Ex Ante Returns and Occupational Choice*

Peter Arcidiacono† V. Joseph Hotz‡ Arnaud Maurel§ Teresa Romano¶

First version: October 2014
This version: March 21, 2020

Abstract

Using panel data on male undergraduates at Duke University, we make three contributions to the literature on subjective expectations. First, we show elicited data on earnings beliefs and probabilities of working in particular occupations are highly informative regarding actual earnings and occupational choices. Second, we show how subjective expectations data can be used to recover *ex ante* treatment effects as well as the relationship between these treatment effects and individual choices. We find large differences in expected earnings across occupations, and substantial heterogeneity across individuals in the corresponding *ex ante* returns. We also find clear evidence of sorting across occupations based on expected earnings; those who report higher probabilities of working in a particular occupation also report higher earnings in that occupation. Finally, while expected earnings influence occupational choice, we show that non-pecuniary factors also play an important role, with a sizable share of individuals expecting to give up substantial amounts of earnings by choosing occupations that are not their highest paying ones.

*We are grateful to the editor, Jim Heckman, and four anonymous referees for helpful comments. We also thank seminar participants at Arizona State University, UC-Berkeley, Carnegie Mellon, CREST, London School of Economics, New York Fed, Stanford University, University College London, University of Chicago, University of Oslo, University of Tennessee, University of Wisconsin - Madison, Uppsala University, Toulouse School of Economics and Washington University in St. Louis, as well as participants at the 2017 AEA Meetings (Chicago, IL), the 2015 AEA Meetings (Boston, MA), the NBER Labor Studies 2014 Spring Meeting, the Workshop on Subjective Expectations and Probabilities in Economics and Psychology (Essex, March 2014), the 2014 European Meeting of the Econometric Society (Toulouse), and the 2014 Duke Empirical Micro Jamboree for helpful comments and suggestions. Luis Candelaria, Ashley DeVore, Kate Maxwell Koegel and Jintao Sun provided excellent research assistance. This paper previously circulated under the title "Recovering *Ex Ante* Returns and Preferences for Occupations using Subjective Expectations Data."

†Duke University, NBER and IZA.
‡Duke University, NBER and IZA.
§Duke University, NBER and IZA.
¶Oxford College of Emory University.

1 Introduction

Subjective expectations data are increasingly being used in economic research. Early work focused on measuring these expectations in a meaningful way (Manski, 1993, 2004; Hurd and McGarry, 1995, 2002; Hurd, 2009; Dominitz and Manski, 1996, 1997). More recent studies have used elicited beliefs about the likelihood of taking particular actions and the potential future payoffs from those actions to analyze individual decision-making under uncertainty. Much of this literature has focused on educational choices, including college enrollment and choice of a college major (Arcidiacono et al., 2012; Zafar, 2013; Stinebrickner and Stinebrickner, 2014; Wiswall and Zafar, 2015, 2018; Kaufmann, 2014; Attanasio and Kaufmann, 2014, 2017).[1] As stressed in a series of papers (Carneiro et al., 2003; Cunha et al., 2005; Cunha and Heckman, 2007, 2008), in the presence of heterogeneity and uncertainty it is *ex ante* differential gains – as opposed to *ex post* ones – that are relevant for agents' decision-making.

In this paper, we examine the use of elicited beliefs about the likelihood of agents going into alternative occupations and their *ex ante* earnings returns to these alternatives to understand how these occupational choices are made. As with educational decision-making, previous studies (Miller, 1984; Siow, 1984; Keane and Wolpin, 1997) have stressed the central role of such *ex ante* returns in modeling occupational choice.

We address three sets of questions. First, do elicited beliefs about future occupational choices and earnings predict subsequent actual choices and realized earnings? While subjective beliefs are of interest in their own right, any evidence that these are predictive of actual labor market outcomes would add to the value of eliciting such data. Second, do elicited beliefs about future earnings and occupational choices indicate selection on *ex ante* earnings gains? Building on the treatment effects literature (Heckman and Vytlacil, 2007), we develop an *ex ante* evaluation framework that allows us to characterize average *ex ante* treatment effects, as well as *ex ante* treatment effects on the treated and untreated. We establish identification conditions and show how to estimate the *ex ante* treatment effect parameters using elicited beliefs data. We then examine how selection across occupations is related to *ex ante* earnings gains. Third, what do students' subjective beliefs and actual occupational choices reveal about the importance of expected earnings and non-pecuniary factors in their occupational choices? We estimate how much money individuals expect to leave on the table by not necessarily choosing their earnings-maximizing occupational option. We then estimate a variant of the generalized Roy model in which agents make occupational choices based on expected monetary returns as well as occupation-specific non-pecuniary factors. We contribute to the literature by leveraging our elicited beliefs data to understand sorting across occupations, allowing us to remain agnostic about the agents' information sets.

