How Early Skill Gaps Contribute to Gender Differences in STEM Major and Occupation Choices: A Random Forest Approach*

Petra E. Todd † Shasha Wang ‡

September 2022, this version January 2023

Abstract

In the US, women go to college at higher rates than men, but they are less likely to choose applied-STEM college majors or occupations. Using the NLSY79 and 97 datasets, this paper assesses the importance of adolescent skill profiles and high school course-taking in explaining gender disparities in four-year college completion, college major, and occupational choices. It considers five cognitive skill measures (math, verbal, science, administrative, and mechanical) and one non-cognitive measure and examines gender skill convergence over a twenty-year time span. Results show that high-school-aged women in the NLSY97 cohort reached parity with men, on average, in mathematics skills and exceed men in verbal and noncognitive skills, but they lag behind in mechanical and, to a lesser extent, science skills. To identify the skill sets, course-taking, and family background characteristics that best predict entry into STEM majors and occupations, we estimate logistic and nonparametric random forest models. The estimates reveal that a combination of mathematics and mechanical skills along with intensive high school exposure to science and math courses are key predictors of choosing STEM majors and careers. A nonparametric decomposition approach is developed and used to quantify how eliminating gender skill disparities would affect entry into STEM fields.

*Petra Todd is Edmund J. and Louise W. Kahn Term Professor of Economics at the University of Pennsylvania and a member of the U Penn Population Studies Group, NBER, HCEO, and IZA. Shasha Wang is an economics graduate student at the University of Pennsylvania. This paper was presented at U Penn, the Southern Economic Association meetings (Fall, 2022), and the 2023 AEA meetings. We thank Amanda Chuan for her helpful comments. This research was supported by a pilot grant from the Population Aging Research Center at the University of Pennsylvania which is funded by the NIH National Institute on Aging under award number AG012836. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

†Petra E. Todd, Department of Economics, University of Pennsylvania; Email: ptodd@econ.upenn.edu

‡Shasha Wang, Department of Economics, University of Pennsylvania; Email: wss@sas.upenn.edu

1 Introduction

The gender earnings gap in the US decreased in the last forty years, but a gap of approximately 18% remains (Goldin 2014). Even though more women than men now complete tertiary education (Becker et al. 2010 and Blau and Kahn 2017), women select disproportionately into lower-paying majors, with very little gender convergence along this margin since the mid-1980s (Bronson 2014). College students majoring in science, technology, engineering, and math (STEM) fields typically receive higher labor market payoffs than those in other fields (Deming and Noray 2018). Yet, according to the Census Bureau's 2019 American Community Survey (ACS), among the 50 million employed college male(female) graduates aged 25 to 64 in 2019, 44.4%(20.9%) reported a bachelor's degree in a STEM major and only 27.8%(7.5%) worked in a STEM occupation (which includes computer occupations, mathematicians and statisticians, engineers, life scientists, and physical and social scientists).[1] The explanations for low female representation in STEM majors and occupations range from biological factors, social/cultural influences, abilities and skill sets (Wiswall and Zafar 2015; Wang et al. 2013; Wang and Degol 2013; Wai et al. 2018; Stoet and Geary 2018; Breda and Napp 2019, etc.), different preferences for non-pecuniary characteristics of STEM occupations (Wiswall and Zafar 2018; Valentino et al. 2016; Bronson 2014; Addison et al. 2020, etc.), to gender discrimination in male-dominated occupations.[2]

This paper examines the contributions of adolescent skill profiles, high school course-taking, and family background to explaining gender differences in four-year college degree attainment, STEM major choices, and occupations using longitudinal data from both the NLSY79 and NLSY97 cohorts. First, we document the degree of convergence by gender in math, verbal, science, mechanical, and administrative skills.[3] Second, we use logistic regression and machine learning methods (Classification and Regression Trees (CART) and random forests) to identify the skill sets that are

[1] See the 2018 Census Occupation Code Lists on https://www.census.gov/topics/employment/industry-occupation/guidance/code-lists.html for further information.

[2] See Goldin (2021) for a historical review and Kahn and Ginther (2017) for a literature review.

[3] Mechanical knowledge involves the ability to understand how mechanical equipment and tools work and to apply that knowledge in practical situations. In the ASVAB test, mechanical skills are tested in automotive and shop information, electronics information, and mechanical comprehension. The automotive and shop information test is designed to assess one's understanding of mechanics in more complex machinery like automobiles as well as simple machines like levers and pulleys; the auto information test is designed to gauge the test taker's overall knowledge of automobiles, including the transmission, electrical systems, and mechanics; the questions in the electronics information ASVAB test are designed to gauge one's overall knowledge of electrical systems, devices, circuits and general concepts related to electricity.

most predictive of four-year college degree attainment, college major choice, and occupational sector. Our analysis distinguishes among three broader sectors in terms of majors and occupations: pure STEM, applied STEM (engineering or computer science), and non-STEM. Third, we develop a nonparametric decomposition approach that accommodates multi-dimensional sequential outcomes and we use it to empirically assess how eliminating pre-college gender skill gaps would alter female representation in STEM majors and occupations.

We build on the prior literature by examining a more comprehensive set of cognitive and noncognitive skill measures as possible determinants of educational and occupational outcomes, investigating the separate determinants of pure-STEM and applied-STEM field choices, and using nonparametric machine learning methods. An advantage of our nonparametric approach is that it allows in a flexible way for interactions between covariates and therefore reveals particular skill and course-taking combinations that are highly predictive of STEM major and occupation choices. In addition, the tree-based estimation procedures endogenously choose test score distribution cut-offs that predict the outcomes. The results are therefore informative on how being in the tails of the skill test score distributions predicts entry into STEM fields. We find that the nonparametric models provide a better fit to the data as measured by the percent correctly classified under the model. We also find important differences between the inferences obtained by logistic regression and the nonparametric analysis. For example, race and gender are generally less significant predictors in the nonparametric models, indicating that their effect on outcomes is mostly mediated through the other covariates.

Our analysis yields a number of insights about the sources of gender disparities in STEM majors and occupations. First, we find that adolescent skill profiles and high school course-taking patterns are key determinants of four-year college completion, college major, and occupation choices. The CART and random forest analysis shows that math skill is the best predictor of four-year college degree attainment. Second, we find that among college graduates, individuals who enter applied STEM majors tend to have a combination of high math and mechanical ability and to have taken more physical science coursework in high school. To our knowledge, the importance of mechanical skills in explaining entrance into applied-STEM majors and occupations has not been previously noted, so this is an important finding.[4] Women have closed the gender gap in math skills and

[4] Chuan and Zhang (2021) argue that men's higher level of mechanical skills leads them to have a higher return in

exceed men in verbal skills by a substantial amount, but they lag behind in mechanical and, to a lesser extent, science skills. In the NLSY cohorts that we analyze, women also take on average fewer high school courses than men in physical science. Third, our nonparametric decomposition analysis indicates that about half of the gender gap in applied-STEM major and occupation choices could be eliminated if women's high school skill profiles and course-taking were aligned with those of men. The results suggest that gender differences in STEM preparation can account for 30-50% of the gender gap in pursuing STEM majors and occupations.

The paper develops as follows. Section 2 reviews related literature that analyzes the determinants of gender differences in STEM participation. Section 3 describes the data and construction of skill measures used in our analysis. Section 4 presents descriptive statistics. Section 5 describes the sequential schooling and occupational choice model that motivates our empirical analysis and how we modify the standard Oaxaca-Blinder decomposition method to accommodate multiple outcomes in our nonparametric setting. Section 6 describes the CART and random forest methods that we use. Section 7 presents the model estimates and the decomposition results. Section 8 concludes.

2 Literature review

The literature exploring why women are underrepresented in certain academic fields of science is vast.[5] A comprehensive survey by Ceci et al. (2014) presents a life-course analysis examining the evidence on different factors in a child's development that are potentially important drivers. The authors note that women are underrepresented in spatially/math-intensive fields that they term GEEMP fields (geoscience, engineering, economics, mathematics/computer science, and physical science, including chemistry), but not in non-math-intensive science fields that they term LPS (life-science, psychology, social science). Some gender differences in cognition emerge at very young ages, suggesting a potential role for biological factors. For example, infant boys as young as 18 months have been shown to exhibit better spatial rotation abilities than similar-age girls.[6] However, there is also strong evidence that gender norms play a role. For example, gender gaps in math test

non-college work, but they do not consider how mechanical ability affects college major and career choices.

[5] It spans diverse disciplines, including psychology, economics, sociology, endocrinology, mathematics, philosophy, bibliometrics, and education.

[6] Miller and Halpern (2014) review the evidence on early male advantage and argue that cognitive sex differences are changing over time.

scores vary substantially across countries and, in the US, across states and across schools within a state. Legewie and DiPrete (2012) find that by middle school, boys are twice as likely as girls to expect to work in science or engineering. Ceci et al. (2014) also document time trends over the years 1992-2011 in the percentage of women graduating high school, majoring in STEM fields in college, matriculating to graduate programs, finishing PhD's and entering tenure track positions. After reviewing the accumulated evidence from economics and other disciplines, they conclude that barriers to women's full participation in mathematically-intensive academic science fields are mostly rooted in pre-college factors and the subsequent likelihood of majoring in these fields and not in gender discrimination.

A central question in the literature examining women's under-representation in STEM is whether women have lower mathematics ability, with most studies focusing on average ability. Hyde et al. (1990) present a meta-analysis that includes 100 studies. They find no sex differences for children at any age, except for a small male advantage in solving complex math problems in high school. A more recent survey by Hyde and Mertz (2009) concludes that by the early twentieth century, girls reached parity with boys, even for more complex math problems, as measured by the National Assessment of Educational Progress (NAEP). They attribute girls' relative improvement over time partly to taking more math courses. Although the gender gap in average math scores has by now been largely eliminated, males are still more likely to be found in the right tail of the math test score distribution. Hyde et al. (2008) finds a 2.09:1 male-female ratio in the top 1% of math scores on the NAEP. Ceci et al. (2014) observe that the 10% quantile score for males on the SAT corresponds to the 5% quantile score for females. Being in the right tail is potentially a highly relevant characteristic for individuals choosing STEM majors and occupations.

Card and Payne (2021) analyze data on high school course-taking and college entrance in the Ontario high school and university system. They find that women leave high school with a lower level of STEM-readiness than men and also that women who are not prepared for STEM are more likely to go to college than similarly situated men. They conclude that gender differences in high school course-taking and in college entry conditional on course-taking are both important in explaining the lower fraction of females choosing STEM majors.[7] Cohodes et al. (2022) present

[7] Another related paper is Neal and Johnson (1996), which shows that the AFQT test score (which is a composite score based on a subset of the ASVAB scores that we analyze in this paper) is an important pre-market factor in explaining racial wage gaps.

randomized control trial (RCT) evidence on the effects of a STEM-focused summer enrichment program, offered at an elite university, targeted at high-achieving STEM-interested high school students between their junior and senior year. The program was aimed at reducing racial/ethnic disparities in STEM degree attainment. The RCT had multiple treatment arms with interventions of different durations, ranging from one to six weeks, either in-person or online. The authors find that the enrichment programs increased four-year college matriculation, increased the quality of the college that students attended, increased college completion rates, and increased STEM degree attainment. The results show the potential benefits from interventions targeted at high school students.

There are also several studies in economics considering how individuals' major choices relate to expectations about future monetary returns from choosing certain occupations. For example, Gemici and Wiswall (2014) study the evolution of human capital investment and labor supply choices for men and women in the 1920-1960 birth cohorts, incorporating college major choice within a dynamic overlapping generations model. They find that men's investment choices respond more to increases in rental rates of science and business majors because they have a lower utility of staying home and anticipate a higher future labor supply. Stinebrickner and Stinebrickner (2014) find that many students enter school optimistic about obtaining a science degree but that few students graduate with one. They argue that students begin school with misperceptions about their ability to perform well academically in science. Mourifie et al. (2020) test whether women's choices to pursue a STEM major in Germany and Canada can be characterized by a Roy model. They reject a Roy model of self-sorting in major choice for most of their data subsamples and argue that the key determinants of STEM major choices are non-labor-market-related.

3 Data

This study makes use of two datasets from the National Longitudinal Survey of Youth: the NLSY79 cohort (born between 1957-64) and the NLSY97 cohort (born between 1980-84), which are nationally representative. Both cohorts were administered the Armed Services Vocational Aptitude Battery (ASVAB) at young ages, from which we derive our skill measures. We restrict our analysis sample to those who took the test at age 19 or earlier to focus on skills acquired prior

to attending college. To make the two cohorts as comparable as possible, we do not include the military and economically disadvantaged oversamples that are available in the NLSY79 but not in the NLSY97. The sample inclusion criteria and the procedures for handling item nonresponse are described in Appendix A.

3.1 Skill measure construction

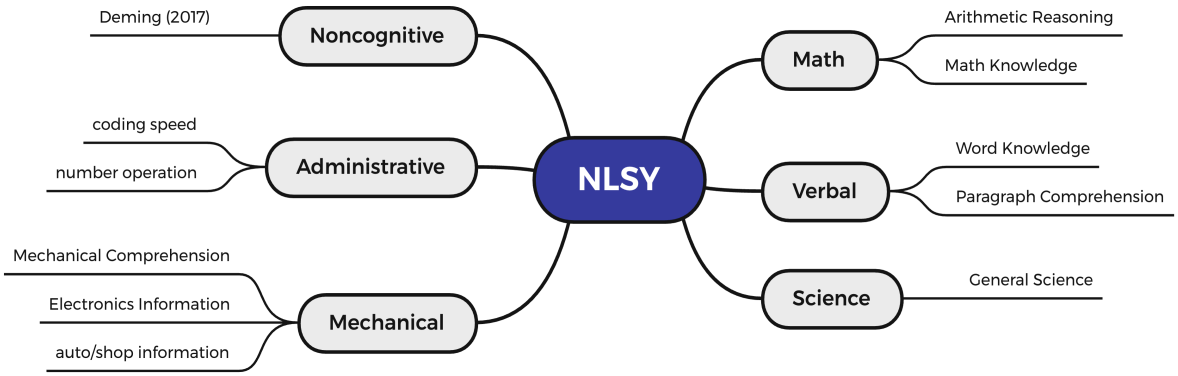
The ASVAB test consists of a battery of ten tests that measure knowledge and skill in the following areas: *general science, arithmetic reasoning, word knowledge, paragraph comprehension, numerical operations, coding speed, auto and shop information, mathematics knowledge, mechanical comprehension*, and *electronics information*.[8] We perform a Principal Component Analysis (PCA) to examine the correlation structure among the ten ASVAB test items and as a guide in constructing a lower-dimensional set of skill composites. Following Prada and Urzúa (2017), we perform a PCA on the two cohorts' ASVAB test scores separately to detect clusters of test scores that to a large extent define one skill component. The PCA procedure is described in detail in Appendix B. The clustering pattern that emerged is robust across cohorts and genders. As indicated by the PCA results, we construct a lower dimensional set of skill measures consisting of five cognitive skills (i.e., math, verbal, science, mechanical, administrative) and one non-cognitive measure, shown in Figure 1.[9] To obtain the math skills composite, we standardize the *arithmetic reasoning* and *math knowledge* subtests to have a mean of 0 and standard deviation equal to 1 and take a simple average. We similarly take the simple average of the standardized *word knowledge* and *paragraph comprehension* subtests for verbal skills. For administrative skills, we take the simple average of the standardized *coding speed* and *number operation* subtests. The PCA analysis indicated that mechanical skills and science skills are distinct components. For the mechanical skill composite, we group together standardized test scores in *mechanical comprehension*, *electronics information*, and *automotive and shop information* and take the average.[10] We use the single standardized *general science* test as the science skills measurement.

[8] There are some minor differences between the ASVAB administered to NLSY79 and to NLSY97. More specifically, for the NLSY97, ASVAB divides the item "auto and shop information" into "auto" and "shop" information separately. In addition, for the NLSY97, ASVAB has a new item called "assembling objects", which we omit to make the results of the two cohorts more comparable.

[9] Our PCA results are largely consistent with similar results reported in previous studies (e.g., Neal and Johnson 1996; Altonji et al. 2012; Deming 2017; Addison et al. 2020; Chuan and Zhang 2021).

[10] In Appendix C, we provide some sample questions that are illustrative of the mechanical skills test.

Figure 1: Skill Score Construction Using NLSY



Note: This graph plots the construction of skill profiles in NLSY79 and NLSY97.

The literature proposes different ways of measuring non-cognitive skills.[11] We follow the approach taken in Heckman et al. (2006) and Deming (2017).[12] In particular, we construct "non-cognitive" skills for the NLSY79 cohort using the standardized Rotter Locus of Control (i.e., a measure of self-control and self-determination, the extent to which individuals believe they exercise control over their lives) and the Rosenberg Self-Esteem Scale (i.e., a measure of self-worth). For the NLSY97 cohort, we use the questions intended to measure conscientiousness (one of the so-called Big Five personality traits) that relate to being organized, conscientious, dependable, thorough, trusting, disciplined, and careful. Thus, the noncognitive skill measures differ somewhat across the two birth cohorts due to differences in the survey instruments, but they relate more generally to the perceived level of control over life and one's own behavior.

