ECONSTOR
Make Your Publications Visible.

A Service of

ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Tchuente, Guy

Working Paper

High school human capital portfolio and college outcomes

School of Economics Discussion Papers, No. 1516

Provided in Cooperation with:
University of Kent, School of Economics

Suggested Citation: Tchuente, Guy (2015) : High school human capital portfolio and college outcomes, School of Economics Discussion Papers, No. 1516, University of Kent, School of Economics, Canterbury

This Version is available at:
https://hdl.handle.net/10419/129980

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

University of Kent

School of Economics Discussion Papers

High school human capital portfolio and college outcomes

Guy Tchuente

July 2015

KDPE 1516

High school human capital portfolio and college outcomes*

Guy Tchuente†

University of Kent

July 2015

Abstract

This paper assesses the relationship between courses taken in high school and college major choice. Using High School and Beyond survey data, I study the empirical relationship between college performance and different types of courses taken during high school. I find that students sort into college majors according to subjects in which they acquired more skills in high school. However, I find a U-shaped relationship between the diversification of high school courses a student takes and their 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.

Keywords: Human capital, Discrete choice, College major.

JEL classification codes: J24, I21.

*I 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 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

Non-technical summary

There has been an emphasis in policy aimed at encouraging enrollment in STEM (Science, Technology, Engineering and Math) majors. In the U.K., for example, the Royal Academy of Engineering reported that the nation will have to produce 100,000 STEM graduate every year until 2020. This paper is interested in understanding how the composition of skills, acquired in high school, affects college field of study. I

The literature suggests that college major choice is associated with ability sorting. This sorting is driven either by variations in the cost of successfully completing degree requirements, or variations in expected returns to different majors by ability in these majors. Arcidiacono (2004) finds that predetermined factors, such as preferences and quantitative skills (preparation), play a larger role in the choice of major than the economic returns. This finding implies that what happens before college could play an important role on the type of major chosen. I therefore investigate the role of high school education in developing quantitative skills and evaluate the potential effectiveness of high school curriculum changes that promote enrollment and success in STEM majors.

Empirical evidence suggests 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 specialize in humanities courses. However, 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) than those who slightly diversify. These results acknowledge the importance of high school curriculum on college major choice and performance.

Given the relationship between high school and college, I estimate a structural model of high school human capital acquisition and college major choice. The explicit modeling of the educational decision-making process, helps both to disentangle the heterogeneous effects of specialization and control for the self-selection inherent in educational outcomes. 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.

1 Introduction

This paper assess 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 specialize in humanities courses. Natural sciences and health majors take a mix of quantitative and humanities courses in high school.

I also find 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 specialize in a different area. 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 major 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 two 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 in a particular subject. Second, I examine the expected outcome if all students faced the same high school curriculum, thus eliminating the possibility to specialize in a particular subject area in high school. Both experiments substantially affect enrollment in STEM majors.

Taking an additional quantitative course in high school increases enrollment in STEM majors by four percentage points. Taking an additional humanities course in high school also has a positive effect on enrollment in STEM majors. An additional life sciences course in high school has the smallest effect on a student's choice of college major. An additional life sciences course increases enrollment in natural science majors by 0.015 percentage points and reduces enrollment in math, physics and engineering majors by the same amount. Imposing a single curriculum on all high school students also boosts enrollment in STEM majors. This suggests that high school specialization plays a key role in influencing what majors students choose.

There is a very extensive literature on college major choice.[1] 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 detail). Turner and Bowen (1999) document the sorting that occurs across majors by 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 outcomes. Silos and Smith (2013) 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

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

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. It also discuss some data irregularities and provides a reduced-form analysis of the relationship between diversification in high school and college performance. The 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[2] 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 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

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

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.[3]

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 the 1980, 1982, 1984 and 1986 surveys.

The HS&B survey contains 14,825 students. A sub-sample of 5,533 students have transcripts encoded for both high school and college. Dropping those who do not have SAT data reduces the sample to 2,064 individuals. Eliminating observations that are missing other control variables reduces the sample to 1,265. Cleaning the data yields a

[3]High school usually runs 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.

final sample of 1,112 students for estimation of my structural model.

Table 2 shows the average characteristics for the unrestricted and restricted samples. In almost all cases, there is no significant difference in mean values between the two samples. This suggests that sample selection issues are not 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 major choice and performance. 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.[4] High school courses are grouped into the following categories: (i) quantitative (mathematics and physics), (ii) reading and writing, (iii) social sciences and humanities, (iv) life sciences, (v) business and communications, (vi) arts, and (vii) other.[5]

Given courses taken in each field (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 proportion of courses taken in subject k.[6] Table 2 displays these portfolio weights by major across the population. For each major, the table displays the mean, 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 humanities classes in high

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

[5]My results are not sensitive to the structure of these categories; I considered other potential categories and obtained similar results.

