High school human capital portfolio and college outcomes[*]

Guy Tchuente[†]

University of Kent

March 2016

Abstract

This paper assesses the relationship between courses taken in high school and college major choice. It considers individuals as holding a portfolio of relative human capital rates that may either be similar to those in their major - specialized - or different from those in their major - diversified. Using High School and Beyond survey data, I find a U-shaped relationship between the diversification of high school courses portfolio, measured by the differences from the typical student in the major, and college performance. The underlying relation linking high school to college is assessed by estimating a structural model of high school human capital acquisition and college major choice. Policy experiments suggest that taking an additional quantitative course in high school increases the probability that a college student chooses a science, technology, engineering, or math major by four percentage points with little effect on college performance.

Keywords: Human capital, discrete choice, college major.

JEL classification codes: J24, I21.

[*] I thank the editor and the two anonymous referees for their helpful comments. I also thank the participants of 9th CIREQ Ph.D. students conference (Montreal, June 2013), CEA 2014 and SCSE 2014 for helpful comments. Comments from Andriana Bellou, Jorgen Hansen, Marc Henry, Joshua Lewis, Rachidi Kotchoni, Iain Fraser, Miguel Leon-Ledesma and Abigail Payne are gratefully acknowledged. I am indebted to Marine Carrasco and Baris Kaymak for their extensive comments and advice.

[†] School of Economics, Email: g.tchuente@kent.ac.uk

1 Introduction

This paper assesses the relationship between courses taken in high school and college major choice. In many countries, there has been an emphasis on encouraging science, technology, engineering and math (STEM) majors. These fields are of critical importance to economic competitiveness in an increasingly global and highly competitive economy. For example, in the U.S., the President's Council of Advisors on Science and Technology promotes the education of future STEM professionals through various grants and programs. The council has stated that over the next decade, a million additional STEM graduates will be needed. In the U.K., the Royal Academy of Engineering reported that the nation will need 100,000 new graduates with STEM majors annually until 2020.

Several studies have shown the existence of ability sorting with respect to college major. This sorting can be driven either by variations in the cost of successfully completing degree requirements or variations in expected returns to different majors by ability in different majors. Arcidiacono (2004) finds that predetermined factors, such as preferences and quantitative skills, play a larger role in major choice than economic returns. Based on these findings, this paper examines the role of high school education in developing quantitative skills and evaluates the potential effectiveness of high school curriculum changes that promote enrollment and success in STEM majors.

I use data from the U.S. High School and Beyond (HS&B) survey, which has detailed information on high school and college students. The first observation is that the types of courses taken in high school vary significantly for each college major. Mathematics and engineering majors take more quantitative courses in high school while business and literature majors have more high school humanities courses. Natural sciences and health majors take a mix of quantitative and humanities courses in high school.

However, there is a U-shaped relationship between the diversity of courses taken in high school and college performance: students who specialize in a particular subject as well as those who broadly diversify across subjects tend to have a higher college grade point average (GPA) in their corresponding major than those who slightly diversify. This result is the consequence of uncertainty about which majors students will pursue in college. Moreover, it suggests that the high school curriculum plays a crucial role in a student's choice of college major and their post-secondary performance.

Based on the link between high school and college, I propose and estimate a structural model of high school human capital acquisition and college major choice. By explicitly modeling the educational decision-making process, I both disentangle the heterogeneous effects of specialization and control for the self-selection inherent in educational outcomes.

Students in the model differ in their abilities in different subjects, as well as their preferences for these subjects. They are endowed with different initial abilities and have two decision periods; in the first period, they choose which high school courses to take, and in the second period, they choose their college major (or decide to not attend college). Students choose high school courses that maximize their expected discounted utility across college majors. Upon graduation from high school, in the second period, they choose their majors and observe their major-specific preferences.

Estimation results suggest that students who specialize in a particular area in high school tend to prefer quantitative majors in college, even after controlling for selection. Particular high school courses also play an important role in influencing a student's choice of college major. More quantitative courses in high school increase the likelihood of majoring in natural sciences, engineering, and math and physics, whereas more humanities courses mean a student is more likely to pursue a major in social sciences and humanities, or business and communications. These results suggest that an appropriate high school quantitative curriculum can increase enrollment in STEM majors.

I examine different counterfactuals to confirm this intuition. First, I examine what we would expect to happen if students were to take one more high school course on a particular subject. Second, I examine the expected outcome if all students faced the same high school curriculum for quantitative, humanities and life sciences courses, thus eliminating the possibility to specialize in a particular subject area in high school.[1] Both experiments substantially affect college major choice and performance. Taking an additional quantitative course in high school increases the probability of enrollment in STEM majors by four percentage points. In contrast, taking an additional humanities

[1] Note that all these simulations are not taking into account general equilibrium effects; they are designed to illustrate how much college major choice and performance is due to high school curriculum.

course in high school almost has no effect on enrollment in STEM majors. An additional life sciences course in high school also has a very small effect on a student's choice of college major. Imposing a single curriculum on all high school students also boosts enrollment in humanities majors. The large effect of standardized curriculum suggests that high school specialization plays a key role in influencing what majors students choose. While high school curriculum plays a non negligible role in the major choice, all experiments suggest a limited effect on college performance. These results contribute to the literature linking high school curriculum to future life achievements.

There is an extensive literature on college major choice.[2] Most of the theoretical frameworks in this literature imply that college major choice is influenced by expectations of future earnings, preferences, ability, and preparation (see Altonji, Blom, and Meghir (2012) for more details). Turner and Bowen (1999) document the sorting that occurs across majors by scholastic aptitude test (SAT) math and verbal scores. Arcidiacono (2004) finds that differences in monetary returns explain little of the ability sorting across majors, and concludes that virtually all ability sorting is a result of preferences for particular majors in college and the workplace, with the former being larger than the latter. I extend the model in Arcidiacono (2004) to add college preparation in high school, where students can choose which subjects to study.

A related strand of the literature studies the causal effect of high school curriculums on labor-market outcomes (see Altonji (1995), Levine and Zimmerman (1995), and Rose and Betts (2004)). More recently, Joensen and Nielsen (2009) and Goodman (2009) use quasi-experiments to estimate the effect of math coursework on earnings. These studies all aim to determine whether skills accumulated in high school matter for college performance and labor-market outcomes.

Unlike these papers, I investigate the effect of the *composition* of skills acquired in high school on college performance. This study, therefore, contributes to existing studies by introducing multi-dimensional endowments of skills and by studying the tension between specialization and diversity. In this sense, this paper is closer to Malamud (2010), Smith (2010), and Malamud (2012), who examine the trade-off between specialized and diversified human capital portfolios in college and their effect on labor-market

[2] See Montmarquette, Cannings, and Mahseredjian (2002), Zafar (2009), Stinebrickner and Stinebrickner (2011), Arcidiacono (2005), and Arcidiacono, Aucejo, and Hotz (2013).

outcomes. Silos and Smith (2014) study how diversification and specialization strategies in college influence income dynamics. They find that diversification generates higher incomes for individuals who switch occupations, whereas specialization benefits those who stick with one type of job. This paper considers the effect of diversification earlier in the educational process, by investigating how specialization in high school affects college major choice and performance.

The paper proceeds as follows. Section 2 provides a brief overview of the U.S. high school system and explains why the U.S. system offers a unique opportunity to investigate the effect of high school course choice on college outcomes. Section 3 describes the data and the sample used for empirical analysis and provides a reduced-form analysis of the relationship between diversification in high school and college performance. A dynamic model of college and major choice as well as the econometric techniques used to estimate the model are described in Section 4. Section 5 provides the empirical and simulation results. Section 6 concludes.

2 Background: High school course choice in the U.S.

The U.S. high school education system provides a particularly appropriate setting to examine almost all aspects of the effect of high school preparation on college. In the U.S., high school students have significant control over their education and are allowed to choose their core classes. This allows us to understand not only how success in each high school subject affects college outcomes, but also how the choice of courses affects college outcomes. The degree of control given to students varies from state to state[3] and from school to school. This leads to a substantial variation in students' academic experiences, both between schools in the same state and across states (Lee, Croninger, and Smith (1997), Allensworth, Nomi, Montgomery, and Lee (2009)). Despite the wide variations in curriculums, many schools require that courses in the "core" areas of English, science, social studies, and math be taken every year. However, some schools

[3] See for example Goodman (2009), Figure 2, for differences in math requirements by state. Graduation requirements also differ by state (see Bruce Daniel (2007)).

set the required number of credits and allow students to choose when the courses will be taken.

The menu of courses available to students depends on a particular school's financial and staffing situation. Thus, the available choices are a direct function of the preferences of teachers, which are usually idiosyncratic. Furthermore, inducements for students to take a particular set of classes may differ between schools, as certain teachers are hired or school administrators decide to place greater emphasis on these subject areas. Thus, there is a substantial element of exogenous variation in course choice across schools due to the idiosyncrasies of teachers, school administrators and states. I take advantage of these exogenous variations to identify how the composition of courses taken (specialized or diversified) in high school affects college performance.

3 Data and descriptive statistics

3.1 Data

To investigate the empirical relationship between courses completed in high school and post-secondary education outcomes, I use data from the 1980 HS&B survey. This panel data set tracks students from high school to post-secondary, and contains detailed information on courses taken in high school as well as post-secondary outcomes. The HS&B survey was conducted by the National Center for Education Statistics. A nationally representative sample of high school sophomores, from 1980, were interviewed once every two years from 1980 to 1986, and again in 1992. These interviews recorded detailed information about the high school courses students took and their grades. This high-quality data provides my measures of human capital and high school preparation.[4]

My data on students' college performance comes from the Post-Secondary Education Data System (PEDS), which contains institutional transcripts from all post-secondary institutions attended for a sub-sample of students present in the HS&B survey. My estimations are performed using data from 1980, 1982, 1984 and 1986 surveys.

