Review of Economic Studies (2014) **81**, 426–472 doi: 10.1093/restud/rdt025

Advance access publication 23 July 2013

A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout

RALPH STINEBRICKNER
Berea College and The University of Western Ontario
and
TODD R. STINEBRICKNER
The University of Western Ontario

First version received April 2011*; final version accepted April* 2013 (*Eds.*)

Taking advantage of unique longitudinal data, we provide the first characterization of what college students believe at the time of entrance about their final major, relate these beliefs to actual major outcomes, and provide an understanding of why students hold the initial beliefs about majors that they do. The data collection and analysis are based directly on a conceptual model in which a student's final major is best viewed as the end result of a learning process. We find that students enter school quite optimistic about obtaining a science degree, but that relatively few students end up graduating with a science degree. The substantial overoptimism about completing a degree in science can be attributed largely to students beginning school with misperceptions about their ability to perform well academically in science.

Key words: Education, College major, Dropout, Science, Learning, Expectations data.

JEL Codes: I2, J24

1. INTRODUCTION

It is well known that lifetime earnings are influenced strongly by a student's college major.[1] Further, certain disciplines, such as math and other sciences, are viewed as being particularly important for the future path of the economy (COSEPUP, 2007). Then it is not surprising that policymakers often express a desire to influence the number of graduates in certain majors.[2] Nonetheless, much remains unknown about how college majors are determined. In this article, we use a unique combination of survey and administrative data to provide new evidence about this issue.

We provide the first characterization of what students believe at the time of college entrance about whether they will graduate and their majors at graduation if they do graduate, relate these

1. See Daymont and Andrisani (1984), Grogger and Eide (1995), Hamermesh and Donald (2008), Loury (1997), Loury and Garman (1995), and James *et al.* (1989).

2. This desire has received much attention. See, for example, "Why Science Majors Change Their Minds (It's just so darn hard)." *NY Times*, November 4, 2011.

beliefs to realized graduation and major outcomes, and, of interest given important differences found between beliefs and outcomes, provide an understanding of why students hold the "initial" beliefs about outcomes that they do. Our motivation for focusing on initial beliefs comes from the reality that, when a student's final major is viewed as the end result of a process in which he learns about the quality of his match with each possible major (Arcidiacono, 2004), an understanding of initial beliefs serves as the necessary foundation for any comprehensive understanding of how college majors are determined.[3] The fundamental importance of initial beliefs is apparent if one considers that institutional rules related to flexibility in major choice, the timing of major declaration, and the nature/number of required and elective courses vary dramatically across institutions (both within and across countries) and represent a natural policy lever. In theory, these institutional rules may have a substantial impact on a student's ability/propensity to experiment with different subject areas and, as a result, may have a large impact on the distribution of major outcomes. However, in practice, the impact of institutional rules will depend critically on beliefs at entrance. If students are well informed about their abilities or other major-specific factors that influence their choice of major, then institutional rules may have little effect. However, if much uncertainty exists about these factors or if systematic bias exists in beliefs about these factors, then institutional rules may have an important impact.

That virtually nothing is known about initial beliefs can be attributed primarily to the difficulty of obtaining data that are ideal for providing the type of evidence described in the previous paragraph. For illustration focusing on college major (rather than dropout), characterizing beliefs at entrance about major outcomes is arguably best achieved through the use of carefully worded survey questions that allow a student to express uncertainty about the object of ultimate interest—his "final" major at the time of graduation. Relating these initial beliefs to realized major outcomes requires administrative data at the time that each student finishes school. Finally, natural guidance for an investigation into why students hold particular initial beliefs about their final major comes from a learning model of major choice in which students enter school with beliefs about major-specific factors (*e.g.* the grade performance or future income associated with each major) and in which uncertainty about one's final major arises because individuals have the opportunity to update their beliefs about these factors before making a final major decision. This conceptual framework suggests the need for substantial additional data which, for example: (1) characterize beliefs about major-specific factors at entrance, (2) describe the extent to which beliefs about major-specific factors are updated during school, (3) provide an understanding of the determinants of these updates, and (4) describe what beliefs about major-specific factors would be consistent with, for example, actual observed grade performance in the sample or actual observed future income in the sample. Thus, a thorough examination of beliefs at the time of entrance requires both detailed survey data and detailed administrative data, with observations being needed at very specific times over a relatively long period.

In this article, we take advantage of unique survey and administrative data that we collected specifically to address the data challenges above. As described in Section 2, the data come from the Berea Panel Study (BPS), a longitudinal survey of students at Berea College that was initiated to allow an in-depth study of a variety of decisions and outcomes in higher education. Generally, the survey data from the BPS are well-suited for the type of analysis in this article because the data collection was guided closely by theoretical models of learning. More specifically, the data contain two unique features that are of central importance for this study. First, the survey is unique among surveys of college students in its frequency of contact with respondents; each student was surveyed approximately twelve times each year while in school and annually thereafter,

3. For other work recognizing the importance of learning in determining educational outcomes see, for example, Manski (1989); Altonji (1993); Carneiro *et al.* (2003); Cunha *et al.* (2005).

with, importantly, the first survey taking place immediately before the beginning of the student's freshman year. Secondly, taking advantage of recent methodological advances in the elicitation of beliefs (Dominitz and Manski, 1996, 1997; Dominitz, 1998; Manski, 2004), the BPS was perhaps the first sustained longitudinal survey whose initiation was motivated primarily by a desire to supplement detailed administrative data with individual-level expectations data.

From a policy standpoint, it is important to identify the stage at which students tend to move away from particular disciplines, such as math and other sciences, that are often viewed as being particularly important. On one hand, at the time of college entrance, many students may have already decided that these disciplines are either not of interest or are not a good match academically. On the other, it is possible that students tend to enter college quite open to these disciplines, but often realize after entrance that these disciplines are not a good match. Existing data limitations imply that very little is known about the relative importance of these two explanations.

Section 3, which characterizes beliefs about outcomes at entrance and compares these beliefs to realized outcomes, provides new evidence about this issue. We consider eight possible outcomes that could be realized as of the time a student leaves college by aggregating individual majors into seven major "groups" (hereafter often referred to simply as "majors") and by, as in Arcidiacono (2004), adding dropout as the eighth possible outcome. Our primary survey questions elicit, at the time of entrance, the individual-specific probability (belief) that each of these final outcomes will be realized (hereafter, referred to simply as outcomes). We find that much uncertainty exists at entrance. For example, on average, a student assigns a probability of only 0.310 to the realized outcome that is observed for him in the administrative data. This occurs, in part, because, while students are very open at entrance to majoring in science, relatively few students end up graduating with a science major. Specifically, while the proportion of students who begin school believing that a degree in science is their most likely outcome (0.20) is higher than the proportion for any other major and while the average perceived probability of completing a science degree (0.16) is almost as high as the average perceived probability for any other major, only 0.07 of students complete a degree in science. This result highlights a common theme found throughout the article—that the science major is an outlier among majors in many fundamental ways.

Then, a natural question is what makes students often realize that their match is worse for degrees in math or science than they expected at entrance. Guided by the conceptual learning model described above, Section 4 provides an understanding of why students begin school with the beliefs about outcomes that they do. The model that we estimate places a particular emphasis on understanding initial beliefs about science and can be used to examine how beliefs about outcomes would change if a student's actual initial beliefs about grade performance and future income were replaced by: (1) the beliefs about grade performance and future income that the student held in the later stages of college or (2) beliefs about grade performance and future income that correspond to actual distributions of grade performance and future income. We find that the substantial overoptimism about completing a degree in science can be attributed largely to students beginning school with misperceptions about their ability to perform well academically in science. An intermediate step in reaching this conclusion involves examining the process by which students update their beliefs about grade performance in each major. We again find evidence that science deviates from other majors. By and large, the primary way to learn about one's ability in science is to take science classes, while, for majors j other than science, students tend to collect much of their information about their ability in j from classes taken outside of the major.

This article builds on work in Stinebrickner and Stinebrickner (2012) and Stinebrickner and Stinebrickner (forthcoming) which studies, in detail, the process by which students update their beliefs about their overall grade performance/academic ability and finds that dropout is strongly related to what they learn about these factors. Thus, our work is related to a literature which

has begun to recognize the usefulness of expectations data in the education context (see, *e.g.* Kaufmann, 2009; Attanasio and Kaufmann, 2009; and Jensen, 2010) and a literature that has recognized the importance of understanding the process by which beliefs evolve (see, *e.g.* Delavande, 2008; Dominitz and Hung, 2009, Armantier *et al.*, 2012). In terms of research on college major, our work is most related to that of Zafar (2008, 2013) and Arcidiacono *et al.* (2012a) who, using data collected much more recently than the data in this article, illustrate that expectations data can allow a useful next step beyond what is possible using traditional data.[4] This article also contributes to a growing body of literature on the STEM (Science, Technology, Engineering, Math) pipeline. Much has been written about the reasons why students leave majors in these fields to pursue other majors, with discussion often focusing on the possibility that students may learn that their grade performance in STEM majors will be worse than expected.[5] This article contributes to this literature by providing the first direct evidence detailing how beliefs about grade performance in various majors actually change over the entire period that a student is in school and how these changes in beliefs relate to one's choice of major.

From a methodological standpoint, our work contributes heavily to an understanding of the value of expectations data. Given the longitudinal nature of our data, we are able to provide new evidence about how beliefs about factors that might influence a decision evolve. Further, the nature of our data/exercise complements Blass *et al.* (2010), which represents some of the only other previous work using survey questions that allow agents to express uncertainty about a choice that will be made in the future.[6] From a descriptive standpoint, because we collected longitudinal data and because we study a real-world situation in which the future decision is actually observed, we are able to, for example: (1) characterize the amount of uncertainty about outcomes that is present at the time of entrance and (2) examine whether, on average, students have correct beliefs about outcomes at the time of entrance. From a modelling standpoint, because in our context it is reasonable to take a stand on the underlying outcome-specific factors that may cause uncertainty about a final outcome (*e.g.* academic performance/ability and future income) and because we can use additional survey questions to characterize the person-specific distributions representing beliefs about these factors, we are able to relax the assumption in Blass *et al.* (2010) that the amount of uncertainty about underlying factors is unobservable and homogeneous across people.

In addition, our work relates very directly to a literature which notes that correct policy decisions often hinge on determining whether differences in outcomes across individuals are indicative of individual uncertainty or are indicative of heterogeneity that is known to individuals (Cunha *et al.*, 2005; Browning and Carro, 2007). Our survey question elicits uncertainty about outcomes directly at the time of college entrance, and our approach for understanding how people

4. These papers do not provide evidence related to the central motivation for this article—that obtaining a comprehensive understanding requires viewing the final major as the end result of a learning process which starts at the time of entrance. This is the case both because these projects do not involve the type of longitudinal aspect (starting at the time of entrance) that is present in the BPS and because the survey instruments used in these other projects did not allow students to express uncertainty about their final major.

5. This could occur for a variety of reasons including differences in faculty expectations between high school and college (see, *e.g.* Daempfle, 2003; Federman, 2007), or because grades tend to be inflated in non-STEM majors (see, *e.g.* Sabot and Wakefield-Linn, 1991; Rask, 2010). Related to this literature, much has also been written concerning why certain groups, for example minorities or females, are less likely to finish with STEM majors. (See, *e.g.* Ost, 2010; Arcidiacono *et al.*, 2012b; Arcidiacono and Koedel, 2012; Arcidiacono *et al.*, 2013)

6. Recent work by Wiswall and Zafar (2011) follows this article in collecting beliefs about major in probabilistic form, with a focus on examining how these beliefs are updated during a survey session when students are provided with new information about the benefits of particular majors. For other work on college major that does not use expectations data see, for example, Beffy *et al.* (2013) and Montmarquette *et al.* (2002).

arrive at these initial beliefs explicitly takes into account uncertainty in the outcome-specific factors that influence decisions.

2. THE BPS AND THE SAMPLE USED IN THIS ARTICLE

Designed and administered by Todd Stinebrickner and Ralph Stinebrickner, the BPS is a multi-purpose longitudinal survey that takes place at Berea College and elicits information of relevance for understanding a wide variety of issues in higher education, including those related to dropout, major, time-use, social networks, peer effects, and transitions to the labour market. The school is unique in certain respects that have been discussed in previous work. For example, the school has a focus on providing an education to students from low income backgrounds and offers a full tuition subsidy to all students. Both family income and tuition costs have the potential to impact major choice. For example, it seems likely that students from low income backgrounds, who are less likely to have attended high-quality schools and are less likely to have college-educated parents, may be especially uninformed about what will take place during college. As such, it is necessary to be appropriately cautious about the exact extent to which the results from our case study would generalize to other demographic groups or to other specific institutions. However, important for the notion that the basic lessons from our work are likely to be useful for thinking about what takes place elsewhere, Berea operates under a standard liberal arts curriculum and the students at Berea are similar in academic quality to, for example, students at the University of Kentucky (S&S, 2008a). Further, in Section 3 we find that patterns of major-switching at Berea are similar in spirit to those found by Arcidiacono (2004) for students in the NLS-72.

The BPS consists of two cohorts. Baseline surveys were administered to the first cohort (the 2000 cohort) immediately before it began its freshman year in the fall of 2000 and baseline surveys were administered to the second cohort (the 2001 cohort) immediately before it began its freshman year in the fall of 2001. In addition to collecting detailed background information, the baseline surveys were designed to take advantage of advances in survey methodology (see, *e.g.* Barsky *et al.*, 1997; Dominitz, 1998; Dominitz and Manski, 1996, 1997) in order to collect expectations towards uncertain outcomes and the factors that might influence these outcomes. Substantial follow-up surveys that were administered at the beginning and end of each subsequent semester document how expectations towards uncertain outcomes and the factors that might influence these outcomes have changed. Students are surveyed annually after leaving school.

Combining the 2000 and 2001 cohorts, our full sample consists of the 655 students (roughly 85% of all entering students in these two cohorts) who answered the baseline survey that took place immediately before the start of the students' freshman year. The BPS survey data are linked to administrative data which serve several purposes in the article. First, the administrative data allow us to document whether a student graduated and what his major was if he graduated. Secondly, the administrative data contain information about a variety of observable characteristics, X_i. Among these, we take advantage of a student's sex and his/her score on the American College Test (ACT). For the full sample, the proportion of students that are male (MALE) is 41.8%. The average (std. deviation) score on the ACT math test (MACT) is 21.81 (4.14), and the average (std. deviation) score on the ACT verbal test (VACT) is 23.37 (4.51). In earlier work, we found that these test scores are generally similar to those observed for students at the University of Kentucky and the University of Tennessee (S&S, 2008a). Finally, as discussed in detail in Section 4, the administrative data contain transcript information which allow us to observe each class that a student took while in school and the grade that he received in that class.

3. BELIEFS ABOUT OUTCOMES AT ENTRANCE

In terms of outcomes, our object of interest is a student's choice at the time of departure from college. The largest set of possible outcomes we consider is given by the choice set $J^D = \{\text{DROP}, \text{SCI}, \text{AG}, \text{BUS}, \text{ED}, \text{HUM}, \text{PRO}, \text{SS}\}$. The first outcome (DROP) indicates whether a person drops out of college before graduation. The remaining outcomes represent the aggregated major groups (hereafter often referred to simply as majors) that could be chosen if a student graduates: Science including Math (SCI), Agriculture and Physical Education (AG), Business (BUS), Education (ED), Humanities (HUM), Professional programs (PRO), and Social Science (SS). The set of individual majors that appear in each of these major groups is shown in survey Question 1 (Note: all survey questions appear in Appendix A). We have used the D superscript when denoting the choice set J^D to indicate that the outcome DROP is part of the choice set. Some of our analyses involve somewhat different specifications of the choice set, including cases where we do not include DROP. We introduce the relevant choice sets as they are utilized.

Understanding how open students are to degrees in certain disciplines, such as math and other sciences, at the time of entrance is of direct policy importance. Blass *et al.* (2010) describe the problems that can arise when a respondent is forced to "state" a choice in a context in which uncertainty exists about a decision that will take place in the future. Then, importantly, at the time of entrance, the first column of survey Question 1 allows a respondent to express uncertainty about his final choice by reporting the per cent chance of ending up with each major (conditional on not dropping out) and Question 2a allows a respondent to express uncertainty about whether he will drop out by reporting the per cent chance of graduation.[7]

We refer to student i's reported probability at the time of entrance of ending up with an outcome of $j \in J^D$ as i's initial (perceived) probability of choosing j and denote this probability $Pr_{i,j}$. $Pr_{i,DROP}$ is calculated by subtracting the per cent chance in Question 2a from 100 (and dividing by 100). Because students can only finish with a particular major if they do not drop out, for all $j \neq$ DROP, $Pr_{i,j}$ is constructing by multiplying the response for j in survey Question 1 by $1 - Pr_{i,DROP}$ (and dividing by 100). Juster (1966) and Manski (1990) reasoned that, when asked to declare the outcome of a future decision in a case where uncertainty will be resolved before the final decision is made, survey respondents will tend to state the alternative with the highest probability as of the time of the survey. Hereafter, we follow this literature by referring to the most likely outcome at the time of entrance (*i.e.* $\arg\max_{j \in J^D} Pr_{i,j}$) as the "stated" outcome $STATE_i$, although we note that this title is somewhat of a misnomer since we construct the most likely outcome ourselves from Questions 1 and 2a.

We denote the random variable representing a student's outcome as J, with j representing a generic possible realization, and j* referring to the student's actual choice (*i.e.* the realization of J that is observed in the administrative data). As discussed in the introduction, the amount of uncertainty that exists about outcomes at entrance is of central importance for policy. Because observing changes in declared major is not sufficient to understand how much uncertainty exists at the time of entrance, the responses to Questions 1 and 2 represent the first opportunity to provide evidence about this uncertainty. One simple measure of the actual amount of uncertainty that exists about outcomes at entrance is $Pr_{i,j*}$, the probability that student i assigned at entrance to the outcome $j^* \in J^D$ that was eventually realized. The sample average of 0.310 indicates that

7. The "per cent chance" questions were answered after students completed classroom training which, among other things, discussed this type of question in non-education contexts. For this article, "illegitimate" responses in the first column of Question 1 are responses where the sum of the per cent chances was more than 110 or less than 90. For sums that were between 90 and 110, but not equal to 100, we adjusted each per cent chance proportionally to make the sum equal 100.

students do not have a particularly good sense of what outcome will be realized, although we note that students are not entirely uninformed since 0.310 is approximately double the probability that would be observed if either: (1) students put equal probability on all eight choices or (2) students randomly chose one of the eight outcomes and assigned a probability of 1.0 to that outcome.

The relatively low sample average of $Pr_{i,j*}$ indicates that an administrator/counselor would not be very certain about a student's final major on the basis of answers to Questions 1 and 2a. One way to examine whether students understand how much uncertainty exists is to compare the sample average of $Pr_{i,j*}$ to the sample average $\sum_j Pr_{i,j}Pr_{i,j}$, which represents the conceptual equivalent of asking students the following survey question: "At the time you leave college we are going to find what outcome you chose (from administrative data) and then see what probability you assigned to this outcome at entrance (from survey data). What do you expect this probability to be?"[8] The fact that the sample average of $\sum_j Pr_{i,j}Pr_{i,j}$ is 40% larger than the sample average of $Pr_{i,j*}$ (0.433 versus 0.310), with a test of the null hypothesis that the population means are equal producing a t-statistic of 10.75, implies that students are not fully aware of how much actual uncertainty exists.[9]

One possible reason for the difference observed between the sample averages in the previous paragraph is that students may be, on average, too optimistic about the probability of having certain outcomes and too pessimistic about the probability of having other outcomes. The existence of systematic biases can be informative for policymakers who are interested in why more students do not complete degrees in certain areas. For example, a finding that students are very optimistic about science at entrance might help alleviate a concern that students lose interest in science majors at earlier (pre-college) stages of schooling, but might raise the possibility that students learn after college entrance that they are not well-prepared for these majors.

The first panel of Table 1 examines systematic biases that exist for the population as a whole. For each outcome $j \in J^D$, Column 1 shows the proportion of students that have j as their stated major and Column 2 shows the sample average of $Pr_{i,j}$. Important for policy, students are very open to a major in science at the time of entrance; the proportion of students in the sample who begin school believing that Science is the most likely major, 0.198, is higher than the proportion in the sample for any other major (Column 1) and the average perceived probability of Science in the sample, 0.156, is higher than the average perceived probability for all but one other major (Column 2). However, despite this optimism, Column 3 shows that only 0.074 of students end up with an outcome of Science, and the only majors with proportions lower than that of SCI are the speciality majors AG and ED. This motivates our work in Section 4 which attempts to understand why individuals hold the beliefs about outcomes that they do at entrance and to understand why these beliefs are found to be systematically wrong for Science.

The results in Table 1 indicate that Science is an outlier among majors in the sense that it is very "popular" at entrance, but is ultimately chosen by only a relatively small number of students. This is formalized in Column 4, which shows the difference between Column 3 and Column 1, and Column 5, which shows the difference between Column 3 and Column 2. Column 4 shows that, for j = SCI, the sample proportion of students who have an outcome of SCI (0.074) is 0.123 smaller than the sample proportion that initially believed that SCI was their most likely major

8. Another measure of relevance is $\max_j Pr_{i,j}$. We find a sample average of 0.61.

9. Both dropout and choice of major conditional on graduating contribute to the differences between the sample averages of $\sum_j Pr_{i,j}Pr_{i,j}$ and $Pr_{i,j*}$. When we consider the choice set {DROP, GRAD}, the sample averages are 0.813 and 0.608, respectively, and a test of mean equality produces a t-statistic of 13.43. When, for the subset of students who graduate, we consider the choice set {AG, BUS, ED, HUM, SCI, PRO, SS}, the sample averages are 0.501 and 0.368, respectively, and a test of mean equality produces a t-statistic of 8.38.

TABLE 1
Outcomes and beliefs about outcomes

	1	2	3	4	5
Full sample, $j \in J^D$ $n = 652$	$1(STATE_i = j)$ Mean (std. err.)	$Pr_{i,j}$ Mean (std. err.)	$1(j^* = j)$ Mean (std. err.)	(3)–(1) Mean (std. err.)	(3)–(2) Mean (std. err.)
j = DROP	0.051 (0.008)	0.134 (0.005)	0.375 (0.018)	0.324 (0.020)	0.241 (0.019)
j = SCI	0.198 (0.015)	0.156 (0.009)	0.074 (0.010)	–0.123 (0.013)	–0.081 (0.010)
j = AG	0.101 (0.011)	0.092 (0.006)	0.047 (0.008)	–0.054 (0.011)	–0.045 (0.008)
j = BUS	0.125 (0.012)	0.122 (0.007)	0.074 (0.010)	–0.050 (0.012)	–0.047 (0.009)
j = ED	0.074 (0.009)	0.084 (0.006)	0.053 (0.008)	–0.021 (0.010)	–0.031 (0.008)
j = HUM	0.185 (0.014)	0.164 (0.009)	0.164 (0.014)	–0.020 (0.014)	–0.0001 (0.012)
j = PRO	0.137 (0.013)	0.121 (0.007)	0.128 (0.013)	–0.009 (0.015)	–0.006 (0.012)
j = SS	0.124 (0.012)	0.124 (0.007)	0.091 (0.011)	–0.033 (0.013)	–0.032 (0.011)
$\sum$ of previous rows	1	1	1.01	0.01	0.01
Graduate sample, $j \in J^{ND}$ $n = 409$	$1(STATE_i = j)$ Mean (std. err.)	$Pr_{i,j}$ Mean (std. err.)	$1(j^* = j)$ Mean (std. err.)	(3)–(1) Mean (std. err.)	(3)–(2) Mean (std. err.)
j = SCI	0.215 (0.020)	0.183 (0.013)	0.119 (0.106)	–0.095 (0.017)	–0.063 (0.013)
j = AG	0.091 (0.014)	0.101 (0.009)	0.075 (0.013)	–0.015 (0.013)	–0.026 (0.010)
j = BUS	0.131 (0.016)	0.135 (0.010)	0.119 (0.016)	–0.011 (0.013)	–0.016 (0.012)
j = ED	0.072 (0.012)	0.091 (0.007)	0.085 (0.013)	0.013 (0.012)	–0.006 (0.011)
j = HUM	0.199 (0.019)	0.191 (0.013)	0.264 (0.021)	0.064 (0.017)	0.072 (0.016)
j = PRO	0.134 (0.016)	0.142 (0.011)	0.205 (0.019)	0.070 (0.019)	0.063 (0.017)
j = SS	0.155 (0.017)	0.152 (0.010)	0.146 (0.017)	–0.008 (0.108)	–0.006 (0.016)
$\sum$ of previous rows	1	1	1.017	0.017	0.017

Notes: Each entry shows a sample mean (standard error of the sample mean). The first panel includes the full sample. The second panel includes only students who graduated and removes the DROP option.
Column 1: This is roughly the proportion of students who have $STATE_i$=j. To be more exact, it is the sample average of a variable which has a value of one if $STATE_i$=j. But, if two majors j and k are tied with the highest probability (*i.e.* $Pr_{i,j} = Pr_{i,k}$) we assume that the probability is 0.50 that $STATE_i$=j and 0.50 that $STATE_i$=k.
Column 2: Sample average of perceived initial probabilities $Pr_{i,j}$.
Column 3: Sample average of indicator variable which has a value of one if student's outcome j* is j. The "$\sum$ of previous rows" value of 1.01 in Column 3 indicates that about 1% of students have majors in two different major groups.
Column 4 is Column 3 – Column 1.
Column 5 is Column 3 – Column 2.