[6]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. I also consider other diversification measures, such as the Gini index. The results obtained are qualitatively the same.

school (0.258) 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 mean in some subjects appears small, the last two rows of Table 2 shows that these differences are statistically significant.

Each student i has a vector of human capital weights, $\omega_{i,k}$, which measure the weight of skill type k in the overall portfolio. 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. To evaluate the level of specialization, I follow Silos and Smith (2012); 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.[7]

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.

3.2.1 Estimation results

Table 3 presents regression estimates linking college GPAs and the portfolio distance measure, ρ. This helps us investigate the data beyond raw mean difference.

I estimate the following reduced-form equation:

$$G_i = \alpha_0 + \alpha_1 \rho_{im} + \alpha_2 \rho_{im}^2 + \alpha_3 X_i + \alpha_m + \alpha_h + \varepsilon_{ih}$$

[7] This measure is related to the diversification index in trade from Krugman (1992), which uses an absolute distance instead of a square root. Additionally, see Palan (2010) for a review of the specialization index in trade.

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.

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, SES, parents' education, ability (measured by SAT-Math and SAT-Verbal scores) and the number of courses taken in each high school subject. It is also robust to regional disparities by including a dummy variable for living in the southern U.S. The major-specific effect is controlled for by including a dummy variable for each major. It is worth noting that the inclusion of more control variables increases the effect of specialization on college performance. This suggests that the effect of specialization on performance may be larger than the estimates reported here.

3.2.2 U-shaped relationship between high school diversification and college performance

The results from Table 3 show a U-shaped relationship between college GPA and diversification of high school courses. This suggests a trade-off between specialization and diversification. This trade-off 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 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.

To formally test for the presence of a U-shaped relationship between diversification and performance, I use the procedure proposed by Lind and Mehlum (2010). The results, in Table 4, show that there is indeed a U-shape relationship. I also perform a non-parametric robustness check. I run a regression on all control variables used in the best regression in Table 3. I perform a non-parametric regression of the residual on ρ. The predicted values of the residual of the non-parametric regression show a U-shaped relationship between diversification and performance. The results are shown in Figure 1.

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. High school course choices are based on these differences.

I assume that students know their ability to acquire imperfectly substitutable skills. They choose their high school courses to maximize their expected utility across college majors. Upon graduation, students choose to pursue a particular college major, or do not enroll in college.

A student's initial ability and their preferred college major provide an incentive for the student to specialize by acquiring skills that reflect their 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 skills.

I suppose that individuals with discount factor $\beta \in (0, 1)$ live for a finite number of discrete periods, $t = 0, 1, 2, ...T$. Individuals 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 three types of skills that are useful for all majors. High school skills are useful in college, but their importance differs from one major to another.[8] I denote an individual's portfolio of human capital by $s = (s_Q, s_H, s_{NS})$, where s_Q is quantitative human capital, s_H is humanities human capital and s_{NS} is natural sciences human capital.[9] Individuals can choose their portfolio composition by selecting more high school courses in a particular skill area.[10]

Before choosing s, 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.

The cost of accumulating s with ability τ is $c^h(s, \tau)$, with c^h convex and twice differentiable. Individuals know how useful each type of human capital is for each college major. However, individuals are unsure about an idiosyncratic component of their college preferences.

Once an individual has acquired a skill set s, they decide whether or not to enter

[8] In the empirical framework, I use seven different fields. Here, I focus on three skills, both for computational ease, as well as to focus on the most important skills.

[9] 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, and business and communication. Natural sciences human capital is the number of high school life sciences courses.

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

college in period $t = 1$. Individuals who choose to enter college also select a major. Although individuals 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 they complete high school and enter college that their true fit in a major becomes known; actual experience in a major reveals an individual's true preference for that major.[11]

The timing of the model is as follows:

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

4.1 High school and college stages

I assume that college GPA (G) is a function of individual abilities, as well as X_G, which represents other demographic characteristics, such as gender and SES.[12] Specifically, performance in college takes the following form:

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

The model also contains a major-specific fixed effect, ε_m, as well idiosyncratic shocks (the ε_1's), which are drawn from distribution $\mathcal{N}(0; \sigma_G^2)$.

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 students to be more predisposed to a particular major.

[11] For simplicity, I assume that students make a one-time decision about their college major; I ignore the possibility that students may do post-graduate work or drop out of college.

[12] Due to data limitations, I do not include wages in the model. Given that GPA has a positive effect on future earnings (see Arcidiacono (2004)), I use it as 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))

The utility from being in high school is given by

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

where ε is normally distributed. High school and college cost functions are $c_m(s, G)$ and $c^h(\tau, s)$, respectively.

