining racial or socioeconomic gaps in STEM choice, but leaves substantial variation unexplained in the case of gender differences.

Table 4.4: Linear probability model: Intend to major in STEM

	(1)	(2)	(3)	(4)	(5)
High SES					
Black female	0.000	−0.011*	0.021***	0.023***	0.027***
	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
White male	0.180***	0.180***	0.190***	0.171***	0.176***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Black male	0.167***	0.157***	0.201***	0.192***	0.200***
	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)
Low SES					
White female	−0.017***	−0.017***	0.002	−0.001	0.004
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Black female	−0.033***	−0.045***	0.001	−0.001	0.007
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
White male	0.154***	0.154***	0.181***	0.161***	0.169***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Black male	0.131***	0.120***	0.175***	0.165***	0.175***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Additional controls					
HS characteristics[a]		Y	Y	Y	Y
Skill proxies: Non-cognitive [b]			Y		Y
Skill proxies: Cognitive[c]				Y	Y
Observations	60,078	60,078	60,078	60,078	60,078
R^2	0.057	0.057	0.071	0.073	0.077

Note: Sample includes all students that completed at least 9th grade. Indicators for race/gender/SES categories are defined relative to high-SES white females.

[a] High school characteristics are taken from a student's 9th grade institution, and include indicators for a magnet high school and being located in a city with 250K+ residents or a rural area, and continuous measures for the share of 9th grade students that are black and white.

[b] Non-cognitive skill proxies include a number of measures from a student's 9th grade records, including: cumulative GPA, number of excused and unexcused absences, number of in school and out of school suspensions, number of tardies, and a categorical variable for a student's amount of time spent on homework during 8th grade.

[c] Cognitive skill proxies include a student's 8th grade End-of-Grade scores in math and reading, normalized relative to statewide test scores for the year a student took a given 8th grade EOG test, along with an indicator for whether a student was identified as academically and intellectually gifted in 9th grade.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.5 examines the effect of cognitive and non-cognitive factors on whether the student applies to at least four postsecondary institutions. Similar to results on high school graduation, disadvantaged whites and black males document the largest disparities relative to advantaged white females. Low SES white males and females are on average 18.3 and 14.5 percentage points less likely to send more than four applications, a significant gap given the baseline 55% of high SES white females that do so. The full model reduces the coefficients by at least one-half, with the preponderance of influence coming from the inclusion of non-cognitive measures. Despite the rich set of control variables, we cannot account for some aspects of disadvantaged white students' behaviors. In contrast, conditioning on either set of measures is sufficient to compensate for the full application deficit for every African American subgroup.

One shortcoming of the present analyses is that we are unable to establish the relative contribution of cognitive vs. non-cognitive skills. In the ensuing section, we construct a structural factor model to more precisely define the two underlying skills and quantify the influence of each.

4.3 Factor model

The differences we observe between students from different demographic and socioeconomic groups, both in descriptive statistics and reduced form evidence, suggest skills may play a role in explaining why their outcomes vary in the transition from secondary to postsecondary education and college major choice. However, the analysis we have presented thus far suffers from two shortcomings with regard to the proxies we use for skills, which are key to explaining the gaps in outcomes we observe by race, gender, and socioeconomic status. First, these proxies measure the underlying skills of interest with error, and hence the parameters we report are likely biased estimates of the effects of skills on student outcomes. Second, our measures of non-cognitive skills, particularly in the case of GPA, are jointly determined by both cognitive and non-cognitive skills. Hence we are unable, by construction, to interpret

Table 4.5: Linear probability model: Apply to at least four institutions

	(1)	(2)	(3)	(4)	(5)
High SES					
Black female	−0.021	−0.041***	0.040***	0.043***	0.053***
	(0.014)	(0.015)	(0.015)	(0.014)	(0.014)
White male	−0.020***	−0.022***	0.003	−0.023***	−0.006
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Black female	−0.022	−0.041***	0.077***	0.056***	0.085***
	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)
Low SES					
White female	−0.145***	−0.134***	−0.077***	−0.095***	−0.072***
	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)
Black female	−0.034**	−0.047***	0.082***	0.071***	0.099***
	(0.014)	(0.015)	(0.015)	(0.014)	(0.014)
White male	−0.183***	−0.172***	−0.094***	−0.136***	−0.096***
	(0.018)	(0.018)	(0.017)	(0.018)	(0.017)
Black male	−0.086***	−0.098***	0.062***	0.031**	0.076***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Additional controls					
HS characteristics[a]		Y	Y	Y	Y
Skill proxies: Non-cognitive[b]			Y		Y
Skill proxies: Cognitive[c]				Y	Y
Observations	32,411	32,411	32,411	32,411	32,411
R^2	0.008	0.014	0.064	0.052	0.071

Note: Sample includes all students that completed 12th grade, indicated on their SAT test that they intend to go to college, and are from the 2007 cohort, for which data on students' test score submissions was available. Indicators for race/gender/SES categories are defined relative to high-SES white females.

[a] High school characteristics are taken from a student's 9th grade institution, and include indicators for a magnet high school and being located in a city with 250K+ residents or a rural area, and continuous measures for the share of 9th grade students that are black and white.

[b] Non-cognitive skill proxies include a number of measures from a student's 9th grade records, including: cumulative GPA, number of excused and unexcused absences, number of in school and out of school suspensions, number of tardies, and a categorical variable for a student's amount of time spent on homework during 8th grade.

[c] Cognitive skill proxies include a student's 8th grade End-of-Grade scores in math and reading, normalized relative to statewide test scores for the year a student took a given 8th grade EOG test, along with an indicator for whether a student was identified as academically and intellectually gifted in 9th grade.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

what the relative importance of either skill is in determining student outcomes, or how the role skills play vary by race, gender, or socioeconomic status.

In this section we outline a structural factor model that overcomes these issues, and allows us to identify students' underlying skills from our noisy proxies for cognitive and non-cognitive skills. We are then able to use these underlying skills to examine the relative roles they play in determining how students progress through and graduate from high school, determine whether to apply for and attend college, and what major to pursue in college. Finally, we allow students' underlying skill distributions to vary by student race, gender, and socioeconomic status. This heterogeneity in skills allows us, in turn, to leverage detailed information from the NCERDC data on student demographics to explore how the effects of these skills vary for each group.

We start by first developing the measurement system for these cognitive and non-cognitive factors and how they vary by demographic characteristics. Second, we take these factor measures and incorporate them into a sequential model of secondary educational attainment, the decision to apply for and attend college, and the decision of what initial major to pursue in college. Finally, we specify the likelihood for this model and detail how we modify it for estimation.

4.3.1 Measurement system

Each student has a given vector of latent abilities $\boldsymbol{\theta} = \{\theta^C, \theta^{NC}\}$, where the elements denote cognitive and non-cognitive factors, respectively. This vector of abilities is unobserved to the econometrician. However, we assume that individuals know their levels of each factor and both latent traits are fixed by the time individuals make the educational choices relevant to our study: the transition between secondary and postsecondary education. Identifying the distribution of these factors is key because they are relevant for students' choices in each stage of the model: from choosing whether to continue to the next grade in high school, to attending college and what major to pursue. We also make the simplifying assumption that these latent abilities are independent random variables with the following distribution:[26]

[26] Carneiro, Hansen, and Heckman (2003) and Cunha and Heckman (2008) extend a similar factor model to allow for correlated factors. However, we maintain the assumption that students' cognitive and non-cognitive skills are independent, given empirical evidence shows that score correlations

$$\begin{pmatrix} \theta^C \\ \theta^{NC} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \tag{4.1}$$

One of the important features of our approach is that we allow the distribution of students' latent skills to vary by their race, gender, and socioeconomic status. We implement this in the model by allowing students' vector of skills to depend both on the above individual-specific independent random variables, and a linear interaction with indicators for their demographic characteristics Z_i. In particular, a student's cognitive and non-cognitive factors, $\left\{F_i^C, F_i^{NC}\right\}$ are specified relative to those of high-SES, white women:[27]

$$\begin{aligned} F_i^C &= \theta_i^C + \alpha_1^C \mathbf{Z_i} + \alpha_2^C \mathbf{Z_i} * \theta_i^C \\ F_i^{NC} &= \theta_i^{NC} + \alpha_1^{NC} \mathbf{Z_i} + \alpha_2^{NC} \mathbf{Z_i} * \theta_i^{NC} \end{aligned} \tag{4.2}$$

$$Z_i = \{male_i, black_i, lowSES_i, male_i * black_i, male_i * lowSES,$$
$$black_i * lowSES_i, male_i * black_i * lowSES_i\}$$

Since we are unable to directly observe any of the components of Equation 4.2, other than Z_i, we use a measurement system that allows us to identify $\left\{F_i^C, F_i^{NC}\right\}$ and $\boldsymbol{\theta_i}$ from our set of proxy measures for cognitive and non-cognitive skills. We also follow Aucejo (2015) in allowing for our proxy measures for non-cognitive skills to be jointly determined by both the latent cognitive and non-cognitive skills F_i^C and F_i^{NC}, but constrain our proxies for cognitive skills to only be a function of latent cognitive skills. With this specification, there are G and H measures of cognitive and non-cognitive skills, such that G_j denotes the jth cognitive test taken and H_k denotes the kth non-cognitive measure for individual i:

within cognitive or non-cognitive tests are stronger than across-test correlations.