The HS&B survey contains 14,825 students. A reduced form and a structural model

[4] High schools usually run from either grade 9 or 10 to grade 12. I restrict my analysis to grades 10 to 12, since this data is available for all students in the sample.

are estimated. For the reduced form model the sample contains individuals who are enrolled in college. These individual should also have information for any variable used in the reduced form model. I started with a sub-sample of 5,533 students having transcripts encoded for both high school and college. Dropping those who do not have SAT and have not participated to the first follow-up and base year test reduces the sample to 1,921 individuals. Eliminating individuals with missing observations for control variables, and dropping outliers reduce the sample to 1,083. The structural model used information from base year tests to college. Starting with a sample of 1,921 individuals cleaning the data yields a final sample of 1,222 students for estimation. This sample contains students who have never attended college and high school dropouts. For both models, it was not possible to take into account college dropouts and major switches, because of data limitation.

There is a huge amount of data dropout. Table 1 shows the average characteristics for restricted samples and for the dropped sample (with the initial sample being of 1,921 individuals). It also reports the results of a t-test for difference in mean in the two samples. Apart from college GPA, all the other tests assume equal variance. There is not a large difference between the dropped sample and the restricted one for all but one variable at the 5% level. The variable with statistically different mean is the proportion of African Americans. However, the under-representation of African Americans is not likely to affect our results since the race is not central to the research questions. Sample selection may not, therefore, be a concern.

3.2 Empirical structure and descriptive statistics

This subsection provides empirical findings that show a possible relationship between high school preparation and college outcomes.

I group subjects studied in high school into different categories (which could be interpreted as types of human capital). Each student has a human capital portfolio based solely on the courses that the student takes in high school. The portfolio contains seven categories of study.[5] High school courses are grouped into the following categories:

[5] The Appendix provides a step-by-step description of the construction of human capital portfolios, as well as college major aggregation.

(i) quantitative (mathematics and physics (Quant.)), (ii) reading and writing (R. & W.), (iii) social sciences and humanities (Humanities), (iv) life sciences (Life sci.), (v) business and communications (Bus. & Com.) , (vi) Arts, and (vii) Other.[6]

Let us study how the composition of high school courses is related to major choice. Given courses taken in each subject (or type of human capital) $k = 1, ...K$, the weights in the human capital portfolio of an individual i are:

$$\omega_{i,k} = \frac{course_{i,k}}{\sum_{j=1}^{K} course_{i,j}},$$

where $K = 7$ and $course_{i,k}$ is the number of courses taken in subject k.[7] Table 2 displays these portfolio weights by major across the population. For each major, the table contains the average, across individuals, of the weights in each of the seven subject areas.

The proportion of quantitative subjects in high school varies from 0.165 for education majors in college to 0.227 for engineering majors. It is not surprising that college students majoring in humanities took a greater proportion of reading and writing classes in high school (0.259) than other college majors. Likewise, business and communications majors took a greater proportion of business and communications courses in high school (0.095) than did other college students. Although the difference in means in some subjects appears small, the last two rows of Table 2 shows that these differences are statistically significant. In summary, students enroll in a college major related to subjects they concentrated on in high school.

Let us now investigate how the composition of high school courses will affect college performance, through a reduced form regression analysis. This helps us investigate the data beyond raw mean difference. To this end, a measure of diversification is defined. With the chosen diversification measure, identification of the courses composition effect is discussed.

Each student i has a vector of human capital weights, $\omega_{i,k}$, which measures the weight of skill type k in his overall portfolio. This student portfolio can be viewed as

[6] My results are robust to the structure of these categories; I considered other potential categories and obtained similar results.

[7] I focus on the distribution of courses by examining the share of total courses in a given subject, rather than the number of courses taken.

specialized or diversified. A skewed or balanced portfolio does not necessarily imply specialization or diversification of human capital investments. Some students may choose a uniform allocation of courses across fields to self-insure against shocks or because a particular major explicitly rewards balanced skills.[8] To evaluate the level of diversification, I follow Silos and Smith (2015); I assess how well tailored an individual's acquired skill set is for a particular college field by comparing human capital investments to a benchmark for that field.[9]

Let us define the measure of diversification as

$$\rho_{i,m} = \sqrt{\sum_{k=1}^{K} (\omega_{i,k} - \bar{\omega}_{k,m})^2}$$

where $\bar{\omega}_{k,m}$ denotes the average portfolio for major m observed in Table 2. I assume that a portfolio is chosen for a given major if that portfolio is "close" to the average portfolio of that major. Self-insurance against shocks is simply the distance between the portfolio weights and the typical portfolio of the college major. Thus, students can specialize in major-related subjects, or hedge with respect to a major by diversifying their portfolios. Small values of ρ thus mean a student has specialized, and large values indicate a student has diversified.

I estimate the following reduced-form equation:

$$G_i = \alpha_0 + \alpha_1 \rho_i + \alpha_2 \rho_i^2 + \alpha_3 X_i + \alpha_m + \alpha_h + \varepsilon_i$$

where α_m and α_h are fixed effects for major and high school, respectively. G_i is the college GPA of individual i in major m from high school h. X represents control variables such as SAT scores, socioeconomic status (SES) and gender.

Identifying the effect of diversification is one of the main challenges of this paper. The identification is based on two assumptions. The first one is that there is a certain amount of exogeneity in high school supply of courses. The second is that students are identical in term of unobserved ability within a major.

[8] The student can for example not be accepted in his first-choice major, which can be viewed as a shock.

[9] This measure is related to the diversification index in the trade literature from Krugman (1992), which uses an absolute distance instead of a square root. See also Palan (2010) for a review of the diversification index in trade. I also consider other diversification measures, such as the Gini index. The results obtained are qualitatively the same.

A key identifying assumption of the diversification effect is that the variation in courses supplied across schools is exogenous. However, course choice can be shaped by school requirements and tracking policies. I use the fact that some HS&B survey students came from the same high schools to control for high school fixed effects. This eliminates concerns about differences in high school curriculums driving the results. Identification of the diversification effect is, therefore, coming from within school variation in course selection. The intuition is that each student will react differently to school requirements and this exogenous variation will help identify the diversification effect.

On the other hand, within school course selection can be affected by individual ability. An interesting feature of the diversification measure is that, under the assumption that students are identical in term of unobserved ability within, it can help reduce the major specific ability bias.[10] Under these assumptions, the variations in the diversification measure within a school is driven by exogenous factors like preference for a particular teacher or individual specific needs.

To evaluate the validity of this assumption, controls for ability are used, like SAT-Math scores, SAT-Verbal scores and the number of courses taken in each high school subject and the base year standardized test scores. The effect of ability variables on the relationship between GPA and diversification will inform us about unobserved ability bias.

Table 3 shows that the relationship between GPA and the measure of diversification ρ is quadratic, large and significant. The results are robust to controlling for gender, race, and SES. It is also robust to regional disparities by including a three regional dummies. The major-specific effect is controlled for by including a dummy variable for each major. It is worth noting that the inclusion of several ability control variables does not change the effect of diversification on college performance. In Table 3, it can be noted that the coefficients on ρ and ρ^2 are not significantly affected by the inclusion of ability measures like SATs and base year tests. This suggests that the effect of diversification did not suffer too much from ability bias.

The U-shape is also robust to the level of course taken in different type of subject

[10] Readers interested in the identification of the diversification effect can refer to the working paper version of this paper.

in high school. This means that individuals with the same level of high school math courses, but that differ in other courses will have different performances. The inclusion of college major dummies has the largest effect on the diversification coefficients (see Table 3 columns 3 and 4). However, this effect is not very large. The fact that most of the controls do not have a very strong effect on diversification is a sign that even if there is some selection on unobservables it is very unlikely that this effect is large enough to cancel the U-shaped relationship between college GPA and diversification (see Altonji, Elder, and Taber (2005) for discussion on selection on observed and unobserved variables). To formally test for the presence of a U-shape, I use the procedure proposed by Lind and Mehlum (2010). The results show that there is indeed a U-shaped relationship.[11]

These empirical findings show the importance of high school preparation in determining students' college majors and performance. However, mean statistics and parameter estimates may be subject to a selection bias due to the presence of unobserved characteristics. I, therefore, propose and estimate a structural model of high school human capital acquisition and college major choice. This enables me to not only to control for potential selection bias on unobserved variables but also to conduct counterfactual experiments to study the potential effects of various curriculum policies in high schools.

4 Structural model of high school human capital choice

This section proposes and estimates a model of high school human capital acquisition and college major choice. In the model, individuals differ in both their innate ability to learn and in their preferences for different college majors. I assume that students know their ability to acquire imperfectly substitutable skills in high school through courses. They choose their high school courses to maximize their expected utility across college majors. Upon high school graduation, students choose to pursue a particular college major or do not enroll in college.

[11] The results of the test for U-shaped relationship are in the footnote of Table 3

Individuals live a finite number of period and have a discount factor $\beta \in (0,1)$. They choose their human capital investments, i.e. a set of high school courses, in the initial period ($t = 0$) to optimize expected discounted utility.