(0.198). This represent a decrease of (0.123/0.198)% = 62%. For all majors j ≠ SCI (and j ≠ DROP), we reject the null hypothesis that the difference in Column 4 is the same for j as it is for SCI, with *t*-statistics from the tests ranging from 3.589 (AG) to 5.576 (ED). Column 5 shows that, for j = SCI, the sample proportion of students who have an outcome of SCI (0.074) is 0.081 smaller than the average perceived initial probability of choosing SCI (0.156). This represents a decrease of (0.081/0.156)% = 52%. For all majors j ≠ SCI (and j ≠ DROP), we reject the null hypothesis that the difference in Column 5 is the same for j as it is for SCI, with *t*-statistics from the tests ranging from 2.377 (j = BUS) to 5.098 (j = PRO).

Row 1 of Table 1 shows that students are also substantially overoptimistic about graduation at entrance. While only 0.051 of students believe that DROP is the most likely outcome and, on average, students assign a probability of 0.134 to DROP, in reality 0.375 of students have DROP as their outcome. This overoptimism about dropout has been explored in our recent work (S&S, 2012; S&S, 2013). In terms of its implications for major choice, we note that this overoptimism contributes to the numbers in Columns 4 and 5 of Table 1 being negative for all majors. It is then natural to wonder whether the fact that the entries in Columns 4 and 5 are substantially more negative for SCI than for the other majors might be an indication of especially large dropout rates for students who start school believing that SCI is very likely. We find no evidence of this

TABLE 2
Proportion with actual outcomes j stratified by stated major*

	1	2	3	4	5	6	7	8
Full sample, j∈J^D	$STATE_i$ = DROP	$STATE_i$ = SCI	$STATE_i$ = AG	$STATE_i$ = BUS	$STATE_i$ = ED	$STATE_i$ = HUM	$STATE_i$ = PRO	$STATE_i$ = SS
Outcome j* $n = 655$	$n = 33.67$	$n = 130.08$	$n = 66.75$	$n = 82.17$	$n = 49$	$n = 121.25$	$n = 90.25$	$n = 81.83$
$j^* =$ DROP	0.529	0.335	0.458	0.389	0.414	0.335	0.407	0.297
$j^* =$ SCI	0.059	0.311	0.014	0	0.02	0	0.038	0.012
$j^* =$ AG	0.044	0.038	0.284	0.01	0	0.01	0.022	0.024
$j^* =$ BUS	0.103	0.042	0.044	0.377	0	0.016	0.011	0.036
$j^* =$ ED	0.029	0.046	0	0.046	0.384	0.016	0.022	0.012
$j^* =$ HUM	0.059	0.049	0.029	0.912	0.057	0.555	0.083	0.15
$j^* =$ PRO	0.074	0.103	0.127	0.066	0.102	0.012	0.387	0.152
$j^* =$ SS	0.099	0.096	0.034	0.022	0.02	0.056	0.049	0.338
$\sum$ of previous rows	1	1.023	1	1	1	1	1.022	1.024
Grad sample, j ∈ JND		$STATE_i$ = SCI	$STATE_i$ = AG	$STATE_i$ = BUS	$STATE_i$ = ED	$STATE_i$ = HUM	$STATE_i$ = PRO	$STATE_i$ = SS
Outcome j^* $n = 409$		$n = 81.66$	$n = 37.41$	$n = 53.58$	$n = 29.67$	$n = 81.67$	$n = 55.0$	$n = 63.41$
SCI		0.464	0.026	0	0.033	0	0.072	0.031
AG		0.056	0.507	0.018	0.033	0.012	0.036	0.031
BUS		0.065	0.086	0.629	0	0.024	0.018	0.051
ED		0.067	0.089	0.071	0.634	0.024	0.036	0.031
HUM		0.084	0.053	0.139	0.095	0.824	0.136	0.21
PRO		0.152	0.24	0.102	0.168	0.03	0.654	0.197
SS		0.141	0.075	0.037	0.033	0.083	0.081	0.478
$\sum$ of previous rows		1.033	1	1	1	1	1.036	1.031

Notes: For the sample stratified by $STATE_i$, the table shows the proportion of students with various possible outcome j*. The sample sizes can be fractions because if two majors j and k are tied with the highest probability (*i.e.* $Pr_{i,j} = Pr_{i,k}$) the person would be counted as 0.5 of an observation in each of the majors. The numbers in the SCI, PRO, and SS columns add up to more than one because at least one person with each of these stated majors has more than one actual major j*. The first panel includes the full sample. The second panel includes only students who graduated and removes the DROP option.

in the first panel of Table 2, which shows the proportion of students with each possible outcome $j \in J^D$ after stratifying the sample by $STATE_i$. Row 1 shows that the sample dropout proportion of students who have a stated major of SCI (0.335) is lower than the sample dropout proportion of students from all but one of the other stated majors. This is relevant for a policy concern that encouraging students to try SCI may lead some students to become overly pessimistic about their general academic ability. The results in the first panel of Table 2 are similar in spirit to those found in Arcidiacono (2004) for the NLS-72.[10]

Thus, the fact that the entries in Columns 4 and 5 of the first panel of Table 1 are larger for SCI than for other majors arises because, despite considerable initial optimism about SCI, graduates very often end up in majors other than SCI. This is formalized in the second panel of Table 1 which, using the subsample of students who graduate, repeats the first panel for a choice

10. An exact comparison is somewhat difficult since Table 3 of Arcidiacono (2004) shows transitions during the first two years of college and also uses a somewhat smaller group of majors. Nonetheless, a rough comparison shows strong similarities. He finds, for example, that, among students who start in science, 28%, 42%, 8%, 19%, and 2%, respectively, have the majors of DROP, SCI, BUS, SS/PRO (combined), and ED, respectively, after two years. If, in Column 2 of Table 2 we aggregate SS and PRO and aggregate AG and SCI, the comparison percentages from our data are 34%, 35%, 4%, 25%, and 5%, respectively. The DROP and SCI percentages would become even closer by the time students in his data complete school.

set $J^{ND} = \{SCI, AG, BUS, ED, HUM, PRO, SS\}$ which does not contain DROP (ND indicates "no dropout"). As expected, the entries in Columns 4 and 5 are, as in the first panel, larger (in magnitude) for SCI than for the other majors. For all $j \neq SCI$, we reject the null hypothesis that the difference in Column 4 is the same for j as it is for SCI (-0.095), with t-statistics from the tests ranging from 3.072 (SS) to 6.285 (HUM). For all $j \neq SCI$, we reject the null hypothesis that the difference in Column 5 is the same for j as it is for SCI (-0.063), with *t*-statistics ranging from 2.049 (AG) to 6.055 (HUM).

It is worth delving further into why students who graduate tend to end up in majors other than SCI. Among those who graduate, the number of students who have a realized major of j depends on both: (1) the true probability of having a realized major of j conditional on having a stated major of j (*i.e.* the probability of "staying" in j) and (2) the true probability of having a realized major of j conditional on having a stated major of $k \neq j$ (*i.e.* the probability of "changing" to j). These numbers can be computed for the sample from the second panel of Table 2 which is analogous to the first panel but uses the graduate subsample and the choice set J^{ND} (as in the second panel of Table 1). With respect to (1), Figure 1A shows that the proportion in the sample who stay in j is lower when j = SCI than when j is any of the other majors. With respect to (2), Figure 2A shows that the proportion in the sample who change to j is lower when j is Science than when j is any of the other majors. Misperceptions in the Pr_{ij}'s will exist if beliefs about the probability of staying in j and beliefs about the probability of changing to j do not correspond to the actual probabilities in Figures 1A and 2A. In contrast to Figure 1A, Figure 1B indicates that, at entrance, students who have $STATE_i$ = SCI believe they are as likely to stay in their stated major as are

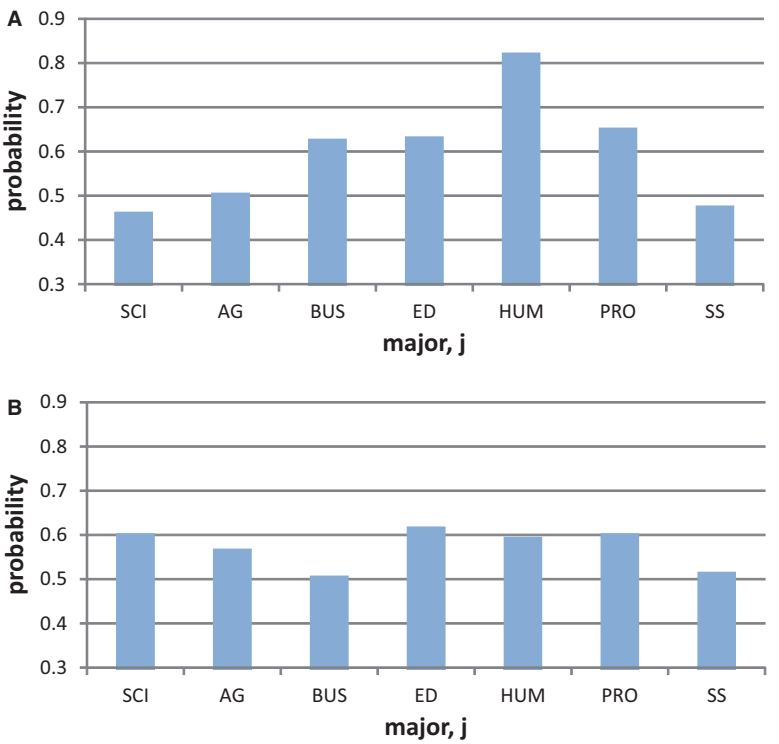

FIGURE 1

(A) True probability of staying in stated major j. (B) Average perceived probability at t=1 of staying in stated major j.

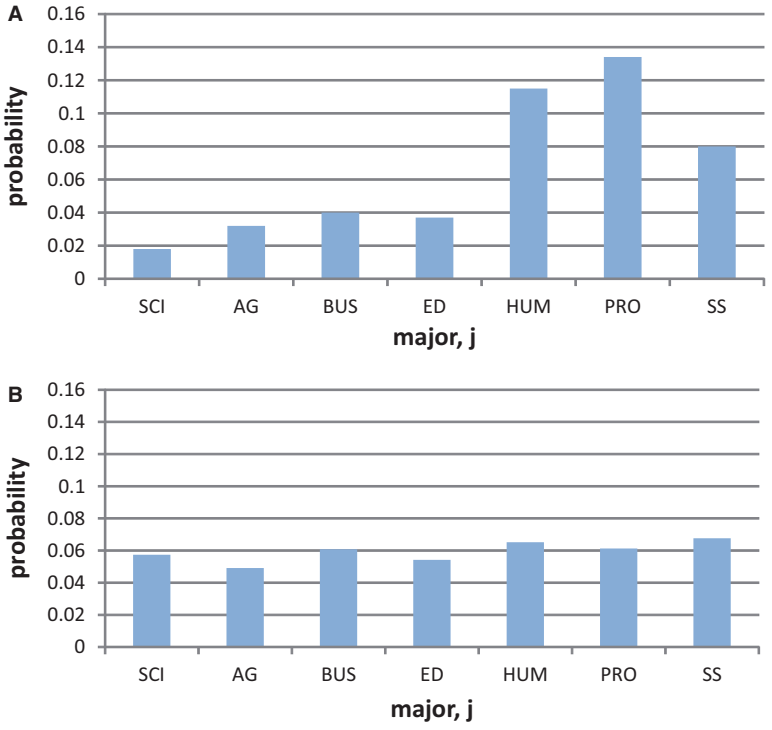

FIGURE 2

(A) True probability of changing to major j. (B) Average perceived probability at t=1 of changing to major j.

students whose stated major is something other than SCI. In contrast to Figure 2A, Figure 2B shows that students believe they are as likely to change into SCI as they are to change into any other major.[11] Thus, Figures 1A and 2A show that starting in Science is close to necessary but far from sufficient for having a final major of Science, but Figures 1B and 2B show that students do not fully realize that this is the case.

There are two primary, potential explanations for why the movements out of SCI seen in Columns 4 and 5 of Table 1 are large relative to movements out of other majors. One possible explanation is that science has a particularly large option value. Under this explanation, students might find it worthwhile to experiment with science at the start of school even in cases where they plan to leave science unless they perform substantially better than expected. Then, the evidence in Figure 1B—that students starting in SCI are not relatively pessimistic about finishing in SCI—does not provide direct support for this explanation. Further, if under an option value explanation majors with high option values tend to be tried first or not at all, then the evidence in Figure 2B—that students are as optimistic about changing into SCI as they are to changing into other majors—also does not provide direct support of the importance of the option value explanation. Section 4 provides evidence about a second possible explanation—that students tend to be especially misinformed about their ability to perform well in science.

11. For each $j \in J^{ND}$, Figure 1B shows the sample average of $Pr_{i,j}$ for all students who have $STATE_i = j$. For each $j \in J^{ND}$, Figure 2B shows the sample average of $Pr_{i,j}$ for all students who have $STATE_i \neq j$.

4. UNDERSTANDING BELIEFS ABOUT OUTCOMES AT ENTRANCE

4.1. *Overview*

Section 3 uncovers three reasons why the actual number of college graduates majoring in SCI (*i.e.* have $j^* = SCI$) is much smaller than the number that would be expected given the beliefs about outcomes ($Pr_{i,j}: j \in J^D$) observed at entrance. First, students are overoptimistic about the probability that they will graduate. Secondly, students who have a stated major of SCI are too optimistic about the probability that they will remain in SCI. Finally, students who have a stated major other than SCI are too optimistic about the probability that they will change to a major in SCI.

To provide an understanding of these findings, we estimate a model describing why individuals hold the initial beliefs about outcomes that they do. In Subsection 4.2, we describe the model of initial beliefs, which takes into account that, before making his final decision about his dropout/major outcome at some future time t^*, student i will have the opportunity to resolve some of the uncertainty about major-specific factors $Factor_{i,j}$ that influence the utility that he receives from each of his alternatives j. When not all uncertainty about $Factor_{i,j}$ can be resolved by t^*, the expected utility that i will attach to alternative j when making his decision at t^* will depend on the mean of the distribution describing i's beliefs at t^* about $Factor_{i,j}$. However, at time t=1, when initial beliefs about outcomes, $Pr_{i,j}$, are reported but uncertainty about $Factor_{i,j}$ has not yet been resolved, this "t^*-mean-for-j" will not be known. This implies that a student cannot know for sure which alternative he will choose at t^*. Instead, he uses the distributions describing his beliefs at t=1 about what the t^*-mean-for-j will turn out to be for each j to construct his report of $Pr_{i,j}$. The empirical difficulty is that, for reasons discussed in Subsection 4.2, these t=1 distributions cannot be observed directly in the data. Instead, we construct these t=1 distributions by taking advantage of several unique features of our data. We first specify the updating process by which a person will arrive at his t^*-mean-for-j. In this process, the t^*-mean-for-j is a function of i's t=1-mean-for-j (*i.e.* the mean of the distribution describing i's beliefs at t=1 about $Factor_{i,j}$) and a noisy signal of $Factor_{i,j}$. Access to data documenting the t^*-mean-for-j, the t=1-mean-for-j, and the observed value of the signal allows the parameters of the updating process to be estimated. Given the estimated updating process, uncertainty at t=1 about what the t^*-mean-for-j will turn out to be comes from uncertainty about what value of the signal will be realized. Then, individual-specific information characterizing beliefs about the distribution of the signal allows us to construct the desired object—the distribution describing i's beliefs at t=1 about what the t^*-mean-for-j will turn out to be.

In practice, we focus on two alternative-specific factors, $Factor_{i,j}$. In Subsection 4.3 we describe the construction of i's beliefs at t=1 about what the t^*-mean-for-j will turn out to be when $Factor_{i,j}$ is i's average semester grade performance associated with j. In Subsection 4.4, we describe the construction of i's beliefs at t=1 about what the t^*-mean-for-j will turn out to be when $Factor_{i,j}$ is i's average annual future income associated with j. Subsection 4.5 provides some descriptive evidence that these two factors are likely to influence a student's decision. Subsection 4.6 describes the identification and estimation of the model. Subsection 4.7 discusses parameter estimates and quantifies how much different initial beliefs about outcomes would have been if students had started school with more accurate beliefs about the major-specific factors. Ultimately, our article provides direct evidence that initial beliefs about an outcome of SCI are, on average, incorrect largely because students learn after entrance that their initial beliefs about the major-specific factors associated with SCI were, on average, incorrect.

4.2. *An estimable model of initial beliefs about major*

Student i enters college (t=1) uncertain about his outcome j^*. He envisions a time t^* at which he will finalize his outcome by choosing the optimal j from the set J^D. Taking as given the

courses a student is enrolled in during the early stages of college,[12] our goal is to specify and estimate a model which explains why a student begins school with his set of beliefs $Pr_i = (Pr_{i,j}: j \in J^D)$.

Denote as $U^j(AGPA_{i,j}, AINC_{i,j}, X_i, v_{i,j}, \varepsilon_{i,j})$ the lifetime utility starting at t^* that student i receives from choosing major j.[13] X_i is a vector including i's observable characteristics discussed in Section 2 and also includes a constant equal to one in order to allow for the estimation of the constant in U^j. $v_{i,j}$ and $\varepsilon_{i,j}$ represent the effect on U^j of individual factors that are not observed by the econometrician, and we define $v_i = (v_{i,j}: j \in J^D)$ and $\varepsilon_i = (\varepsilon_{i,j}: j \in J^D)$. We assume that $\varepsilon_{i,j}$ is the portion of unobservable utility that is known to the student at the time of entrance, while $v_{i,j}$ is the portion that is not known at entrance but will be known by the decision time t^*. The remaining two elements of the function U^j are the major-specific factors (denoted $Factor_{i,j}$ in Subsection 4.1) that we utilize. $AGPA_{i,j}$ is the average GPA that person i would receive in a future semester if he had outcome j.[14] Noting that grade performance does not enter the utility from the option DROP, we define $AGPA_i = (AGPA_{i,j}: j \in J^{ND})$. $AINC_{i,j}$ is the average annual income that person i would receive after leaving school if he left school with outcome j, and we define $AINC_i = (AINC_{i,j}: j \in J^D)$. We assume a simple reduced form for the utility function.[15]

$$U^j\left(AGPA_{i,j}, AINC_{i,j}, X_i, v_{i,j}, \varepsilon_{i,j}\right) = \alpha_j X_i + \beta_1 AGPA_{i,j} + \beta_2 AINC_{i,j} + v_{i,j} + \varepsilon_{i,j}. \quad (1)$$

The central issue for the estimation of (1) is that, while $AGPA_{i,j}$ and $AINC_{i,j}$ are constants, their true values may not be fully known by person i at any particular time t. In addition, a student will be uncertain about $v_{i,j}$ at entrance. At time t^* a student is forced to make his final choice. If uncertainty about $AGPA_{i,j}$ and $AINC_{i,j}$ remains "unresolved" at t^*, the student makes his choice at t^* by choosing the option with the highest expected utility. Denoting the expected utility at t^*

12. Our modelling and estimation exercise involves understanding why students have particular $Pr_{i,j}$'s *taking as given* the courses they are enrolled in at the early stages of college. As such, we do not provide evidence about students' strategies for selecting courses in an effort to find a major with a good match. While understanding course choice is undoubtedly important, the structural model that would be needed to understand course choice would require a variety of additional assumptions that are not needed here given our objectives. For example, for our work we need to ascertain how a student updates his beliefs about underlying factors given the specific courses he is taking. To understand course choice it would be necessary to understand how the student would update beliefs for every possible set of classes that he might consider taking.

13. For sake of discussion think of t^* as occurring relatively quickly and abstract from issues related to the utility obtained while in college but before t^*.

14. Assume that i's grade point average (GPA) in major j in some future semester t' is given by

$$GPA_{i,j,t'} = AGPA_{i,j} + \gamma_{i,j,t'},$$

where $\gamma_{i,j,t'}$ is a mean-zero random variable representing the transitory portion of grades in t'. Technically speaking, lifetime utility associated with j might depend on not only $AGPA_{i,j}$ but also on $\gamma_{i,j,t'}$. However, the simplifying focus on the average can be motivated by the reality that knowing $AGPA_{i,j}$ is close to sufficient for knowing one's cumulative grade point average at the end of college for j since the sum of $\gamma_{i,j,t'}$ will tend towards zero with the number of semesters. The motivation for focusing on average income in a year is similar.

15. In theory, $AGPA_{i,j}$ and $AINC_{i,j}$ may influence both the utility received from j while in school and the utility received from j after leaving school. Then, if one wanted to understand why one of these factors mattered or did not matter at its most basic level, it would be necessary to identify the impact of the factor on utility in both the schooling and post-schooling periods. Our simple specification is motivated, in large part, by the difficulty of this identification task. If, as in Arcidiacono *et al.* (2012a), one assumes that some major-specific factors only influence utility in school and other major-specific factors only influence utility after school, then it is possible to put a stronger interpretation on individual coefficients.

of option j as $E_{t^*}U^j()$, the person is observed to choose j if

$$E_{t^*}U^j() - E_{t^*}U^k() > 0 \quad \text{for all} \quad k \neq j. \tag{2}$$

We let $AGPA^t_{i,j}$ and $AINC^t_{i,j}$ denote random variables which represent a student's beliefs at t about the constants $AGPA_{i,j}$ and $AINC_{i,j}$, so that $AGPA^{t^*}_{i,j}$ and $AINC^{t^*}_{i,j}$ are random variables characterizing the unresolvable uncertainty that remains at t^*. At t^*, to compute $E_{t^*}U^j()$ for each j, the student integrates $U^j(AGPA_{i,j}, AINC_{i,j}, X_i, v_{i,j}, \varepsilon_{i,j})$ over the distribution of the random variables $AGPA^{t^*}_{i,j}$ and $AINC^{t^*}_{i,j}$. Given the linear specification in Equation (1), this integration results in

$$E_{t^*}U^j\left(AGPA_{i,j}, AINC_{i,j}, X_i, v_{i,j}, \varepsilon_{i,j}\right) = \alpha_j X_i + \beta_1 E\left(AGPA^{t^*}_{i,j}\right) + \beta_2 E\left(AINC^{t^*}_{i,j}\right) + v_{i,j} + \varepsilon_{i,j}, \tag{3}$$

so that, from the student's perspective, $E_{t^*}U^j$ can be viewed as a function of the constants $E(AGPA^{t^*}_{i,j})$, $E(AINC^{t^*}_{i,j})$ and the value of $v_{i,j}$. In terms of the wording used in Subsection 4.1, $E(AGPA^{t^*}_{i,j})$ is the mean of the distribution describing i's beliefs at t^* about the average grade point average associated with j and $E(AINCOME^{t^*}_{i,j})$ is the mean of the distribution describing i's beliefs at t^* about the average future income associated with j.

