Racial and Gender Differences in School-College-Career Paths*

Katarína Borovičková Alessia Leibert Anusha Nath

September 2024

Abstract

This paper examines the factors affecting early career gender and racial wage gaps. Using a rich longitudinal administrative data from Minnesota and a dynamic discrete choice model, we decompose the wage gap into three components: *endowments* of school quality, family background, ability, and unobserved types; the *decision path* of sequential human capital investment made by individuals in high school, college, and labor market; and a *within-path* component, given endowments. We find that 41% of the gender wage gap can be explained by women choosing paths with lower earnings while differences in endowments explain about 15% of the gap. In contrast, for racial gaps, only 10% of the gap is explained differences in decision paths while endowments explain 37%. The remaining within-path gender and racial gaps are mostly driven by within firm variation for high skilled paths and across firm variation for low skilled paths. Our results show that education policies that encourage individuals to take STEM paths can reduce close to half of the early career gender wage gap while it would have little effect on reducing racial wage gap. Instead, education interventions in schools before individuals reach high school can be most effective in reducing early career White-Black wage gap.

JEL Classifications: I23, I26, J16, J24, J31

Keywords: Human Capital Investments, Wage Gaps, Racial Gap, Gender Gap, School Quality

*Borovičková: Federal Reserve Bank of Richmond, Leibert: Minnesota Department of Employment and Economic Development, and Nath: Federal Reserve Bank of Minneapolis. The findings, conclusions, views, and opinions are solely those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis, the Federal Reserve Bank of Richmond, or the Federal Reserve System.

1 Introduction

Despite decades of progress, gender and racial disparities in wages still exist in the United States: females earn less than males and Black and Hispanic workers earn less than White workers. At the same time, there is increasing evidence that differences in learning ability and initial conditions in the labor market account for more of the variation in lifetime earnings compared to differences in shocks received over the working lifetime of an individual (Huggett et al. (2006), Huggett et al. (2011)).

This paper examines the importance of human capital investments in high school and college for explaining early career wage differences and opens the black box of what goes into these initial labor market conditions. We focus our analysis on early career outcomes of individuals for two reasons. First, the wage gaps across demographic groups increase, not close, over workers' life cycle (Goldin et al. (2017), Rauh and Valladares-Esteban (2018)). It is therefore important to understand why there is an initial gap. Second, factors affecting wage gaps later in life, such as differences in home responsibilities (Goldin (2014), Blau and Kahn (2017), Erosa et al. (2022)), promotion standards (Lazear and Rosen (1990)), or wage bargaining abilities (Babcock and Laschever (2003)), are not likely to affect wage differences in the first jobs. Policy responses to factors that affect initial wage gaps would therefore be likely different from policy responses that explain gaps over the working life cycle.

We use a novel data set from the state of Minnesota which contains information on students from kindergarten through post secondary education into the work force. A unique feature of the labor market information in our administrative dataset is that for each worker, we observe quarterly hours worked in addition to quarterly earnings. This allows us to construct hourly wages, hence removing the differences in hours worked as a source of inequality. Our main outcome of interest is average hourly wages of individuals in the first three years of the labor market after their terminal degree.

We use information on students before they enter high school to measure differences in their endowments or initial states. We use standardized test scores and attendance to construct measures of cognitive and non-cognitive ability measures respectively. We use enrollment into free or reduced price lunch as a proxy for socioeconomic background and language spoken at home as a proxy for cultural factors. We also control for the quality of high school students attend. We use a Value Added Model to obtain the distribution of school quality and Empirical Bayes Methods to improve the estimates for each school.

While the differences in students' abilities, demographics, and family background affect their decisions in high school and beyond, there can be important unobserved heterogeneity driving the choices. We estimate individual unobserved types from labor market history data using group fixed

effect estimator as in Bonhomme and Manresa (2015).

Given the initial endowments and unobserved types, we examine how individuals make sequential decisions in high school, college, and the labor market. The first choice individuals make is what type of advanced courses they enroll into in high school. In order to ensure that the enrollment differences are driven by choices and not due to availability of advanced courses, we restrict our analysis to schools which offer them (most public schools in Minnesota offer such courses).

At the end of high school, individuals decide whether to attend a two-year, a four-year college, or not go to a college. If they go to college, they choose one of 10 majors and then decide whether to drop out, graduate and go to the labor market, or graduate and continue further studies. Students who did not go to college do not make any choices at these points in the sequence. They go directly to the labor market. Once in the labor market, the individuals choose one of nine industries to work in.

We refer to this sequence of choices made at different stages in life as the school-college-career path (or simply the path). We use a dynamic discrete choice model to analyze the decisions that students make on each node of the path. Our set up allows us to estimate the model as a sequence of static multinomial logit models.

We first analyze whether the decision to take Advanced Placement (AP) classes and Career and Technical Education (CTE) classes depends on gender or race, while controlling for ability, socioeconomic background, and school quality. We find significant differences, with males and Whites more likely to take special classes focused on STEM than females and racial minorities, respectively. This is especially true for high-ability students in low quality high schools.

This early choice of special classes affects later choices. Students who took AP or CTE classes are more likely to go to college than students who did not take special classes. This is true for any gender and race. We find that conditional on special classes, males and Blacks are more likely to go to college than females and other racial groups.

Zooming now on students who decided to go to college, the choice of STEM versus non-STEM AP and CTE classes is closely related to the choice of major. Not surprisingly, students with STEM classes tend to choose STEM majors. This is another important node where students make different choices which later affect the choice of occupation and the wage. Students are more likely to drop out of college if college quality is below median or if they are from disadvantaged socioeconomic background. This holds for both males and females as well as all of the racial and ethnic groups.

We use the estimated model to analyze factors affecting the gender wage gap. We decompose the wage gap into three components: the initial endowments, the decision paths, and a within-path component conditional on endowments. We find that 41% of the gender wage gap can be explained by women choosing paths with lower wages while differences in endowments explain about 15% of the gap. In contrast, for racial gaps, only 10% of the gap is explained differences in decision

paths while endowments explain 37%. The remaining 44% for gender and 53% of White-Back wage gap is explained by within-path differences.

To further investigate which endowment characteristics matter and when differences in choice of paths arise, we carry out two counterfactual exercises. We first examine how the aggregate gender and White-Black gap changes when we equalize the distribution of one endowment at a time between the analayzed group and the benchmark group. We find that each of the endowments has a small contribution towards reduction in the wage gaps, especially of White-Black gap. Changing all endowments together closes the gap with a magnitude that is consistent with our decomposition exercise. This finding shows that the joint distribution of the initial characteristics matter and that education policy interventions that target only one factor would not be very successful in reducing wage gaps. For example, simply moving students from lower quality schools to higher quality schools will not improve outcomes unless students also get the resources needed to improve their test scores.

Next, we conduct a counterfactual exercise where we modify behaviour of one group in a specific node while keeping the behaviour in the other nodes and the distribution of endowments unchanged. We find that in this case, the gender gap would decrease by half if females choose the same college majors as men. Interestingly, if we assign females the decisions of men on all nodes of the decision path, the reduction in wage gap is less than when only college major choice is changed. The patterns are similar for race - college major is the most critical node. However, in contrast, there is no reduction in racial wage gap if we change all the decision nodes together.

These results show that education policies that encourage individuals to take STEM paths can reduce close to half of the early career gender wage gap while it would have little effect on reducing racial wage gap. Instead, education interventions in schools before individuals reach high school can be most effective in reducing early career White-Black wage gap.

Relation to the literature. Our paper contributes to a rich literature on gender and racial wage gaps. In explaining gender wage gaps over the lifecycle, recent studies have documented that human-capital factors are relatively unimportant (Altonji and Blank (1999), Goldin (2014), Blau and Kahn (2017)) and that gender-specific value of job flexibility, differential occupational choice, and differential hours worked are more important factors. Our paper focuses on early careers rather than wage gaps over the life cycle and shows that the sequence of choices made along the school-college-career paths are important factors for explaining wage gaps in initial jobs. Since gender wage gaps increase over the life cycle (Goldin et al. (2017)), our results suggest that human-capital factors should not be ignored.

While the importance of pre-labor-market factors in White-Black wage gaps and labor market outcomes has been documented in the economics literature (Altonji and Blank (1999), Cameron

and Heckman (2001), Carneiro et al. (2005), Fryer (2011)), there is a debate about the relative importance of various pre-labor market factors. In 1987, years of schooling completed and scores on the Armed Forces Qualifications Test (AFQT) accounted for 75% of the hourly earnings difference between Whites and Blacks, and differences in work experience accounted for most of the remaining gap (O'Neill (1990)). The remaining White-Black wage differential among young men was around 7%, and statistically insignificant (Neal and Johnson (1996)). Black et al. (2006) argue that the AFQT is a better measure of pre-market human capital than years of schooling. They additionally control for non-test college degree-level characteristics, such as field of highest degree. They find that after controlling for these characteristics, no racial differences in wages emerges.