There are several types of skills that are differently useful for all majors. In other words, the high school skills are useful in college, but their importance differs from one major to another. I denote an individual's high school set of course by $s = (s_Q, s_H, s_{NS})$, where s_Q is the number of quantitative courses, s_H is the number of humanities courses and s_{NS} is the number of natural sciences courses.[12] Individuals can choose their portfolio composition by selecting more high school courses in a particular area.[13]

Before choosing their high school courses, individuals draw abilities $\tau = (\tau_Q, \tau_H, \tau_{NS})$ from distribution $H(\tau)$, where τ_{NS} represents the ability to accumulate natural sciences human capital. Individuals know how useful each type of human capital is for each college major. However, they are unsure about an idiosyncratic component of their college preferences. A student's initial or innate abilities and his preferred college major provide an incentive for the student to specialize by acquiring skills that reflect his personal circumstances. In contrast, the risk of low utility draws in each college major provides an incentive to acquire a more widely applicable portfolio of human capital.[14]

Once an individual has acquired a skill set of high school courses (s), they decide whether or not to enter college in the second period ($t = 1$). Those who choose to enter college also select a major. Although they have a general idea, before they invest in their portfolio of skills, of how well they are likely to fit into a given major, it is only after the completion of high school and when they enter to college that their true fit in a major becomes known; actual experience in a major reveals an individual's true preference for that major.[15]

[12] In the estimation I used seven types of courses or human capital like in the reduced form part. However to ease the presentation I have used only three types of courses. Quantitative human capital is measured by the number of high school courses taken in math and physics. Humanities human capital is measured by high school courses taken in reading and writing, humanities or business and communication. Natural sciences human capital is the number of high school biology and chemistry courses.

[13] Students could also change their portfolio by doing more homework or tutoring in a particular area, but this behavior is not observed in the data.

[14] The student may discover that he is not very good at a major, or that he does not like a major as much as he thought he would.

[15] For simplicity, I assume that students make a one-time decision about their college major; I ignore the

The timing of the model is as follows:

- **In period 1**: Individuals draw abilities τ from distribution $H(\tau)$.
- **In period 2**: They choose the number of courses to take in each high school subject.
- **In period 3**: Individuals choose a major. They receive new information about their abilities and preferences in that major and accumulate human capital.

4.1 College major choice

In this subsection, I specify the model of college major choice given high school outcome. Once individual decide their college major, there is no decision left. Base on results of the reduced-form in Section 3, the college performance measured by GPA (G) is a function of individual observed abilities, the level of diversification as well as demographic characteristics, such as gender and SES.[16] Specifically, performance in college takes the following form:

$$G = \eta_0 + \eta_1 \rho + \eta_2 \rho^2 + \eta_3 s + \eta_4 X_G + \eta_m + \varepsilon_1$$

The model also contains a major-specific fixed effect, η_m, as well idiosyncratic shocks (the ε_1's), which are drawn from distribution $\mathcal{N}(0; \sigma_G^2)$. I assume that the human capital gains in high school by attending courses (s_k) affect college performance directly and through the level of diversification (ρ) of the student. I also use SAT math and verbal scores as a measure of observable ability.

The utility of choosing a college major m is given by

$$u_m^c = \vartheta_{0c} s + \vartheta_{1c} X_{cm} - c_m(s, G) + v_m + \varepsilon_m$$

where ε_m is a generalized extreme value (GEV) distribution. The fixed intercept (v_m) represents the combined effect of all omitted major-specific covariates that cause some

possibility that students may do post-graduate work or drop out of college.

[16] Due to data limitations, I do not include a wage equation in the model. However, given that GPA has a positive effect on future earnings (see Arcidiacono (2004)), it can be used as a proxy for future wages. Moreover, several recent studies suggest that monetary factors are not the main driver of college major choice (see Beffy, Fougère, and Maurel (2012), Carneiro, Hansen, and Heckman (2003), and Delavande and Zafar (2014))

students to be more predisposed to a particular major. The variables X_{cm} are individual characteristics that could influence college major preference like gender, performance in college, SATs, state wage in manufacturing and courses taken in high school. This college major utility includes the cost of effort needed for a particular combination of high school courses (s) and major to achieve a performance level (G). It reflects the fact that high school preparation can make some major more enjoyable and also affect the effort needed to succeed.

The utility from being in high school is given by

$$u^h = -c^h(\tau, s) + \varepsilon$$

where ε is normally distributed. Acquiring high school human capital has a cost $c^h(\tau, s)$. This high school utility implies that the cost of attending a course in a particular area depends on individual initial ability.

Let us now specify the costs functions using marginal costs. I assume that the marginal high school cost of acquiring a specific high school human capital depend of your initial ability and the quantity of courses already taken in the area. The intuition is that, more advanced courses require more effort and initial abilities reduce the marginal cost. The marginal cost of acquiring high school human capital k is:

$$Mc_k^h(\tau, s) \quad = \quad \vartheta_{3hk} + \vartheta_{4hk} s_k + \vartheta_{5hk}\tau$$

where Mc_k^h is the marginal cost of acquiring skill k in high school. $\vartheta_{.hk}$ are parameter of the cost function contribution of producing human capital (k).

The college cost functions is $c_m(s, G)$. The marginal benefit of acquiring a specific high school human capital (k) for a student entering major m is:

$$Mc_{mk}(G) = \vartheta_{4mk} + \vartheta_{5mk} G$$

where Mc_{mk} is the marginal benefit of acquiring high school skill k for major m. $\vartheta_{.mk}$ are parameters observed with error; that is why I control for major-specific fixed effects, v_m. Having appropriate high school skills make the subject more enjoyable for the student and may also reduce the effort needed to perform well.

Introducing cost of effort in college may imply that even if an individual were allowed to enroll in any major, the individual may not choose to attend the highest-paying major because of the effort required or their lack of preparation.

Individuals also have the option not to attend college. In this scenario, the individual receives a utility u_o, where the o subscript indicates that the individual chooses an outside option.

College students choose the major with the highest u^c_m, i.e. the major that yields the highest utility. I assume that ε_m follows a GEV distribution. Special cases of the GEV distribution require the use of a multinomial logit or nested logit model. I use a nested logit model; this GEV distribution as set out in McFadden (1978), allows for errors to be correlated across multiple nests while still being consistent with random utility maximization.[17]

I assume that majors are grouped into four nests:

- **Nest 1**: Quantitative majors (math, physics and engineering)
- **Nest 2**: Business & communications, humanities, education and military majors
- **Nest 3**: Health and natural sciences majors
- **Nest 4**: No college

Let $u^{c'}_m$ be the net present value of the indirect utility for completing major m.

$$F\left(e^{u^{c'}}\right) = \sum_m \left(\sum_N exp\left(\frac{u^{c'}_{mn}}{\eta}\right)\right)^\eta + exp\left(u_o\right)$$

The error terms are known to the individuals, but they are not observed by the econometrician. Therefore, from the econometrician's perspective, the probability of choosing

[17] The framework from McFadden (1978) is as follows. Let $r = 1, ..., R$ be an index of all possible choices. Define a function $G(y_1, ..., y_R)$ on y_r for all r. If G is non negative, homogeneous of degree 1, approaches $+\infty$ as one of its arguments approaches $+\infty$, it has non-negative n^{th} cross-partial derivatives for odd values of n, and non-positive cross-partial derivatives for even values of n, then McFadden (1978) shows that

$$F(\epsilon_1, ..., \epsilon_R) = exp\{-G(e^{-\epsilon_1}, ..., e^{-\epsilon_R})\}$$

is the cumulative distribution function for a multivariate extreme value distribution. Furthermore, the probability of choosing the r^{th} alternative conditional on the observed characteristics of the individual is given by

$$P(r) = \frac{y_r G_r(y_1, ..., y_R)}{G(y_1, ..., y_R)}$$

where G_r is the partial derivative of G with respect to the r^{th} argument. This is the same as in Arcidiacono (2005).

a major m is given by

$$Pr(m) = \frac{exp\left(\frac{u^{c'}_{mn}}{\eta}\right)\left(\sum_N exp\left(\frac{u^{c'}_{mn}}{\eta}\right)\right)^{\eta-1}}{F\left(e^{u^{c'}}\right)}.$$

Before choosing a major, individuals first choose their high school human capital. The net utility from the outside option, which is not going to college, is normalized to zero.

4.2 Choice of high school human capital

After deciding on a college major, there are no decisions left. Let u^c_1 indicate an individual's optimal choice of college major. Individuals need to choose how much of the different types of human capital to accumulate in high school. They choose the level of high school human capital (s) that yields the highest discounted utility $V_0(s,\tau)$:

$$V_0(s,\tau) = u^h + \beta E_0(u^c_1|\tau)$$

For each type of human capital, s^*_k is the optimal level of human capital in the area (k).[18] Let us consider

$$\tilde{s}^*_k = s^*_k + \varepsilon_k = \theta_{0k} + \theta_{1k}\tau + \theta_{2k}G + \theta_{3k}m + \varepsilon_k$$

where ε_k is the normally distributed forecast error. It is independent of initial abilities, college performance and major. The observed chosen level of high school courses (s_k) is:

$$s_k = \begin{cases} \tilde{s}^*_k & if \quad \tilde{s}^*_k > 0 \\ 0 & if \quad s^*_k \leq 0 \end{cases}$$

The optimal high school human capital choice implies that the number of courses chosen in high school depends on the initial abilities, the expected major choice and the expected level of GPA in college. I estimate the coefficients of the model with a Tobit model. The parameters of this equation are functions of deep structural parameters. But we will not try to recover them because they are not going to be useful for our counterfactual analysis.

[18] The solution solves the Euler equation $Mc_k = \beta E_0(u^c_{1k}|\tau)$. If I apply the envelope theorem to u^c_1, I get $E_0(u^c_{1m}|\tau) = \beta\vartheta_{0ck} - \beta E_0(Mc_{mk}(s,G)))$ and $\vartheta_{4hk}s_k + \vartheta_{5hk}\tau + \vartheta_{3hk} = \beta\vartheta_{0ck} - \beta Mc_{mk}(G)$. Thus, $s^*_k = \theta_{0\hat{m}k} + \theta_{1\hat{m}k}\tau + \theta_{2\hat{m}k}G$

4.3 Identification and estimation strategy

In this subsection, I discuss how several key parameters of the model are identified and how they are going to be estimated. Two versions of the model are estimated. In the first case, I assume that there is no selection on unobservables for high school courses and major choice as well as no unobserved omitted variable affecting college performance. This assumption is relaxed in the second set of estimations where the selection on unobservables is allowed.