In the NLSY survey, individuals take the ASVAB tests once and the age at which they take the test varies somewhat, but not in any systematic way. To make the scores comparable across individuals, we need to account for possible age effects. First, we restrict our sample to individuals who are 19 or younger at the time of taking the test. Then, we regress each of the test score measures on age indicator variables and obtain fitted values at a common age (age 19). Lastly, we

[11] For example, in Guvenen et al. (2020), the noncognitive skill set is measured using the Rotter Locus of Control Scale and the Rosenberg Self-Esteem Scale. Addison et al. (2020) use for the NLSY79 cohort the respondents' self-report on their then-current sociability and (retrospective) sociability at age six along with their rank on the Rotter Locus of Control Scale and the Rosenberg Self-Esteem Scale, while for the NLSY97 cohort, they use two questions on extroversion and two questions on conscientiousness.

[12] Deming (2017) considers both a measure of non-cognitive skills and of social skills that he argues have increased in labor market importance over the last two decades. We only consider the non-cognitive skill measure in this paper. Participation in high school sports and clubs, which is used to measure social skills, seems less directly tied to STEM preparation, although its relevance might be examined in future work.

convert each of the test score measures into a percentile score that is used in our analysis.[13]

3.2 A finer classification - non-STEM, pure STEM, applied STEM

A frequently used classification in discussions over gender inequality is the STEM v.s. non-STEM dichotomy, where STEM stands for the fields of science, technology, engineering, and mathematics, and such classification pertains to both majors and occupations.[14] Admittedly, STEM fields have attracted more men than women overall, but a closer look reveals that not every STEM field repels women. According to Sloane et al. (2021), a number of historically male-dominated majors have seen female representation surge over the past four decades, e.g., bio-sciences, physical sciences, business, etc., but fields such as computer science, engineering, and engineering technology continue to have low female representation (less than 20%). Such stark differences in female representation by subfield call for a finer classification of the STEM fields in consideration of gender inequality.

This paper employs a finer taxonomy by dividing the STEM fields into the applied-STEM and the pure-STEM fields. The applied-STEM fields consist of the "T" (technology) and the "E" (engineering) in STEM, i.e., computer science, engineering, and engineering technologies, while the pure-STEM fields consist of the other two sub-fields "S" (science) and "M" (mathematics), i.e., biological and physical sciences, environmental science, mathematics, some social sciences. Appendix E provides a detailed description of how we classify majors and occupations into non-STEM, pure-STEM, and applied-STEM. Compared with pure-STEM fields, the applied-STEM fields are more subject to technological change, have higher initial returns, experience flatter wage growth, and show more frequent job exit over time (Deming and Noray 2018). The applied-STEM fields have experienced very little growth in female representation over the last two decades, whereas the pure-STEM fields have attracted more and more women, in terms of both major (see Figure 2) and occupation (see Figure 3).

[13] Appendix D provides further information on the age adjustment procedure.

[14] Following Peri et al. (2015), we adopt the definition of STEM major used by the U.S. Department of Homeland Security in determining visitor eligibility for an F-1 Optional Practical Training (OPT) extension at https://www.ice.gov/sites/default/files/documents/Document/2016/stem-list.pdf. The classification for occupations is taken from the website of the Bureau of Labor Statistics for defining STEM occupations under the 2018 Standard Occupational Classification (SOC) system at https://www.bls.gov/soc/attachmentastem2018.pdf.

Figure 2: College Major Composition (2008-2010, 2017-2019 ACS)

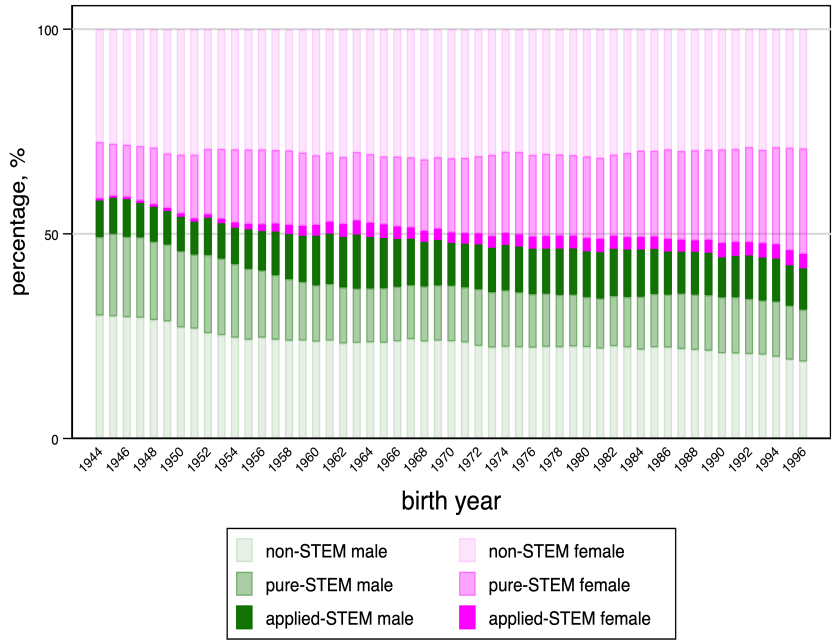
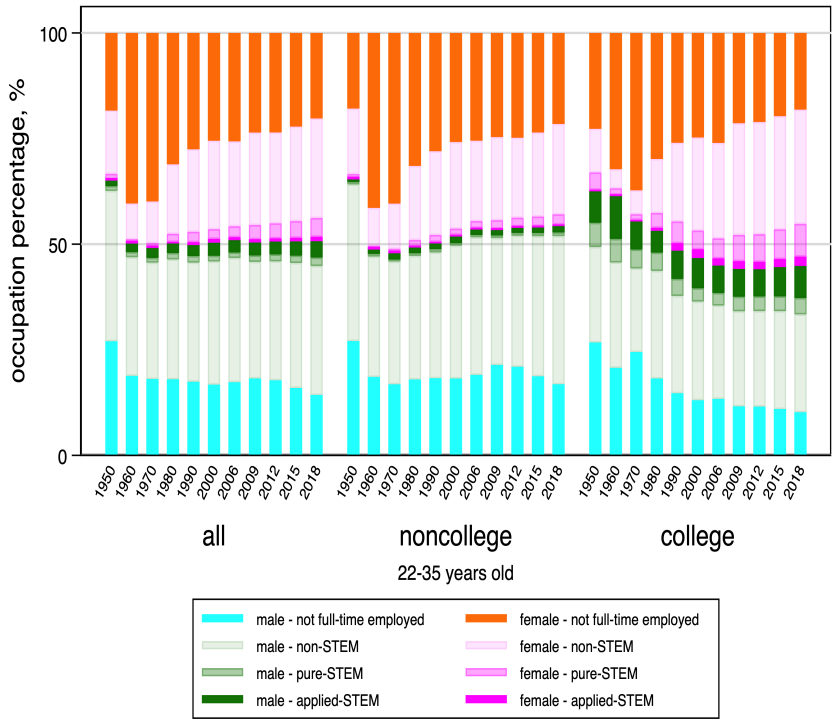


Figure 3: Occupation composition by education groups (1950-2019 ACS)



4 Descriptive statistics

Table 1 shows summary statistics for the variables used in our analysis for the two NLSY birth cohorts by gender. Roughly 13-16% of the sample is African American. The average age at which individuals entered the NLSY sample is 17.6 in 1979 and 15.1 in 1997. In 1979, mothers and fathers of the NLSY respondents have on average less than a high school degree, and mothers have slightly lower education levels than fathers. For fathers, the average years of education increased by roughly one year from 1979 to 1997. For mothers, the average years of education shows a greater increase, from 11.7 to 13.1. The majority of NLSY respondents do not complete a four-year college degree, but the percentages in the "< four-year college" category decreases from 69-72% in 1979 to 56-69% in 1997. Females complete four-year college degrees at higher rates than males with a widening gender gap over time (44% versus 31% in 1997). With regard to major choices, females are more likely to major in pure-STEM fields, and males are more likely to major in applied-STEM fields. Women are more likely than men to work in pure-STEM occupations (15% versus 6% in 1997) and less likely to work in applied-STEM occupations (5% versus 14% in 1997).

Our analysis considers STEM preparation through high school course-taking as a potentially important determinant of college major and occupation choices. Gender differences in course-taking may reflect interests developed at earlier ages, parental influence, and/or high school graduation requirements that typically differ by state of residence. Table 1 shows the average number of courses taken per year.[15] Students in the 97 cohort report taking fewer non-STEM courses than those in the 79 cohort. The number of math courses has fallen slightly for males from 0.84 to 0.71, but the number of math courses for women has held roughly constant (0.75 to 0.72). The number of high school physical science courses shows a similar pattern, falling slightly for males (0.42 to 0.29) and less so for females (0.34 to 0.29). Thus, the gender gaps in math and physical science course-taking have declined largely because males decrease their coursework. Males in both cohorts take more mechanics courses than females, although the initial gender gap of 0.41 reduces substantially to 0.15. Males were taking fewer computer science courses than females in the 79 cohort, but the gap reverses for the 97 cohort. The numbers of biology/health science/social studies courses are roughly equivalent by gender for the 79 cohort, but females take slightly more courses on average in

[15]The numbers of courses reported in the data are adjusted to take into account whether the course term was a quarter, semester, or year-long course.

the 97 cohort. Males in the 79 cohort take substantially more vocational coursework, but there is no gender difference for the 97 cohort. The data show a general pattern of both males and females in the younger 97 cohort taking fewer courses in physical sciences, mechanics, biological sciences, and computer science than the older 79 cohort.

The last six rows of Table 1 show the average test scores for the six different skill measures, measured on a scale of 0-100. For the 79 cohort, there is a male-female gap in the average math score of around 2 points on average; for the 97 cohort, females have a slightly higher math score. The variance in the math score is higher for males in both years, consistent with the literature (cited previously) that finds that males are more likely to be found in the math test score distribution tails. With regard to verbal skills, females have a higher average score than men in both the 79 and 97 cohorts - a gap of roughly three points. On the other hand, males have an advantage in science. The gap is 10 points for the 79 cohort but falls to 5 points for the 97 cohort. Males also have a substantially higher mechanical score, with a gap of 24 points for the 79 cohort that narrows to 16 points for the 97 cohort. Lastly, females show a substantial advantage in administrative skills, although the female-male gap in the average administrative score narrows from 14 points to 7 points.

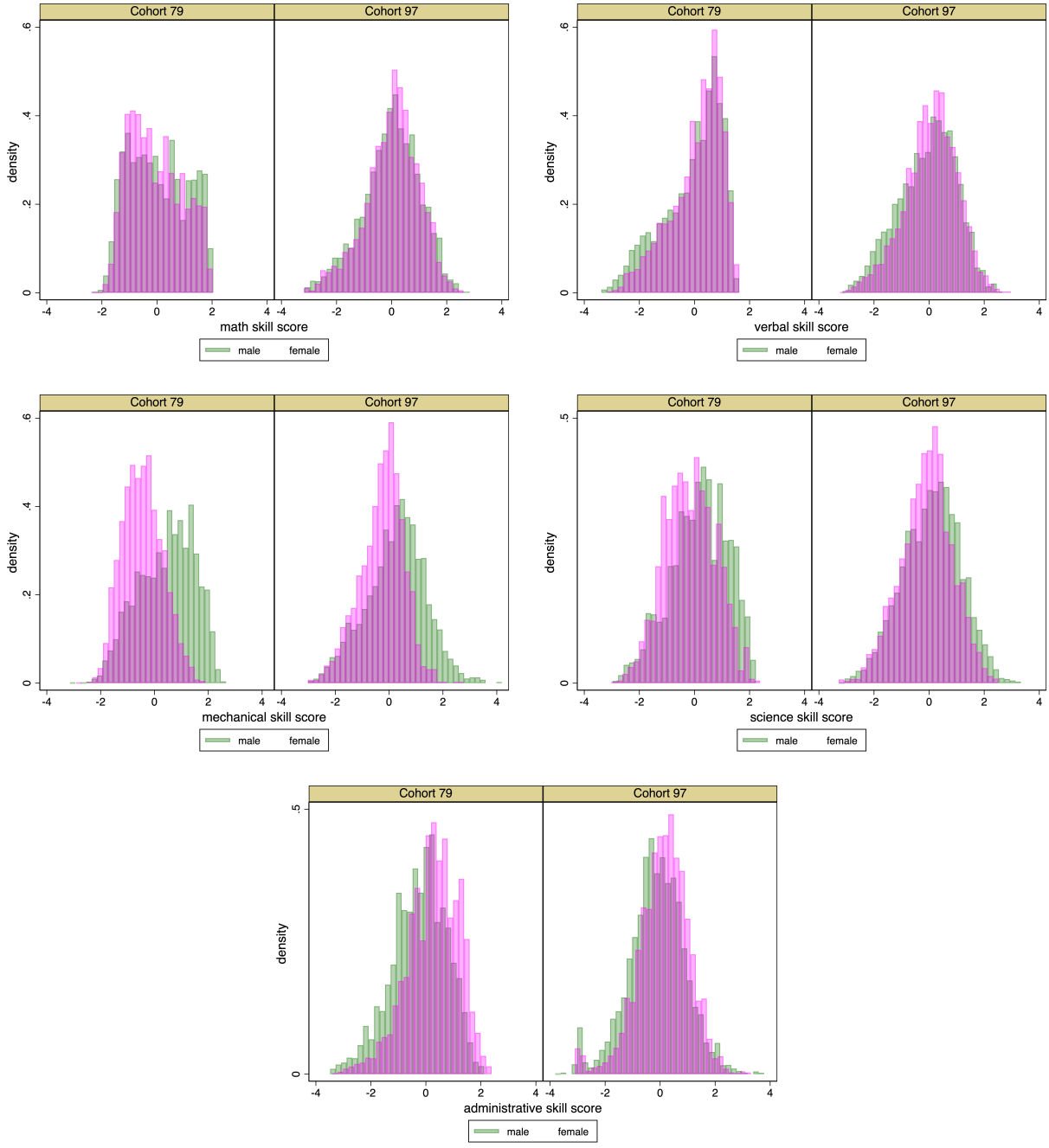
Figure 4 plots the test score distributions for males and females in math, verbal, mechanical, science, and administrative skill areas for both the 79 and 97 cohorts. The tests themselves differ somewhat over time, so we do not draw inferences about temporal changes in the shape of the distribution. For math, it is easy to see that the female and male test score distributions largely converged for the younger cohort, although men have a slightly larger probability of being in the tails. For verbal, the female distribution stochastically dominates the male distribution. In the mechanical realm, however, the male distribution stochastically dominates the female distribution. The gender gap reduces over time, but the mechanical skill gap is still substantial for the 97 cohort, with the female distribution also exhibiting smaller variance. Males also have an advantage in science, although the gap is not nearly as large as for mechanical skills. Lastly, as seen in the bottom panel, females have an advantage in administrative skills.