In subsequent studies, the wage differential between black and white men was found to be larger than Neal and Johnson (1996) estimate (Carneiro et al. (2005), Rodgers III and Spriggs (1996)). Upon controlling for years of schooling at the time the AFQT test was taken, a larger wage differential emerges (Lang and Manove (2011)). Lang and Manove reject the hypothesis that this is due to differences in school quality between blacks and whites. Instead, they argue that the wage gap is likely due to differential treatment in the labor market, such as statistical discrimination in the search process.

Our paper sheds light on the debate by bringing a rich longitudinal administrative data that allows us to examine these various factors in early careers of individuals. Our results suggest that while both these factors are at play, within-path component is significantly larger than the across-path components in explaining racial wage gaps. In other words, our results are closer to the findings of Lang and Manove (2011).

We also contribute to the literature which examines the impact of high school curriculum on the choice of major and earnings. Joensen and Nielsen (2009) exploit educational reform in Denmark in 1988 that allowed students to combine advanced math with chemistry rather than only with physics. They find that taking advanced math course in combination with advanced chemistry course increases earnings by 0.20 log points. However, controlling for post-secondary educational attainment eliminates most of the effect. In contrast, our paper shows that post-secondary choices are path dependent - the courses chosen in high school affects college degree and college major choices and the sequence of choices in turn explains earnings outcomes. Moreover, similar to Joensen and Nielsen (2009), many papers in this area focus on math classes in high school due to limited data availability. The richness of our data allows us to examine the choice of various AP and CTE classes, splitting them into STEM and non-STEM related, hence using a wider set of classes.

Arcidiacono (2004) estimates returns to majors using a structural model to control for selection into majors. He finds large differences in earning premia across majors. Math ability is important both for labor market returns and sorting into particular majors. We similarly find that the choice

of classes in high school is important for choosing the path, including the major, as well as the earnings.

Hoekstra (2009) looks at contribution of college quality to earnings. He estimates that attending the most selective state university increases earnings by around 20% for white men. In our data set, we observe the college name and hence are able to control for college quality. We find that attending a high-quality college increases men's wages but decreases women's wages.

Arcidiacono et al. (2016) try to understand why minority students are underrepresented in STEM. They examine data for students who applied to, were admitted to, enrolled at, and/or graduated with a BA degree from one of the University of California campuses in the late '90s. They find that minority students at top ranked campuses would, on average, have significantly higher probabilities of graduating in STEM if they had attended lower ranked campuses. In our estimation, we allow the decision of which major to choose and whether to graduate depend on the college quality, and differently so for different demographic groups. In light of Arcidiacono et al. (2016), it is important to have such flexibility.

2 Model of Education Choices

We use a dynamic discrete choice model to analyze decisions that students make.

2.1 Model

Students make a sequence of education and labor market decisions. We use t to denote times at which these decisions are made, $t = 1, \dots T$.

At time t, a student can choose one of J_t options, indexed by $j = 1, \dots J_t$. The flow utility of a student with characteristics X_t at time t is from making a choice j is $u_{tj}(X_t) + \varepsilon_{tj}$. We let d_{tj} be the binary variable which equals one of a student makes chooses option j at time t, and zero otherwise. Finally, ε_{tj} is the stochastic utility component associated with choice j. We use $\varepsilon_t = (\varepsilon_{t1}, \dots \varepsilon_{tJ_t})$ to denote the vector of stochastic utility components at time t. We assume that ε_t are iid across students, choices and time.

Student's optimization problem at time t

$$\max_{d_{tj} \in \{0,1\}} E_0 \left(\sum_{t=1}^{T} \beta^t \sum_{j=1}^{J_t} d_{tj} \left(u_{tj}(X_t) + \varepsilon_{tj} \right) \right),$$

where β is the discount factor and expectation is taken over realizations of ε as well as evolution of state variable X_t.

We formulate the Bellman equation for the problem. At time t, a student The Bellman equation $V(S_{it};\theta)$ is

$$V_t(X_t,\varepsilon_t) = \max_{d_{tj}\in\{0,1\}} \left(\sum_{j=1}^{J_t} d_{tj}\left(u_{tj}(X_t)+\varepsilon_{tj}\right) + \beta E_t\left(V_{t+1}(X_{t+1},\varepsilon_{t+1})\right) \right).$$

We next define the conditional valuation function associated with choosing j in period t as

$$v_{tj}(X_t) = u_{tj}(X_t) + \beta E\left[V_{t+1}(X_{t+1})|X_t, d_{tj}=1\right].$$

Then, at time t, we can represent student's choice as

$$V_t(X_t,\varepsilon_t) = \max_j \left(v_{tj}(X_t)+\varepsilon_{tj}\right).$$

Hence, if we can evaluate $v_{tj}(X_t)$, then we can solve decision at time t as a static decision.

To proceed, we impose more structure on the problem. We assume that ε_t is extreme value, and that the state X_t evolves according to $X_{t+1} = [X_t, \{d_{tj}\}_j]$ with X_0 given as initial condition.

The assumption that ε is extreme value has two implications. First, conditional on knowing continuation value, we can estimate decision at time t as a multinomial logit model. Second, we can rewrite $v_{tj}(X_t)$ in a useful form.

$$\begin{aligned}
v_{tj}(X_t) &= u_{tj}(X_t) + \beta E\left[V_{t+1}(X_{t+1})|X_t, d_{tj}=1\right] \\
&= u_{tj}(X_t) + \beta E\left[\max_k \left(v_{t+1k}(X_{t+1})+\varepsilon_{t+1k}\right)\right] \\
&= u_{tj}(X_t) + \beta \int\int \max_k \left(v_{t+1k}(X_{t+1})+\varepsilon_{t+1k}\right) dG_\varepsilon(\varepsilon_{t+1}) dG_X(X_{t+1}|X_t, d_{tj}=1),
\end{aligned}$$

where in the last line we specify that the expectation is taken with respect to realization of ε_{t+1} as well as evolution of the state variable.

We first deal with expectation with respect to ε_{t+1}. Our distributional assumptions on ε imply that

$$\int \max_k \left(v_{t+1k}(X_{t+1})+\varepsilon_{t+1k}\right) dG_\varepsilon(\varepsilon_{t+1}) = \log\left(\sum_{k=1}^{J_{t+1}} e^{v_{t+1k}(X_{t+1})}\right) + \gamma,$$

where γ is Euler constant. This can be simplified further using the insight from Hotz and Miller (1993). The logit equation for some arbitrary $k*$ implies that

$$Pr(d_{t+1,k^*}=1|X_{t+1}) = \frac{e^{v_{t+1,k^*}(X_{t+1})}}{\sum_{k=1}^{J_{t+1}} e^{v_{t+1,k}(X_{t+1})}},$$

and hence

$$\log\left(\sum_{k=1}^{J_{t+1}} e^{v_{t+1k}(X_{t+1})}\right) = v_{t+1,k^*}(X_{t+1}) - \log Pr(d_{t+1,k^*}=1|X_{t+1}).$$

Thus, the expected value with respect to stochastic utility realization can be expressed as the conditional valuation function associated with one choice k^* (it can be any choice) and a term which can be directly measured in the data, as long as the state X_{t+1} is observable.

In our framework, it is straightforward to deal the conditional expectation with respect to state variable, $dG_X(X_{t+1}|X_t, d_{tj}=1)$. Recall that the evolution of the state variable is given by $X_{t+1} = [X_t, \{d_{tj}\}_j]$.

Putting all these results together, $v_{tj}(X_t)$ can be written as

$$v_{tj}(X_t) = u_{tj}(X_t) + \beta v_{t+1k^*}(\tilde{X}_{t+1}) - \beta \log\left(Pr(d_{t+1,k^*}=1|X_t, d_{tj}=1)\right) + \beta\gamma,$$

where $\tilde{X}_{t+1} = [X_t, \{d_{tk}\}_k]$, with $d_{tk}=1$ if $k=j$ and zero otherwise.

Finally, we assume that the period utility is linear in state x_t, so that the utility flow from choice j at time t is

$$u_{tj}(X_t) = \theta_{jt} X_t.$$

We recover terms $\log\left(Pr(d_{t+1,k^*}=1|X_t, d_{tj}=1)\right)$ in the data and take linear projection on $[X_t, d_{tj}=1]$. That is, we represent it as

$$\log Pr(d_{t+1,k^*}=1|X_t, d_{tj}=1) = \alpha_{t+1,j} + \beta_{t+1,k^*} X_t + \eta_j,$$

with η_j is a zero mean error. This implies that $v_{tj}(X_t)$ is linear in X_t,

$$v_{tj}(X_t) = \gamma_{t,j} X_t.$$

Our goal is to estimate the utility functions, that is, parameter vectors θ_{jt} for all t and all $j = 1, \ldots J_t$. We proceed as follows. We estimate the multinomial logit model period by period to recover $\hat{\gamma}_{t,j}$. We then recover θ_{jt} backward, starting from $t=T$.