4.3.1 Identification and estimation without unobservables

In the version of the model without unobservables, all individual characteristics are assumed to be exogenous, including gender, SES and 10^{th}-grade standardized test score. Selection into college, irrespective of the major, and state difference in major preference is controlled for by the hourly state wage in manufacturing. One of the main advantages of HS&B data is that individuals in the sample have base-year test scores in different subjects. These scores are in math, science, civics, reading and writing and are my main exogenous variables useful to estimate high school human capital acquisition. I assume that there is no correlation across the various stages of the model. Therefore, selection into majors is controlled for by these exogenous characteristics.

The log-likelihood function is the sum of three pieces:

- $L_1(\eta)$ – the log-likelihood contribution of grade point averages,
- $L_2(\vartheta_c, \eta)$ – the log-likelihood contribution of major decisions, and
- $L_3(\vartheta_h, \vartheta_c, \eta)$ – the log-likelihood contribution of high school human capital decisions.

The total log-likelihood function is then $L = L_1 + L_2 + L_3$.

Consistent estimates of η can be found by maximizing L_1 separately conditional on knowing your major and high school courses choices. Then, the η is replaced by consistent estimates in L_2. A consistent estimate of ϑ_c can then be obtained by maximizing L_2. I estimate ϑ_h using L_3 and all other estimates. This procedure provides us with consistent estimates of the model parameters.

4.3.2 Identification and estimation with unobservables

It is unreasonable to assume that preference parameters are uncorrelated over time (that is, if one has a strong preference for high school math initially, he is just as likely as someone who has a weak preference for high school math to choose any major in college). This is likely not the case. Furthermore, it is unreasonable to assume that there is no unobserved (to the econometrician) ability that is known to the individual.

To account for unobservable characteristics affecting students' choice of majors, I use a mixture distribution that allows errors to be correlated across the various stages it also provides a way of controlling selection base on unobserved characteristics. More precisely, I assume that there are two types of individuals. Types remain the same throughout all stages and individuals know their type. Preferences for particular fields in college and high school courses may vary across types. For computational simplicity, in all equations estimated, I assume that the parameters do not vary across types except for the constant term. Some variables can be used to identify types: initial ability (here measured by base-year standardized test scores), the level of high school human capital and college major choice.

The log-likelihood function for a data set with N observations is then given by

$$L(\eta, \vartheta) = \sum_{i=1}^{N} ln(\sum_{r=1}^{R} \pi_r \mathcal{L}_{ir1} \mathcal{L}_{ir2} \mathcal{L}_{ir3})$$

where π_r is the proportion of type r in the data and $\mathcal{L}_{ir.}$ refers to the likelihood (as opposed to the log likelihood L) and R is the number of types.[19]

The log-likelihood function is no longer additively separable. I use the expectation-maximization (EM) algorithm to solve the problem. The EM algorithm has two steps:

- **First**, calculate the expected log-likelihood function given the conditional probabilities at the current parameter estimates, and
- **Second**, maximize the expected likelihood function holding the conditional probabilities fixed.

[19] The proportion of each type is estimated using the expectation-maximization (EM) algorithm. The Type 1 individuals make up 60% of the population, while the Type 2s make up 40%. See Arcidiacono (2004, 2005) for other examples of using mixture distributions to control for unobserved heterogeneity in college major choice models.

These steps are repeated until there is convergence.

The expected log-likelihood function is:

$$L(\eta, \vartheta) = \sum_{i=1}^{N} \sum_{r=1}^{R} P_i(r|X_i, \alpha, \eta, \vartheta)[L_{ir1}(\eta) + L_{ir2}(\eta, \vartheta_c) + L_{ir3}(\eta, \vartheta_{c,h})]$$

with $P_i(r|X_i, \eta, \vartheta) = \dfrac{\pi_r \mathcal{L}_{ir1} \mathcal{L}_{ir2} \mathcal{L}_{ir3}}{\sum_{r=1}^{R} \pi_r \mathcal{L}_{ir1} \mathcal{L}_{ir2} \mathcal{L}_{ir3}}$

Using the EM algorithm helps to recover the additivity of the log-likelihood function. Parameters can also be estimated at each step, as in the case without unobservable heterogeneity. Note that all pieces of the likelihood function are still linked through the conditional probabilities, where the conditional probabilities are updated at each iteration of the EM algorithm. Arcidiacono and Jones (2003) show that it is possible to estimate parameters sequentially during each maximization step. Using this sequential estimator generates large computational savings with little loss of efficiency.

5 Structural model estimation results

This section presents and discusses the results from estimating the parameters of the performance equations, the structural parameters of the utility function and high school course choice equations. Results of the model with unobserved heterogeneity are presented in the estimation of each equation separately.

5.1 College performance regressions

Estimates of the performance equation for the college period are given in Table 4. The first column displays the coefficient estimates without unobserved heterogeneity, while the second presents estimates with unobserved heterogeneity approximated by two types of students.

There is a U-shaped relationship between college performance and diversification in high school. The size of the coefficients is the same with or without unobserved heterogeneity. Figure 1 represents college GPA as a function of diversification measure. The visual representation also suggests the presence of a U-shaped relationship between GPA and the measure of diversification. This means people with broader minds and those with very focused minds will both do well in college.[20]

The U-shaped relationship suggests a tension between specialization and diversification. This tension is driven by two opposing forces implied by the diversification strategy. On the one hand, diversification reduces human capital in the targeted college major, but on the other hand, it increases knowledge in other subjects. When the diversification starts, the negative effect is stronger. As the level of diversification increases, more knowledge in other subjects is accumulated. At a turning point, other skills acquired compensate the losses through complementarity, and diversification's positives outweigh its negatives.

The tension between specialization and diversification is not new in economics. Usually, in modern labor markets, workers specialize in specific occupations. Likewise, before entering college, individuals may acquire particular skills in high school. Every field of study requires a specific set of skills. Conversely, many skills are useful, to different degrees, in a wide variety of fields. Psychology, law and biology students all require some reading, writing and arithmetic ability, albeit in different amounts. Moreover, some fields appear to more heavily emphasize a small subset of particular skills, whereas other majors more or less weigh skills evenly.

In high school, individuals are uncertain about their future college major. As a result, a high school graduate may study science courses and end up majoring in an unrelated field. Faced with uncertainty, a high school student may want to balance their efforts in case their intended major does not pan out. However, if students specialize in a particular skill, they may be more productive in a related field – this is why we first observe a positive effect of specialization. But if they diversify, they will acquire skills that have some use, even if they are rarely used. As such, there is a certain point at which diversification has a positive effect on performance. This explanation is based on substitutability of skill acquired in different high school subjects. There is another explanation of the U-shaped relationship based on learning.

Indeed, while in high school, individuals could have a guess of their expected major in high school and some make appropriate investments while others make a slightly different one. In college, they receive new information. In order to switch against their early investments and course choices, they have to have new information suggesting

[20] I thank Paul Oyer for suggesting this interpretation.

that the destination major is a better fit. This could explain why individuals whose investments cut against their eventual decision do so well. While those who do not receive new information and were slightly different do not do as well as others.[21]

Other variables coefficient are qualitatively similar to those obtain in the literature. For example, females earn higher grades than their males counterparts. All of the ability coefficients are positive, with smaller coefficients for SAT-Verbal scores. Without unobserved heterogeneity, ability in math is particularly useful. Once the mixture distribution is added, the differences in ability coefficients dissipate. The results with unobserved heterogeneity show that type 1s receive substantially higher grades.

5.2 Estimate of the utility function parameters

I use the estimates of performance to obtain the second-stage maximum likelihood estimates of the utility function parameters. Table 5 and Table 6 display the maximum likelihood estimates for the parameters of the utility function.

The first three sections of Table 5 display the preferences for the three types of high school courses, depending on a college student's major. More quantitative courses are attractive for college majors in natural sciences, engineering, and math and physics, while more humanities courses are preferred for social science and humanities majors, as well as business and communications majors.[22]

Females are more likely to enroll in education or health majors, and less likely to enroll in quantitative majors.[23] There is a sizeable literature on college major choice and the gender gap,which has documented differences in males' and females' college major choices that are in line with my findings.[24] However, the investigation of the effect of high school choices on the college gender gap is beyond the scope of this paper.

Types 1s are more likely to enroll in science majors in the model with the mixture distribution. Ability measures (SAT-Math and SAT-Verbal scores), GPA, and $GPA \times HScourses$ interact with major, along with major-specific constants that were included.

[21] I thank the editor for suggesting this alternative explanation of the U-shaped relationship. Developing a model integrating these intuitions is beyond the scope of this paper.

[22] Controlling for unobserved heterogeneity does not change these results.

[23] Taking unobserved heterogeneity into account does not change this result.

[24] See Zafar (2009) for more information.

Consistent with Arcidiacono (2004), I also find that students' comparative advantages in their abilities for different majors play a very important role in the choice of a major.

The nesting parameters, for the models with and without unobserved heterogeneity, are both relatively small for all models. The estimates that are less than one suggest that preferences for different majors are correlated. Indeed, these nesting parameters measure the cross-major component of the variance. In particular, had these coefficients been estimated to be one, then a multinomial logit would have resulted.

5.3 Course choice equations regressions

Estimates of the course equations Tobit model are presented in Tables 7 to 12. As with performance results, adding controls for unobserved heterogeneity does not significantly affect other parameter estimates. Those who have high math and science scores from the grade 10 standardized test tend to accumulate more skills in quantitative and life sciences subjects. Those with high scores in civics and writing are more likely to accumulate humanities skills. Type 1s tend to take more life sciences than quantitative courses or humanities courses in high school.

5.4 Model fit

In order to see how the model matches some key features and trends of the data, Table 13 compares actual data with the predictions of the model. I show two sets of parameter estimates from the model: one with unobserved heterogeneity, and the other without.