We address these questions using panel data on individuals first interviewed as undergraduates at Duke University. In the initial wave, Phase 1, we elicited their beliefs about their expected earnings and probabilities of entering each of a set of occupations after they graduate from college. We collected information on their occupational choices, actual earnings, and again elicited their beliefs about future occupations and earnings in the Phase 2 follow-up six years later. These latter data allow us to examine

[1] Several studies also have incorporated subjective expectations about objective events in the estimation of structural dynamic models (Delavande, 2008; van der Klaauw and Wolpin, 2008; van der Klaauw, 2012).

their *ex post* choices and earnings realizations, as well as how young men updated their beliefs at an early stage of their labor market careers.

With respect to predictive validity, we find that our Phase 1 elicited beliefs data are highly correlated with actual occupational choices and actual earnings collected in Phase 2. For example, those working in a health occupation reported Phase 1 probabilities for that occupation that were over four times larger than the probabilities reported by those not working in a health occupation. We also show that Phase 1 beliefs about earnings are predictive of what these individuals actually earned seven years later, even after controlling for students' chosen majors and occupations.

We next turn to what these beliefs tell us about the expected earning premia associated with different occupations. In contrast to the literature that uses *ex post* observational data to identify *ex ante* gains (Carneiro et al., 2003; Cunha et al., 2005; Cunha and Heckman, 2007, 2008), we directly elicit agents' *ex ante* gains.[2] We find that the average *ex ante* gains (treating the education occupation as the baseline) range from 30% higher earnings (Science) to as much as 122% higher earnings (Business) ten years after graduation.

We also find that *ex ante* earnings treatment effects for occupations change from Phase 1 to Phase 2, possibly as a result of occupation-specific aggregate shocks that occurred between the two surveys. The *ex ante* treatment effects for business occupations increase from Phase 1 to Phase 2, especially for those who indicated a high probability of choosing a business occupation at Phase 2. But as the *ex ante* treatment effects evolve so do the probabilities of choosing particular occupations. Namely, occupations that see increases in *ex ante* treatment effects over time, such as occupations in Business, also see increases in the probabilities of choosing them. The reverse holds for occupations such as Law where both the *ex ante* treatment effects and the probability of choosing that occupation fall over time.

While we find significant effects of expected monetary returns in our specifications of occupational choice, we also find evidence that the individuals in our sample still expect to give up sizable amounts of money by not always choosing their highest paying occupation. These findings are consistent with previous evidence using *ex post* observational data that pure income-maximizing models do not adequately capture individual choices (see, e.g., Heckman and Sedlacek, 1990 in the context of sectoral choices).

Given that agents are not pure income-maximizers, we estimate a variant of the generalized Roy model of occupational choice that allows for sorting on *ex ante* earnings differences across occupations but also incorporates occupation-specific non-pecuniary factors affecting these choices. We find a positive, statistically and economically significant effect of earnings beliefs on occupational choices. This finding holds true whether we use beliefs about *ex ante* occupational choices elicited at Phase 1 or Phase 2, or students' actual occupational choices at Phase 2, and whether the expected earnings measure is taken from Phase 1 or Phase 2. Our findings also are robust to number of alternative assumptions regarding the relationship between the unobserved factors affecting occupational choice and *ex ante* earnings beliefs.

The rest of the paper is organized as follows. Section 2 discusses the data, while Section 3 shows that

[2]Most of our analysis focuses on sorting across occupations based on expected, as opposed to *ex post*, returns. As such, our paper complements the literature using observational data to show that individuals sort on *ex post* returns. Notable recent examples in the schooling context include Heckman et al. (2018) and Kirkeboen et al. (2016).

subjective beliefs are predictive of actual labor market outcomes. Section 4 discusses the estimation of the *ex ante* treatment effect parameters and the estimation results. Section 5 quantifies the importance of sorting across occupations on expected earnings, and discusses the role that non-pecuniary factors play in occupational choice. Finally, Section 6 concludes.

2 Data

2.1 Phase 1 Data

The data used in this paper come from the Duke College Major and Expectations Survey (DuCMES). The DuCMES first collected data from a sample of male undergraduate students at Duke University between February and April 2009. We refer to this as Phase 1 of the DuCMES. Gender was the only restriction on sample recruitment; male students from any major or year in school were eligible to participate in the study. Sample members were recruited by posting flyers around the Duke campus and were asked to complete a short survey.[3] All 173 students who completed it were paid $20.

