College attrition and the dynamics of information revelation *

Preliminary

Peter Arcidiacono † Esteban Aucejo‡ Arnaud Maurel§
Tyler Ransom ¶

October 2015

Abstract

This paper investigates the determinants of college attrition in a setting where individuals have imperfect information about their schooling ability and labor market productivity. We estimate a dynamic structural model of schooling and work decisions, where high school graduates choose a bundle of education and work combinations. We take into account the heterogeneity in schooling investments by distinguishing between two- and four-year colleges and graduate school, as well as science and non-science majors for four-year colleges. Individuals may also choose whether to work full-time, part-time, or not at all. A key feature of our approach is to account for correlated learning through college grades and wages, thus implying that individuals may leave or re-enter college as a result of the arrival of new information on their ability and

*We thank the attendees of the 2012 Conference in Honor of John Kennan, the 2012 Cowles Foundation Conference on Structural Microeconomics, the 2012 North American Summer Meeting of the Econometric Society (Northwestern), the 2014 SED conference (Toronto), the 2014 SITE Workshop on Empirical Implementation of Theoretical Models of Strategic Interaction and Dynamic Behavior (Stanford), the 2014 WEA Annual Meetings (Denver), the 2014 Swiss Leading House conference on the Economics of Study Choices, the 2015 Barcelona GSE Structural Microeconometrics Workshop, the 2015 Labour Workshop (Warwick), the 11^{th} World Congress of the Econometric Society (Montreal), as well as seminar participants at Arizona State, Carnegie Mellon, CREST, IDB, Toulouse School of Economics, University of Colorado-Boulder, University of North Carolina at Chapel Hill, University of Pennsylvania and Yale University for helpful comments and suggestions.

†Duke University and NBER. E-mail address: psarcidi@econ.duke.edu
‡London School of Economics. E-mail address: E.M.Aucejo@lse.ac.uk
§Duke University, IZA and NBER. E-mail address: apm16@duke.edu
¶Duke University. E-mail address:tyler.ransom@duke.edu

productivity. We use our results to quantify the importance of informational frictions in explaining the observed school-to-work transitions and to examine sorting patterns.

JEL Classification: C35; D83; J24

1 Introduction

About half of students entering college in the United States do not earn a bachelor's degree within five years, a proportion that has been increasing since the 1970's (Bound et al. 2010). To the extent that there is a large wage premium to receiving a four-year college degree (Heckman et al., 2006, Goldin & Katz, 2008, Bound & Turner, 2011), this suggests that imperfect information and learning may be important to the decision to leave college. In this paper, we focus on the role of learning about academic ability as well as labor market productivity as an explanation for the rate of college attrition and re-entry that is observed in the U.S. In the current environment where high college attrition rates are considered a major issue, addressing these issues is important to understand (*i*) whether these attrition rates *should* be a concern, and (*ii*) which type of policies would be effective in reducing attrition rates.

In order to quantify the importance of information frictions in the decision to leave or return to college, we estimate a dynamic model of schooling and work decisions where such decisions depend on the arrival of information on schooling ability and work productivity. A key feature of the model is that students have imperfect information about their ability and productivity. After graduating from high school or receiving a GED, individuals decide whether to attend college and/or work part-time or full-time, or engage in home production in each period. When entering college, individuals have some beliefs about their ability and productivity. At the end of each school year, they learn about their ability, using their grades to update their beliefs. Since schooling ability and productivity will in general be correlated, individuals will also use their grades to update their productivity belief. Likewise, employed individuals update both their productivity and ability beliefs after receiving a wage (see Miller 1984).

We estimate a richer model than previously possible by making use of recent innovations in the computation of dynamic models of correlated learning. James (2011) shows that (*i*) integrating out over actual abilities as opposed to the signals and (*ii*) using the EM algorithm where at the maximization step ability is treated as known, results in models that are computationally very fast. James (2011) builds on the results from Arcidiacono & Miller

(2011) to show that estimation is computationally simple even in the presence of unobserved heterogeneity that is known to the individual. Using this approach in our current context makes estimation of our correlated learning model both feasible and fast. Importantly, it also allows us to easily take into account heterogeneity in schooling investments by distinguishing between two- and four-year colleges, as well as science and non-science majors for four-year colleges.[1]

We then use our model estimates to quantify the importance of informational frictions in explaining the observed school-to-work transitions, and to evaluate the value of information in this context. We find that a sizable share of the dispersion in college grades and wages is accounted for by the ability components that are initially unknown to the individuals. Allowing for correlated learning is important in this context. Focusing on the ability components which are unknown to the individuals at the time of high school graduation, we find that schooling abilities are highly correlated across college types and majors (2-year college, 4-year college science, 4-year college non-science). The correlation between productivities in the unskilled and skilled sectors is also large. On the other hand, schooling abilities are only weakly correlated with productivity in both sectors, thus suggesting that grades earned in college actually reveal little information about future labor market performance.

We also simulate our model under a counterfactual scenario where all individuals would have perfect information on their abilities by the end of high school. We find that four-year college graduation rates would increase substantially (by 40%) relative to the baseline imperfect information scenario, mostly through a decrease in dropout rates. We further provide evidence that imperfect information on ability has important implications regarding the composition of college graduates, dropouts and stopouts in the different types of colleges and majors. Simulations reveal that ability sorting would be much stronger in the perfect information scenario. In particular, our results show that imperfect information on ability significantly limits the extent to which individuals can pursue their comparative advantage.

Our analysis builds on seminal research by Manski & Wise (1983) and Manski (1989), which argued that college entry can be seen as an experiment that may not lead to a college degree. According to these authors, an important determinant of college attrition lies in the fact that, after entering college, students get new information and thus learn about their ability. More recently, several other papers in the literature on college completion stress the importance of learning about schooling ability to account for college attrition (see, e.g., Altonji 1993, Light & Strayer 2000, Arcidiacono 2004, and Heckman & Urzua 2009). Of

[1] See the recent surveys by Altonji et al. (2015) and Altonji et al. (2012), who discuss the importance of heterogeneity in human capital investments.

particular relevance to us is the work by Stinebrickner & Stinebrickner (2012), who provide direct evidence, using subjective expectations data from Berea College, that learning about schooling ability is a major determinant of the college drop-out decision.

Much of the learning literature assumes that the labor market is an absorbing state, implying that the decision to leave college is irreversible (Stange 2012, Stinebrickner & Stinebrickner 2012). By relaxing this assumption, we are able to predict the substantial college re-entry rates of over 25% which are observed in the data.[2] This is an important step towards a comprehensive analysis of school-to-work transitions, building on the insights of Pugatch (2012) who provides evidence from South African data that the option to re-enroll in high school is a key determinant of the decision to leave school and enter the labor market.

The remainder of the paper is organized as follows. Section 2 presents the data. Section 3 describes a dynamic model of schooling and work decisions, where individuals have imperfect information about their schooling ability and labor market productivity, and update their beliefs through the observation of grades and wages. Section 4 discusses the identification of the model, with Section 5 detailing the estimation procedure. Section 6 presents our estimation results. Finally, Section 7 concludes.

2 Data

The model is estimated using data from the National Longitudinal Survey of Youth 1997 (NLSY97). The NLSY97 is a longitudinal, nationally representative survey of 8,984 American youth who were born between January 1, 1980 and December 31, 1984. Respondents were first interviewed in 1997 and have continued to be interviewed annually (for a total of 15 Rounds as of 2011, which corresponds to the most recent data used in the paper) on such topics as labor force activities, education, and marriage and fertility, among many others.

Of particular importance for our analysis is the choice variable, d_t, which is constructed at each period as follows:

1. Any individual attending a college in the month of October is classified as being in college for this year (either in a two- or a four-year college). For four-year colleges, our definition of "Science" majors includes majors in Sciences, Technology, Engineering, and Mathematics (STEM).[3] See Table A.1 for details on the exact majors in each

[2] In our sample, 26% of the individuals leaving college before graduation are observed to re-enroll at some point. Note that because of right-censoring, this underestimates the actual re-entry rate.

[3] Hereafter, we use "science" and "STEM" interchangeably.

category.

2. Any individual reporting college attendance who also reports working at least four weeks in October and at least 10 hours per week is classified as working part-time while in school, with full-time work requiring at least 35 hours per week and four weeks worked in October.

3. Any individual not in college (according to the criterion above) is classified as working part-time or full-time according to the criteria above.[4]

4. Finally, all other cases are classified as home production.[5]

The other dependent variables in the analysis are college GPA and wages. College GPA is measured on a four-point scale and calculated as the average GPA across all semesters in the calendar year. Wages are calculated as follows:

1. We compute the hourly compensation (i.e. wage plus tips and bonuses) for the self-reported main job, converted to 1996 dollars.

2. If a person does not report hourly compensation, we impute earnings as annual income divided by annual hours worked (7.2% of our sample).

3. Finally, we top- and bottom-code the resulting earnings distribution at the 99.5 percentile and 2.5 percentile.

It is worth noting that GPA and college major are missing (or reported as "don't know") quite frequently in our data. GPAs are missing for 28% of college students (50% of first-year students), and majors are missing for 25% of four-year college students (48% of first-year students). There is a high rate of missing data on these outcomes for two main reasons: (*i*) Dropout rates are highest in the first year of college, and students interviewed after dropping out are less likely to answer survey questions about their brief college tenure; and (*ii*) Interviews primarily occur between October and March when academic outcomes have yet to be realized.[6] These two effects result in a significant fraction of students reporting

[4]These criteria for labor force participation resemble those of Keane and Wolpin (1997).

[5]Following this criterion, any individual who is unemployed in October is classified in the home production sector. Our results do not appear to be sensitive to the inclusion of unemployment in the home production alternative.

[6]The modal interview months are November and December. Data are collected for the period of time between interviews, so if, for example, a respondent is interviewed in October 2004 and again in November 2005, college information related to the Fall 2004 and Spring 2005 semesters would both be reported in the November 2005 interview. This lag in reporting is also likely to contribute to the missing data problem.

college attendance but not answering college-specific questions. We address this missing data issue as discussed in Section 5.

We estimate the model on males who have graduated high school and have valid Armed Services Vocational Aptitude Battery (ASVAB) test scores. As measures of academic preparation, we use the SAT score where observed.[7] If an individual did not take the SAT, we predict the individual's SAT score using each of his ASVAB component test scores.[8] This predicted SAT score is used synonymously with actual SAT score throughout. We drop all current and future observations for any respondents missing wage observations while choosing a work activity, or missing GPA or college major while enrolled in college. Our final estimation subsample includes 20,882 person-year observations for 2,712 males. Table A.2 in Appendix section A gives further details on our sample selection.

Tables 1 through 10 present some descriptive statistics for our subsample, by college enrollment, major and completion status. Table 1 shows that individuals who attend college at some point and start at a four-year institution have, on average, higher SAT test scores, with science majors having higher scores than other majors. The proportion of blacks and Hispanics is also lower among four-year college attendees, with white males disproportionately choosing science majors. Conversely, it is worth noting that those starting at a two-year college tend to have a lower SAT score, and disproportionately come from minorities. Overall, this difference in composition between two and four-year colleges (and between majors in four-year colleges) stresses the need to distinguish between college and major type when modeling college enrollment decisions.

Table 2 reports the mean GPA (on a four-point scale) by type of college attended, major and period of enrollment.[9] Looking at the individuals enrolled in a four-year college with a science major, the evolution of the GPA provides clear evidence of selection over time. Individuals who leave college or switch to a two-year institution or other type of major tend to have lower GPA than those who stay enrolled in a four-year college science major. We find a similar pattern for two-year college enrollees, with the GPA being on average lower for these students than for either type of four-year college enrollees. Overall, these descriptive findings are consistent with two stories, which may not be mutually exclusive: (*i*) individuals decide to leave or switch college/major as they learn about their schooling ability, or (*ii*)

[7]The distribution of raw SAT scores is standardized to be zero-mean and standard deviation 1 for the NLSY97 population who took the SAT.

[8]Each ASVAB component test score is standardized to be zero-mean and standard deviation 1 for the NLSY97 population. The sub-tests used are Arithmetic Reasoning, Mathematical Knowledge, Numerical Operations, Paragraph Comprehension and Word Knowledge.

[9]Note that these periods of college enrollment may not be consecutive.

those who leave or switch college/major tend to have a lower ability, that they observe perfectly even before starting college. Telling apart these two explanations is a key objective of our structural estimation, which will be discussed in the following section.

Table 3 lists the frequencies of continuous enrollment until graduation (either in four or two-year colleges), stopping out (i.e. leaving college before graduation and returning to school at some point) and dropping out (i.e. permanently leaving college, before four-year graduation) in the NLSY97 full sample, our estimation subsample, and type of college/major first enrolled in. Our subsample slightly understates dropping out because we discarded observations in right-censored missing interview spells, and missing an interview is positively correlated with dropping out of college. Also evident from Table 3 is the fact that dropping out and stopping out are more common in two-year colleges than four-year colleges. Four-year science majors have the lowest proportions of dropping out and stopping out. This again points at the need to distinguish between these two types of colleges and majors in our model. Due to the ongoing nature of the survey and the fact that some respondents are still in college, Table 4 aims to identify the lower bound of the stopout rate. For example, of those who had graduated with a four-year college degree by round 15 of the survey, 17.4% were stopouts. For those beginning college in a four-year university science major, this number is 11.0%, compared with 11.3% for humanities majors. For those originating in a two-year college, the figure is 40.7%.