In period $t=T$, there is no continuation value and hence $v_{Tj}(X_T) = u_{Tj}(X_T)$, and so $\hat{\theta}_{T,j} = \hat{\gamma}_{T,j}$. At $t=T-1$, we use

$$\begin{aligned} v_{tj}(X_t) &= u_{tj}(X_t) + \beta v_{t+1k^*}(\tilde{X}_{t+1}) - \beta \log\left(Pr(d_{t+1,k^*}=1|X_t, d_{tj}=1)\right) + \beta\gamma \\ \hat{\gamma}_{t,j} X_t &= \hat{\theta}_{jt} X_t + \beta\left(\hat{\gamma}_{T,k^*}\tilde{X}_{t+1}\right) - \beta\left(\hat{\alpha}_{T,k^*} + \hat{\beta}_{T,k^*} X_t\right) + \beta\gamma \end{aligned}$$

where $\tilde{X}_{t+1} = [X_t, \{d_{tk}\}_k]$. The only unknown in this equation is $\hat{\theta}_{jt}$ and so we can recover it.

2.2 Unobserved Heterogeneity

Students are heterogeneous. They differ in their abilities, demographics as well as preferences for different classes. While there are several variables in X_{it} that allow us to capture these differences, there are still an important unobserved heterogeneity component, which we denote by ψ_i. This captures heterogeneity in individual's productivity. We estimate ψ from the pre-high school longitudinal data using group fixed effect estimator as in Bonhomme and Manresa (2015).

To recover ψ, we proceed as follows. The math test score of individual i at time t is

$$M_{it} = \alpha_{g_i t} + X_{it}'\beta_x + \nu_{it}, \quad i = 1, \dots T$$

where $\alpha_{g_i t}$ is the time specific group effect and X_{it} are observable characteristics of i at time t consisting of age, demographics, FRPL status, homelessness indicator, migrant indicator, and language spoken at home. In particular, we assume that there are G groups and use $g_i \in \{1, \dots G\}$ to denote which group worker i belongs to. The group fixed effect estimator is defined as the solution to the following problem

$$(\hat{\beta}_x, \hat{\beta}_d, \hat{\alpha}, \hat{\gamma}) = \arg \min_{\beta_x, \beta_d, \psi, \gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(M_{it} - X_{it}'\beta_x - D_{it}\beta_d - u_{it}\beta_u - \alpha_{g_i t} \right)^2$$

where γ denotes particular grouping of workers. This specification groups workers based on similarity in evolution of test scores from elementary to high school.

We use assignment of workers into groups, $\hat{\gamma}$, as the unobserved productivity type ψ_i. In particular, we represent ψ as a G-dimensional vector of dummy variables $\psi = (\psi^1, \psi^2, \dots \psi^G)$ where $\psi_i^g = 1$ if student i belongs to group $g \in G$, and zero otherwise. We treat the vector ψ as known in the estimation of multinomial logit models.

3 Data

We use a novel Minnesota statewide longitudinal education data (SLEDS) which matches student data from kindergarten through completion of post secondary education and into the workforce. The period of our study is 2003-2019.

The first block of the data comes from Minnesota Department of Education and follow students in K-12 grades. All students in Minnesota public schools are included. The data contains school and district identifiers of the elementary, middle and high school attended by the student, together with standardised test scores in grades 4, 8, and 11. We further see which Advanced Placement (AP) classes and Career and Technical Education (CTE) classes students take in high school. These

are advanced courses students can enroll into on top of basic course requirements. We observe several demographic characteristics, such as gender, age, race, language spoken at home, school district of residence, and whether a students receives free or reduced lunch (FRL). In addition, we observe annual attendance for students.

The second block comprises the universe of students attending a college in the state of Minnesota maintained by Minnesota Office of Higher Education. We observe college identifiers, degree that a student enrolls for and achieves, and the choice of major. The pot secondary files also consist of demographic characteristics such as age, gender, race and ethnicity, and citizenship status. We also obtained the National Clearinghouse Data which helps us track students who left Minnesota to attend college in other states. This data set also consists of information on college identifiers, major choice, and degree type.

Finally, the third data set comes from Minnesota Department of Employment and Economic Development and contains the labor market histories of individuals who have ever worked in Minnesota since 2003. For employed workers, we observe quarterly earnings and hours worked, together with establishment and firm identifiers, industry, and location. If an individual is not in the data set for any amount of time, they could either have had an unemployment spell or been self employed for the duration of the absence.

We select cohorts of students who graduated between 2010 and 2014 since these are the cohorts for which we observe test scores in grade 8, high school and college choices, as well as earnings in the first couple of years of their careers. We consider earnings in the first three years after graduation for students who graduate from college and are employed. For high school dropouts or college dropouts, we consider the first three years of employment after the last recorded year of study.

The male-female wage gap in the raw data is 3%. The White-Black, White-Hispanic, and White-Asian wage gaps are 25%, 15%, and 10% respectively. These are the differences in hourly wages we aim to explain in this paper.

3.1 Decision Nodes

We consider 5 time periods where workers make educational and labor market decisions. We refer to these time points as nodes. They are: beginning of high school, end of high school, beginning of college, end of college, and the labor market.

In Node 1, the beginning of high school, students choose what type of specialized classes, if any, they want to attend. There are five options, advanced placement (AP) STEM classes, AP non-STEM classes, career and technical education (CTE) STEM classes, and CTE non-STEM classes, or no specialized classes. AP classes are targeted toward students who plan to continue

their studies in college, while CTE classes teach specific career skills to high school students. We classify these classes into STEM and non-STEM.

One challenge is that not all high schools offer these specialized classes. To address this issue, we use data from cohorts of high school students in the years before our period of study to identify which schools offered these courses. We then restrict our analysis to these schools.

In node 2, the end of high school, students decide whether to attend a two-year, a four-year college, or not go to a college. We do not distinguish between students who finish high school but do not go to college, and those who drop out of high school because the latter category is very small.

Node 3 and 4 describe decisions of students who chose to go to college. In node 3, beginning of college, students choose one of 10 majors, while in Node 4, end of college, their choice set includes dropping out, graduating and going to the labor market, or graduating and continuing further studies. Students who did not go to college do not make any decisions in Nodes 3 and 4.

Finally, in Node 5, students enter the labor market and choose one of 9 industries. If an individual does not get a job within the year of graduating, we use the industry of the first job they are employed in. If individuals hold multiple jobs spanning across industries, we use the industry corresponding to their primary job, that is the job where they devote the majority of their work hours.

3.2 Observable States

The decision of student i in node n is affected by observable state vector X_{in}, $n = 1,\ldots N$. At node $n = 1$, the vector contains five state variables: student cognitive skills, non-cognitive skills, socio-economic status, cultural background, as well as as a measure of quality of student's high school.

We use math test score from middle or elementary school as a proxy for cognitive ability. In particular, within each cohort, we sort students into quartiles using their eighth-grade math scores and use a vector of dummy variables representing this assignment. We proceed similarly for constructing a proxy for non-cognitive skills, using absenteeism instead of math test scores. If eighth grade math score or absenteeism measure is missing, we use other middle or elementary school scores.

We use eligibility for FRPL program as a proxy for socio-economic status and language spoken at home as a proxy for cultural background. There are over 200 different languages represented in our dataset and we categorize them into six language categories: English, Hmong, other Asian, Somali, Spanish, and all other languages.

Finally, we use Value Added Model (VAM) to obtain the distribution of school quality and

Empirical Bayes Method to improve the estimates for each school. The ACT scores of students are used to obtain the VAM estimates. We sort high schools into quartiles based on the value added and use the corresponding dummy variables in X. Appendix A provides further details of how we construct school quality measurs.

For $n > 1$, the observable state X_{in} is constructed as $X_{in} = [X_{in-1}, D_{in-1}]$, where D_{in} encodes decision of student i at node n. That is, we use vector of fixed characterstics X_{i1} and all decisions in previous nodes as a state variable in node n.

At node 3, we add one additional state variable which is the college quality. We observe the name of the college in the data and link it to average math test score of admitted students, the data that are collected by Integrated Post-secondary Education Data System (IPEDS). This data is publicly available. We use this to classify colleges into high-quality (above median) and low-quality (below median).

Table 1: **Summary Statistics: Demographic Characteristics**

Demographic Characteristic		N	Percent Share
Gender	Female	85,516	51
	Male	80,956	49
Race/Ethnicity	White	132,210	79
	Black	13,718	8
	Hispanic	8,942	6
	Asian	11,602	7
Total N		166,472	100

3.3 Descriptive Statistics

There are 166,472 unique individuals born between 1991 and 1996 for whom we observe high school decisions, college decisions, and labor market outcomes. These individuals were in Grade 12 in high school between years 2010 and 2014. Table 1 shows the gender and racial/ethnic composition of our sample: there are 51% females and 49% males. The racial and ethnic composition is as follows: 79% whites, 8% blacks, 6% Hispanic, 7% Asians.