The DuCMES Phase 1 survey collected information on students' background characteristics and their current or intended major, where majors were categorized into six broad groups: Natural Sciences, Humanities, Engineering, Social Sciences, Economics, and Public Policy.[4] Students also were asked about the likelihood of choosing future careers, and how much they expected to earn in them. Namely, for each major, we asked students the probability that they would enter a particular career and the earnings they would expect to receive in that career ten years after graduation. We used the following six broad categories to characterize possible careers: Science/Technology, Health, Business, Government/Non-Profit, Education and Law.[5] It is important to note that, for all students in the sample, these probabilities and expected earnings were elicited for all possible occupation-major combinations.[6] Specifically, to elicit career probabilities, students were asked:

> "Suppose you majored in each of the following academic fields [Sciences, Humanities, Engineering, Social Sciences, Economics, Public Policy]. What are the probabilities that you will pursue the following career field [Science, Health, Business, Government/Non-Profit, Education, Law] after majoring in this academic field?"

Throughout the paper, we let p_{ij1} denote the probability elicited from individual i of their choosing occupation j conditional on their chosen major and where the last entry, 1, denotes that the elicitation was in Phase 1 of our study.

To elicit expected earnings associated with different careers and majors, students were asked:

[3] A copy of the questionnaire used in the Phase 1 survey can be found at `public.econ.duke.edu/~vjh3/working_papers/college_major_questionnaire_ph1.pdf`.

[4] The mapping of students' actual college majors into the major groups is reported in Table A.1 of Appendix A.1.

[5] In most of the paper, we simply refer to these six career groups as occupations. We chose this classification based on the main groups of careers in which Duke graduates worked upon graduation, using data from the Duke Senior Exit Survey of 2007.

[6] For more details on the data collected in Phase 1 of the DuCMES, see Arcidiacono et al. (2012).

> "For the following questions regarding future income, please answer them in pre-tax, per-year, US dollar term, ignoring the inflation effect. Suppose you majored in the following academic field. How much do you think you will make working in the following career ten years after graduation?"

We let $\mathcal{Y}_{ij1}$ denote individual i's elicited future income belief at Phase 1 if he worked in occupation j ten years after graduating from Duke.

Descriptive statistics for the Phase 1 data are shown in Tables A.2 – A.6 of Appendix A.2. Table A.3 shows that there are substantial earnings variation across major/occupation pairs in ways that one would expect. For example, higher premiums in Business are found if one majors in economics; higher premiums in Health are found if one majors in the natural sciences. Table A.4 shows that higher earnings in a major/occupation pair are associated with higher probabilities of choosing an occupation conditional on major.[7]

2.2 Phase 2 Data

In order to assess whether beliefs about future labor market outcomes are predictive of the actual choices and outcomes of sample members after they graduated from college, we collected data on the actual occupational choices and earnings of our sample members several years after all of them completed their BA degrees.[8] These data were collected in Phase 2, from two different sources. We describe each source below.

To collect data on occupational choices and additional demographic data on the DuCMES sample, we used information obtained from the social network *LinkedIn* in July 2015, six years after the Phase 1 data was collected. *LinkedIn* contains information on current job titles and companies, graduate degrees, as well as demographic and contact information. Using information on individual's name, major, and graduation year from the Duke Alumni Database we were able to find the occupations of 143 out of the 173 individuals from our original sample on *LinkedIn*.[9] For another 18 individuals, occupations were obtained from an internet search, where we matched on at least two pieces of information from our initial survey and/or the Duke Alumni Database to ensure an accurate match. Finally, occupations were subsequently gathered for 4 more respondents directly from updated information in the Alumni Directory. Thus, our Phase 2 data collection produced current occupations for 165 of the 173 members of our original sample. The occupation data obtained from these sources were then mapped into each of the six occupation classifications used in Phase 1: Science, Health, Business, Government, Education and Law.[10] Let $d_{ij}, j = 1, ..., 6$, denote indicator variables for whether individual i's *actual occupation* is j at Phase 2.

[7] Arcidiacono et al. (2012) used the Phase 1 DuCMES data on subjective probabilities and expected earnings for all alternative occupation-major pairs to model students' college major choice. In this paper, we focus on occupational choice and, for most of the analyses presented in this paper, we *only* use data on occupation probabilities and earnings expectations elicited for a student's *chosen*, or *declared* major, recorded at the time of the Phase 1 survey.

[8] This was made possible through the collection of names and contact information during Phase 1.

[9] The Duke Alumni Database is maintained by the Duke Alumni Association.

[10] For example, engineers and software developers were mapped into Science careers; doctors, residents and medical stu-

We also collected additional data on *ex post* labor market outcomes and updated our sample members' expectations about careers on the DuCMES sample in a follow-up survey administered between February and April of 2016,[11] using information from *LinkedIn* and the Alumni database to contact the Phase 1 sample members and solicit their participation in the Phase 2 survey.[12] A total of 117 individuals – about 68% of the initial sample – participated in the follow-up survey and 112 individuals completed it. Table A.6 of Appendix A.2 compares the characteristics of the individuals who completed the Phase 2 follow-up survey with those of the baseline Phase 1 sample. Individuals who responded to the Phase 2 survey have very similar characteristics to the Phase 1 sample, including having similar Phase 1 occupation-specific earnings beliefs, $\mathcal{Y}_{ij1}$, and subjective probabilities of choosing each type of occupation, p_{ij1}. Overall, the comparison in Table A.6 suggests that the non-response for the Phase 2 follow-up survey appears to be largely ignorable.