Denote $E(AGPA^t_i) = (E(AGPA^t_{i,j}): j \in J^{ND})$ and $E(AINC^t_i) = (E(AINC^t_{i,j}): j \in J^D)$. At t=1 when the student reports Pr_i, he knows that his decision at t^* will be made taking into account the vector of constants $E(AGPA^{t^*}_i)$, the vector of constants $E(AINC^{t^*}_i)$, and the realization of v_i. Then, if at t=1 the student knew what $E(AGPA^{t^*}_i)$, $E(AINC^{t^*}_i)$, and v_i would turn out to be, there would be no uncertainty about what choice he would make at t^*. Uncertainty arises because there is likely to exist uncertainty about $AGPA_{i,j}$ and $AINC_{i,j}$ that will be "resolved" by t^* (and uncertainty about $v_{i,j}$ will be resolved by t^*). The perceived probability at t=1 of having a final outcome of j, $Pr_{i,j}$, is the probability that i will arrive at t^* with values of $E(AGPA^{t^*}_i)$, $E(AINC^{t^*}_i)$, and v_i such that, given ε_i, j is the optimal choice. Let $E(AGPA^{t^*}_i)^1$ and $E(AINC^{t^*}_i)^1$ be random variables describing i's beliefs at t=1 about what $E(AGPA^{t^*}_i)$ and $E(AINC^{t^*}_i)$ will turn out to be and let g denote a density function. Recall from Equation (3) that $E_{t^*}U^j$ can be viewed as a function of $E(AGPA^{t^*}_i)$, $E(AINC^{t^*}_i)$, and v_i. Then for any $j \in J$, $Pr_{i,j}$ is determined by utilizing Equation (2) as

$$\begin{aligned} Pr(E_{t^*}U^j() - E_{t^*}U^k() > 0 \ \text{ for all } \ k \neq j) &= \int 1(E_{t^*}U^j() - E_{t^*}U^k() > 0 \text{ for all } k \neq j) \\ &\times g\left(E\left(AGPA^{t^*}_i\right)^1, E\left(AINC^{t^*}_i\right)^1, v_i\right) dE\left(AGPA^{t^*}_i\right)^1 dE\left(AINC^{t^*}_i\right)^1 dv_i. \end{aligned} \tag{4}$$

The econometrician is assumed to know the distributions of $E(AGPA^{t^*}_i)^1$, $E(AINC^{t^*}_i)^1$, and v_i that are being used by student i in the calculation of (4). However, unlike the student, the econometrician does not know the individual-specific realization of the vector ε_i. Given the known distribution of ε_i, the likelihood contribution for i, L_i, is the probability of obtaining ε_i such that i would report the set $(Pr_{i,j}: j \in J^D)$ that he did:

$$\begin{aligned} L_i &= Prob\left(i \text{ reports } \left(Pr_{i,j}: j \in J^D\right)\right) \\ &= Prob(\varepsilon_i : \forall j \in J^D, \ Pr(E_{t^*}U^j() - E_{t^*}U^k() > 0 \text{ for all } k \neq j) = Pr_{i,j}) \end{aligned}$$

$$= \text{Prob}(\varepsilon_i : \forall j \in J^D, \tag{5}$$

$$\int 1(E_{t^*}U^j() - E_{t^*}U^k() > 0 \text{ for all } k \neq j)\, g\left(E\left(AGPA_i^{t^*}\right)^1,\ E\left(AINC_i^{t^*}\right)^1,\ v_i\right)$$

$$\times\ dE\left(AGPA_i^{t^*}\right)^1 dE\left(AINC_i^{t^*}\right)^1 dv_i = Pr_{i,j}).$$

Then, estimation of the model requires knowledge of the distributions of $E(AGPA_i^{t^*})^1$ and $E(AINC_i^{t^*})^1$. These distributions are not observed directly in our data, primarily because it would not be easy to construct a survey question that would credibly elicit this information.[16] Subsection 4.3 describes construction of $E(AGPA_i^{t^*})^1$ and Subsection 4.4 describes construction of $E(AINC_i^{t^*})^1$. Our effort to be explicit regarding the source of uncertainty about a student's major and to allow this uncertainty to be heterogeneous across students represents a natural next step in the very small literature that allows agents to express uncertainty about a future choice (Blass *et al.*, 2010).

Also needed to estimate the model is a value of t^*. It is important to stress that t^* is not observed in institutional rules or in administrative data. This is the case because t^* should reflect a student's views, at the time he reports his initial beliefs, about when he will likely choose an outcome. These views may be quite different from, for example, the time at which institutional rules require students to declare an outcome (roughly the end of the 2nd year at Berea). Institutional rules in U.S. schools are typically designed to allow much flexibility in the timing of choices; a student is allowed to drop out at any point, rules about the timing of major declaration only specify the latest date at which a student can declare a major, and even this declaration is typically not truly binding. Waiting as long as possible to finalize a decision allows increased experimentation. However, the costs of delaying a decision could be considerable; a student who waits unnecessarily long to leave school foregoes earnings, and a student who waits unnecessarily long to choose a major risks unnecessary delays in graduation and may also not end up with an optimal mix of major and non-major courses. As such, there do exist incentives that push a student towards making a decision relatively early, even if institutional rules do not require this.[17]

Students will tend to feel comfortable making a decision early if they are quite certain about their abilities and interests. Then, given that overconfidence at entrance is a strong, common theme throughout this article and also in our other work related to dropout (S&S, 2012; S&S, forthcoming), our intuition is that students in our data will tend to believe at entrance that they will make decisions substantially earlier than: (1) what would be required by institutional rules and (2) what might be suggested by the timing of actual decisions observed in the data. We have some direct evidence in support of our intuition for one of our outcomes—DROP. This is the case because our baseline survey elicited not only a student's beliefs about the probability of graduating, but also the student's beliefs about the probability of returning for the second year (Question 2b). Combining these questions we find that, on average, students believe that there is

16. The survey question related to AGPA would have to be something along the lines of, "please tell us the per cent chance that, at the time you make your major decision, the mean of the distribution describing your beliefs about $AGPA_{i,j}$ will fall in each of the following grade categories …" Note that this is not the same as asking a respondent at entrance about the per cent chance that his GPA would be in various grade categories in the future if he had major j. One reason is that, in addition to capturing uncertainty about $AGPA_{i,j}$, the answers to this grade question would also reflect period-specific transitory variation in a person's grade performance. A second reason is that, even if no transitory variation existed, this grade question would capture both resolvable and unresolvable uncertainty about $AGPA_{i,j}$ when what is needed for estimation is to isolate the unresolvable portion.

17. For recent work examining the timing of specialization see Malamud (2010, 2011).

only a 0.04 chance of dropping out after the start of the second year. Thus, the evidence suggests that students believe that the dropout decision will take place quickly even though institutional details do not require this and even though S&S (forthcoming) finds that a substantial amount of dropout takes place after the first year.[18] Of course, it is possible that students believe that the major decision takes place on a different timing than the dropout decision. Looking at our outcome of particular interest, SCI, the data seem to support the notion that a decision also tends to be made rather quickly. At the middle of the second year, 90% of students who have an outcome of SCI indicate that SCI is the most likely major, and, among graduates, only 3% of students who have an outcome other than SCI list SCI as the most likely major.

Given this empirical evidence related to DROP and SCI and given that Section 3 found that students do not fully appreciate how much uncertainty exists about college major, we use $t^*=3$ (the beginning of the 3rd semester) to obtain our main results, although we also examine the robustness of our results to different choices about t^*. An additional appeal of using $t^*=3$ for our primary results is that this choice is most consistent with our implicit assumption that students know the courses they will take between $t=1$ and t^* when reporting their initial beliefs. At $t=1$, students will know the set of classes that are mandatory under the general studies curriculum, will know the elective courses they will take in the first semester, and may also often have a good sense of the elective courses they will take in the second semester.

We note that the choice of $t^*=3$ may not be as appealing for all other majors as it is for SCI and DROP. For example, at the middle of the second year, 73% of students in the sample who have an outcome of PRO indicate that PRO is the most likely major compared with the 90% that was found above for SCI. This percentage is 80%, 92%, 84%, 75%, and 88%, respectively, for the remaining majors, AG, BUS, ED, HUM, and SS, respectively. Aggregating the non-science majors yields a proportion of 0.80, and a test that the proportion for the aggregated non-science major is the same as the proportion found earlier for SCI (0.90) is rejected at a 0.05 level of significance. The difference between SCI and some of the other majors is perhaps not surprising because there may be differences in the timing of when classes in particular majors tend to be taken. For example, using transcript data, we find that students in the overall sample take, on average, 1.74 SCI classes in the first year and students who have $STATE_i = SCI$ take, on average, 1.86 SCI classes in the first year. However, students in the overall sample take, on average, only 0.29 PRO classes in the first year and students who have $STATE_i = PRO$ take, on average, only 0.32 PRO classes in the first year.[19]

The differences observed between SCI and certain other majors suggests that it may be appealing to assume that at an early stage of school students choose between three options {DROP, SCI, and NON-SCI} where NON-SCI is an aggregated category that contains all majors other than SCI. Modelling the decision in this way would not allow us to understand how students

18. Given that, on average, students believe that the overall probability of dropout is 0.134 (Column 2 Table 1), they believe that over 70% of dropout will take place in the first year. S&S (2012) find that roughly 38% of dropout takes place before the start of the second year. This highlights the reality that, in the presence of overoptimism, examining the timing of actual outcomes is not particularly useful for understanding what students think about when they will make decisions.

19. The non-trivial number of science courses arises, in part, because the curriculum requires students to take some science in order to graduate. On average, students take 9.51 classes in the first year. The average number of courses taken is 1.53, 0.21, 0.11, 2.22, 1.74, 0.29, and 0.73 in AG, BUS, ED, HUM, SCI, PRO,and SS, respectively. In addition, students, on average, take 2.65 classes that do not fit within any specific major area but are required under the school's liberal arts (General Studies) curriculum. A small number of the classes in other areas may also fit the description of satisfying liberal arts requirements. For example, the large average number for AG arises, in large part, due to required physical education classes and students would often take at least one humanities class in the first year as part of the liberal arts curriculum.

choose among the individual non-science majors, but may be most realistic given our specific objective of understanding the choice of SCI major. We return to a discussion of issues related to the choice set when we discuss the identification and estimation of the model in Subsection 4.6. However, we first describe in Subsection 4.3. how we construct the distribution of $E(AGPA_i^{t^*})^1$, describe in Subsection 4.4. how we construct the distribution of $E(AINC_i^{t^*})^1$, and provide some descriptive evidence in Subsection 4.5. that future grades and future income are likely play an important role in determining a student's choice of major.

4.3. *Constructing $E(AGPA_i^{t^*})^1$*

In this section we detail the construction of the distribution of $E(AGPA_i^{t^*})^1$, which describes a student's beliefs at t=1 about what $E(AGPA_i^{t^*})$ will turn out to be. Our motivation for taking into account that students may be uncertain about $E(AGPA_i^{t^*})$ at entrance comes from the fact that much updating is observed in the data. For the 428 students for which $E(AGPA_{i,j}^{t})$ is observed at both t=1 and at our choice of t^*=3, the sample average (std. dev.) of $E(AGPA_{i,j}^{t^*}) - E(AGPA_{i,j}^{1})$ is −0.18 (0.75), −0.02 (0.72), −0.08 (0.68), −0.01 (0.67), −0.04 (0.69), −0.03 (0.67), and −0.04 (0.71), respectively, for j = SCI, AG, BUS, ED, HUM, PRO, and SS, respectively. The null hypothesis that the average updating is the same for SCI as it is for $j \neq SCI$ is rejected at 0.05 for all $j \neq SCI$.[20] This suggests that taking into account uncertainty may be especially important given our particular interest in SCI. However, the non-trivial standard deviations observed for all majors imply that substantial updating is present in all majors, even those with means close to zero.

As described in Subsection 4.1, the first step in characterizing the distribution of $E(AGPA_i^{t^*})^1$ is to specify the updating process by which i will arrive at the mean of the distribution describing his beliefs at t^* about the average grade point average he will receive in major j (*i.e.* how he will arrive at $E(AGPA_i^{t^*})$). Understanding why substantial heterogeneity in updating is observed across students is important for thinking about the specification of this updating process. Focusing on SCI, we find evidence that heterogeneity in updating exists, in part, because students who have $STATE_i = SCI$ have significantly (at 0.05) different mean updates than students with $STATE_i \neq SCI$, but that much heterogeneity in updating remains even within each of these groups; the sample mean (std. dev.) of $E(AGPA_{i,SCI}^{t^*}) - E(AGPA_{i,SCI}^{1})$ is −0.30 (0.46) for students with $STATE_i = SCI$ and is −0.13 (0.82) for students with $STATE_i \neq SCI$.

Our updating process involves viewing $E(AGPA_{i,j}^{t^*})$ as a posterior mean which is arrived at from a prior mean $E(AGPA_{i,j}^{1})$ through an updating process that also takes into account an imperfect signal $Sig_{i,j}$ of $AGPA_{i,j}$:

$$E\left(AGPA_{i,j}^{t^*}\right) = \alpha_j + \beta_j E\left(AGPA_{i,j}^{1}\right) + \gamma_j Sig_{i,j} + u_{i,j}, \tag{6}$$

where $u_{i,j}$ is a mean-zero unobservable.

For the full sample (Column 1) and for the sample stratified by $STATE_i$ (Columns 2–9), Table 3 shows the sample average of $E(AGPA_{i,j}^{1})$ for each j. The results in the first column

20. Students appear to be learning about their ability in SCI, not that they are unwilling to put effort into science. Indeed, the average daily number of hours that students expect to study if they had SCI as their major increases between t=1 and t=3. The change in the average expected number of daily study hours between t=1 and t=3 is 0.17, −0.03, 0.06, −0.03, −0.05, 0.17, and −0.001 for SCI, AG, BUS, ED, HUM, PRO, and SS, respectively. See S&S (2004, 2008b) for work examining the role of study effort in determining grade performance.

TABLE 3
Sample mean and standard deviation of $E(AGPA^1_{i,j})$

	Sample mean (std. dev.)								
	1	2	3	4	5	6	7	8	9
Full sample	Full Sample $n = 638$	$STATE_i =$ DROP $n = 38$	$STATE_i =$ SCI $n = 139$	$STATE_i =$ AG $n = 73$	$STATE_i =$ BUS $n = 92$	$STATE_i =$ ED $n = 55$	$STATE_i =$ HUM $n = 134$	$STATE_i =$ PRO $n = 99$	$STATE_i =$ SS $n = 94$
j = SCI	2.91 (0.84)	2.65 (0.65)	3.53 (0.33)	2.76 (0.74)	2.86 (0.79)	2.71 (0.88)	2.65 (1.04)	2.83 (0.83)	2.57 (0.86)
j = AG	3.05 (0.70)	2.99 (0.63)	3.28 (0.56)	3.32 (0.44)	3.01 (0.80)	3.04 (0.60)	2.74 (0.86)	3.04 (0.62)	2.90 (0.69)
j = BUS	3.05 (0.69)	3.02 (0.59)	3.27 (0.47)	2.78 (0.67)	3.48 (0.36)	2.89 (0.66)	2.83 (0.87)	2.94 (0.77)	2.96 (0.62)
j = ED	3.18 (0.65)	3.15 (0.52)	3.32 (0.53)	3.02 (0.58)	3.09 (0.76)	3.56 (0.271)	3.13 (0.78)	3.08 (0.69)	3.09 (0.54)
j = HUM	3.18 (0.61)	3.03 (0.53)	3.19 (0.60)	2.88 (0.74)	3.10 (0.674)	3.13 (0.653)	3.61 (0.30)	3.05 (0.67)	3.15 (0.49)
j = PRO	3.07 (0.68)	2.93 (0.62)	3.20 (0.54)	2.84 (0.65)	3.05 (0.684)	3.02 (0.68)	2.89 (0.86)	3.45 (0.38)	2.95 (0.65)
j = SS	3.14 (0.67)	3.01 (0.54)	3.22 (0.59)	2.89 (0.65)	3.01 (0.831)	2.94 (0.71)	3.20 (0.66)	3.00 (0.73)	3.49 (0.34)
		$STATE_i =$ DROP $n = 38$	$STATE_i =$ SCI $n = 139$	$STATE_i = k$ $k \neq SCI$ $k \neq$ DROP $n = 486$					
j = SCI		2.65 (0.65)	3.53 (0.33)	2.74 (0.87)					
j = NON-SCI (measure 1)		3.26 (0.35)	3.42 (0.31)	3.44 (0.33)					
j = NON-SCI (measure 2)		3.46 (0.28)	3.61 (0.29)	3.59 (0.32)					
Grad sample	All graduates $n = 403$		$STATE_i =$ SCI $n = 93$	$STATE_i =$ AG $n = 40$	$STATE_i =$ BUS $n = 59$	$STATE_i =$ ED $n = 33$	$STATE_i =$ HUM $n = 89$	$STATE_i =$ PRO $n = 59$	$STATE_i =$ SS $n = 72$
j = SCI	2.93 (0.83)		3.54 (0.31)	2.74 (0.76)	2.79 (0.76)	2.84 (0.81)	2.71 (1.05)	2.70 (0.87)	2.72 (0.79)
j = AG	3.03 (0.66)		3.28 (0.58)	3.29 (0.41)	2.92 (0.74)	3.09 (0.55)	2.72 (0.77)	3.01 (0.66)	2.96 (0.67)
j = BUS	3.09 (0.65)		3.33 (0.39)	2.81 (0.75)	3.46 (0.357)	3.00 (0.46)	2.88 (0.80)	2.92 (0.74)	3.02 (0.59)
j = ED	3.22 (0.58)		3.36 (0.44)	3.07 (0.62)	3.06 (0.72)	3.52 (0.27)	3.21 (0.64)	3.14 (0.66)	3.13 (0.51)
j = HUM	3.21 (0.59)		3.16 (0.63)	2.97 (0.60)	2.88 (0.72)	3.19 (0.47)	3.63 (0.298)	3.08 (0.65)	3.15 (0.50)
j = PRO	3.12 (0.62)		3.25 (0.44)	2.80 (0.70)	3.00 (0.62)	3.12 (0.53)	2.95 (0.784)	3.46 (0.37)	3.01 (0.62)
j = SS	3.16 (0.66)		3.28 (0.54)	2.96 (0.69)	2.91 (0.87)	2.96 (0.536	3.23 (0.65)	2.94 (0.84)	3.48 (0.34)
			$STATE_i =$ SCI $n = 93$	$STATE_i =$ $k \neq SCI$ $n = 316$					
j = SCI			3.54 (0.31)	2.75 (0.85)					
j = NON-SCI (measure 1)			3.42 (0.32)	3.43 (0.32)					
j = NON-SCI (measure 2)			3.62 (0.29)	3.59 (0.31)					

Notes: For the overall sample (Column 1) and for the sample stratified by $STATE_i$ (Columns 2–9) the table shows the sample mean (sample standard deviation) $E(AGPA_{i,j})$ for each j. If $Pr_{i,j} = Pr_{i,k}$ are tied as the highest probabilities, then beliefs enter in both the column associated with j and the column associated with k. The first panel includes all students with legitimate responses. The second panel includes only students who graduated.

indicate that students begin school believing that grade performance may depend on major. For example, on average, students believe that SCI will be more difficult from a grade perspective; students believe that their grades will be, on average, between 0.14 (BUS) and 0.27 (ED, HUM) lower in SCI than in the other majors, and, for each $j \neq SCI$ the null hypothesis that the average $E(AGPA^1_{i,j})$ is equal to the average $E(AGPA^1_{i,SCI})$ is rejected at all traditional

significance levels.[21] However, the non-trivial standard deviations in Column 1 indicate that there is substantial heterogeneity across students in $E(AGPA^1_{i,j})$, with the largest sample standard deviation of 0.84 being associated with j = SCI. We note that the low (relative) sample average of $E(AGPA^1_{i,SCI})$ comes despite the fact that students believe they would study more in SCI.[22]

Grade information obtained from administrative records is used to construct the signal, $Sig_{i,j}$, of $AGPA_{i,j}$ in Equation (6). We check robustness to two signals. The first signal is student i's overall grade point average in the period between entrance and t*, which we denote as GPA_i. GPA_i is observed directly in the administrative data. For our choice of t*=3 we find a mean (standard deviation) of 2.88 (0.72). As described in more detail later in this section, the second signal is a multi-dimensional signal which disaggregates GPA_i to account for differences in grade performance in different types of courses. The course-specific grade information is obtained from transcript data. For our choice of t*=3, we find that the sample average (std. dev.) grade is 2.58 (1.19), 3.23 (0.88), 2.69 (1.23), 3.48 (0.90), 3.13 (0.97), 3.04 (0.82), 2.75 (1.10), and 2.99 (0.96), respectively, in SCI, AG, BUS, ED, HUM, PRO, SS, and GEN courses, respectively, where GEN represents courses that do not fit within any specific major area but are required under the school's liberal arts (General Studies) curriculum.

The two signals have different strengths and weaknesses. In a textbook Bayesian model, the weight assigned to a signal should vary across signals that contain different amounts of information (noise). Grade performance in a particular class may be more informative about $AGPA_{i,j}$ for some majors j than other majors j. The disaggregation associated with the second signal is appealing because it attempts to group together courses that might be equally informative about $AGPA_{i,j}$. As discussed in more detail later in this section, the strength of the first signal comes from the fact that, as described in Subsection 4.1, the construction of $E(AGPA^{t^*}_{i,j})^1$ requires not only the realized signal $Sig_{i,j}$ (which is used to estimate the updating equation) but also the distribution describing beliefs about $Sig_{i,j}$ (which is used to characterize i's uncertainty about $E(AGPA^{t^*}_{i,j})$ given estimates of the updating equation). While we directly elicited each student's belief about the distribution of the first signal GPA_i using survey Question 3,[23] the construction of beliefs about the distribution of the second signal requires some assumptions.

With $E(AGPA^1_{i,j})$, $Sig_{i,j}$, and $E(AGPA^{t^*}_{i,j})$ observed for each j, we are able to estimate Equation (6) for each j. We begin by examining the determination of $E(AGPA^{t^*}_{i,SCI})$. The first column of Table 4 shows results from the estimation of Equation (6) using the overall GPA signal (*i.e.* $Sig_{i,SCI} = GPA_i$). The results in the first panel of Column 1 indicate that the "posterior" belief $E(AGPA^{t^*}_{i,SCI})$ is strongly related to both the "prior" belief $E(AGPA^1_{i,SCI})$ and the signal GPA_i with the coefficients having *t*-statistics of 11.72 and 5.29, respectively. As noted earlier, this signal does not take into account that the informativeness of GPA_i for understanding future

21. Without information describing what each student believes about the distribution of the actual grades that are assigned in each major (across students), it is not possible to know whether the low average $E(AGPA^1_{i,SCI})$ at entrance indicates that (1) students tend to think they are not good at science relative to other students or (2) students simply realize that the grades assigned in science classes tend to be lower. However, we will eventually show that, among students with $STATE_i$ = SCI, students who have $j^* \neq$ SCI start school with beliefs $E(AGPA^1_{i,SCI})$ that are similar to students who have j^* = SCI but end school with beliefs that are much different. Then, students who are leaving SCI are learning that they are not particularly well-suited for SCI relative to other students.

22. The expected study effort is 3.64, 2.26, 2.93, 2.17, 2.89, 3.23, and 3.06, respectively, for SCI, AG, BUS, ED, HUM, PRO, and SS, respectively. Then, constructing ability-type measures that are analogous to $E(AGPA^1_{i,j})$ but holding study effort constant leads to even larger differences between SCI and the other majors (S&S, 2011).

23. In reality, although this question technically reflected beliefs about the first semester, we assume this distribution is the relevant distribution for the entire first year.