[27] In practice we define a student as "high-SES" if they do not qualify for either the free, or reduced-price lunch program.

$$
\begin{aligned}
G_{ij} &= \gamma_{1j}^{C} + \gamma_{2j}^{C} F_i^{C} + \gamma_{3j}^{C} \mathbf{X_i} + \epsilon_{ij}^{G} \\
H_{ik} &= \gamma_{1k}^{CNC} + \gamma_{2k}^{CNC} F_i^{C} + \gamma_{3k}^{CNC} F_i^{NC} + \gamma_{4k}^{CNC} \mathbf{X_i} + \epsilon_{ik}^{H}
\end{aligned}
\tag{4.3}
$$

Equation 4.3 is asymmetric in the sense that the latent cognitive factor is a determinant of all chosen measures of non-cognitive ability, but the latent non-cognitive factor is not an input into cognitive measures. Since we assume the proxy measures for non-cognitive skills H_k are jointly determined by both latent cognitive and non-cognitive skills, we allow for measures like high school GPA to be determined both by a student's academic ability and features like their ability to complete tasks on time, or personality traits like conscientiousness. However, this added flexibility in the model, along with our lack of "pure" measures of non-cognitive skills like a Rotter scale score, implies that "non-cognitive" skills are orthogonal to cognitive skills. In other words, "non-cognitive" skills are all other skills aside from cognitive skills.

We rely on cognitive and non-cognitive measures taken from no later than the early high school years, since we assume that students' skills are fixed at the time they make decisions relevant to their transition from secondary to postsecondary education.[28] Our three cognitive measures G_j include standardized test scores on the North Carolina 8th grade End-of-Grade (EOG) math and reading exams, and an indicator for whether a student was identified as academically and intellectually gifted (AIG).[29] The five non-cognitive measures H_k include 9th grade cumulative grade point average (GPA), the number of excused and unexcused absences, indicators for whether a student was suspended or tardy, and a categorical variable for the effort a student put into homework during 8th grade.

[28] In particular, since our sample is comprised of students that complete at least 9th grade, we take measures from no later than 9th grade.

[29] The 8th grade EOG math and reading exams are independently standardized relative to the statewide distribution of test scores for a given exam year. AIG is a designation conferred by a student's teacher starting as early as 3rd grade.

These measures may also be driven in part by other covariates. For example, discrimination could make black students more likely to be suspended from school. Hence, a measure that doesn't control for such discrimination - for example, with the proportion of black students in a students' high school - would incorrectly find that black students have lower non-cognitive skills. We capture these features by including in $\mathbf{X}$ a vector of covariates that affect the values of these measures, such as high school characteristics. Lastly, we assume that once we condition on these covariates, the idiosyncratic error terms are mutually independent and independent of the factor variables and $\mathbf{X}$, such that $\epsilon^{C,NC} \perp \left\{\theta^{C,NC}, \mathbf{X}\right\}$, $\epsilon^{C}_{ja} \perp \epsilon^{C}_{jb}$, and $\epsilon^{NC}_{ka} \perp \epsilon^{NC}_{kb}$ for all $a \neq b$. Furthermore, all ϵ are assumed to be normally distributed with mean 0 for continuous variables, and Logit distributed for binary variables.

Anderson and Rubin (1956), and Carneiro, Hansen, and Heckman (2003) establish that this sort of measurement system is non-parametrically identified under three conditions.[30] First, factors must be orthogonal to each other such that $\theta^C \perp \theta^{NC}$, and as we described earlier, our specification of the measurement system implies the non-cognitive factor is orthogonal to the cognitive factor by construction. Second, if we have M factors we must have at least $2M+1$ measures, $G+H$, in order to identify the factors. Since we use two factors in this model, $M = 2$, we require a combination of at least five cognitive and non-cognitive measures. Third, the underlying factor structure must exhibit the following pattern: first, least two measurements must depend exclusively on $\theta_1, ..., \theta_{M-1}$, and second, at least three measurements must depend on $\theta_1, ..., \theta_M$. Applied to our model, two or more measurements must depend solely on θ^C and three or more on $\left\{\theta^C, \theta^{NC}\right\}$. The intuition for this restriction is that, in order to identify the non-cognitive factor separately from the cognitive factor, we need at least one additional measure that depends on both factors H, instead of just

[30] This only holds up to one normalization, or when one of the loadings in each measurement equation is set to 1.

the cognitive factor G. Otherwise, we would be unable to distinguish the effect of the unobserved non-cognitive factor from the intercept parameters in the measures that only depend on the cognitive measures.

4.3.2 Student outcomes

Once we have established our measurement system for students' unobserved cognitive and non-cognitive factors, we can incorporate how these skills influence their decisions in the transition between secondary and postsecondary education. In this component we model three decisions related to this transition period: first, the sequential decision to continue in, and graduate from, high school; second, the decision of whether to apply to college conditional on finishing high school, and if so, how many colleges to apply to; and third, what initial major to pursue in college, conditional on deciding to apply to college.

In the first sequence of decisions, student i begins at the end of their 9th grade year and chooses whether to continue in high school by completing 10th grade. If the student chooses to drop out at any point they enter the labor market, which is an absorbing state.[31] However, if a student chooses to continue, they decide at the end of each grade s whether to complete that year of education, up to the point of graduating from high school S.[32] Furthermore, students only choose to continue in

[31] Some papers like Arcidiacono et al. (2014) allow students to drop- or stop-out of schooling, particularly at the college level. However we follow papers like Stange (2012) in restricting reentry for two reasons. First, modeling reentry is not the primary focus of this paper, as fewer than 8% of our sample drops out of high school at any point, and an even smaller fraction chooses to reenter. Second, allowing students to reenter schooling greatly increases the complexity and computational burden of the model.

[32] We also assume that all students that choose to complete 12th grade, $S = 12$ graduate, and hence do not model the stochastic process that determines whether students meet graduation requirements and are allowed to graduate. Instead, these individuals are modeled as choosing to drop out at the end of 11th grade. This may be an issue if a significant number of students are promoted through 11th grade, but not able to ultimately graduate. However, Table 4.1 shows this "bunching" of students completing 11th grade but not 12th does not appear to occur in our sample. Instead, we see proportional drops in the percentage of the original 9th grade sample that complete 10th, 11th, and 12th grade.

school at level s if the associated latent utility is sufficiently high:

$$V_{is}^{HS} = \beta_{0s}^{HS} + \beta_{1s}^{HS} F_i^C + \beta_{2s}^{HS} F_i^{NC} + \beta_{3s}^{HS} \mathbf{Z_i} + \beta_{4s}^{HS} \mathbf{X_i} + \eta_{is}^{HS} \tag{4.4}$$

$$D_{is}^{HS} = \begin{cases} 1, \text{if } V_{is}^{HS} \geqslant 0 \\ 0, \text{otherwise} \end{cases}$$