The availability of *LinkedIn* data also makes it possible to compare respondents and non-respondents in terms of realized labor market outcomes. In particular, conditional on being on *LinkedIn*, the average number of jobs held since college graduation, which we collect from *LinkedIn*, is equal to 2.0 for those who responded to the follow-up survey, against 1.9 for those who did not respond. More generally, the distribution of number of jobs held since graduation is very similar for those who respond versus those who do not respond to the follow-up survey, limiting the importance of survey non-response.

The Phase 2 follow-up survey collected information on sample members' past and current occupations, as well as their actual current earnings, which we denote by Y_{i2}. Respondents also were asked to update their expectations about what they expect their occupations and earnings to be ten years after college graduation.[13] Let p_{ij2} and $\mathcal{Y}_{ij2}$ denote these occupational choice probabilities and expected earnings, respectively, that were elicited in the Phase 2 follow-up survey for their chosen major.

Finally, we used data collected on respondents' current occupation in our Phase 2 follow-up survey to supplement and adjust the information on chosen occupations obtained from *LinkedIn* and the Duke Alumni database. In particular, 19 individuals declared an occupation in the Phase 2 survey that did not match the occupation from *LinkedIn* or the Alumni database.[14] For those cases, we used the occupation

dents into Health; teachers, instructors, and school administrators into Education; Law clerks and Lawyers into Law; and lieutenants and policy analysts at Government organizations into Government. The Business classification contained the largest variety of reported occupations including associate, account executive, analyst, manager, and CEO. In each case, both the current job title as well as the employer were considered in constructing the mapping from reported occupation to the six broad occupational classifications.

[11] A copy of the questionnaire used in the Phase 2 follow-up survey can be found at `public.econ.duke.edu/~vjh3/working_papers/college_major_questionnaire_ph2.pdf`.

[12] All individuals who completed the follow-up survey received a coupon for a Duke Basketball Championship T-shirt that could be redeemed through the Duke University Bookstore's website.

[13] 8.3% of the expected earnings elicited in the Phase 2 survey are missing for the 112 individuals who completed this follow-up survey. For these cases, occupation-specific expected earnings are imputed as the predicted earnings computed from a linear regression of log expected earnings on chosen major and occupation indicators, interaction between major and occupation, individual-specific average log expected earnings in Phase 2 across all occupations, occupation-specific log expected earnings in Phase 1, and an indicator for whether the subjective probability of working in this occupation is equal to zero.

[14] Not all of these discrepancies are misclassifications, since occupations obtained from *LinkedIn* or the Alumni database and Phase 2 survey were not measured at the exact same time (July 2015 vs. Feb - Apr. 2016), and some individuals switch across occupations over this period of time.

given in the Phase 2 survey. We also obtained the occupations for two additional individuals from this survey that were not found in *LinkedIn* or the Alumni database. Overall, we have data on the chosen occupations at Phase 2 of 167 of the 173 (96.5%) original sample members. We use this augmented data on students' chosen occupations at Phase 2 in all of the tabulations and analysis of chosen occupations presented below.

3 Predictive Validity

3.1 Subjective Choice Probabilities versus Chosen Occupations

In this section, we assess the predictive validity of our elicited occupational choice probabilities at Phases 1 and 2 conditional on chosen majors, and the occupations individuals actually chose, measured 4-7 years after they completed their undergraduate degrees.

Columns (1) and (2) of Table 1 display the average probabilities for occupations elicited at Phase 1 $[p_{ij1}]$ and the frequency of actual occupations observed at Phase 2 $[d_{ij}]$. A much greater share of our sample has ended up in Business than what they predicted at the time they were undergraduates, while smaller shares are seen in occupations such as Government and Law. These differences do not rule out rational expectations as they may be the result of intervening aggregate shocks to the labor market.[15] This possibility is especially relevant for our analysis as the Great Recession began in December 2007 and our expectations were elicited in early 2009. For example, there is evidence that entry into the legal profession was affected by a post-Great Recession negative shock that may have not been fully anticipated.[16]

A comparison of elicited beliefs about the likelihood of choosing particular occupations at Phase 1 for those who chose it $[d_{ij} = 1]$ in column (4) with those that did not choose $[d_{ij} = 0]$ in column (5) indicate that the elicited probabilities do appear to have informational content. For example, among those who actually chose Science, the average subjective probability of choosing Science was about 35%, while it is only 13.7% among those who did not. More generally, the average probabilities in column (4) are more than double those in column (5) for all occupations but Education, indicating a tight association between actual occupational choice and elicited probabilities. At the same time, the average probabilities in column (4) are all smaller than 0.5, indicating that, at the time of the initial survey, students had quite a bit of uncertainty about their future choice of occupation.