TABLE 4
Determinants of $E(AGPA^{j^*}_{i,j})$ from equation (6) using $Sig_{i,j} = GPA_i$

	1	2	3	4	5	6	7
	j = SCI estimate (std. error) Full sample $n=420$	j = AG estimate (std. error) Full sample $n=412$	j = BUS estimate (std. error) Full sample $n=423$	j = ED estimate (std. error) Full sample $n=416$	j = HUM estimate (std. error) Full sample $n=423$	j = PRO estimate (std. error) Full sample $n=421$	j = SS estimate (std. error) Full sample $n=425$
GPA_i	0.227 (0.051)*	0.246 (0.047)*	0.226 (0.047)*	0.293 (0.042)*	0.344 (0.044)*	0.226 (0.043)*	0.252 (0.043)*
$E(AGPA^1_{i,j})$	0.434 (0.037)*	0.361 (0.041)*	0.413 (0.041)*	0.238 (0.040)*	0.314 (0.039)*	0.296 (0.039)*	0.262 (0.037)*
Constant	0.810 (0.183)*	1.190 (0.187)*	1.032 (0.177)*	1.543 (0.165)**	1.077 (0.157)*	1.464 (0.166)*	1.502 (0.163)*
R^2	0.281	0.209	0.253	0.199	0.29	0.191	0.189
	$State_i = SCI$ $n=106$	$State_i = AG$ $n=47$	$State_i = BUS$ $n=61$	$State_i = ED$ $n=40$	$State_i = HUM$ $n=83$	$State_i = PRO$ $n=73$	$State_i = SS$ $n=65$
GPA_i	0.376 (0.071)*	0.275 (0.104)*	0.320 (0.077)*	0.268 (0.111)	0.280 (0.074)*	0.315 (0.082)*	0.250 (0.113)*
$E(AGPA^1_{i,j})$	0.630 (0.142)*	0.470(0.132)*	0.296 (0.114)*	0.359 (0.218)*	0.496 (0.120)*	0.414 (0.128)*	0.246 (0.175)
Constant	–0.073 (0.454)*	0.753 (0.476)	1.377 (0.391)*	1.301 (0.756)	0.762 (0.467)	0.886 (0.442)*	1.638 (0.650)*
R^2	0.281	0.362	0.367	0.242	0.315	0.325	0.117
	$State_i \neq SCI$ $n=314$	$State_i \neq AG$ $n=365$	$State_i \neq BUS$ $n=362$	$State_i \neq ED$ $n=376$	$State_i \neq HUM$ $n=340$	$State_i \neq PRO$ $n=348$	$State_i \neq SS$ $n=360$
GPA_i	0.131 (0.062)*	0.241 (0.052)*	0.227 (0.053)*	0.297 (0.045)*	0.335 (0.050)*	0.209 (0.507)*	0.250 (0.047)*
$E(AGPA^1_{i,j})$	0.347 (0.04)*	0.355 (0.044)*	0.382 (0.045)*	0.223 (0.042)*	0.266 (0.044)*	0.272 (0.043)*	0.253 (0.039)*
Constant	1.274 (0.21)*	1.225 (0.203)*	1.086 (0.195)*	1.568 (0.174)*	1.215 (0.181)*	1.574 (0.186)*	1.525 (0.174)*
R^2	0.208	0.199	0.225	0.189	0.232	0.166	0.183

Notes: Results show OLS estimate (std. error) of parameters from updating equation (6) using $Sig_{i,j} = GPA_i$. Sample includes all students for which both $E(AGPA^1_{i,j})$ and $E(AGPA^3_{i,j})$ are observed.
* Significant at 0.05.

grade performance in SCI is likely to vary with, for example, the number of science classes a student is taking. Because a student's courses will tend to vary with his stated major, we estimate Equation (6) separately for students with $STATE_i = SCI$ (panel 2, Column 1) and for students with $STATE_i \neq SCI$ (Panel 3, Column 1). We find that GPA_i is a substantially stronger predictor of $E(AGPA^{t^*}_{i,SCI})$ when $STATE_i = SCI$ than when $STATE_i \neq SCI$, with estimates (t-stats) of 0.376 (5.29) and 0.131 (2.11), respectively.

Even after stratifying on the basis of $STATE_i$, the overall GPA signal is less than ideal. The number of science classes might vary within the $STATE_i = SCI$ or $STATE_i \neq SCI$ groups, and two students who obtain the same overall GPA_i in the same set of classes may have different grades in, for example, their science classes. This serves as our motivation for our second signal which disaggregates GPA_i to account for grade performance in different types of courses. Specifically, we construct the terms $GPA_{i,SCI} \bullet 1(0 < CL_{i,SCI} \leq 2)$, $GPA_{i,SCI} \bullet 1(CL_{i,SCI} > 3)$, and $GPA_{i,NON\text{-}SCI}$, where $1(\bullet)$ is an indicator function, $CL_{i,SCI}$ is the number of classes taken in SCI between t=1 and t^*, $GPA_{i,SCI}$ is the GPA in these SCI classes, and $GPA_{i,NON\text{-}SCI}$ is the GPA in all non-science classes. Related to the first two terms, the possibility that some students will take zero science classes implies the need to include an additional term $1(CL_{i,SCI} = 0)$. With respect to $GPA_{i,NON\text{-}SCI}$, we note two things: (1) we have simplified our specification by aggregating the classes in all non-science majors and (2) while the informativeness of the signal from the non-science classes will depend on the number of non-science courses, differentiating between different numbers of non-science classes is not overly important because students typically have a sizeable number of non-science classes. Finally, in the textbook Bayesian context, the informativeness of the signal will also influence the weight assigned to the prior mean. As such, we also add to equation (6) interaction terms allowing the importance of the prior $E(AGPA^1_{i,j})$ to depend on the values of $1(CL_{i,SCI} = 0)$, $1(0 < CL_{i,SCI} \leq 2)$, and $1(CL_{i,SCI} > 3)$.

Column 1 of Table 5 shows the estimates of equation (6) for the dependent variable $E(AGPA^{t^*}_{i,SCI})$ using the second signal as described in the previous paragraph. The results indicate that courses in SCI play a particularly important role in the determination of $E(AGPA^{t^*}_{i,SCI})$. For a student who takes only one or two courses in SCI, the estimated effect of $GPA_{i,SCI}$ is slightly larger than the estimated effect of $GPA_{i,NON\text{-}SCI}$ (0.109 versus 0.087) even though students are taking, on average, between seven and eight non-science classes. The conclusion that an individual SCI course is substantially more informative than an individual non-science course can be seen even more clearly by noting that the estimated effect of $GPA_{i,SCI}$ increases in a statistically significant (at a 0.05 level) manner (from 0.109 to 0.257) when a student takes three or more science classes.[24]

Columns 2–7 of Table 5 estimate the updating specification in Column 1 (from equation (6)) for majors j other than SCI. Most striking is that the effect of GPA in courses outside of j, which we denote $GPA_{i,NON\text{-}j}$, is substantially more important in Columns 2–7 than it was for SCI in Column 1. The t-statistics for $GPA_{i,NON\text{-}j}$ range from 4.43 to 8.47 in Columns 2–7 compared with 1.64 in Column 1, and the estimated effects of $GPA_{i,NON\text{-}j}$ range from 0.201 to 0.288 in Columns 2–7 compared with 0.087 in Column 1. Then, noting that the estimated effects of $GPA_{i,j}$ (interacted with number of classes taken in j) are often not significant in Columns 2–7, our results indicate that, in majors j other than SCI, students collect much of their information from classes outside of the major j. Roughly speaking, students seem to believe that the set of majors other than SCI require a somewhat common set of abilities/knowledge, with this set somewhat different than what is required for SCI. Looking back at the second two panels of Table 4, we find additional evidence in support of this conclusion. Unlike what was found for SCI in Column 1, the effect

24. It is possible that this increase occurs, in part, because students who are taking more science classes tend to also be taking somewhat different types of science classes as part of, for example, a major in science.

TABLE 5
Determinants of $E(AGPA^{t^*}_{i,j})$ from equation (6) using $Sig_{i,j} = (GPA_{ij}, GPA_{i,NON\text{-}j})$

Variable	1 j = SCI $n = 422$	2 j = AG $n = 414$	3 j = BUS $n = 426$	4 j = ED $n = 418$	5 j = HUM $n = 427$	6 j = PRO $n = 423$	7 j = SS $n = 428$
$Sig_{i,j}$							
$GPA_{i,j} \times 1(CL_{i,j} \geq 3)$	0.257 (0.070)*	0.054(0.092)	NA	NA	0.020 (0.064)	0.223 (0.040)*	0.292 (0.153)
$GPA_{i,j} \times 1(0 < CL_{i,j} \leq 2)$	0.109 (0.038)*	0.074 (0.041)	0.104 (0.078)	0.005 (0.092)	0.184 (0.039)*	0.024 (0.079)	0.109 (0.039)*
$1(CL_{i,j} = 0)$	−0.163 (0.262)	0.114 (0.403)	−0.983 (0.433)*	0.103 (0.82)	−0.104 (0.273)	−0.408 (0.418)*	0.297 (0.248)
$GPA_{i,NON\text{-}j}$	0.087 (0.053)	0.201 (0.044)*	0.216 (0.043)*	0.288 (0.038)*	0.218 (0.042)*	0.232 (0.404)*	0.219 (0.043)*
$Prior_{i,j}$							
$E(AGPA^1_{i,j}) \times 1(CL_{i,j} = 0)$	0.402 (0.076)*	0.340 (0.117)*	0.405 (0.043)*	0.223 (0.404)*	0.375 (0.071)*	0.306 (0.042)*	0.254 (0.054)*
$E(AGPA^1_{i,j}) \times 1(0 < CL_{i,j} \leq 2)$	0.490 (0.047)*	0.350 (0.046)*	0.123 (0.126)	0.295 (0.214)	0.193 (0.049)*	0.167 (0.099)	0.255 (0.051)*
$E(AGPA^1_{i,j}) \times 1(CL_{i,j} \geq 3)$	0.348 (0.066)*	0.054 (0.092)	NA	NA	0.020 (0.064)	0.223 (0.040)*	0.292 (0.153)
constant	1.271* (0.199)**	1.188 (0.188)*	2.034 (0.423)*	1.502 (0.817)	1.28 (0.181)*	1.830 (0.390)*	1.403 (0.187)*
R^2	0.316	0.229	0.303	0.251	0.354	0.211	0.228

Notes: Results show OLS estimate (std. error) of parameters from updating equation (6) using $Sig_{i,j} = (GPA_{ij}, GPA_{i,NON\text{-}j})$.
Sample includes all students for which both $E(AGPA^1_{i,j})$ and $E(AGPA^3_{i,j})$ are observed.
* Significant at 0.05.

of GPA_i on $E(AGPA_{i,j}^{t^*})$ for the other majors does not tend to depend much on whether or not a person has $STATE_i = j$ (in which case the person would likely be taking more classes in j); the difference between the estimated effect of GPA_i in Panel 2 and the estimated effect of GPA_i in Panel 3 is 0.245 for SCI and is, on average, 0.025 for the other six majors (with estimates ranging from −0.055 to 0.106).

We note one more practical issue related to updating. During the estimation of our model for the reported Pr_{ij}'s in equations (1)–(5), we obtain the distribution of $E(AGPA_i^{t^*})^1$ by considering the likelihood of different possible values of $\alpha_j + \beta_j E(AGPA_{i,j}^1) + \gamma_j Sig_{i,j}$ (from equation (6)) given the distribution describing person i's beliefs at entrance about $Sig_{i,j}$ (constructed as described in subsection 4.3). S&S (forthcoming) finds that, in terms of grades assigned, courses taken in later years tend to be of similar difficulty as courses taken in earlier years.[25] Then, internal consistency would seem to suggest that the mean of $E(AGPA_{i,j}^{t^*})^1$ should be equal to $E(AGPA_{i,j}^1)$.[26] Our approach is to allow the constant α_j to vary across i for each j and to choose these constants in a way that satisfies this internal consistency constraint. In practice, for the first signal in which $Sig_{i,j} = GPA_i$, this is equivalent to the updating rule:

$$E\left(AGPA_{i,j}^{t^*}\right)^1 = E\left(AGPA_{i,j}^1\right) + \gamma_j\left(GPA_i^1 - E\left(GPA_i^1\right)\right), \tag{7}$$

where GPA_i^1 is a random variable describing i's beliefs about GPA_i. Similarly, the analogous internally consistent updating rule for the second (disaggregated) signal can be written as

$$\begin{aligned} E\left(AGPA_{i,j}^{t^*}\right)^1 = {} & E\left(AGPA_{i,j}^1\right) + \gamma_{j,1}\left(GPA_{i,j}^1 - E\left(GPA_{i,j}^1\right)\right) \\ & + \gamma_{j,2}\left(GPA_{i,NON\text{-}j}^1 - E\left(GPA_{i,NON\text{-}j}^1\right)\right), \end{aligned} \tag{8}$$

where $GPA_{i,j}^1$ is a random variable describing beliefs about $GPA_{i,j}$, and $GPA_{i,NON\text{-}j}^1$ is a random variable describing beliefs about $GPA_{i,NON\text{-}j}$ (i's grade performance in all classes taken outside of major j). We take our estimates for γ_j from Table 4 and our estimates for $\gamma_{j,1}$ and $\gamma_{j,2}$ from Table 5.[27] Note from Table 5 that the value we use for $\gamma_{j,1}$ will depend on how many classes in j a person is taking.

Given estimates of the parameters in equations (7) and (8), the distribution of $E(AGPA_{i,j}^{t^*})^1$ can be constructed for each j if one knows the distribution describing the person's beliefs about $Sig_{i,j}$. A desirable feature of the first signal GPA_i is that the needed distribution describing beliefs about GPA_i is directly observed using survey Question 3. In the sample, students believe that the average probability is 0.405, 0.350, 0.173, 0.048, 0.016, and 0.006, respectively, that GPA_i will be between 3.5 and 4.0, 3.0 and 3.5, 2.5 and 3.0, 2.0 and 2.5, 1.0 and 2.0, and 0.0 and 1.0, respectively. Thus, on average, students are optimistic; they believe that GPA_i will be approximately 3.27, while the actual average GPA_i is 2.88.

25. For example, Row 2 of Table 1 in S&S (2013) finds that, for a composition constant sample, the sample average GPA remains remarkably constant over time.

26. By internal consistency we have in mind that: a student's best guess (expected value) at t=1 of what the mean of the distribution describing his beliefs at t^* about $AGPA_{i,j}$ will turn out to be (*i.e.* what $E(AGPA_{i,j}^{t^*})$ will turn out to be) should be the mean of the distribution describing his beliefs at t=1 about $AGPA_{i,j}$ (*i.e.* $E(AGPA_{i,j}^1)$). This internal consistency would not be appealing if students believed that courses in t^* assigned systematically higher or lower grades in the future as compared to t=1.

27. Estimating a version of equations (7) and (8) directly uses different variation, but leads to little change.

As described earlier in this Subsection 4.3, for each $j \in J^{ND}$ the second signal disaggregates GPA_i into $GPA_{i,j}$ and $GPA_{i,NON\text{-}j}$, where the former term is i's GPA in all classes in major j and the latter term is i's GPA in all classes outside of j. Thus, for any student i, characterizing beliefs about the second signal for all possible j requires knowledge of the distribution describing beliefs about $GPA_{i,j}$ for all j. For each j, a natural assumption is that the mean of the distribution describing beliefs about $GPA_{i,j}$ is given by the reported value $E(AGPA^1_{i,j})$.[28] Then, the reason we noted earlier that the distribution describing beliefs about $GPA_{i,j}$ is not fully observed is that the variance of this distribution is not directly observed. What can be observed from survey Question 3 is unique information about the variance of the distribution describing beliefs about overall GPA_i. Observing this overall variance allows us to take into account that some students may generally be more uncertain about grade performance than others. However, given this overall variance and knowledge of the set of classes that student i took, constructing the variance of the distribution describing beliefs about $GPA_{i,j}$ for each j requires additional assumptions related to what student i believes about the relative amount of uncertainty across j. Roughly speaking, we make the assumption that i's uncertainty about grade performance in a particular class in major j is proportional to the variance of the grades observed in the data for classes in major j. While this assumption is untestable, it does have the virtue of allowing students to realize that there may be more grade uncertainty in some majors than in other majors. In Appendix B, we describe this assumption in more detail and describe the latent variable model that we use to characterize the distribution describing beliefs about $GPA_{i,j}$ for each j.

4.4. *Constructing $E(AINC_i^{t^*})^1$*

In this section we detail the construction of the distribution of $E(AINC_i^{t^*})^1$, which describes a student's beliefs at t=1 about what $E(AINC_i^{t^*})$ will turn out to be. Our approach in Subsection 4.3 for constructing the distribution of $E(AGPA_i^{t^*})^1$ involved: (1) describing the updating rule for beliefs about $AGPA_{i,j}$ as a function of $E(AGPA^1_{i,j})$ and a signal of $AGPA_{i,j}$ (constructed from actual grade performance), and (2) characterizing the person-specific amount of uncertainty that exists about the signal. From equations (7) and (8) we see that the prior $E(AGPA^1_{i,j})$ represents the mean of $E(AGPA_i^{t^*})^1$ and that the distribution of possible updates to $E(AGPA^1_{i,j})$ is determined by how much person-specific uncertainty exists about the signal.

It would be appealing to take an analogous approach to construct the distribution of $E(AINC_i^{t^*})^1$. For the age of 28 years, the expected value of the distribution describing beliefs about the average annual income i would receive if he had major j, $E(AINC^1_{i,j})$, is observed in the third column of Question 1. However, a difficulty arises because there does not exist a direct signal of $AINC_{i,j}$ that is analogous to the grade performance signals used for $AGPA_{i,j}$. One might consider an updating rule that is a function of other indirect information that might influence beliefs. However, we find that the most obvious of this other information—actual grade

28. This says that a person's belief about the average grade he would receive in, say, a science class in the first year corresponds to the expected grade that the person would receive in the future if he had SCI as a major. This assumption abstracts from the reality that the survey question for $E(AGPA^1_{i,j})$ does not ask a person to consider only the SCI classes that would be taken as part of the major. Nonetheless, to a rough approximation it seems reasonable to believe that a student thinking about what classes he would take in a future semester if he had a specific major within our SCI category would believe that most of his classes would fall in this specific major or in other majors within SCI. This would be especially true after the first couple of semesters when students take the majority of their liberal arts (General Studies) requirements.

performance—does not have a quantitatively strong relationship to updating. This implies that the use of such an updating rule would amount to imposing an assumption that very little updating about average income would take place during school. That is, a person would expect to update very little from $E(AINC^1_{i,j})$, a situation that seems less than ideal since we simply may not be able to observe the same signals that are observed by the agent.

Analogous to the approach in Subsection 4.3, we use the prior $E(AINC^1_{i,j})$ as the mean of $E(AINC^{t^*}_i)^1$. However, we take a somewhat different approach for characterizing the distribution of possible updates to $E(AINC^1_{i,j})$ that could take place by t^*. At various times during school, Survey Question 4 allows us to observe a student's total uncertainty about his income at the age of 28 years—by eliciting the lower bound, upper bound, 25th percentile, 50th percentile, and 75th percentile of the distribution describing beliefs about income at age of 28 years under a graduation scenario. This total uncertainty comes from three sources: (1) uncertainty about updates to beliefs about average income that will take place as of t^* (resolvable uncertainty), (2) uncertainty about updates to beliefs about average income that will take place after t^* (unresolvable uncertainty), and (3) transitory yearly fluctuations in earnings that exist for a person with a given value of $AINC_{i,j}$. What is needed to characterize the distribution of possible updates is to isolate the first source. To do this, we make an assumption that the proportion of the uncertainty that person i believes will be resolved about j before he makes his decision at t^* corresponds to the proportion of uncertainty that is actually resolved about j, on average, in the sample. Information about the proportion of uncertainty that is actually resolved, on average, in the sample, can be obtained by comparing how much total uncertainty exists, on average, at the time of entrance to how much total uncertainty exists, on average, at t^*. Appendix C describes the approach in more detail, including the manner in which we deal with the reality that we do not elicit the 25th, 50th, and 75th percentiles or bounds separately by major.

4.5. *Descriptive evidence that AGPA and AINC are likely to matter*

Our estimation of equation (6) in Subsection 4.3 showed that $E(AGPA^1_{i,j})$ plays a prominent role in the construction of $E(AGPA^{t^*}_i)^1$, the random variable that directly enters our full model of Subsection 4.2. Then, before estimating our full model, we can provide some informal, descriptive evidence that beliefs about $AGPA_i$ will likely turn out to be important determinants of $(Pr_{i,j}: j \in J^D)$. In Columns 3–9 of Table 3, we see that, for every j and every $k \neq j$, the average $E(AGPA^1_{i,j}) - E(AGPA^1_{i,k})$ is substantially larger for students with $STATE_i = j$ than it is for students with any other stated major. For example, Column 3 shows that students with $STATE_i = SCI$ have an average $E(AGPA^1_{i,SCI})$ that is quantitatively and significantly larger than the average $E(AGPA^1_{i,k})$ for each $k \neq SCI$. In contrast, Columns 4–9 show that, for students with any stated major j other than SCI, the average $E(AGPA^1_{i,SCI})$ is substantially lower than the average $E(AGPA^1_{i,j})$. To facilitate the description further, the second panel of Table 3 includes two measures which aggregate the outcomes AG, BUS, ED, HUM, PRO, SS into a single non-science (NON-SCI) major grouping. The first measure (measure 1) constructs $E(AGPA^1_{i,NON\text{-}SCI})$ for person i as a weighted average of $E(AGPA^1_{i,j})$ across all $j \neq SCI$, where the weight associated with j is the probability $Pr_{i,j}$ conditional on $j \neq SCI$. Students with $STATE_i = SCI$ have average $E(AGPA^1_{i,SCI}) - E(AGPA^1_{i,NON\text{-}SCI}) = 0.11$, whereas students with a stated major other than SCI (and also not equal to DROP) have average $E(AGPA^1_{i,SCI}) - E(AGPA^1_{i,NON\text{-}SCI}) = -0.70$. This difference arises because students with $STATE_i = SCI$ believe that they are much stronger academically in SCI than other students (average $E(AGPA^1_{i,SCI})$ of 3.53 versus 2.74), rather than because they believe

they are weaker in NON-SCI subjects than other students (average $E(AGPA^1_{i,NON\text{-}SCI})$ of 3.42 versus 3.44). Similar conclusions are reached in the last row of the second panel using the second NON-SCI measure (measure 2) in which $E(AGPA^1_{i,NON\text{-}SCI})$ is constructed as $\max_j\{E(AGPA^1_{i,j}): j\neq SCI\}$.

The results in the second panel of Table 3 also present a convenient way to see that beliefs about $AGPA_i$ may play an important role in determining $Pr_{i,DROP}$. Students with $STATE_i = DROP$ have a sample average of $E(AGPA^1_{i,SCI})$ that is 0.88 lower (2.65 versus 3.53) than the sample average $E(AGPA^1_{i,SCI})$ for students with $STATE_i = SCI$ and is 0.09 lower (2.65 versus 2.74) than the sample average for students with $STATE_i = NON\text{-}SCI$. Similarly, the last two rows of Panel two show that students with $STATE_i = DROP$ have a sample average of $E(AGPA^1_{i,NON\text{-}SCI})$ that is lower than the sample average of $E(AGPA^1_{i,NON\text{-}SCI})$ for students with $STATE_i = SCI$ or for students with $STATE_i = NON\text{-}SCI$.

Similar to what was seen in Table 3 for $AGPA_{i,j}$, Table 6 suggests that beliefs about $AINC_{i,j}$ may be an important determinant of the initial beliefs ($Pr_{i,j}: j\in J^D$). Focusing again on the major outcomes in Columns 3–9, we again see that, for every major j and every major $k\neq j$, the sample average of $E(AINC^1_{i,j})$–$E(AINC^1_{i,k})$ is larger for students with a stated major of j than for students with any other stated major. The second panel of Table 6 again facilitates comparisons by including two measures which aggregate the outcomes AG, BUS, ED, HUM, PRO, SS into a single non-science (NON-SCI) major grouping. For either of the two measures of $E(AINC^1_{i,NON\text{-}SCI})$ we find that, as compared with individuals with $STATE_i = j$, $j \neq SCI$, students with $STATE_i = SCI$ expect substantially higher future earnings from a SCI degree but very similar earnings from a NON-SCI degree. Thus, as before we see that students with $STATE_i = SCI$ believe they have a comparative advantage in SCI, but not necessarily an absolute disadvantage in NON-SCI. Beliefs about income are also related to beliefs about dropout. Compared with other students, students with $STATE_i = DROP$ have (statistically) more optimistic beliefs about their future income if they leave school immediately; the sample average $E(AINC^1_{i,DROP})$ is \$36,400 for $STATE_i = DROP$, is \$29,600 for $STATE_i = SCI$, and is \$29,900 for $STATE_i = NON\text{-}SCI$.

4.6. *Identification and estimation*

4.6.1. The choice set. For our primary estimation of the model in equations (1)–(5), we consider a further aggregated (Agg) choice set in which all non-science majors are combined into a single NON-SCI group: $J^{D,Agg} = \{DROP, SCI, NON\text{-}SCI\}$ with $Pr_{i,NON\text{-}SCI} = \sum_{j\in\{AG, BUS,ED,HUM,PRO,SS\}} Pr_{i,j}$. Subsection 4.2 suggested reasons related to the timing of choices that may make this appealing and Subsection 4.3 suggested that, at least in terms of updating, there are substantial similarities among the non-science majors. In addition, there are practical considerations that make this specification appealing. First, given that our model implies that i's probability of choosing each major j will be strictly between 0 and 1, we must adjust reported probabilities $Pr_{i,j}$ that are either 0.0 or 1.0.[29] The need to do this becomes less common with the further aggregated choice set. Second, as will be discussed in more detail, equation (5) shows that, for any set of parameter values, computing the likelihood contribution for i requires solving for a set of zeros. Since the number of zeros depends on the number of elements of the choice set, the burden of doing this is reduced non-trivially with a reduction in the choice set.

29. In practice we assume that $Pr_{i,j} = 0.99$ if i reports 1.0 and we assume that $Pr_{i,j} = 0.01$ if i reports 0.0. Changing these assumptions slightly has little effect. This issue is described in Blass et al. (2010) who implement a strategy that does not require this type of adjustment.