Finally, Equation 4.4 illustrates another assumption and benefit of using a factor model to examine students' choices. Note that the value term for pursuing a particular grade s in high school does not explicitly contain a future value term that captures the option value of remaining in school for another academic year. This is because, as is detailed in Cunha, Heckman, and Navarro (2007), factor models are able to capture the characteristics of a dynamic discrete models like this one, where students make a sequence of educational choices. Essentially, the unobserved factors $\{F_i^C, F_i^{NC}\}$ drive the dynamic selection process, whereby students drop out of high school in successive grades, or choose not to apply for and attend college.[33]

In the second stage, graduating high school students choose whether to attend college in one of M major categories, or to not attend college. We apply a multinomial choice framework and fit a model that is linear in the same set of two latent factors and a vector of observed controls as in Equation 4.4. Let V_{im} be the utility associated with individual i pursuing field $m \in \{1, ..., 5\}$ corresponding to STEM; arts, humanities, education, and other; health; business, law, and social sciences; or an undeclared major respectively, or not attending college $m = 0$. The corresponding system of latent utilities is:[34]

[33] Cunha, Heckman, and Navarro (2007) find that a factor model must meet two requirements to accurately model a dynamic discrete choice model. First, as we explicitly state in this section, the unobservable factors $\{F_i^C, F_i^{NC}\}$ must be separable from observables like X_i and Z_i. Second, we assume that the local optimum students find by comparing choices across proximal states is also the global optimum.

[34] We collect data from students on their major intentions from their SAT test reports, which are

$$\begin{aligned} V_{i0} &= \beta_{00} + \beta_{10}F_i^C + \beta_{20}F_i^{NC} + \beta_{30}\mathbf{Z_i} + \beta_{40}\mathbf{X_i} + \eta_{i0} \\ &\vdots \\ V_{iM} &= \beta_{0M} + \beta_{1M}F_i^C + \beta_{2M}F_i^{NC} + \beta_{3M}\mathbf{Z_i} + \beta_{4M}\mathbf{X_i} + \eta_{iM} \end{aligned} \tag{4.5}$$

Given this system of equations students choose whether to attend college in a major D_{ik} or to not attend college D_{i0} according to the following process:

$$D_{im} = \underset{m \in \{0,...,5\}}{\arg\max} \{V_{im}\} \tag{4.6}$$

Since we are using a multinomial choice framework, we normalize a students' choice probabilities for each of the five possible majors relative to the decision not to attend college $D_{i0} = 1$. Hence we can express the probability of student i choosing to pursue major $m \in \{1, ..., 5\}$, conditional on graduating from high school as:

$$\begin{aligned} &\Pr(D_{im} = 1 | F_i^C, F_i^{NC}, \mathbf{Z_i}, \mathbf{X_i}, D_{iS}^{HS} = 1) \\ &\qquad = \frac{\exp\left(\beta_{0m} + \beta_{1m}F_i^C + \beta_{2m}F_i^{NC} + \beta_{3m}\mathbf{Z_i} + \beta_{4m}\mathbf{X_i}\right)}{1 + \sum_l \exp\left(\beta_{0l} + \beta_{1l}F_i^C + \beta_{2l}F_i^{NC} + \beta_{3l}\mathbf{Z_i} + \beta_{4l}\mathbf{X_i}\right)} \end{aligned} \tag{4.7}$$

Lastly we incorporate students' college application decisions as a Logit model of whether students choose to apply to four or more colleges, which we proxy with them choosing to send SAT score reports to four or more colleges.[35] As with the components corresponding to high school continuation and major choice, we use the same vector of factors and controls.

typically taken in the last 1-2 years of a student's high school years. Since the model assumes that major choice decisions are made after high school graduation, we assume that students' decisions about college applications and initial majors are stable between their SAT test date and graduation.

[35] We also estimated a specification where we modeled the number of score reports sent as a Tobit, since students are able to choose to not send their score reports to any colleges. While these results were qualitatively similar to the Logit specification reported here, estimating the Tobit model was simply more computationally-burdensome.

$$\Pr(D_i^A | F_i^C, F_i^{NC}, \mathbf{Z_i}, \mathbf{X_i}, D_{iS}^{HS} = 1, D_{i0} = 0) \tag{4.8}$$
$$= \frac{\exp\left(\beta_0^A + \beta_1^A F_i^C + \beta_2^A F_i^{NC} + \beta_3^A \mathbf{Z_i} + \beta_4^A \mathbf{X_i}\right)}{1 + \exp\left(\beta_0^A + \beta_1^A F_i^C + \beta_2^A F_i^{NC} + \beta_3^A \mathbf{Z_i} + \beta_4^A \mathbf{X_i}\right)}$$

Finally, since the intensity of students' college major applications is dependent on choosing to attend college in the first place, this Logit is conditional on students choosing a college major $D_{i0} = 0$.

4.3.3 Estimation

Although our model contains a number of components corresponding to the factor measurement system and student outcomes, we can express the likelihood of any sequence of schooling and major intentions for our sample as a simple expression:

$$\mathcal{L} = \prod_{i=1}^{N} \Pr\left(\mathbf{D_i}, \mathbf{G_i}, \mathbf{H_i} | \mathbf{Z_i}, \mathbf{X_i}, \mathbf{D_{i,-1}}\right) \tag{4.9}$$

We can then expand this function to express the likelihood as a function of five components: first, the probability of choosing to progress through high school D_{is}; second, choosing a college major D_{im}; third, whether to apply to four or more colleges D_i^A; fourth, observing our measures of cognitive and non-cognitive skills G_{ij} and H_{ik}; and fifth, the probability density functions (PDFs) for cognitive and non-cognitive skills $f\left(\theta_i^C\right)$ and $f\left(\theta_i^{NC}\right)$.

$$\prod_{i=1}^{N} \iint_{\theta_i} \left[\prod_{s=1}^{S} \Pr\left(D_{is} | F_i^C, F_i^{NC}, \mathbf{Z_i}, \mathbf{X_i}, D_{i,s-1}\right) \prod_{m=0}^{M} \Pr\left(D_{im} | F_i^C, F_i^{NC}, \mathbf{Z_i}, \mathbf{X_i}, D_{iS} = 1\right) \right] \times$$

$$\Pr\left(D_i^A | F_i^C, F_i^{NC}, \mathbf{Z_i}, \mathbf{X_i}, D_{iS} = 1\right) \times \tag{4.10}$$

$$\left[\prod_{j=1}^{J} \Pr\left(G_{ij} | F_i^C, \mathbf{X_i}\right) \right] \left[\prod_{k=1}^{K} \Pr\left(H_{ik} | F^C, F_i^{NC}, \mathbf{X_i}\right) \right] \times$$

$$f\left(\theta_i^C\right) f\left(\theta_i^{NC}\right) d\theta_i^C d\theta_i^{NC}$$

Since the vector of latent abilities $\boldsymbol{\theta}$ is unobserved, we estimate our parameters by integrating over the individual component of the factors, $\left\{\theta_i^C, \theta_i^{NC}\right\}$, which follow the independent standard normal distributions we specify in Equation 4.1. However, since integrating over each unobserved factor is computationally burdensome, we follow Aguirregabiria and Mira (2007) by discretizing the distributions of the individual-specific components of the latent factors as Normal$(0, 1)$ distributions with $T = 21$ points of support.

4.4 Results

In this section we report parameter estimates corresponding to the three components of the factor model that capture student outcomes: first, a series of Logits for progressing through postsecondary education; second, a multinomial Logit that captures students' decision to pursue postsecondary education and their initial choice of major, conditional on graduating from high school; and third, a Logit that captures the intensity of their college application process, conditional on choosing to attend college.[36]

[36] We also report parameter estimates for the remaining two components of the model that comprise the measurement system for students' unobserved cognitive and non-cognitive skills in Appendix B: first, parameter estimates for the factors, and second, parameter estimates for the cognitive and non-cognitive measurement equations.

4.4.1 High school continuation

Table 4.6 reports the parameter estimates for the component of the model corresponding to the high school continuation Logit. The key result that this table shows is that cognitive and non-cognitive skills explain all of the gaps in high school graduation across race, gender, and socioeconomic status, except for white males, and finds that high SES black females are more likely to graduate from high school. This shows that the factor model we specified in Section 4.3 improves upon the results of our reduced form linear probability models in Table 4.3. Even after flexibly controlling for high school characteristics and nonlinear functions of cognitive and non-cognitive skill proxies, there was still a statistically significant gap in the probability of high school completion for high and low SES black females, high SES white males, and low SES white females, in addition to low SES white males.