Table 5 shows that those who continuously complete college have higher SAT scores, higher high school GPA, and come from families with higher income and mothers who are more educated. It is also interesting to note that stopouts on average straddle the continuous completion and dropout categories. This highlights the importance of studying stopping out as a third category of college completion. The descriptive evidence presented in Table 5 also points to the fact that family background variables are important to include in an analysis of college completion.

Table 6 breaks out Table 2 by college completion status. Similar to Table 2, there is evidence of selection over time and over (eventual) completion status. This further supports the idea that those who leave college may do so because of a bad signal on ability in the form of low grades. To illustrate this more fully, Table 7 presents differences between actual and expected grades in period t of college broken out by $t+1$ enrollment decision, where expected grades are taken as a function of family background, race, SAT scores, work status, and age. Interestingly, this shows that the selection patterns discussed above still hold after controlling for this set of observed characteristics.

Table 8 shows the origin and destination majors and college types for those who stop out. Stopouts are most likely to return to their original major or college type, but they are also likely to move to four-year humanities or two-year college upon returning to school.

Table 9 further describes the evolution of GPA over time by in-school work status. For both major types in four-year colleges, average GPA is roughly decreasing in work intensity, and increasing over time within each work intensity category. For two-year colleges, GPA is decreasing in work intensity only in the first period. By periods 3 and 4, the opposite is true. This illustrates a substitution effect between school and work intensity—those working hardest take longer than two years to complete a two-year degree.

Finally, to illustrate learning on wages as a reason for stopouts to return to college, Table 10 lists the difference between actual and expected log wages for those who have stopped out, broken out by next-period decision. Those who have left college for the labor force and choose to return to school have 5% lower wages on average the year before returning to school, even when controlling for a rich set of individual, family background, and schooling and labor force experience characteristics. This provides suggestive evidence that learning on wages contributes to the decision to return to college.

3 The model

3.1 Overview

After graduating from high school, individuals in each period make a joint schooling and work decision. For those who have not graduated from a four-year college, their schooling options include whether to attend a two-year institution, a four-year institution as a science major, or a four-year institution as a non-science major. After graduating from a four-year college, the schooling option includes whether or not to enroll in graduate school.

In addition to choosing among the different schooling options, individuals also choose whether to work full-time, part-time, or not at all. All three of these decisions are available regardless of their schooling choice.[10] Working while in college may be detrimental to academic performance (see, e.g., Stinebrickner & Stinebrickner, 2003) but is also likely to be a channel through which individuals learn about their productivity. Our framework incorporates this tradeoff.

Individuals only have imperfect information about (*i*) their schooling ability and (*ii*)

[10]See also Joensen (2009) who estimates a dynamic structural model of schooling and work decisions allowing for work while in college.

their labor market productivity. If they attend college, they learn about their ability by observing their schooling performance, as measured by their Grade Point Average (GPA) at the end of the academic year. The gap between the observed and predicted GPA then provides a noisy signal for their ability, which is used to update their belief in a Bayesian fashion. Since schooling ability and productivity will in general be correlated, the GPA also provides some information about labor market productivity. Individuals will therefore use their GPA to update their productivity belief. Similarly, those who participate in the labor market update their beliefs about their labor market productivity as well as their beliefs about their schooling ability after receiving their wage.

Individuals are forward-looking and choose the sequence of actions yielding the highest value of expected lifetime utility. Hence, when making their schooling and labor market decisions, individuals take into account the option value associated with the new information acquired on different choice paths. Individuals who choose to work while in college will get two signals, through their GPA and their wage, on their ability and productivity. Interestingly, without the need to invoke a credit constraint argument, the value of information implies that working while in college may be optimal for some students in spite of a detrimental impact on academic performance.

We now detail the main components of the model, namely the grade and wage equations, together with the learning process and the flow utility functions for each alternative.

3.2 Grades

We denote by $j \in \{a, bs, bn\}$ the type of college and major attended, where a (for *Associate*) denotes a two-year college, bs (for *Bachelor*, Science) a four-year college Science major, and bn (for *Bachelor*, non-Science) a four-year college non-Science major. We assume that grades depend on A_{ij} where A_{ij} is the unobserved schooling ability about which individuals have some beliefs initially given by the prior distribution $\mathcal{N}(0, \sigma^2_{Aj})$. Grades also depend on a set of covariates, X_{ict}, that are known to the individual and include observed ability measures and past decisions.

Denoting by t calendar time and τ the period of college enrollment, grades in two-year colleges and in the first two years of four-year colleges are given by:

$$G_{ij\tau} \quad = \quad \gamma_{0j} + X_{ict}\gamma_{1j} + A_{ij} + \varepsilon_{ij\tau}$$

The idiosyncratic shocks $\varepsilon_{ij\tau}$ are distributed $\mathcal{N}(0, \sigma^2_{j\tau})$ and are independent from the other state variables. Define the type-j (college, major) academic index of i at time t, AI_{ijt}, as:

$$AI_{ijt} = \gamma_{0j} + X_{ict}\gamma_{1j} + A_{ij}$$

The academic index AI_{ijt} gives expected grades conditional on knowing A_{ij} but not the idiosyncratic shock $\varepsilon_{ij\tau}$.

For four-year colleges and periods $\tau > 2$, we express grades relative to AI_{ijt} as follows:[11]

$$G_{ij\tau} \quad = \quad \lambda_{0j} + \lambda_{1j} AI_{ijt} + \varepsilon_{ij\tau}$$

Hence, the return to the academic index varies over period of college enrollment and across majors. In particular, consistent with Hansen et al. (2004), our specification allows the effect of latent ability on grades to vary with the number of years spent in college.

3.3 A two-sector labor market

Individuals who choose one of the work options (either full-time or part-time) receive an hourly wage that depends on graduation status. We assume that there are two sectors in the labor market, which are referred to in the following as *skilled* (four-year college graduates and individuals with a graduate school degree) and *unskilled* (all other labor market participants, including high school graduates or GED recipients, college dropouts and stopouts, as well as two-year college graduates).

Wages in sector l depend on productivity A_{il}, a set of observed characteristics X_{ilt}, time dummies δ_{lt}, and idiosyncratic shocks ε_{ilt}:

$$\ln(W_{ilt}) \quad = \quad \delta_{lt} + X_{ilt}\gamma_{1l} + A_{il} + \varepsilon_{ilt}$$

We account for nonstationarity in wages by including calendar year dummies, δ_{lt}, thus incorporating business cycle effects. The time dummies at t are observed in period t but individuals must form expectations over this variable for periods $t+1$ and beyond. We will come back to this in Section 3.6.2. The idiosyncratic shocks, ε_{lt}, are assumed to be distributed $\mathcal{N}(0, \sigma_l^2)$ and are independent over time and independent of the other state variables.

3.4 Flow utilities

We denote in the following by $d_{it} = (j, k)$ the choice for individual i at time t over school, where $j \in \{a, bs, bn, 0\}$ (respectively $j \in \{gs, 0\}$) before (resp. after) graduation from a four-

[11] See Arcidiacono (2004) for a similar ability index specification.

year college, and work $k \in \{p, f, 0\}$, where gs refers to graduate school, and p and f refer to part-time and full-time work. The choice $d_{it} = (0, 0)$ then indicates the home production option: no work and no school.

Up to an intercept term and an idiosyncratic preference shock, we assume that the utility of the choice (j, k) is additively separable. Let Z_{1it} denote variables that affect the utility of school and Z_{2it} denote the variables that affect the utility of working. The flow payoff for choice (j, k) is given by:

$$U_{ijkt}(Z_{it}, \varepsilon_{ijk}) = \alpha_{jk} + Z_{1it}\alpha_j + Z_{2it}\alpha_k + \varepsilon_{ijkt} \tag{3.1}$$

$$= u_{jk}(Z_{it}) + \varepsilon_{ijkt} \tag{3.2}$$

where Z_{it} includes characteristics such as SAT scores (Math and Verbal), race, and previous choice. Controlling for the previous choice allows for switching costs, in a similar spirit as in Keane & Wolpin (1997). The idiosyncratic preference shocks ε_{ijkt} are assumed to follow a (standard) Type-I extreme value distribution. Embedded in Z_{1it} are the expected abilities in sector j. The flow payoff for graduate school depends on the expected ability in four-year college corresponding to the college major the individual has graduated from. Embedded in Z_{2it} are expected log-wages in sector k.[12]

Finally, the home production sector ($d_{it} = (0, 0)$) is chosen as a reference alternative, and we normalize the corresponding flow utility to zero. The flow utility parameters therefore need to be interpreted relative to this alternative.

3.5 The optimization problem

Individuals are forward-looking, choosing the sequence of college enrollment and labor market participation decisions yielding the highest present value of expected lifetime utility. The individual chooses d_{it} to sequentially maximize the discounted sum of payoffs:

$$E\left[\sum_{t=1}^{T} \beta^{t-1} \sum_{j} \sum_{k} (u_{jkt}(Z_{it}) + \varepsilon_{ijkt}) 1\{d_{it} = (j, k)\}\right]$$

where $\beta \in (0, 1)$ is the discount factor. The expectation is taken with respect to the distribution of the future idiosyncratic shocks as well as the signals associated with the different choice paths.

[12]Both of these covariates will vary given the choices. However, to conserve on notation we do not put jk subscripts on the Z's.

Let $V_t(Z_{it})$ denote the ex ante value function at the beginning of period t, the expected discounted sum of current and future payoffs just before ε_t is revealed. The conditional value function v_{ijkt} is given by:

$$v_{jkt}(Z_{it}) = u_{jkt}(Z_{it}) + \beta E_t(V_{t+1}(Z_{t+1})|Z_{it}, d_{it} = (j,k))$$

Given the assumption that the ε's are i.i.d. Type 1 extreme value,

$$v_{jkt}(Z_{it}) = u_{jkt}(Z_{it}) + \beta E_t \left[\ln \left(\sum_j \sum_k \exp(v_{jkt+1}(Z_{it+1})) \right) \middle| Z_{it}, d_{it} = (j,k) \right] + \beta\gamma$$

where γ denotes Euler's constant.

3.6 Beliefs

Individuals are uncertain about (*i*) their future preference shocks, (*ii*) their schooling ability and labor market productivity, and (*iii*) the evolution of the market shocks (the δ_{lt}'s). The first components, which we have discussed as expectations over future preference shocks, are encompassed in the ex ante value function. We next describe beliefs over abilities and productivities as well as the market shocks.

3.6.1 Beliefs over schooling ability and labor market productivity

We denote A_i as the five dimensional ability vector, $A_i \equiv (A_{ia}, A_{ibs}, A_{ibn}, A_{is}, A_{iu})'$ (simply referred to as *ability* in the following). Individuals update their beliefs in a Bayesian fashion. Their initial ability beliefs are given by the population distribution of A, which is supposed to be multivariate normal with mean zero and covariance matrix Δ. Importantly, we do not restrict Δ to be diagonal, thus allowing for correlated learning across the five different ability components.[13]

Namely, at each period τ of college attendance, individuals use their GPA to update their belief about their schooling ability in all college options $(A_{ia}, A_{ibs}, A_{ibn})$, as well as their labor market productivity in both sectors (A_{is}, A_{iu}). The GPA provides a noisy signal for their ability, which is denoted by $S_{ij\tau}$ for type-j college option and period of enrollment τ. For two-year colleges and the first two years of four-year colleges, the signal is given by:

$$S_{ij\tau} \quad = \quad G_{ij\tau} - \gamma_{0j} - X_{ict}\gamma_{1j}$$

[13]See also Antonovics & Golan (2012), James (2011) and Sanders (2010) who estimate occupational choice models with correlated learning.

For four-year colleges and subsequent periods ($\tau > 2$), the index specification yields:

$$S_{ij\tau} = \frac{G_{ij\tau} - \lambda_{0j} - \lambda_{1j}(\gamma_{0j} + X_{ict}\gamma_{1j})}{\lambda_{1j}}$$

Similarly, individuals who participate to the labor market update their ability beliefs after receiving their wages. The signal is given by, for sector l and period t:

$$S_{ilt} = \ln(W_{ilt}) - \gamma_{0l} - X_{ilt}\gamma_{1l}$$

Finally, individuals may choose to work while in college, in which case they will receive two ability signals $(S_{ij\tau}, S_{ilt})$.