I assume that marginal cost of acquiring human capital k is:

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

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

where Mc_k^h is the marginal cost of acquiring skill k in high school. Mc_{mk} is the marginal cost of acquiring skill k in major m. ϑ_{4mk} and ϑ_{5mk} are the cost elasticity contribution of producing grade in major m of human capital type k. $\vartheta_{.mk}$ is observed with error; that is why I control for major-specific fixed effects, v_m. Integrating on different dimensions of human capital will give the effort cost function. This cost of effort 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.

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_m^c, 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.[13]

[13]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 non0negative, homogeneous of degree 1, approaches $+\infty$ as one of its arguments approaches $+\infty$, 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

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

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

The error terms are known to the individual, but they are not observed by the econometrician. Therefore, from the econometrician's perspective, the probability of choosing a major m is given by

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

Before choosing a major, individuals first choose their high school human capital acquisition. 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_1^c 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 s that yields the highest utility $V_0(s,\tau)$:

$$V_0(s,\tau) = u^h + \beta E_0(u_1^c|\tau)$$

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).

For each type of human capital, s_k^* is the optimal value of s_k that solves the Euler equation

$$Mc_k = \beta_1 E_0(u_{1k}^c|\tau).$$

If I apply the envelope theorem on u_1^c, I get $E_0(u_{1m}^c|\tau) = \beta\vartheta_{0ck} - \beta E_0(MC_k(s,G)))$ and

$$\vartheta_{4hk}s_k + \vartheta_{5hk}\tau_k = \beta\vartheta_{0ck} - \beta(MC_k(s,G))$$

Thus,

$$s_k^* = \theta_{0\hat{m}k} + \theta_{1\hat{m}k}\tau_k + \theta_{2\hat{m}k}G$$

Let $\tilde{s}_j^*$ be a latent variable with

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

where ε_k is the normal forecast error.

The observed chosen basket of high school courses, s_k, is

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

The forecast error, ε_k, is independent of τ, G and m. I estimate the coefficients of the model with a Tobit model.

4.3 Identification and estimation strategy

In this section, I discuss how several key parameters of the model are identified.

4.3.1 Identification without unobservables

All individual characteristics are exogenous, including test scores and GPA in college, high school courses and 10^{th}-grade standardized test score. One of the main advantage of HS&B data is that for all individuals in the sample, there are base-year test scores in different subjects. These scores are in math, science, civics, reading and writing and are my main exogenous variables. 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.

4.3.2 Identification 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 initially, he is just as likely as someone who has a weak preference for high school 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. Some variables can be used to identify types: initial ability (here measured by base-year standardized test scores), the level of human capital and college major choice.

4.3.3 Estimation method

I first estimate a model with independent errors across grades and choice processes. 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. Then, the η are replaced by consistent estimates of 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.

Following Arcidiacono (2004, 2005), I assume that there are $R = 2$ types of people.[14] 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.

Types remain the same throughout all stages, and individuals know their type. Preferences and abilities may vary across types.

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

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

[14]Type 1 individuals make up 33% of the population, while Type 2s make up 67%.

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).

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.

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 5. 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 are the same with or without unobserved heterogeneity. 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 2s 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 6 displays the maximum likelihood estimates for the parameters of the utility function.

The first three sections of Table 6 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.[15]

The level of diversification in high school also affects a student's choice of college major. Being more diversified (i.e. having a large diversification index) is better for business and communications majors than for math and physics majors.[16] Diversification has larger negative effects for quantitative majors (engineering and math and physics) than for other majors. This suggests that specialization in high school is particularly useful for potential STEM majors.

Females are more likely to enroll in education or health majors, and less likely to

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

[16] This effect is the same after controlling for unobserved heterogeneity.

enroll in quantitative majors.[17] There is a sizeable literature on college major choice and the gender gap,[18] which has documented differences in males' and females' college major choices that are in line with my findings. 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 quantitative 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. 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 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-school 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 in Tables 8, 9 and 10.

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 and quantitative courses than 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 11 compares actual data with the predictions of the model. I show two sets of parameter estimates from the model: one with unobserved heterogeneity, and one without.

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

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

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 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 (which implies more specialization in math and sciences). The second experiment is an increase in high school humanities course requirements, while the third simulation increases high school life sciences course requirements. The last simulation assumes that there is no specialization in high school.

Increasing the enrollment in STEM majors is of considerable interest for 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 earning substantially more than other college graduates, with the potential exception of business graduates (see Arcidiacono (2004), Pavan and Kinsler (2012), andArcidiacono, 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 are designed to 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 are only designed to illustrate how much of the current major choice is due to high school courses or specialization.

Table 12 shows that quantitative courses and specializations affect choices about 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, but it also decreases overall college enrollment. It is interesting to note that when I use the model without unobserved heterogeneity, one more high school quantitative course increases enrollment in STEM majors by five percentage points. In the model with unobserved heterogeneity, the enrolment in STEM majors only increases by four percentage points, which suggests a correction of the unobserved ability bias. An increase in one high school humanities course does not heavily decrease enrollment in STEM majors. One more life sciences course in high school increases enrollment in natural sciences majors by 0.015 percentage points and reduces enrollment in STEM majors by about the same amount.