Column (3) of Table 1 shows beliefs about future occupations that were elicited in Phase 2. Many individuals in Phase 2 are already in their preferred occupations. This is supported by the fact that Phase 2 expectations about the occupations they will be in 10 years after graduation are very similar to the occupations that individuals were already in at Phase 2. On average, individuals reported a higher probability of working in Business at Phase 2 and lower probabilities of working in Law or Government,

[15] See D'Haultfoeuille et al. (2018) for a detailed discussion of the importance of accounting for unanticipated aggregate shocks when evaluating deviations from rational expectations.

[16] As noted in Barton (2015), while the number of LSAT takers was increasing prior to the Great Recession, this number peaked in 2009-10 and has fallen by 45% between then and 2014-15.

Table 1: Comparison of Elicited Occupational Probabilities at Phases 1 & 2 with Frequency of Actual Occupations at Phase 2

	Phase 1 Beliefs: p_{ij1} (1)	Phase 2 Chosen: d_{ij} (2)	Phase 2 Beliefs: p_{ij2} (3)	p_{ij1}, given: $d_{ij}=1$ (4)	p_{ij1}, given: $d_{ij}=0$ (5)	p_{ij2}, given: $d_{ij}=1$ (6)	p_{ij2}, given: $d_{ij}=0$ (7)
Science	0.177	0.156	0.170	0.350	0.137	0.662	0.082
Health	0.165	0.210	0.226	0.424	0.098	0.893	0.014
Business	0.261	0.437	0.414	0.374	0.186	0.791	0.120
Government	0.143	0.054	0.062	0.301	0.134	0.536	0.039
Law	0.169	0.090	0.078	0.391	0.148	0.761	0.018
Education	0.086	0.054	0.051	0.122	0.087	0.690	0.021

DATA: Columns (1), (2), (4) and (5) are based on 167 individuals for whom we obtained their current occupation from *LinkedIn* and the Alumni database at Phase 2, augmented with occupations reported in the Phase 2 follow-up survey for some individuals. Columns (3), (6) and (7) are based on 112 individuals who completed the Phase 1 and 2 follow-up surveys. The differences between Columns (1) and (2) are significant at the 1% level for Business, Government, and Law and are significant at the 10% level for the Health and Education.

patterns that are consistent with the observed occupational choices of the sample at Phase 2. Finally, the last two columns of Table 1 show the expected probability of working in each occupation elicited in the Phase 2 follow-up survey, conditional on currently working and not working in that occupation. In all cases, the average perceived probability of working in their current occupations in three to six years is over fifty percent which is significantly higher than the correspondingly probabilities that were elicited at Phase 1 [Column (4)]. This suggests that by Phase 2, when our sample members are 4-7 years out from graduation, much of the uncertainty regarding occupational choices has been resolved. The discrepancy between the conditional means in columns (6) and (7) is particularly large for occupations such as Health (89.3% conditional on working in Health versus 1.4% conditional on not working in that occupation) or Law (76.1% versus 1.8%). These findings are consistent with a very high cost of switching into these two occupations. Nevertheless, the elicited probabilities are all significantly less than one, suggesting that, while uncertainty has been reduced, some of the individuals in our sample still perceive a significant chance of moving to another occupation in the future.

3.2 Decision to Work in Business

While the previous results show that the subjective probabilities in Phase 1 have informational content, a natural question is whether they have informational content beyond the majors of the former students in our sample. While this comparison cannot be made for all possible major-occupation pairs, given that several pairs were not chosen by our sample, it can be made for the decision to enter Business.

Table 2 shows estimates of a linear probability model of choosing an occupation in Business. Column (1) controls for the elicited probability of choosing Business (indexed by $j = 3$) at Phase 1 conditional on the student's actual major, i.e., p_{i31}. Conditioning only on this one variable results in an R^2 close to 0.19 and the coefficient on it (0.936) is not statistically different from one. Column (2) estimates the

Table 2: Linear probability model of whether Phase 2 occupational choice is Business ($d_{i3} = 1$)