It follows from the normality assumptions on the initial prior ability distribution and on the idiosyncratic shocks that the posterior ability distributions are also normally distributed. Specifically, denoting by $E_t(A_i)$ and $\Sigma_t(A_i)$ the posterior ability mean and covariance at the end of period t, we have (see DeGroot, 1970):

$$E_t(A_i) = (\Sigma_{t-1}^{-1}(A_i) + \Omega_{it})^{-1}(\Sigma_{t-1}^{-1}(A_i)E_{t-1}(A_i) + \Omega_{it}\widetilde{S}_{it}) \quad (3.3)$$

$$\Sigma_t(A_i) = (\Sigma_{t-1}^{-1}(A_i) + \Omega_{it})^{-1} \quad (3.4)$$

where Ω_{it} is a (5×5) matrix with zeros everywhere except for the diagonal terms corresponding to the occupations of the individual in period t (namely two-year college, four-year college Science major, four-year college non-Science major, skilled or unskilled labor), which are given by the inverse of the idiosyncratic shock variances (multiplied by λ_{1j}^2 for four-year colleges in junior and senior years). $\widetilde{S}_{it}$ a (5×1) vector with zeros everywhere except for the elements corresponding to the occupations of the individual in period t, which are given by the ability signals received in this period. Individuals then integrate out over the possible signals they could receive for each possible decision.

3.6.2 Beliefs over market shocks

We finally need to specify how individuals form beliefs about the labor market. Individuals observe the current values of δ_{st} and δ_{ut}. We assume that the process governing the aggregate sector-specific shocks δ_t's is an AR(1):

$$\delta_{lt} = \phi_{0l} + \phi_1\delta_{lt-1} + \zeta_{lt} \quad (3.5)$$

where the ζ_{lt} are assumed to be distributed N(0,σ_ζ). The assumption that the aggregate shocks follow an AR(1) process, or a discretized version of it (Markov process of order 1) is very common in the literature (see, e.g., Adda et al., 2010, Robin, 2011). Given the realizations of the δ_{lt}'s, individuals then integrate over possible realizations of the ζ_{lt}'s when forming their expectations over the future.

4 Identification

Before turning to the estimation procedure, we discuss below the identification of the model.[14] As is common for these types of dynamic discrete choice models (see, e.g., Rust,1994, Magnac & Thesmar, 2002, and Arcidiacono & Miller, 2013), identification of the flow utility parameters hinges on the distributional assumptions imposed on the idiosyncratic shocks, the normalization of the home production utility and the discount factor β, which is set equal to 0.9 in the following.

Let us consider the identification of the outcome equations (grades and log-wages). The GPA $G_{ij\tau}$ is only observed for the individuals who are enrolled in a type-j (college, major) in their τ-th period of college enrollment. To the extent that college enrollment decisions depend on the ability (A_i), this raises a selection issue. We show the identification of the grade equation parameters by using, for each period τ, the prior ability at the beginning of the period ($E_{t-1}(A_{ij})$) as a control function in the grade equation (see Navarro, 2008, for an insightful review of the control function approach). Specifically, we consider the following augmented regression for $j \in \{bs, bn\}$ and $\tau > 2$:

$$G_{ij\tau} \quad = \quad \lambda_{0j} + \lambda_{1j}(\gamma_{0j} + X_{ict}\gamma_{1j}) + \lambda_{1j}E_{t-1}(A_{ij}) + \nu_{ij\tau}$$

where it follows from the Bayesian updating rule (see Equation (3.3), p.14) that $E_{t-1}(A_{ij})$ can be expressed as a weighted sum of all the past ability signals. Under the key assumption, consistent with the specification of the flow utilities in Subsection 3.4, that college enrollment decisions only depend on ability through the ability beliefs, application of ordinary least squares to this equation identifies the parameters $(\lambda_{0j}, \lambda_{1j})$, with the ability index coefficients $(\gamma_{0j}, \gamma_{1j})$ being identified from the first and second period grades.

Identification of the ability index coefficients also follows from the assumption that enrollment decisions only depend on ability through the past ability signals. Specifically, grades in the first two years of four-year college as well as in two-year colleges can be expressed as

[14] For the sake of exposition, we first consider the case of the model without type-specific unobserved heterogeneity, before discussing the identification of the unobserved heterogeneity parameters.

follows:

$$G_{ij\tau} = \gamma_{0j} + X_{ict}\gamma_{1j} + E_{t-1}(A_{ij}) + \nu_{ij\tau}$$

Application of ordinary least squares therefore directly identifies $(\gamma_{0j}, \gamma_{1j})$. Similar arguments can be used for the identification of the log-wage equations in each sector.

Finally, the signal-to-noise ratios as well as the ability covariance matrix Δ are identified from the past ability signal coefficients. Of particular interest here are the correlations between the different ability components, which are identified from individuals switching occupations.

In our specification with latent heterogeneity types, one also needs to tell apart the type-specific unobserved (to the econometrician only) heterogeneity components from the ability beliefs. Initial choices made before the individuals learn about their ability are a key source of identification. For instance, low-SAT individuals who choose to enroll in a four-year college right after high school graduation would be predicted to have high unobserved preference for four-year college. Besides, low-SAT individuals who are enrolled in college may decide to leave college after receiving a high GPA. Individuals exhibiting these types of behavior would be predicted to have a high type-specific schooling ability.

5 Estimation

We first detail the estimation procedure for the specification without type-specific unobserved heterogeneity. Assuming that the idiosyncratic shocks are mutually and serially uncorrelated, estimation proceeds in two stages, which consists of (*i*) estimation of the grade and log-wage equations and (*ii*) estimation of the flow utility parameters. The validity of this sequential approach follows from the key assumption that choices only depend on ability through the (observed) sequence of signals. This results in the likelihood being separable in the outcome and choice contributions.

5.1 Additive separability

Specifically, we consider the case of an individual i attending college during T_c periods, who participate to the unskilled (resp. skilled) labor market during T_u (resp. T_s) periods and for whom we observe a sequence of T_d decisions. We write the individual contributions to the likelihood of the grades, log-wages and choices by integrating out the unobserved ability terms $A = (A_a, A_{bs}, A_{bn}, A_s, A_u)'$, which breaks down the dependence across the grades, log-wages, choices and between all of these variables. The contribution to the likelihood writes, denoting by G_i the grades, w_{iu} (resp. w_{is}) the unskilled (resp. skilled) log-wages and d_i the decisions, as a five-dimensional integral:

$$\begin{aligned} & l(d_{i1}, \ldots, d_{iT_d}, G_{i1}, \ldots, G_{iT_c}, w_{iu1}, \ldots, w_{iuT_u}, w_{is1}, \ldots, w_{isT_s}) \\ = & \int l(d_{i1}, \ldots, d_{iT_d}, G_{i1}, \ldots, G_{iT_c}, w_{iu1}, \ldots, w_{iuT_u}, w_{is1}, \ldots, w_{isT_s}|A) l(A) dA \end{aligned}$$

where $l(A)$ is the pdf. of the ability distribution $\mathcal{N}(0, \Delta)$.

From the law of successive conditioning, and using the fact that choices depend on A only through the signals, we obtain the following partially separable expression:

$$l(d_{i1}, \ldots, d_{iT_d}, G_{i1}, \ldots, G_{iT_c}, w_{iu1}, \ldots, w_{iuT_u}, w_{is1}, \ldots, w_{isT_s}) = L_{d_i} \times L_{G_i, w_{iu}, w_{is}}$$

Where the contribution of the sequence of decisions is given by:

$$L_{di} = l(d_{i1}) l(d_{i2}|d_{i1}, G_{i1}) \ldots l(d_{iT_d}|d_{i1}, d_{i2}, \ldots, d_{iT_d-1}, G_{i1}, G_{i2}, \ldots w_{iu1}, w_{iu2}, \ldots, w_{is1}, w_{is2}, \ldots)$$

This simply corresponds to the product over T_d periods of the type-1 extreme value choice probabilities obtained from the dynamic discrete choice model.

The contribution of the observed sequence of grades, unskilled and skilled log-wages is given by:

$$\begin{aligned} L_{G_i, w_{iu}, w_{is}} = & \int l(G_{i1}|d_{i1}, A) \ldots l(G_{iT_c}|d_{i1}, d_{i2}, \ldots, A) l(w_{iu1}|d_{i1}, A) \ldots l(w_{iuT_u}|d_{i1}, d_{i2}, \ldots, A) \\ & \times l(w_{is1}|d_{i1}, A) \ldots l(w_{isT_s}|d_{i1}, d_{i2}, \ldots, A) l(A) dA \end{aligned}$$

Where $l(w_{iut}|d_{i1}, \ldots, A)$, $l(w_{ist}|d_{i1}, \ldots, A)$, and $l(G_{it}|A)$ are Gaussian pdf of respectively, the unskilled, skilled log-wage and GPA distributions.

5.2 Estimation of grade and wage parameters

Estimation of the parameters of the outcome equations proceeds as follows. Instead of directly maximizing the likelihood of the outcomes, which would be computationally costly because of the ability integration, we compute the parameter estimates using the EM algorithm (Dempster et al., 1977). The estimation procedure iterates over the following two steps, until convergence:

- E-step: update the posterior ability distribution from all the observed outcome data (log-wages and grades), using the outcome equation parameters obtained from the previous iteration. This follows from the Bayesian updating formulas (3.3)-(3.4), for the posterior ability mean and covariance, given in Section 3.6.1. The (population) variance of the ability distribution is then updated as follows, for each iteration k of the EM estimation:

$$\Delta^k = \frac{1}{N}\sum_{i=1}^{N}\left(\Sigma_i^k(A) + E_i^k(A)E_i^k(A)'\right)$$

  where N denotes the number of individuals in the sample, $E_i^k(A)$ the posterior ability mean ($E_i^k(A)'$ its transposed) and $\Sigma_i^k(A)$ the posterior ability covariance computed at the beginning of the E-step.

- M-step: given the posterior ability distribution obtained at the E-step, maximize the expected complete log-likelihood of the outcome data, which is separable across sectors (two-year college, four-year college Science major, four-year college non-Science major, skilled or unskilled labor).

Namely, at the M-step of each iteration k of the EM estimation, denoting by $l_{ik}(A)$ the posterior ability distribution computed at the E-step, we maximize the expected complete log-likelihood El_{ik}:

$$\begin{aligned} El_{ik} &= \int \ln(l(G_{i1}|d_{i1}, A)\ldots l(G_{iT_c}|d_{i1}, d_{i2}, \ldots, A)l(w_{iu1}|d_{i1}, A)\ldots l(w_{uiT_u}|d_{i1}, d_{i2}, \ldots, A))l_{ik}(A)dA \\ &= El_{ik,a} + El_{ik,bs} + El_{ik,bn} + El_{ik,s} + El_{ik,u} \end{aligned}$$

For instance, the parameters of the unskilled wage equation are updated by maximizing the contribution $El_{ik,u}$, which writes, denoting by $l_{ik}(A_u)$ the marginal posterior distribution of A_u:

$$El_{ik,u} = \int(\ln(l(w_{iu1}|d_{i1}, A_u)) + \ldots + \ln(l(w_{iuT_u}|d_{i1}, d_{i2}, \ldots, A_u)))l_{ik}(A_u)dA_u$$

Note that this term is additively separable over time. For any given period τ of unskilled labor market participation, it follows from the normality assumptions on the idiosyncratic productivity shocks and the unobserved ability that:

$$\begin{aligned} \int \ln(l(w_{iu\tau}|d_{i1}, d_{i2}, \ldots, A_u))l_{ik}(A_u)dA_u &= \\ -\frac{1}{2}\ln(2\pi\sigma_u^2) - \frac{1}{2\sigma_u^2}\left(\Sigma_{iuu}^k(A) + (w_{iu\tau} - X_{iut}\gamma_{1u} - \delta_{ut} - E_{iu}^k(A))^2\right) \end{aligned}$$

where t refers to calendar time (which should be understood as individual-specific here), $E^k_{iu}(A)$ and $\Sigma^k_{iuu}(A)$ denote respectively the posterior mean and variance of the ability in the unskilled sector (computed at the E-step). This equality implies that the wage equation parameters $(\gamma_{1u}, \delta_{ut})$ can be simply updated by regressing (via OLS) the log-wages in the unskilled sector on the set of observed characteristics, calendar time dummies, and the posterior (unskilled) ability mean which plays the role of a selection correction term. The idiosyncratic shock variance (σ^2_u) is then updated as follows:

$$\sigma^2_{u,k+1} = \frac{\sum_{i,\tau}\left(\Sigma^k_{iuu}(A) + (w_{iu\tau} - X_{iut}\gamma_{1u} - \delta_{ut} - E^k_{iu}(A))^2\right)}{N^{\text{obs}}_u}$$

where N^{obs}_u is the total number of wage observations in the unskilled sector. Skilled wage equation parameters are updated similarly.