We see in Table B.1 that most of these groups have levels of cognitive and non-cognitive skills that are statistically significantly lower than that of the reference group of high SES white females. These cognitive and non-cognitive skills also play an important role in explaining why students choose to progress to higher levels of secondary education, where students with higher levels of skills are more likely to complete a given grade. Hence, these results suggest that the bulk of the gaps we saw in educational attainment in the reduced form linear probability model reported in Table 4.3 were due to the effects of these underlying factors. Where this effect could operate directly, causing students with low skills to choose not to complete the next grade, or indirectly, by selecting out students with low skills before they choose whether to complete 12th grade and graduate from high school.

We also find, like Aucejo (2015), that the importance of both sets of skills for completing a given level of education declines as students progress through higher levels of secondary education. However, these results also suggest that the relative importance of non-cognitive skills increases as students progress, particularly by the time they are choosing whether to complete their high school education. This also implies that selection on these unobserved skills operates more heavily in earlier grades, particularly on cognitive skills. Hence, by the time students get to 12th grade, these skills are less important, since low-skilled students have already chosen to exit secondary education.

Table 4.6: Structural high school continuation parameter estimates

	10th Grade		11th Grade		12th Grade	
Intercept	3.610***	(0.8454)	3.342***	(0.8096)	2.886***	(0.7587)
Cognitive factor	2.396***	(0.1703)	1.850***	(0.1305)	1.222***	(0.1111)
Non-cognitive factor	1.369***	(0.1664)	1.236***	(0.1525)	1.282***	(0.1556)
Demographic characteristics[a]						
High SES						
White Male	0.107	(0.3034)	−0.551**	(0.2384)	−0.542***	(0.1981)
Black Female	0.402	(0.4427)	0.357	(0.3811)	1.540***	(0.5079)
Black Male	0.022	(0.4125)	0.230	(0.3841)	0.561	(0.3796)
Low SES						
White Female	−0.695**	(0.3250)	−0.964***	(0.2794)	−0.398	(0.2797)
White Male	−0.973***	(0.3016)	−0.903***	(0.2800)	−0.854***	(0.2483)
Black Female	0.464	(0.3683)	0.729**	(0.3486)	0.554*	(0.2975)
Black Male	−0.315	(0.3215)	−0.044	(0.3023)	0.124	(0.2814)
High school characteristics[b]						
Magnet high school	0.849	(0.5428)	−0.174	(0.3808)	0.371	(0.4106)
City with 250K+ residents	0.015	(0.2809)	0.130	(0.2580)	0.174	(0.2454)
Rural area	0.328*	(0.1859)	0.232	(0.1662)	0.158	(0.1527)
Share of black students	2.664***	(0.9593)	1.925**	(0.9215)	1.274	(0.8644)
Share of white students	2.574***	(0.8754)	2.287***	(0.8539)	2.201***	(0.8092)
2007 Cohort	−0.157	(0.1729)	−0.433***	(0.1562)	−0.534***	(0.1453)

Note: Logits for each grade are estimated on a sample of students that completed the prior grade.
[a] Indicators for race/gender/SES categories are defined relative to high-SES white females.
[b] High school characteristics are taken from a student's 9th grade institution.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

What is particularly interesting about these results, however, is that these skills are still unable to explain the persistent gap in educational attainment for low SES white males. The parameter estimates across 10th-12th grade show that the size of the gap decreases as these students progress through education. This suggests that selection on these unobservable skills play an important role in explaining why this gap decreases. Understanding why this gap is particularly persistent, even for low SES white males relative to low SES black males and low SES white females is an important policy question, given the wage penalty for high school dropouts relative to high school graduates. However, one potential explanation for these relative gaps is that low SES white males have better outside options for high school dropout employment than low SES black males or low SES white females.

4.4.2 Major intentions

Next we turn to the parameter estimates on college major choice for students that graduate from high school in Table 4.7. First, these results show a continuation of a pattern present in the Logits for high school continuation: cognitive and non-cognitive factors play less of a role in students' decisions to attend college, and non-cognitive skills play an even more limited role in their decision making. Furthermore, we also see that higher levels of skills, in particular cognitive skills, are associated with a higher probability of choosing to pursue any of the five possible majors, relative to not pursuing postsecondary education.

However, we do see some significant differences in how high levels of each skill are associated with the probability of choosing a particular major. High levels of cognitive skills are associated with a higher probability of choosing a STEM, health, or business, law, or social sciences degree. We also see that students with high levels of non-cognitive skills are more likely to choose a health major. This result could suggest that these majors require and select on high levels of non-cognitive ability: for example, students in a pre-med track may need to juggle a demanding course load and simultaneously prepare for medical school applications. However, since these data report major intentions, investigating this pattern further would require data on students postsecondary major choice and switching behavior. Doing so would allow us to identify whether high levels of non-cognitive skills actually determine success in these fields.

Table 4.7: Structural major choice parameter estimates

	STEM		Arts / Humanities Education / Other		Health		Business / Law Social Sciences		Undeclared Major	
Intercept	−2.627***	(0.6263)	−1.317**	(0.5339)	−0.737	(0.5407)	−0.786	(0.5421)	−0.560	(0.5374)
Cognitive factor	2.164***	(0.0777)	1.493***	(0.0631)	1.734***	(0.0674)	1.654***	(0.0702)	1.450***	(0.0674)
Non-cognitive factor	0.868***	(0.1138)	0.748***	(0.0901)	0.992***	(0.0986)	0.770***	(0.1053)	0.901***	(0.1040)
Demographic characteristics[a]										
High SES										
White Male	0.778***	(0.1131)	−0.643***	(0.0907)	−1.385***	(0.1067)	−0.552***	(0.1064)	−0.129	(0.0979)
Black Female	0.454*	(0.2333)	−0.310*	(0.1793)	0.534***	(0.1538)	0.710***	(0.1692)	0.185	(0.1843)
Black Male	1.334***	(0.1964)	−0.157	(0.1777)	−0.742***	(0.2182)	0.557***	(0.1790)	0.097	(0.1946)
Low SES										
White Female	−0.839***	(0.2767)	−0.997***	(0.1659)	−0.732***	(0.1569)	−1.033***	(0.2137)	−1.213***	(0.2190)
White Male	−0.423*	(0.2237)	−1.505***	(0.1900)	−2.106***	(0.2424)	−1.726***	(0.2674)	−1.395***	(0.2268)
Black Female	−0.425	(0.2646)	−0.670***	(0.1643)	0.169	(0.1409)	0.405***	(0.1560)	−0.163	(0.1683)
Black Male	1.181***	(0.1892)	−0.445***	(0.1691)	−1.245***	(0.2269)	0.186	(0.1773)	0.162	(0.1692)
High school characteristics[b]										
Magnet high school	0.314	(0.2268)	0.494**	(0.1946)	0.025	(0.2134)	0.552***	(0.1941)	0.495**	(0.2028)
City with 250K+ residents	0.364**	(0.1587)	0.307**	(0.1344)	0.677***	(0.1415)	0.575***	(0.1385)	0.267*	(0.1410)
Rural area	−0.097	(0.1053)	−0.059	(0.0869)	0.242**	(0.0941)	−0.188*	(0.0974)	−0.167*	(0.0925)
Share of black students	0.132	(0.7067)	0.147	(0.6091)	−0.626	(0.6136)	−1.078*	(0.6151)	−0.806	(0.6138)
Share of white students	−0.015	(0.6680)	0.395	(0.5738)	−0.654	(0.5835)	−0.691	(0.5847)	−0.615	(0.5786)
2007 Cohort	0.423***	(0.0969)	0.398***	(0.0797)	0.393***	(0.0849)	0.193**	(0.0881)	−0.226***	(0.0852)

Note: The multinomial Logit for major choice is estimated on a sample of students that completed high school, and parameters are defined relative to choosing not to continue on to college.

[a] Indicators for race/gender/SES categories are defined relative to high-SES white females.