	Full Sample			Excluding Seniors			Excluding Econ Majors		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
p_{i31}	0.936		0.913	0.797		0.790	0.931		0.918
	(0.152)		(0.205)	(0.201)		(0.260)	(0.233)		(0.246)
Chosen Major:									
Engineering		0.017	-0.106		0.126	-0.021		0.017	-0.107
		(0.121)	(0.118)		(0.160)	(0.162)		(0.122)	(0.121)
Humanities		0.305	0.239		0.318	0.266		0.305	0.239
		(0.158)	(0.150)		(0.183)	(0.177)		(0.160)	(0.153)
Social Science		0.211	0.069		0.299	0.161		0.211	0.068
		(0.126)	(0.124)		(0.155)	(0.156)		(0.128)	(0.128)
Economics		0.485	0.055		0.423	0.075			
		(0.121)	(0.150)		(0.153)	(0.187)			
Public Policy		0.252	0.114		0.273	0.131		0.252	0.114
		(0.120)	(0.118)		(0.147)	(0.149)		(0.121)	(0.121)
R^2	0.187	0.125	0.223	0.126	0.082	0.157	0.109	0.063	0.155

DATA: Full sample includes 167 individuals for whom we obtained their current occupation from information obtained from respondents' *LinkedIn* accounts and the Duke Alumni database, as well as information collected in the Phase 2 follow-up survey. The Excluding Seniors (Econ Majors) sample consists of the 113 (135) respondents who were not seniors (Econ majors).

NOTES: Subjective probability of choosing Business is conditional on their chosen major. Standard errors in parentheses. All specifications include a constant term.

differences in choosing a Business occupation by one's chosen major. Compared to having graduated with a major in the natural sciences, all other majors have a higher probability of being in a Business occupation as of Phase 2, with economics majors having the highest relative probability (48.5 percentage points higher than for Science majors). However, accounting for one's major results in a lower R^2 (0.125) than conditioning on the elicited probability at Phase 1 of choosing Business. Column (3) includes both the Phase 1 elicited probability of Business and one's chosen major. While the coefficient on the elicited probability declines relative to column (1), the difference is not significant and the coefficient remains large in magnitude. Interestingly, the coefficient on being an economics major falls substantially (from 0.485 to 0.055) and is no longer statistically significant. These results provide additional evidence that the subjective probabilities are quite informative about future occupation decisions beyond the choice of major.

The findings in Table 2 may be driven by the inclusion of those who were seniors at Phase 1, as some of them may have already lined jobs at the time of our initial survey. In columns (4)-(6) of Table 2, we perform the same analysis as in columns (1)-(3) but remove seniors from the sample.[17] The same patterns emerge: the elicited probability of choosing Business has more explanatory power than major dummies and its inclusion renders the coefficient on being an economics major insignificant. That the R^2 decreases

[17]Throughout the rest of the paper, we will show results both with and without seniors. However, since removing senior changes none of the qualitative results, we will not comment on them.

when we exclude seniors from the sample (0.187 for Specification (1) against 0.126 for Specification (4)) also indicates that the beliefs about the probabilities of choosing Business are more predictive of the actual choice of occupation for senior students. Estimation results obtained on the subsample of students who were not enrolled in economics similarly yield a lower R^2 than in the full sample [0.187 in column (1) vs. 0.109 in column (7)]. These results indicate that students who are closer to graduation as well as economics majors, who may have acquired more information about Business, are better at forecasting whether they will work in Business.[18]

3.3 Expected versus Actual Earnings

Finally, we conclude this section by examining the relationship between the actual earnings of the members of our study collected in the Phase 2 follow-up survey, Y_{i2}, and expected earnings elicited in the initial Phase 1 survey for the (eventually) chosen occupation, $\mathcal{Y}_{i1}$. To do so, we use data on 81 individuals who reported having positive current annual earnings in the Phase 2 follow-up survey, $Y_{i2} > 0$.[19]

In Table 3, we report estimation results from a linear regression of log actual earnings on log expected earnings in chosen occupation. In the spirit of the analysis conducted in the previous section, column (1) displays estimates when we restrict the sample to the individuals who work in Business and control for chosen majors (Table 2). The estimated elasticity (0.64) is positive, sizable, and statistically significant at the 5% level.[20] As noted above, cell sizes prevent us from estimating separate regressions for each of the other occupations. Nonetheless, in column (2) we estimate the relationship between expected and actual earnings for all occupations, controlling for both chosen major and chosen occupation. While the estimated elasticity (0.423) is smaller than the one for those who chose business careers, it remains positive and significant at the 1% level.[21]

Taken together, the results presented in this section provide evidence that beliefs about future occupations, as well as about earnings, are predictive of the future labor market outcomes for our sample.

[18] In Appendix A.3 we complement this analysis by using responses to questions about what the average Duke student would make in different major and occupations, separately for upper- versus lower-classmen (Table A.7) and chosen versus non-chosen majors (Table A.8). For the majority of occupations and majors, beliefs are less dispersed among upper-classmen than lower-classmen. This is consistent with students forming more accurate beliefs about labor market outcomes as they acquire more education. A similar pattern also holds, for the majority of occupations and majors, when comparing between chosen and counterfactual majors the beliefs about the earnings of the average Duke student for each occupation and major. This offers suggestive evidence that students who enrolled in particular major tend to make more precise forecasts for the earnings associated with that particular major.