Forcing every student to take the same courses (see simulation 4) also boosts enrollment. The share of students choosing STEM majors moves up by 18 to 20 percentage points. This suggests that high school specialization plays a key role in major choice.

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 STEM enrollment, but this policy is less feasible in practice, since there is likely a significant demand for a certain amount of choice by students and educators. Requiring high school students to take more quantitative courses is, therefore, the most appealing policy for increasing enrollment in STEM majors.

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 their choice of college major. I find that students usually choose a major in which they acquired more related skills in high school, suggesting that specialization occurs in high school. However, I find a U-shaped relationship between the diversity of courses taken in high school and college performance.

This result suggests that there is a trade-off 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 quantitative majors are preferred by specialized students. I also find that high school course selection plays an important role in determining college major choice.

More quantitative high school courses makes natural sciences, engineering and 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.

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.

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?,” *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.

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 (2012): "Human capital portfolios," Discussion paper.

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 1,265 individual that are used in the reduced-form analysis. To estimate the structural model, I reduce the sample to 1,112 to eliminate observations that are missing other variables used in certain estimations.

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.[19]

- 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, educa-

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

tion, and health. The criteria for aggregation is the degree of similarity in field topics. Here is a list of majors by category:

- 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.

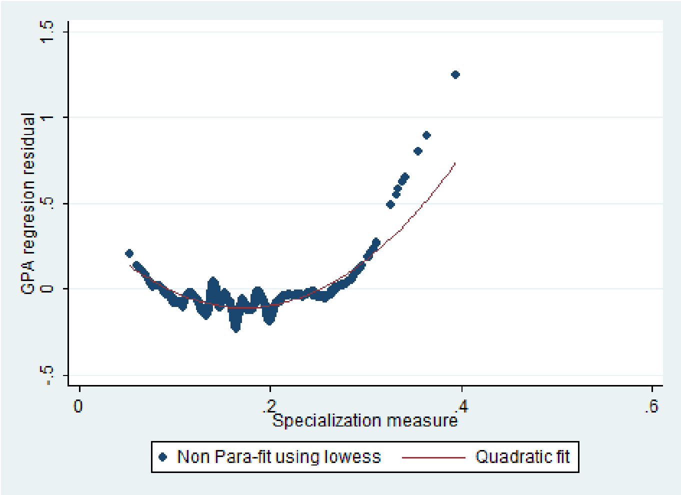

Figure 1: U-shaped relationship between GPA residual and ρ using non-parametric regression.

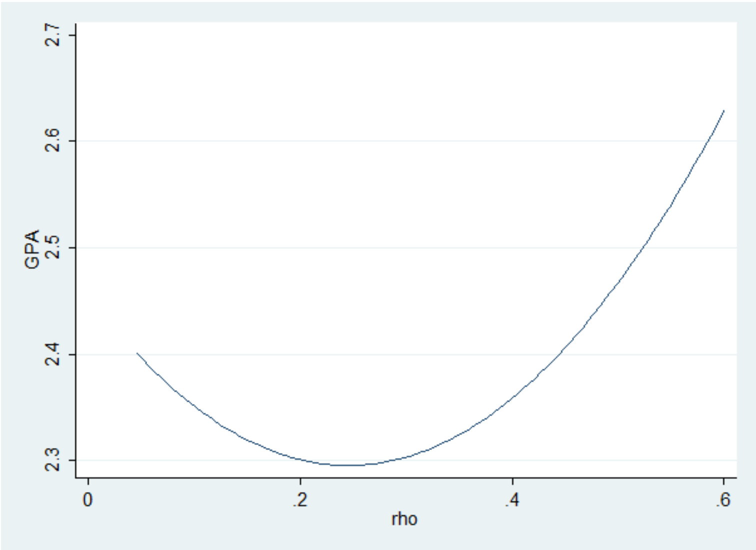

Figure 2: U-shaped relationship between GPA and ρ with quadratic fit.