For each of the three groups of high school courses (quantitative, humanities and life sciences), I show the average number of these courses that different college majors took while in high school. The actual number of quantitative courses chosen in high school is very close to what is predicted by the model. The models with and without unobserved heterogeneity predict the trends in the data extremely well. The predictions with the mixture model are better than those without.

5.5 Simulations

Since the model matches the data reasonably well, I can use the model to simulate how decisions about majors and college performance would vary in different environments. The purpose of the simulations is to compare policies that may increase enrollment in STEM majors.

The first policy I examine is an increase in high school quantitative course requirements. The second experiment is an increase in high school humanities course requirements while the third simulation increases high school life sciences course requirements. These three simulations are designed to evaluate the impact of a change in high school curriculum on college outcomes. However, they are not helpful in evaluating the effect of diversification with respect to performance or to major choice; because they leave the level of diversification unchanged.

The last simulation assumes that there is no course choice in high school. This means that students are required to take same courses. I consider the situation in which they are required to take the average number of courses in the sample for all subjects. The aim of this counterfactual experiment is to measure the effect of forced specialization or high school standardization on college outcomes. The level of diversification in this simulation is zero. The impact of the single curriculum experiment depends on which curriculum is imposed. Moreover, such a standardization could have general equilibrium implications. Indeed, standardization may lead to the production of students that are more suitable for specific majors. Implying excess demand for these majors and shortage in others majors. This simulation does not take general equilibrium nature of the problem into account. It, therefore, provides limited evidence and should be interpreted with caution. However, it can be viewed as the partial effect of standardization in one school.[25]

Increasing the enrollment in STEM majors is of considerable interest to many countries, given that the economy is increasingly driven by complex knowledge and advanced cognitive skills. Thus, STEM workers are a key component to ensuring competitiveness in a global economy. The shortage of STEM majors occurs despite STEM majors

[25] Taking into account the general equilibrium nature of the question and providing alternative standardization are left for future research.

earning substantially more than other college graduates, with the potential exception of business graduates (see Arcidiacono (2004), Pavan and Kinsler (2012), and Arcidiacono, Aucejo, and Hotz (2013)).

The first, second and third simulations assume that students each take one more quantitative course, one more humanities course and one more life sciences course, respectively, in high school. These simulations show the extent to which the choice to pursue a STEM major is a result of high school course choice. The last simulation eliminates specialization in high school. The results of the simulation show how much specialization in high school affects enrollment in STEM majors. Note that these simulations do not account for general equilibrium effects; the simulations illustrate how much of the current major choice is due to high school courses choice or specialization.

Table 18 shows that high school quantitative courses affect the decision of pursuing STEM majors. When students take one more high school quantitative course, the share of people in STEM and natural sciences majors increases (see Simulation 1). One more high school quantitative course increases enrollment in STEM majors by four percentage points.[26] The adoption of such a policy for one decade, with an initial number of STEM graduates of 300,000, will lead to an additional 161,836 STEM graduates.

An increase in one high school humanities course does not decrease enrollment in STEM majors. One more life sciences course in high school increases enrollment in natural sciences majors by very small percentage points and reduces enrollment in other STEM majors. Forcing every student to take the same courses (see Simulation 4) also boosts enrollment in math, physics and engineering majors. The share of students choosing math, physics and engineering majors moves up by 8 percentage points. However, I observed the same amount of reduction is natural science majors. This suggests that high school specialization plays a key role in major choice. The adoption of a standardized curriculum for one decade, with an initial number of math physics and engineering majors graduates of 200,000, will lead to an additional 266,327 math physics and engineering majors graduates.

These results suggest that increasing high school quantitative course requirements would improve enrollment in STEM majors. Imposing a uniform curriculum in high school can also lead to a major increase in some STEM enrollment, however, this

[26] These results are similar to those obtain by Ning (2014)

depends on the curriculum imposed. Another aspect of college outcomes investigated by simulation is college performance.

The simulations one to four show little impact of the change in high school curriculum on college performance. In all majors, the change in performance is less than 3.35% for all four experiments. It is interesting to note that when the model without unobserved heterogeneity is used, one more high school quantitative course slightly decreases performance in all but engineering major. In the model with unobserved heterogeneity, the college performance slightly increases in all but engineering major. Suggesting a correction of the unobserved ability bias. An increase of one high school humanities course does have a larger effect on college performance than others changes.

To summarize, simulations suggest that increased of enrollment in STEM can be achieved by increasing quantitative requirements in high school. However, changes in high school curriculum will have a slight effect on college performance.

6 Conclusion

This paper investigates how the high school curriculum influences future college major choices and performance.

I establish panel data evidence linking an individual's high school skill sets with his choice of college major. I find that students usually choose a major in which they acquired more related skills in high school. However, I find a U-shaped relationship between diversification and college performance.

This result suggests that there is a tension between specialization and diversification. The link between high school and college is assessed through a model of high school human capital acquisition and college major choice. In the model, individuals with different initial abilities and preferences, who are uncertain about their preferences for particular college majors, choose a set of high school courses and a college major. Estimation of the structural parameters of the model suggests that high school course choice plays an important role in determining college major choice.

More quantitative high school courses make natural sciences, engineering, math and physics majors more attractive while more humanities courses are preferred by social sciences, humanities and business and communications majors. Moreover, the estimated model remarkably matches some central tendencies in the data.

I then exploit the model to evaluate and quantify the impact of education policies on enrollment in STEM majors. Policy experiments suggest that requiring students to take an additional high school quantitative course would boost enrollment in STEM majors by four percentage points. For the U.S., it means that one additional quantitative course in high school will contribute for around 16% of the additional STEM graduates needed by the President's Council of Advisors on Science and Technology to maintain U.S. competitiveness.

In this paper, I restrict my attention to the role played by high school specialization on college major choice and performance. Possible future research could investigate the effect of high school specialization on labor-market outcomes (e.g. unemployment and income). It would also be interesting to compare systems with forced specialization in high school (European-style systems) with more flexible systems (U.S.-style systems).

References

ALLENSWORTH, E., T. NOMI, N. MONTGOMERY, AND V. E. LEE (2009): "College Preparatory Curriculum for All: Academic Consequences of Requiring Algebra and English I for Ninth Graders in Chicago," *Educational Evaluation and Policy Analysis*, 31(4), 367–391.

ALTONJI, J. G. (1995): "The Effects of High School Curriculum on Education and Labor Market Outcomes," *Journal of Human Resources*, 30(3), 409–438.

ALTONJI, J. G., E. BLOM, AND C. MEGHIR (2012): "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers," *Annual Review of Economics*, 4(1), 185–223.

ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113(1), 151–184.

ARCIDIACONO, P. (2004): "Ability sorting and the returns to college major," *Journal of Econometrics*, 121(1-2), 343–375.

——— (2005): "Affirmative Action in Higher Education: How Do Admission and Financial Aid Rules Affect Future Earnings&quest;," *Econometrica*, 73(5), 1477–1524.

ARCIDIACONO, P., E. AUCEJO, AND V. J. HOTZ (2013): "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California," IZA Discussion Papers 7227, Institute for the Study of Labor (IZA).

ARCIDIACONO, P., AND J. B. JONES (2003): "Finite mixture distributions, sequential likelihood and the em algorithm," *Econometrica*, 71(3), 933–946.

BEFFY, M., D. FOUGÈRE, AND A. MAUREL (2012): "Choosing the Field of Study in Postsecondary Education: Do Expected Earnings Matter?," *The Review of Economics and Statistics*, 94(1), 334–347.

Bruce Daniel, S. (2007): *High school coursetaking findings from the Condition of education, 2007.* DIANE Publishing.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice*," *International Economic Review*, 44(2), 361–422.

Delavande, A., and B. Zafar (2014): "University Choice: The Role of Expected Earnings, Non-pecuniary Outcomes and Financial Constraints," .

Goodman, J. (2009): "The labor of division: returns to compulsory math coursework," Working paper series, Harvard University, John F. Kennedy School of Government.

Joensen, J. S., and H. S. Nielsen (2009): "Is there a Causal Effect of High School Math on Labor Market Outcomes?," *Journal of Human Resources*, 44(1).

Krugman, P. (1992): *Geography and Trade*, vol. 1 of *MIT Press Books.* The MIT Press.

Lee, V. E., R. G. Croninger, and J. B. Smith (1997): "Course-Taking, Equity, and Mathematics Learning: Testing the Constrained Curriculum Hypothesis in U.S. Secondary Schools," *Educational Evaluation and Policy Analysis*, 19(2), 99–121.

Levine, P. B., and D. J. Zimmerman (1995): "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business & Economic Statistics*, 13(2), 137–49.

Lind, J. T., and H. Mehlum (2010): "With or Without U? The Appropriate Test for a U-Shaped Relationship*," *Oxford Bulletin of Economics and Statistics*, 72(1), 109–118.

Malamud, O. (2010): "Breadth versus Depth: The Timing of Specialization in Higher Education," *LABOUR*, 24(4), 359–390.

——— (2012): "The Effect of Curriculum Breadth and General Skills on Unemployment," Discussion paper, University of Chicago and NBER.

McFadden, D. L. (1978): "Modelling the Choice of Residential Location," *Spatial Interaction Theory and Planning Models, ed. by A. Karlqvist, L. Lundqvist, F. Snikcars, and J.Weibull. New York: North-Holland,*, pp. 75–96.

Montmarquette, C., K. Cannings, and S. Mahseredjian (2002): "How do young people choose college majors?," *Economics of Education Review*, 21(6), 543–556.

Ning, J. (2014): "Do Stricter High School Math Requirements Raise College STEM Attainment?," Working paper series, Department of Economics, University of Notre Dame.

Palan, N. (2010): "Measurement of Specialization - The Choice of Indices," FIW Working Paper series 062, FIW.

Pavan, R., and J. Kinsler (2012): "The Specificity of General Human Capital: Evidence from College Major Choice," Discussion paper.