[19] Some 30 individuals out of the 112 individuals who completed the Phase 2 survey indicated that they did not have a current job and, thus, were not asked about their current annual earnings. The vast majority (more than 80%) of those individuals were medical interns or residents, who did not consider these positions as jobs, or were enrolled in an MBA program at the time of the survey.

[20] Note these estimated elasticities are under-estimates due to attenuation bias if log expected earnings are subject to classical measurement error.

[21] It is interesting to compare these results with Wiswall and Zafar (2018), who estimate, in a different context, the association between log realized earnings and log expected earnings. Among males, they find a positive but insignificant relationship between these two quantities, with an estimated elasticity of 0.167 without controlling for majors or occupations. (They find that beliefs are much more predictive of the actual earnings of females.)

Table 3: Relationship between actual and expected earnings in chosen occupation

	Log Actual Earnings ($\ln Y_{i2}$):	
	Business Only	All Occupations
	(1)	(2)
Log Expected Earnings ($\ln \mathcal{Y}_{i1}$)	0.640	0.423
	(0.257)	(0.149)
Other variables included:		
Chosen Major	Y	Y
Chosen Occupation	–	Y
R^2	0.396	0.521

DATA: Column (1) (Column (2)) is based on 37 (81) individuals who reported they had a current job and provided their current annual earnings for that job.
NOTES: Standard errors in parentheses. All specifications include a constant term. $\mathcal{Y}_{i1}$ refers to expected earnings at Phase 1 in the chosen occupation at Phase 2.

4 *Ex Ante* Treatment Effects

Given that our expectations data have empirical content, we develop a framework for analyzing *ex-ante* treatment effects using subjective expectations data. We focus in particular on the average *ex-ante* treatment effect (ATE) and the *ex-ante* treatment effect on the treated (TT) where occupations are the treatment.[22] We define the average *ex-ante* treatment effect of an occupation as a weighted average of the *ex-ante* treatment effects of each occupation, which upweights (downweights) the occupations agents are more (less) likely to choose. This parameter, which combines two sources of uncertainty - the future earnings of the occupation, and the uncertainty associated with the probability of working in that occupation - provides an aggregate look at how individuals expect to sort into occupations based on expected earnings. In Section 4.1 we discuss the identification and estimation of the *ex ante* treatment effect parameters using subjective expectations data on earnings and occupational choice. In Section 4.2, we use the estimated *ex ante* treatment effects to quantify the extent of selection across occupational choices on *ex ante* earnings gains.

4.1 Definitions, Identification and Estimators

4.1.1 *Ex Ante* Treatment Effect Parameters

To proceed, we note that earnings beliefs 10 years after college, conditional on occupation j and i's information set at time t, $\mathcal{I}_{it}$, are defined to be

$$\mathcal{Y}_{ijt} := \mathcal{E}\left[Y_{ij}|\mathcal{I}_{it}\right] \tag{4.1}$$

[22]Recent papers by Wiswall and Zafar (2018) and Giustinelli and Shapiro (2019) build on the methodology developed in this section to estimate expected treatment effects of college major choices on future earnings, labor supply and spousal earnings, and of health status on retirement, respectively.

i.e., denoting by $\mathcal{E}[\cdot|\mathcal{I}_{it}]$ the subjective expectation operator, agent i's future expected earnings if he was to enter occupation j, where these beliefs are taken at point t.[23] In our context, beliefs are elicited at two points: Phase 1 ($t = 1$), when the agent is an undergraduate, and Phase 2 ($t = 2$), 4-7 years after graduation. In the following, we denote the *elicited* earnings beliefs by $\widetilde{\mathcal{Y}}_{ijt}$ to reflect the fact that elicited beliefs may differ from the true beliefs $\mathcal{Y}_{ijt}$.[24] Agents were asked at both points to provide what they expected their earnings to be ten years after graduation for each possible occupation.

For any given individual i, the *ex ante* earnings gain from being in occupation $j \in \{2, ..., 6\}$ relative to a baseline occupation, $j = 1$, is given by $\Delta\mathcal{Y}_{ijt} := \mathcal{Y}_{ijt} - \mathcal{Y}_{i1t}$. Throughout, we use Education as the baseline occupation.[25] We define the average *ex ante* treatment effect of occupation j to be:

$$ATE_{jt} := E(\Delta\mathcal{Y}_{ijt}). \tag{4.2}$$

The average *ex ante* treatment effect in our context represents what the mean expected earnings gain would be if agents were randomly assigned to occupation j relative to occupation 1, where expectations are taken at point t.[26]

It follows that the average *ex ante* treatment effect parameters are identified directly from our elicited data of expected earnings for each occupation, provided there is no measurement error on the elicited earnings beliefs. In fact, even in the presence of (possibly non-classical) measurement errors, ATE_{jt} is still identified from these data so long as such errors have the same mean across occupations (see Appendix A.4). Under this assumption, ATE_{jt} is consistently estimated as the sample average of the elicited *ex ante* earnings gain from being in occupation j relative to occupation 1:

$$\widehat{ATE}_{jt} = N^{-1} \sum_i \Delta\widetilde{\mathcal{Y}}_{ijt}, \tag{4.3}$$

where $\Delta\widetilde{\mathcal{Y}}_{ijt}$ is the elicited *ex ante* earnings gain from occupation j relative to occupation 1, and N is the sample size.