TABLE 6
Sample mean and standard deviation of $E(AINC^{1}_{i,j})$

	Average (std. dev.)								
	1	2	3	4	5	6	7	8	9
Full sample	Full sample $n = 632$	$STATE_i$ = DROP $n = 38$	$STATE_i$ = SCI $n = 138$	$STATE_i$ = AG $n = 72$	$STATE_i$ = BUS $n = 90$	$STATE_i$ = ED $n = 55$	$STATE_i$ = HUM $n = 134$	$STATE_i$ = PRO $n = 99$	$STATE_i$ = SS $n = 94$
j = DROP	29.8 (23.7)	36.4 (23.2)	29.6 (29.4).	32.8 (25.0)	37.0 (32.7)	27.4 (21.8)	26.5 (15.9)	26.4 (16.8)	28.2 (17.4)
j = SCI	52.0 (29.3)	47.8 (16.8)	65.8 (41.3)	41.7 (18.2)	52.0 (23.9)	41.9 (17.7)	50.8 (25.8)	51.5 (26.2)	48.1 (23.5)
j = AG	34.8 (15.3)	33.4 (12.9)	36.1 (17.3)	40.1 (19.4)	38.1 (13.9)	31.7 (10.5)	31.4 (12.3)	35.7 (15.3)	31.4 (13.7)
j = BUS	51.6 (27.5)	50.0 (19.5)	50.2 (35.0)	44.4 (19.2)	63.3 (29.2)	49.3 (27.4)	54.4 (35.0)	49.6 (23.3)	50.3 (30.3)
j = ED	33.3 (14.9)	34.8 (16.8)	33.5 (18.5)	30.3 (11.8)	35.4 (13.9)	34.3 (10.1)	31.9 (12.0)	32.6 (14.4)	33.0 (15.5)
j = HUM	37.8 (18.6)	38.0 (15.0)	35.8 (18.8)	32.6 (12.7)	40.6 (19.4)	34.6 (11.2)	43.5 (21.7)	37.1 (18.7)	35.5 (15.5)
j = PRO	53.5 (38.5)	52.2 (23.8)	55.5 (47.4)	47.5 (24.8)	55.8 (28.0)	45.2 (20.4)	53.4 (46.9)	64.7 (42.5)	45.3 (21.1)
j = SS	48.5 (31.8)	48.6 (21.6)	49.9 (32.9)	41.4 (19.4)	49.7 (25.2)	43.2 (17.9)	51.0 (47.9)	46.8 (24.7)	51.1 (22.8)
		$STATE_i$ = DROP $n = 38$	$STATE_i$ = SCI $n = 138$	$STATE_i = k$ $k \neq SCI$ $k \neq DROP$ $n = 486$					
j = DROP		36.4 (0.23)	29.6 (29.4)	29.9 (21.9)					
j = SCI		47.8 (16.8)	65.8 (41.3)	48.2 (24.0)					
j = NON-SCI (Measure 1)		46.1 (16.5)	47.0 (24.0)	48.9 (24.4)					
j = NON-SCI (Measure 2)		61.9 (23.9)	63.6 (48.6)	66.5 (57.1)					
Graduates	Full sample $n = 398$		$STATE_i$ = SCI $n = 92$	$STATE_i$ = AG $n = 40$	$STATE_i$ = BUS $n = 57$	$STATE_i$ = ED $n = 32$	$STATE_i$ = HUM $n = 86$	$STATE_i$ = PRO $n = 56$	$STATE_i$ = SS $n = 65$
j = SCI	51.3 (27.9)		60.9 (36.5)	38.3 (16.9)	52.2 (23.1)	41.2 (14.6)	51.2 (26.9)	51.3 (27.3)	49.7 (21.8)
j = AG	33.9 (13.6)		34.5 (15.5)	37.6 (18.6)	36.6 (13.0)	33.2 (11.5)	30.8 (10.2)	35.0 (11.5)	31.0 (11.1)
j = BUS	50.6 (26.4)		50.1 (23.9)	41.5 (18.3)	58.4 (23.3)	51.7 (32.1)	52.9 (33.8)	49.0 (22.7)	49.2 (25.7)
j = ED	33.1 (14.7)		33.9 (19.2)	28.1 (11.8)	35.0 (11.1)	35.5 (11.2)	31.2 (10.2)	33.9 (16.4)	33.5 (14.2)
j = HUM	37.7 (19.0)		34.5 (18.3)	30.7 (12.6)	41.9 (21.2)	35.8 (11.7)	43.4 (21.8)	37.2 (20.1)	35.2 (13.7)
j = PRO	53.7 (41.3)		57.0 (55.0)	43.6 (20.1)	55.5 (24.3)	47.4 (20.9)	56.5 (54.7)	58.2 (30.6)	47.6 (20.6)
j = SS	48.6 (34.8)		49.7 (33.1)	38.2 (17.3)	49.7 (25.3)	44.6 (17.5)	53.3 (56.1)	43.7 (17.7)	50.5 (22.6)
			$STATE_i$ = SCI $n = 92$	$STATE_i = k$ $k \neq SCI$ $n = 302$					
j = SCI			60.9 (36.5)	48.4 (24.1)					
j = NON-SCI			46.6 (23.8)	47.3 (21.5)					
j = NON-SCI (Measure 2)			65.1 (54.8)	64.0 (44.5)					

For the overall sample (Column 1) and for the sample stratified by $STATE_i$ (Columns 2–9) the table shows the sample mean (sample std. dev.) $E(AINC_{i,j})$ for each j. If $Pr_{i,j} = Pr_{i,k}$ are tied as the highest probabilities, then beliefs enter in both the column associated with j and the column associated with k.

4.6.2. Identification. With all of the information needed to characterize the distributions of $E(AGPA_i^{[*]})^1$ and $E(AINC_i^{[*]})^1$ constructed outside the model as described in Subsections 4.3 and 4.4, the parameters to be estimated are those in equation (1). In terms of the coefficients in equation (1), we make a standard normalization by setting $\alpha_{DROP} = 0$ so that α_{SCI} and $\alpha_{NON\text{-}SCI}$ are interpreted as effects of X_i on SCI and NON-SCI relative to the effects of X_i on the base case DROP. In terms of the unobservables, in the standard discrete choice case, $v_{i,j}$ and $\varepsilon_{i,j}$ would

be combined by necessity into a single random variable and the scale of the model would be fixed by normalizing the variance of this random variable. To illustrate how observing person-specific reports of the $Pr_{i,j}$'s influences identification issues related to $v_{i,j}$ and $\varepsilon_{i,j}$, we simplify the discussion by: (1) considering a binary choice set $J^{ND,AGG} = \{SCI, NON\text{-}SCI\}$ and (2) ignoring the effects of X_i and $AINC_{i,j}$ in equation (1). For concreteness, we assume that i reports $Pr_{i,SCI} = 0.20$.

Equation (5) shows that we must first find values of $\varepsilon_{i,SCI}$ and $\varepsilon_{i,NON\text{-}SCI}$ so that $Pr_{i,SCI} = 0.20$ and then evaluate how likely these values are. Rewriting equation (5) and inserting equation (3) for our simplified case,

$$\begin{aligned}&\text{Prob}(\varepsilon_{i,SCI}, \varepsilon_{i,NON\text{-}SCI} : \text{Prob}(E_{t^*}U^{SCI} > E_{t^*}U^{NON-SCI}) = 0.20\\ &= \text{Prob}(\varepsilon_{i,SCI}, \varepsilon_{i,NON\text{-}SCI} : \text{Prob}(\beta_1 E(AGPA^{t^*}_{i,SCI})^1 + v_{i,SCI} + \varepsilon_{i,SCI} > \beta_1 E(AGPA^{t^*}_{i,NON\text{-}SCI})^1 \\ &\quad + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}) = 0.20).\end{aligned} \tag{9}$$

Recalling that $\varepsilon_{i,SCI}$ and $\varepsilon_{i,NON\text{-}SCI}$ are known to the student and rearranging so that all random variables capturing the uncertainty of i appear on the left side,

$$\begin{aligned}&\text{Prob}(\varepsilon_{i,SCI}, \varepsilon_{i,NON\text{-}SCI} : \text{Prob}(\beta_1 E(AGPA^{t^*}_{i,SCI})^1 - \beta_1 E(AGPA^{t^*}_{i,NON\text{-}SCI})^1 \\ &\quad + v_{i,SCI} - v_{i,NON\text{-}SCI} > \varepsilon_{i,NON\text{-}SCI} - \varepsilon_{i,SCI}) = 0.20).\end{aligned} \tag{10}$$

Equation (10) shows that, as is standard in discrete choice models, it is the differenced unobservables $v_{i,diffSN} = v_{i,SCI} - v_{i,NON\text{-}SCI}$ and $\varepsilon_{i,diffSN} = \varepsilon_{i,SCI} - \varepsilon_{i,NON\text{-}SCI}$ that are of relevance. In practice, we will assume that differenced unobservables such as $v_{i,diffSN}$ and $\varepsilon_{i,diffSN}$ are normal, with means of zero and standard deviations $\sigma_{vdiffSN}$ and $\sigma_{\varepsilon diffSN}$ that do not vary across i. We assume for the discussion here that $E(AGPA^{t^*}_{i,SCI})^1$ and $E(AGPA^{t^*}_{i,NON\text{-}SCI})^1$ are also normal, although our estimation strategy does not always require/utilize this assumption. With equation (8) indicating that the mean of $E(AGPA^{t^*}_{i,SCI})^1 = E(AGPA^1_{i,SCI})$ and that the mean of $E(AGPA^{t^*}_{i,NON\text{-}SCI})^1 = E(AGPA^1_{i,NON\text{-}SCI})$, equation (10) can be rewritten as:

$$\text{Prob}\left(\varepsilon_{i,diffSN} : \text{Prob}\left(Z > \frac{-\varepsilon_{i,diffSN} + \beta_1 E(AGPA^1_{i,NON\text{-}SCI}) - \beta_1 E(AGPA^1_{i,SCI})}{\sqrt{\beta_1^2 \text{Var}(E(AGPA^{t^*}_{i,SCI})^1) + \beta_1^2 \text{Var}(E(AGPA^{t^*}_{i,NON\text{-}SCI})^1) + \sigma^2_{vdiffSN}}}\right) = 0.20\right), \tag{11}$$

where Z is a standard normal random variable. Noting that $\Pr(Z > 0.84) \approx 0.20$, denoting the ratio in equation (11) as • and setting this ratio equal to 0.84 allows us to solve for $\varepsilon^*_{i,diffSN}$, the value of $\varepsilon_{i,diffSN}$ that satisfies $\Pr(Z>\bullet) = 0.20$, as a function of β_1, $\sigma^2_{vdiffSN}$, $E(AGPA^1_{i,SCI})$, $E(AGPA^1_{i,NON\text{-}SCI})$, $\text{Var}(E(AGPA^{t^*}_{i,NON\text{-}SCI})^1)$, and $\text{Var}(E(AGPA^{t^*}_{i,SCI})^1)$. The likelihood contribution involves evaluating the density of $\varepsilon_{i,diffSN}$ at $\varepsilon^*_{i,diffSN}$. With $\varepsilon_{i,diffSN} \sim N(0, \sigma^2_{\varepsilon diff})$, the likelihood for person i is proportional to:

$$\exp\left(-\frac{1}{2}\left[\frac{\beta_1 E(AGPA^1_{i,NON\text{-}SCI})}{\sigma_{\varepsilon diffSN}} - \frac{\beta_1 E(AGPA^1_{i,SCI})}{\sigma_{\varepsilon diffSN}} - 0.84\sqrt{\frac{\beta_1^2 \text{Var}(E(AGPA^{t^*}_{i,SCI})^1))}{\sigma^2_{\varepsilon diffSN}} + \frac{\beta_1^2 \text{Var}(E(AGPA^{t^*}_{i,NON\text{-}SCI}))^1}{\sigma^2_{\varepsilon diffSN}} + \frac{\sigma^2_{vdiffSN}}{\sigma^2_{\varepsilon diffSN}}}\right]^2\right). \tag{12}$$

$E(AGPA^1_{i,SCI})$, $E(AGPA^1_{i,NON\text{-}SCI})$, $Var(E(AGPA^{t^*}_{i,NON\text{-}SCI})^1)$, and $Var(E(AGPA^{t^*}_{i,SCI})^1)$ can be thought of as data since, as described in Subsection 4.3, they are constructed from survey information. The parameters to be estimated are $\sigma_{vdiffSN}$, $\sigma_{\varepsilon diffSN}$, and β_1. From equation (12), it is evident that these parameters are not separately identified since doubling each of the parameters would leave the equation (12) unchanged. However, normalizing $\sigma_{vdiffSN}$ allows both β_1 and $\sigma_{\varepsilon diffSN}$ to be identified.

Returning to our primary case which uses $J^{D,Agg}$ instead of $J^{ND,Agg}$, equation (5) can now be written as a function of any two of $Pr_{i,DROP}$, $Pr_{i,SCI}$, and $Pr_{i,NON\text{-}SCI}$. Using the latter two, equation (5) is

$$\text{Prob}(\varepsilon_{i,SCI}, \varepsilon_{i,NON\text{-}SCI}, \varepsilon_{i,DROP} : \text{Prob}(E_{t^*}U^{SCI} > E_{t^*}U^{NON-SCI}, E_{t^*}U^{SCI} > E_{t^*}U^{DROP}) = \text{Prob}_{i,SCI} \text{ and} \quad (13)$$

$$\text{Prob}(E_{t^*}U^{NON-SCI} > E_{t^*}U^{SCI}, E_{t^*}U^{NON-SCI} > E_{t^*}U^{DROP}) = \text{Prob}_{i,NON\text{-}SCI}).$$

Expanding analogously to equation (9), $\varepsilon_{i,SCI} - \varepsilon_{i,NON\text{-}SCI}$, $\varepsilon_{i,SCI} - \varepsilon_{i,DROP}$, $\varepsilon_{i,NON\text{-}SCI} - \varepsilon_{i,SCI}$, and $\varepsilon_{i,NON\text{-}SCI} - \varepsilon_{i,DROP}$ appear in the inner probability expressions. However, $\varepsilon_{i,diffSN} = \varepsilon_{i,SCI} - \varepsilon_{i,NON\text{-}SCI}$ and $\varepsilon_{i,diffSD} = \varepsilon_{i,SCI} - \varepsilon_{i,DROP}$ are sufficient for determining all four of these differences. Similarly, the terms involving $v_{i,SCI}$, $v_{i,NON\text{-}SCI}$, and $v_{i,DROP}$ can be written as functions of the differences $v_{i,diffSN} = v_{i,SCI} - v_{i,NON\text{-}SCI}$ and $v_{i,diffSD} = v_{i,SCI} - v_{i,DROP}$. We normalize $Var(v_{i,diffSN}) = Var(v_{i,diffSD}) = 1$ and $corr(v_{i,diffSN}, v_{i,diffSD}) = 0$ and assume that $v_{i,diffSN}$ and $v_{i,diffSD}$ are normal. We assume that $\varepsilon_{i,diffSN}$ and $\varepsilon_{i,diffSD}$ have a joint normal distribution and estimate the standard deviation of $\varepsilon_{i,diffSN}$, denoted $\sigma_{\varepsilon diffSN}$, the standard deviation of $\varepsilon_{i,diffSD}$, denoted $\sigma_{\varepsilon diffSD}$, and the correlation between $\varepsilon_{i,diffSN}$ and $\varepsilon_{i,diffSD}$.

Estimation involves maximizing the likelihood function that comes from equation (5). As is standard, the parameters are updated using a Newton–Raphson algorithm. However, within this updating process, there is a second need for Newton–Raphson since, for each guess of the parameters during the updating process, it is necessary to find, for each person i, the values of $\varepsilon_{i,diffSN}$ and $\varepsilon_{i,diffSD}$ that satisfy the two Pr conditions in equation (13). From a computational standpoint, finding these zeros can be burdensome because, for each possible value of $\varepsilon_{i,diffSN}$ and $\varepsilon_{i,diffSD}$, computing $Pr(E_{t^*}U^{SCI} > E_{t^*}U^{NON\text{-}SCI}, E_{t^*}U^{SCI} > E_{t^*}U^{DROP})$ and $Pr(E_{t^*}U^{NON\text{-}SCI} > E_{t^*}U^{SCI}, E_{t^*}U^{NON\text{-}SCI} > E_{t^*}U^{DROP})$ requires the evaluation of the multi-dimensional integral seen in equation (5) over the random variables $E(AGPA^{t^*}_i)^1$, $E(AINC^{t^*}_i)^1$, $v_{i,diffSN}$, and $v_{i,diffSD}$. We evaluate these integrals by simulation, taking advantage of the results in Subsections 4.3 and 4.4 in order to draw from the distributions of $E(AGPA^{t^*}_i)^1$ and $E(AINC^{t^*}_i)^1$.

4.7. *Results*

4.7.1. Estimates. With DROP as the base choice, the parameters to be estimated are the equation (1) coefficients α_{SCI}, $\alpha_{NON\text{-}SCI}$, β_1, β_2 and the variance terms $\sigma_{\varepsilon diffSN}$, $\sigma_{\varepsilon diffSD}$ and $corr(\varepsilon_{i,diffSN}\varepsilon_{i,diffSD})$. α_{SCI} and $\alpha_{NON\text{-}SCI}$ are four-dimensional vectors with $X_i = \{1.0, MALE_i, MACT_i, \text{and } VACT_i\}$. Our estimation sample contains the 572 members of the full sample who have no missing information of any type.[30]

Given a desire to ensure robustness, Subsection 4.3 discussed two possible ways of aggregating the non-science majors when constructing the utility from NON-SCI and two possible signals

30. One reason the sample is somewhat smaller than what is seen in, for example, Table 3 is that some individuals in our first cohort did not provide $E(AGPA^1_{i,j})$ when they assigned $Pr_{i,j} = 0$. In Table 3, we used whatever information was provided by the "partial" cases (with sample sizes reflecting the total number of people contributing information) But, partial cases are not included in our estimation sample.

TABLE 7
Estimates of models of Pr_i from equations (1) to (5)

	(1) $n = 572$	(2) $n = 572$	(3) $n = 572$	(4) $n = 366$
MAJOR-SPECIFIC FACTORS				
$AGPA_{i,j}$	0.520 (0.034)*	0.511 (0.034)*	0.521 (0.032)	0.910 (0.085)*
$AINC_{i,j}/10$	0.060 (008)*	0.060 (0.008)*	0.062 (0.007)*	0.116 (0.027)*
SCIENCE				
Constant	−1.162 (0.115)*	−1.142 (0.117)*	−1.177 (0.107)*	−.948 (0.088)*
Male	−0.011 (0.069)	−0.006 (0.069)	−0.003 (0.069)	0.185 (0.128)
Math_ACT	0.015 (0.009)	0.017 (0.009)	0.016 (0.009)	0.029 (0.019)
Verbal_ACT	−0.021 (0.009)*	−0.019 (0.009)*	−0.019 (0.009)	−2.529 (0.018)*
NON-SCIENCE				
Constant	−0.490 (0.124)*	−0.465 (0.126)*	−0.384 (0.110)*	0.0 (normalization)
Male	−0.116 (0.066)	−0.109 (0.066)	−0.068 (0.067)	0.0 (normalization)
Math_ACT	−0.007 (0.009)	−0.005 (0.009)	−0.007 (0.009)	0.0 (normalization)
Verbal_ACT	−0.009 (0.008)	−0.008 (0.008)	−0.011 (0.008)	0.0 (normalization)
DROP				
Constant	0.0 (normalization)	0.0 (normalization)	0.0 (normalization)	NA
Male	0.0 (normalization)	0.0 (normalization)	0.0 (normalization)	NA
Math_ACT	0.0 (normalization)	0.0 (normalization)	0.0 (normalization)	NA
Verbal_ACT	0.0 (normalization)	0.0 (normalization)	0.0 (normalization)	NA
UNOBSERVABLES				
$\sigma_{\varepsilon diffSN}$	0.627 (0.018)*	0.626 (0.018)*	0.617 (0.108)*	1.139 (0.03)*
$\sigma_{\varepsilon iffSD}$	1.028 (0.039)*	1.034 (0.039)*	1.031 (0.039)*	
corr ($\varepsilon_{i,diffSN}$, $\varepsilon_{i,diffSd}$)	0.733 (0.022)*	0.731 (0.022)*	0.731 (0.022)*	
Log Likelihood	–1151.66	–1165.35	–1155.91	–566.97

Notes: Table shows estimate (std. error) from simulated maximum likelihood.
* Significant at 5%.
Column 1: $J^{D,AGG}$; $Sig_{i,j} = (GPA_{i,j}, GPA_{i,NON\text{-}j}.)$;
$E_{t^*}U^{NON\text{-}SCI} = Max_{j\in\{AG,BUS,ED,HUM,PRO,SS\}}\alpha_{NON\text{-}SCI}X_i + \beta_1 E(AGPA^{t^*}_{i,j}) + \beta_2(AINC^{t^*}_{i,j}) + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}$
Column 2: $J^{D,AGG}$; $Sig_{i,j} = (GPA_i)$;
$E_{t^*}U^{NON\text{-}SCI} = Max_{j\in\{AG,BUS,ED,HUM,PRO,SS\}}\alpha_{NON\text{-}SCI}X_i + \beta_1 E(AGPA^{t^*}_{i,j}) + \beta_2(AINC^{t^*}_{i,j}) + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}$
Column 3: $J^{D,AGG}$; $Sig_{i,j} = (GPA_{i,j}, GPA_{i,NON\text{-}j}.)$;
$E_{t^*}U^{NON\text{-}SCI}() = \alpha_{NON\text{-}SCI}X_i + \beta_1 E(AGPA^{t^*}_{i,NON\text{-}SCI}) + \beta_2 E(AINC^{t^*}_{i,NON\text{-}SCI}) + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}$ where $E(AGPA^{t^*}_{i,NON\text{-}SCI})$ and $E(AINC^{t^*}_{i,NON\text{-}SCI})$ are weighted averages of $E(AGPA^{t^*}_{i,j})$ and $E(AINC^{t^*}_{i,j})$ with weights given by $Pr_{i,j}$'s
Column 4: $J^{ND,AGG}$; $Sig_{i,j} = (GPA_{i,j}, GPA_{i,NON\text{-}j}.)$;
$E_{t^*}U^{NON\text{-}SCI} = Max_{j\in\{AG,BUS,ED,HUM,PRO,SS\}}\alpha_{NON\text{-}SCI}X_i + \beta_1 E(AGPA^{t^*}_{i,j}) + \beta_2(AINC^{t^*}_{i,j}) + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}$

that could enter the updating process related to $APGA_{i,j}$. Column 1 of Table 7 shows results in which: (1) a student knows that at t^* he will compare $E_{t^*}U^{DROP}()$ and $E_{t^*}U^{SCI}()$ to $E_{t^*}U^{NON\text{-}SCI} = Max_{j\in\{AG,BUS,ED,HUM,PRO,SS\}}[\alpha_{NON\text{-}SCI}X_i + \beta_1 E(AGPA^{t^*}_{i,j}) + \beta_2 E(AINC^{t^*}_{i,j}) + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}]$ and (2) the updated belief $E(AGPA^{t^*}_{i,j})$ for any $j\in J^{ND}$ uses the results from Table 5 where the signal involves both $GPA_{i,j}$ and $GPA_{i,NON\text{-}j}$. Note that in the definition of $E_{t^*}U^{NON\text{-}SCI}$ in (1), $AGPA_{i,j}$ and $AINC_{i,j}$ receive an identical treatment as would be present in a full (non-aggregated) model in the sense that i learns about these factors separately for each j before t^* and compares $\beta_1 E(AGPA^{t^*}_{i,j}) + \beta_2 E(AINC^{t^*}_{i,j})$ across j when making a decision.