Table 1: Summary statistics

	Unrestricted sample					Restricted sample				
	Mean	SD	SD in HS	Frac in HS	Obs.	Mean	SD	SD in HS	Frac. in HS	Obs.
Female	0.541	0.498	0.211	0.821	5072	0.530	0.499	0.266	0.716	1265
Black	0.125	0.331	0.210	0.597	5072	0.089	0.284	0.151	0.717	1265
SAT-Math	477.747	115.141	47.416	0.830	2064	483.075	110.838	43.172	0.848	1265
SAT-Verbal	440.612	107.143	43.714	0.834	2042	447.249	103.189	40.650	0.845	1265
College GPA	2.316	0.803	0.211	0.931	4686	2.453	0.689	0.190	0.924	1265
SES	0.224	0.738	0.412	0.688	4912	0.403	0.687	0.398	0.664	1265
HS Share of Courses					5072					
Reading and writing	0.232	0.066	0.045	0.530	5072	0.246	0.062	0.046	0.455	1265
Math	0.122	0.041	0.023	0.686	5072	0.132	0.037	0.022	0.640	1265
Life sciences	0.168	0.067	0.055	0.325	5072	0.168	0.066	0.059	0.221	1265
Physics	0.054	0.041	0.021	0.746	5072	0.065	0.041	0.022	0.718	1265
Humanities	0.186	0.074	0.065	0.235	5072	0.200	0.079	0.071	0.195	1265
Business and communications	0.074	0.065	0.031	0.768	5072	0.060	0.054	0.027	0.744	1265
Art	0.070	0.071	0.036	0.747	5072	0.059	0.066	0.035	0.709	1265
Other	0.050	0.058	0.044	0.422	5072	0.042	0.054	0.047	0.260	1265

NB: This table provides the mean and standard deviation of some variables in the full sample and in the restricted one. There is not a large difference between the two samples, 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. and W.	Life sci.	Hum.	Bus./Com.	Arts	Others
Bus. & comms.	0.169	0.236	0.167	0.190	**0.095**	0.063	0.081
Natural sciences	0.219	0.251	0.186	0.178	0.040	0.065	0.061
Math and physics	**0.225**	0.245	0.167	0.187	0.056	0.056	0.064
Education	0.165	0.232	0.169	0.180	0.075	**0.092**	0.088
Engineering	**0.227**	0.227	0.171	0.172	0.050	0.063	0.089
Social sci./hum.	0.185	**0.258**	0.163	**0.198**	0.056	0.066	0.074
Health	0.172	0.232	0.181	0.188	0.075	0.073	0.078
Other	0.169	0.228	0.170	0.176	0.066	**0.093**	0.098
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. 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: Estimation results for college performance (GPA as the dependent variable)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ρ	-6.834***	-5.222***	-5.824***	-5.766***	-5.861***	-5.923***	-7.154***
	(2.08)	(1.98)	(1.90)	(1.90)	(1.89)	(1.96)	(2.59)
ρ^2	17.477***	14.707***	16.128***	15.985***	16.195***	16.740***	21.376***
	(5.49)	(5.25)	(5.05)	(5.05)	(5.03)	(5.34)	(6.87)
Female		0.127***	0.179***	0.179***	0.178***	0.164***	0.072
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
Black		-0.281***	-0.141**	-0.136**	-0.141**	-0.114*	-0.047
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.09)
SES		0.039	-0.057**	-0.050	-0.054	-0.045	-0.051
		(0.03)	(0.03)	(0.05)	(0.05)	(0.05)	(0.06)
SAT-Math			0.109***	0.109***	0.111***	0.126***	0.096***
			(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
SAT-Verbal			0.127***	0.126***	0.126***	0.126***	0.144***
			(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Father's education				0.002	0.002	-0.000	-0.005
				(0.01)	(0.01)	(0.01)	(0.01)
Mother's education				-0.006	-0.005	-0.006	0.007
				(0.01)	(0.01)	(0.01)	(0.01)
Plan college Dad					0.034	0.034	0.093
					(0.08)	(0.08)	(0.08)
Plan college Mom					-0.037	-0.036	0.009
					(0.08)	(0.08)	(0.09)
Dummy for south					0.047	0.071	
					(0.04)	(0.04)	
Plan HS Dad						-0.018	-0.001
						(0.07)	(0.07)
Plan HS Mom						-0.134	-0.096
						(0.10)	(0.10)
Majors		Yes	Yes	Yes	Yes	Yes	Yes
High school courses						Yes	Yes
Constant	3.067***	2.591***	1.604***	1.612***	1.607***	1.646***	1.864***
	(0.19)	(0.18)	(0.20)	(0.21)	(0.22)	(0.24)	(0.37)
Observations	1265	1265	1265	1265	1265	1265	1265
R^2	0.01	0.11	0.20	0.20	0.20	0.21	0.19
Number of groups							389
R^2 overall							0.16

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 for column 1 to 6. Column 7 estimates ordinary least squares with a high school fixed effect. Background characteristics include parents' education and parents' participation in the college enrollment decision. High school courses are formal courses taken in high school and gathered from high school transcripts.