Table 2 shows the summary statistics for observable characteristics by gender and race in our sample. Note that the ability variable is calculated based on test sccores of the entire student population in Minnesota within a cohort, not just the test scores of individuals in our sample. We see that 58% of females and 56% of males in our sample score above median in Grade 8 Maths test in the Minnesota Comprehensive Assessments (MCA) exams.

Table 2: **Observable Characteristics by Gender and Race in Our Sample**

	Females	Males	Whites	Blacks	Hispanics	Asians
Cognitive ability Q4	0.28	0.29	0.33	0.08	0.13	0.20
Q3	0.28	0.29	0.31	0.16	0.21	0.23
Q2	0.25	0.28	0.24	0.28	0.30	0.28
Q1	0.19	0.17	0.13	0.49	0.37	0.29
Non-cognitive ability Q4	0.29	0.28	0.28	0.23	0.22	0.40
Q3	0.28	0.28	0.30	0.22	0.24	0.22
Q2	0.25	0.26	0.27	0.22	0.25	0.18
Q1	0.18	0.18	0.16	0.32	0.29	0.20
School quality Q4	0.26	0.25	0.28	0.13	0.16	0.19
Q3	0.24	0.25	0.27	0.15	0.20	0.15
Q2	0.25	0.25	0.25	0.23	0.22	0.20
Q1	0.25	0.25	0.20	0.49	0.42	0.46
Poor	0.36	0.37	0.27	0.84	0.70	0.71
Non-English speakers	0.10	0.10	0.01	0.21	0.52	0.74
High Quality College	0.26	0.16	0.17	0.09	0.09	0.22

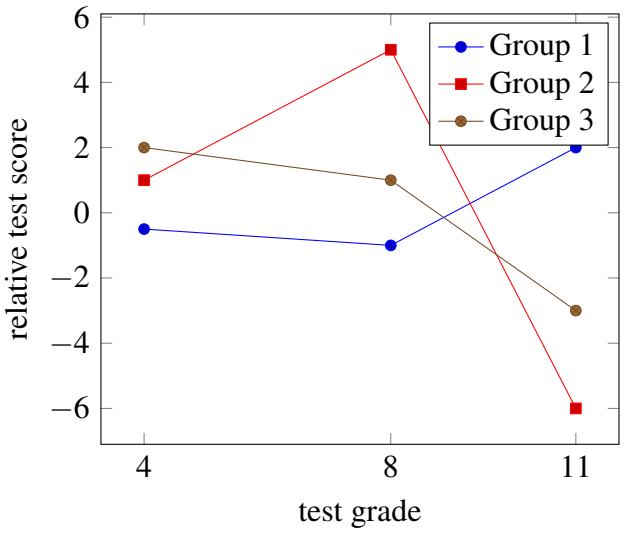
Students, both male and female, are almost uniformly distributed across high schools with different quality, with around 25% of students in each quality quartile. There are however differences across race - 28% of white students is in the highest quality schools compared to just 13% of blacks; conversely, almost 50% of the black students are enrolled in the low-quality schools compared to 20% of white students. Within the non-White group, the average quality of high schools Blacks attend is the lowest. A similar pattern emerges for proportion of students who are poor - male and female proportions are similar at 36% and 37% respectively while the difference between Whites and non-Whites is large. While 27% of Whites are eligible for free and reduced price lunch, the corresponding numbers for Blacks, Hispanics, and Asians are 84%, 70% and 71%.

3.4 Unobserved Productivity Types

We use Math test scores for individuals in Grades 4, 8, and 11 and group them based on the evolution of these test scores. We control for age, demographics, FRPL status, homelessness indicator, migrant indicator and language spoken at home. We choose $G = 3$, that is, there are three different productivity types. We then use group fixed effects described earlier to estimate assignment of workers into groups.

Figure 1 shows the evolution of the common component of the test score evolution for the three estimated types. For group 1 we observe that, on average, the test scores are relatively stable between elementary and middle schools and then increase sharply in high school. We refer to this

Figure 1: Evolution of Test Scores for Three Estimated Types



group as the 'increasing' productivity type. For group 2, the test scores increase in middle school and then decline sharply in high school. This group will be referred to as the 'hump shape' type. Test scores for Group 3 decrease over time and they are thus referred to as the 'decreasing' type.

Table 3 shows the distribution of types by race and gender. The share of males and females that of decreasing type are very similar at 13%. The difference comes from the fact that males are more likely to be of increasing type (48%) while women are more likely to be the hump shaped type (46%). In contrast, for White and Black individuals, the likelihood of being a hump shaped type is similar at 44% and 43% respectively. Relatively, Blacks are more likely to be the decreasing type while Whites are more likely to be increasing type.

For the rest of the paper, we focus our analysis on four demographic groups: females, males, Black individuals, and White individuals. In the next section, we present the results of the node-by-node multinomial logit estimations for each of these groups.

4 Estimation Results: Choices and Transitions

4.1 Node 1: choice of courses

At node $n = 1$, the students can choose whether to take advanced courses or not and which types of courses to take. More specifically, the choices individuals can make are $d_1 \in \{$AP STEM, AP non-STEM, CTE STEM, CTE non-STEM, no advanced course$\}$. The state variables s_1 are factors

Table 3: **Unobserved Productivity Types and Demographics**

Characteristic	Unobserved types		
	Increasing	Hump Shape	Decreasing
Female	40.9	46.3	12.8
Male	48.2	39.2	12.6
White	45.3	43.9	10.8
Black	37.9	42.7	19.4
Hispanic	40.9	39.4	19.7
Asian	44.5	33.4	22.1

that can affect the students' choice of courses: cognitive ability, non-cognitive ability, productivity type, school quality, cultural background, and socioeconomic status.

Figure 2 shows the results of the multinomial logit regression for node 1 decisions by gender. The left panel presents results for females and the right panel for males. The base outcome is the decision of individuals to not take any advanced courses in high school.[1] The results show that individuals with higher test scores are more likely to choose AP courses and less likely to take CTE courses compared to individuals with lower test scores. The log odds for high test score males choosing AP STEM courses is higher than log odds for high test score females. While individuals of higher non-cognitive ability are more likely to take all types of advanced courses, females of higher types are significantly more likely to take AP STEM compared to males.

The role of school quality,socioeconomic status, productivity type, and cultural background is similar across gender. Individuals enrolled in FRPL programs are significantly less likely to take AP courses. There is no significant difference in log odds of taking CTE courses compared to taking no advanced courses. Individuals of lower productive types are less likely to take any of the advanced courses with log odds being significantly lower for AP courses. Interestingly, there is a U-shaped relationship between school quality and log odds of choosing advanced courses. This is more pronounced for choice of AP courses.