Our primary interest is in the influence of $AGPA_{i,j}$ and $AINC_{i,j}$. These factors have a statistically significant effect on utility, with the coefficients β_1 and β_2, respectively, having *t*-statistics of 15.15 and 7.42, respectively. In terms of other observable influences on utility, a particular element of X_i may have an effect on utility from major j beyond its influence

through $AGPA_{i,j}$ and $AINC_{i,j}$ if, for example, the characteristic is related to how much a person enjoys major j. We find some evidence of this. For example, being male and having higher math ACT scores increases utility from SCI relative to NON-SCI; $\alpha_{SCI,MALE} - \alpha_{NON\text{-}SCI,MALE}$ has an estimate (t-stat) of 0.105 (1.88) and $\alpha_{SCI,MACT} - \alpha_{NON\text{-}SCI,MACT}$ has an estimate (t-stat) of 0.022 (2.78). Having higher verbal ACT scores increases utility in NON-SCI relative to SCI; $\alpha_{SCI,VACT} - \alpha_{NON\text{-}SCI,VACT}$ has an estimate (t-stat) of -0.012 (-2.88). In terms of unobservable influences on utility, recall that the model is estimated under a normalizing assumption about the variance of the resolvable portions of unobserved utility: $Var(v_{i,diffSN}) = Var(v_{i,diffSD}) = 1$. We find that the variance across students in the known (to students) unobservable portions of utility is somewhat similar in size with an estimate for $\sigma_{\varepsilon diffSN}$ of 0.627 and an estimate for $\sigma_{\varepsilon diffSD}$ of 1.028. Finally, we find that people who like SCI relative to NON-SCI also tend to like SCI relative to DROP, with $corr(\varepsilon_{i,diffSN}, \varepsilon_{i,diffSD})$ having an estimate (t-stat) of 0.73 (33.31).

As one robustness check, Column 2 of Table 7 is the same as Column 1 except that the updated belief $E(AGPA^{t^*}_{i,j})$ for any $j \in J^{ND}$ uses the results from Table 4 where the signal involves only the overall GPA_i. More specifically, we utilize the second and third panels of Table 4 so that the effect of GPA_i depends on whether or not $STATE_i = j$. As a second robustness check, Column 3 of Table 7 is the same as Column 1 except that a student knows that at t^* he will compare $E_{t^*}U^{DROP}()$ and $E_{t^*}U^{SCI}()$ to $E_{t^*}U^{NON\text{-}SCI}() = \alpha_{NON\text{-}SCI}X_i + \beta_1 E(AGPA^{t^*}_{i,NON\text{-}SCI}) + \beta_2 E(AINC^{t^*}_{i,NON\text{-}SCI}) + v_{i,NON\text{-}SCI} + \varepsilon_{i,NON\text{-}SCI}$ where $E(AGPA^{t^*}_{i,NON\text{-}SCI})$ and $E(AINC^{t^*}_{i,NON\text{-}SCI})$, respectively, are the weighted averages of $E(AGPA^{t^*}_{i,j})$ and $E(AINC^{t^*}_{i,j})$, respectively, across all $j \neq SCI$, where the weight associated with j is the student's reported probability $Pr_{i,j}$ conditional on $j \neq SCI$. The benefit of this specification is that, to the extent that decisions are influenced heavily by the $\varepsilon_{i,j}$'s, the weights contain some information about which values of $E(AGPA^{t^*}_{i,j})$ and $E(AINC^{t^*}_{i,j})$ are likely to be particularly important to i. Regardless, the estimates in Columns 2 and 3 are very similar to the estimates in Column 1 so, in what follows, we focus primarily on the results from Column 1.[31]

4.7.2 Are effects of $AGPA_{i,j}$ and $AINC_{i,j}$ quantitatively important? Whether misperceptions about $AGPA_i$ and $AINC_i$, if they exist, have the potential to explain the substantial differences in Columns 2 and 3 of Table 1 between perceived probabilities ($Pr_{i,j}$'s) and actual outcomes (j*'s) depends on the quantitative importance of $AGPA_{i,j}$ and $AINC_{i,j}$.[32] Given that it is hard to understand the quantitative importance of $AGPA_{i,j}$ and $AINC_{i,j}$ from examining parameter estimates alone, we provide some evidence about this issue here. For given distributions of $E(AGPA^{t^*}_i)^1$, $E(AINC^{t^*}_i)^1$ (and for the given normalized distributions of $v_{i,diffSN}$, and $v_{i,diffSD}$), i's values of $Pr_{i,SCI}$, $Pr_{i,NON\text{-}SCI}$, and $Pr_{i,DROP}$ are determined by his realizations of $\varepsilon_{i,diffSN}$ and $\varepsilon_{i,diffSD}$. Using simulation to integrate over the joint distribution $\varepsilon_{i,diffSN}$ and $\varepsilon_{i,diffSD}$ yields

31. We have also found that the estimates here and the conclusions of the paper, more generally, are not very sensitive to changing t^* (*e.g.* to $t^*=5$). Changing t^* involves reestimating the updating equations in Table 4 and Table 5. We find that the changes to Table 4 and Table 5 are not particularly large. In addition, the effect of any changes to Tables 4 and 5 on the estimates in Table 7 tends to be mitigated by the fact that at entrance students tend to understate the amount of uncertainty that exists about the grade performance signal (Subsection 4.3. showed that students tend to not anticipate bad grade outcomes that do actually sometimes occur). Recall that the distribution of $E(AGPA^{t^*}_i)^1$ is determined by the set of values of $E(AGPA^{t^*}_i)$ that could arise as a person uses Tables 4 and 5 to update his beliefs for each possible value of $Sig_{i,j}$ that he thinks is relevant. Then, changes to Tables 4 and 5 will tend to have less effect on the distribution of $E(AGPA^{t^*}_i)^1$ if a person is not very uncertain about $Sig_{i,j}$.

32. Note that we use the word misperception to describe a case where a student's beliefs are not accurate given what tends to happen in college. A different question, not examined here, is whether a student formed these beliefs in a reasonable way given the information that he received before arriving at college.

TABLE 8

Examining the quantitative importance of $AGPA_{i,j}$ and $AINCOME_{i,j}$: sample averages under various scenarios for $E(AGPA_i^1)$ and $E(AINC_i^1)$

Major j	(1) Actual PROB $j^* = j$	(2) Pr_{ij}	(3) Simulated Pr_{ij} Pr_{ij} baseline	(4) Simulated Pr_{ij} $E(AGPA^1_{i,SCI})$ High $E(AGPA^1_{i,k}),k\neq SCI$ High	(5) Simulated Pr_{ij} $E(AGPA^1_{i,SCI})$ Low $E(AGPA^1_{i,k}),k\neq SCI$ Low	(6) Simulated Pr_{ij} $E(AGPA^1_{i,SCI})$ Low $E(AGPA^1_{i,k}),k\neq SCI$ High	(7) Simulated Pr_{ij} $E(AGPA^1_{i,SCI})$ High $E(AGPA^1_{i,k}),k\neq SCI$ Low
j = SCI	0.082	0.166	0.172	0.230	0.169	0.093	0.285
j = NON-SCI	0.558	0.697	0.692	0.660	0.628	0.787	0.531
j = DROP	0.360	0.137	0.136	0.110	0.203	0.120	0.184
				Simulated Pr_{ij} $E(AINC^1_{i,SCI})$ High $E(AINC^1_{i,k}),k\neq SCI$ High	Simulated Pr_{ij} $E(AINC^1_{i,SCI})$ Low $E(AINC^1_{i,k}),k\neq SCI$ Low	Simulated Pr_{ij} $E(AINC^1_{i,SCI})$ Low $E(AINC^1_{i,k}),k\neq SCI$ High	Simulated Pr_{ij} $E(AINC^1_{i,SCI})$ High $E(AINC^1_{i,k}),k\neq SCI$ Low
j = SCI				0.181	0.172	0.149	0.214
j = NON-SCI				0.692	0.673	0.722	0.636
j = DROP				0.127	0.155	0.129	0.150

Notes: For the estimation sample from Table 7:

Column 1 shows the proportion of students with each outcome (from administrative data).

Column 2 shows the sample average of the $Pr_{i,j}$, j = SCI, NON-SCI,DROP (from survey data).

Column 3 shows simulated sample average values of $Pr_{i,j}$ under the baseline model in which no changes have been made.

Column 4 (1st panel): simulated sample average values of $Pr_{i,j}$ when $E(AGPA^1_{i,k})$, k = SCI, AG, BUS, ED, HUM, PRO, SS are each one-half of a standard deviation above their mean from Column 1 of Table 3.

Column 5 (1st panel): simulated sample average values of $Pr_{i,j}$ when $E(AGPA^1_{i,k})$, k = SCI, AG, BUS, ED, HUM, PRO, SS are each one-half of a standard deviation below their mean from Column 1 of Table 3.

Column 6 (1st panel): $E(AGPA^1_{i,SCI})$ one-half standard deviation below mean. $E(AGPA^1_{i,k})$, k = AG, BUS, ED, HUM, PRO, SS one-half standard deviation above mean.

Column 7 (1st panel): $E(AGPA^1_{i,SCI})$ one-half standard deviation above mean. $E(AGPA^1_{i,k})$, k = AG, BUS, ED, HUM, PRO, SS one-half standard deviation below mean.

The second panel of Columns 4–7 are the same as the first panel of Columns 4–7 except that changes are to $E(AINC^1_{i,k})$ rather than $E(AGPA^1_{i,k})$.

average values of $Pr_{i,SCI}$, $Pr_{i,NON\text{-}SCI}$, and $Pr_{i,DROP}$ for i. Averaging across i yields the simulated sample average probabilities in Table 8. Of interest is the sensitivity of the simulated sample average probabilities to changes in $E(AGPA_i^{t^*})^1$ and $E(AINC_i^{t^*})^1$.

For the estimation sample, the sample average values of $Pr_{i,SCI}$, $Pr_{i,NON\text{-}SCI}$, and $Pr_{i,DROP}$, respectively, from our survey data are 0.166, 0.697, and 0.137, respectively (Column 2, Table 8). As a check of our model, the baseline results in Column 3 of Table 8 show the simulated sample average probabilities under a scenario in which the distributions of $E(AGPA_i^{t^*})^1$ and $E(AINC_i^{t^*})^1$ are left unchanged (*i.e.* the distributions are constructed from the actual data as described in Subsections 4.3. and 4.4.).[33] The simulated sample average probabilities, 0.172, 0.692, and 0.136, are close to the actual sample averages.

Columns 4–7 of the first panel of Table 8 show results when we vary the mean $E(AGPA_i^1)$ of the distribution of $E(AGPA_i^{t^*})^1$. In Column 4 we assume that, for each $k\in J^{ND}$, $E(AGPA^1_{i,k})$ is one-half of a standard deviation above the sample mean observed in the first column, first panel of Table 3. In Column 5 we assume that, for each $k\in J^{ND}$, $E(AGPA^1_{i,k})$ is one-half of a standard deviation below

33. With the estimates coming from Column 1 of Table 7, the Column 1 note in Table 7 indicates that the construction uses $Sig_{i,j} = (GPA_{i,j}, GPA_{i,NON\text{-}j})$.

its observed sample mean. In Column 6 we assume that, for each k≠SCI, $E(AGPA^1_{i,k})$ is one-half of a standard deviation above its sample mean while $E(AGPA^1_{i,SCI})$ is one-half of a standard deviation below its sample mean. In Column 7 we assume that, for each k≠SCI, $E(AGPA^1_{i,k})$ is one-half of a standard deviation below its sample mean while $E(AGPA^1_{i,SCI})$ is one-half of a standard deviation above its sample mean. The results indicate that $Pr_{i,SCI}$ is very sensitive to changes in beliefs about grade performance. $Pr_{i,SCI}$ ranges from 0.093 to 0.285 across the four scenarios. Increasing $E(AGPA^1_{i,SCI})$ by one standard deviation while leaving $E(AGPA^1_{i,k})$ unchanged for all k≠SCI leads to an increase in $Pr_{i,SCI}$ of $0.285-0.169=0.116$ (69%) if $E(AGPA^1_{i,k})$ is one-half of a standard deviation below its sample mean for all k≠SCI (Col. 7 versus Col. 5) and leads to an increase in $Pr_{i,SCI}$ of $0.230-0.093=0.137$ (147%) if $E(AGPA^1_{i,k})$ is one-half of a standard deviation above its sample mean for all k≠SCI (Col. 4 versus Col. 6). These changes in $Pr_{i,SCI}$ are of similar magnitude as the difference observed between the average $Pr_{i,SCI}$ observed in the survey data (0.166, Col. 2) and the actual proportion of students who have the outcome of SCI (0.082, Col. 1). This leaves open the possibility that substantial misperceptions, if they exist, about $E(AGPA^1_i)$ could generate the type of overoptimism about SCI observed at entrance.

Columns 4–7 also show that $Pr_{i,DROP}$ can be influenced by changes in beliefs about grade performance, although substantial increases in $Pr_{i,DROP}$ require that beliefs about grade performance become considerably less positive in all non-science majors. For example, decreasing $E(AGPA^1_{i,k})$ for all k≠SCI by one standard deviation while leaving $E(AGPA^1_{i,SCI})$ unchanged leads to an increase in $Pr_{i,DROP}$ of 0.083 (69%) if $E(AGPA^1_{i,SCI})$ is one-half of a standard deviation below its sample mean (Col. 6 versus Col. 5) and leads to an increase in $Pr_{i,DROP}$ of 0.074 (67%) if $E(AGPA^1_{i,SCI})$ is one-half of a standard deviation above its sample mean (Col. 4 versus Col. 7). However, unlike what was seen for $Pr_{i,SCI}$, the changes that are generated to $Pr_{i,DROP}$, while non-trivial, are small relative to the difference between the average $Pr_{i,DROP}$ observed in the survey data (0.137, Col. 2) and the actual proportion of people who drop out (0.360, Col. 1). Misperceptions about $E(AGPA^1_i)$, if they exist, may have difficulty explaining the overoptimism about graduation observed at entrance.

Columns 4–7 of the second panel of Table 8 show results analogous to those in the first panel when we vary the mean $E(AINC^1_i)$ of the distribution of $E(AINC^{t^*}_i)^1$. There exists a non-trivial role for $AINC_{i,j}$, but one that is substantially smaller than what was seen for $AGPA_{i,j}$. For example, increasing $E(AINC^1_{i,SCI})$ by one standard deviation while leaving $E(AINC^1_{i,k})$ unchanged for all k≠SCI leads to an increase in $Pr_{i,SCI}$ of 0.042 (24%) if $E(AINC^1_{i,k})$ is one-half of a standard deviation below its sample mean for all k≠SCI (Col. 7 versus Col. 5) and leads to an increase in $Pr_{i,SCI}$ of 0.032 (21%) if $E(AINC^1_{i,k})$ is one-half of a standard deviation above its sample mean for all k≠SCI (Col. 4 versus Col. 6).

4.7.3. What would $Pr_{i,j}$'s be under true distributions of $AGPA_i$ and $AINC_i$. Subsection 4.7.2 showed that, generally speaking, the reported probabilities (Pr_i's) are sensitive to beliefs about $AGPA_i$ and $AINC_i$. Here we show that misperceptions about $AGPA_i$ and $AINC_i$ exist at entrance, and quantify the extent to which these misperceptions can explain the substantial difference between the average $Pr_{i,SCI}$, $Pr_{i,NON\text{-}SCI}$, and $Pr_{i,DROP}$ observed in the survey data (Col. 2, Table 8) and the proportion of students who have the outcomes SCI, NON-SCI, and DROP (Col. 1, Table 8).

We start by focusing on $AGPA_i$. Uncertainty about one's final outcome arises, in part, because i faces uncertainty about $AGPA_i$ that will be resolved by time t^*. For any $j \in J^{ND}$, this uncertainty is given by the distribution of the random variable $E(AGPA^{t^*}_{i,j})^1$ from Subsection 4.3. The distribution

of $E(AGPA^{t^*}_{i,j})^1$ may be biased when compared with the true distribution of $AGPA_{i,j}$ in the population. Our objective here is to simulate how the reported probabilities $Pr_{i,SCI}$, $Pr_{i,NON\text{-}SCI}$, and $Pr_{i,DROP}$ would have been different if, for each i and j, the distribution of $E(AGPA^{t^*}_{i,j})^1$ were replaced during estimation by a distribution representing the true values of $AGPA_{i,j}$ across students who are similar to i in observable ways.

To characterize this true distribution we again take advantage of transcript data, now utilizing course and grade information over a student's entire time in school. As in Subsection 4.3, we make the simplifying assumption that a student forms beliefs about future grades in major j by primarily considering his classes in j, not the classes he might take that would fall in some other major.[34] We assume that i's grade G in a particular class k in a particular major $j \in J^{ND}$ is given by the permanent/transitory structure

$$G_{i,j,k} = \text{constant}^G_j + \eta^G X_i + \theta^G_{i,j} + u^G_{i,j,k}. \tag{14}$$

With $\theta^G_{i,j}$ representing the mean-zero permanent portion of unobserved performance and $u^G_{i,j,k}$ representing the mean-zero transitory portion of unobserved performance, we view equation (14) as a random effects specification. Then, the object of interest—i's long-run average GPA in major j—is given by

$$AGPA_{i,j} = \text{constant}^G_j + \eta^G X_i + \theta^G_{i,j}. \tag{15}$$

With one equation for each major, we estimate the system of the seven equations defined by $\{G_{i,j,k}: j \in J^{ND}\}$. To allow for possible correlations in $\theta^G_{i,j}$ across majors j, we assume that $\theta^G_{i,j}$: $j \in J^{ND}$) has a multivariate normal distribution. We assume that $u^G_{i,j,k}$ is independent across j and k and that $u^G_{i,j,k} \sim N(0, Var(u^G_{i,j}))$.

We are able to separate the importance of the permanent and transitory components for a particular major j because, for some students i, we observe $G_{i,j,k}$ for multiple courses k. Our estimates of the parameters in equation (14) associated with a particular major j will be identified by individuals who take classes in j. Given that we are interested in the distribution of $AGPA_{i,j}$ across all students, one might worry, for example, that students who are good at major j will be more likely to select into taking classes in j. However, this concern is seemingly mitigated to a non-trivial extent in our context because: (1) with respect to our major of primary interest SCI, due (in part) to SCI requirements that are part of the liberal arts curriculum, almost all students take at least one SCI course and the majority of students take two or more SCI courses and (2) with respect to the other majors, we show momentarily that there exists a very strong correlation across the set of non-science majors (which suggests that grade performance in $j \neq SCI$ for those who take classes in j may be similar to what grade performance in j would have been for those who did not take classes in j).[35] Further, if students who would receive bad grades in SCI tend

34. Footnote 28 discusses this assumption. Here we have done some experimentation to make sure that this assumption is not driving our subsequent finding that misperceptions are able to explain why students are much less likely to end up in SCI than expected. As we will discuss, this result depends to a large extent on the misperceptions of students who have $STATE_i = SCI$ but have an outcome $j^* \neq SCI$. Table 3 shows that students with $STATE_i = SCI$ have high values of $E(AGPA^1_{i,SCI})$, but also have high values of $E(AGPA^1_{i,j})$, $j \neq SCI$. Then the beliefs about performance in SCI of this crucial group is not very sensitive to exactly how many classes this group considers taking outside of SCI as part of a SCI major.

35. For the full sample we find that only 11% of students take zero SCI classes and 63% take two or more SCI classes. For the sample of graduates we examine in Subsection 4.7.4 we find that only 5% of students take zero SCI classes and 73% take two or more SCI classes.

to avoid science classes, our estimates of equation (15) would be too positive about the true distribution of $AGPA_{i,SCI}$. Thus, adjusting the distribution might further strengthen our finding that misperceptions are important. Regardless, it is necessary to view the estimates of equation (15) with appropriate caution.

Column 1 of Table 9 shows simulated maximum likelihood estimates of the parameters defining the system $\{G_{i,j,k}: j \in J^{ND}\}$ assuming that $\theta^G_{i,j}$ is uncorrelated across j and that $\eta^G = 0$. Two particular results stand out. First, the variance of $AGPA_{i,j}$ as measured by $Var(\theta^G_{i,j})$ is substantially higher for j = SCI than for the other majors; $Var\theta^G_{i,j}) = 0.72, 0.19, 0.50, 0.30, 0.41, 0.26$, and 0.56,

TABLE 9

Random effects estimation of grade equations from equation (14) (Cols. 1–5) and income equations from equation (16) (Cols. 6–7)

	Average (standard error)						
Note: x = G in Cols. 1–5 and x = INC in Cols. 6–7.	(1) Dependent variable $G_{i,j,k}$	(2) Dependent variable $G_{i,j,k}$	(3) Dependent variable $G_{i,j,k}$	(4) Dependent variable $G_{i,j,k}$	(5) Dependent variable $G_{i,j,k}$	(6) Dependent variable $INC_{i,j,k}/1000$	(7) Dependent variable $INC_{i,j,k}/1000$
$constant^x_{DROP}$						28.61 (1464)*	25.71 (2.37)*
$constant^x_{SCI}$	2.368 (0.05)*	2.555 (0.081)*	2.519 (0.085)*	2.658 (0.058)*	2.656 (0.068)*	40.21 (2.639)*	37.90 (3.75)*
$constant^x_{AG}$	3.392 (0.04)*	3.514 (0.052)*	3.498 (0.059)*	3.559 (0.037)*	3.611 (0.044)*	29.28 (3.09)*	31.51 (4.25)*
$constant^x_{BUS}$	2.722 (0.08)*	2.922 (0.125)*	2.925 (0.015)*	2.901 (0.072)*	3.027 (0.077)*	38.84 (2.57)*	34.69 (3.98)*
$constant^x_{ED}$	3.374 (0.11)*	3.540 (0.144)*	3.445 (0.106)*	3.516 (0.092)*	3.579 (0.074)*	33.01 (1.83)*	32.71 (2.13)*
$constant^x_{HUM}$	3.077 (0.04)*	3.230 (0.060)*	3.230 (0.070)*	3.296 (0.037)*	3.362 (0.050)*	26.68 (1.69)*	26.90 (2.50)*
$constant^x_{PRO}$	3.084 (0.04)*	3.233 (0.064)*	3.202 (0.065)*	3.202 (0.045)*	3.284 (0.054)*	40.07 (2.81)*	37.32 (3.30)*
$constant^x_{SS}$	2.737 (0.04)*	2.920 (0.081)*	2.871 (0.083)*	2.936 (0.053)*	3.042 (0.061)*	31.88 (2.31)*	32.79 (2.69)*
$Var(\theta^x_{i,DROP})$						152.34	147.37
$Var(\theta^x_{i,SCI})$	0.718 (0.073)*	0.583 (0.059)*	0.722 (0.075)*	0.449 (0.058)*	0.345 (0.049)*	257.05	271.59
$Var(\theta^x_{i,AG})$	0.191 (0.018)*	0.169 (0.018)*	0.228 (0.024)*	0.079 (0.010)*	0.082 (0.014)*	216.97	228.31
$Var(\theta^x_{i,BUS})$	0.503 (0.085)*	0.468 (0.077)*	0.615 (0.099)*	0.245 (0.059)*	0.195 (0.047)*	350.81	354.47
$Var(\theta^x_{i,ED})$	0.296 (0.044)*	0.225 (0.067)*	0.402 (0.084)*	0.135 (0.045)*	0.119 (0.035)*	90.11	81.88
$Var(\theta^x_{i,HUM})$	0.411 (0.041)*	0.316 (0.033)*	0.410 (0.041)*	0.156 (0.022)*	0.151 (0.021)*	228.94	234.09
$Var(\theta^x_{i,PRO})$	0.263 (0.039)*	0.193 (0.031)*	0.302 (0.042)*	0.145 (0.026)*	0.146 (0.025)*	519.38	511.21
$Var(\theta^x_{i,SS})$	0.557 (0.065)*	0.384 (0.044)*	0.598 (0.067)*	0.283 (0.037)*	0.219 (0.038)*	186.73	151.21
$Var(u^x_{i,DROP})$						85.72	85.72
$Var(u^x_{i,SCI})$	0.672 (0.017)*	0.668 (0.017)*	0.658 (0.016)*	0.607 (0.017)*	0.604 (0.016)*	102.41	102.41
$Var(u^x_{i,AG})$	0.484 (0.013)*	0.482 (0.012)*	0.480 (0.012)*	0.378 (0.012)*	0.377 (0.012)*	63.69	63.69
$Var(u^x_{i,BUS})$	0.467 (0.030)*	0.465 (0.030)*	0.458 (0.028)*	0.451 (0.031)*	0.456 (0.033)*	54.16	54.16
$Var(u^x_{i,ED})$	0.431 (0.023)*	0.426 (0.022)*	0.404 (0.016)*	0.313 (0.020)*	0.308 (0.018)*	44.89	44.89
$Var(u^x_{i,HUM})$	0.582 (0.007)*	0.583 (0.007)*	0.575 (0.007)*	0.506 (0.007)*	0.504 (0.007)*	46.24	46.24
$Var(u^x_{i,PRO})$	0.430 (0.015)*	0.428 (0.015)*	0.418 (0.015)*	0.404 (0.015)*	0.398 (0.016)*	65.69	65.69
$Var(u^x_{i,SS})$	0.508 (0.017)*	0.505 (0.016)*	0.469 (0.015)*	0.442 (0.016)*	0.432 (0.016)*	24.31	24.31
MALE-DROP							2.79 (0.30)
MALE-SCI		−0.310 (0.102)*	−0.361 (0.090)*		0.038 (0.095)		3.55 (4.4)
MALE-AG		−0.227 (0.056)*	−0.250 (0.056)*		−0.056 (0.050)		−1.14 (6.86)
MALE-BUS		−0.234 (0.142)*	−0.509 (0.122)*		−0.179 (0.116)		7.11 (4.52)
MALE-ED		−0.447 (0.179)*	−0.388 (0.148)*		−0.236 (0.121)*		9.33 (5.17)*
MALE-HUM		−0.374 (0.071)*	−0.465 (0.069)*		−0.166 (0.058)*		1.12 (3.67)
MALE-PRO		−0.379 (0.079)*	−0.374 (0.077)*		−0.106 (0.078)		1.28 (6.66)*
MALE-SS		−0.547 (0.098)*	−0.575 (0.091)*		−0.212 (0.082)*		1.15 (4.92)
MACT-DROP							−0.93 (0.50)