The updating rule above needs to be adjusted to account for the ability index specification of the grade equations along with the time-varying variances of the idiosyncratic shocks. For instance, for four-year colleges (period of enrollment $\tau > 2$), the contribution to the log-likelihood writes:

$$\int \ln(l(G_{ij\tau}|d_{i1}, d_{i2}, \ldots, A_j))l_{ik}(A_j)dA_j = -\frac{1}{2}\ln(2\pi\sigma^2_{j\tau}) - \frac{1}{2\sigma^2_{j\tau}}\left(\lambda^2_{1j}\Sigma^k_{ijj}(A) + (G_{ij\tau} - \lambda_{0j} - \lambda_{1j}AI^k_{ijt})^2\right)$$

where $j \in \{bs, bn\}$, $\Sigma^k_{ijj}(A)$ denotes the posterior variance of the college-j ability (computed at the E-step), and $AI^k_{ijt} = \gamma_{0j} + X_{ict}\gamma_{1j} + E^k_{ij}(A)$ is the posterior mean of the ability index in college j. It follows that the parameters $(\gamma_{0j}, \gamma_{1j}, \lambda_{0j}, \lambda_{1j}, (\sigma^2_{j\tau})_\tau)$ are updated by solving the following minimization problem:

$$\min \sum_{i,\tau}\left(\ln(\sigma^2_{j\tau}) + \frac{1}{\sigma^2_{j\tau}}\left(\lambda^2_{1j\tau}\Sigma^k_{ijj}(A) + (G_{ij\tau} - \lambda_{0j\tau} - \lambda_{1j\tau}AI^k_{ijt})^2\right)\right)$$

where $(\lambda_{0j\tau}, \lambda_{1j\tau}) = (0, 1)$ for $\tau \leq 2$, and $(\lambda_{0j\tau}, \lambda_{1j\tau}) = (\lambda_{0j}, \lambda_{1j})$ otherwise.

5.3 Estimation of the flow payoffs

With the estimates of the grade and wage parameters taken as given, we estimate the flow payoffs in a second stage. Following Arcidiacono & Miller (2011), we express the future payoffs in such a way that avoids solving the full backwards recursion problem. Namely, the expected value function at time $t + 1$ can be expressed relative to the conditional value function for one of the choices plus a function of the conditional choice probabilities. With the assumption that the preference shocks are distributed Type 1 extreme value, the expected value function can be expressed as:

$$E_t \left[V_{t+1} \left(Z_{it+1} | d_{it} = (j,k)\right)\right] \quad = \quad E_t \left[v_{j'k't+1}(Z_{it+1}) - \ln\left(p_{j'k't+1}(Z_{it+1}) | d_{it} = (j,k)\right]\right)\right]$$

for any choice (j', k'), where $p_{j'k't+1}(Z_{it+1})$ is the conditional choice probability (CCP) of choosing $d_{it+1} = (j', k')$.

Recall that in estimation it is difference in the conditional value functions that are relevant, not the conditional value functions themselves. Consider any choice (j', k') as well as the choice $(0, 0)$ (home). Given these initial choices, it is straightforward to show that there exists a sequence of choices such that, in expectation, individuals will be in the same state three periods ahead, namely:

$$\begin{aligned} E_t \left[V_{t+3}(Z_{it+3}) | d_{it} = (0,0), d_{it+1} = (j',k'), d_{it+2} = (0,0)\right] \quad &= \\ E_t \left[V_{t+3}(Z_{it+3}) | d_{it} = (j',k'), d_{it+1} = (0,0), d_{it+2} = (0,0)\right] & \end{aligned}$$

We can then reformulate the problem in terms of two-period ahead flow payoffs and conditional choice probabilities and then estimate the conditional choice probabilities (CCPs) in a first stage. The differenced conditional value function is then:

$$v_{jkt}(Z_{it}) - v_{00t}(Z_{it}) = \begin{pmatrix} u_{jk}(Z_{it}) - \beta E_t \left(\ln\left[p_{00t+1}(Z_{t+1})\right] | Z_{it}, d_{it} = (j,k)\right) \\ +\beta E_t \left(\ln\left[p_{jkt+1}(Z_{t+1})\right] - u_{jk}(Z_{t+1}) | Z_{it}, d_{it} = (0,0)\right) \\ +\beta^2 E_t \left(\ln\left[p_{00t+2}(Z_{t+2})\right] | Z_{it}, d_{it} = (0,0), d_{it+1} = (j,k)\right) \\ -\beta^2 E_t \left(\ln\left[p_{00t+2}(Z_{t+2})\right] | Z_{it}, d_{it} = (j,k), d_{it+1} = (0,0)\right) \end{pmatrix}$$

Estimation of the flow utility parameters then involves the following steps:

1. Estimate the CCPs via a flexible multinomial logit model.[15]

2. Calculate the expected differenced future value terms along the finite dependence paths.

3. Estimate the flow utility parameters after expressing the future value function as a function of the CCPs. Having estimated the CCPs in a first step, this simply amounts to estimating a multinomial logit with an offset term.

[15] The CCPs are identified from the data and could in principle be estimated nonparametrically. However, we choose to estimate them using a parametric specification to avoid the curse of dimensionality.

5.4 Estimation with permanent unobserved heterogeneity

We account for permanent unobserved unobserved heterogeneity by assuming that individuals are one of R types where type is orthogonal to the covariates at $t = 1$. Accounting for type-specific unobserved heterogeneity breaks down the separability between the choice and outcome components of the likelihood described above as our full likelihood function is:

$$\sum_i \ln\left[\sum_{r=1}^{R} \pi_r L_{d,i|r} L_{G,w_u,w_s,i|r}\right] \tag{5.1}$$

Following Arcidiacono & Miller (2011), we use an adaptation of the EM algorithm that restores the additive separability of the likelihood function. Rather than updating the structural parameters of the decision process at each step, we use their two stage approach and approximate the decision process with a reduced form. Specifically, let $L^*_{d,i|r}$ give the reduced form likelihood conditional on being of type r. The probability of i being the rth type follows from Bayes rule:

$$q_{ir} = \frac{\pi_r L^*_{d,i|r} L_{G,w_u,w_s,i|r}}{\sum_{r'=1}^{R} \pi_{r'} L^*_{d,i|r'} L_{G,w_u,w_s,i|r'}} \tag{5.2}$$

In the first stage we recover the parameters of the grade and wage processes, the (type-specific) CCPs, and the conditional probabilities of being each type.

The second stage boils down to a weighted multinomial logit with an offset term. Note that this is identical to the case without unobserved heterogeneity except that now the q_{ir}'s are used as weights. Relative to full solution methods, this estimation procedure yields very substantial computational savings, and only uses the CCPs two periods ahead. Thanks to the latter feature, our estimates do not hinge on any behavioral assumptions of the model far into the future.

5.5 Missing college majors and GPAs

In our data, college GPAs and four-year college majors are each missing at a fairly high rate. This is especially true for the first period of college enrollment. We take this issue into account within our estimation procedure, by treating the first instance of unobserved GPA or major as another unobserved latent variable. We discretize the observed GPA distribution into quartiles so that unobserved GPA can be treated as a discrete unobserved type. The estimation procedure discussed above can be easily adjusted to allow for these additional latent variables.

Specifically, along with the type-specific unobserved heterogeneity distribution, the distribution of (unobserved) GPAs and majors, conditional on each heterogeneity type, is estimated within the first stage of our estimation procedure. The distribution of the unobserved majors is then taken as given in the second stage of the estimation, which still corresponds to a weighted multinomial logit where the weights are given by $\Pr(\text{Type}, \text{Major}|\text{data})$, $\Pr(\text{Type}, \text{GPA}|\text{data})$, or $\Pr(\text{Type}, \text{Major}, \text{GPA}|\text{data})$ (depending on which outcomes are missing) instead of $\Pr(\text{Type}|\text{data})$. The log-likelihood which is maximized at the M-step is now conditional on both the heterogeneity type as well as the major or GPA quartile.

6 Results

In this section, we present and discuss the estimation results. All of the results discussed below were obtained assuming the existence of $R = 2$ unobserved heterogeneity types. Type 1 (respectively Type 2) individuals account for 50.3% (resp. 49.7%) of the overall population.

6.1 Grade parameters

The parameter estimates for the grade equations are presented in Table 11. All else equal, blacks are found to have lower GPA than whites across the board, particularly so in 2-year colleges. While both grades in high school as well as SAT scores are significant predictors of grades, it is worth noting that high school GPA plays a particularly important role for all types of schools and majors. SAT Math and Verbal play a similar role in predicting grades in 4-year college science majors, while the returns to SAT Verbal are substantially larger than the returns to SAT Math in 2-year colleges and, to a lesser extent, in 4-year college non-science majors. This pattern contrasts with the wage returns to SAT scores (discussed in Subsection 6.2), with the estimated returns to SAT Verbal being negative after controlling for SAT Math. Working while in college is associated with lower grades for both types of colleges and majors, although the effects are modest. Interestingly and consistent with Hansen et al. (2004), returns to the ability index are found to be smaller after sophomore year for both groups of majors in 4-year colleges. Finally, turning to the type-specific unobserved ability (known to the agent), students in the Type 1 group have lower GPA in two-year colleges, and higher GPA in four-year colleges.[16]

[16]These type-specific coefficients appear to be imprecisely estimated. However the standard errors, which are currently estimated using bootstrap, are likely conservative here due to the existence of multiple local

6.2 Wage parameters

Estimates of the wage equations are given in Table 12. All else being equal, blacks have significantly lower wages in the unskilled sector. For both sectors we find significant and sizeable returns to SAT math scores (6% and 11% in the unskilled and skilled sectors, respectively). The returns to SAT verbal are negative for both sectors. The latter finding echoes a number of other papers finding a negative effect of SAT verbal after controlling for SAT math (see, e.g., Arcidiacono, 2004, Kinsler & Pavan, 2015 and Sanders, 2015). Returns to experience are slightly larger in the skilled sector (8%) than in the unskilled sector (6%), although both returns are of similar order of magnitude. Experience in the unskilled sector does not translate into higher labor market earnings in the skilled sector. One possible explanation is that, after graduating from college, individuals who have accumulated more experience in the unskilled sector might also be more likely to end up working in a relatively low-paying occupation. Investigating this issue would require modeling the choice of occupation. This does highlight the importance of accounting for sector-specific labor market experiences in this context, as both types of experiences are rewarded very differently on the labor market. Returns to schooling in the unskilled sector are positive and significant, even though they are quantitatively pretty small. Working while in school (as opposed to working without being enrolled in college) results in a substantial wage loss, particularly for part-time work in both two-year and four-year colleges. Turning to the skilled sector, we find positive returns to graduate schooling. Besides, all else equal, graduating from a science (as opposed to non-science) major increases the wage in the skilled sector by 14%. It is interesting to note that the science premium, while significant and sizable, is on the low side of the range of the premia which have been estimated in earlier studies on this question (Altonji et al., 2015). This may be partly driven by the fact that we control for selection into college majors in a more thorough way than most of these previous analyses. Finally, type 1 individuals have higher wages than type 2 individuals in the skilled sector (17% premium) but lower wages in the unskilled sector (11% penalty), consistent with the existence of a sector-specific comparative advantage.

---

optima of the objective function in the first step of the estimation procedure.

6.3 Learning

Table 13 presents the correlation matrix for the unobserved abilities (initially unknown to the individual) in each sector, along with their variances. Schooling ability is highly correlated across college types and majors. The correlations across college types and majors range from 0.703 (for four-year college science majors and two-year colleges) to 0.902 (for four-year colleges non-science majors and four-year colleges science majors). A similar picture emerges across skilled and unskilled sectors within the labor market, which are strongly correlated (estimated correlation coefficient of 0.781).

The correlations between schooling abilities and labor market productivity are all positive, but markedly lower than the correlations across college types and majors. Notably, the correlations between ability in four-year college non-science or science major, and productivity in the skilled sector are very small (between 0.05 and 0.06). Correlations between schooling abilities in four-year colleges and labor market productivity are larger for the unskilled sector, although both of them remain below 0.21. Taken together, these patterns provide clear indication that grades earned in college, in science as well as in non-science majors, reveal little information about future labor market performance.

Finally, it is worth noting that the unobserved ability variance is much larger for 4-year sciences (even so after rescaling by the variance of the corresponding outcomes), which suggests that the role played by unobserved ability is more important in those majors. Table 14 further shows that, even though our approach allows us to account for both types of unobserved ability (known and unknown to the individuals), residual variation in log-wages and GPA does remain sizeable. Our estimates also show that grades in 2-year colleges are noisier signals of ability than in 4-year colleges. Finally, for all types of colleges and majors, the ability signals are less precise in the initial relative to the subsequent periods of college enrollment.

6.4 Flow payoffs

Finally, Table 18 reports the structural parameter estimates obtained from the procedure described in Subsection 5.3. The results indicate that individuals with higher prior ability have a higher utility for two and four-year colleges (relative to home production), with a larger coefficient for non-Science majors. Similarly, individuals with higher SAT in Math have a higher utility for all schooling options, particularly so in four-year college science majors. On the other hand, SAT verbal has a negligible effect on the utility of two-year college as well as four-year college science majors.

The same holds true for high school grades, which are positively associated with the utility for all (undergraduate) schooling options, especially so in four-year colleges. Overall, this pattern is consistent with a cost of effort decreasing with these ability measures. Consistent with the existence of higher monetary costs of attending a four-year college (as opposed to a two-year college), individuals whose parents went to college also have a substantially higher utility for four-year colleges relative to two-year colleges. As expected, individuals with higher expected log-wages have a higher utility for work.[17] Furthermore, the estimated coefficients on previous activities point to the existence of large switching costs across types of colleges and majors. Finally, Type-1 individuals are found to have higher preferences for four-year colleges and a relative distaste for two-year colleges. Together with the estimation results obtained for the grade equations, the latter estimates point to a positive correlation between preferences and abilities for both types of colleges.