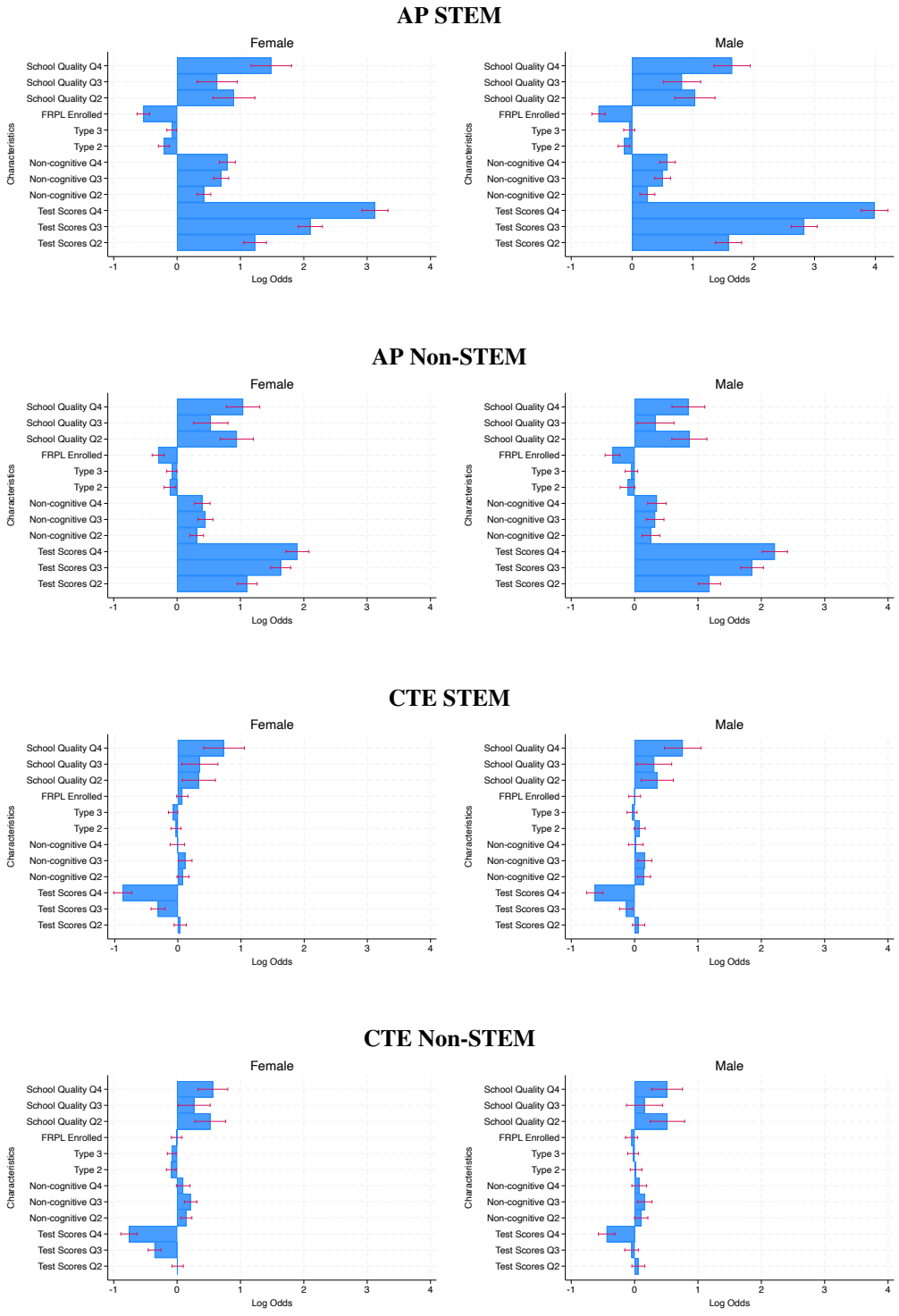
Table 3 shows the results of the multinomial logit regression for node 1 decisions by race. We suppress the estimates of cultural factors for brevity. The results show that the students with higher

[1]The coefficients in mlogit model are interpreted as follows: for a unit increase in the predictor state variable, the multinomial log-odds for decision d_i relative to the base outcome will change by the estimated amount while holding all other variables in the model constant.

test scores have higher odds of taking AP courses. The magnitudes are higher for AP STEM courses compared to AP non-STEM. Black students with high test scores have the lowest odds of taking AP courses. On the other hand, Blacks with high non-cognitive ability have a higher likelihood of taking AP courses compared to White students.

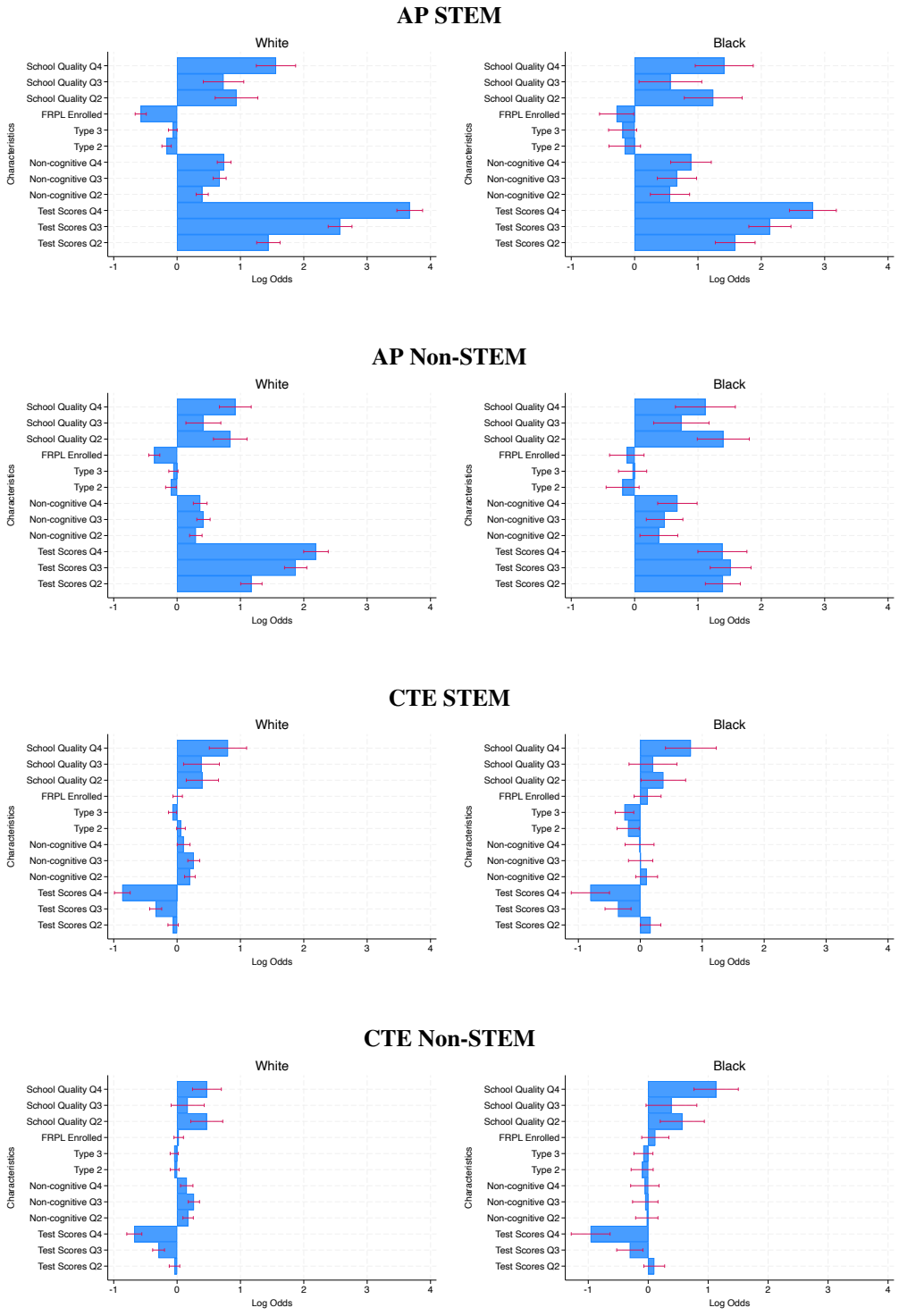
The odds of poorer students taking any AP courses is lower than not taking any advanced courses. We see a U-shaped pattern across school quality and odds of taking advanced courses for both the racial groups. Conditional on all other factors, the odds of choosing advanced courses is no different than not taking an advanced courses for low productivity types.

Figure 2: **Node 1 Decision: Mlogit results by Gender**



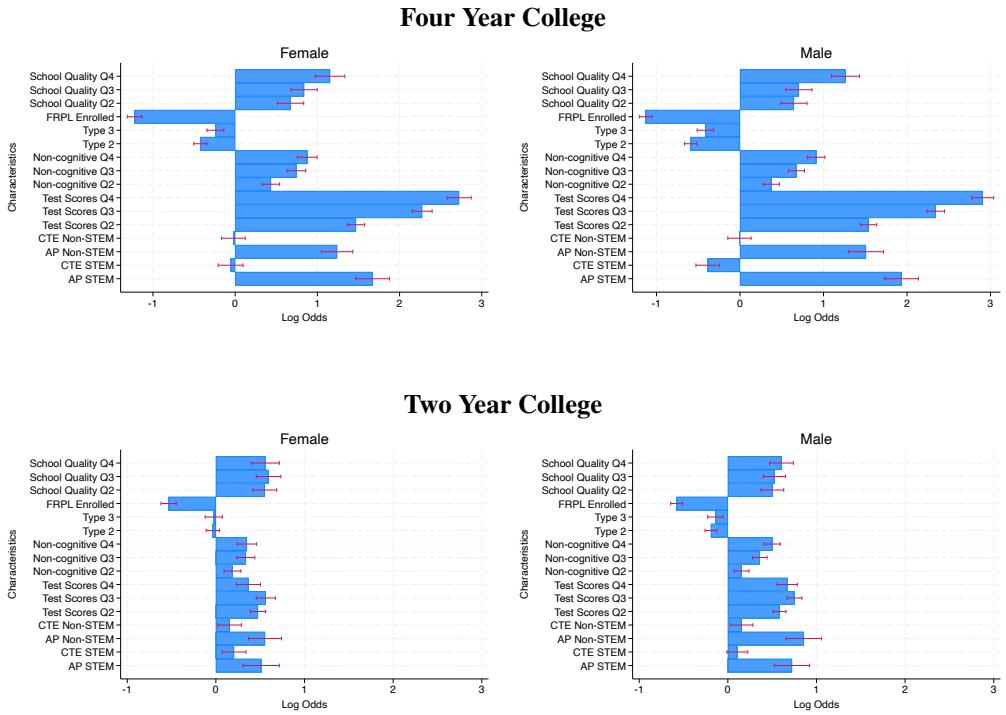
Note. The base outcome: no advanced courses are taken. Other controls not shown in this figure include language spoken at home and cohort fixed effects.