Table 4: Lind and Mehlum (2010) test for U-shape

Specification: $f(x) = x^2$		
Extreme Point: 0.1699134		
	H_1: U-shape vs. H_0: Monotone or inverse U-shape	
	Lower bound	Upper bound
Interval	.04625	.6006787
Slope	-4.066305	14.16445
t-value	-2.559242	3.20296
P>|t|	.0054755	.0007488
Overall test for presence of a U-shape:		
	t-value=2.56	
	P>|t|=0.0054	

Table 5: Performance regressions

	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
ρ	-7.6077	2.2671	-8.3000	2.0323
ρ^2	22.2215	6.0855	27.6217	5.4638
Female	0.1563	0.0483	0.1418	0.0433
Dummy South	-0.0105	0.0159	0.0014	0.0143
SAT-Math	0.1192	0.0279	0.1110	0.0250
SAT-Verbal	0.0859	0.0283	0.0929	0.0254
SES	-0.0326	0.0333	-0.0688	0.0300
Black	-0.0814	0.0748	-0.0604	0.0671
Type 1			-0.7128	0.0430
Const.	2.1672	0.3110	1.9749	0.2790
Variance	0.7225	0.0153	0.6476	0.0137

NB:Major-specific constant terms are included along with courses taken in high school.

Table 6: Utility parameter estimates

		One type		Two types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
Life sciences courses					
	Business and communications	0.1041	0.0979	0.1204	0.0665
	Natural sciences	0.2068	0.1301	0.1694	0.0996
	Math and physics	0.0511	0.1247	-0.0242	0.0813
	Education	0.1038	0.1316	0.1256	0.0705
	Engineering	0.0658	0.1217	-0.0147	0.0812
	Humanities	0.0648	0.0994	0.0764	0.0687
	Health	0.1603	0.1199	0.1600	0.0917
Quantitative courses					
	Business and communications	4.8964	1.0766	-0.0443	0.0879
	Natural sciences	5.0620	1.0837	0.0421	0.1277
	Math and physics	5.2565	1.0777	0.2146	0.1057
	Education	4.8938	1.0670	-0.0359	0.0945
	Engineering	5.2959	1.0805	0.2558	0.1058
	Humanities	4.8925	1.0742	-0.0770	0.0907
	Health	5.0410	1.0862	0.0662	0.1180
Humanities courses					
	Business and communications	0.1223	0.0763	0.1031	0.0369
	Natural science	0.0548	0.1064	-0.0086	0.0717
	Math and Physics	0.0597	0.1026	-0.0464	0.0499
	Education	0.0657	0.1123	0.0764	0.0417
	Engineering	0.0538	0.0984	-0.0413	0.0496
	Humanities	0.0793	0.0777	0.0600	0.0391
	Health	0.0385	0.0965	0.0118	0.0613
GPA					
	Business and communications	1.5436	0.9165	0.9756	0.3705
	Natural sciences	0.8338	1.1941	-0.4789	0.8508
	Math and physics	2.0673	1.1337	0.7189	0.4828
	Education	1.7819	1.2066	1.1058	0.3971
	Engineering	2.0723	1.1096	0.7165	0.4878
	Humanities	1.1457	0.9292	0.5134	0.4223
	Health	0.9436	1.1081	0.1823	0.6781
ρ					
	Business and communications	-12.9912	2.6876	-4.2077	2.8318
	Natural science	-17.4797	3.7017	-4.6484	3.5321
	Math and physics	-21.0977	3.3094	-14.1069	3.3499
	Education	-14.8482	3.1440	-4.5911	2.9716
	Engineering	-20.5266	3.2969	-13.5531	3.3480
	Humanities	-13.4522	2.7094	-4.4163	2.8372
	Health	-11.8757	3.3616	-1.4088	3.4043

Table 7: Utility parameter estimates (cont.)

		One type		Two types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
GPA × high school courses					
	Business and communications	-0.0279	0.0353	-0.0147	0.0147
	Natural sciences	-0.0145	0.0451	0.0274	0.0328
	Math and physics	-0.0487	0.0436	-0.0041	0.0192
	Education	-0.0305	0.0472	-0.0162	0.0163
	Engineering	-0.0561	0.0427	-0.0097	0.0194
	Humanities	-0.0162	0.0359	0.0018	0.0169
	Health	-0.0100	0.0421	0.0132	0.0273
SAT-Math					
	Business and communications	0.2981	0.1769	0.2984	0.1721
	Natural sciences	0.5880	0.2189	0.5464	0.2084
	Math and physics	0.7487	0.1986	0.7520	0.1914
	Education	0.1307	0.2003	0.2090	0.1798
	Engineering	0.6788	0.1981	0.7199	0.1916
	Humanities	0.1598	0.1771	0.2223	0.1719
	Health	0.1296	0.2111	0.2267	0.2015
SAT-Verbal					
	Business and communications	0.4005	0.1723	0.4078	0.1660
	Natural sciences	0.5737	0.2123	0.4268	0.1988
	Math and physics	0.1994	0.1945	0.1636	0.1848
	Education	0.3943	0.1949	0.4086	0.1734
	Engineering	0.2815	0.1937	0.1999	0.1845
	Humanities	0.6591	0.1729	0.5390	0.1659
	Health	0.3215	0.2049	0.3063	0.1963
Female					
	Business and communications	0.3113	0.2789	0.4923	0.2721
	Natural sciences	0.4354	0.3483	0.7970	0.3313
	Math and physics	-0.0143	0.3175	-0.0038	0.3067
	Education	1.0980	0.3605	0.9397	0.3023
	Engineering	-0.6547	0.3223	-0.3526	0.3075
	Humanities	0.4578	0.2800	0.5848	0.2726
	Health	1.3096	0.3554	1.2890	0.3391
Type 1					
	Business and communications			-0.4613	0.3148
	Natural sciences			-0.8189	0.3875
	Math and physics			-0.2081	0.3527
	Education			-0.4716	0.3296
	Engineering			-0.3901	0.3543
	Humanities			-0.4986	0.3154
	Health			-0.2955	0.3712
Nesting Parameter		0.4834	0.0111	0.2653	0.0088