6.5 Model fit and Sorting

Tables 19 and 20 report the fit of the model in several relevant dimensions. Table 19 reports, for each period, the empirical frequency (Data column) and the choice frequency computed using the model (Model column) for four different events, namely college entry, college attrition, college re-entry and graduation. The choice frequencies in the model column are computed through forward simulation, using the structural parameter estimates presented above along with the reduced-form CCPs. Overall, although there are some non-trivial discrepancies in several cases, most of the predicted choice frequencies are reasonably close to the empirical ones. Importantly, our model does a good job in predicting the dynamics of these choices across all periods. Table 20 shows the fit of the model in terms of choice frequencies, pooled across all periods. The fit is pretty good, with the predicted choice frequencies being in most cases very close to the empirical frequencies.

Turning to the sorting patterns, Table 21 shows the posterior mean for each unobserved ability for different choice paths. These results are obtained by forward simulating 100 times, for each individual in the sample, the outcomes (grades and log-wages) and sequences of choices.[18]

[17] The expected log-wages utility coefficients are restricted to be the same for part-time and full-time work.

[18] Note that one could alternatively construct a similar table by using the learning estimates only and then compute the average posterior abilities for those who chose particular paths. Using forward simulations

Though sorting effects are relatively small, the signs are generally in the expected direction. Those who go continuously to college and graduate with a degree in science have relatively high (posterior) ability in science, but also, to a lesser extent, in non-science majors and two-year colleges. Individuals who are continuously enrolled in college and graduate with a non-science degree also have larger abilities in all types of schools and majors than dropouts and stopouts, as well as individuals who never attend college. However, it is interesting to note that, among the group of individuals who do not work while in college, science college graduates are on average not only better than non-science college graduates in science, but they also have slightly higher posterior ability in non-science majors. This pattern points to college graduates in sciences having an absolute advantage in both types of majors.

Those who go continuously to college and work while in school at all periods have higher unskilled (and skilled) labor market productivities than those who do not work while in school at all periods. Individuals who stop out, but then graduate from college in science have lower schooling ability in four-year sciences than the continuous enrollees who work two years or less in college and obtain a degree in science. On the other hand, these individuals have substantially larger schooling abilities, for all types of college and majors, than the individuals who stopped out but then left college again without graduating. Individuals for whom dropping out was an absorbing state have on average higher schooling abilities than those who stopout and dropout from science or non-science majors, but substantially smaller abilities than those who graduate after stopping out.

Table 22 reports the posterior ability variances for different choice paths, at the time of (permanent) labor market entry. Focussing on the first two panels (continuous enrollees graduating in science or non-science), the results show that a fair amount of ability learning takes place while in college. On average among college graduates in science, the posterior ability variance in science is about six times smaller upon graduation than when they graduated from high school. Consistent with the existence of a smaller signal-to-noise ratio in non-science majors, learning is somewhat slower but nonetheless sizable in non-science fields, with the posterior ability variance being about five times smaller upon graduation. Individuals also learn about their abilities in other types of colleges and majors than the one they graduate from. For instance, the posterior non-science ability variance is three times and a half smaller upon graduation from science major.

enables to work with larger cell sizes.

Table 23 reports the counterfactual sorting patterns which are obtained by simulating the model in a scenario where all the individuals would have perfect information on their abilities from the initial period. Comparing how much ability sorting there is in this counterfactual full information scenario with the baseline sorting patterns discussed above (Table 21) speaks to the cost of imperfect information in this context. In most cases, the sorting effects go in the same direction as the ones obtained earlier without assuming perfect information. However, the magnitudes of these effects are generally much larger, and we find much stronger evidence of sorting on comparative advantage in the perfect information scenario. For instance, individuals who do not work while in school, are continuously enrolled and graduate from a science major in a four-year college have on average a 0.86 standard deviation higher science ability than those who never enroll in college. The ability differences are much smaller in the baseline scenario, where the continuous enrollees in four-year science have a 0.21 standard deviation higher science ability that those who never enroll in college.

It is also interesting to compare sorting on science and non-science abilities across both types of majors, in the baseline and in the counterfactual scenario. In the full information scenario, among those who do not work while in school, continuous enrollees in science majors have a 0.21 standard deviation higher science ability than the continuous enrollees in non-science majors. Those differences are much smaller in the baseline scenario (0.04 standard deviation), which provides evidence that imperfect information severely limits the extent to which individuals sort across college majors based on their comparative advantage.

Finally, Table 24 reports the college completion status frequencies in the baseline and counterfactual full-information scenario. The predicted share of individuals graduating from four-year college would increase by about 10 points (a 40% relative increase) if they had perfect information on their ability by the end of high school. This substantial increase in college graduation rates in the full information scenario is primarily driven by a 22% decrease in dropout rates. On the other hand, the proportion of individuals who never enroll in college remains very stable under both scenarios. Importantly the sorting estimates discussed above show that, although the share of individuals enrolling in college would not change much, providing high school students will full information on their ability would significantly affect the composition of college enrollees.

7 Conclusion

This paper examines the determinants of college attrition, in a situation where individuals have imperfect information about their schooling ability and labor market productivity. Using longitudinal data from the NLSY97, we estimate a dynamic model of college attendance, major choice and work decisions. A key feature of our framework is to account for correlated learning about ability and productivity through college grades and wages. Estimation results show that a sizable fraction of the dispersion in college grades as well as log-wages is attributable to the ability components which are gradually revealed to individuals as they accumulate more signals. These ability components are highly correlated across college types and majors, and across the skilled and unskilled labor market. In contrast, grades earned in college, in science as well as in non-science majors, turn out to reveal little information about future labor market performance. To the extent that part of the mission of higher education is to help prepare students for the labor market, this finding suggests that there is room for improvement in the screening mechanisms in place in college. Finally, simulations conducted under a counterfactual full information scenario indicate that four-year college graduation rates would increase substantially relative to the baseline imperfect information scenario, mostly through a decrease in dropout rates. Imperfect information on ability also has significant implications regarding the composition of college graduates, dropouts and stopouts. We find in particular evidence that imperfect information on ability acts as a barrier to the pursuit of comparative advantage through schooling choices.

References

Adda, J., Dustmann, C., Meghir, C. & Robin, J.-M. (2010), Career progression and formal versus on-the-job training. Working paper.

Altonji, J. (1993), 'The demand for and return to education when education outcomes are uncertain', *Journal of Labor Economics* **11**, 48–83.

Altonji, J., Arcidiacono, P. & Maurel, A. (2015), The analysis of field choice in college and graduate school: Determinants and wage effects, *in* E. Hanushek, S. Machin & L.Woessmann, eds, 'Forthcoming in the Handbook of the Economics of Education', Vol. 5, Elsevier.

Altonji, J., Blom, E. & Meghir, C. (2012), 'Heterogeneity in human capital investments: High school curriculum, college majors, and careers', *Annual Review of Economics* **4**, 185–223.

Antonovics, K. & Golan, L. (2012), 'Experimentation and job choice', *Journal of Labor Economics* **30**, 333–366.

Arcidiacono, P. (2004), 'Ability sorting and the returns to college major', *Journal of Econometrics* **121**, 343–375.

Arcidiacono, P. & Miller, R. (2011), 'Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity', *Econometrica* **79**, 1823–1867.

Arcidiacono, P. & Miller, R. (2013), Identifying dynamic discrete choice models off short panels. Working paper.

Bound, J., Lovenheim, M. & Turner, S. (2010), 'Why have college completion rates declined? an analysis of changing student preparation and collegiate resources', *American Economic Journal: Applied Economics* **2**, 129–157.

Bound, J. & Turner, S. (2011), Dropouts and diplomas: The divergence in collegiate outcomes, *in* E. Hanushek, S. Machin & L.Woessmann, eds, 'Handbook of the Economics of Education', Vol. 4, Elsevier.

DeGroot, M. (1970), *Optimal Statistical Decisions*, New York: McGraw Hill.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data with the em algorithm', *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Goldin, C. & Katz, L. (2008), *The Race Between Education and Technology*, Belknap/Harvard University Press, Cambridge, Mass.

Hansen, K., Heckman, J. & Mullen, K. (2004), 'The effect of schooling and ability on achievement test scores', *Journal of Econometrics* **121**, 39–98.

Heckman, J., Lochner, L. & Todd, P. (2006), Earnings functions, rates of return and treament effects: the Mincer equation and beyond, *in* E. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 1, Elsevier.

Heckman, J. & Urzua, S. (2009), The option value of educational choices and the rate of return to educational choices. Working paper.

James, J. (2011), Ability matching and occupational choice. Working paper.

Joensen, J. (2009), Academic and labor market success: the impact of student employment, abilities and preferences. Working paper.

Keane, M. & Wolpin, K. (1997), 'The career decisions of young men', *The Journal of Political Economy* **105**, 473–522.

Kinsler, J. & Pavan, R. (2015), 'The specificity of general human capital: Evidence from college major choice', *Forthcoming in the Journal of Labor Economics* .

Light, A. & Strayer, W. (2000), 'Determinants of college completion: School quality or student ability?', *Journal of Human Resources* **35**, 299–332.

Magnac, T. & Thesmar, D. (2002), 'Identifying dynamic discrete decision processes', *Econometrica* **70**, 801–816.

Manski, C. F. (1989), 'Schooling as experimentation: a reappraisal of the postsecondary dropout phenomenon', *Economics of Education Review* **8**, 305–312.

Manski, C. & Wise, D. (1983), *College Choice in America*, Harvard University Press.

Miller, R. (1984), 'Job matching and occupational choice', *The Journal of Political Economy* **92**, 1086–1120.

Navarro, S. (2008), Control functions, *in* S. Durlauf & L. Blume, eds, 'The New Palgrave Dictionary of Economics', London: Palgrave Macmillan Press.

Pugatch, T. (2012), Bumpy rides: School to work transitions in south africa. IZA Discussion Paper No. 6305.

Robin, J.-M. (2011), 'On the dynamics of unemployment and wage distributions', *Econometrica* **79**, 1327–1355.

Rust, J. (1994), Estimation of dynamic structural models, problems and prospects: Discrete decision processess, *in* C. Sims, ed., 'Advances in Econometrics: Sixth World Congress', Vol. 2, Cambridge University Press, New York, pp. 5–33.

Sanders, C. (2010), Skill uncertainty, skill accumulation, and occupational choice. Working paper.

Sanders, C. (2015), Reading skills and earnings: Why do doing words good hurt you're wages? Working paper.

Stange, K. (2012), 'An empirical investigation of the option value of college enrollment', *American Economic Journal: Applied Economics* **4**, 49–84.

Stinebrickner, T. & Stinebrickner, R. (2003), 'Working during school and academic performance', *Journal of Labor Economics* **21**, 473–491.

Stinebrickner, T. & Stinebrickner, R. (2012), 'Learning about academic ability and the college drop-out decision', *Journal of Labor Economics* **30**, 707–748.

A Data

This appendix section details the criteria we use to define science majors, as well as how we select our estimation subsample. Table A.1 lists the majors in each category. Table A.2 outlines each of the criteria used to construct our estimation subsample.

Table A.1: Major Definitions

Science (STEM) Majors	Non-STEM Majors
Agriculture and natural resource sciences	All other majors
Biological sciences	
Computer/Information science	
Engineering	
Mathematics	
Physical sciences	
Nutrition/Dietetics/Food Science	

Table A.2: Sample Selection

Selection criterion	Resultant persons	Resultant person-years
Full NLSY97 sample	8,984	134,760
Drop females	4,599	68,985
Drop other race	4,559	68,385
Drop missing test scores, HS grades or Parental education	3,327	49,905
Drop HS Dropouts (or those not receiving GED)	2,959	44,385
Drop observations before HS graduation	2,843	31,573
Drop right-censored missing interview spells	2,841	29,749
Drop any who attend college at a young age or graduate college in 2 or fewer years	2,841	29,097
Drop any who are not in HS at age 15 or under or have other outlying data	2,841	29,061
Drop observations after someone has a missing 4-year college major	2,454	23,083
Drop observations after someone has a missing wage while working	2,245	18,665
Drop observations after someone has a missing GPA while in college	1,914	14,104
Final estimation subsample (descriptive statistics)	1,914	14,104
Final estimation subsample (structural estimation)[a]	2,713	21,343

[a] The structural estimation incorporates integration of missing GPA and major observations, as discussed in Section 5.

Table A.3: Cell sizes of ability covariance matrix

	Skilled	Unskilled	Science	Humanities	2-year
Correlation matrix					
Skilled	352	273	233	300	67
Unskilled	273	2,257	555	731	767
4 year Science	233	555	796	660	143
4 year Humanities	300	731	660	990	197
2 year	67	767	143	197	843

Note: Numbers reflect the number of individuals that ever participate in the given combinations of sectors.