Figure 3: **Node 1 Decision: Mlogit results by Race and Ethnicity**



Note. The base outcome: no advanced courses are taken. Other controls not shown in this figure include language spoken at home and cohort fixed effects.

Figure 4: **Node 2 Decision: Mlogit results by Gender**



Note. The base outcome is 'no college'. Other controls not shown in this figure include language spoken at home and cohort fixed effects.

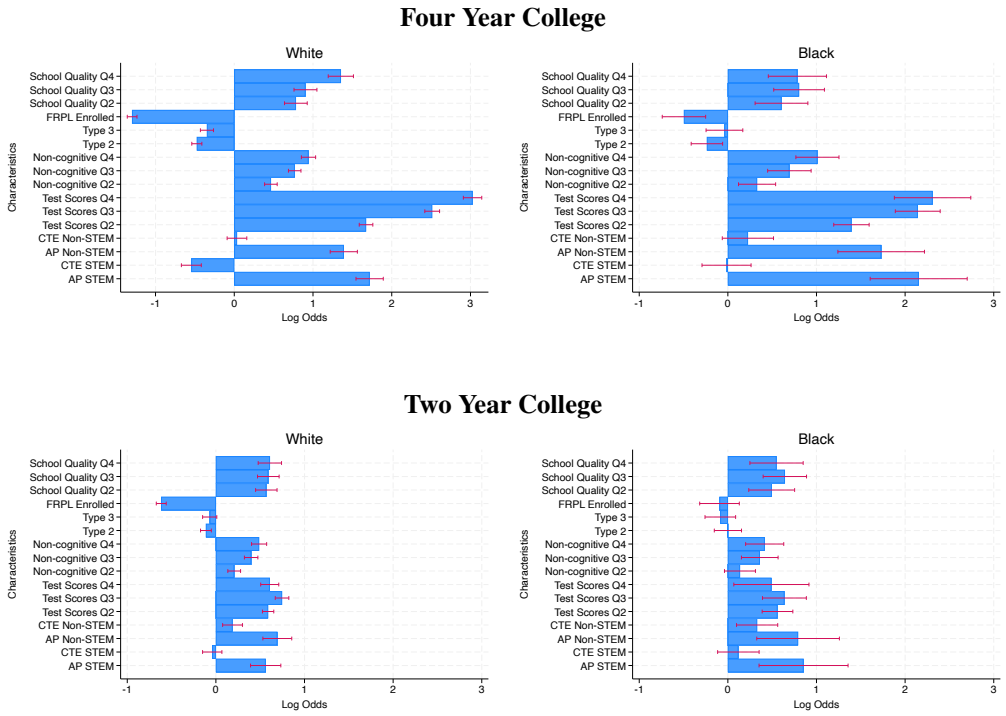
4.2 Node 2: high school-college transition decision

At node $n = 2$, the students can choose to enroll into a two-year college, enroll in four year college, or not to go to college at all. So, $d_2 \in \{\text{Two-year college, Four-year college, No college}\}$. The state variables s_2 consists of all state variables in s_0 and s_1 as well as d_1. In the estimation, we consider no college as the base outcome such that the estimated coefficients can be interpreted as log-odds compared to the no college option.

Figure 4 shows the results of multinomial logit estimation by gender. Having taken AP courses, the odds of females enrolling into four year or two year college is significantly lower compared to males. Males and females who take CTE courses are significantly less likely to go to four year college. Test scores, non-cognitive ability, school quality, productivity, and socioeconomic status play a similar role across gender in determining the odds of enrolling into two year or four year college.

Figure 5 shows the estimation results by race. Among students who are in the highest test score quartile and those from the best quality schools, Black students have lower odds of enrolling in four year colleges. In contrast, for the students in the highest quartile of non-cognitive ability, Black

Figure 5: **Node 2 Decision: Mlogit results by Race and Ethnicity**



Note. The base outcome is 'no college'. Other controls not shown in this figure include language spoken at home and cohort fixed effects.

students are have the highest odds of applying to four year colleges. Also, having taken AP courses, Blacks have higher odds of applying to four year colleges. Black students from disadvantaged families are more likely to apply to four year college compared to their White counterparts.

4.3 Node 3: major choice

At node $n = 3$, the students who enroll into a two-year or four-year college choose a major. Those who chose not to go to college at Node 2 do not choose a major and we label them 'not applicable.' The state variables $\boldsymbol{s}_2$ consists of all state variables in s_0, s_1, s_2 as well as d_2. In the estimation, we consider the Liberal Arts and Humanities majors as the base outcome. We combine all majors into 10 categories based on CIP codes. Showing results for each major is too cumbersome, so we discuss the main results for Engineering and Education.

The most important factor that increases the likelihood of females choosing engineering as a major, compared to liberal arts and humanities, is whether they are in the top quartile of the test score distribution. This factor is important for males as well but the magnitude is significantly smaller. While likelihood of choosing engineering is higher for all individuals who take AP STEM,

the role of CTE STEM is more important for males compared to females.

White and Black students who take AP STEM are more than twice as likely to take Engineering than Liberal Arts and Humanities. Among students taking CTE STEM in high school, Black students have lower likelihood of choosing Engineering over Liberal Arts majors. For most students who enroll into a two year college, the likelihood of choosing engineering majors is lower than choosing liberal arts and humanities.

The patterns we observe for Engineering also hold for biological and physical sciences. The main difference there is that Black students who are above median in the test score distribution are significantly more likely to choose Science than Liberal Arts. While they are about one and a half times more likely to choose engineering compared to Liberal Arts, the relative likelihood of choosing Science majors is close to three times.

For all groups, individuals with higher test scores have a lower likelihood of choosing Education as a major compared to Liberal Arts and Humanities as their major. This is especially pronounced for Black students in the highest test score quartile. For students enrolled in a lower quality college, Females and Whites are twice as likely to choose Education than Liberal Arts and Humanities. Students enrolled in FRPL programs during K-12 are equally likely to choose either of these majors.

4.4 Node 4: college graduation

At node $n = 4$, the students who enroll into a two-year or four-year college and choose a major decide whether to drop out, to graduate and go to the labor market, or to graduate and go for higher studies (masters or higher). Those who chose not to go to college at Node 2 and do not choose a major on node 3, we label them 'not applicable' here as well. The state variables $\boldsymbol{s}_2$ consists of all state variables in s_0, s_1, s_2, s_3 as well as d_3. In the estimation, we consider 'did not complete college' as the base outcome.

The likelihood of graduating and going directly into the labor market is higher than dropping out. Moreover, everything else constant, if a student is of high ability and in a high quality college, they are twice as likely to to graduate than not complete college. These patterns hold for both males and females as well as for racial groups.

The relative likelihood of graduating varies across majors. For all demographic groups, students enrolled in Health Professional majors are about four times more likely to graduate than to drop out of college. Students from Social Sciences and Business majors also have a high relative likelihood of graduating. Black students from Engineering have the lowest relative likelihood of graduating compared to both other demographic groups in Engineering and to any group in any other major.

Students are more likely to drop out of college if college quality is below median or if they are from disadvantaged socioeconomic background. These patterns hold for both gender and racial groups. Students who took CTE courses in high school are less likely to graduate college.

4.5 Summary

What we have found so far is that, conditional on observable characteristics and unobserved types, the observed decisions differ for females and minorities. Moreover, we saw that early choices predict future decisions. In the next section, we examine how much of the wage gaps can be explained by the characteristics of individuals before they reach high school and how much is driven by the decision path. We use the model estimated in this section to analyze these factors affecting the gender wage gap.

5 Factors Affecting Initial Wage Gap

5.1 Decomposition of Wage Gap

We start with the wage gap decomposition. The goal of this exercise is to understand how the decisions that students make affect the wage gap.

We want to distinguish between endowments, that is, the observable characteristics at node 1 X_1, and the decisions the decisions that students make along the way. Recall that all observable characteristics were discrete, and so we can index the vector of them by $x = 1, \dots X$. We refer to the sequence of decisions in nodes 1–5 as a path, and index it by k, $k = 1, \dots K$.