Table 8: High school course choice estimations

	Humanities courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base-year test score				
Vocabulary	0.0610	0.0172	0.0568	0.0175
Reading	0.0104	0.0166	-0.0005	0.0168
Math	0.0031	0.0194	-0.0377	0.0193
Science	-0.0357	0.0171	-0.0402	0.0173
Writing	0.0587	0.0184	0.0410	0.0184
Civics	0.0137	0.0142	0.0193	0.0143
Expected GPA	2.8965	0.5722	4.4575	0.4860
Expected GPA interacted with major:				
Business and communications	-6.9553	0.8609	-4.3865	0.6309
Natural sciences	-13.1673	2.0744	-8.5054	1.5168
Math and physics	-9.1344	2.4134	-3.0334	0.9578
Education	-4.6856	1.4340	-5.5758	1.4804
Engineering	-6.4705	1.3250	-5.0364	0.9241
Humanities	-6.9276	0.9197	-5.7341	0.6865
Health	-8.4818	1.9966	-4.2283	1.0902
Major:				
Business and communications	17.7216	2.1358	11.1098	1.5040
Natural sciences	32.4560	5.1415	21.3923	3.7467
Math and physics	20.9101	5.7522	6.6542	2.5020
Education	11.1049	3.7968	12.6932	3.5860
Engineering	14.7227	3.2620	11.5303	2.2438
Humanities	17.5508	2.2610	14.6686	1.6314
Health	20.3698	4.7245	10.3561	2.5678
Type 1			0.6906	0.3993
Variance	3.3723	0.0712	3.4201	0.0723

Table 9: High school course choice estimations

	Life sciences courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base-year test score				
Vocabulary	-0.0311	0.0107	-0.0318	0.0108
Reading	-0.0124	0.0103	-0.0077	0.0104
Math	-0.0337	0.0121	-0.0247	0.0119
Science	0.0142	0.0107	0.0166	0.0107
Writing	-0.0223	0.0115	-0.0179	0.0114
Civics	0.0119	0.0089	0.0160	0.0088
Expected GPA	3.7839	0.3567	2.8776	0.2998
Expected GPA interacted with major:				
Business and communications	-0.3033	0.5366	-0.2213	0.3892
Natural sciences	0.6705	1.2927	-1.0865	0.9356
Math and physics	3.0931	1.5043	0.1327	0.5908
Education	-1.3200	0.8938	1.3837	0.9131
Engineering	-1.2856	0.8259	-0.4680	0.5700
Humanities	-0.9882	0.5733	-1.1866	0.4235
Health	-0.2333	1.2446	-2.3845	0.6725
Major:				
Business and communications	0.4262	1.3312	0.7125	0.9277
Natural sciences	-0.5294	3.2042	4.1990	2.3111
Math and physics	-7.0656	3.5854	-0.6851	1.5433
Education	2.7386	2.3666	-2.7487	2.2120
Engineering	3.1075	2.0332	1.4399	1.3841
Humanities	2.2836	1.4094	3.2095	1.0063
Health	1.2013	2.9451	6.5287	1.5839
Type 1			1.4862	0.2463
Variance	2.1020	0.0445	2.1096	0.0447

Table 10: High school course choice estimations

	Quantitative courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base-year test score				
Vocabulary	-0.0055	0.0084	-0.0031	0.0083
Reading	0.0112	0.0081	0.0068	0.0080
Math	0.0531	0.0095	0.0451	0.0092
Science	0.0289	0.0084	0.0263	0.0082
Writing	-0.0162	0.0090	-0.0208	0.0087
Civics	-0.0046	0.0070	-0.0051	0.0068
Expected GPA	0.8426	0.2795	0.9916	0.2301
Expected GPA interacted with major:				
Business and communications	-0.1068	0.4205	0.8343	0.2987
Natural sciences	1.3458	1.0131	0.9615	0.7181
Math and physics	-2.3925	1.1788	-0.2132	0.4534
Education	-1.0858	0.7004	-1.6626	0.7008
Engineering	1.2274	0.6471	0.6087	0.4375
Humanities	-0.0793	0.4492	0.6151	0.3250
Health	-0.3827	0.9752	1.0640	0.5161
Major:				
Business and communications	-0.1636	1.0431	-2.3126	0.7120
Natural sciences	-2.9809	2.5111	-1.6040	1.7737
Math and physics	5.2358	2.8096	1.2428	1.1844
Education	3.3402	1.8545	3.8587	1.6976
Engineering	-2.1985	1.5932	-0.3277	1.0622
Humanities	-0.0869	1.1043	-1.5347	0.7723
Health	0.9063	2.3076	-2.3652	1.2156
Type 1			1.1915	0.1890
Variance	1.6471	0.0349	1.6191	0.0343