Table 1: SAT scores and race, broken out by college enrollment status

(a) Full Descriptive Sample

Variable	Obs	Mean	Std. Dev.
SAT math	1,913	-0.266	0.705
SAT verbal	1,913	-0.280	0.748
black	1,913	0.256	0.436
hispanic	1,913	0.187	0.390

(b) Start in four-year science

Variable	Obs	Mean	Std. Dev.
SAT math	128	0.480	0.788
SAT verbal	128	0.195	0.783
black	128	0.141	0.349
hispanic	128	0.141	0.349

(c) Start in four-year humanities

Variable	Obs	Mean	Std. Dev.
SAT math	235	0.130	0.743
SAT verbal	235	0.128	0.795
black	235	0.234	0.424
hispanic	235	0.098	0.298

(d) Start in two-year college

Variable	Obs	Mean	Std. Dev.
SAT math	326	-0.266	0.587
SAT verbal	326	-0.251	0.647
black	326	0.212	0.409
hispanic	326	0.199	0.400

Note: Raw SAT scores are standardized to be mean-zero unit-variance for the NLSY97 population. Predicted SAT scores (reported as "SAT" throughout) are not mean-zero because the population that takes the SAT is higher ability. The process used to construct predicted SAT scores for non-test-takers is described in the text.

Table 2: GPA over time by type of college attended

(a) Four-year college Science

Variable	Obs	Mean	Std. Dev.
year 1	128	2.891	0.899
year 2	98	3.069	0.571
year 3	76	3.156	0.556
year 4	81	3.174	0.561

(b) Four-year college Humanities

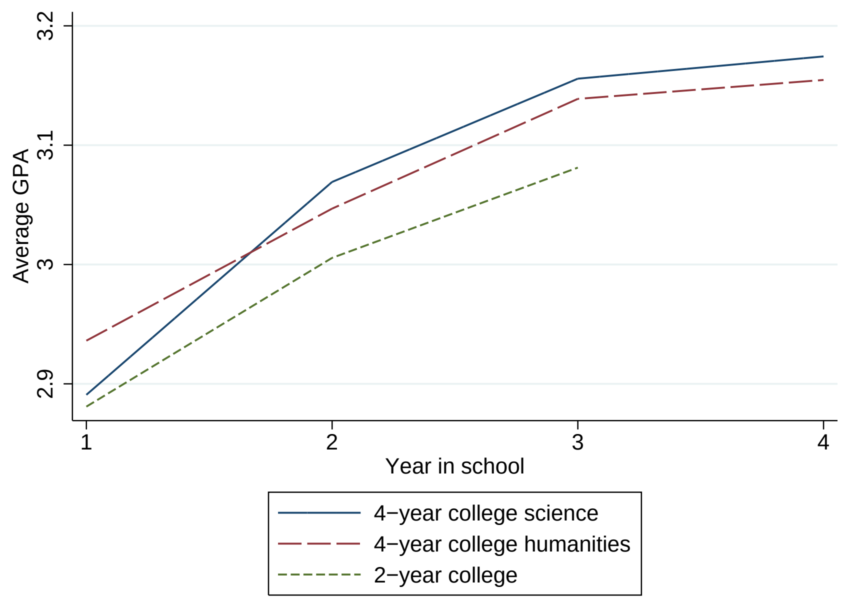
Variable	Obs	Mean	Std. Dev.
year 1	235	2.936	0.751
year 2	183	3.047	0.544
year 3	163	3.139	0.511
year 4	131	3.155	0.477

(c) Two-year college

Variable	Obs	Mean	Std. Dev.
year 1	326	2.881	0.895
year 2	189	3.006	0.656
year 3	83	3.081	0.653
year 4	21	2.811	0.813

Figure 1: GPA over time by type of college attended

(a) Average GPA



(b) Standard Deviation of GPA

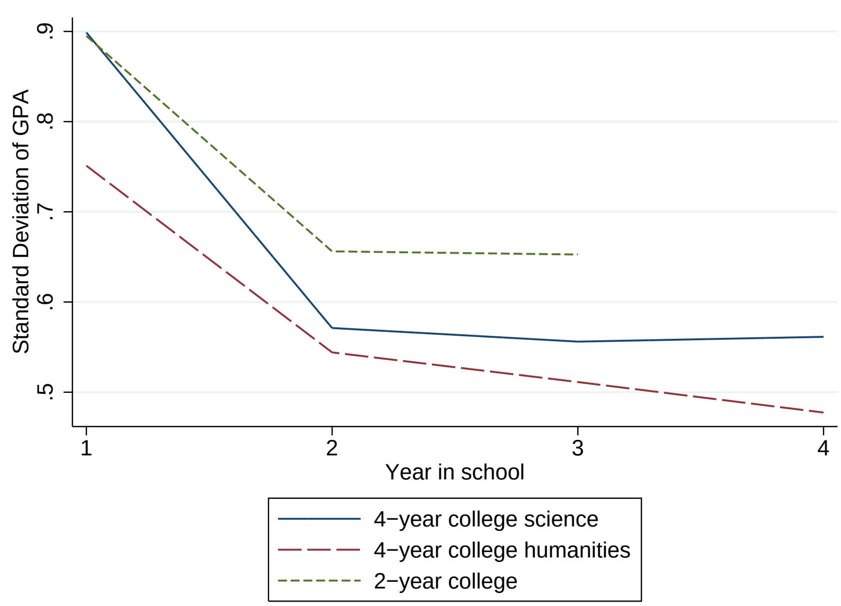


Table 3: Outcomes of college enrollees

	Full Sample	Estimation Subsample	Starting College Type Two-Year	Four-Year Sci	Four-year Hum
Continuous completion (CC)	30.48%	33.38%	10.12%	57.03%	52.77%
Stopped out (SO) but graduated	6.26%	7.11%	7.36%	7.03%	6.81%
Stopped out (SO) then dropped out	11.16%	10.16%	15.03%	3.91%	6.81%
Dropped out (DO)	39.91%	37.45%	52.76%	20.31%	25.53%
CC right censored	4.53%	3.19%	1.53%	4.69%	4.68%
SO right censored	7.66%	8.71%	13.19%	7.03%	3.40%
Total N	2,428	689	326	128	235

Notes: Full sample refers to all males in the NLSY97. Estimation subsample refers to individuals included in the structural estimation (however, completion status is still determined using the full data, so in the structural estimation some categories may be underrepresented because of missing outcomes early in the college career). Students who begin two-year college but never enroll in a four-year college are considered as dropouts.

Table 4: Outcomes of college enrollees, conditional on having graduated by Round 15 of NLSY97, by type of college first attended

	Two-year		Four-Year Sci		Four-year Hum		Any	
	N	%	N	%	N	%	N	%
CC	35	59.3%	73	89.0%	125	88.7%	233	82.6%
SO	24	40.7%	9	11.0%	16	11.3%	49	17.4%
Total	59	100.0%	82	100.0%	141	100.0%	282	100.0%

Table 5: Background characteristics of college enrollees

	SAT math	SAT verbal	HS GPA	Mother with BA	Family Income ($1996)	N
CC	0.33	0.26	0.68	44.44%	59,254	252
SO	-0.03	-0.10	0.04	21.79%	39,140	179
DO	-0.28	-0.29	-0.12	18.99%	38,887	258
Total	0.01	-0.04	0.21	29.03%	46,402	689

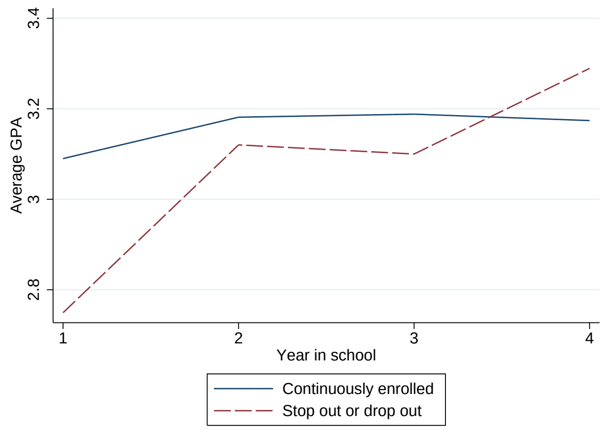
Note: SAT scores and high school GPA are each standardized to be mean-zero unit-variance for NLSY97 population. See text for a description of how SAT scores are constructed.

Table 6: Average GPA over time by college completion status

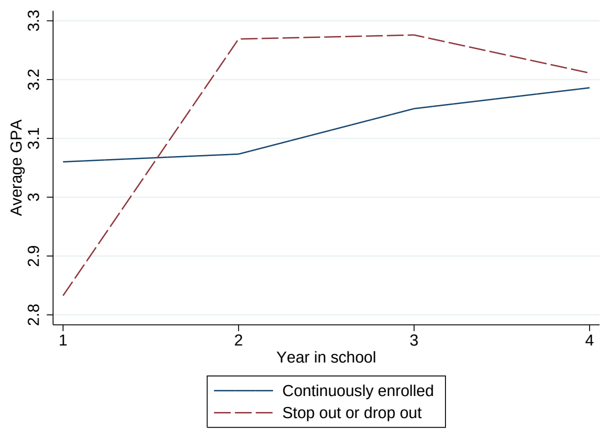
	Four-Year Sci		Four-Year Hum		Two-Year	
Period	CC	DO/SO	CC	DO/SO	CC	DO/SO
1	3.09	2.75	3.06	2.83	3.22	2.83
2	3.18	3.12	3.07	3.27	3.08	3.01
3	3.19	3.10	3.15	3.28	3.27	3.17
4	3.17	3.29	3.19	3.21	3.30	2.32

Figure 2: Average GPA over time by college completion status

(a) 4-year college science



(b) 4-year college humanities



(c) 2-year college

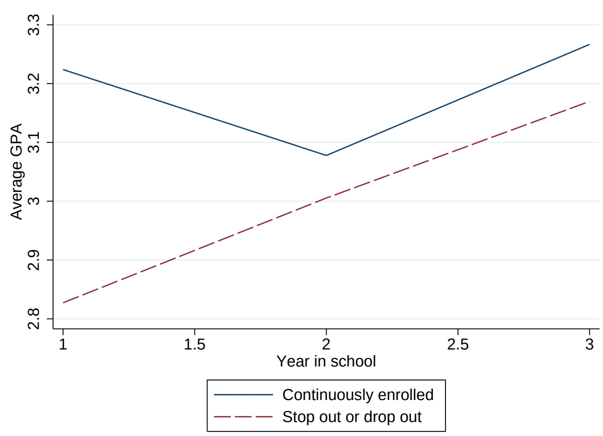


Table 7: Difference between actual and expected period-t grades (by $t+1$ period college decision)

(a) 4-year Science Majors

	Mean residual	Std Dev	N	Mean diff T (p-val)
Drop out from 4-year college & science	-0.203	0.629	28	1.87
Complement	0.014	0.593	411	(0.06)
Switch to 4-year college & humanities	-0.021	0.535	42	0.24
Complement	0.002	0.604	397	(0.81)
Switch to 2-year college	0.189	0.794	8	0.90
Complement	-0.004	0.593	431	(0.37)

(b) 4-year Humanities Majors

	Mean residual	Std Dev	N	Mean diff T (p-val)
Drop out from 4-year college & humanities	-0.104	0.686	99	2.00
Complement	0.014	0.530	722	(0.05)
Switch to 4-year college & science	-0.140	0.659	35	1.54
Complement	0.006	0.547	786	(0.13)
Switch to 2-year college	-0.108	0.498	13	0.71
Complement	0.002	0.553	808	(0.48)

(c) 2-year Students

	Mean residual	Std Dev	N	Mean diff T (p-val)
Drop out from 2-year college	-0.163	0.901	201	3.78
Complement	0.076	0.654	431	(0.00)
Switch to 4-year college (any major)	-0.026	0.599	42	0.23
Complement	0.002	0.759	590	(0.82)
Switch to 4-year college & science	-0.291	0.558	11	1.30
Complement	0.005	0.751	621	(0.19)
Switch to 4-year college & humanities	0.068	0.593	31	0.52
Complement	-0.004	0.756	601	(0.60)

Note: regression covariates include race dummies, SAT scores, parental education, high school GPA, age dummies, and work intensity dummies.