Consider a group of students indexed by $i = 1, \dots N$, and denote log wage of student i as w_i. Each student is characterized by a combination of endowments with index x and path with index k. We let Z_x denote the set of students with characteristics x, and $Z_{x,k}$ the set of students with characteristic x who chose path k. Then the average log wage is

$$\begin{aligned}
\bar{w} &= \frac{1}{N} \sum_{x=1} \sum_{k=1}^{K} \sum_{i \in Z_{x,k}} w_i \\
&= \sum_{x=1} \frac{N_x}{N} \sum_{k=1}^{K} \frac{N_{x,k}}{N_x} \frac{1}{N_{x,k}} \sum_{i \in Z_{x,k}} w_i \\
&= \sum_{x=1} s_x \sum_{k=1}^{K} s_{x,k} \bar{w}_{x,k}
\end{aligned}$$

where $s_x = \frac{N_x}{N}$, $s_{x,k} = \frac{N_{x,k}}{N_x}$ and $\bar{w}_{x,k} = \frac{1}{N_{x,k}} \sum_{i \in Z_{x,k}} w_i$ is the average wage.

Using upper script a and b on all these objects to denote two different groups a and b, for example, males and females, we can express the wage gap as

$$WG = \bar{w}^a - \bar{w}^b,$$

and decompose it into three terms:

$$\begin{aligned} WG = & \sum_x \frac{s^a_x + s^b_x}{2} \sum_k \frac{s^a_{xk} + s^b_{xk}}{2} (\bar{w}^a_{x,k} - \bar{w}^b_{x,k}) \\ & + \sum_x \frac{s^a_x + s^b_x}{2} \sum_k \frac{w^a_{xk} + w^b_{xk}}{2} (\bar{s}^a_{x,k} - \bar{s}^b_{x,k}) \\ & + \sum_x \frac{m^a_x + m^b_x}{2} (\bar{s}^a_x - \bar{s}^b_x), \end{aligned}$$

where $m^a_x = \sum_k s^a_{x,k} \bar{w}^a_{x,k}$. The first term captures the within-path wage gap: it is a weighted average of wage gaps on path k conditional on characteristics x, $\bar{w}^a_{x,k} - \bar{w}^b_{x,k}$. The second term is the between path gap: it is a weighted average of differences in shares on path k, conditional on characteristics x, $s^a_{x,k} - s^b_{x,k}$. The third is is the differences in share of students with given characteristics x.

We view the second term, the across-path wage conditional x, as the one that a student can affect by his or her decisions at different nodes.

Our estimated model implies that the total gender wage gap is 4.6 percentage points (pp). The within path gap is 2.0 pp, across path 1.9 pp and finally the differences in endowments is 0.7 pp. In other words, if females made different decisions, they could close 41% of the wage gap.

The decomposition for black-white gap is different. The gap is 29.6 pp, and the decomposition is 15.6 pp within-path, 3.1 pp across-path and 10.9 pp differences in endowments. This means that by changing their decisions, they can only close just about 10% of the wage gap.

5.2 Counterfactual Analysis

We now present the results from our counterfactual exercises. In panel [b] of Table 4, we first examine how the aggregate gender and White-Black gap changes when we equalize the distribution of one endowment at a time between the analyzed group and the benchmark group. We find that each of the endowments has a small contribution towards reduction in the wage gap. For gender gap, unobserved productivity types closes the gap with 1.2 percentage points. For White-Black gap, changing the cognitive ability distribution closes the gap the most. Changing all endowments together closes the gap with a magnitude that is consistent with the our decomposition exercise. We find that the wage gap reduces by 1.5 percentage points for gender wage gap and 9.2 percentage

Table 4: Counterfactual Analysis: Changing Pre-labor Market Factors

		Hourly Wage Gap	
		Gender	White-Black
[a] Benchmark Gaps		4.7	29.9
[b] Changing Endowments	Cognitive	3.8	27.1
	Non-cognitive	3.6	29.4
	High-school quality	3.6	28.8
	FRLP eligibility	3.4	27.8
	Language	3.6	29.7
	Unobserved types	3.5	29.6
	All endowments	3.2	20.4
[c] Changing Decisions	High school courses	4.6	29.7
	College decision	4.9	29.9
	College major	2.3	27.3
	College completion	5.9	32.1
	Industry	4.7	29.3
	All Decisions	3.8	29.4

points for the black-white wage gap.

Next, we conduct a counterfactual exercise where we modify behaviour of one group in a specific node. We keep the behaviour in the other nodes and the distribution of endowments unchanged. The results are shown in panel [c] of Table 4. If we modify female behaviour in Node 1 (high school advanced courses) to mimic that of males, the gender gap would be 4.6 pp, so barely affected. If females made the same decisions as males in Node 3 (choosing the college majors), then the gender gap would be 2.3 pp, so a significant reduction. The other nodes do not reduce the gaps, and in fact the gap increases if females make the same college going and college completion choices as men. Note that changing just the major choices for females closes the gender wage gap even more compared to changing all endowments at once.

The second column of panel [c] shows the results for black-white wage gap. Even though the magnitudes differ from those presented for gender wage gap, the patterns are similar. We again find that Node 3 is the most critical. However, in contrast to gender gap, there is no reduction in racial wage gap if we change all the decision nodes together.

5.3 Examining the Within-Path Gaps: Role of Firms

The results of our decomposition exercise showed that the within-path gap plays an important role for both gender and racial wage gaps. This decomposition exercise takes as given the wage equation. In other words, it takes as given the log wage as a function of student's observable characteristics, types and decisions. A rich labor market literature argues that firm's productivity,

Table 5: Role of Firms in Within-Path Gaps

Groups	High Skill Path		Low Skill Path	
	Wage Gap	Wage Gap with Firm FE	Wage Gap	Wage Gap with Firm FE
	(a)	(b)	(c)	(d)
Male - Female	16.5	9.8	12.7	3.9
White - Black	13.0	11.5	15.5	5.2

typically unobserved, explains an important part of the wage variation. To understand how much this plays a role in our specification, we estimate the wage equations again by adding firm specific fixed effects and then proceed as before.

We focus on two extreme paths to illustrate our main result. We define 'high skill' path as one where individuals were in the top quartile of the cognitive ability distribution, took STEM courses in high school, and took a STEM major in college. The 'low skilled' path is defined as one where individuals were below median in the cognitive ability distribution and did not go to college.

Table 5 presents the results of our analysis. Column (a) shows the wage gap for individuals on the high skill path. Compared to the entire sample where gender gap was 4.7%, we find that the gender gap on the high skill path is significantly higher at 16.5%. In contrast, the White-Black gap is significantly lower at 13%. Column (b) presents the wage gap on this path when we control for firm fixed effects. The gender gap reduces, but there remains a gap of about 10 percentage points. For White-Black gap, there is very little change in the wage gap. These results suggest that majority of the variation in gender gap and almost all of the racial gap in wages for high skill path come from within-firm differences across these groups.

The results for low skilled path tell a different story. Column (c) shows the wage gap on this path, which is again higher than the aggregate gap gender and lower than the aggregate for White-Black differences. Unlike the case of high skilled path, in this case the gaps become substantially lower for both gender and race when we control for firm fixed effects. This suggests that most of the gaps for the unskilled path are driven by across-firm differences across these groups. In other words, females and minority individuals are more likely to work for lower wage firms while males and Whites are more likely to work for firms that pay higher for the same skill type.

Why are low skill females and minorities more likely to work in firms that pay lower wages on average? To answer this question, we now classify firms into types based on two characteristics: where the firms are in the wage distribution and how much work flexibility the firms offer to their

Table 6: Distribution of Workers Across Firm Type

Firm Type (Wage-flexibility)	Proportion of Individuals			
	Female (a)	Male (b)	White (c)	Black (d)
High - High	10.16	10.69	11.04	11.13
High -Low	33.41	42.61	39.75	32.18
Low - High	31.74	26.28	26.94	34.61
Low - Low	24.70	20.43	22.27	22.07

employees. Our measure of work flexibility is based on fraction of workers that are employed part-time. We first calculate the average wage and share of workers working part-time for each firm. We then compare these two numbers to the average in the industry firm is in. If the firm has higher wages than industry median wage, then we classify that firm as a high-wage firm. Similarly, if a firm has a share of part-time workers that is higher than the industry average, we classify that firm as a high flexibility firm.

We get four types of firms on the basis of wage-flexibility classifications defined above: high-high, high-low, low-high, and low-low. Table 6 documents the distribution of workers across firm types. Comparing columns (a) and (b), we find that the proportion of males and females in high-high type firm is very similar. The main difference comes from females being more likely to match with low-high firms and less likely to match with high-low firms. In columns (c) and (d), we see exactly the same patterns and they are even starker: White and Black individuals are equally likely to be in high-high firms (11%) and low-low firms (22%). The difference comes from Blacks sorting into low wage firms with more flexibility.

To summarize: we find that, for high skilled paths, most of the variation in wages across groups comes from within-firm differences. For low skilled paths, almost all of the variation comes from females and Black individuals sorting into firms with lower wages but larger share of part-time jobs.

6 Conclusion

This paper examines the early career differences across gender and race in hourly wages. Using a dynamic discrete choice model and a rich longitudinal administrative data from Minnesota, we

find that pre-labor market factors explain about half of the wage gaps. We find that 41% of the gender gap can be explained by women being on paths with lower hourly wages. In contract, this factor contributes only 10% of Black-White wage gap while differences in initial states across these groups explain close to 40% of the gap.