Rose, H., and J. R. Betts (2004): "The Effect of High School Courses on Earnings," *The Review of Economics and Statistics*, 86(2), 497–513.

Silos, P., and E. Smith (2015): "Human capital portfolios," *Review of Economic Dynamics*, 18(3), 635–652.

Smith, E. (2010): "Sector-Specific Human Capital and the Distribution of Earnings," *Journal of Human Capital*, 4(1), 35–61.

Stinebrickner, T. R., and R. Stinebrickner (2011): "Math or Science? Using Longitudinal Expectations Data to Examine the Process of Choosing a College Major," NBER Working Papers 16869, National Bureau of Economic Research, Inc.

Turner, S. E., and W. G. Bowen (1999): "Choice of major: The changing (unchanging) gender gap," *Industrial and Labor Relations Review*, 52(2), 289–313.

Zafar, B. (2009): "College major choice and the gender gap," Discussion paper.

A Appendix

A.1 Data

This appendix section describes the data used for estimations. First, I describe the sample selection. Second, I show how different high school courses are aggregated into human capital portfolios. Finally, I describe how I aggregate college majors.

Data used for estimations are obtained by merging the PEDS, Sophomores in 1980 - HS&B and high school transcript data sets. This first aggregation reduces the initial sample of 11,391 to 5,533 students who have both high school and college transcripts. Dropping students for whom there is no SAT data reduces the sample to 2,064 individuals, which includes students who did no enroll in college. Eliminating observations that are missing other control variables reduces the sample to 1,083 individuals that are used in the reduced-form analysis. To estimate the structural model the sample is 1,222 when accounting for non-college.

To construct high school course portfolios, courses are classified into seven broad areas of knowledge using the National Center for Education Statistics' Classification of Secondary School Courses (CSSC). The measure of human capital in each of these areas is the sum of courses taken in all subjects belonging to the same group of knowledge.[27]

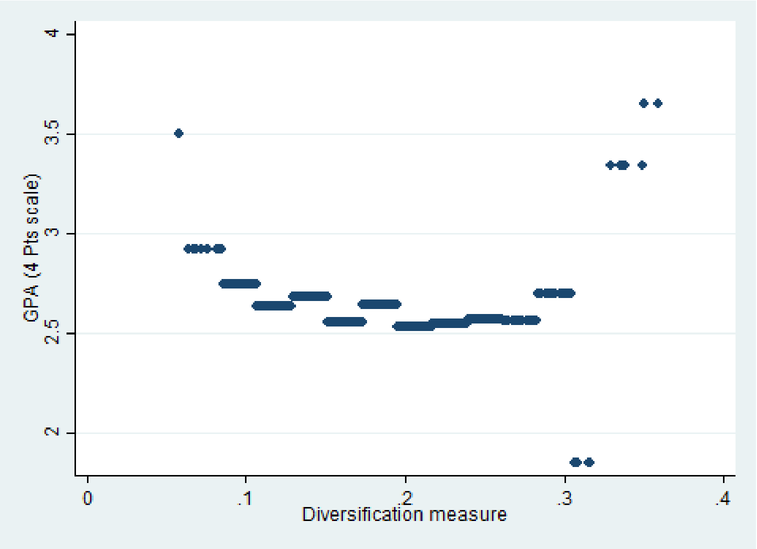
- Quantitative (math and physics): 04, 11, 15, 14, 27, 40,41
- Reading and writing:16, 23
- Social sciences and humanities: 05, 13, 19, 24, 37, 38, 39, 42, 43, 44, 45
- Natural and life sciences: 02,17, 18, 26, 34
- Business and communications: 01, 06, 22, 07, 08, 09,10
- Art: 21, 50
- Other: 03, 12, 20, 25, 28, 29, 30, 31, 32, 33, 35, 36, 46, 47, 48, 49, 54, 51, 55, 56

I also aggregate college majors into seven categories: math and physics, engineering, business and communications, social sciences and humanities, natural sciences, education, and health. The criteria for aggregation is the degree of similarity in field topics. Here is a list of majors by category:

[27] The number for each field corresponds to CSSC codes.

- Math and physics: Physics, science technologies, mathematics, Calculus, communication technologies, computer and information sciences, and computer programming.
- Engineering: Engineering, civil engineering, electrical and communications engineering, mechanical engineering, and architecture and environmental design.
- Business and communications: Construction trades, business and management, accounting, banking and finance, business and office, secretarial and related programs, marketing and distribution, communications, journalism, precision production, and transportation and material moving.
- Natural and life sciences: Geology, life sciences, geography, and renewable natural resources, biology, chemistry.
- Social sciences and humanities: Area and ethnic studies, foreign languages, home economics, vocational home economics, law, letters, composition, American literature, English literature, philosophy and religion, theology, psychology, protective services, public affairs, social work, social sciences, anthropology, economics, geography, history, political science & government, sociology, visual and performing Arts, dance, fine arts, music, and liberal/general studies.
- Education: Education, adult and continuing education, elementary education, junior high education, pre-elementary education, secondary education.
- Health: Allied health, practical nursing, health sciences, nursing.

Figure 1: Relationship between GPA and diversification measure.



NB: This figure shows collapse mean by bins of 0.022 length of the diversification measure. It suggests a quadratic relationship.

Table 1: Summary statistics

	Dropped sample			Restricted sample				
	Mean	Std. Err.	Obs.	Mean	Std. Err.	Obs.	t	Pr(|T| > |t|)
Female	.523	.0184	737	.559	.014	1,184	-1.510	0.130
Black	.135	.012	737	.098	.008	1,184	2.482	0.013
SAT-Math	478.23	4.423	737	477.56	3.247	1,184	0.120	0.901
SAT-verbal	442.79	4.176	737	440.55	3.017	1,184	0.440	0.657
College GPA	2.413	.103	109	2.620	.020	1,184	-1.96	0.052
High school share of courses								
Reading and Writing	.246	.002	737	.249	.001	1,184	-1.056	0.290
Math	.132	.001	737	.130	.001	1,184	1.177	0.239
Life Science	.164	.002	737	.167	.001	1,184	-1.049	0.294
Physics	.062	.001	737	.064	.001	1,184	-0.996	0.310
Humanities	.199	.002	737	.197	.002	1,184	0.649	0.515
Bussiness and Communication	.058	.001	737	.060	.001	1,184	-0.956	0.338

NB: This table provides the mean and standard error of some variables in the dropped sample and in the restricted one. The initial sample is the sample are those who have participated to the first follow-up and have their SATs scores. t represents the two-sample difference in mean. t statistic and Pr(|T| > |t|) the p-value of the test. There is not a large difference between most of the variable in both sample, suggesting that sample selection may not be an issue.

Table 2: High school human capital portfolios by college major

College Major \ Share HS courses	Quant.	R. & W.	Life sci.	Humanities	Bus. & Com.	Arts	Others
Bus. & com.	.166	.233	.167	.189	**.095**	.065	.086
Natural sciences	.222	.252	**.188**	.176	.039	.063	.060
Math and physics	**.227**	.244	.164	.186	.057	.058	.063
Education	.165	.230	.173	.182	.075	**.094**	.081
Engineering	**.227**	.226	.173	.171	.050	.065	.088
Social sci./hum.	.188	**.259**	.162	**.199**	.055	.066	.071
Health	.171	.232	.181	.190	.075	.071	.081
Other	.162	.222	.173	.176	.062	.089	.116
F	50.218	12.651	3.385	5.174	37.233	9.784	6.410
P-value	0.000	0.000	0.001	0.000	0.000	0.000	0.000

NB: This table shows the mean share of high school subjects by college major. In bold are the largest, relative to other majors, the share of a specific high school subject. The last two rows show the F statistics and p-values for the test of significance for the difference in means. For all the subjects, the null hypothesis of mean equality is rejected at 1%.

Table 3: Reduced form estimation results for college performance (GPA as the dependent variable)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ρ	-7.756***	-9.108***	-9.638***	-6.783**	-5.647**	-5.216*	-5.216*	-5.001*
	(2.63)	(3.20)	(3.15)	(3.07)	(2.83)	(2.90)	(2.90)	(2.68)
ρ^2	19.004***	20.252**	24.099***	19.535**	16.057**	14.830*	14.830*	14.018*
	(6.81)	(8.82)	(8.94)	(8.56)	(7.67)	(7.87)	(7.87)	(7.33)
Female		0.145**	0.204***	0.161***	0.184***	0.183***	0.183***	0.245***
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)
Black		-0.333***	-0.195**	-0.179**	-0.158*	-0.172*	-0.172*	-0.237***
		(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)	(0.07)
SES		0.008	-0.060	-0.056	-0.040	-0.040	-0.040	-0.019
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)
SAT Math			0.171***	0.157***	0.095**	0.086**	0.086**	0.090***
			(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)
SAT verbal			0.113***	0.100***	0.046	0.037	0.037	0.091**
			(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Major fixed-effect				X	X	X	X	X
Soph. Test				X	X	X	X	X
High School course level						X	X	X
High school region							X	X
Constant	3.349***	3.490***	2.126***	1.723***	1.300***	1.798***	1.798***	1.403***
	(0.24)	(0.29)	(0.31)	(0.30)	(0.36)	(0.50)	(0.50)	(0.34)
Observations	1157	1157	1157	1157	1083	1083	1083	1083
R2	0.01	0.04	0.17	0.21	0.21	0.23	0.23	0.26
Numberofgroups		366	366	366	345	345	345	
R2overall		0.05	0.20	0.24	0.24	0.23	0.23	

NB: *** denotes significance at the 1% level, ** denotes significance at the 5% level, and * denotes significance at the 10% level. Heteroskedasticity-robust standard errors are clustered by high school in parentheses. Column (1) and (8) are ordinary least square. While Column (2) to (7) estimate ordinary least squares with a high school fixed-effect. Lind and Mehlum (2010) test for U-shape, Overall test for presence of a U-shape: t-value=2.10 $P > |t| = .0182$

Table 4: Performance regressions: Structural model.