Table 8: Major transition matrix for stopouts

	Major when returned after stopping out			
Major before stopping out	science	humanities	2-year college	N
science	25.00%	25.00%	50.00%	8
humanities	9.09%	54.55%	36.36%	22
2-year college	15.00%	23.33%	61.67%	60
Total	14.44%	31.11%	54.44%	90

Note: Table only includes first instance of stopping out

Table 9: Average GPA over time by college type and work intensity

(a) 4-year Sciences

	Period			
	1	2	3	4
Work FT	2.72	3.29	3.07	3.21
N	15	8	9	19
Work PT	2.87	2.99	3.10	3.12
N	27	27	23	25
No Work	2.93	3.07	3.20	3.20
N	86	63	44	37
Total	2.89	3.07	3.16	3.17
N	128	98	76	81

(b) 4-year Humanities

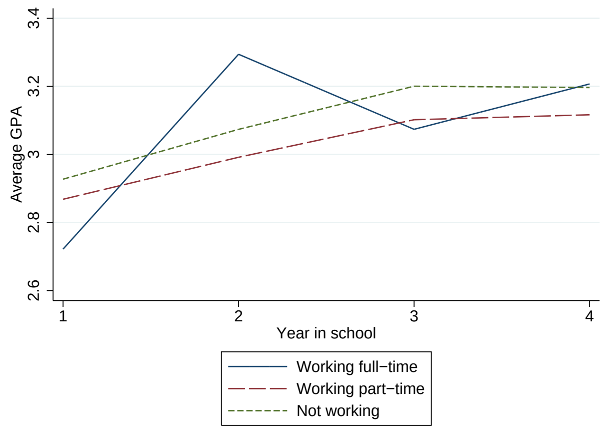
	Period			
	1	2	3	4
Work FT	2.78	3.08	3.22	3.02
N	22	30	28	19
Work PT	2.87	2.99	3.07	3.18
N	51	52	46	46
No Work	2.98	3.07	3.15	3.17
N	162	101	89	66
Total	2.94	3.05	3.14	3.15
N	235	183	163	131

(c) 2-year college

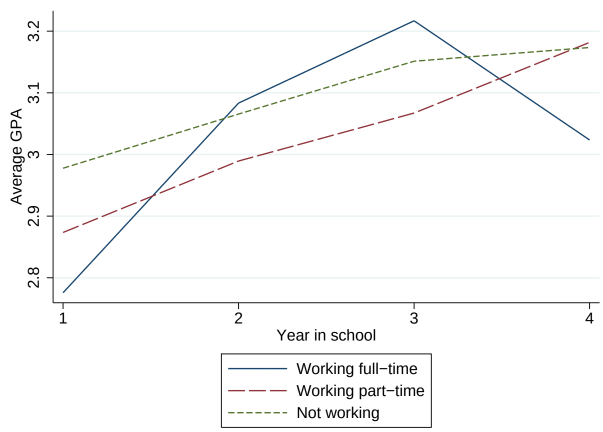
	Period			
	1	2	3	4
Work FT	2.90	3.06	3.18	3.10
N	99	65	27	7
Work PT	2.79	3.04	3.01	2.60
N	99	55	30	7
No Work	2.93	2.93	3.06	2.73
N	128	69	26	7
Total	2.88	3.01	3.08	2.81
N	326	189	83	21

Figure 3: Average GPA over time by college type and work intensity

(a) 4-year college science



(b) 4-year college humanities



(c) 2-year college

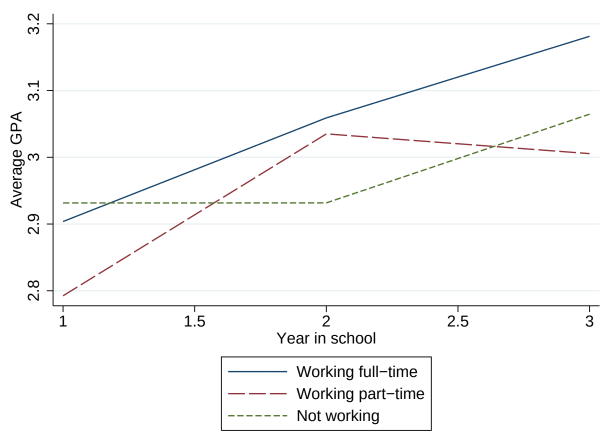


Table 10: Difference between actual and expected wages in time t for stopouts (by $t+1$ decision)

	Mean residual	Std Dev	N	Mean diff T (p-val)
Stay in work	0.100	0.533	1,129	2.24
Return to school	-0.054	0.346	61	(0.03)
Total	0.092	0.526	1,190	

Note: Residuals do not average to zero here because we compute the residuals using all wage observations in the estimation subsample of the data, but we condition the table on those who have attended at least one year of college, are currently working, and have not yet graduated from college. Regression covariates include levels and interactions of the following variables: race and year dummies; SAT scores; experience; age; in-school work dummies; and work intensity dummies.

Table 11: Estimates of 2- and 4-year GPA Parameters

	4 year Science		4 year Humanities		2 year	
	Coeff.	Std. Error	Coeff.	Std. Error	Coeff.	Std. Error
Constant	3.07	(0.168)	2.91	(0.133)	2.78	(0.144)
Black	-0.14	(0.139)	-0.07	(0.069)	-0.20	(0.060)
Hispanic	0.01	(0.128)	-0.09	(0.082)	-0.03	(0.049)
SAT math	0.13	(0.077)	0.08	(0.048)	0.00	(0.038)
SAT verbal	0.12	(0.071)	0.12	(0.050)	0.09	(0.035)
Parent graduated college	0.01	(0.107)	0.07	(0.056)	0.02	(0.051)
HS Grades	0.21	(0.056)	0.21	(0.033)	0.17	(0.027)
Work FT	-0.10	(0.073)	-0.07	(0.040)	-0.02	(0.053)
Work PT	-0.04	(0.058)	-0.05	(0.036)	-0.08	(0.056)
Age 18 and under	-1.00	(0.098)	-0.65	(0.057)	-0.39	(0.061)
Age 19	-0.59	(0.085)	-0.30	(0.048)	-0.25	(0.066)
Age 20	-0.25	(0.080)	-0.19	(0.044)	-0.33	(0.076)
Age 21	-0.11	(0.064)	-0.14	(0.038)	-0.09	(0.066)
Year 2+					0.43	(0.041)
λ_0 (ability index intercept)	0.12	(0.219)	0.25	(0.144)	0.00	(—)
λ_1 (ability index loading)	0.90	(0.070)	0.93	(0.047)	1.00	(—)
Unobserved type 1	0.06	(0.249)	0.05	(0.217)	-0.08	(0.218)
Person-years obs.	1,045		2,248		1,570	

Notes: Bootstrapped standard errors in parentheses. Age 22 and older is the reference category for the age dummies. Not working while in school is the reference category for the work intensity dummies.

Table 12: Estimates of Skilled and Unskilled Wage Parameters

	Skilled		Unskilled	
	Coeff.	Std. Error	Coeff.	Std. Error
Constant	2.08	(0.182)	1.98	(0.122)
Black	0.03	(0.052)	-0.09	(0.020)
Hispanic	-0.02	(0.052)	-0.01	(0.018)
SAT math	0.11	(0.025)	0.06	(0.012)
SAT verbal	-0.05	(0.026)	-0.02	(0.012)
Parent graduated college	-0.01	(0.039)	0.00	(0.016)
HS Grades	0.08	(0.028)	-0.01	(0.009)
Age	0.00	(0.011)	0.01	(0.005)
Unskilled Experience	0.00	(0.011)	0.06	(0.003)
Skilled Experience	0.08	(0.008)		
PT work	-0.07	(0.039)	-0.02	(0.010)
PT 2 year			-0.11	(0.017)
PT 4 year			-0.16	(0.024)
FT 2 year			-0.05	(0.019)
FT 4 year			-0.07	(0.020)
PT graduate school	0.00	(0.065)		
FT graduate school	-0.08	(0.039)		
1 year graduate school	0.11	(0.045)		
2 years graduate school	0.06	(0.057)		
3 years graduate school	0.10	(0.089)		
4+ years graduate school	0.09	(0.090)		
1 year college			0.06	(0.012)
2 years college			0.06	(0.015)
3 years college			0.11	(0.021)
4+ years college			0.15	(0.020)
Science major	0.14	(0.048)		
Unobserved type 1	0.17	(0.166)	-0.11	(0.148)
person-years	1,700		12,372	

Notes: Bootstrapped standard errors in parentheses. Full-time outside of school is the reference category for the work intensity dummies. Controls for calendar year dummies were also included.

Table 13: Correlation Matrix and Variances for Unobserved Abilities

	Skilled	Unskilled	Science	Humanities	2-year
Correlation matrix					
Skilled	1.000	0.781	0.057	0.053	0.321
	(—)	(0.054)	(0.099)	(0.071)	(0.151)
Unskilled	0.781	1.000	0.212	0.163	0.180
	(0.054)	(—)	(0.078)	(0.064)	(0.129)
4 year Science	0.057	0.212	1.000	0.902	0.703
	(0.099)	(0.078)	(—)	(0.058)	(0.118)
4 year Humanities	0.053	0.163	0.902	1.000	0.883
	(0.071)	(0.064)	(0.058)	(—)	(0.071)
2 year	0.321	0.180	0.703	0.883	1.000
	(0.151)	(0.129)	(0.118)	(0.071)	(—)
Variances	0.127	0.074	0.235	0.165	0.085
	(0.013)	(0.005)	(0.047)	(0.023)	(0.015)

Note: Bootstrapped standard errors in parentheses.

Table 14: Idiosyncratic Variances

Period	Skilled	Unskilled	Science	Humanities	2-year
1	0.148	0.161	0.646	0.670	0.897
	(0.005)	(0.002)	(0.051)	(0.035)	(0.047)
2			0.160	0.212	0.352
			(0.024)	(0.016)	(0.028)
3			0.140	0.135	0.372
			(0.054)	(0.025)	(0.021)
4			0.120	0.111	
			(0.045)	(0.021)	
5+			0.321	0.199	
			(0.107)	(0.031)	

Notes: Bootstrapped standard errors in parentheses. The third variance in 2-year college is the same for all periods after period 3.

Table 15: College wage premium

	skilled wage	unskilled wage	Δ
Overall premium	2.31	2.08	0.23
Sheepskin effect	2.31	2.23	0.09

Note: Wages for a 22 year-old college graduate born in 1983 with no work experience, and average background characteristics and test scores.

Table 16: Wage market shock forecasting estimates

Parameter	Coeff.	Std. Error
Drift (common)	0.0334	(0.0201)
Drift (graduates)	-0.0206	(0.0192)
Autocorrelation	0.6390	(0.1641)
SD of shock (non-graduates)	0.0428	(0.0095)
SD of shock (graduates)	0.0349	(0.0058)

Notes: Estimates come from a pooled AR(1) regression with sector-specific drift and shock variance, but common autocorrelation coefficient. Bootstrapped standard errors in parentheses.

Table 17: Estimates of Probability of Graduation

	Coeff.	Std. Error
Constant	-3.856	(0.250)
SAT math	0.083	(0.104)
SAT verbal	0.115	(0.094)
Black	-0.679	(0.188)
Hispanic	-0.664	(0.198)
HS grades	0.191	(0.078)
Parent college	0.038	(0.148)
Years in 2yr college	0.479	(0.060)
Years in 4yr college	0.867	(0.044)
Science major	-0.302	(0.129)
Prior ability science × Science major	1.247	(0.229)
Prior ability hum. × Hum. major	0.994	(0.201)
Currently working part-time	0.090	(0.123)
Currently working full-time	-0.296	(0.171)
Unobserved type 1	0.170	(0.163)
Person-year obs.	13,390	

Notes: Parameter estimates from a logit predicting probability of graduating in the following period. Estimated only on students in their junior year and above. Bootstrapped standard errors in parentheses

Table 18: Flow Utility Estimates

	2-year	4-year Sci	4-year Hum	Work PT	Work FT	Grad Sch.
Constant	**-3.622**	**-6.621**	**-5.194**	**-4.332**	**-3.467**	**-3.571**
SAT math	**0.067**	**0.453**	**0.180**	**-0.077**	-0.011	-0.002
SAT verbal	0.016	-0.004	**0.207**	**0.103**	-0.018	0.021
Black	-0.023	0.107	0.159	**-0.171**	**-0.188**	0.146
Hispanic	0.071	0.015	-0.048	-0.066	-0.038	0.016
HS grades	**0.098**	**0.436**	**0.324**	0.038	**0.020**	0.012
Parent graduated college	**0.215**	**0.552**	**0.611**	0.097	**-0.123**	**0.104**
Prior Academic Ability	0.115	0.868	0.943			0.074
Expected Log Wage				**0.471**	**0.471**	
Previous HS	**1.595**	**2.791**	**2.234**	**1.131**	**0.888**	
Previous 2-year	**2.932**	**1.769**	**1.487**	**0.321**	**0.307**	
Previous 4-year Sci	**1.314**	**5.806**	**3.185**	**0.797**	**0.580**	0.049
Previous 4-year Hum	**0.608**	**2.671**	**4.426**	**0.664**	**0.616**	0.324
Previous Work PT	0.084	0.224	**0.206**	**1.869**	**1.443**	0.116
Previous Work FT	**-0.214**	-0.010	0.095	**0.959**	**2.431**	0.299
Previous Grad School				0.257	0.436	**3.444**
Graduated 4-year college				-0.339	0.481	
Work PT	**0.519**	-0.318	-0.136			-0.338
Work FT	**-0.866**	**-1.408**	**-1.426**			**-3.651**
Unobserved type 1	-0.127	0.346	0.467	0.836	0.230	-0.085
Pr (Unobserved type = 1)	0.5028					
Pr (Unobserved major = science)	0.3139					
Pr (Unobserved GPA $\in [0.0, 2.5]$)	0.4660					
Pr (Unobserved GPA $\in (2.5, 3.0]$)	0.2211					
Pr (Unobserved GPA $\in (3.0, 3.6]$)	0.2223					
Pr (Unobserved GPA $\in [3.6, 4.0]$)	0.0905					
log likelihood	-25,839					
Person-year obs.	21,343					

Notes: Bold indicates bootstrap statistical significance at the 5% level. Beliefs on labor market productivity are included in the expected log wage term.