Our results suggest that pre-labor market interventions that encourage students to take more quantitative college majors have a large potential in reducing gender gaps while having little effect on racial wage gaps. Instead, education interventions in schools before individuals reach high school can be most effective in reducing White-Black wage gap. In future work, we will further analyze these factors and assess which type of policy instruments can be most effective in reducing these gaps.

References

ALTONJI, J. G. AND R. M. BLANK (1999): "Race and gender in the labor market," *Handbook of labor economics*, 3, 3143–3259.

ARCIDIACONO, P. (2004): "Ability sorting and the returns to college major," *Journal of econometrics*, 121, 343–375.

ARCIDIACONO, P., E. M. AUCEJO, AND V. J. HOTZ (2016): "University differences in the graduation of minorities in STEM fields: Evidence from California," *American Economic Review*, 106, 525–562.

ARCIDIACONO, P. AND R. A. MILLER (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79, 1823–1867.

BABCOCK, L. AND S. LASCHEVER (2003): *Women don't ask: Negotiation and the gender divide*, Princeton University Press.

BLACK, D., A. HAVILAND, S. SANDERS, AND L. TAYLOR (2006): "Why do minority men earn less? A study of wage differentials among the highly educated," *The Review of Economics and Statistics*, 88, 300–313.

BLAU, F. D. AND L. M. KAHN (2017): "The gender wage gap: Extent, trends, and explanations," *Journal of economic literature*, 55, 789–865.

BONHOMME, S. AND E. MANRESA (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83, 1147–1184.

CAMERON, S. V. AND J. J. HECKMAN (2001): "The dynamics of educational attainment for black, hispanic, and white males," *Journal of political Economy*, 109, 455–499.

CARNEIRO, P., J. J. HECKMAN, AND D. V. MASTEROV (2005): "Labor market discrimination and racial differences in premarket factors," *The Journal of Law and Economics*, 48, 1–39.

EROSA, A., L. FUSTER, G. KAMBOUROV, AND R. ROGERSON (2022): "Hours, occupations, and gender differences in labor market outcomes," *American Economic Journal: Macroeconomics*, 14, 543–90.

FRYER, R. (2011): "Racial inequality in the 21st century: The declining significance of discrimination," in *Handbook of labor economics*, Elsevier, vol. 4, 855–971.

GOLDIN, C. (2014): "A grand gender convergence: Its last chapter," *American Economic Review*, 104, 1091–1119.

GOLDIN, C., S. P. KERR, C. OLIVETTI, AND E. BARTH (2017): "The expanding gender earnings gap: Evidence from the LEHD-2000 Census," *American Economic Review*, 107, 110–114.

HOEKSTRA, M. (2009): "The effect of attending the flagship state university on earnings: A discontinuity-based approach," *The review of economics and statistics*, 91, 717–724.

HOTZ, V. J. AND R. A. MILLER (1993): "Conditional choice probabilities and the estimation of dynamic models," *The Review of Economic Studies*, 60, 497–529.

HUGGETT, M., G. VENTURA, AND A. YARON (2006): "Human capital and earnings distribution dynamics," *Journal of Monetary Economics*, 53, 265–290.

——— (2011): "Sources of lifetime inequality," *American Economic Review*, 101, 2923–2954.

JOENSEN, J. S. AND H. S. NIELSEN (2009): "Is there a causal effect of high school math on labor market outcomes?" *Journal of Human Resources*, 44, 171–198.

LANG, K. AND M. MANOVE (2011): "Education and labor market discrimination," *American Economic Review*, 101, 1467–1496.

LAZEAR, E. P. AND S. ROSEN (1990): "Male-female wage differentials in job ladders," *Journal of Labor Economics*, 8, S106–S123.

NEAL, D. A. AND W. R. JOHNSON (1996): "The role of premarket factors in black-white wage differences," *Journal of political Economy*, 104, 869–895.

O'NEILL, J. (1990): "The role of human capital in earnings differences between black and white men," *Journal of economic Perspectives*, 4, 25–45.

RAUH, C. AND A. VALLADARES-ESTEBAN (2018): "Wage and employment gaps over the lifecycle–the case of black and white males in the US," .

RODGERS III, W. M. AND W. E. SPRIGGS (1996): "What does the AFQT really measure: Race, wages, schooling and the AFQT score," *The Review of Black Political Economy*, 24, 13–46.

Appendix

A. Measure of School Quality

In this section, we describe the measure of school quality used in our analysis. Since we have observational data from all school districts in Minnesota, where none of the districts use a centralized school assignment or fully use a lottery based system for seats, we use a Value Added Model (VAM) to obtain the distribution of school quality and Empirical Bayes Method to improve the estimates for each school. Deming (2014) finds that VAMs have accurate out-of-sample prediction of student achievement and he fails to reject the hypothesis that school effects estimated using VAM are unbiased.

Consider a population of students indexed by i, each attending one of J schools in a district in Minnesota. Let $Y_{i(j)}$ denote student i's potential academic achievement if she attends school $j \in \{1,...,J\}$. Let D_{ij} indicate attendance at school j and X_i denote a vector of covariates. The observed outcome is:

$$Y_i = \sum_j \alpha_j D_{ij} + \varepsilon_i \tag{1}$$

where, α_j is the value-added of school j and ε_i is the unobserved student heterogeneity such that $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The identifying assumption is that X_i captures all selection bias, such that $E[D_{ij}\varepsilon_i] = 0 \; \forall j$. This allows us to use ordinary least squares to estimate value added for each school $\hat{\alpha}_j$ and its standard error s_j.

The estimates obtained above can have substantial sampling errors for small schools with relatively few student observations. We use Empirical Bayes method that uses the full set of VAM estimates $\hat{\alpha}_j$ to reduce sampling variance in individual school quality estimates. Suppose we have the following hierarchical model:

$$\hat{\alpha}_j | \alpha_j, s_j \sim \mathcal{N}(\alpha_j, s_j^2)$$

$$\alpha_j \sim G(\alpha), \qquad j = 1,...,J$$

where G is a mixing distribution. We will first use $\{\hat{\alpha}_j, s_j\}_{j=1}^J$ to get an estimate of the mixing distribution $\hat{G}$. We will then treat $\hat{G}$ as a prior and update $\{\hat{\alpha}_j, s_j\}_{j=1}^J$ to get the posterior estimate of our school value added measure $\{\hat{\alpha}_j^*\}_{j=1}^J$.

Under the parametric assumption that $\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$, the posterior value added measure for school j is

$$\hat{\alpha}_j^{PEB} = \hat{\alpha}_j \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + s_j^2} + \hat{\mu}_\alpha \frac{s_j^2}{\hat{\sigma}_\alpha^2 + s_j^2};$$

where, $\hat{\mu}_\alpha$ and $\hat{\sigma}_\alpha^2$ can be obtained using method of moments: $\hat{\mu}_\alpha = \frac{1}{J}\sum_{j=1}^J \hat{\alpha}_j$ and $\hat{\sigma}_\alpha^2 = \frac{1}{J}\sum_{j=1}^J [(\hat{\alpha}_j - \hat{\mu}_\alpha)^2 - s_j^2]$.

Imposing normality assumption on G can lead to specification error. To avoid this, we obtain nonparametric Bayes estimator for school value added using methodology similar to Gilraine, Gu, and McMillan (2020). The distribution G can be estimated nonparametrically from the data via the following optimization:

$$\hat{G} \equiv \underset{G \in \mathbb{G}}{\operatorname{argmax}} \left\{ \sum_{j=1}^J log \int \varphi_j(y_j - \alpha) dG(\alpha) \right\} \tag{2}$$

where φ_j is a normal density with mean zero and variance $\sigma_\varepsilon^2 / n_j$. The state space $\mathbb{G}$ is the set of all probability distributions in $\mathbb{R}$. We follow Koenker and Mizera (2014) by estimating $\hat{G}$ from a finite-dimension convex approximation of the above problem. Let M be a positive integer and let $\mathbb{G}_M$ be the class of probability distribution functions supported on M grid points given by $min_j\{\hat{\alpha}_j\} < \alpha_1 < \alpha_2 < ... < \alpha_M < max_j\{\hat{\alpha}_j\}$. The nonparametric maximum likelihood estimator is then defined to be the maximizer of equation 2 when $\mathbb{G}$ is replaced by $\mathbb{G}_M$. We take M to be the square-root

of the sample size.

We use $\hat{G}$ estimated from the above exercise to obtain the nonparametric empirical Bayes estimator for the school value added:

$$\hat{\alpha}_j^{NPEB} = \frac{\int \alpha \varphi_j(y_j - \alpha) d\hat{G}(\alpha)}{\int \varphi_j(y_j - \alpha) d\hat{G}(\alpha)}.$$