[b] High school characteristics are taken from a student's 9th grade institution.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

These results also reveal interesting patterns relative to the linear probability model of majoring in STEM that we reported in Table 4.4. As in the reduced form results, we find that after controlling for dynamic selection on unobservable skills, we find the same high preference for STEM majors among all high SES groups and low SES black males, relative to high SES white females. However, unlike the reduced form results, we find that low SES white females are even less likely to choose STEM majors than high SES white females, and that these cognitive and non-cognitive skills largely explain the gap between high SES white females and low SES white males. Turning to other majors, these results suggest that relative to high SES white females, white males are less likely to choose a health major or a major in the arts, humanities, or education; and that black males and high SES black females are more likely to choose a major in business, law, or the social sciences. Finally, the parameter results across all majors suggest that low SES whites are less likely to choose to continue on to postsecondary education, conditional on completing high school.

4.4.3 College applications

The final set of student outcomes we report parameter estimates for is the Logit on applying to four or more colleges in Table 4.8. As was the case with both the high school continuation Logits and the major choice multinomial Logit, cognitive and non-cognitive factors play a significant role in determining the intensity with which students apply to colleges, where students with high levels of each factor are more likely to apply to four or more institutions. However, as was the case with these other results, both factors are less relevant and non-cognitive skills are relatively less important than for the 10th grade continuation Logit.

Similarly, controlling for dynamic selection on these two unobservable skills largely explains the race, gender and, socioeconomic status gaps we see in the rate of stu-

Table 4.8: Structural college application parameter estimates

Intercept	−0.517	(0.4971)
Cognitive factor	0.551***	(0.0683)
Non-cognitive factor	0.304***	(0.1021)
Demographic characteristics[a]		
High SES		
White Male	−0.147	(0.0981)
Black Female	0.448***	(0.1736)
Black Male	−0.149	(0.1895)
Low SES		
White Female	−0.417*	(0.2192)
White Male	−0.434*	(0.2633)
Black Female	0.228	(0.1692)
Black Male	0.016	(0.1911)
High school characteristics[b]		
Magnet high school	0.060	(0.1656)
City with 250K+ residents	0.538***	(0.1267)
Rural area	0.053	(0.0941)
Share of black students	0.141	(0.5886)
Share of white students	0.174	(0.5528)

Note: The Logit for applying to four or more colleges is estimated relative to a sample of students that completed high school and declared a major.
[a] Indicators for race/gender/SES categories are defined relative to high-SES white females.
[b] High school characteristics are taken from a student's 9th grade institution.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

dents applying to four or more institutions in the summary statistics (Table 4.1). This is particularly striking relative to the results of the linear probability model on applying to four or more institutions presented in Table 4.5. The reduced form results were unable to explain any of the gaps relative to high SES white females, with the exception of high SES white males, even with the inclusion of controls for high school characteristics and nonlinear functions of cognitive and non-cognitive skill proxies. However, the structural results suggest that only high SES black females are significantly more likely to apply to four or more institutions than high

SES white females, while low SES whites are only marginally less likely to apply to four or more institutions.

4.5 Conclusion

An individual's educational decisions, from level of schooling attained to major choice among college-goers, demarcate sizable differences in labor market outcomes. While ample research shows substantial returns to additional years of education, a simple unidimensional view of schooling is no longer sufficient, as evidenced by substantial differences in earnings premiums by college quality and across fields of study. Efforts to improve outcomes and bridge economic inequalities for disadvantaged groups require a better understanding of the process by which these groups navigate the path spanning secondary to postsecondary education, including college applications and major intentions.

We examine this crucial transition period for students in North Carolina. Statewide administrative data enables us to segment students by race, gender, and socioeconomic status, which provides a richer characterization of cross-group disparities than previously available in the US. This three-way interaction makes clear that collapsing across socioeconomic status masks significant heterogeneity within race and gender groups. Strikingly, low SES white females consistently perform worse than their disadvantaged black female peers in attainment and college applications. Meanwhile, high SES white females outperform all peers in attainment and several dimensions of college application behavior. SES has an influence on academic outcomes independent of the mediating circumstances of gender and race that is documented throughout this paper.

The characterization of family socioeconomic status using a binary indicator of free and reduced lunch eligibility obscures the nuances of family background, in particular the placement of individuals along the income distribution on both sides of the 185% poverty threshold. However, evidence exists to suggest that much of the empirical regularities observed here are not driven by only income differences. Disadvantaged white females in the NLSY97, for example, graduate high school at lower rates than disadvantaged African American females despite being financially better-off. Barriers to attainment for this group arise for reasons other than parental income.

We estimate the contribution of secondary school inputs, cognitive, and non-cognitive skills to observed gaps. We begin with a set of reduced form regressions that are increasingly augmented using measures for high school quality and student ability. Variables such as test scores or disciplinary infractions measure “pure” abilities with error, and can be themselves affected by schooling attainment. Furthermore, some schooling measures are jointly determined by underlying cognitive and socioemotional abilities. We address these issues by specifying a factor model of educational attainment, college major choice, and college application intensity. The structural model allows us to separately identify the effects of cognitive and non-cognitive skills across our eight socio-demographic groups. We find that these unobserved skills play a significant role in explaining the remaining gaps in educational outcomes across race, gender, and socioeconomic status. In particular, controlling for dynamic selection on cognitive and non-cognitive factors explains every gap in secondary education attainment other than low SES white males, and every gap but for high SES black females for college application intensity. While our factor model is unable to completely explain the remaining gaps in college major intentions, we are able to explain the gap in STEM major intention between high SES white females and low SES white males.

The characterization of observed ability measures or unobserved factors as explanatory accounts should not obscure the fact that they are shaped over time by familial, neighborhood, and schooling attributes. The factor model addresses this complexity in part by allowing measures to vary by school and individual characteristics. To the extent that latent skills are determined by contextual factors not considered in the model, the contributions of these skills to educational disparities reflect the collective impact of these influences. Whether they reflect the benefits reaped from experienced teaching (Clotfelter, Ladd, and Vigdor, 2006), peer group spillovers (Fruehwirth, 2013), parental investments (Murnane et al., 2006), or other inputs are paths for future research.

Finally, the results of our model corresponding to college major choice identified a number of interesting patterns, including a positive effect of high non-cognitive skills on the probability of a student majoring in health. In order to determine whether success in this and other majors depends on non-cognitive skills, we would need information on students' performance, major choice and switching behavior while in college. And a particularly interesting extension would be to connect these or similar data on primary and secondary education to administrative data on students' college careers. Doing so would then allow researchers to address the question of whether the significant gaps in labor market outcomes we observe across majors are due primary to treatment or selection. In other words: do engineers earn high wages because of the intrinsic value of the engineering-specific human capital they accumulate, or because employers value the high levels of skills they needed to succeed as an undergraduate engineering major?

Appendix A

CPS Data

The 1989-1991 monthly CPS survey has a sample target of 50,000 households split into eight representative subsamples, each of which is interviewed for the first and last four months of a 16-month period. In any given month, a new sample of 6,250 households is surveyed for the first time. As a result, the pooled monthly CPS data from January of 1989 through December of 1991 contain 268,750 unique households. From these pooled cross-sections we keep only observations in the age range spanned by the NLSY79 cohort, which leaves 795,631 observations. Then we drop observations where an individual is unemployed, does not report a Census occupation code, has a missing level of education, did not complete the reported level of education, or is enrolled in college. After making these cuts, we are left with a sample of 506,930 occupation and education level pairs, where the education level is defined as the highest grade achieved by the surveyed individual.

From this sample we estimate the required level of education for each occupation identified by its 3-digit Census occupation code. The required level of education for a given occupation is defined as the sample mode of the distribution of education levels among workers in the occupation. Then we match observed occupations in the NLSY79 to a required level of education based on their 3-digit occupation code. 125 out of 488 occupations are observed less than 100 times in our CPS pooled sample. In order to reduce the sampling variance of the corresponding required levels of education, we collapse these low-frequency occupations using two-digit codes rather than three-digit codes before applying the procedure described above. Less than 2% of our NLSY79 observations are in such occupations.