Table 1: Summary Statistics

	(1) Cohort 79 Male	(2) Cohort 79 Female	(3) Cohort 97 Male	(4) Cohort 97 Female
Black	0.13	0.14	0.15	0.16
	(0.334)	(0.343)	(0.358)	(0.371)
age	17.58	17.58	15.07	15.13
	(1.117)	(1.114)	(1.434)	(1.422)
father's highest grade completed	12.01	11.88	12.96	13.15
	(3.372)	(3.272)	(3.138)	(3.007)
mother's highest grade completed	11.81	11.73	13.06	13.08
	(2.568)	(2.545)	(2.694)	(2.710)
< four year college	0.72	0.69	0.69	0.56
	(0.447)	(0.461)	(0.462)	(0.496)
college major: non-STEM	0.14	0.19	0.15	0.23
	(0.346)	(0.388)	(0.358)	(0.423)
college major: pure-STEM	0.07	0.09	0.03	0.05
	(0.247)	(0.290)	(0.178)	(0.221)
college major: applied-STEM	0.07	0.03	0.03	0.01
	(0.252)	(0.160)	(0.164)	(0.0719)
Not working full-time	0.04	0.06	0.13	0.16
	(0.185)	(0.236)	(0.335)	(0.364)
full-time occupation: non-STEM	0.68	0.64	0.67	0.64
	(0.466)	(0.479)	(0.469)	(0.479)
full-time occupation: pure-STEM	0.09	0.19	0.06	0.15
	(0.292)	(0.396)	(0.244)	(0.357)
full-time occupation: applied-STEM	0.19	0.10	0.14	0.05
	(0.391)	(0.304)	(0.342)	(0.217)
# of high school non-STEM courses taken	3.93	4.46	2.83	3.07
	(1.323)	(1.285)	(1.823)	(1.879)
# of high school math courses taken	0.84	0.75	0.71	0.72
	(0.363)	(0.325)	(0.453)	(0.454)
# of high school physical sciences courses taken	0.42	0.34	0.29	0.29
	(0.298)	(0.272)	(0.259)	(0.242)
# of high school mechanics courses taken	0.64	0.23	0.21	0.06
	(0.542)	(0.245)	(0.355)	(0.166)
# of high school computer courses taken	0.13	0.36	0.14	0.11
	(0.188)	(0.329)	(0.217)	(0.176)
# of high school biology/health/soc. sciences courses taken	0.91	0.89	0.51	0.57
	(0.534)	(0.456)	(0.401)	(0.431)
# of high school vocational courses taken	0.33	0.19	0.16	0.16
	(0.447)	(0.331)	(0.236)	(0.257)
math score	51.39	48.84	49.74	50.34
	(29.86)	(27.67)	(29.49)	(28.22)
verb score	48.46	51.95	48.54	51.55
	(29.73)	(28.06)	(29.78)	(27.84)
science score	55.67	45.53	52.80	47.23
	(29.60)	(27.22)	(29.99)	(27.40)
mechanical score	62.37	37.56	57.81	42.16
	(28.69)	(22.99)	(30.52)	(24.72)
admin score	43.05	57.45	46.22	53.90
	(27.98)	(27.99)	(29.31)	(27.88)
noncog index	51.64	48.43	47.77	52.80
	(28.23)	(29.34)	(28.81)	(29.10)
Observations	2030	2010	2538	2557

mean coefficients; standard deviation in parentheses

Figure 4: Score Distributions by Gender - NLSY 79 and 97



5 Model

Our empirical analysis explores whether and to what the extent gender differences in high school skill profiles and course-taking can explain gender disparities in four-year college degree attainment, college major, and occupation choice. We next define some notation and describe the choice models that we estimate for the sequential decisions to pursue a college degree, to major in a STEM field, and to enter a STEM occupation. Let $E_i = 1$ if an individual completes at least four years of college, else $E_i = 0$. Let $M_i \in \{0, 1, 2\}$ denote the major choice, which can be non-STEM, pure-STEM, or applied-STEM. Let $O_i \in \{1, 2, 3\}$ denote the occupation choice, which can be non-STEM, pure-STEM, or applied-STEM. Let X_i denote individual characteristics determined prior to entering college, which includes the five cognitive skill measures, the one non-cognitive skill measure, variables describing high school course-taking, and family background. Let g_i denote gender, where $g_i \in \{M, F\}$.

5.1 Conditional choice probabilities

The likelihood of observing a particular set of major and occupational outcomes for a college graduate ($E_i = 1$) with characteristics (X_i, g_i) can be written sequentially as

$$Pr(E_i = 1, M_i, O_i | X_i, g_i) = Pr(O_i | M_i, E_i = 1, X_i, g_i) \times Pr(M_i | E_i = 1, X_i, g_i) \times Pr(E_i = 1 | X_i, g_i)$$

For an individual who does not complete a four-year college degree, the likelihood of observing their education and occupation outcome is:

$$Pr(E_i = 0, O_i | X_i, g_i) = Pr(O_i | E_i = 0, X_i, g_i) \times Pr(E_i = 0 | X_i, g_i)$$

We can write the joint likelihood of observing a set of outcomes (college four-year degree sta-

tus,college major and occupation choice) for N individuals as

$$L = \prod_{i=1}^{N} \Bigg\{ \Bigg(\prod_{k=1}^{K} \prod_{j=1}^{J} \Big[Pr(O_i = k | M_i = j, E_i = 1, X_i, g_i)^{I(O_i=k, M_i=j)} \times Pr(M_i = j | E_i = 1, X_i, g_i)^{I(M_i=j)} \Big] \times Pr(E_i = 1 | X_i, g_i) \Bigg)^{I(E_i=1)} \times \Bigg(\prod_{k=1}^{K} \Big[Pr(O_i = k | E_i = 0, X_i, g_i)^{I(O_i=k)} \Big] \times Pr(E_i = 0 | X_i, g_i) \Bigg)^{I(E_i=0)} \Bigg\},$$

where the index k denotes occupation and j the college major, with $sK = 3$ and $J = 3$.

Our empirical strategy is as follows. First, we estimate each of the components of the above likelihood equation parametrically, using multinomial and binary choice modes. Then, we estimate the components nonparametrically, using classification and regression tree analysis (CART) and random forest methods. Our decomposition analysis, described below, combines the different estimated likelihood components to obtain the probability that an individual of gender g with some initial skill and family background profile X will end up working in an applied-STEM occupation:

$$\begin{aligned} Pr(O_i = 3 | X_i, g_i) = & \\ & \left(\sum_{j=1}^{J} Pr(O_i = 3 | M_i = j, E_i = 1, X_i, g_i) \times Pr(M_i = j | E_i = 1, X_i, g_i) \right) \times Pr(E_i = 1 | X_i, g_i) \\ & + Pr(O_i = 3 | E_i = 0, X_i, g_i) \times Pr(E_i = 0 | X_i, g_i). \end{aligned}$$

In the above equation, the first term sums over different possible college major choices. The second term is the proportion of people who enter STEM occupations without a four-year college degree (such as lab technicians).

5.2 Nonparametric decomposition approach

After estimating the choice probabilities, we perform a decomposition to quantify the relative contribution of adolescent skill profiles (included in the X variables) in accounting for gender gaps. Our decomposition approach is a nonparametric analogue to the linear regression approach pioneered by Oaxaca (1973) and Blinder (1973). It is also similar to the decomposition approach of DiNardo et al. (1995), except that we use random forest methods in estimating the conditional

probabilities and we analyze multiple discrete outcomes that are realized sequentially (four-year college degree attainment, major and occupation choices).

For example, we can decompose the gender gap in the four-year college degree attainment ($E_i = 1$) probability into the following:

$$\begin{aligned}
\Delta^1 &= Pr(E=1|g=M) - Pr(E=1|g=F) \\
&= \int Pr(E=1|X,g=M) f(X|g=M) dX - \int Pr(E=1|X,g=F) f(X|g=F) dX + \\
&\quad \int Pr(E=1|X,g=F) f(X|g=M) dX - \int Pr(E=1|X,g=F) f(X|g=M) dX \\
&= \int [Pr(E=1|X,g=F) - Pr(E=1|X,g=M)] f(X|g=M) dX + \\
&\quad \int Pr(E=1|X,g=F) [f(X|g=F) - f(X|g=M)] dX \\
&= \Delta^2 + \Delta^3
\end{aligned}$$

In the above equation, $f(X|g=M)$ and $f(X|g=F)$ denote the distribution of characteristics for males and females. Note that the second line in the equation adds a term that sums to zero. The percentage of the college completion gap that is explained by gender differences in the X distribution is Δ^2/Δ^1. We can similarly decompose gender gaps in other outcomes. For example, to analyze the determinants of college major disparities among college graduates, we replace the $Pr(E=1|X,g)$ in the above equations with $Pr(M=m|E=1,X,g) \times Pr(E=1|X,g)$.

Δ^3 consists of two terms: $\int Pr(E=1|X,g=F) f(X|g=F) dX$ and $\int Pr(E=1|X,g=F) f(X|g=M) dX$. The first term is obtained by estimating the predicted probability for each female in the data (from the random forest estimation) and averaging the predicted probabilities across the individuals. The second term is obtained by plugging the X characteristics observed for the male subsample into the random forest tree models, with the gender indicator set equal to female, to get predicted probabilities and then averaging. The latter quantity gives the predicted four-year college degree attainment probabilities if females would have the X characteristics that are distribution observed for males.

6 CART and random forest estimation methods

We use logistic regression, CART (Classification And Regression Trees), and random forest methods to estimate each of the likelihood components: (i) the probability of attaining at least a four-year college degree, (ii) the probability of majoring in non-STEM, pure STEM, or applied STEM fields conditional on college completion, and (iii) the probability of choosing a non-STEM, pure STEM, or applied STEM occupation conditional on attaining a four-year college degree and conditional on major. Below is a brief description of the CART and random forest methods that we use and how we choose relevant tuning parameters.

6.1 Tree and random forest methods

CART is a nonparametric modeling approach used in statistics, data mining, and machine learning for classification and prediction (Loh 2011). Our implementation estimates models that best predict four-year college degree attainment, college major, and occupational choice, where the covariates are adolescent skill profiles (our five skill measures and the noncognitive measure), family background, and high school course-taking. CART models can be represented as binary trees that sequentially partition the space of covariates, which are also sometimes called input variables or features. The CART procedure selects variables and cut-points on those variables until a tree is constructed to minimize a cost function using a greedy algorithm.[16]

CART is considered a model-free approach, but the procedure does require specifying some tuning parameters. We use the "sklearn" package of Python 3.8.3 to estimate and plot the CART trees. The procedure sets aside a random 20% subset of the data as the test data with the rest constituting the training data. We make use of a "cost-complexity" pruning strategy combined with cross-validation to start from an overly complicated tree (trained on the training data) and then prune it back by recursively removing the weakest link. Specifically, the cost of a tree is defined as

$$R_\alpha(T) = R(T) + \alpha|T|$$

where $R(T)$ is the sum of squared error,[17] $|T|$ is the number of leaves (i.e., terminal nodes) in the

[16] A greedy algorithm solves a problem by sequentially selecting the best option at each node of the tree.

[17] For continuous outcome variables, the error is measured by mean squared error, and for categorical outcome variables, the error is measured by *Gini Impurity*, a measure of how often a randomly chosen element from the set

tree, and α is a penalty parameter for tree complexity that measures the cost of adding another node to the tree. For each value of α, there is a tree, estimated on the training data, that minimizes the cost $R_\alpha(T)$. Call it T^*_α. Then we select among all the $\{T^*_\alpha\}_{\alpha \in \mathcal{A}}$ the one that has the lowest prediction error on the test set.

We also use random forest methods, implemented using the *rforest* function available in STATA SE 16.0.[18] Random forest methods first obtain a large number (in our case 500) of random bootstrap samples from the data, sampling with replacement. For each bootstrap sample, a decision tree is built using a random subset of the covariates. Lastly, the predictions obtained from each tree are averaged in a procedure known as *bagging* to create a final estimated conditional probability.[19] We estimate separate random forests for the probability of attaining a four-year college degree, choosing a non-STEM, pure-STEM, or applied-STEM major, or entering a non-STEM, pure-STEM, or applied-STEM occupation given the same set of X covariates.

6.2 A single tree versus random forest

There are pros and cons both to estimating a single tree versus estimating a random forest.[20] A single tree can be graphically visualized even with multi-dimensions and interactions, while a random forest less so. The main disadvantages of single tree estimation is the potential risk of overfitting and sensitivity to outliers. Random forest methods generally perform better than single tree estimation methods (like CART) on test data that were not used in estimation. Because random forest methods typically require estimating hundreds of trees, they can be computationally intensive. We plot single trees below in Figure 5, 7 - 9 for easy visualization, but we use random forest to obtain the variable importance rankings and to estimate the conditional choice probabilities used in the decomposition analysis.

would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset, i.e., $\mathrm{I}_G(p) = 1 - \sum_{i=1}^{J} p_i^2$, where p_i is the fraction of items labeled with class i in the set, and J is the number of classes of a set of items $i \in \{1, 2, \ldots, J\}$. Please refer to Raileanu and Stoffel (2004) for more technical details.

[18] Please refer to Breiman (2001) for technical details.

[19] We use the default of the STATA routine, which sets the number of randomly chosen variables at each split to equal the square root of the total number of independent variables. Before growing a random forest, we tune the parameter for the number of trees bagged. We choose 500 eventually.

[20] Please refer to Ho (1995) for related discussion.

7 Empirical results

As previously described, we estimate models for the probability of attaining at least a four-year college degree, choosing a type of major (from among three types) and choosing an occupation (from three types). For comparison purposes, we estimate the probabilities using both parametric methods (multinomial logistic) and nonparametric methods (random forest). Table 2 shows the performance of the logistic model and of the random forest model in classifying individuals. As seen in the table, the logistic model, when estimated using the NLSY 79 data, classifies 22% of individuals incorrectly and, when estimated using the NLSY 97 data, classifies 26% incorrectly.[21] In comparison, the percentage misclassified under the random forest model is very small (1%).

In random forest modeling, the out-of-bag (OOB) error is a commonly reported measure of the model's prediction error. During the estimation process, each tree is trained on a bootstrap sample of the data.[22] Some observations will be included in the training set for some trees, while others will not. The observations that are not included in the training set for any given tree are referred to as the out-of-bag (OOB) observations. The model can be used to make predictions on the OOB observations and these predictions can then be compared to the actual outcomes as a goodness of fit measure. As seen in Table 2, the OOB error measure is lower for the college degree outcome than for the major or occupation outcomes, indicating higher prediction accuracy for that outcome. Also, for the first three outcome measures, the model fit is better for the 97 cohort data than for the 79 data.

Table 2: Proportion Incorrectly Classified and OOB Error (Source: NLSY 79 and 97)

Test Error	college grad		major		occupation - college grad.		occupation - noncollege	
	Cohort 79	Cohort 97	Cohort 79	Cohort 97	Cohort 79	Cohort 97	Cohort 79	Cohort 97
Random Forest - PIC	.010	.002	.025	.106	.009	.005	.017	.100
Random Forest - OOB	.175	.230	.418	.222	.452	.423	.254	.158
Logit - PIC	.225	.258	.585	.602	.475	.538	.557	.563
Obs	4040	5095	953	1151	953	1150	3060	3353

Note: This table compares the model fit between the random forest method and the logistic regressions. The entries are the proportion incorrectly classified using the trained model.

[21] If the estimated conditional probability of completing a four-year college degree exceeds the fraction of individuals who complete such degree in the data, the person is classified as a college graduate.

[22] Bootstrapping is done with replacement.

7.1 Logistic regression results

Table 3 shows estimates from a logistic regression of the probability of attaining at least a four-year college degree, using both the 79 and 97 cohort datasets. The math score is the best predictor of college degree attainment followed by the verbal score. Higher test scores are generally associated with a greater probability of graduating college with the exception of the mechanical score, which has a negative associated coefficient. Not surprisingly, youth whose parents have higher education levels are more likely to graduate college. With regard to course-taking, a greater number of math courses increases the probability of graduating college for the 79 cohort and a greater number of physical sciences courses increases the probability for the 97 cohort. Greater numbers of courses in mechanics or vocational education decrease the probability. The estimated coefficient on $female$ is positive for the 97 cohort, consistent with the much higher college-going rate for females. Finally, being African American, ceteris paribus, increases the probability of completing college for the 79 cohort but is not a significant predictor for the 97 cohort.

Table 4 shows the estimated coefficients for a multinomial logit model of college major choice, where non-STEM is the left-out category and the model is estimated using the subsample that completes college. A higher math score increases the likelihood of choosing an applied-STEM major for individuals in the 97 cohort; A higher mechanical score increases the same likelihood for students in the 79 cohort. A higher administrative score decreases the likelihood. High school course-taking patterns also emerge as important determinants of college major choices. Students who take more physical science courses are more likely to major in STEM fields. Taking more mechanics courses also increases the probability of majoring in an applied-STEM field for both the 79 and 97 cohorts. Taking more biology/social science/health courses increases the probability of choosing a pure-STEM major for the 79 cohort. Being female lowers the probability of majoring in applied-STEM fields for the 97 cohort, indicating that gender is relevant even conditional on pre-college skills and coursework preparation. Being African American increases the probability of choosing an applied-STEM major for the 79 cohort.

Table 5 reports regression estimates of the probability of entering a pure-STEM or applied-STEM occupation, for the subsample of college graduates.[23] The omitted category is non-STEM.

[23]Individuals are classified as applied-STEM or pure-STEM if they are observed to work in such occupation after leaving school. Individuals who never report any occupation are excluded from the occupation choice analysis.