(*continued*)

TABLE 9
continued

	Average (standard error)						
Note: x = G in Cols. 1–5 and x = INC in Cols. 6–7.	(1) Dependent variable $G_{i,j,k}$	(2) Dependent variable $G_{i,j,k}$	(3) Dependent variable $G_{i,j,k}$	(4) Dependent variable $G_{i,j,k}$	(5) Dependent variable $G_{i,j,k}$	(6) Dependent variable $INC_{i,j,k}/1000$	(7) Dependent variable $INC_{i,j,k}/1000$
MACT-SCI		0.066 (0.013)*	0.075 (0.013)*		0.047 (0.012)*		0.35 (0.69)
MACT-AG		0.030 (0.007)*	0.031 (0.007)*		0.023 (0.006)*		0.84 (0.96)
MACT-BUS		0.012 (0.020)	0.057 (0.018)*		0.013 (0.017)		−0.17 (9.22)
MACT-ED		0.043 (0.019)*	0.017 (0.018)		−0.013 (0.014)		−0.43 (0.51)
MACT-HUM		0.036 (0.101)*	0.039 (0.009)*		0.022 (0.007)*		−0.29 (0.46)
MACT-PRO		0.011 (0.011)	0.018 (0.010)		0.006 (0.009)		0.04 (0.83)
MACT-SS		0.046 (0.013)*	0.049 (0.012)*		0.028 (0.010)*		2.33 (0.72)*
VACT-DROP							0.84 (0.40)*
VACT-SCI		0.006 (0.014)	0.101 (0.011)		0.031 (0.011)*		0.23 (0.80)
VACT-AG		−0.005 (0.008)	−0.009 (0.007)		0.0003 (0.005)		−1.58 (1.12)
VACT-BUS		0.027 (0.017)	0.026 (0.015)		0.043 (0.012)*		0.49 (0.72)
VACT-ED		−0.021 (0.023)	0.003 (0.022)		0.035 (0.020)		−0.67 (0.60)
VACT-HUM		0.006 (0.008)	0.0001 (0.008)		0.012 (0.007)		−0.18 (0.52)
VACT-PRO		0.022 (0.011)*	0.015 (0.009)		0.026 (0.009)*		0.71 (0.74)
VACT-SS		0.032 (0.012)*	0.038 (0.01)*		0.048 (0.009)*		−1.06 (0.66)
$Cov(\theta^x_{i,SCI}, \theta^x_{i,AG})$			0.362 (0.036)*		0.133 (0.022)*		
			Corr 0.146				
$Cov(\theta^x_{i,SCI}, \theta^x_{i,BUS})$			0.616 (0.075)*		0.245 (0.037)*		
			Corr 0.410				
$Cov(\theta^x_{i,SCI}, \theta^x_{i,ED})$			0.444 (0.075)*		0.149 (0.037)*		
			Corr 0.238				
$Cov(\theta^x_{i,SCI}, \theta^x_{i,HUM})$			0.441 (0.045)*		0.140 (0.025)*		
			Corr 0.239				
$Cov(\theta^x_{i,SCI}, \theta^x_{i,PRO})$			0.395 (0.047)*		0.137 (0.027)*		
			corr 0.184				
$Cov(\theta^x_{i,SCI}, \theta^x_{i,SS})$			0.302 (0.042)*		0.212 (0.040)*		
			corr 0.387				
$Corr\theta^x_{i,j}, \theta^x_{i,k}) k \neq j$			corr 0.838(0.019)*		0.653 (0.042)*		
Log Like	−15463.71	−15284.16	−1487.79	−11817.18	−11542.43		

*Significant at 0.05.
Columns 1–3 show estimates of equation (14) for full sample.
Column 4–5 show estimates of equation (14) for graduate subsample.
Columns 6–7 show estimates of equation (16).

respectively, for j = SCI, AG, BUS, ED, HUM, PRO, SS, respectively. Second, comparing the average value of $AGPA_{i,j}$ for each major j given by $constant^G_j$ to the sample average perceived value of $E(AGPA^1_{i,j})$ for each major j as given in Column 1 of Table 3, we see that students were particularly overoptimistic at entrance about SCI. The difference between $constant^G_j$ (Table 9) and $E(AGPA^1_{i,SCI})$ (Col. 1, Table 3) is −0.54 for SCI and is, on average, −0.05 for the other six majors (ranging between −0.40 and 0.34). Column 2 of Table 9 shows results including X_i. We see evidence in support of the importance of MACT and see that, conditional on MACT and VACT, women perform significantly better than males in all majors.[36] Finally, Column 3 relaxes the assumption that $\theta_{i,j}$ is uncorrelated across j. We reduce the dimension of the covariance matrix

36. A student's math ACT score is a statistically significant predictor of performance in five of the seven majors, having the largest *t*-statistic (and largest estimate) for SCI. The estimated effects across the seven majors have, on average, a *t*-statistic of 2.79. In contrast, a student's verbal ACT score is a statistically significant predictor of grade performance

TABLE 10
Examining the quantitative importance of $AGPA_{i,j}$ and $AINC_{i,j}$. Sample averages under various scenarios for $E(AGPA_i^1)$ and $E(AINC_i^1)$

	(1) actual Prob(j* = j) Note: same as Col. 1 Table 8	(2) Pr_{ij} Note: same as Col. 2 Table 8	(3) Simulated Pr_{ij} baseline Note: same as Col. 3 Table 8	(4) Simulated Pr_{ij} Note: $E(AGPA_i^{t^*})^1$ replaced by true dist. of $AGPA_i$ from equation (15)	(5) Simulated Pr_{ij} Note: $E(AINC_i^{t^*})^1$ replaced by true dist. of $AINC_i$ from equation (17)	(6) Simulated Pr_{ij} Note: $E(AGPA_i^{t^*})^1$ replaced by dist. of $AGPA_i$ from equation (15) & $E(AINC_i^{t^*})^1$ replaced by dist. of $AINC_i$ from equation (17)
Full sample $j \in J^{D,AGG}$						
SCI	0.082	0.166	0.172	0.121	0.164	0.104
NON-SCI	0.558	0.697	0.692	0.717	0.677	0.717
DROP	0.36	0.137	0.136	0.161	0.157	0.178
Grad sample $j \in J^{ND,AGG}$						
SCI	0.119	0.189	0.183	0.133	0.181	0.114
NON-SCI	0.881	0.811	0.816	0.867	0.818	0.886

Notes: For the estimation sample from Table 7:
Column 1 shows the proportion of students with each outcome (from administrative data).
Column 2 shows the sample average of the $Pr_{i,j}$, j = SCI, NON-SCI,DROP (from survey data).
Column 3 shows simulated sample average values of $Pr_{i,j}$ under the baseline model in which no changes have been made.
Column 4 shows simulated sample average probabilities when $E(AGPA_i^{t^*})^1$ is replaced by the "true"distribution of $AGPA_i$ from equation (15).
Column 5 shows simulated sample average probabilities when $E(AINC_i^{t^*})^1$ is replaced by the true distribution of $AINC_i$ from equation (17).
Column 6 shows simulated sample average probabilities when $E(AGPA_i^{t^*})^1$ is replaced by the "true" distribution of $AGPA_i$ from equation (15) and $E(AINC_i^{t^*})^1$ is replaced by the true distribution of $AINC_i$ from equation (17).

by estimating $COV(\theta_{i,SCI}, \theta_{i,j})$ for each $j \in \{AG, BUS, ED, HUM, PRO, SS\}$, but by constraining the correlation to be the same across all pairs of non-science majors: $corr(\theta_{i,j}, \theta_{i,k}) = corr(\theta_{i,m}, \theta_{i,n})$ for all $j,k,m,n \in \{AG, BUS, ED, HUM, PRO, SS\}$, $j \neq k$, $m \neq n$.[37] In Subsection 4.3, when examining the determinants of the update $E(AGPA_{i,j}^{t^*})$, we found that grades in classes outside of j tend to be more informative when $j \neq SCI$ than when j=SCI and concluded that students believe that the set of majors other than SCI require a somewhat common set of abilities/knowledge, with this set somewhat different than what is required for SCI. We find direct support for this notion here; $corr(\theta_{i,SCI}, \theta_{i,j})$ ranges from 0.146 to 0.410 across $j \neq SCI$, but the common $corr(\theta_{i,j}, \theta_{i,k})$, $j,k \neq SCI$, $j \neq k$ is 0.84.

Using the joint distribution of $\{AGPA_{i,j}: j \in \{SCI, AG, BUS, ED, HUM, PRO, SS\}\}$ constructed from equation (15) in place of the belief distribution $E(AGPA_i^{t^*})^1$ constructed from observed beliefs, Column 4 of Table 10 shows simulated sample average probabilities for $Pr_{i,SCI}$, $Pr_{i,NON\text{-}SCI}$, and $Pr_{i,DROP}$ constructed as in Subsection 4.7.2. Compared to the baseline simulations in Column 3, we see a decrease in $Pr_{i,SCI}$ from 0.172 to 0.121. Thus, replacing perceptions about the average grade performance with the measure of the true distribution of average grade performance explains about $56\% = 0.172 - 0.121/(0.172 - 0.082)$ of the difference

in only two of the seven majors and the estimated effects across the seven majors have, on average, a *t*-statistic of 0.829. Part of the explanation for the gender difference is the higher level of effort of females as detailed in S&S (2012).

37. The models in this subsection are estimated by simulated maximum likelihood.

between the baseline simulated average value of $Pr_{i,SCI}$ (Column 3, Table 10) and the actual proportion of matriculants who have an outcome of SCI (column 1, Table 10).

That the simulation in Column 4 does not explain even more of the difference between the SCI numbers in Columns 1 and 3 of Table 10 can be attributed to the fact that the majority of the dramatic difference in DROP between Columns 1 and 2 is left unexplained (we simulate a change in $Pr_{i,DROP}$ of $0.161-0.136=0.025$, whereas the difference is $0.360-0.136=0.224$). Indeed, Column 4 is quite successful in correcting misperceptions about the choice of SCI relative to the choice of NON-SCIENCE. The probability of SCI conditional on not dropping out, $Pr_{i,SCI}/(Pr_{i,SCI} + Pr_{i,NON\text{-}SCI})$, is 0.199 in our baseline simulation (Column 3), is 0.128 in reality (Column 1), and is 0.144 in our Column 4 simulation. In Subsection 4.7.4, we discuss what our previous research suggests about why the expectations data collected at entrance may be better suited for studying major choice than for understanding dropout. However, before doing this, we examine whether misperceptions exist about $AINC_{i,j}$, and, if so, whether these misperceptions can further explain the differences between Columns 1 and 3 of Table 10.

Survey Question 1 eliciting $E(AINC^1_{i,j})$ asks about earnings at the age of 28 years. We use our post-college annual survey to examine actual earnings around the age of 28 years. Analogous to equations (14) and (15), we use $INC_{i,j,k}$, i's actual earnings observed at the ages of k = 26, 27, 28 to estimate, for each j, the random effects specification

$$INC_{i,j,k} = constant^{INC}_j + \eta^{INC} X_i + \theta^{INC}_{i,j} + u^{INC}_{i,j,k}. \quad (16)$$

The object of interest—i's average income in major j—is given by

$$AINC_{i,j} = constant^{INC}_j + \eta^{INC} X_i + \theta^{INC}_{i,j}. \quad (17)$$

Since students are only observed with income in a single major, it is not possible to allow $\theta^{INC}_{i,j}$ to be correlated across j. More generally, from the standpoint of selection, our estimation of equation (16) is less appealing than our estimation of equation (14) (where students are observed with grades in different majors). Indeed, concerns about selection are one of the primary motivations for collecting expectations data because such data allow one to observe beliefs about all potential outcomes. Nonetheless, little is known about the relationship between income expectations and actual long-run income outcomes.

Column 6 of Table 9 shows results with $\eta^{INC}=0$. Looking across majors for those who graduate, the mean of $AINC_{i,j}$ as measured by $constant^{INC}_j$ is largest for SCI, PRO, and BUS ($40,210, $40,070, and $38, 840) and is smallest for HUM ($26,680). Comparing these $constant^{INC}_j$'s to the perceptions $E(AINC^1_{i,j})$ in Col. 1 of Table 6 (and ignoring selection), we see that students entered school with a reasonable perception about the income ordering across the seven majors. For example, consistent with reality, they believed that SCI, PRO, and BUS are most lucrative. However, they may have somewhat overstated the returns to a degree. For example, in Column 1 of Table 6 we see a sample average $E(AINC^1_{i,DROP})$ of $29,800, whereas the sample average $E(AINC^1_{i,j})$ over j≠DROP is $44,500. In Column 6 of Table 9 we see an estimate for $constant^{INC}_{DROP}$ of $28,610, whereas the average $constant^{INC}_j$ over j≠DROP is only $34,280.

Intuitively, the fall in the return to schooling that is present when we simulate the model using the distribution of $(AINC_{i,j}: j\in J^D)$, constructed from equation (17) (using estimates in Column 7 of Table 9 with $\eta^{INC}\neq 0$), in place of the distribution of $E(AINC^{t^*}_i)^1$, constructed from survey data, should increase dropout. However, the simulation results from Subsection 4.7.2 suggest that the effect will likely be small (relative to the very large amount of unexpected dropout). The results

in Column 6 of Table 10, where we modify the specification in Column 4 by using the distribution of ($AINC_{i,j}$: $j \in J^D$), constructed from equation (17), in place of the distribution of $E(AINC_i^{t^*})^1$, constructed from survey data, indicate that this is the case. Compared to Column 4, the simulated average $Pr_{i,DROP}$ increases from 0.161 to 0.178. The simulated average $Pr_{i,SCI}$ decreases further to 0.104 so that our results are now fully successful in correcting misperceptions about the choice of SCI relative to the choice of NON-SCIENCE; the probability of SCI conditional on not dropping out, $Pr_{i,SCI}/(Pr_{i,SCI}+Pr_{i,NON\text{-}SCI})$, is 0.127 (Column 6) compared to 0.128 in reality (Column 1). More generally, the results in Columns 4–6 show that misperceptions about grade performance play a more important role than misperceptions about income.

4.7.4. Direct evidence of learning using $J^{ND,AGG}$ = {SCIENCE, NON-SCIENCE}. Subsection 4.7.3 shows that our model is successful in explaining why beliefs about $Pr_{i,SCI}/(Pr_{i,SCI}+Pr_{i,NON\text{-}SCI})$ are much different than the actual proportion of graduating students who have the outcome SCI. However, our model has difficulty explaining the large amount of dropout that is present relative to what students expected. There are natural reasons for this. One issue detailed in S&S (2012) is that students assign little probability to the very poor grade outcomes that are most strongly tied to dropout. In this case, examining in Subsection 4.7.3 how the $Pr_{i,j}$'s would be influenced by replacing the distribution of $E(AGPA_{i,j}^{t^*})^1$ with the distribution describing the true values of $AGPA_{i,j}$ (as measured by equations (14) and (15)), may, in essence, involve some extrapolation which could have difficulty capturing fully the particularly large effects of very poor grade performance. In addition, students may simply not anticipate the full set of avenues through which poor grade performance will influence them. In S&S (forthcoming) we use a dynamic programming model to understand the relative importance of various explanations for the strong relationship between grade performance and dropout. We find that grade performance operates to a large extent by influencing how enjoyable it is to be in school. It seems quite plausible that students may not anticipate how stressful/unenjoyable school might turn out to be when grade performance is bad, and this may lead them to believe incorrectly at entrance that they will "stick it out" no matter what happens.

Given that this discussion suggests that expectations data collected at entrance seems particularly useful for understanding choices among majors, it seems worthwhile to estimate the model using the choice set $J^{ND,AGG}$ = {SCI,NON-SCI}. We do this for the subset of our estimation sample who graduate and show the results in Column 4 of Table 7. The estimated effects of $AGPA_{i,j}$ and $AINC_{i,j}$ are almost twice as large as in Columns 1–3, but that is primarily because the estimate of $\sigma_{\varepsilon diffSN}$ is also almost twice as large. Indeed, the t-statistics associated with $AGPA_{i,j}$ and $AINC_{i,j}$ remain large (10.71 and 4.29). In the second panel of Table 10, we show simulations for this choice set that are analogous to the results for $J^{D,AGG}$ in the first panel of Table 10. Column 3 shows that the baseline simulated average $Pr_{i,SCI}$ of 0.183 is similar to the average reported $Pr_{i,SCI}$ of 0.189 (Column 2), but much different than the actual proportion who have SCI of 0.119 (Column 1). Column 6 shows that replacing $E(AGPA_i^{t^*})^1$ by the constructed true distribution of $AGPA_i$ from equation (15) (using estimates for the grad sample from Column 5 of Table 9) and $E(AINC_i^{t^*})^1$ by the constructed true distribution of $AINC_i$ from equation (17) (using estimates from Column 7 of Table 9) leads to a simulated average $Pr_{i,SCI}$ of 0.114. Thus, our model finds that misconceptions about $AGPA_i$ and $AINC_i$ can fully explain why graduating students are substantially overoptimistic about completing a degree in SCI. Columns 3–5 (2^{nd} panel Table 10) reveal that most of the movement in $Pr_{i,SCI}$ comes from misperceptions about $AGPA_i$, rather than misperceptions $AINC_i$. Comparing $constant_{SCI}^G$,…,$constant_{SS}^G$ in Column 4 of Table 9 to Col. 1, Panel 2 of Table 3 reveals that, as was seen before, misperceptions tend to be about grade performance in SCI rather than about grade performance in other majors.

The results in the previous paragraph suggest that misperceptions about $Pr_{i,SCI}$ are likely to exist because students start school with misperceptions about $AGPA_i$ and, to a lesser extent, $AINC_i$. A particular advantage of examining the subsample of graduates is that we are able to use surveys taken later in school to provide direct evidence about whether students actually update their beliefs about $AGPA_i$ and $AINC_i$ substantially by the end of school.[38] While earlier we suggested that it seems reasonable to assume that students believe at entrance that they will make a major choice rather quickly, beliefs near the end of school are likely to be informative about outcomes because students who do not fully appreciate how much uncertainty exists about their major (Section 3) are likely to end up choosing a major later than they expected.

We focus on the 335 (out of 366) graduating students for which, in addition to observing $E(AGPA_i^1)$ and $E(AINC_i^1)$, we also observe $E(AGPA_i^t)$ and $E(AINC_i^t)$ for some time after the start of the third year. We refer to the last date at which this information is observed for i as t^{**}.[39] The first column of Table 11 shows $E(AGPA_i^1)$ in the first panel and $E(AINC_i^1)$ in the second panel. The second column of Table 11 shows $E(AGPA_i^{t^{**}}) - E(AGPA_i^1)$ and $E(AINC_i^{t^{**}}) - E(AGPA_i^1)$. The sample mean of -0.15 for SCI in the first panel of Column 2 indicates that, on average, students do adjust their beliefs about $AGPA_{i,SCI}^1$ downward during school. The means associated with the other majors $j \neq SCI$ range from -0.049 to 0.107 and for each of these majors j we reject, at a 5% level of significance, the null that $E(AGPA_{i,SCI}^{t^{**}}) - E(AGPA_{i,SCI}^1)$ is the same as $E(AGPA_{i,j}^{t^{**}}) - E(AGPA_{i,j}^1)$. The evidence in Subsection 4.7.3 showed that students enter school overly optimistic about the future income associated with all majors. The sample means in the second panel of Column 2 show that students revise their beliefs about earnings downward over time in all majors.

Section 3 showed that many students with $STATE_i = SCI$ have an outcome other than SCI and that very few students who have $STATE_i \neq SCI$ end up with an outcome of SCI. The descriptive statistics in Columns 3–8 of the first panel of Table 11 examine whether heterogeneity in updating about $AGPA_i$ is consistent with these patterns by stratifying the sample into three groups: those who have $STATE_i = SCI$ and $j^* = SCI$, those who have $STATE_i = SCI$ but $j^* \neq SCI$, and those who have $STATE_i \neq SCI$.[40] Consistent with the finding that many students leave SCI, the first row of Table 11 shows that the $STATE_i = SCI$, $j^* \neq SCI$ group start school with a sample average $E(AGPA_i^1)$ that is very similar to the $STATE_i = SCI$, $j^* = SCI$ group (3.53 versus 3.63) but, as a result of having an average $E(AGPA_{i,SCI}^{t^{**}}) - E(AGPA_{i,SCI}^1)$ of -0.61 (Col. 6), end school with a sample average $E(AGPA_i^{t^{**}})$ that is much more similar to the $STATE_i \neq SCI$ group (2.92 versus 2.71). Of relevance for understanding the finding that few people move into SCI after entrance, Column 8 shows that the $STATE_i \neq SCI$ group has an average (std. dev.) $E(AGPA_{i,SCI}^{t^{**}}) - E(AGPA_{i,SCI}^1)$ of -0.066 (0.785). Although the sample mean is slightly negative, one might think that, given the very large standard deviation, some students might learn that they are very good at SCI. To understand why we do not see more changes into SCI, we disaggregated Column 8 further by stratifying on whether a students was in the bottom quartile or top three quartiles in terms of $E(AGPA_{i,SCI}^1)$. We find that the positive updating tends to be concentrated to a large extent in the (former) group of students who had very low initial expectations, and, as a result, were likely not close to the margin of choosing SCI. For example, the sample average $E(AGPA_{i,SCI}^{t^{**}})$

38. One could attempt to characterize beliefs at the time of exit for those who drop out. But, in practice, this is difficult because, for example, our surveys eliciting major information take place at the beginning of semesters and much dropout occurs between semesters.

39. Given high response rates, the majority of observations come from the second semester of the 4th year.

40. Note that it might be desirable to further divide the last group on the basis of a person's major j^*. However, in practice this is not useful since so few people who have $STATE_i \neq SCI$ have $j^* = SCI$.