Appendix B

Factor and Measure Parameter Estimates

In this appendix I report the parameter estimates for four of the following sets of results from the cognitive and non-cognitive factor measurement system I use in Chapter 4: first, parameters that determine the relative distribution of unobserved skills among race, gender, and socioeconomic groups; second, parameters that govern measures of cognitive skills; and third, parameters that govern measures of non-cognitive skills, including continuous and binary measures in Table B.3 and a multinomial measure of homework intensity in Table B.4.

Table B.1: Structural factor parameter estimates

	Cognitive		Non-cognitive	
High SES				
White Male	−0.026	(0.0237)	−0.062**	(0.0265)
Black Female	−0.015	(0.0347)	−0.065*	(0.0369)
Black Male	−0.061	(0.0394)	−0.020	(0.0416)
Low SES				
White Female	−0.078*	(0.0441)	−0.010	(0.0435)
White Male	0.112**	(0.0505)	0.032	(0.0498)
Black Female	0.036	(0.0344)	−0.041	(0.0381)
Black Male	0.047	(0.0426)	−0.046	(0.0442)
Interactions				
High SES				
White Male	−0.124***	(0.0117)	−1.566***	(0.0196)
Black Female	−0.116***	(0.0227)	−1.574***	(0.0344)
Black Male	−0.144***	(0.0257)	−0.445***	(0.0420)
Low SES				
White Female	−0.117***	(0.0272)	−1.508***	(0.0463)
White Male	−1.861***	(0.0312)	−1.462***	(0.0584)
Black Female	−1.870***	(0.0210)	−1.638***	(0.0342)
Black Male	−0.135***	(0.0277)	−0.411***	(0.0436)

Note: Indicators for race/gender/SES categories are defined relative to high-SES white females.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Structural cognitive measure parameter estimates

	Continuous					
	8th Grade EOG Scores[a]				**Binary (Logit)**	
	Math		Reading		Academically Gifted[b]	
Intercept	0.029	(0.1243)	−0.101	(0.1183)	−2.401***	(0.6651)
Cognitive factor[c]	1		0.856***	(0.0088)	3.548***	(0.1173)
High school characteristics[d]						
Magnet high school	0.003	(0.0455)	0.058	(0.0441)	0.688***	(0.2329)
City with 250K+ residents	0.253***	(0.0320)	0.162***	(0.0308)	0.135	(0.1642)
Rural area	−0.042**	(0.0214)	−0.083***	(0.0204)	−0.504***	(0.1117)
Share of black students	−0.324**	(0.1420)	−0.102	(0.1353)	−1.186	(0.7657)
Share of white students	0.447***	(0.1333)	0.588***	(0.1271)	−0.140	(0.7132)
2007 Cohort	−0.033*	(0.0198)	−0.017	(0.0189)	−0.106	(0.1028)
σ^2	0.172	(1.0320)	0.285	(1.0198)		

[a] End-of-Grade scores in math and reading are normalized relative to statewide test scores for the year a student took a given EOG test.
[b] Academically / intellectually gifted is defined as whether a student was identified as academically and intellectually gifted up to 8th grade.
[c] In order to identify the cognitive factor, we must normalize the parameter for the cognitive factor to 1 on at least one of the measures.
[d] High school characteristics are taken from a student's 9th grade institution.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.3: Structural non-cognitive measure parameter estimates

	Continuous						Binary (Logit)			
			Absences							
	9th Grade GPA		Excused		Unexcused		Suspended		Tardy	
Intercept	3.138***	(0.1553)	1.740***	(0.5642)	3.059***	(0.6336)	−1.231***	(0.4176)	−0.599**	(0.2864)
Cognitive factor	0.728***	(0.0136)	−0.800***	(0.0546)	−1.399***	(0.0609)	−1.131***	(0.0516)	−0.077***	(0.0294)
Non-cognitive factor[a]	1		−1.631***	(0.0986)	−1.790***	(0.1029)	−1.272***	(0.0876)	0.014	(0.0529)
High school characteristics[b]										
Magnet high school	−0.094	(0.0596)	−0.531**	(0.2233)	−0.979***	(0.2502)	−1.290***	(0.3084)	0.024	(0.1450)
City with 250K+ residents	0.162***	(0.0409)	−1.565***	(0.1523)	−0.756***	(0.1709)	−1.354***	(0.1642)	−1.513***	(0.1129)
Rural area	−0.209***	(0.0265)	0.251**	(0.0981)	−0.477***	(0.1102)	−0.078	(0.0774)	0.119**	(0.0507)
Share of black students	−0.255	(0.1768)	0.536	(0.6466)	2.039***	(0.7262)	−0.435	(0.4763)	0.514	(0.3261)
Share of white students	−0.241	(0.1663)	1.430**	(0.6101)	0.022	(0.6849)	−1.429***	(0.4555)	−0.743**	(0.3125)
2007 Cohort	−0.031	(0.0243)	−0.028	(0.0895)	0.118	(0.1006)	−0.115	(0.0731)	−0.118***	(0.0476)
σ^2	0.528	(1.0283)	16.991***	(1.0152)	20.703***	(1.0151)				

[a] In order to identify the non-cognitive factor, we must normalize the parameter for the non-cognitive factor to 1 on at least one of the measures.
[b] High school characteristics are taken from a student's 9th grade institution.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.4: Structural cognitive measure parameter estimates (homework time /week)

	No HW assigned		Doesn't do HW		Less than 1 hr		More than 3 hrs	
Intercept	−2.574***	(0.9632)	−3.059***	(0.9977)	−0.391	(0.3136)	−0.979***	(0.3553)
Cognitive factor	−0.971***	(0.1252)	−0.782***	(0.1247)	−0.381***	(0.0330)	0.595***	(0.0353)
Non-cognitive factor	0.196	(0.2013)	−0.125	(0.1938)	−0.208***	(0.0547)	−0.009	(0.0599)
High school characteristics[a]								
Magnet high school	0.063	(0.5554)	0.205	(0.4160)	−0.177	(0.1431)	0.174	(0.1253)
City with 250K+ residents	−1.074**	(0.4186)	−0.064	(0.3074)	−0.418***	(0.0940)	0.260***	(0.0876)
Rural area	−0.084	(0.1940)	−0.007	(0.2077)	0.024	(0.0556)	−0.078	(0.0604)
Share of black students	−1.072	(1.1108)	−0.680	(1.1417)	−0.235	(0.3592)	−0.184	(0.4086)
Share of white students	−1.294	(1.0458)	−1.193	(1.0857)	−0.314	(0.3401)	0.233	(0.3833)
2007 Cohort	−0.220	(0.1835)	0.223	(0.1895)	0.110**	(0.0512)	0.149***	(0.0547)

Note: The multinomial Logit for the amount of time spent per week on homework is defined relative to the modal level of time spent on homework: 1-3 hours per week.
[a] High school characteristics are taken from a student's 9th grade institution.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Bibliography

Acemoglu, Daron and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, vol. 4, Part B, edited by David Card and Orley Ashenfelter. Elsevier, 1043 – 1171.

Aguirregabiria, Victor and Pedro Mira. 2007. "Sequential Estimation of Dynamic Discrete Games." *Econometrica* 75 (1):1 – 53.

Aït-Sahalia, Yacine and Andrew W. Lo. 2000. "Nonparametric Risk Management and Implied Risk Aversion." *Journal of Econometrics* 94 (1):9 – 51.

Alba-Ramirez, Alfonso. 1993. "Mismatch in the Spanish Labor Market: Overeducation?" *The Journal of Human Resources* 28 (2):259 – 278.

Allen, Jim and Rolf van der Velden. 2001. "Educational Mismatches Versus Skill Mismatches: Effects on Wages, Job Satisfaction, and on-the-Job Search." *Oxford Economic Papers* 53 (3):434 – 452.

Altonji, Joseph G. 1993. "The Demand for and Return to Education When Education Outcomes are Uncertain." *Journal of Labor Economics* 11:48 – 83.

Altonji, Joseph G., Peter Arcidiacono, and Arnaud Maurel. 2015. "The Analysis of Field Choice in College and Graduate School: Determinants and Wage Effects." Tech. rep., National Bureau of Economic Research.