Table 3: Logistic regression for the probability of four-year college degree attainment

	(1) Cohort 79	(2) Cohort 97
math score	0.031***	0.021***
	(0.003)	(0.002)
verb score	0.012***	0.015***
	(0.003)	(0.002)
science score	0.004	0.003
	(0.003)	(0.002)
mechanical score	-0.012***	-0.010***
	(0.003)	(0.002)
admin score	0.009***	0.004**
	(0.002)	(0.002)
noncog index	0.005***	0.010***
	(0.002)	(0.001)
father's highest grade completed	0.071***	0.114***
	(0.018)	(0.015)
mother's highest grade completed	0.084***	0.112***
	(0.022)	(0.016)
# of high school non-STEM courses taken	-0.018	-0.072**
	(0.036)	(0.034)
# of high school math courses taken	0.386***	0.142
	(0.142)	(0.141)
# of high school physical sciences courses taken	0.197	1.046***
	(0.167)	(0.198)
# of high school mechanics courses taken	-0.509***	-0.349**
	(0.128)	(0.147)
# of high school comp sci courses taken	-0.315*	0.176
	(0.168)	(0.191)
# of high school biology/health/soc. sciences courses taken	-0.163*	0.011
	(0.094)	(0.111)
# of high school vocational courses taken	-0.511***	-0.434**
	(0.134)	(0.187)
female = 1	0.00002	0.493***
	(0.122)	(0.082)
African American	0.587***	0.054
	(0.125)	(0.094)
Observations	4040	5095
chi2	1268.990	1690.584

Table 4: Multinomial logit model for college major choice among college graduates

	(1) Cohort 79		(2) Cohort 97	
	pure STEM	applied STEM	pure STEM	applied STEM
math score	0.004	0.000	0.008	0.020**
	(0.005)	(0.007)	(0.005)	(0.009)
verb score	-0.012**	-0.017***	-0.002	-0.009
	(0.005)	(0.007)	(0.006)	(0.009)
science score	0.013**	0.006	0.003	0.000
	(0.005)	(0.007)	(0.006)	(0.009)
mechanical score	0.001	0.032***	0.004	0.005
	(0.006)	(0.008)	(0.006)	(0.009)
admin score	-0.002	0.002	-0.003	-0.010*
	(0.004)	(0.005)	(0.004)	(0.006)
noncog index	-0.005	-0.003	0.001	0.003
	(0.003)	(0.004)	(0.003)	(0.005)
father's highest grade completed	0.023	-0.047	-0.002	-0.064
	(0.030)	(0.037)	(0.035)	(0.053)
mother's highest grade completed	0.002	-0.012	0.002	-0.075
	(0.035)	(0.044)	(0.036)	(0.053)
# of high school non-STEM courses taken	-0.090	0.076	-0.141*	-0.309**
	(0.072)	(0.092)	(0.083)	(0.140)
# of high school math courses taken	0.344	0.554	0.162	-0.217
	(0.289)	(0.359)	(0.372)	(0.634)
# of high school physical sciences courses taken	0.706**	1.589***	0.036	1.784***
	(0.337)	(0.416)	(0.446)	(0.675)
# of high school mechanics courses taken	-0.314	0.718***	-0.346	0.840**
	(0.277)	(0.254)	(0.452)	(0.349)
# of high school comp sci courses taken	-0.277	0.197	-0.476	0.396
	(0.325)	(0.423)	(0.481)	(0.579)
# of high school biology/health/soc. sciences courses taken	0.396**	-0.337	0.267	-0.878
	(0.180)	(0.234)	(0.288)	(0.536)
# of high school vocational courses taken	0.352	0.459	-0.049	1.034
	(0.270)	(0.334)	(0.507)	(0.660)
female = 1	0.331	0.045	0.138	-1.716***
	(0.205)	(0.269)	(0.194)	(0.354)
African American	-0.025	0.767***	0.015	-0.245
	(0.233)	(0.290)	(0.242)	(0.443)
Observations	953		1151	
chi2	147.924		115.845	

Not surprisingly, college major predicts occupational sector for both the 79 and 97 cohorts. Individuals with applied-STEM majors are much more likely to enter applied-STEM occupations. Having a pure-STEM major increases likelihood of entering either a pure-STEM or applied-STEM occupation. For the 79 cohort, the only test score that is a statistically significant predictor of occupation sector, conditional on major, is the mechanical score, which increases the probability of entering an applied-STEM occupation. For the 97 cohort, higher math, science and mechanical test scores all increase the probability of entering an applied-STEM occupation. Conditional on college major, high school coursework also predicts future occupational choices. More courses in mechanics, computer science, and biology/health sciences increase the likelihood of entering an applied-STEM occupation. Being female increases the likelihood of entering a pure-STEM occupation. For the 79 cohort, being female also increases the probability of entering an applied-STEM field, but for the 97 cohort, it lowers it. Conditional on the other included covariates, race is not a significant factor.

Lastly, Table 6 shows analogous occupational choice regression results for the subsample without a four-year college degree.[24] The test scores and high school course-taking patterns are important predictors of occupational choices. Higher verbal, science, and mechanical scores all increase the probability of having a STEM occupation. Taking a greater number of math courses also increases the probability of having an applied-STEM occupation for the 97 cohort. Females, ceteris paribus, have a higher likelihood of working in a pure-STEM field and a lower likelihood of working in an applied-STEM field. Race is not a statistically significant predictor, conditional on the other covariates.

7.2 Classification tree results

We next use CART and random forest methods to estimate the probabilities of attaining a four-year degree and of choosing non-STEM or STEM college majors and occupations. These methods allow for potential interactions between covariates and endogenously choose cut-off values for covariates that best predict the outcome. As previously noted, being in the tails of one or more test score distributions may be important to pursuing STEM majors.

[24]These individuals may have two-year degrees. We did not attempt to model their higher education choices, because the two-year degree and certification programs are so varied.

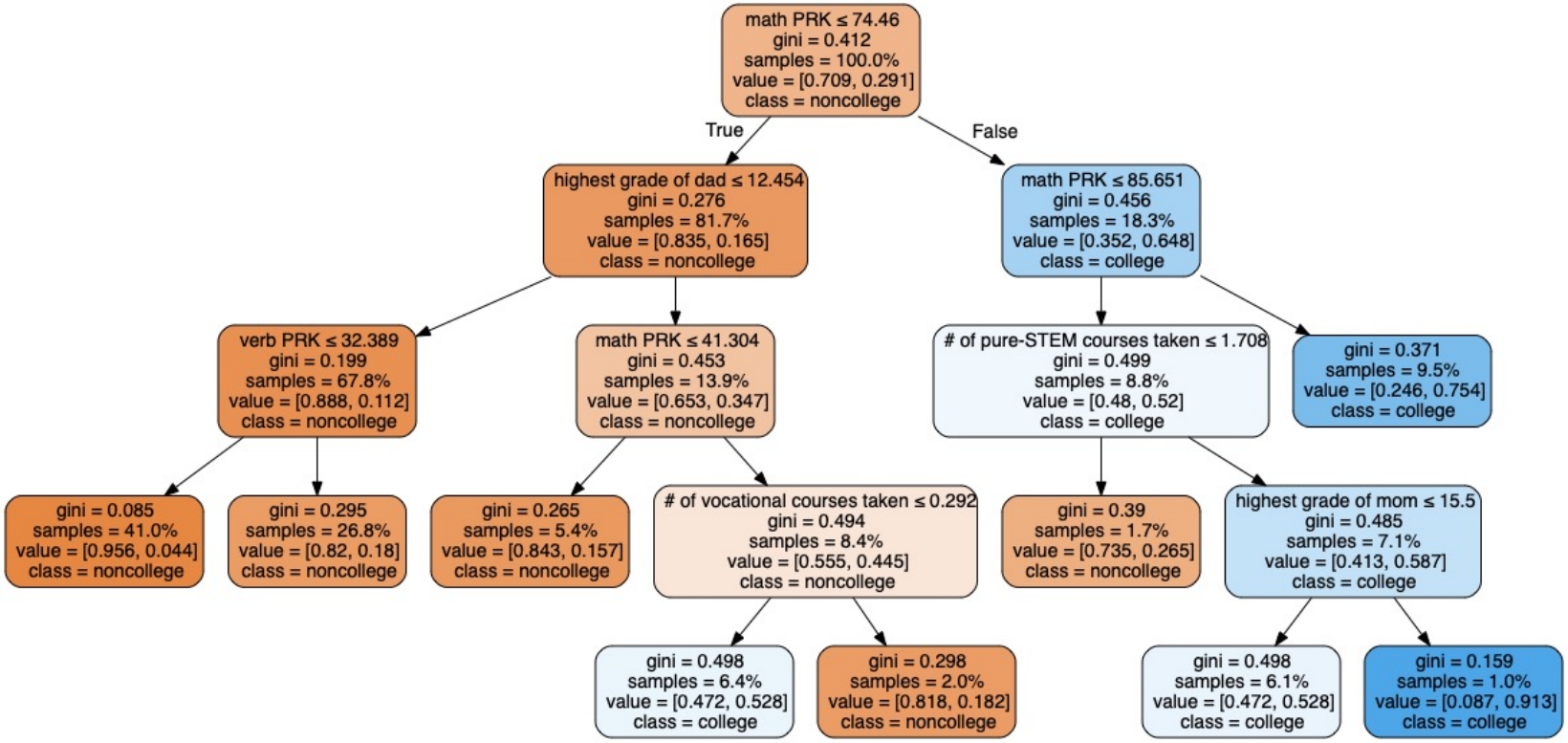
Table 5: Multinomial logit model for occupational sector of four-year college graduates

	(1) Cohort 79		(2) Cohort 97	
	pure STEM	applied STEM	pure STEM	applied STEM
pure-STEM major	1.343***	0.367	1.204***	0.799***
	(0.185)	(0.225)	(0.192)	(0.213)
applied-STEM major	-0.071	2.033***	-1.150	1.467***
	(0.367)	(0.254)	(0.748)	(0.283)
math score	-0.011**	0.003	0.001	0.011***
	(0.006)	(0.006)	(0.004)	(0.004)
verb score	0.009*	0.002	0.005	-0.003
	(0.006)	(0.006)	(0.004)	(0.005)
science score	0.005	0.003	-0.003	0.006
	(0.005)	(0.006)	(0.004)	(0.004)
mechanical score	0.002	0.025***	0.003	0.012***
	(0.006)	(0.007)	(0.004)	(0.004)
admin score	0.001	0.005	0.002	-0.005*
	(0.004)	(0.004)	(0.003)	(0.003)
noncog index	0.003	-0.001	0.004*	0.001
	(0.003)	(0.003)	(0.002)	(0.002)
father's highest grade completed	-0.045	0.021	0.081***	0.039
	(0.030)	(0.034)	(0.027)	(0.027)
mother's highest grade completed	0.022	-0.081**	-0.017	0.049*
	(0.037)	(0.040)	(0.028)	(0.029)
# of high school non-STEM courses taken	-0.061	-0.028	-0.168***	-0.095
	(0.076)	(0.082)	(0.065)	(0.061)
# of high school math courses taken	0.472	0.323	-0.032	-0.078
	(0.310)	(0.332)	(0.274)	(0.274)
# of high school physical sciences courses taken	0.497	-0.009	0.471	0.150
	(0.365)	(0.386)	(0.355)	(0.338)
# of high school mechanics courses taken	0.418	0.555**	-0.574	0.278
	(0.287)	(0.273)	(0.381)	(0.253)
# of high school comp sci courses taken	0.247	0.346	-0.561	0.809***
	(0.333)	(0.370)	(0.385)	(0.306)
# of high school biology/health/soc. sciences courses taken	0.153	0.024	0.438**	0.440**
	(0.194)	(0.207)	(0.217)	(0.212)
# of high school vocational courses taken	-0.123	0.151	0.069	0.245
	(0.294)	(0.308)	(0.363)	(0.357)
female = 1	0.726***	0.394*	0.379**	-0.615***
	(0.218)	(0.238)	(0.156)	(0.149)
African American	-0.258	0.430	0.191	0.208
	(0.244)	(0.266)	(0.175)	(0.195)
Observations	965		1721	
chi2	334.181		353.318	

Table 6: Multinomial logit model for occupational sector of < 4 year college sample

	(1) Cohort 79		(2) Cohort 97	
	pure STEM	applied STEM	pure STEM	applied STEM
math score	-0.004	0.012***	0.001	0.009**
	(0.004)	(0.004)	(0.004)	(0.004)
verb score	0.007*	-0.001	0.002	0.009*
	(0.004)	(0.004)	(0.005)	(0.005)
science score	0.010***	0.009**	0.011***	-0.001
	(0.003)	(0.004)	(0.004)	(0.005)
mechanical score	-0.003	0.003	-0.004	0.009**
	(0.004)	(0.004)	(0.004)	(0.004)
admin score	0.003	0.001	0.005	-0.006*
	(0.003)	(0.003)	(0.003)	(0.003)
noncog index	0.004**	0.005**	0.007***	-0.001
	(0.002)	(0.002)	(0.002)	(0.003)
father's highest grade completed	0.006	0.052**	-0.006	0.033
	(0.021)	(0.023)	(0.026)	(0.028)
mother's highest grade completed	0.001	-0.018	0.065**	0.086***
	(0.024)	(0.027)	(0.029)	(0.031)
# of high school non-STEM courses taken	-0.067	0.043	-0.142**	-0.135**
	(0.042)	(0.045)	(0.065)	(0.069)
# of high school math courses taken	-0.029	-0.198	0.121	0.471*
	(0.161)	(0.181)	(0.246)	(0.245)
# of high school physical sciences courses taken	-0.002	0.148	0.165	0.144
	(0.190)	(0.211)	(0.370)	(0.384)
# of high school mechanics courses taken	-0.330**	-0.133	-0.088	-0.232
	(0.156)	(0.140)	(0.280)	(0.230)
# of high school comp sci courses taken	-0.310	0.461**	0.464	1.125***
	(0.194)	(0.212)	(0.367)	(0.328)
# of high school biology/health/soc. sciences courses taken	0.194*	-0.164	0.355*	-0.143
	(0.102)	(0.118)	(0.181)	(0.222)
# of high school vocational courses taken	0.042	-0.024	-0.183	0.212
	(0.127)	(0.139)	(0.306)	(0.305)
female = 1	0.903***	-0.279	1.229***	-0.745***
	(0.159)	(0.171)	(0.159)	(0.182)
African American	0.056	-0.012	-0.160	-0.460**
	(0.138)	(0.163)	(0.159)	(0.211)
Observations	3060		3353	
chi2	291.985		357.788	

Figure 5: Classification Tree for Completing Four-Year College (NLSY 79)



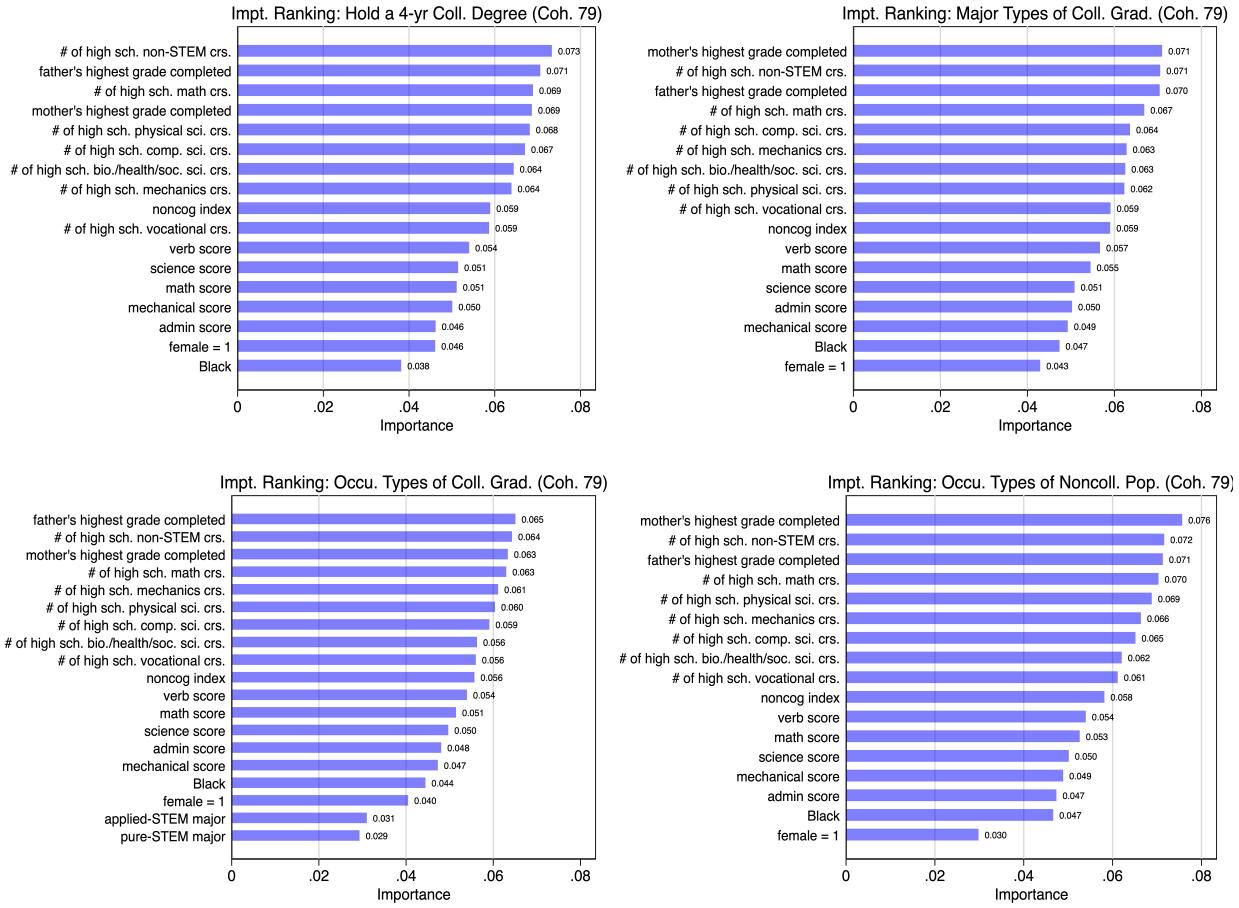
Random forest methods estimate hundreds of trees and combine the estimates through a bagging procedure, so the results can be difficult to visualize. For this reason, we also show results from a single tree estimated using CART, which is similar and illustrative of the tree structures obtained using the random forest methods. We estimate the CART models using data from the NLSY79 cohort. However, the decomposition results (reported below) are based on random forest analysis performed using both the 79 and 97 cohort data.