Table 11: Comparing model predictions of high school course selection with the data

	Data	One type	Two types
		Quantitative	
Business and communications	5.5044	5.5068	5.4890
Natural sciences	6.7018	6.7304	6.6787
Education	6.7570	6.7757	6.7561
Math and physics	5.1628	5.1810	5.1848
Engineering	7.1130	7.1161	7.1001
Humanities	5.6754	5.6703	5.6627
Health	5.6582	5.6384	5.6624
		Humanities	
Business and communications	13.7609	13.9761	13.8079
Natural sciences	13.7368	13.5100	13.5951
Education	12.8505	12.8022	12.7235
Math and physics	12.5349	12.6713	12.5517
Engineering	12.3826	12.4544	12.4572
Humanities	14.0623	13.8463	13.9706
Health	13.4177	13.3346	13.4069
		Life sciences	
Business and communications	5.0612	5.0585	5.0410
Natural sciences	5.9298	5.9030	6.0344
Education	4.8879	4.9064	4.9184
Math and physics	5.0465	5.0528	5.0619
Engineering	4.8261	4.8231	4.7790
Humanities	4.9377	4.9399	4.9283
Health	5.5063	5.5292	5.4045

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 12: Simulations of the change in major choice distribution

		Simulations			
		(1)	(2)	(3)	(4)
One type	Math, phys. & eng. majors	0.036	-0.014	-0.015	0.186
	Natural sciences	0.011	0.000	0.015	0.000
	Humanities	-0.05	0.015	0.002	-0.127
	No college	0.003	0.001	-0.002	-0.059
Two types	Math, phys. & eng.	0.027	-0.017	-0.021	0.231
	Natural sciences	0.012	0.009	0.033	-0.039
	Humanities	-0.05	0.009	-0.01	-0.168
	No college	0.01	-0.001	0.0019	-0.023

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.

Recent Kent Discussion Papers in Economics

15/15: 'Regularized LIML for many instruments', Guy Tchuente

15/14: 'Agglomeration Economies and Productivity Growth: U.S. Cities, 1880-1930', Alexander Klein and Nicholas Crafts

15/13: 'Microcredit with Voluntary Contributions and Zero Interest Rate - Evidence from Pakistan', Mahreen Mahmud

15/12: 'Act Now: The Effects of the 2008 Spanish Disability Reform', Matthew J. Hill, Jose Silva and Judit Vall

15/11: 'Testing for Level Shifts in Fractionally Integrated Processes: a State Space Approach', Davide Delle Monache, Stefano Grassi and Paolo Santucci de Magistris

15/10: 'German Wage Moderation and European Imbalances: Feeding the Global VAR with Theory', Timo Bettendorf and Miguel A. León-Ledesma

15/09: 'Repaying Microcredit Loans: A Natural Experiment on Liability Structure', Mahreen Mahmud

15/08: 'Fundamental shock selection in DSGE models', Filippo Ferroni, Stefano Grassi and Miguel A. León-Ledesma

15/07: 'Direct calibration and comparison of agent-based herding models of financial markets', Sylvain Barde

15/06: 'Efficiency in a forced contribution threshold public good game', Edward Cartwright and Anna Stepanova

15/05: 'Military Aid, Direct Intervention and Counterterrorism', María D.C. García-Alonso, Paul Levine and Ron Smith

15/04: 'A Practical, Universal, Information Criterion over Nth Order Markov Processes', Sylvain Barde

15/03: 'Public Good Provision in Indian Rural Areas: the Returns to Collective Action by Microfinance Groups', Paolo Casini, Lore Vandewalle and Zaki Wahhaj

15/02: 'Fiscal multipliers in a two-sector search and matching model', Konstantinos Angelopoulos, Wei Jiang and James Malley

15/01: 'Military Aid, Direct Intervention and Counterterrorism', María D.C. García-Alonso, Paul Levine and Ron Smith

14/18: 'Applying a Macro-Finance Yield Curve to UK Quantitative Easing', Jagjit S. Chadha and Alex Waters

14/17: 'On the Interaction Between Economic Growth and Business Cycles', Ivan Mendieta-Muñoz