Altonji, Joseph G., Erica Blom, and Costas Meghir. 2012. "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers." *Annual Review of Economics* 4:185 – 223.

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer. 2014. "Cashier or Consultant? Entry Labor Market Conditions, Field of Study, and Career Success." Forthcoming in the *Journal of Labor Economics*.

Anderson, Theodore W. and Herman Rubin. 1956. "Statistical Inference in Factor Analysis." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5.

Arcidiacono, Peter. 2004. "Ability Sorting and the Returns to College Major." *Journal of Econometrics* 121 (1-2):343 – 375.

Arcidiacono, Peter, Esteban Aucejo, Arnaud Maurel, and Tyler Ransom. 2014. "College Attrition and the Dynamics of Information Revelation." In *2014 Meeting Papers*, 529. Society for Economic Dynamics.

Arcidiacono, Peter and Paul B. Ellickson. 2011. "Practical Methods for Estimation of Dynamic Discrete Choice Models." *Annual Review of Economics* 3 (1):363 – 394.

Arcidiacono, Peter, V. Joseph Hotz, and Songman Kang. 2012. "Modeling College Major Choices using Elicited Measures of Expectations and Counterfactuals." *Journal of Econometrics* 166 (1):3 – 16.

Arcidiacono, Peter and Robert A. Miller. 2011. "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity." *Econometrica* 79 (6):1823 – 1867.

Arulampalam, Wiji, Paul Gregg, and Mary Gregory. 2001. "Unemployment Scarring." *The Economic Journal* 111 (475):577 – 584.

Aucejo, Esteban. 2015. "Explaining Cross-Racial Differences in the Educational Gender Gap." Working Paper, Centre for Economic Performance, London School of Economics and Political Science.

Autor, David H., David N. Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman. 2016. "School Quality and the Gender Gap in Educational Achievement." Working Paper 21908, National Bureau of Economic Research.

Bauer, Thomas K. 2002. "Educational Mismatch and Wages: A Panel Analysis." *Economics of Education Review* 21 (3):221 – 229.

Beffy, Magali, Denis Fougere, and Arnaud Maurel. 2012. "Choosing the Field of Study in Postsecondary Education: Do Expected Earnings Matter?" *The Review of Economics and Statistics* 94 (1):334 – 347.

Belzil, Christian and Jörgen Hansen. 2002. "Unobserved Ability and the Return to Schooling." *Econometrica* 70 (5):2075 – 2091.

Berger, Mark C. 1988. "Predicted Future Earnings and Choice of College Major." *Industrial and Labor Relations Review* 41 (3):418 – 429.

Bettinger, Eric P., Bridget T. Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment." *The Quarterly Journal of Economics* 127 (3):1205 – 1242.

Betts, Julian R. 1996. "What do Students Know About Wages? Evidence From a Survey of Undergraduates." *The Journal of Human Resources* :27 – 56.

Bils, M. and K. McLaughlin. 2001. "Interindustry Mobility and the Cyclical Upgrading of Labor." *Journal of Labor Economics* 19 (1):94 – 135.

Blanchard, Olivier J., Lawrence F. Katz, Robert E. Hall, and Barry Eichengreen. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* 1992 (1):1 – 75.

Bliss, Robert R. and Nikolaos Panigirtzoglou. 2004. "Option-Implied Risk Aversion Estimates." *The Journal of Finance* 59 (1):407 – 446.

Blom, Erica, Brian C. Cadena, and Benjamin J. Keys. 2015. "Investment Over the Business Cycle: Insights From College Major Choice." Tech. rep., IZA Discussion Papers.

Buchinsky, Moshe and Phillip Leslie. 2010. "Educational Attainment and the Changing US Wage Structure: Dynamic Implications on Young Individuals Choices." *Journal of Labor Economics* 28 (3):541 – 594.

Cameron, Stephen V. and James J. Heckman. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy* 109 (3):455 – 499.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, vol. 3, Part A, edited by Orley C. Ashenfelter and David Card. Elsevier, 1801 – 1863.

Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman. 2003. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice." *International Economic Review* 44:361 – 422.

Chevalier, Arnaud and Joanne Lindley. 2009. "Overeducation and the Skills of UK Graduates." *Journal of the Royal Statistical Society: Series A* 172:307 – 337.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *The Journal of Human Resources* 41 (4):778 – 820.

———. 2009. "The Academic Achievement Gap in Grades 3 to 8." *The Review of Economics and Statistics* 91 (2):398 – 419.

Cook, Michael D. and William N. Evans. 2000. "Families or Schools? Explaining the Convergence in White and Black Academic Performance." *Journal of Labor Economics* 18 (4):729 – 754.

Cunha, Flavio and James J. Heckman. 2007. "Identifying and Estimating the Distributions of *ex post* and *ex ante* Returns to Schooling." *Labour Economics* 14 (6):870 – 893.

———. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Non-cognitive Skill Formation." *The Journal of Human Resources* 43 (4):738 – 782.

Cunha, Flavio, James J. Heckman, and Salvador Navarro. 2007. "The Identification and Economic Content of Ordered Choice Models With Stochastic Thresholds." *International Economic Review* 48 (4):1273 – 1309.

Cunha, Flavio, James J. Heckman, and Susanne Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3):883 – 931.

Dolton, Peter and Anna Vignoles. 2000. "The Incidence and Effects of Overeducation in the UK Graduate Labour Market." *Economics of Education Review* 19 (2):179 – 198.

Duncan, J. and S. Hoffman. 1981. "The Incidence and Effects of Overeducation." *Economics of Education Review* 1:75 – 86.

Dutta, Santanu and Alok Goswami. 2010. "Mode Estimation for Discrete Distributions." *Mathematical Methods of Statistics* 19 (4):374 – 384.

Elbers, Chris and Geert Ridder. 1982. "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model." *The Review of Economic Studies* 49:403 – 409.

Epstein, Larry G. and Stanley E. Zin. 1991. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *Journal of Political Economy* 99 (2):263 – 286.

Frank, Robert H. 1978. "Why Women Earn Less: The Theory and Estimation of Differential Overqualification." *The American Economic Review* 68:360 – 373.

Freeman, Richard B. 1976a. "A Cobweb Model of the Supply and Starting Salary of New Engineers." *Industrial and Labor Relations Review* :236 – 248.

———. 1976b. *The Overeducated American*. New York: Academic Press.

Frenette, Marc. 2004. "The Overqualified Canadian Graduate: The Role of the Academic Program in the Incidence, Persistence, and Economic Returns to Overqualification." *Economics of Education Review* 23 (1):29 – 45.

Fruehwirth, Jane C. 2013. “Identifying Peer Achievement Spillovers: Implications for Desegregation and the Achievement Gap.” *Quantitative Economics* 4 (1):85 – 124.

Fryer, Roland G. and Steven D. Levitt. 2004. “Understanding the Black-White Test Score Gap in the First Two Years of School.” *The Review of Economics and Statistics* 86 (2):447 – 464.

———. 2006. “The Black-White Test Score Gap Through Third Grade.” *American Law and Economics Review* 8 (2):249 – 281.

Gemici, Ahu and Matthew Wiswall. 2014. “Evolution of Gender Differences In Post-Secondary Human Capital Investments: College Majors.” *International Economic Review* 55 (1):23 – 56.

Gottschalk, Peter and Michael Hansen. 2003. “Is the Proportion of College Workers in Noncollege Jobs Increasing?” *Journal of Labor Economics* 21 (2):449 – 471.

Gregory, Anne, Russell J. Skiba, and Pedro A. Noguera. 2010. “The Achievement Gap and the Discipline Gap: Two Sides of the Same Coin?” *Educational Researcher* 39 (1):59 – 68.

Grogger, Jeff and Eric Eide. 1995. “Changes in College Skills and the Rise in the College Wage Premium.” *The Journal of Human Resources* 30 (2):280 – 310.

Guvenen, Fatih. 2007. “Learning Your Earning: Are Labor Income Shocks Really Very Persistent?” *The American Economic Review* :687 – 712.

Hartog, Joop. 1980. “Earnings and Capability Requirements.” *The Review of Economics and Statistics* 62:230 – 240.