Table 19: Model fit: College entry, attrition, re-entry, and graduation rates by period

	College Entry		College attrition		College re-entry		Ever graduated	
Period	Data	Model	Data	Model	Data	Model	Data	Model
1	43.27	46.01	0.00	0.00	0.00	0.00	0.00	0.00
2	8.51	4.52	8.71	11.81	0.00	0.00	0.00	0.00
3	4.03	3.77	8.37	8.66	1.34	1.04	0.00	0.00
4	3.04	3.23	7.17	6.85	2.33	1.61	0.19	2.92
5	1.80	2.88	6.26	5.69	2.57	2.00	7.39	7.57
6	1.48	2.54	6.02	4.85	2.14	2.23	12.65	12.98
7	1.93	2.27	2.93	4.14	2.63	2.44	16.50	17.43
8	2.23	2.13	3.29	3.84	1.92	2.41	18.55	20.38
9	1.74	1.96	2.47	3.59	1.27	2.47	20.29	22.32
10	1.52	1.74	2.58	3.53	1.44	2.44	21.56	23.85
11	1.16	1.58	1.90	3.37	1.90	2.45	21.48	25.19
12	1.15	1.49	2.30	3.32	1.64	2.48	22.70	26.43
13	0.61	1.40	1.23	3.20	0.61	2.44	21.78	27.54
14	0.00	1.30	1.27	3.07	1.27	2.56	32.91	28.57
15	0.00	1.20	0.00	3.10	0.00	2.48	32.91	29.58

Notes: Model rates are constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 20: Model fit: Choice frequencies

Choice alternative	Data Frequency (%)	Model Frequency (%)
2yr & work FT	2.35	2.07
2yr & work PT	2.43	2.17
2yr only	2.57	2.09
4yr sci & work FT	0.68	0.54
4yr sci & work PT	1.39	0.96
4yr sci only	2.83	2.00
4yr hum & work FT	1.56	1.54
4yr hum & work PT	3.03	3.01
4yr hum only	5.94	4.96
work PT only	7.50	7.76
work FT only	46.35	49.26
home production	22.29	21.60
grad school & work FT	0.42	0.62
grad school & work PT	0.23	0.42
grad school only	0.43	0.99

Notes: Model frequencies are constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 21: Average abilities for different choice paths

Choice Path	Skilled	Unskilled	Science	Humanities	2-year	N
Continuous enrollment, graduate in science with x years of in-school work experience						
$x = 0$	0.012	0.044	0.210	0.194	0.153	2,901
$x = 1$	-0.003	0.022	0.189	0.174	0.137	3,742
$x = 2$	0.023	0.053	0.187	0.170	0.136	2,989
$x = 3$	0.048	0.077	0.140	0.127	0.104	1,659
$x = 4$	0.111	0.151	0.081	0.063	0.052	663
$x \geq 5$	0.092	0.119	0.063	0.060	0.058	285
Continuous enrollment, graduate in humanities with x years of in-school work experience						
$x = 0$	0.012	0.032	0.174	0.189	0.167	5,854
$x = 1$	-0.013	0.000	0.164	0.181	0.159	10,420
$x = 2$	0.020	0.039	0.151	0.166	0.148	10,051
$x = 3$	0.045	0.068	0.147	0.160	0.146	5,652
$x = 4$	0.070	0.095	0.130	0.140	0.131	2,131
$x \geq 5$	0.094	0.126	0.200	0.223	0.211	770
Stop out (SO)						
SO, graduate in science	0.026	0.046	0.145	0.137	0.115	1,487
SO, graduate in humanities	0.000	0.004	0.130	0.148	0.137	6,062
SO then DO, start in 2yr	-0.021	-0.033	-0.088	-0.093	-0.083	7,441
SO then DO, start in science	-0.008	-0.046	-0.263	-0.253	-0.205	2,083
SO then DO, start in humanities	-0.009	-0.031	-0.203	-0.221	-0.194	4,722
Drop out (DO) after x years of school						
$x = 1$	-0.010	-0.015	-0.050	-0.053	-0.049	17,848
$x = 2$	-0.011	-0.024	-0.135	-0.144	-0.127	10,969
$x = 3$	-0.019	-0.041	-0.194	-0.208	-0.183	6,879
$x = 4$	-0.016	-0.039	-0.215	-0.228	-0.200	4,537
$x \geq 5$	0.001	-0.025	-0.260	-0.272	-0.234	4,980
Never attended college						
never attend college	0.000	0.000	0.000	0.000	0.000	78,514

Notes: This table is constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 22: Average posterior variance after last year of college for different choice paths

Choice Path	Skilled	Unskilled	Science	Humanities	2-year	N
Continuous enrollment, graduate in science with x years of in-school work experience						
$x = 0$	0.126	0.072	0.042	0.050	0.047	2,901
$x = 1$	0.097	0.045	0.040	0.049	0.046	3,742
$x = 2$	0.083	0.033	0.040	0.048	0.045	2,989
$x = 3$	0.077	0.026	0.038	0.046	0.044	1,659
$x = 4$	0.073	0.023	0.037	0.044	0.043	663
$x \geq 5$	0.069	0.020	0.039	0.043	0.041	285
Continuous enrollment, graduate in humanities with x years of in-school work experience						
$x = 0$	0.126	0.073	0.081	0.038	0.034	5,854
$x = 1$	0.097	0.045	0.079	0.036	0.033	10,420
$x = 2$	0.083	0.032	0.076	0.034	0.032	10,051
$x = 3$	0.076	0.026	0.072	0.031	0.031	5,652
$x = 4$	0.072	0.022	0.069	0.029	0.030	2,131
$x \geq 5$	0.068	0.019	0.066	0.028	0.029	770
Stop out (SO)						
SO, graduate in science	0.082	0.032	0.049	0.044	0.040	1,487
SO, graduate in humanities	0.080	0.030	0.077	0.035	0.032	6,062
SO then DO, start in 2yr	0.078	0.030	0.156	0.095	0.048	7,441
SO then DO, start in science	0.080	0.031	0.093	0.067	0.044	2,083
SO then DO, start in humanities	0.079	0.030	0.121	0.069	0.042	4,722
Dropout (DO) after x years of school						
$x = 1$	0.113	0.062	0.209	0.145	0.075	17,848
$x = 2$	0.104	0.054	0.164	0.107	0.058	10,969
$x = 3$	0.098	0.048	0.129	0.079	0.046	6,879
$x = 4$	0.092	0.042	0.105	0.061	0.039	4,537
$x \geq 5$	0.084	0.035	0.091	0.052	0.035	4,980
Never attended college						
never attend college	0.071	0.020	0.227	0.162	0.083	78,514

Notes: This table is constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 23: Average abilities for different choice paths in full-information counterfactual scenario

Choice Path	Skilled	Unskilled	Science	Humanities	2-year	N
Continuous enrollment, graduate in science with x years of in-school work experience						
$x = 0$	0.034	0.015	0.607	0.474	0.381	6,131
$x = 1$	0.167	0.163	0.536	0.402	0.331	7,389
$x = 2$	0.237	0.277	0.500	0.368	0.300	5,275
$x = 3$	0.300	0.331	0.435	0.302	0.256	2,184
$x = 4$	0.358	0.368	0.328	0.192	0.191	664
$x \geq 5$	0.516	0.541	0.391	0.246	0.250	204
Continuous enrollment, graduate in humanities with x years of in-school work experience						
$x = 0$	0.026	-0.068	0.400	0.501	0.520	10,621
$x = 1$	0.144	0.080	0.358	0.451	0.484	14,623
$x = 2$	0.266	0.216	0.311	0.389	0.445	11,496
$x = 3$	0.345	0.288	0.284	0.350	0.424	5,065
$x = 4$	0.450	0.399	0.238	0.316	0.410	1,596
$x \geq 5$	0.478	0.365	0.212	0.252	0.369	452
Stop out (SO)						
SO, graduate in science	0.097	0.104	0.427	0.331	0.258	2,486
SO, graduate in humanities	0.120	0.062	0.264	0.345	0.378	6,998
SO then DO, start in 2yr	-0.115	-0.090	-0.207	-0.197	-0.196	5,874
SO then DO, start in science	-0.158	-0.050	0.080	0.020	-0.088	1,267
SO then DO, start in humanities	-0.142	-0.065	-0.014	0.015	-0.034	2,828
Drop out (DO) after x years of school						
$x = 1$	-0.064	-0.039	-0.217	-0.226	-0.226	14,922
$x = 2$	-0.133	-0.070	-0.125	-0.135	-0.175	7,597
$x = 3$	-0.130	-0.024	-0.057	-0.061	-0.131	4,575
$x = 4$	-0.143	-0.048	-0.029	-0.033	-0.103	2,637
$x \geq 5$	-0.121	-0.078	-0.158	-0.162	-0.185	2,358
Never attended college						
never attend college	-0.067	-0.056	-0.263	-0.284	-0.275	78,626

Notes: This table is constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 24: College completion status frequencies: baseline and counterfactual

	Baseline (%)	Counterfactual (%)
Continuous completion (CC)	20.71	29.27
Stop out (SO) but graduated	3.15	3.96
Graduate from four-year college	23.86	33.23
Stop out (SO) then drop out	6.67	4.90
Drop out (DO)	29.65	23.42
Never went to college	28.94	28.95
N	271,300	271,300

Notes: Figures constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 25: Average period-T posterior variances by demographic characteristics

Choice Path	Skilled	Unskilled	Science	Humanities	2-year	N
Race & Ethnicity						
Black	0.069	0.024	0.173	0.116	0.063	61,900
Hispanic	0.065	0.020	0.173	0.117	0.062	47,900
White	0.058	0.022	0.144	0.095	0.054	162,000
Parental college status						
Parent did not graduate	0.065	0.021	0.174	0.118	0.063	191,000
Parent graduated	0.053	0.025	0.111	0.070	0.043	80,300
SAT math quartile						
1st quartile	0.068	0.022	0.185	0.126	0.066	81,100
2nd quartile	0.066	0.021	0.176	0.119	0.063	54,600
3rd quartile	0.060	0.022	0.146	0.096	0.054	69,600
4th quartile	0.051	0.024	0.111	0.072	0.045	66,000
SAT verbal quartile						
1st quartile	0.067	0.022	0.180	0.123	0.065	76,800
2nd quartile	0.066	0.021	0.177	0.120	0.063	58,900
3rd quartile	0.060	0.022	0.145	0.096	0.054	67,800
4th quartile	0.053	0.023	0.119	0.076	0.046	67,800
High school GPA quartile						
1st quartile	0.068	0.021	0.187	0.127	0.066	137,000
2nd quartile	0.061	0.022	0.147	0.096	0.054	54,900
3rd quartile	0.055	0.023	0.120	0.077	0.047	52,300
4th quartile	0.043	0.025	0.083	0.053	0.037	27,500

Notes: This table is constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 26: Major switching probabilities: baseline and counterfactual

	Baseline (%)	Counterfactual (%)
Pr(Grad science | started science, grad 4yr)	54.76	59.73
Pr(Grad hum | started hum, grad 4yr)	89.51	85.84
Pr(Grad hum | started science, grad 4yr)	45.24	40.27
Pr(Grad science | started hum, grad 4yr)	10.49	14.16
Pr(Grad science | started 2yr, grad 4yr)	23.32	28.44
Pr(Grad hum | started 2yr, grad 4yr)	76.68	71.56
Pr(Grad science | grad 4yr)	24.39	31.32
Pr(Grad hum | grad 4yr)	75.61	68.68
N	271,300	271,300

Notes: Figures constructed using 100 simulations of the structural model for each individual included in the estimation.

Table 27: Difference in abilities at $t = 2$ for stopouts (by $t = 3$ decision)

(a) Posterior ability

Path	Unskilled	Sci	Hum	2yr	N
Stay in Work	-0.019	-0.008	-0.005	-0.001	1,611
Return to school	0.007	-0.021	-0.022	-0.018	5,825
Total	0.002	-0.018	-0.018	-0.015	7,436

(b) True ability

Path	Unskilled	Sci	Hum	2yr	N
Stay in Work	0.016	0.007	0.013	0.035	1,611
Return to school	0.021	-0.031	-0.031	-0.024	5,825
Total	0.020	-0.023	-0.022	-0.011	7,436

(c) Counterfactual true ability

Path	Unskilled	Sci	Hum	2yr	N
Stay in Work	0.122	0.103	0.109	0.104	1,174
Return to school	0.138	0.021	0.034	0.036	3,485
Total	0.134	0.042	0.053	0.053	4,659