Figure 5 shows the estimated CART model for the four-year college degree attainment probability. The predicted values conditional on the covariate values denoted in the previous binary split are designated by *value*=. For example, without any conditioning, the probability of being a four-year college graduate is 29.1% and the probability of not being a college graduate is 70.9%. After conditioning only on a binary indicator of whether the percentile math score is less than or equal to 74.46, those with values above the threshold are predicted to go to college with 64.8 probability and those with values below are predicted to go to college with 16.5 probability. The cut-off value (74.46) that best distinguishes between college and non-college groups is chosen by the CART procedure. The overall final cell classification, shown by *class*, is designated *college* whenever the predicted four-year college probability is higher than the predicted "< four-year college" probability.

Having a math score above the 74.46 percentile cut-off is the most important predictor of attaining a four-year college degree. Other variables that also increase the probability of four-year college completion are the number of pure-STEM courses taken in high school and father's highest grade completed. Individuals who score below the 74.46 percentile in math are likely to go to college only if their father has more than 12 years of education. This shows that parental education is an important factor in encouraging youth with lower skill levels to nonetheless attend college. Other important variables are the highest grade of the mother and the verbal test score. Interestingly, the CART procedure does not find gender and race to be key determinants.

The first panel of Figure 6 summarizes the relative contribution of each of the covariates in order of importance in explaining the outcome variable (attaining a four-year college degree) based on the random forest analysis. The most important is the math score, followed by the admin score, verbal score, noncognitive skill index, mechanical score, and science score. In addition, mother's and father's education are important, as was also seen in the logistic regression analysis. Lastly,

Figure 6: Importance Ranking of Variables according to random forests - NLSY 79

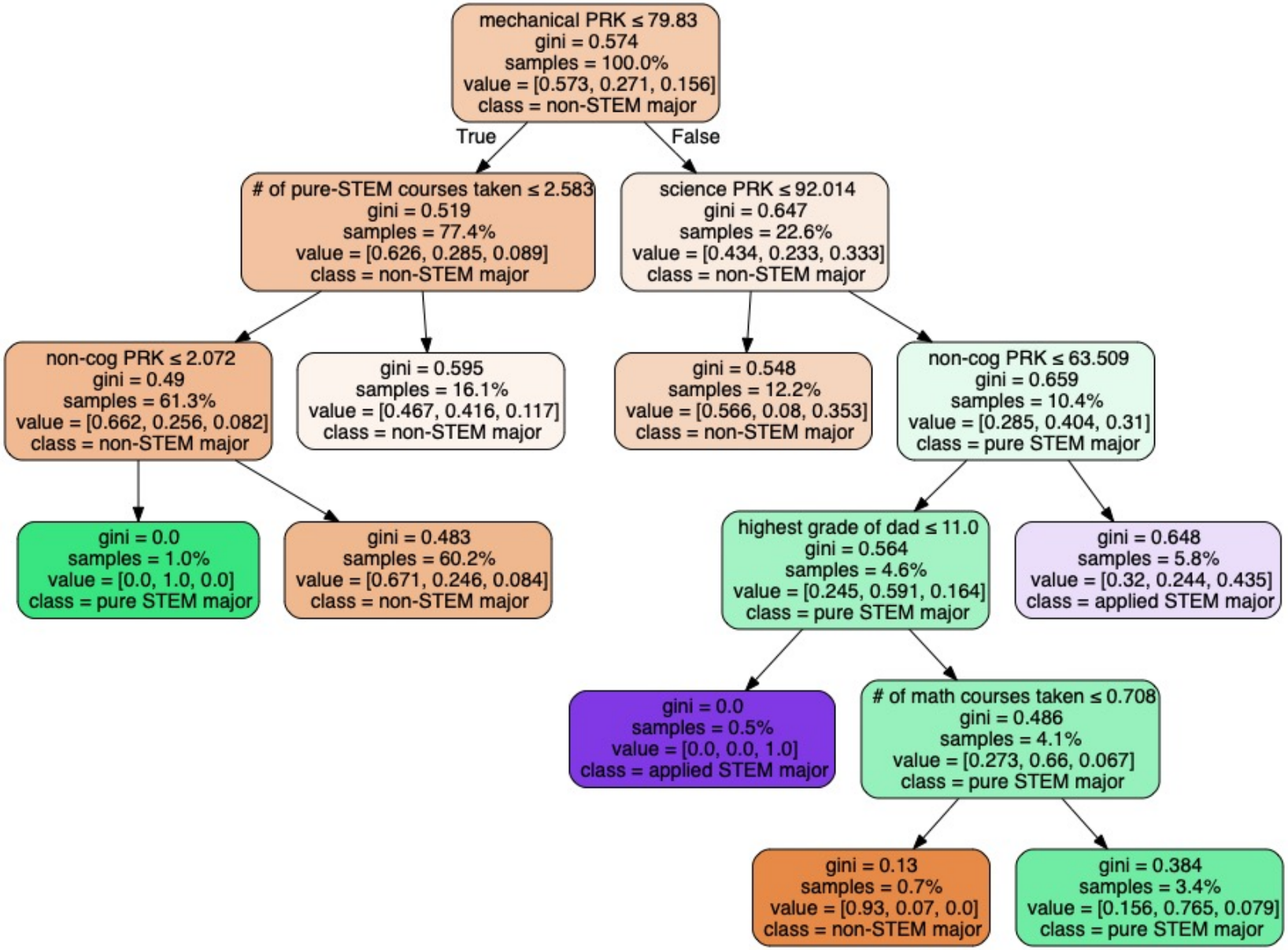


taking more high school courses in science and math and fewer non-STEM courses increases the probability of being a four-year college graduate. Race and gender are relatively less important, conditional on the other covariates, and appear at the bottom of the ranking.

In Figure 7, we examine the relative importance of the same set of variables in predicting college major choices (non-STEM, pure-STEM, or applied-STEM). The *value*= line shown in each box gives the predicted probability of majoring in non-STEM, applied-STEM, or pure-STEM (in order). Having high mechanical and science scores increases the probability of majoring in STEM. Interestingly, the noncognitive skill index is also an important predictor for the non-STEM, pure-STEM or applied-STEM major choice. Conditional on the other covariates, gender and race are not key factors.[25]

[25]It would be interesting to know if a person's parents are in STEM occupations, but the data do not contain this information.

Figure 7: Classification Tree of College Majors



The right upper panel of Figure 6 shows the importance ranking for the college major choice based on the random forest analysis. The key predictors are the test score measures, with mechanical being the most important, followed by administrative, noncognitive, math, verbal, and science skills. Parental education levels and high school course-taking, particularly in science classes, are also highly relevant predictors. Gender and race appear at the bottom of the importance ranking.

Figure 8 shows the estimated CART model for the probability of entering a non-, pure-, or applied-STEM occupation for the subsample with a four-year degree, where the covariates now also include college major choice. The first two binary splits are on college major choice and mechanical test scores. Test scores, especially science and verbal scores, increase the likelihood of entering a STEM occupation. The highest grade of the mom is also important in promoting STEM career choice.

The bottom left panel of Figure 6 shows the importance ranking. The primary predictors are the test score measures, followed by parents' education and high school course-taking in science-related fields. The ranking order is similar to that observed for the college major determinants. Remarkably, the high school test scores appear higher in the importance ranking than having a pure-STEM or applied-STEM major. Race and gender are at the bottom of the rankings.

Lastly, Figure 9 shows the estimated CART model for the occupational choice for the subsample without a four-year degree. The verbal, math, and science test scores are the best predictors of whether a non-college graduate goes into a STEM field. Gender is also important, with females having a higher probability of entering a pure-STEM occupation.

The bottom right panel of Figure 6 shows the importance ranking. A comparison of the right and left panels shows that similar factors predict STEM occupation choices for both the less than four-year college and four-year college graduate subsamples. The test score measures are the primary predictors, particularly the mechanical score, followed by high school course-taking and father's and mother's education. For all three outcome variables (college completion, major choice, occupation choice), gender and race have less predictive power than the other covariates. This suggests that observed gender and racial disparities in entering STEM majors and occupations can be explained in large part by different pre-college skill and course-taking profiles.

Figure 8: Classification Tree for Occupation Choices (NLSY 79 College Graduates)

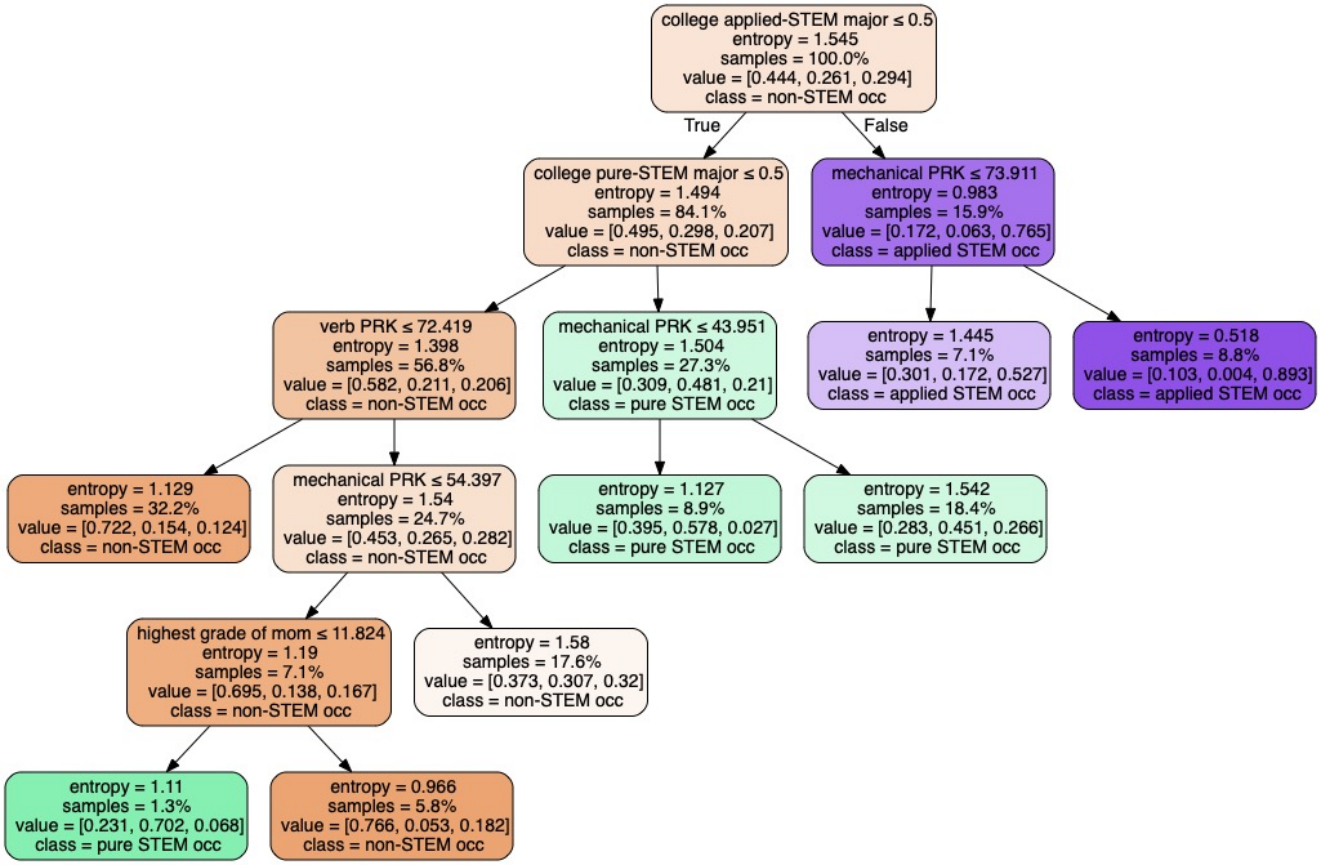
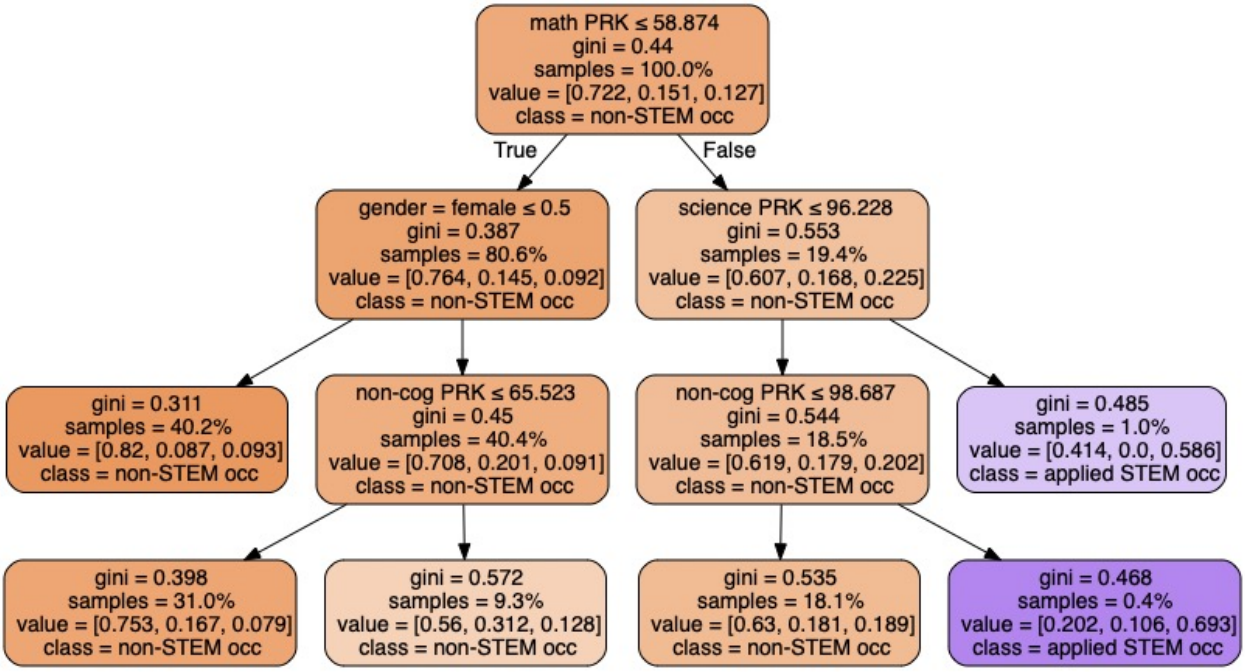


Figure 9: Classification tree for occupation choices (NLSY 79), < four-year college sample)



7.3 Decomposition of gender differences by cohort

We next quantify the extent to which gender disparities in outcomes (four-year college degree attainment, college major, and occupational choice) can be explained by gender differences in pre-college skills and high school course preparation.[26] As previously noted, our decomposition approach extends the classic Oaxaca-Blinder decomposition approach to our nonlinear, multi-period, and nonparametric setting. We perform the decomposition using both the 79 and 97 NLSY datasets. The estimates used to perform the 79 (97) decomposition are obtained using the 79 (97) datasets.

Table 7 shows the decomposition results for the college degree completion outcome. In the data, 28% of males and 31% of females in the 79 birth cohort complete at least a four-year degree. As seen in the table, the predicted values derived from the random forest estimation are the same as those in the data. When we equate the skill distribution of females to be equal to that of males, as shown in the row labeled *Counterfactual*, the predicted degree attainment gap in 79 is unchanged, indicating that the covariates do not explain the gap for the 79 cohort. For the 97 birth cohort, the college completion gap is 14 pp (favoring females) and the random forest estimates predict a

[26]The covariates also include mothers' and fathers' education and race but these variables are similar on average for males and females.

12 pp female advantage. If the covariate distribution were the same for females as for males, then the predicted gap would be 7pp instead of 12pp. Thus, (5/12) or 42% of the gap is explained by gender differences in the pre-college skill and highschool course-taking profiles.

Table 7: Decomposition of four-year college degree attainment gap

	Cohort 79			Cohort 97		
Prop (%) completing college	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	28	31	-3	30	44	-14
Pred. by Model	28	31	-3	31	43	-12
Counterfactual		31	-3		38	-7
Percentage explained by covariates			0			42

Table 8 shows analogous decomposition results when the outcome is college major choice. For the 79 cohort, 51% of male and 61% of female college graduates choose non-STEM majors. Comparing the 79 and 97 cohorts, we see that the proportion choosing non-STEM majors increases by 20 pp for both males and females, with the female-male gap narrowing from 10pp to 8 pp. The decomposition results show that all of the 79 cohort gap and 63% of the 97 cohort gap can be explained by gender differences in early skill and course-taking profiles.