	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
ρ	-4.216**	1.888	-3.019*	1.738
ρ^2	12.603**	4.956	11.234**	4.559
Female	0.241***	0.038	0.238***	0.035
SATM	0.119***	0.023	0.101***	0.021
SATV	0.123***	0.023	0.129***	0.021
SES	-0.011	0.027	0.006	0.025
Business and Communication	0.351***	0.050	0.332***	0.046
Natural science	0.360***	0.102	0.358***	0.093
Math & Physics	0.363***	0.078	0.353***	0.072
Education and Military	0.411***	0.103	0.394***	0.095
Engineering	0.161*	0.087	0.185**	0.080
Humanities	0.382***	0.054	0.357***	0.049
Health	0.408***	0.086	0.332***	0.079
High school courses				
Humanities	-0.003	0.007	-0.005	0.007
Reading and Writing	-0.004	0.009	0.033	0.009
Math	-0.012	0.018	-0.003	0.016
Physics	0.005	0.016	0.019	0.015
Life science	0.005	0.008	0.005	0.008
Business and Com.	0.004	0.011	0.024	0.011
Other	-0.008	0.006	-0.003	0.005
Type 1			0.51***	0.03
Variance	0.59***	0.012	0.549***	0.011

NB: Major-specific constant terms are included along with courses taken in high school. *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 5: Utility parameters estimates (1/2)

		One type		Two types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
Math & Physics courses					
	Business and Communication	-0.058**	0.027	0.005	0.052
	Natural science	0.011	0.134	0.107	0.121
	Math & Physics	0.115	0.101	0.329***	0.092
	Education and Military	-0.131	0.058	-0.025	0.060
	Engineering	0.139	0.102	0.298***	0.095
	Humanities	-0.125**	0.056	-0.023	0.056
	Health	0.010	0.130	-0.045	0.084
Life science courses					
	Business and Communication	-0.022	0.019	0.071*	0.041
	Natural science	-0.025	0.120	0.064	0.108
	Math & Physics	-0.123	0.090	0.084	0.085
	Education and Military	-0.045	0.044	0.056	0.049
	Engineering	-0.076	0.092	0.083	0.086
	Humanities	-0.068	0.040	0.022	0.047
	Health	-0.028	0.115	-0.092	0.069
Humanities courses					
	Business and Communication	-0.054***	0.017	-0.002	0.041
	Natural science	-0.204*	0.119	-0.130	0.106
	Math & Physics	-0.190**	0.087	0.006	0.081
	Education and Military	-0.102**	0.041	-0.010	0.048
	Engineering	-0.170**	0.088	-0.027	0.082
	Humanities	-0.120***	0.037	-0.037	0.046
	Health	-0.194*	0.115	-0.269***	0.067
Biss. courses					
	Business and Communication	-0.004	0.027	0.112**	0.050
	Natural science	-0.269*	0.143	-0.207	0.133
	Math & Physics	-0.177	0.112	0.028	0.104
	Education and Military	-0.047	0.057	0.050	0.060
	Engineering	-0.173	0.115	-0.031	0.109
	Humanities	-0.100*	0.053	-0.019	0.060
	Health	-0.178	0.134	-0.233**	0.089
Languages courses					
	Business and Communication	-0.010	0.020	0.087**	0.044
	Natural science	-0.132	0.125	-0.036	0.112
	Math & Physics	-0.169*	0.096	0.044	0.089
	Education and Military	-0.046	0.045	0.044	0.053
	Engineering	-0.139	0.098	0.015	0.091
	Humanities	-0.016	0.043	0.078*	0.047
	Health	-0.095	0.121	-0.129*	0.073
Arts courses					
	Business and Communication	-0.031	0.020	0.039	0.044
	Natural science	-0.238*	0.126	-0.180	0.113
	Math & Physics	-0.180*	0.096	0.010	0.086
	Education and Military	0.001	0.048	0.103**	0.046
	Engineering	-0.206**	0.098	-0.070	0.090
	Humanities	-0.093**	0.041	-0.018	0.051
	Health	-0.191	0.122	-0.278***	0.077

Table 6: Utility parameters estimates (2/2)

		One type		Two types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
Others courses					
	Business and Communication	-0.047**	0.019	0.008	0.037
	Natural science	-0.275**	0.124	-0.197*	0.111
	Math & Physics	-0.156*	0.087	0.038	0.076
	Education and Military	-0.090**	0.043	0.001	0.043
	Engineering	-0.137	0.088	0.006	0.077
	Humanities	-0.118***	0.040	-0.039	0.043
	Health	-0.245**	0.119	-0.313***	0.073
Female					
	Business and Communication	0.033	0.083	0.111	0.168
	Natural science	0.318	0.281	0.384	0.289
	Math & Physics	-0.488**	0.246	-0.335	0.249
	Education and Military	0.100	0.172	0.163	0.177
	Engineering	-0.916***	0.258	-0.843***	0.266
	Humanities	0.088	0.167	0.142	0.169
	Health	0.711**	0.283	0.884***	0.292
State Wage 80					
	Business and Communication	0.013	0.038	0.140**	0.073
	Natural science	-0.077	0.131	0.062	0.132
	Math & Physics	-0.048	0.111	0.104	0.110
	Education and Military	-0.018	0.080	0.081	0.079
	Engineering	-0.108	0.114	0.004	0.115
	Humanities	0.014	0.077	0.119	0.074
	Health	-0.035	0.127	0.117	0.124
GPA					
	Business and Communication	0.187	0.136	1.712***	0.292
	Natural science	-0.465	1.120	0.331	0.973
	Math & Physics	-0.145	0.839	2.141***	0.778
	Education and Military	0.559	0.361	1.770***	0.362
	Engineering	-0.110	0.862	1.546**	0.808
	Humanities	0.284	0.308	1.373***	0.359
	Health	0.160	1.099	-0.495	0.509
SATM					
	Business and Communication	-0.007	0.051	0.014	0.103
	Natural science	-0.043	0.170	0.082	0.177
	Math & Physics	0.386***	0.144	0.419***	0.147
	Education and Military	-0.086	0.105	-0.077	0.108
	Engineering	0.467***	0.146	0.525***	0.149
	Humanities	-0.120	0.101	-0.121	0.103
	Health	-0.358**	0.166	-0.320*	0.171
SATV					
	Business and Communication	0.071	0.051	0.185	0.106
	Natural science	0.225	0.171	0.281	0.180
	Math & Physics	-0.073	0.147	0.004	0.151
	Education and Military	0.136	0.106	0.183	0.112
	Engineering	-0.176	0.150	-0.132	0.154
	Humanities	0.248**	0.102	0.319***	0.106
	Health	0.133	0.166	0.158	0.172
GPA × Hs Courses					
	Business and Communication	0.116***	0.039	-0.191**	0.088
	Natural science	0.533	0.370	0.262	0.311
	Math & Physics	0.355	0.281	-0.391	0.251
	Education and Military	0.187	0.108	-0.197	0.114
	Engineering	0.285	0.290	-0.268	0.260
	Humanities	0.271***	0.091	-0.075	0.113
	Health	0.396	0.364	0.605***	0.158
Type 1					
	Business and Communication			-0.119	0.185
	Natural science			0.029	0.314
	Math & Physics			0.007	0.274
	Education and Military			-0.095	0.193
	Engineering			-0.029	0.281
	Humanities			-0.080	0.187
	Health			0.186	0.303
Nesting Parameter		0.266***	0.012	0.331***	0.023

NB:major-specific constant terms were also included. *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 7: High school courses choices estimations

	Quantitative courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	-0.002	0.002	-0.002	0.002
Reading	0.001	0.002	0.001	0.002
Math	0.000	0.001	0.000	0.001
Science	0.003***	0.001	0.003***	0.001
Writing	-0.002**	0.001	-0.002**	0.001
Civic	-0.000	0.001	-0.000	0.001
Expected GPA	2.403***	0.041	2.436***	0.048
Expected GPA interacted with major				
Business and Communication	-0.850**	0.440	-0.906**	0.442
Natural science	-0.352	1.283	-0.411	1.282
Math & Physics	-1.504**	0.619	-1.529**	0.619
Education and Military	-0.965	0.725	-0.968	0.725
Engineering	-1.027	0.799	-1.062	0.799
Humanities	-1.397***	0.455	-1.412***	0.455
Health	-0.906	0.999	-0.906	0.999
Major				
Business and Communication	1.300	1.213	1.448	1.217
Natural science	0.918	3.729	1.076	3.727
Math & Physics	2.664	1.723	2.734	1.723
Education and Military	2.680	2.074	2.684	2.072
Engineering	3.547	2.038	3.635	2.038
Humanities	2.910***	1.286	2.946***	1.285
Health	1.438	2.815	1.447	2.814
Type 1			-0.141	0.104
Variance	1.766***	0.037	1.765***	0.037

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 8: High school courses choices estimations

	Reading and Writing courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	-0.001	0.002	-0.001	0.002
Reading	0.001	0.002	0.001	0.002
Math	0.000	0.001	0.000	0.001
Science	0.001	0.001	0.001	0.001
Writing	0.000	0.001	-0.001	0.001
Civic	-0.000	0.001	-0.000	0.001
Expected GPA	3.127***	0.047	3.393***	0.052
Expected GPA interacted with major				
Business and Communication	-1.571***	0.502	-2.029***	0.484
Natural science	-3.504**	1.463	-3.980***	1.406
Math & Physics	-2.301***	0.706	-2.503***	0.679
Education and Military	-2.754***	0.827	-2.775***	0.794
Engineering	-4.493***	0.912	-4.773***	0.876
Humanities	-1.311	0.519	-1.434***	0.499
Health	-2.486**	1.140	-2.483**	1.095
Major				
Business and Communication	3.144**	1.383	4.356***	1.334
Natural science	8.789**	4.252	10.079**	4.087
Math & Physics	4.747**	1.965	5.322***	1.889
Education and Military	6.095**	2.365	6.128**	2.272
Engineering	10.581***	2.324	11.29***	2.234
Humanities	2.753*	1.466	3.046**	1.409
Health	5.633*	3.211	5.702*	3.085
Type 1			-1.147***	0.114
Variance	2.014***	0.042	1.936***	0.041