TABLE 11
Beliefs at entrance (t=1) and end of school (t^{**}): sample mean and standard deviation

	Sample mean (std. dev.)							
	1	2	3	4	5	6	7	8
	$E(AGPA^1_{i,j})$	$E(AGPA^{t^{**}}_{i,j}) - E(AGPA^1_{i,j})$	$STATE_i = SCI$ & $j^* = SCI$ $E(AGPA^1_{i,j})$	$STATE_i = SCI$ & $j^* = SCI$ $E(AGPA^{t^{**}}_{i,j}) - E(AGPA^1_{i,j})$	$STATE_i = SCI$ & $j^* \neq SCI$ $E(AGPA^1_{i,j})$	$STATE_i = SCI$ & $j^* \neq SCI$ $E(AGPA^{t^{**}}_{i,j}) - E(AGPA^1_{i,j})$	$STATE_i \neq SCI$ $E(AGPA^1_{i,j})$	$STATE_i \neq SCI$ $E(AGPA^{t^{**}}_{i,j}) - E(AGPA^1_{i,j})$
	$n = 335$	$n = 335$	$n = 32$	$n = 32$	$n = 40$	$n = 40$	$n = 263$	$n = 263$
j = SCI	2.944 (0.824)	−0.150 (0.750)	3.628 (0.294)	−0.270 (0.384)	3.532 (0.252)	−0.606 (0.535)	2.771 (0.841)	−0.066 (0.785)
j = AG	3.031 (0.680)	0.107 (0.710)	3.281 (0.491)	0.116 (0.500)	3.325 (0.493)	−0.100 (0.649)	2.956 (0.706)	0.138 (0.736)
j = BUS	3.079 (0.661)	−0.030 (0.688)	3.371 (0.423)	0.070 (0.402)	3.293 (0.410)	−0.161 (0.662)	3.011 (0.700)	2.988 (0.544)
j = ED	3.227 (0.599)	0.048 (0.677)	3.453 (0.392)	0.029 (0.442)	3.313 (0.420)	−0.063 (0.523)	3.186 (0.636)	0.068 (0.720)
j = HUM	3.190 (0.613)	−0.049 (0.645)	3.261 (0.506)	0.038 (0.659)	3.182 (0.543)	−0.065 (0.561)	3.182 (0.636)	−0.057 (0.656)
j = PRO	3.121 (0.616)	0.041 (0.639)	3.310 (0.400)	0.017 (0.483)	3.216 (0.499)	0.0025 (0.623)	3.084 (0.648)	0.049 (0.659)
j = SS	3.132 (0.691)	0.030 (0.688)	3.339 (0.452)	0.115 (0.510)	3.302 (0.400)	−0.155 (0.553)	3.082 (0.740)	0.048 (0.721)
	1	2						
	$E(AINC^1_{i,j})$	$E(AINC^{t^{**}}_{i,j}) - E(AINC^1_{i,j})$						
	$n = 335$	$n = 335$						
j = SCI	51.0 (38.1)	−12.9 (25.6)						
j = AG	33.3 (12.8)	−4.5 (13.0)						
j = BUS	49.9 (24.3)	−8.9 (27.3)						
j = ED	32.7 (13.3)	−3.8 (14.1)						
j = HUM	37.0 (17.7)	−6.8 (20.7)						
j = PRO	51.4 (25.7)	−12.5 (25.9)						
j = SS	46.2 (22.1)	−11.2 (21.8)						

– $E(AGPA^{1}_{i,SCI})$ is 0.761 for these students and is −0.326 for students in the top three quartiles. Thus, consistent with our earlier results, it seems difficult for students who initially believe they are good at science to learn that they are extremely good at SCI without taking many science classes.[41]

Finally, we examine whether the changes in beliefs about $AGPA_i$ and $AINC_i$ over time are consistent with the fact that, in our graduate estimation sample, Table 10 showed an average $Pr_{i,SCI}$ of 0.189 but that only 0.119 of students had $j^* = SCI$.[42] Our baseline simulation using the 335 student subsample finds a simulated average value for $Pr_{i,SCI}$ of 0.184. To simulate the actual proportion making the choice of SCI, we use our model to ask the question "what proportion of students would have chosen SCI if they were forced to choose their major at entrance and had the beliefs given by $E(AGPA^{t^{**}}_{i,j})$, $j \in J^{ND}$ and $E(AINC^{t^{**}}_{i,j})$, $j \in J^{ND}$ at that time?" Using the estimates in Column 4 of Table 7, our model finds that the proportion choosing SCI would be 0.104.[43] Thus, our model, when used with the revised beliefs, is able to capture the reality that the sample proportion with a major of SCI is substantially smaller than what was expected at entrance.

5. CONCLUSION

We find that students enter college as open to a major in science as to any other major, but that relatively few students finish school with science as their outcome. This occurs because, relative to other majors, students are both more likely to leave science (if they started in science) and are less likely to change into science (if they started in a major other than science). In terms of major-specific factors that influence the major decision, we find a particularly important role for future grade performance, with future income playing a statistically significant, but smaller role. As to why students leave science, departing students have typically learned that their future grade performance in science would be substantially worse than expected, with this reflecting learning about ability rather than learning about their willingness to study. As to why students do not often change into science, students who do not start in science typically do not learn that they are especially talented in science.

These patterns are consistent with our findings when we use transcript data to examine the process by which students update their beliefs about future major-specific grade performance. We find that taking courses in science is the primary way to learn about one's ability in science, with science being an outlier in that grades in courses taken outside of the major play a relatively uninformative role in what one learns about his ability in the major. It is possible that requiring additional courses in science during college might lead to more science graduates. However, simulations of our model show that, when we replace each student's beliefs at entrance about major-specific average grade performance (from survey data) with a distribution representing true major-specific average grade performance (computed from transcript data), students' beliefs about the probability of choosing science become consistent with the relatively low proportion of students who actually choose science. This suggests that, by and large, students are ultimately

41. For the stratified sample, our focus on $AGPA_i$ is motivated by our finding in previous sections that this factor is particularly important for explaining major choice. S&S (2012) and S&S (2013) examines $AINC_i$ for these same stratification groups—finding differences between groups that are similar in spirit, but smaller in magnitude.

42. These numbers are very similar for the 335 person subsample.

43. If a person is forced at entrance to choose a major he will not have the opportunity to learn about $v_{idiffSN} = v_{i,SCI} - v_{i,NON\text{-}SCI}$ so, analogous to the treatment of $AGPA^{t^*}_{i,SCI}$ and $AINC^{t^*}_{i,SCI}$ in equation (3), what will now be relevant for $v_{i,diffSN}$ will be the expectation $E(v_{diffSN}) = 0$. $\varepsilon_{i,diffSN} = \varepsilon_{i,SCI} - \varepsilon_{i,NON\text{-}SCI}$ is known to the individual so, as before, our simulations integrate out with respect to $\varepsilon_{i,diffSN}$. Thus, what is calculated is the average probability that each person will choose SCI. Averaging over the sample gives us our simulated sample proportion.

choosing science in numbers that are roughly consistent with their abilities at the time of college entrance. As such, if more science graduates are desired, the findings suggest the importance of policies at younger ages that lead students to enter college better prepared to study science.

APPENDIX A: SURVEY QUESTIONS

Question 1 (beginning of first semester). We realize that you may not be sure what area of study you will eventually graduate with. In the first column below are listed possible areas of study. In the second column write down the per cent chance that you will end up with this area of study. In the third column, please write down the grade point average (GPA) you would expect to receive in a typical semester in the future if you had each of these areas of study. In the fourth column write down the **yearly** income you would expect to earn at age of 28 years (or 10 years from now if you are now 20 years of age or older) if you graduated with each of these areas of study. In the fifth column, write down how many hours you would expect to study/do homework (outside of class) on a typical day in the future if you had each of these areas of study.

Please fill out all remaining columns even if you have a zero in the per cent chance column for a particular area of study.

Humanities include Art, English, Foreign Languages, History, Music, Philosophy, Religion, and Theatre.

Natural Science and Math includes Biology, Chemistry, Computer Science, Physics, and Mathematics.

Professional Programmes include Industrial Arts, Industrial Technology, Child Development, Dietetics, Home Economics, Nutrition, and Nursing.

Social Sciences include Economics, Political Science, Psychology, and Sociology.

Area of study	Per cent chance (See above)	Expected GPA (0.00–4.00) **	Expected Yearly Income Age 28 (in dollars)	Expected study/ homework on a typical (in hours)
1. Agriculture (and Natural Resources)	_____	_____	_____	_____
2. Business	_____	_____	_____	_____
3. Elementary Education	_____	_____	_____	_____
4. Humanities	_____	_____	_____	_____
5. Natural Science & Math	_____	_____	_____	_____
6. Professional Programs	_____	_____	_____	_____
7. Social Sciences	_____	_____	_____	_____

Note: Numbers in the second column (per cent chance) should each be between 0 and 100 and should add up to 100.
Note: A = 4.0, B = 3.0, C = 2.0, D = 1.0, F = 0.0. So numbers in third column (GPA) should be between 0.00 and 4.00.

Question 2a.

What it the per cent chance that you will eventually graduate from Berea College?_____

Question 2b.

What is the per cent chance that you will complete more than one year of school at Berea College? That is, what is the per cent chance that you will return to Berea for at least the beginning of your second year? _____

Question 3.

We realize that you do not know exactly how well you will do in classes. However, we would like to have you describe your beliefs about the grade point average that you expect to receive in the first semester. Given the amount of study-time you indicated above, please tell us the per cent chance that your grade point average will

be in each of the following intervals. That is, for each interval, write the number of chances out of 100 that your final grade point average will be in that interval.

Note: The numbers on the six lines must add up to 100.

Interval	**Per cent Chance (number of chances out of 100).**
[3.5, 4.00]	_____
[3.0, 3.49]	_____
[2.5, 2.99]	_____
[2.0, 2.49]	_____
[1.0, 1.99]	_____
[0.0, .99]	_____

Note: A = 4.0, B = 3.0, C = 2.0, D = 1.0, F = 0.0

APPENDIX B: CONSTRUCTING BELIEFS ABOUT THE SECOND (DISAGGREGATED) SIGNAL OF AGPA

In order to construct beliefs about $GPA_{i,j}$ for all j, we characterize i's beliefs about his grade performance in each of the M_i individual classes that he takes. We note that these beliefs can be aggregated up to form beliefs about $GPA_{i,j}$ for any j.

We refer to the random variable which describes i's beliefs about the grade he will receive in class m, m=1,…,M_i as $GPACL_{i,m}$. For each class m, we define a latent analog to $GPACL_{i,m}$.

$$GPACL^*_{i,m} = E\left(GPACL^*_{i,m}\right) + v^*_{i,m}, \tag{A.1}$$

where $v^*_{i,m} \sim N(0, Var(v^*_{i,m}))$ and the relationship between the latent $GPACL^*_{i,m}$ and the actual $GPACL_{i,m}$ is given by

$$\begin{aligned} GPACL_{i,m} &= GPACL^*_{i,m} \text{ if } 0 < GPACL^*_{i,m} < 4.0 \\ &4.0 \text{ if } GPACL^*_{i,m} \geq 4.0 \\ &0.0 \text{ if } GPACL^*_{i,m} \leq 0.0. \end{aligned} \tag{A.2}$$

Given the relationship in (A.2), i's beliefs about $GPACL_{i,m}$, m = 1,…,M_i are fully determined by $E(GPACL^*_{i,m})$ and $Var(v^*_{i,m})$, m = 1,…,M_i. The objective here is to find values of $E(GPACL^*_{i,m})$ and $Var(v^*_{i,m})$, m = 1,…,M_i that are most consistent with what is observed about $GPACL_{i,m}$, m = 1,…,M_i in the survey data.

To describe what is observed about $GPACL_{i,m}$, m = 1,…,M_i in the survey data let Major(m) be a function which identifies the major associated with any particular class m = 1,2,…,M_i. For any m such that Major(m)$\in J^{ND}$ a natural assumption described in Subsection 4.3. is that $E(GPACL_{i,m}) = E(AGPA^1_{i,Major(m)})$. There are some classes such that Major(m)$\notin J^{ND}$. These classes fall under Berea's General Studies (liberal arts) heading. We refer to these general studies courses as GEN so that for classes m from the general studies curriculum Major(m) = GEN. Survey question 1 provides no direct information about i's beliefs about his expected grade performance in a given General Studies course. However, this expectation can be computed because it is the only unknown in the equation $E(GPA_i) = (1/M_i)\sum E(GPACL_{i,m})$ since $E(GPA_i)$, a person's expectation about overall GPA, is observed in survey Question 3. Thus, we can think of $E(GPACL_{i,m})$ as being observed for each m. Var($GPACL_{i,m}$), m = 1,…,M_i is not observed directly. However, beliefs about var(GPA_i) as given by var([$GPACL_{i,1} + GPACL_{i,2} + \ldots + GPACL_{i,M}$]/$M_i$) is observed in survey Question 3.

Then, roughly speaking, our goal is to choose $E(GPACL^*_{i,m})$, m = 1,…,M_i and $Var(v^*_{i,m})$, m = 1,…,M_i so that equations A.1 and A.2 produce values that satisfy, as closely as possible: (1) $E(GPACL_{i,m}) = E(AGPA^1_{i,Major(m)})$, m = 1,…,$M_i$ and (2) var([$GPACL_{i,1} + GPACL_{i,2} + \ldots + GPACL_{i,M}$]/$M_i$).

For any choice of $Var(v^*_{i,m})$, m = 1,…,M_i there exist values of $E(GPACL^*_{i,m})$, m = 1,…,M_i such that, given A.1 and A.2, $E(GPACL_{i,m}) = E(AGPA^1_{i,Major(m)})$ for all m. Then, imposing that $E(GPACL_{i,m}) = E(AGPA^1_{i,Major(m)})$ for all m, the question is how to choose $Var(v^*_{i,m})$, m = 1,…,M_i. We begin by imposing restrictions to reduce the dimensionality of the problem. In terms of the covariance terms, purely for simplicity we assume that $COV(v^*_{i,m}, v^*_{i,n}) = 0$ for all courses

m and n. In terms of the variance terms, we assume that, if course m is from major j and course n is from major k, then $\text{var}(\text{GPACL}^*_{i,m})/\text{var}(\text{GPACL}^*_{i,n}) = a/b$ where a represents the variance of actual course grades in major j and b represents the variance of actual course grades in major k. Pooling all first year classes we observe that the variances of SCI, AG, BUS, ED, HUM, PRO, SS, and GEN are 1.48, 0.83, 1.38, 0.63, 1.02, 0.74, 1.25, and 0.98, respectively. Thus, we set $\text{Var}(v^*_{i,m})$ equal to $1.48\tau_i$, $.83\tau_i$, $1.38\tau_i$, $0.63\tau_i$, $1.02\tau_i$, $0.74\tau_i$, $1.25\tau_i$, and $0.982\tau_i$, respectively, if Major(m)=SCI, AG, BUS, ED, HUM, SS, PRO, and GEN, respectively. With these restrictions in place we choose the person-specific value of τ_i that makes the belief distribution of GPA_i coming from (A.1) and (A.2) (which is given by $(1/M_i)\sum\text{GPACL}_{i,m}$) as close to the reported belief distribution of GPA_i from survey Question 3.

APPENDIX C: CONSTRUCTING $\text{E}(\text{AINC}_i^{t^*})^1$

To illustrate the construction of $\text{E}(\text{AINC}_i^{t^*})^1$, we use an example that provides some concreteness. Consider a person i who has $\text{Pr}_{i,\text{SCI}} = 0.4$ and $\text{Pr}_{i,\text{SS}} = 0.6$. We assume that, when i reports, in Survey Question 4, the upper bound, lower bound, and quartiles associated with the distribution describing his beliefs about income at the age of 28 years, he does so assuming that there is a 0.40 chance of ending up with SCI and that, if this turns out to be the case, his beliefs about the income he will receive at the age of 28 years are given by $\text{E}(\text{AINC}^1_{i,\text{SCI}}) + v^1_{i,\text{SCI}}$. Here, $v^1_{i,\text{SCI}}$ is a mean-zero random variable that captures all sources of uncertainty about the income at the age of 28 years associated with SCI: resolvable uncertainty about $\text{AINC}_{i,\text{SCI}}$, unresolvable uncertainty about $\text{AINC}_{i,\text{SCI}}$, and transitory yearly fluctuations in earnings that exist for a person with average income $\text{AINC}_{i,\text{SCI}}$. Similarly, we assume that, when i provides the upper bound, lower bound, and quartiles associated with the distribution describing his beliefs about income at the age of 28 years, he does so assuming that there is a 0.60 chance of ending up with SS and that, if this turns out to be the case, his beliefs about the income he will receive at the age of 28 years are given by $\text{E}(\text{AINC}^1_{i,\text{SS}}) + v_{i,\text{SS}}$.

Since the quartiles above are not observed for individual majors, an additional assumption is needed about the relative importance of uncertainty across majors. We assume that the variance of the resolvable portion of $v^1_{i,j}$ is the same for person i across majors j. We make the same assumption about the unresolvable and transitory portions of $v^1_{i,j}$. Then, $\text{Var}(v^1_{i,j}) = \text{Var}(v^1_{i,k})$ for all j and k. Referring to this total variance as $\sigma^2_{v,i,1}$ and making the assumption that the $v^1_{i,j}$ are normally distributed, we can find the value of $\sigma^2_{v,i,1}$ that represents the best fit to the income distribution given by i for time t (Survey Question 4). Taking advantage of the fact that upper bound, lower bound, and quartiles associated with the distribution describing beliefs about income at the age of 28 years were also collected at $t^*=3$ allows us to compute a value of σ^2_{v,i,t^*}, the total variation at time t^*. $1-(\sigma^2_{v,i,t^*}/\sigma^2_{v,i,1})$ shows the percentage of the total variation that has been resolved (for each major) between t=1 and t^* for person i. On average, we find this ratio to be 0.21. Then, for each person i we assume that the variance of the resolvable uncertainty portion $v^1_{i,j}$ is given by $0.21\sigma^2_{v,i,1}$.

Acknowledgments. We are grateful for support from The Mellon Foundation, The Spencer Foundation, The National Science Foundation, SSHRC, and Berea College. This work would not have been possible without the extraordinary work of Lori Scafidi and the assistance of Diana Stinebrickner, Pam Thomas, and Albert Conley.

REFERENCES

ALTONJI, J. (1993), "The Demand for and Return to Education When Education Outcomes are Uncertain," *Journal of Labor Economics*, **11**, 48–83.

ARCIDIACONO, P. (2004), "Ability Sorting and the Returns to College Major", *Journal of Econometrics*, **121**, 343–375.

ARCIDIACONO, P., HOTZ, J. and KANG, S. (2012a), "Modeling College Major Choices Using Elicited Measures of Expectations and Counterfactuals", *Journal of Econometrics*, **166**, 3–16.

ARCIDIACONO, P., AUCEJO, E. and SPENNER, K. (2012b), "What Happens After Enrollment? An Analysis of the Time Path of Racial Differences in GPA and Major Choice", *IZA Journal of Labor Economics*, **1**, Article 5.

ARCIDIACONO, P. and KOEDEL, C. (2012), "Race and College Success: Evidence from Missouri" (NBER Working Paper 19188).

ARCIDIACONO, P., AUCEJO, E. and HOTZ, V. J. (2013), "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California" (Working Paper).

ARMANTIER, O., NELSON, S., TOPA, G., VAN DER KLAAUW, W. and BASIT, Z. (2012), "The Price is Right: Updating Inflation Expectations in a Randomized Price Information Experiment" *Federal Reserve Bank of New York*, (Staff Report, no. 543).

ATTANASIO, O. and KAUFMANN, K. (2009), "Educational Choices, Subjective Expectations, and Credit Constraints", (Working Paper 15087), National Bureau of Research.

BARSKY, R., KIMBALL, M., JUSTER, F. T., and SHAPIRO, M. (1997), "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Survey", *The Quarterly Journal of Economics*, 537–579.

BEFFY, M., FOUGERE, D. and MAUREL, A. (2013), "Choosing the Field of Study in Post-Secondary Education: Do Expected Earnings Matter?" *The Review of Economics and Statistics*, forthcoming.
BLASS, A., LACH, S., and MANSKI, C. (2010), "Using Elicited Choice Probabilities to Estimate Random Utility Models: Preferences for Electricity Reliability", *International Economic Review*.
BROWNING, M. and CARRO, J. (2007), "Heterogeneity and Microeconometrics Modelling" in Blundell, R. Newey W. and T. Perssson, *Advances in Economics and Econometrics*, Vol. 3, (Cambridge University Press).
CARNEIRO, P., HANSEN, K. and HECKMAN, J. (2003), "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effect of Uncertainty on Schooling Choice", *International Economic Review*.
CUNHA, F., HECKMAN, J. and NAVARRO, S. (2005), "Separating Uncertainty from Heterogeneity in Life Cycle Earnings", *Oxford Economic Papers*, **57**, 191–261.
COSEPUP (Committee on Science, Engineering, and Public Policy). (2007), "Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future", *The National Academies Press*.
DAEMPFLE, P. (2003), "An Analysis of the High Attrition Rates among First Year College Science, Math, and Engineering Majors", *Journal of College Student Retention: Research, Theory and Practice*, **5**, 37–52.
DAYMONT, T. and ANDRISANI, P. (1984), "Job Preferences, College Major, and the Gender Gap in Earnings", *Journal of Human Resources*, **19**, 408–428.
DELAVANDE, A. (2008), "Measuring Revisions to Subjective Expectations", *Journal of Risk and Uncertainty*, **36**.
DOMINITZ, J. (1998), "Earnings Expectations, Revisions, and Realizations", *The Review of Economics and Statistics*, 374–388.
DOMINITZ, J. and HUNG, A. (2009), "Empirical Models of Discrete Choice and Belief Updating in Observational Learning Experiments", *Journal of Economic Behavior and Organization*, **69**, 94–109.
DOMINITZ, J. and MANSKI, C. (1996), "Eliciting Student Expectations of the Returns to Schooling", 1–26.
DOMINITZ, J. and MANSKI, C. (1997), "Using Expectations Data to Study Subjective Income Expectations", *Journal of American Statistical Association*, 855–867.
FEDERMAN, M. (2007), "State Graduation Requirements, High School Course Taking, and Choosing a Technical College Major", *The B.E. Journal of Economic Analysis & Policy*, **7**, 1–32.
GROGGER, J. and EIDE, E. (1995), "Changes in College Skills and the Rise in the College Wage Premium", *Journal of Human Resources*, **30**, 280–310.
HAMERMESH, D. and DONALD, S. (2008), "The Effect of College Curriculum on Earnings: An Affinity Identifier for Non-Ignorable Non-Response Bias", *Journal of Econometrics*, **144**, 479–491.
JAMES, E., NABEEL, A., CONATY, J. and TO, D. (1989), "College Quality and Future Earnings: Where Should you Send Your Child to College?" *American Economic Review*, **79**, 247–252.
JENSEN, R. (2010), "The (perceived) Returns to Education and the Demand for Schooling", *Quarterly Journal of Economics*, **125**, 515–548.
JUSTER, T. (1966), "Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design", *Journal of the American Statistical Association*, **61**, 658–96.
KAUFMANN, K. (2009), "Understanding the Income Gradient in College Attendance in Mexico: The Role of Heterogeneity in Expected Returns" (Working Paper), Bocconi University.
LOURY, L. (1997), "The Gender-Earnings Gap Among College-Educated Workers", *Industrial and Labor Relations Review*, **50**, 580–493.
LOURY, L. and GARMAN, D. (1995) "College Selectivity and Earnings", *Journal of Labor Economics*, **13**, 289–308.
MALAMUD, O. (2010), "Breadth vs. Depth: The Timing of Specialization in Higher Education", *Labour*, **24**, 359–390.
MALAMUD, O. (2011), "Discovering One's Talent: Learning from Academic Specialization", *Industrial and Labor Relations Review*, **62**, 375–405.
MANSKI, C. (1989), "Schooling as Experimentation: A Reappraisal of the Post-secondary Drop-out Phenomenon", *Economics of Education Review*, **8**, 305–312.
MANSKI, C. (1990), "The Use of Intentions Data to Predict Behavior: A Best Case Analysis", *Journal of the American Statistical Association*, **85**, 934–940.
MANSKI, C. (2004), "Measuring Expectations", *Econometrica*, **72**, 1329–1376.
MONTMARQUETTE, C., CANNINGS, K. and MAHSEREDJIAN, S. (2002), "How Do Young People Choose College Majors?", *Economics of Education Review*, Elsevier, **21**, 543–556.
OST, B. (2010), "The Role of Peers and Grades in Determining Major Persistence in the Sciences", *Economics of Education Review*, **29**, 923–934.
RASK, K. (2010), "Attrition in STEM Fields at a Liberal Arts College: The Importance of Grades and Pre-Collegiate Preferences." *Economics of Education Review*, **29**, 892–900.
SABOT, R. and WAKEMAN-LINN, J. (1991), "Grade Inflation and Course Choice," *Journal of Economic Perspectives*, **5**, 159–70.
STINEBRICKNER, T. and STINEBRICKNER, R. (2004), "Time-Use and College Outcomes", *Journal of Econometrics*, **121**, 243–269.
STINEBRICKNER, T. and STINEBRICKNER, R. (2008a), "The Effect of Credit Constraints on the College Drop-Out Decision: A Direct Approach Using a New Panel Study", *American Economic Review*, **98**, 2163–2184.
STINEBRICKNER, T. and STINEBRICKNER, R. (2008b), "The Causal Effect of Studying on Academic Performance", *Frontiers in Economic Policy and Analysis* (*Frontiers*), Berkeley Electronic Press, **8**, 1–53.

STINEBRICKNER, T. and STINEBRICKNER, R. (2011), "Math or Science? Using Longitudinal Expectations Data to Examine the Process of Choosing a College Major", (NBER Working Paper, w16869).
STINEBRICKNER, T. and STINEBRICKNER, R. (2012), "Learning about Academic Ability and the College Drop-Out Decision", *Journal of Labor Economics*.
STINEBRICKNER, T. and STINEBRICKNER, R. (2013), "Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model", *Journal of Labor Economics*, forthcoming.
WISWALL, M. and ZAFAR, B. (2011), "Determinants of College Major Choice: Identification Using an Information Experiment", *Federal Reserve Bank of New York*, (Staff Report, no. 500).
ZAFAR, B. (2008), "College Major Choice and the Gender Gap" (Working Paper).
ZAFAR, B. (2013), "How Do College Students Form Expectations?" *Journal of Labor Economics*, forthcoming.