The bottom two panels of the table show the decomposition results for the pure-STEM and applied-STEM major choices. Comparing the two birth cohorts, we see that the percentages of individuals choosing applied-STEM or pure-STEM majors have fallen substantially. In applied-STEM fields, where females are underrepresented, the drop for males was greater than the drop for females. This led to a narrowing of the male-female gap from 16 pp for the 79 cohort to 11 pp for the 97 cohort. The decomposition results show that 64% of the applied-STEM major gap for the 79 cohort and 44% for the 97 cohort can be explained by gender covariate differences. Thus, pre-college skill profiles and course-taking can account for a substantial portion of the gender gap in the applied-STEM major category.

Table 9 presents the decomposition results for occupation sector choices for college graduates. Almost twice as many women as men go into pure-STEM occupations, but about half as many women as men go into applied-STEM occupations. The gender gaps in occupation sectors tend to be larger than the gaps we observed in college major choices. The decomposition results show that pre-college skill profiles and course-taking patterns explain around 33% of the gender gap in the percentage choosing the pure-STEM occupation sector (where females are overrepresented) for the

Table 8: Decomposition of college major gap (among four-year college graduates)

Perc. Majoring	Cohort 79			Cohort 97		
non-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	51	61	-10	71	81	-10
Pred. by Model	51	61	-10	72	80	-8
Counterfactual		50	1		75	-3
Percentage explained by covariates			110			63
pure-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	24	30	-6	15	18	-3
Pred. by Model	25	30	-5	16	17	-1
Counterfactual		31	-6		18	-2
Percentage explained by covariates			-20			-100
applied-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	25	9	16	13	2	10
Pred. by Model	24	10	14	12	3	9
Counterfactual		19	5		7	5
Percentage explained by covariates			64			44

79 cohort and 44% of the gap for the 97 cohort. For the applied-STEM occupation sector (where females are underrepresented), the covariates explain 69% of the gender gap for the 79 cohort and 47% of the gender gap for the 97 cohort. To summarize, more than a third of gender disparities in occupational choices can be predicted from pre-college skill profiles and course-taking backgrounds.

Table 9: Decomposition of the occupational sector gaps - college graduates

Occupation sector	Cohort 79			Cohort 97		
non-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	41	43	-2	46	57	-11
Pred. by Model	40	44	-4	47	56	-9
Counterfactual		36	4		52	-5
Percentage explained by covariates			200			44
pure-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	18	35	-17	15	25	-10
Pred. by Model	21	33	-12	16	25	-9
Counterfactual		29	-8		21	-5
Percentage explained by covariates			33			44
applied-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	40	22	18	39	18	21
Pred. by Model	39	23	16	36	19	17
Counterfactual		34	5		27	9
Percentage explained by covariates			69			47

Lastly, Table 10 shows analogous decomposition occupation sector results for individuals who

do not complete a four-year college degree. Again, there is a pattern that women are more likely than men to enter pure-STEM and less likely to enter applied-STEM occupations. Comparing the 79 and 97 cohorts, we see that the gender gap in the applied-STEM category increases from 4pp for the 79 cohort to 9pp for the 97 cohort. Adolescent skill and course-taking profiles explain roughly a similar portion of the gender gap (30-50%) for individuals that do not complete four years of college as for the subsample that completes a four-year degree.

Table 10: Decomposition of the occupational sector gaps, $<$ 4 years college

Occupation sector	Cohort 79			Cohort 97		
non-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	76	69	7	84	82	2
Pred. by Model	75	69	6	83	81	2
Counterfactual		70	5		81	2
Percentage explained by covariates			17			0
pure-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	9	20	-11	4	15	-11
Pred. by Model	10	20	-10	5	14	-9
Counterfactual		17	-7		11	-6
Percentage explained by covariates			30			33
applied-STEM	Male	Female	Δ_{M-F}	Male	Female	Δ_{M-F}
Data	15	11	4	12	3	9
Pred. by Model	15	11	4	11	4	7
Counterfactual		13	2		7	4
Percentage explained by covariates			50			43

8 Conclusions

Women are less likely than men to graduate with a degree or to enter an occupation in an applied STEM field. As STEM majors and occupations are some of the highest paying, the under-representation of women is thought to be a significant factor underlying the gender wage gap. For example, Cortes and Pan (2018) show that occupation and industry segregation continues to be an important driver of gender wage gaps.

Previous studies showed that gender differences are evident in high school course-taking patterns and in skills measured prior to entering college. For example, Card and Payne (2021) find that women have lower levels of STEM-readiness in Ontario, Canada. Gender differences in STEM

readiness upon leaving high school can affect college degree attainment, college major choice, and occupational sector choice. Using data from the NLSY79 and 97 cohorts, we present evidence on how five cognitive skill measures and one non-cognitive measure compare for males and females and analyze how skill disparities change over a twenty-year time span. Although women made progress in closing skill gaps in mathematics and in reversing skill gaps in verbal, there remains a substantial gender gap in mechanical skills and, to a lesser extent, science skills. High school course-taking patterns also differ by gender, with female high school students taking more courses in biology/social/health science fields and fewer courses in physics and computer science. Females generally have greater levels of preparation for college than males in the NLSY97 cohort and go to college at substantially higher rates (44% versus 30%). At the same time, though, female high school graduates, on average, show some deficiencies in mechanical skills and in physical science coursework. We find these characteristics to be important determinants for entry into applied-STEM majors and occupations.

To analyze how gender differences in high school preparation contribute to later educational and employment outcomes, we developed a statistical model of the sequential decisions to obtain a four-year college degree, major in STEM or non-STEM fields, and to enter a STEM or non-STEM occupation. Each of these conditional probabilities is estimated using logistic regression and using CART and random forest methods, which yielded interesting insights about what kinds of adolescent skill profiles and course-taking backgrounds lead to STEM majors and careers. Although both the logistic regression and the nonparametric models identify many of the same key predictor variables, there are also some important differences. For example, race and gender tend to be less important in the random forest estimation, and they appear at the bottom of the importance rankings for most of the outcomes. We also find that the fit of the nonparametric models is better than that of the logistic regression models.

The CART and random forest results show that math skills are the primary predictors of completing a four-year college degree, followed by other test score measures (administrative, verbal, mechanical, and science) and the noncognitive skill index. Father's and mother's education and high school course-taking in science-related areas are also important. An interesting pattern that becomes apparent from the tree analysis is that higher parental education facilitates college degree attainment for individuals with lower test scores.

The random forest analysis for college major choice, from among three options (non-STEM, pure-STEM, and applied-STEM), shows that the mechanical score is the most important predictor, followed by the admin, math, verbal and science scores. Noncognitive skills also predict college majors. Individuals who have high mechanical ability and also take a larger number of physical science courses in high school are more likely to enter applied STEM majors. Thus, women's lower average mechanical ability and lower average numbers of high school physical sciences courses can partly explain their underrepresentation in applied-STEM majors.

Our analysis of occupational choices shows that, for college graduates, college major type and high school test scores, especially the mechanical score, predict whether individuals work in a STEM occupation. Not surprisingly, individuals who major in pure-STEM or applied-STEM fields are much more likely to enter occupations in these areas. For individuals who do not complete four-year college degrees, high school math and science test scores are key predictors of entering a STEM occupation.

Our decomposition analysis examined how much of the gender gap in educational and occupational outcomes can be explained by pre-college skill profiles/course-taking. It extends the Oxaca-Blinder decomposition method to our setting involving multiple discrete, sequential outcomes (college degree, major, occupation), where the conditional probability of each outcome is estimated nonparametrically. The results show that the gender gap in four-year college degree attainment, which favors women, widened substantially between the 79 and 97 cohorts and that about 42% of the gap can be explained by differences in skill profiles/course-taking. With regard to college major choice, the gender gap is large in the applied-STEM category, with 11% of male college graduates choosing such majors in comparison to 2% of women. About 44% of the disparity can be explained by the early skills/course-taking, notably women taking fewer physical science and mechanics courses and having lower mechanical test scores. Lastly, women and men have different occupational sector distributions, with more women entering pure-STEM fields and fewer women entering applied-STEM fields. About 30-50% of the STEM occupational sector differences can be explained by the observed covariates.[27]

A key finding from our analysis is that mechanical skills play an important role in the deci-

[27] Although the focus of this paper is on the gender gap, the results indicated that African American race was often not a significant predictor. It would be useful to investigate whether the under-representation of African Americans in STEM fields may also be partly explained by differences in skills, course-taking and family background.

sion to major in an applied-STEM field and to enter an applied-STEM occupation. The prior literature mostly emphasizes math skills. Our analysis shows that although math skill is an important determinant of four-year college completion, it is not the primary factor in explaining entry into applied-STEM fields. Individuals who pursue applied-STEM majors and careers tend also to have strong mechanical skills. Although women have largely closed the gender gap in math and outperform men in verbal, administrative, and noncognitive domains, there remains a gender gap favoring men in mechanical and, to a lesser extent, in science skills. Efforts to increase women's participation in applied-STEM might therefore consider ways of fostering skill acquisition in these areas.

References

Addison, John T, Liwen Chen, and Orgul D Ozturk, "Occupational Skill Mismatch: Differences by Gender and Cohort," *ILR Review*, 2020, *73* (3), 730–767.

Altonji, Joseph G, Prashant Bharadwaj, and Fabian Lange, "Changes in the characteristics of American youth: Implications for adult outcomes," *Journal of Labor Economics*, 2012, *30* (4), 783–828.

Autor, David and David Dorn, "The growth of low-skill service jobs and the polarization of the US labor market," *American economic review*, 2013, *103* (5), 1553–97.

Becker, Gary S, William HJ Hubbard, and Kevin M Murphy, "Explaining the worldwide boom in higher education of women," *Journal of Human Capital*, 2010, *4* (3), 203–241.

Blau, Francine D and Lawrence M Kahn, "The gender wage gap: Extent, trends, and explanations," *Journal of economic literature*, 2017, *55* (3), 789–865.

Blinder, Alan S, "Wage discrimination: reduced form and structural estimates," *Journal of Human resources*, 1973, pp. 436–455.

Breda, Thomas and Clotilde Napp, "Girls' comparative advantage in reading can largely explain the gender gap in math-related fields," *Proceedings of the National Academy of Sciences*, 2019, *116* (31), 15435–15440.

Breiman, Leo, "Random forests," *Machine learning*, 2001, *45* (1), 5–32.

Bronson, Mary Ann, "Degrees are forever: Marriage, educational investment, and lifecycle labor decisions of men and women," *Unpublished manuscript*, 2014, *2*.

Card, David and A Abigail Payne, "High school choices and the gender gap in STEM," *Economic Inquiry*, 2021, *59* (1), 9–28.

Ceci, Stephen J, Donna K Ginther, Shulamit Kahn, and Wendy M Williams, "Women in academic science: A changing landscape," *Psychological science in the public interest*, 2014, *15* (3), 75–141.

Chuan, Amanda and Weilong Zhang, "Non-College Occupations, Workplace Routinization, and the Gender Gap in College Enrollment," 2021.

Cohodes, Sarah R, Helen Ho, and Silvia C Robles, "Stem summer programs for underrepresented youth increase stem degrees," Technical Report, National Bureau of Economic Research 2022.

Deming, David J, "The growing importance of social skills in the labor market," *The Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640.

_ **and Kadeem L Noray**, "STEM careers and technological change," 2018.

Diekhoff, George, *Statistics for the social and behavioral sciences: Univariante, bivariate, multivariante* number HA29. D46 1992. 1992.

DiNardo, John, Nicole Fortin, and Thomas Lemieux, "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach," 1995.

Dorn, David, "Essays on inequality, spatial interaction, and the demand for skills." PhD dissertation, Verlag nicht ermittelbar 2009.

Flood, Sarah M, "Harmonizing the 2010 and 2002 Census Occupation Coding Schemes," 2019.

Gemici, Ahu and Matthew Wiswall, "Evolution of gender differences in post-secondary human capital investments: College majors," *International Economic Review*, 2014, *55* (1), 23–56.

Gnecco, Giorgio, Sara Landi, and Massimo Riccaboni, "Can Machines Learn Creativity Needs? An Approach Based on Matrix Completion," *Italian Economic Journal*, 2022, pp. 1–41.

Goldin, Claudia, "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, April 2014, *104* (4), 1091–1119.

_ , *Career and Family: Women's Century-Long Journey toward Equity*, Princeton University Press, 2021.

Guan, Xin, Chang-Tsun Li, and Yu Guan, "Matrix factorization with rating completion: An enhanced SVD model for collaborative filtering recommender systems," *IEEE access*, 2017, *5*, 27668–27678.

Guvenen, Fatih, Burhan Kuruscu, Satoshi Tanaka, and David Wiczer, "Multidimensional skill mismatch," *American Economic Journal: Macroeconomics*, 2020, *12* (1), 210–44.

Hastie, Trevor and Rahul Mazumder, "softImpute: Matrix Completion via Iterative Soft-Thresholded SVD, 2021," *URL https://cran.r-project.org/web/packages/softImpute/softImpute.pdf*, 2021, *1*.

_ , _ , Jason D Lee, and Reza Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares," *The Journal of Machine Learning Research*, 2015, *16* (1), 3367–3402.

Heckman, James J, Jora Stixrud, and Sergio Urzua, "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor economics*, 2006, *24* (3), 411–482.

Ho, Tin Kam, "Random decision forests," in "Proceedings of 3rd international conference on document analysis and recognition," Vol. 1 IEEE 1995, pp. 278–282.

Hyde, Janet S and Janet E Mertz, "Gender, culture, and mathematics performance," *Proceedings of the National Academy of Sciences*, 2009, *106* (22), 8801–8807.

_ , Elizabeth Fennema, and Susan J Lamon, "Gender differences in mathematics performance: a meta-analysis.," *Psychological bulletin*, 1990, *107* (2), 139.

_ , Sara M Lindberg, Marcia C Linn, Amy B Ellis, and Caroline C Williams, "Gender similarities characterize math performance," *Science*, 2008, *321* (5888), 494–495.

Kahn, Shulamit and Donna Ginther, "Women and STEM," Technical Report, National Bureau of Economic Research 2017.

Legewie, Joscha and Thomas A DiPrete, "School context and the gender gap in educational achievement," *American Sociological Review*, 2012, *77* (3), 463–485.

Loh, Wei-Yin, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 2011, *1* (1), 14–23.

Miller, David I and Diane F Halpern, "The new science of cognitive sex differences," *Trends in cognitive sciences*, 2014, *18* (1), 37–45.

Mourifie, Ismael, Marc Henry, and Romuald Meango, "Sharp bounds and testability of a Roy model of STEM major choices," *Journal of Political Economy*, 2020, *128* (8), 3220–3283.

Neal, Derek A and William R Johnson, "The role of premarket factors in black-white wage differences," *Journal of political Economy*, 1996, *104* (5), 869–895.

Oaxaca, Ronald, "Male-female wage differentials in urban labor markets," *International economic review*, 1973, pp. 693–709.

Peri, Giovanni, Kevin Shih, and Chad Sparber, "STEM workers, H-1B visas, and productivity in US cities," *Journal of Labor Economics*, 2015, *33* (S1), S225–S255.

Prada, María F and Sergio Urzúa, "One size does not fit all: Multiple dimensions of ability, college attendance, and earnings," *Journal of Labor Economics*, 2017, *35* (4), 953–991.

Raileanu, Laura Elena and Kilian Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, 2004, *41* (1), 77–93.

Ricci, Francesco, Lior Rokach, and Bracha Shapira, "Introduction to recommender systems handbook," in "Recommender systems handbook," Springer, 2011, pp. 1–35.

Sloane, Carolyn M, Erik G Hurst, and Dan A Black, "College Majors, Occupations, and the Gender Wage Gap," *Journal of Economic Perspectives*, 2021, *35* (4), 223–48.

Stinebrickner, Ralph and Todd R Stinebrickner, "A major in science? Initial beliefs and final outcomes for college major and dropout," *Review of Economic Studies*, 2014, *81* (1), 426–472.

Stoet, Gijsbert and David C Geary, "The gender-equality paradox in science, technology, engineering, and mathematics education," *Psychological science*, 2018, *29* (4), 581–593.

Valentino, Lauren, Stephanie Moller, Elizabeth Stearns, and Roslyn Mickelson, "Perceptions of future career family flexibility as a deterrent from majoring in STEM," *Social Currents*, 2016, *3* (3), 273–292.

Wai, Jonathan, Jaret Hodges, and Matthew C Makel, "Sex differences in ability tilt in the right tail of cognitive abilities: A 35-year examination," *Intelligence*, 2018, *67*, 76–83.