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 9: High school courses choices estimations

	Humanities			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	0.002	0.003	0.002	0.003
Reading	-0.002	0.003	-0.002	0.003
Math	0.001	0.001	0.001	0.001
Science	0.002	0.002	0.002	0.002
Writing	-0.003	0.002	-0.003	0.002
Civic	-0.001	0.001	-0.001	0.001
Expected GPA	2.730***	0.058	2.749***	0.068
Expected GPA interacted with major				
Business and Communication	-3.749***	0.626	-3.782***	0.629
Natural science	-7.188***	1.826	-7.222 ***	1.827
Math & Physics	-3.718***	0.881	-3.733***	0.882
Education and Military	-2.875***	1.032	-2.877***	1.032
Engineering	-2.898***	1.138	-2.919**	1.138
Humanities	-4.934***	0.648	-4.943***	0.648
Health	-3.198***	1.423	-3.197***	1.422
Major				
Business and Communication	8.919***	1.726	9.007***	1.733
Natural science	18.863***	5.308	18.956***	5.310
Math & Physics	8.393***	2.453	8.434***	2.454
Education and Military	6.071**	2.952	6.073**	2.952
Engineering	5.963**	2.901	6.014***	2.902
Humanities	12.468***	1.830	12.489***	1.830
Health	7.315*	4.008	7.320*	4.008
Type 1			-0.083	0.148
Variance	2.515***	0.052	2.514***	0.052

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 10: High school courses choices estimations

	Business and Com.			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	0.000	0.002	0.000	0.002
Reading	-0.000	0.002	-0.000	0.002
Math	-0.001	0.001	-0.001	0.001
Science	0.001	0.001	0.001	0.001
Writing	0.001	0.001	0.001	0.001
Civic	-0.001	0.001	-0.001	0.001
Expected GPA	0.761***	0.046	0.810***	0.053
Expected GPA interacted with major				
Business and Communication	-1.780***	0.490	-1.866***	0.491
Natural science	-1.251	1.497	-1.342	1.498
Math & Physics	-1.333**	0.693	-1.372	0.692
Education and Military	-1.535*	0.827	-1.541	0.826
Engineering	-1.395	0.915	-1.448	0.915
Humanities	-2.599***	0.519	-2.621***	0.519
Health	-1.361	1.104	-1.359	1.103
Major				
Business and Communication	4.922***	1.348	5.150***	1.351
Natural science	2.042	4.348	2.285	4.348
Math & Physics	3.246	1.927	3.353	1.926
Education and Military	3.455	2.363	3.460	2.359
Engineering	2.494	2.331	2.628	2.330
Humanities	6.366***	1.461	6.417***	1.459
Health	3.598	3.110	3.611	3.105
Type 1			-0.208*	0.116
Variance	1.935***	0.047	1.932***	0.047

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 11: High school courses choices estimations

	Life science courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	-0.001	0.002	-0.001	0.002
Reading	-0.000	0.002	-0.000	0.002
Math	0.001	0.001	0.001	0.001
Science	-0.002**	0.001	-0.002**	0.001
Writing	0.002**	0.001	0.002**	0.001
Civic	-0.002**	0.001	-0.002**	0.001
Expected GPA	2.152***	0.051	2.101***	0.059
Expected GPA interacted with major				
Business and Communication	-2.637***	0.550	-2.548***	0.551
Natural science	0.182	1.602	0.274	1.601
Math & Physics	-1.738**	0.774	-1.699**	0.773
Education and Military	-3.119***	0.906	-3.114***	0.905
Engineering	-3.700***	0.999	-3.646***	0.998
Humanities	-2.046***	0.569	-2.022***	0.568
Health	-1.856	1.248	-1.857	1.247
Major				
Business and Communication	6.652***	1.515	6.417***	1.519
Natural science	-0.764	4.656	-1.014	4.653
Math & Physics	3.945*	2.152	3.834*	2.151
Education and Military	7.687***	2.590	7.679***	2.587
Engineering	9.172***	2.546	9.034***	2.544
Humanities	4.579***	1.606	4.522***	1.604
Health	4.585	3.517	4.571	3.513
Type 1			0.217*	0.130
Variance	2.207***	0.046	2.207***	0.046

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 12: High school courses choices estimations

	Art			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	0.000	0.003	0.000	0.003
Reading	-0.001	0.003	-0.001	0.003
Math	0.000	0.001	0.000	0.001
Science	0.002	0.002	0.002	0.002
Writing	-0.003	0.002	-0.003	0.002
Civic	0.002**	0.001	0.002**	0.001
Expected GPA	0.481***	0.076	0.525***	0.087
Expected GPA interacted with major				
Business and Communication	-0.830	0.793	-0.877	0.796
Natural science	-0.387	2.348	-0.439	2.350
Math & Physics	0.645	1.092	0.624	1.092
Education and Military	-2.341*	1.313	-2.344*	1.312
Engineering	-3.094**	1.486	-3.122**	1.485
Humanities	-0.097	0.831	-0.113	0.831
Health	-0.518	1.775	-0.515	1.774
Major				
Business and Communication	2.076	2.183	2.203	2.191
Natural science	0.610	6.834	0.743	6.840
Math & Physics	-0.426	3.038	-0.366	3.038
Education and Military	6.171*	3.745	6.173*	3.745
Engineering	6.744*	3.757	6.818*	3.756
Humanities	-0.051	2.348	-0.022	2.348
Health	1.175	5.002	1.184	5.000
Type 1			-0.189	0.189
Variance	3.045***	0.084	3.044***	0.084

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10%

Table 13: Comparing model predictions of high school course selection with the data

	Quantitative			GPA		
	Data	One type	Two types	Data	One type	Two types
Business and Communication	5.5843	5.4969	5.4969	2.7376	2.6819	2.6813
Natural science	6.8537	6.814	6.814	2.9	2.8024	2.8186
Math & Physics	6.7722	6.6724	6.6724	2.8519	2.7864	2.7977
Education and Military	5.1712	5.1171	5.1171	2.7757	2.7042	2.6974
Engineering	7.0323	6.9576	6.9576	2.5371	2.4679	2.4722
Humanities	5.7111	5.6854	5.6854	2.8004	2.7504	2.7548
Health	5.6379	5.4762	5.4762	2.8103	2.7795	2.7802
	Humanities			Female		
	Data	One type	Two types	Data	One type	Two types
Business and Communication	7.4392	7.386	7.386	0.5882	0.5925	0.5916
Natural science	7.6829	7.5517	7.5517	0.439	0.4416	0.4504
Math & Physics	7.1519	7.1254	7.1254	0.4557	0.4539	0.4592
Education and Military	7.0541	6.9898	6.9898	0.6577	0.653	0.6516
Engineering	7.1129	6.9732	6.9732	0.1452	0.1292	0.1319
Humanities	7.8622	7.8389	7.8389	0.6089	0.6048	0.6045
Health	7.431	7.3262	7.3262	0.8448	0.8211	0.8249
	Life sciences					
	Data	One type	Two types			
Business and Communication	5.2275	5.1408	5.16			
Natural science	5.8293	5.6701	5.77			
Math & Physics	4.8608	4.7313	4.77			
Education and Military	4.982	4.9422	4.95			
Engineering	5.1452	5.1113	5.17			
Humanities	4.7778	4.7711	4.7711			
Health	5.3103	5.1526	5.19			

NB: The data column contains the actual mean from the data. One type refers to estimates using one type of individual, and two types refers to estimates using two types of individuals.

Table 14: Simulations of the change in major choice distribution and changes in college performance (GPA)

		Simulations			
One type		(1)	(2)	(3)	(4)
Major Choice	Math, phys. & eng. majors	0.026	-0.004	-0.006	-0.051
	Natural sciences & Health	0.012	0.008	-0.002	-0.049
	Humanities	-0.038	-0.003	-0.000	-0.035
	No college	-0.000	-0.000	0.008	0.135
GPA (% Changes)					
	Business and Communication	-1.139	-1.337	-1.124	-0.974
	Natural science	-0.760	-1.038	-0.849	-1.377
	Math & Physics	-0.914	-1.143	-0.852	-1.959
	Education and Military	-1.707	-1.930	-1.710	-1.226
	Engineering	0.260	0.097	0.373	-0.882
	Humanities	-1.062	-1.275	-1.054	-1.018
	Health	-1.831	-2.041	-1.865	-2.483
Two types					
Major choice	Math, phys. & eng.	0.027	-0.005	-0.005	0.089
	Natural sciences & Health	0.013	0.007	-0.002	-0.082
	Humanities	-0.036	-0.003	0.000	0.042
	No college	-0.003	-0.003	0.007	-0.049
GPA(% changes)					
	Business and Communication	0.409	1.109	0.655	1.793
	Natural science	0.741	1.281	0.862	-0.175
	Math & Physics	0.891	1.502	1.158	0.899
	Education and Military	1.205	1.881	1.432	1.900
	Engineering	-0.726	-0.004	-0.394	-1.194
	Humanities	0.962	1.627	1.192	0.390
	Health	2.664	3.343	2.854	1.702

NB: Simulation (1): One additional quantitative course in high school. Simulation (2): One additional life sciences course in high school. Simulation (3): One additional humanities course in high school. Simulation (4): The same curriculum imposed to all high school students.