———. 2000. “Over-education and Earnings: Where are We, Where Should We go?” *Economics of Education Review* 19 (2):131 – 147.

Heckman, James J. and Alan B. Krueger. 2005. *Inequality in America: What Role for Human Capital Policies?* Cambridge, Mass.; London: The MIT Press, revised ed.

Heckman, James J., Lance J. Lochner, and Petra E. Todd. 2006. “Earnings Functions, Rates of Return and Treatment Effects: the Mincer Equation and Beyond.” In *Handbook of the Economics of Education*, vol. 1, edited by Eric A. Hanushek and Finis Welch. Elsevier, 307 – 458.

Heckman, James J. and Burton Singer. 1984. “A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data.” *Econometrica* 52:271 – 320.

Heckman, James J., Jora Stixrud, and Sergio Urzúa. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3):411 – 482.

Hersch, Joni. 1991. "Education Match and Job Match." *The Review of Economics and Statistics* :140 – 144.

Hotz, V. Joseph and Robert A. Miller. 1993. "Conditional Choice Probabilities and the Estimation of Dynamic Models." *The Review of Economic Studies* 60 (3):497 – 529.

Hoxby, Caroline and Sarah Turner. 2013. "Expanding College Opportunities for High-Achieving, Low Income Students." *Stanford Institute for Economic Policy Research Discussion Paper* (12-014).

Jacob, Brian A. 2002. "Where the Boys Aren't: Non-cognitive Skills, Returns to School and the Gender Gap in Higher Education." *Economics of Education Review* 21 (6):589 – 598.

Kahn, Lisa B. 2010. "The Long-Term Labor Market Consequences of Graduating From College in a Bad Economy." *Labour Economics* 17 (2):303 – 316.

Katz, Lawrence F. and David H. Autor. 1999. "Changes in the Wage Structure and Earnings Inequality." In *Handbook of Labor Economics*, vol. 3, Part A, edited by Orley C. Ashenfelter and David Card. Elsevier, 1463 – 1555.

Kiker, B.F., Maria C. Santos, and M. Mendes de Oliveira. 1997. "Overeducation and Undereducation: Evidence for Portugal." *Economics of Education Review* 16 (2):111 – 125.

Kinsler, Josh. 2011. "Understanding the Black–White School Discipline Gap." *Economics of Education Review* 30 (6):1370 – 1383.

Kinsler, Josh and Ronni Pavan. 2014. "The Specificity of General Human Capital: Evidence from College Major Choice." Tech. rep.

Korpi, Tomas and Michael Tåhlin. 2009. "Educational Mismatch, Wages, and Wage Growth: Overeducation in Sweden, 1974–2000." *Labour Economics* 16 (2):183 – 193.

Kroft, Kory, Fabian Lange, and Matthew J. Notowidigdo. 2013. "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment." *The Quarterly Journal of Economics* 128 (3):1123 – 1167.

Lee, Donghoon, Sang Yoon (Tim) Lee, and Yongseok Shin. 2014. "The Option Value of Human Capital." Unpublished manuscript.

Leuven, Edwin and Hessel Oosterbeek. 2011. "Overeducation and Mismatch in the Labor Market." In *Handbook of the Economics of Education*, vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier.

Liu, Kai, Kjell G. Salvanes, and Erik Ø. Sorensen. 2012. "Good Skills in Bad Times: Cyclical Skill Mismatch and the Long-Term Effects of Graduating in a Recession." *IZA Discussion Paper No. 6820* .

Manski, Charles F. 1993. "Adolescent Econometricians: How do Youth Infer the Returns to Schooling?" In *Studies of Supply and Demand in Higher Education*. University of Chicago Press, 43 – 60.

———. 2004. "Measuring Expectations." *Econometrica* :1329 – 1376.

Mavromaras, Kostas, Stéphane Mahuteau, Peter Sloane, and Zhang Wei. 2013. "The Effect of Overskilling Dynamics on Wages." *Education Economics* 21 (3):281 – 303.

Montmarquette, Claude, Kathy Cannings, and Sophie Mahseredjian. 2002. "How do Young People Choose College Majors?" *Economics of Education Review* 21 (6):543 – 556.

Murnane, Richard J., John B. Willett, Kristen L. Bub, Kathleen McCartney, Eric Hanushek, and Rebecca Maynard. 2006. "Understanding Trends in the Black-White Achievement Gaps During the First Years of School." *Brookings-Wharton Papers on Urban Affairs* :97 – 135.

Oreopoulos, Philip, Till von Wachter, and Andrew Heisz. 2012. "The Short- and Long-Term Career Effects of Graduating in a Recession." *American Economic Journal: Applied Economics* 4:1–29.

Paglin, Morton and Anthony M. Rufolo. 1990. "Heterogeneous Human Capital, Occupational Choice, and Male-Female Earnings Differences." *Journal of Labor Economics* :123–144.

Reardon, Sean F. and Joseph P. Robinson. 2008. "Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps." *Handbook of Research in Education Finance and Policy* :497 – 516.

Rubb, Stephen. 2003. "Overeducation: A Short or Long Run Phenomenon for Individuals?" *Economics of Education Review* 22 (4):389 – 394.

Rumberger, Russell W. 1987. "The Impact of Surplus Schooling on Productivity and Earnings." *The Journal of Human Resources* 22:24 – 50.

Saporta-Eksten, Itay. 2014. "Job Loss, Consumption and Unemployment Insurance." Unpublished manuscript.

Schmieder, Johannes F., Till von Wachter, and Stefan Bender. 2013. "The Causal Effect of Unemployment Duration on Wages: Evidence from Unemployment Insurance Extensions." *NBER Working Paper No. 19772* .

Sicherman, Nachum and Oded Galor. 1990. "A Theory of Career Mobility." *Journal of Political Economy* 98 (1):169 – 192.

Stange, Kevin M. 2012. "An Empirical Investigation of the Option Value of College Enrollment." *American Economic Journal: Applied Economics* 4 (1):49 – 84.

Stinebrickner, Ralph and Todd R. Stinebrickner. 2008. "The Effect of Credit Constraints on the College Drop-Out Decision: A Direct Approach Using a New Panel Study." *The American Economic Review* 98 (5):2163 – 2184.

———. 2011. "Math or Science? Using Longitudinal Expectations Data to Examine the Process of Choosing a College Major." Tech. rep., National Bureau of Economic Research.

Strand, Steve. 2014. "School Effects and Ethnic, Gender and Socio-Economic Gaps in Educational Achievement at Age 11." *Oxford Review of Education* 40 (2):223 – 245.

Turner, Sarah E. and William G. Bowen. 1999. "Choice of Major: The Changing (Unchanging) Gender Gap." *Industrial and Labor Relations Review* 52 (2):289 – 313.

Urzúa, Sergio. 2008. "Racial Labor Market Gaps: the Role of Abilities and Schooling Choices." *The Journal of Human Resources* 43 (4):919 – 971.

Verdugo, Richard R. and Naomi T. Verdugo. 1989. "The Impact of Surplus Schooling on Earnings: Some Additional Findings." *The Journal of Human Resources* 24 (4):629 – 643.

Weinberger, Catherine J. 1999. "Mathematical College Majors and the Gender Gap in Wages." *Industrial Relations: A Journal of Economy and Society* 38 (3):407 – 413.

Wiswall, Matthew and Basit Zafar. 2015a. "Determinants of College Major Choice: Identification Using an Information Experiment." *The Review of Economic Studies* 82 (2):791 – 824.

———. 2015b. "How Do College Students Respond to Public Information about Earnings?" *Journal of Human Capital* 9 (2):117 – 169.

Zafar, Basit. 2011. "How do College Students Form Expectations?" *Journal of Labor Economics* 29 (2):301 – 348.

———. 2013. “College Major Choice and the Gender Gap.” *The Journal of Human Resources* 48 (3):545 – 595.

Biography

Brian Christopher Clark was born on November 4th, 1987 in Newark, Ohio. He earned a B.S. in Economics and Decision Sciences with distinction from Miami University in 2010. He also earned a M.A. in Economics in 2011 and a Ph.D. in Economics in 2016 from Duke University. After graduation, he will join the Federal Trade Commission's Bureau of Economics in Washington D.C.