Wang, Ming-Te and Jessica Degol, "Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields," *Developmental Review*, 2013, *33* (4), 304–340.

_ , Jacquelynne S Eccles, and Sarah Kenny, "Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics," *Psychological science*, 2013, *24* (5), 770–775.

Wiswall, Matthew and Basit Zafar, "Determinants of college major choice: Identification using an information experiment," *The Review of Economic Studies*, 2015, *82* (2), 791–824.

_ and _ , "Preference for the workplace, investment in human capital, and gender," *The Quarterly Journal of Economics*, 2018, *133* (1), 457–507.

Zhou, Xun, Jing He, Guangyan Huang, and Yanchun Zhang, "SVD-based incremental approaches for recommender systems," *Journal of Computer and System Sciences*, 2015, *81* (4), 717–733.

A Sample inclusion criteria and treatment of missing variables

The NLSY79 is a nationally representative sample of 12,686 young men and women born during the years 1957 through 1964 and living in the United States at the time of the initial survey in 1979. The NLSY97 consists of a nationally representative sample of 8,984 men and women born during the years 1980 through 1984 and living in the United States at the time of the initial survey in 1997. For the NLSY79, we exclude individuals who took the tests after age 19, because the 97 cohort does not include such individuals and to measure test scores that are free from post-college skill accumulation effects. Also, to make the two cohorts more comparable, the two extra sub-samples of the NLSY79 that are not available in the NLSY97 are also omitted (a military sample and an economically disadvantaged non-black, non-Hispanic oversample). After these exclusions, there are 5246 observations remaining in the 79 cohort.

Table 11 provides information on missing variables, some of which were imputed to avoid dropping too many observations. As skill measures are central to our analysis, we exclude individuals

who are missing more than nine scores on the ASVAB tests. We also exclude individuals for whom high school course-taking information is not available. For individuals who are missing partial information on the ASVAB tests or some of the parental background information, we impute missing data values using the method of Matrix Completion via Iterative Soft-Thresholded SVD (Hastie et al. 2015, Hastie and Mazumder 2021).[28] This method is widely used in websites' recommender systems where the algorithm comes up with informative recommendations by inferring users' preferences through matrix completion and then suggesting items to them, based on the inferred preferences (e.g., Ricci et al. 2011, Zhou et al. 2015, Guan et al. 2017, Gnecco et al. 2022, etc.). This technique is motivated by the ability of matrix completion to capture possibly hidden relationships among elements of a rating matrix (e.g., in the case of a movie-rating matrix, to capture the similarity between movies, or between users' preferences). A similar argument can be applied in the case of our matrix of early-life skills.

The final analysis samples comprise 4051 individuals from the 1979 cohort and 5100 individuals from the 1997 cohort who are followed longitudinally, which permits examination of their educational and occupational choices.

[28] We use the R *softImpute* package.

Table 11: Numbers of observations with missing data items on the explanatory variables (Source: NLSY 79 and 97)

	Cohort 79	Cohort 97
Obs All	5246	8984
Obs removed because of missing 9 ASVAB test items	225	1857
Obs removed because of lacking high school course info.	970	2027
Obs missing in the dataset to be imputed		
ASVAB tests		
arithmetic reasoning	0	8
mathematics knowledge	0	21
word knowledge	0	10
paragraph comprehension	0	13
autoshop information	0	37
mechanical comprehension	0	42
electronics information	0	34
general science	0	0
numerical operations	0	87
coding speed	0	86
non-cog index	3	397
age when taking the test	0	0
father's highest grade completed	562	905
mother's highest grade completed	213	318
Obs used	4051	5100

B Principal Component Analysis (PCA)

This section explains how we use principal component analysis (PCA) to aggregate the ten ASVAB test scores into five broader categories (as shown in Figure 1). Following Prada and Urzúa (2017), we perform a PCA analysis on the two NLSY cohorts' ASVAB test scores separately to detect clusters of variables (i.e., exam items) that, to a large extent, define only one skill component.[29] A difference between our approach and that of the previous literature is that we perform PCA with component rotation, which yields fewer factor loadings that are large in absolute value.[30]

Our PCA analysis shows that it is necessary to separate the item "general science" from the other test items relating to practical skills, e.g., auto/shop information, mechanical comprehension, etc. The resulting taxonomy of skill profiles refines the category of practical skills into "science" and "mechanical," which is consistent with the finer classification of STEM fields into pure STEM and applied STEM fields. Below, we explain the procedure we use to construct the skill measures.

B.1 Scree plot

PCA analysis in previous literature (i.e., Prada and Urzúa 2017, Chuan and Zhang 2021) suggests that two separate skills (i.e., components), mainly cognitive and practical, are necessary to explain the ASVAB score variation, but we go beyond two components for two reasons. First, the goal of identifying the key determinants of choosing STEM majors and occupations calls for a finer skill classification. Second, although the Kaiser's rule criterion recommends retaining only components with eigenvalues exceeding one, according to which only the first two components should be kept (as shown in Figure 10), there are alternative approaches for choosing the number of components. One the so-called "elbow method" that checks whether there is a "break" in the scree plot with the remaining components explaining considerably less variation. Examining the scree plot in Figure 10, we can see that the most discernible elbow is at the third component and that the cumulative percentage of variation explained continues to grow almost linearly along with the second and third components until after the fourth one. This provides a justification for using more than two components to capture the variation in the original skill measures.

[29] See also other studies in the previous literature applying PCA analysis in this context, such as Deming 2017, Chuan and Zhang 2021, and Prada and Urzúa 2017)

[30] We employ the Kaiser's varimax rotation, an orthogonal rotation (preserving the independence of the components) that maximizes the squared loadings variance across variables summed over all factors.

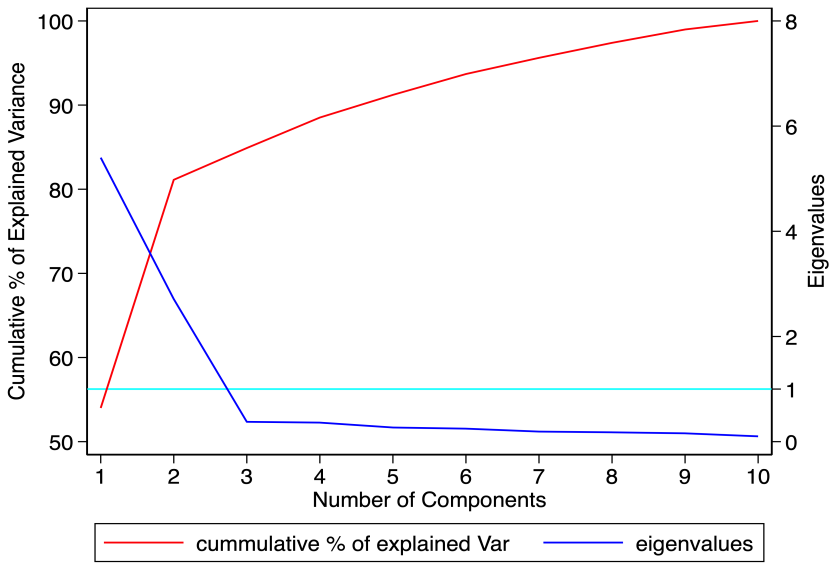


Figure 10: Scree Plot of Eigenvalues and Cumulative Variance Explained After PCA

Note: This graph plots the cumulative percentage of explained variance as well as eigenvalues according to the number of components.

B.2 Rotated loadings of principle components

Figure 11 plots the rotated loadings of principal components for cases where the number of components rotated is $2, 3, 4, 5$ (out of 10) respectively. Note that the two lines $x = 0.3$ and $y = 0.3$ are added on each graph to demarcate the statistically significant and cleanly distinguishable regions as the upper left and lower right corners.[31] There are two points worth mentioning regarding Figure 11. First, once more than two components are considered, the subtest science clusters with sub-tests different from those considered to test mechanical skills (i.e., automotive and shop information, electronics information, and mechanical comprehension). Such clustering pattern justifies our departure from previous literature in distinguishing science from other ASVAB sub-tests that are regarded as tests on mechanical skills.

Second, clusters of other variables are easily distinguishable as there are barely any dots contained in the upper right regions. Specifically, number operation and coding speed are clustered together, and so are math knowledge and arithmetic reasoning, word knowledge and paragraph comprehension along with general science, mechanical comprehension, and electronics information

[31]Note that factor loadings exceeding 0.3 can be regarded as statistically significant (see Diekhoff 1992) and that the goal of factor rotation is to make most factor loadings on any specific factor small while only a few factor loadings large in absolute value. Hence through rotation, we hope to see mutually exclusive clusters of components that have loadings above 0.3.

along with automotive and shop information. Hence, we cluster the ten subtests into the five skill categories shown in Figure 1.

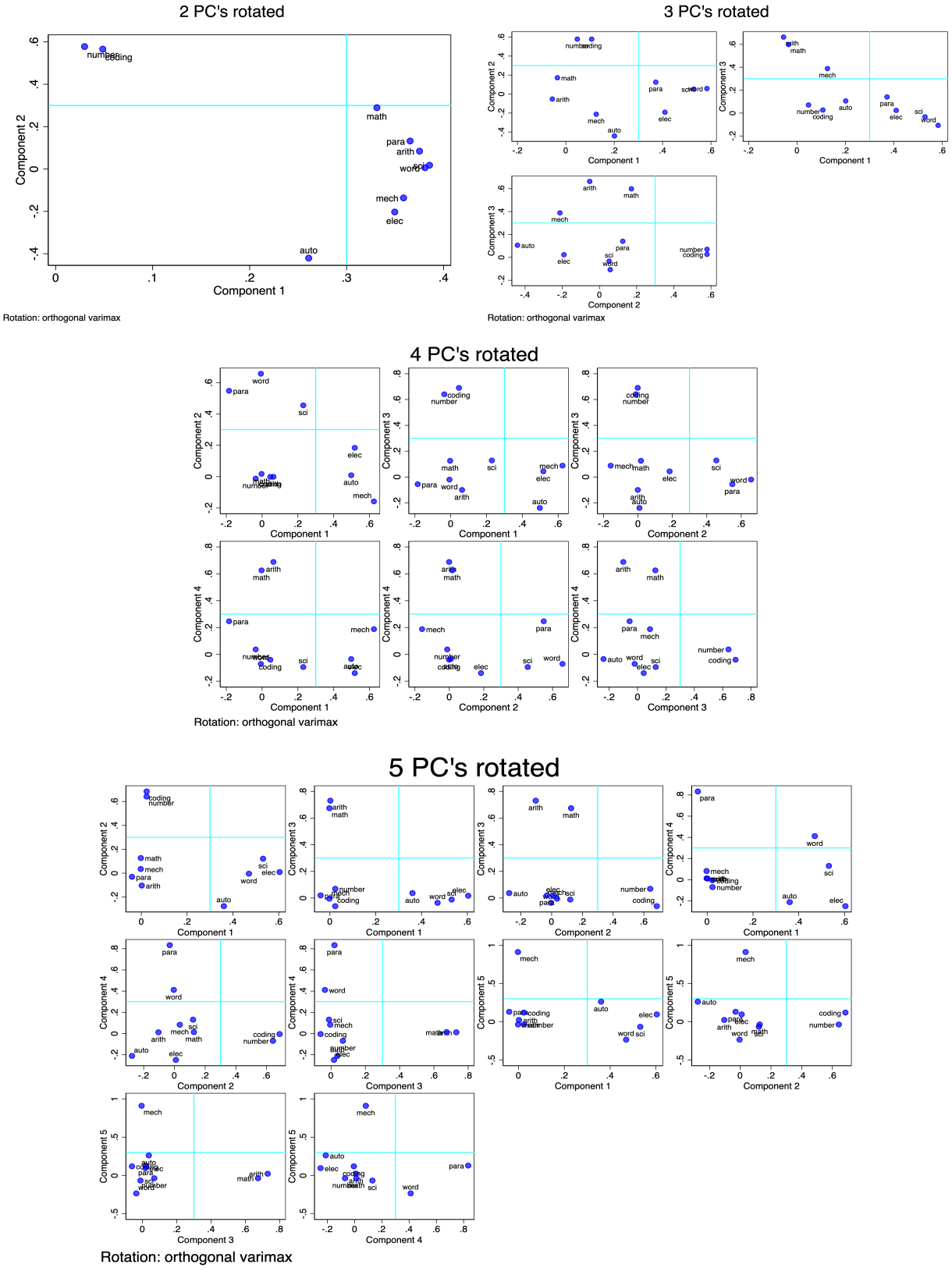


Figure 11: Scatter Plot of Component Loadings

Note: This graph plots the rotated weighting of principal components (PC) with respect to different numbers of PC's applied.

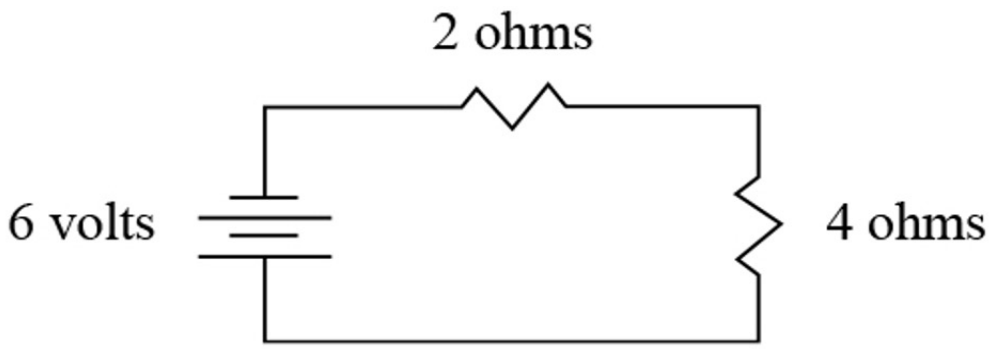
C Sample questions of the mechanical skill measurements in ASVAB

In light of the results from the PCA analysis, we take the simple average of the scores on *mechanical comprehension*, *electronics information*, and *auto/shop information* in the ASVAB test to represent one's mechanical skills. Below we present sample questions for each of the items.[32]

C.1 Electronics Information

Below are a few sample questions for the Electronics Information portion of the ASVAB, focused on electricity and electronics. The correct answers are underlined.

1. The current in the 2-ohm resistor is



   a) 1/3 amp.

   b) 1/2 amp.

   c) 1 amp.

   d) 3 amps.

2. Because solid state diodes have no filament, they

   a) don't work.

   b) are less efficient than tubes.

   c) require less operating power.

   d) require more operating power.

[32] Please refer to https://www.airforce.com/ for further information.

3. Which of the following is measured using an ohmmeter?

    a) voltage

    b) resistance

    c) inductance

    d) capacitance

4. What does the abbreviation AC stand for?

    a) additional charge

    b) alternating coil

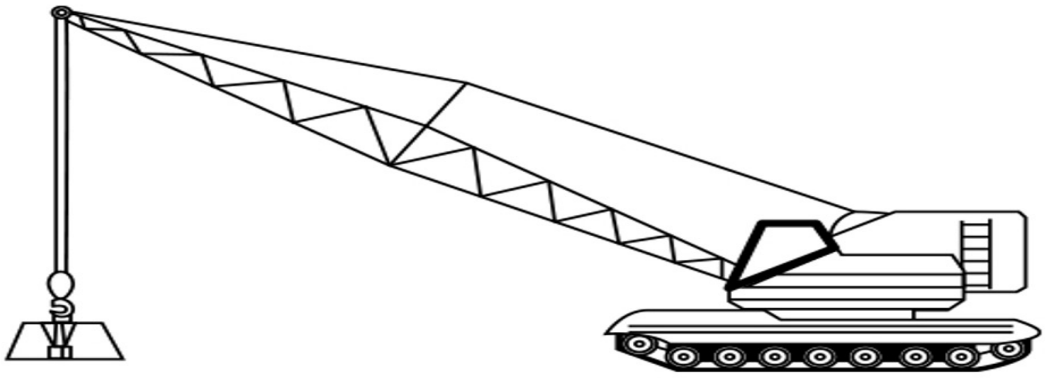
    c) alternating current

    d) ampere current

C.2 Mechanical Comprehension

Below are a few sample questions for the Mechanical Comprehension portion of the ASVAB, focused on mechanical and physical principles. The correct answers are underlined.

1. Extending the reach of this crane will shift its

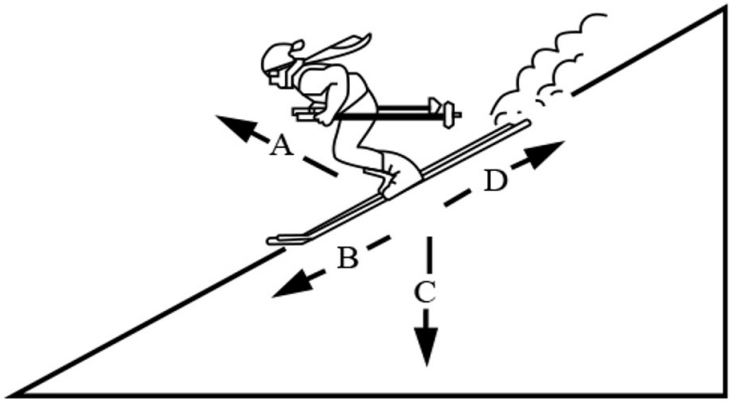


    a) total weight.

    b) allowable speed.

    c) center of gravity.

    d) center of buoyancy.



2. In which direction does friction act on this skier?

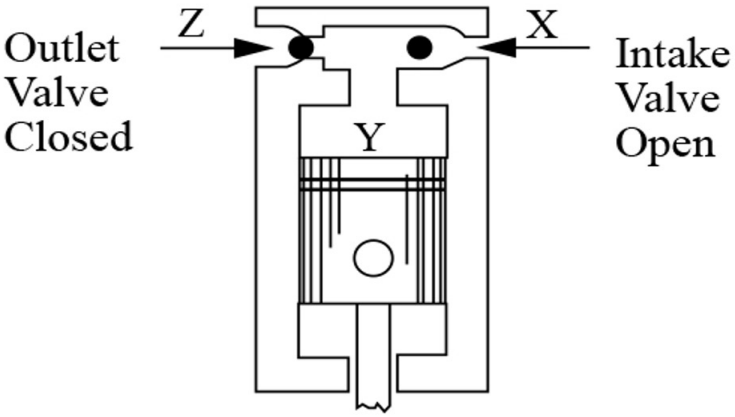
   a) A

   b) B

   c) C

   d) D

3. Why does the intake valve open on this pump when the piston goes down?



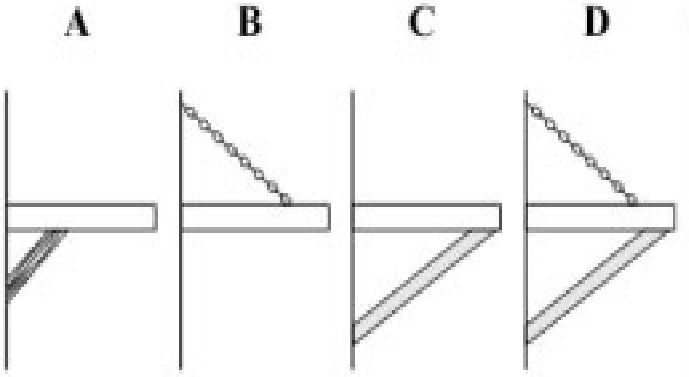
   a) Air pressure at X is less than air pressure at Y.

   b) Air pressure at Z is less than air pressure at X.

   c) Air pressure at X is greater than air pressure at Y.

   d) Air pressure at Y is greater than air pressure at Z.

4. Which shelf will support the most weight?



a) A

b) B

c) C

d) <u>D</u>

C.3 Auto Information

Below are a few sample questions for the Auto Information portion of the ASVAB, focused on automobile technology. The correct answers are underlined.

1. Shock absorbers on a car connect the axle to the

   a) wheel.

   b) <u>chassis</u>.

   c) drive shaft.

   d) exhaust pipe.

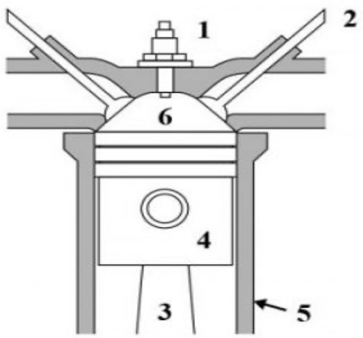
2. Where does combustion take place?

   a) 3

   b) 4

   c) 5

   d) <u>6</u>



3. If a car is driven mostly in a city so that the spark plug tip never gets hot enough to burn off excess carbon in the cylinder, then

   a) lower voltage should be applied to the spark plug.

   b) the tip should be pulled farther out of the cylinder.

   c) the tip should be pushed farther into the cylinder.

   d) a thicker conducter in the plug must be used.

4. The radiator on a car is part of the

   a) cooling system.

   b) ignition system.

   c) electrical system.

   d) lubrication system.

C.4 Shop Information

Below are a few sample questions for the Shop Information portion of the ASVAB, focused on tools and shop terminology and practices. The correct answers are underlined.

1. Sanding blocks are used to

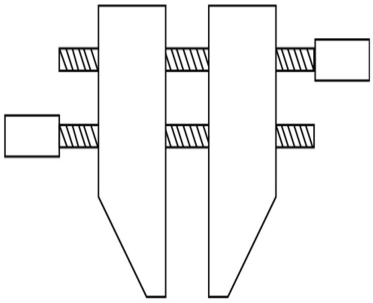
   a) prevent high spots and ridges on sanded surfaces.

   b) prevent dirt from collecting on the sandpaper.

   c) stretch the length of the sandpaper.

   d) prolong the use of the sandpaper.

2. The clamp shown is called a



a) bar clamp.

b) web clamp.

c) spring clamp.

d) parallel clamp.

3. Which handsaw should be used to make a rough cut with the grain in a thick board?

a) ripsaw

b) backsaw

c) coping saw

d) crosscut saw

D Adjustment for age effects

In the NLSY survey, individuals take the tests only once and the age at which they take the test varies. To make the scores comparable across individuals, we account for age effects. To do so, we estimate OLS regressions of the raw scores on age indicators from age 14 (the youngest age) up to age 19 (the oldest age), separately for which NLSY cohort (79 and 97):

$$y_i = \alpha_0 + \alpha_1 \times age14_i + \alpha_2 \times age15_i + \alpha_3 \times age16_i + \alpha_4 \times age17_i + \alpha_5 \times age18_i + \alpha_6 \times age19_i + \epsilon_i,$$

where y_i is the test score. If someone was age 14 when they took the test, we subtract the estimated coefficient on the $age14$ indicator and add the estimated coefficient on the $age19$ indicator, i.e.,

$$y_i^{new} = y_i - \hat{\alpha}_1 \times age14_i + \hat{\alpha}_6 \times age19_i.$$

In this way, test scores are being compared across individuals at a comparable age.

E Categorizing high school courses, college majors, and occupations

High school courses We categorize high school courses into seven categories - (1) non-STEM courses (such as English and history), (2) math, (3) physical sciences such as physics and chemistry, (4) mechanics courses (such as farm and agricultural mechanics), (5) computer courses (such as computer operations and programming), (6) biology/health/social science (related courses such as biology and agriculture), and (7) vocational training courses. Among these, (4) and (5) belong to the applied-STEM category, and (2), (3), and (6) belong to the pure-STEM category.

College majors Following Peri et al. (2015), we adopt the definition of STEM major used by the U.S. Department of Homeland Security in determining visitor eligibility for an F-1 Optional Practical Training (OPT) extension.[33] We further divide the STEM majors into pure and applied STEM subcategories. Table 12 and 13 shows the categorization of college majors of NLSY79 and NLSY97 respectively.[34]

Occupations The classification for occupations is taken from the Bureau of Labor Statistics' definition of STEM occupations under the 2010 Standard Occupational Classification (SOC) system.[35] Under this scheme, there are four types of STEM occupations, (1) Life and physical science, engineering, mathematics, and information technology occupations, (2) social science occupations, (3) architecture occupations, and (4) health occupations. We manually coded the first category's engineering and information technology occupations as "applied-STEM" occupations and the rest as "pure-STEM". Table 14 lists all the applied-STEM occupations under the 2010 Standard Occupational Classification (SOC) system.

We further collapse the categorization on the *SOC2010* level to the *occ1990dd_stem* level using the crosswalk obtained from the IPUMS website.[36] Note that *occ1990dd_stem* builds upon *occ1990dd*[37] from Dorn (2009) and is a slightly finer occupation classification than *occ1990dd* to

[33] https://www.ice.gov/sites/default/files/documents/Document/2016/stem-list.pdf

[34] Please refer to https://www.nlsinfo.org/content/cohorts/nlsy79/other-documentation/codebook-supplement/nlsy79-attachment-4-fields-study for further details of college majors.

[35] https://www.bls.gov/soc/AttachmentCSTEM.pdf

[36] https://forum.ipums.org/t/occ1990-to-soc-crosswalk/2098

[37] Please refer to https://www.ddorn.net/data/DornThesisAppendix.pdf for the construction of *occ1990dd*.

Table 12: Categories of College Majors (NLSY79)

Major	Category
0000 None, General Studies	non-STEM
0100 Agriculture and Natural Resources	pure-STEM
0200 Architecture and Environmental Design	pure-STEM
0300 Area Studies	non-STEM
0400 Biological Sciences	pure-STEM
0500 Business and Management	non-STEM
0600 Communications	non-STEM
0700 Computer and Information Sciences	applied-STEM
0800 Education	non-STEM
0900 Engineering	applied-STEM
1000 Fine and Applied Arts	non-STEM
1100 Foreign Languages	non-STEM
1200 Health Professions	pure-STEM
1300 Home Economics	non-STEM
1400 Law	non-STEM
1500 Letters	non-STEM
1600 Library Science	non-STEM
1700 Mathematics	pure-STEM
1800 Military Sciences	applied-STEM
1900 Physical Sciences	pure-STEM
2000 Psychology	non-STEM
2100 Public Affairs and Services	non-STEM
2200 Social Sciences	pure-STEM
2300 Theology	non-STEM
4900 Interdisciplinary Studies	non-STEM
9994 9995 9996	non-STEM

Table 13: Categories of College Majors (NLSY97)

Major	Category
01 Agriculture, Agriculture Operations, and Related Services	pure-STEM
03 Natural Resources and Conservation	pure-STEM
04 Architecture and Related Services	pure-STEM
05 Area, Ethnic, Cultural, Gender, and Group Studies	non-STEM
09 Communication, Journalism, and Related Programs	non-STEM
10 Communications Technologies/Technicians and Support Services	applied-STEM
11 Computer and Information Sciences and Support Services	applied-STEM
12 Personal and Culinary Services	non-STEM
13 Education	non-STEM
14 Engineering	applied-STEM
15 Engineering Technologies and Engineering-Related Fields	applied-STEM
16 Foreign Languages, Literatures, and Linguistics	non-STEM
19 Family and Consumer Sciences/Human Sciences	non-STEM
22 Legal Professions and Studies	non-STEM
23 English Language and Literature/Letters	non-STEM
24 Liberal Arts and Sciences, General Studies and Humanities	non-STEM
25 Library Science	non-STEM
26 Biological and Biomedical Sciences	pure-STEM
27 Mathematics and Statistics	pure-STEM
28 Military Science, Leadership and Operational Art	applied-STEM
29 Military Technologies and Applied Sciences	applied-STEM
30 Multi/Interdisciplinary Studies	non-STEM
31 Parks, Recreation and Leisure Studies	non-STEM
32 Basic Skills and Developmental/Remedial Education	non-STEM
33 Citizenship Activities	non-STEM
34 Health-Related Knowledge and Skills	pure-STEM
35 Interpersonal and Social Skills	non-STEM
36 Leisure and Recreational Activities	non-STEM
37 Personal Awareness and Self-Improvement	non-STEM
38 Philosophy and Religious Studies	non-STEM
39 Theology and Religious Vocations	non-STEM
40 Physical Sciences	pure-STEM
41 Science and Technologies/Technicians	applied-STEM
42 Psychology	non-STEM
43 Homeland Security, Law Enforcement, Firefighting and Related Protective Services	non-STEM
44 Public Administration and Social Service Professions	non-STEM
45 Social Sciences	pure-STEM
46 Construction Trades	applied-STEM
47 Mechanic and Repair Technologies/Technicians	applied-STEM
48 Precision Production	applied-STEM
49 Transportation and Materials Moving	applied-STEM
50 Visual and Performing Arts	non-STEM
51 Health Professions and Related Programs	pure-STEM
52 Business, Management, Marketing, and Related Support Services	non-STEM
53 High School/Secondary Diplomas and Certificates	non-STEM
54 History	non-STEM
60 Residency Programs	pure-STEM
90 Other Courses	non-STEM

Table 14: Applied-STEM occupations (SOC 2010)

SOC 2010	Job Title
11-3021	Computer and Information Systems Managers
11-9041	Architectural and Engineering Managers
15-1111	Computer and Information Research Scientists
15-1121	Computer Systems Analysts
15-1122	Information Security Analysts
15-1131	Computer Programmers
15-1132	Software Developers, Applications
15-1133	Software Developers, Systems Software
15-1134	Web Developers
15-1141	Database Administrators
15-1142	Network and Computer Systems Administrators
15-1143	Computer Network Architects
15-1151	Computer User Support Specialists
15-1152	Computer Network Support Specialists
15-1199	Computer Occupations, All Other
17-2011	Aerospace Engineers
17-2021	Agricultural Engineers
17-2031	Biomedical Engineers
17-2041	Chemical Engineers
17-2051	Civil Engineers
17-2061	Computer Hardware Engineers
17-2071	Electrical Engineers
17-2072	Electronics Engineers, Except Computer
17-2081	Environmental Engineers
17-2111	Health and Safety Engineers, Except Mining Safety Engineers and Inspectors
17-2112	Industrial Engineers
17-2121	Marine Engineers and Naval Architects
17-2131	Materials Engineers
17-2141	Mechanical Engineers
17-2151	Mining and Geological Engineers, Including Mining Safety Engineers
17-2161	Nuclear Engineers
17-2171	Petroleum Engineers
17-2199	Engineers, All Other
17-3011	Architectural and Civil Drafters
17-3012	Electrical and Electronics Drafters
17-3013	Mechanical Drafters
17-3019	Drafters, All Other
17-3021	Aerospace Engineering and Operations Technicians
17-3022	Civil Engineering Technicians
17-3023	Electrical and Electronics Engineering Technicians
17-3024	Electro-Mechanical Technicians
17-3025	Environmental Engineering Technicians
17-3026	Industrial Engineering Technicians
17-3027	Mechanical Engineering Technicians
17-3029	Engineering Technicians, Except Drafters, All Other
17-3031	Surveying and Mapping Technicians
25-1031	Architecture Teachers, Postsecondary
25-1032	Engineering Teachers, Postsecondary
41-9031	Sales Engineers

account for the fact that an *occ1990dd* category might contain different types of occupations in terms of non-STEM, pure-STEM, and applied-STEM. All that remains then is to link all the occupation observations from NLSY79 and NLSY97 to *occ1990dd* so that we get to know each individual's occupation type.

For the 79 cohort, the pre-2000 observations of occupations are in *occ1970* codes while the post-2000 ones are in *occ2000* codes. Thus we build upon the crosswalks "occ1970_occ1990dd" and "occ2000_occ1990dd" from the replication package of Autor and Dorn (2013)[38] to create the crosswalks from *OCC1970* and *OCC2000* to *occ1990dd_stem*. We also build upon the crosswalk from *SOC2010* to *occ1990dd* at https://forum.ipums.org/t/occ1990-to-soc-crosswalk/2098 to create one from *SOC2010* to *occ1990dd_stem*. Hence we have *occ1990dd_stem* to map the NLSY79 reported occupations to the three occupation types.

For the 97 cohort, the occupation information is recorded in the *occ2002* codes. Building upon Flood (2019), we made a crosswalk between *occ2002* and *occ1990dd_stem*. Then by linking *occ1990dd_stem* and the three categories as we do to the 79 cohort, we have *occ1990dd_stem* to classify the NLSY97 occupation observations to the three occupation types.

F Top 10 occupations in each category

Tables 15-17 present the Top 10 most prevalent occupations by field, taking from the full-time working-for-wage population between the age of 22 and of 35 in the American Community Survey (2017-2019).

[38] https://www.ddorn.net/data.htm#Occupation%20Codes

Table 15: Top 10 most prevalent non-STEM occupations (occ1990d_stem)

Rank	Job Title
1	Managers and administrators
2	Sales supervisors and proprietors
3	Primary school teachers
4	Health and nursing aides
5	Customer service reps, invest., adjusters, excl. insur.
6	Accountants and auditors
7	Driver/sales workers and truck Drivers
8	Sales workers, other commodities
9	Sales workers
10	Cooks

Table 16: Top 10 most prevalent pure-STEM occupations (occ1990d_stem)

Rank	Job Title
1	Registered nurses
2	Health technologists and technicians, n.e.c.
3	Physicians
4	Licensed practical nurses
5	Economists, market and survey researchers
6	Managers of medicine and health occupations
7	Physical scientists, n.e.c.
8	Pharmacists
9	Radiologic technologists and technicians
10	Clinical laboratory technologies and technicians

Table 17: Top 10 most prevalent applied-STEM occupations (occ1990d_stem)

Rank	Job Title
1	Computer systems analysts and computer scientists
2	Engineers and other professionals, n.e.c.
3	Civil engineers
4	Computer managers and administrators
5	Mechanical engineers
6	Computer programmers
7	Engineering technicians, except drafters
8	Electrical engineers
9	Industrial engineers
10	Aerospace engineers