We are interested in evaluating whether students sort across occupations according to differences in what they expect to earn in different occupations. In the following, we take a first look at this question by examining whether those who expect to work in a given occupation have higher expected earnings in that occupation than those who do not. In the *ex-post* case, the treatment on the treated would be defined for those in a particular occupation. What is different in the *ex-ante* context is the uncertainty

[23] Under rational expectations, $\mathcal{Y}_{ijt} = E[Y_{ij}|\mathcal{I}_{it}]$, where $E[.|\mathcal{I}_{it}]$ denotes the conditional expectation operator generated by the true data generating process. Except when we mention otherwise, we do not need to maintain this assumption for the results in this section.

[24] In order to alleviate the notational burden and since there is no ambiguity in the rest of the paper, we use distinct notations for elicited and true subjective beliefs throughout this section only.

[25] We use Education as the base occupational category as there is less variance overall in expected earnings for Education than for the other occupations, which makes it a more natural reference alternative.

[26] Note that the ATE_{jt} defined in (4.2) is not the *internal rate of return* that equilibrates the differences between working in occupations j and 1, since it does not incorporate differences in direct and opportunity costs of working in these occupations. (The same will be true of the other *ex ante* treatment effects defined below.)

associated with the future treatment. Specifically, we define the *ex-ante* TT as the weighted average of the individual-specific treatment effects where the weights are the expected probabilities of working in that occupation given the individual's information set. Formally, we define the *ex ante* TT for occupation j as:

$$TT_{jt} := E\left(\omega_{ijt}\Delta\mathcal{Y}_{ijt}\right), \tag{4.4}$$

where $\omega_{ijt} := p_{ijt}/E(p_{ijt})$, p_{ijt} is agent i's subjective probability (at point t) that he would work in occupation j ten years after graduation, and $E(p_{ijt})$ is its mean taken over all agents. The TT_{jt} will be larger than ATE_{jt} in (4.2) if individuals expect to sort into occupations with higher returns.

Clearly, a sufficient condition for TT_{jt} to be identified from the data on earnings beliefs and subjective probabilities of choosing occupation j is that both sets of elicited beliefs are measured without error. In Appendix A.4 we show that this remains true if earnings beliefs are measured with error, under the assumption that measurement errors on the earnings beliefs have the same mean across occupations, and are uncorrelated with the subjective probabilities. TT_{jt} further remains identified if both earnings beliefs and subjective probabilities are measured with error, as long as measurement errors on the subjective probabilities are mean-zero and uncorrelated with the earnings beliefs, measurement errors on earnings beliefs are uncorrelated with the subjective probabilities, and both types of measurement errors are mutually uncorrelated. Under these assumptions, a consistent estimator of TT_{jt} is given by:

$$\widehat{TT}_{jt} = N^{-1}\sum_i \widehat{\widetilde{\omega}}_{ijt}\Delta\widetilde{\mathcal{Y}}_{ijt}, \tag{4.5}$$

where $\widehat{\widetilde{\omega}}_{ijt} = \widetilde{p}_{ijt}/(N^{-1}\sum_i \widetilde{p}_{ijt})$, denoting by $\widetilde{p}_{ijt}$ the *elicited* subjective probability of choosing occupation j. Finally, the *ex ante* treatment on the untreated, TUT_{jt}, is defined similarly by replacing p_{ijt} with $1 - p_{ijt}$. Identification from subjective expectations data requires similar assumptions on the form of measurement error.[27]

4.1.2 Connection with Alternative Treatment Effect Parameters

One can alternatively define a treatment effect that is *ex-ante* with respect to earnings but *ex-post* with respect to the choice of occupation. This treatment on the treated takes the beliefs about earnings at time t but for those who actually are treated. As such, it does not incorporate the uncertainty associated with the choice itself. We denote this alternative treatment on the treated as TT^*_{jt} and is given by:

$$TT^*_{jt} := E\left(\Delta\mathcal{Y}_{ijt}|d_{ij}=1\right), \tag{4.6}$$

where $\{d_{ij}=1\}$ denotes those who *ex-post* chose to work in occupation j. While this parameter, when contrasted with the ATE, is informative about selection on expected gains, an important limitation is

[27] In Appendix A.5 we show that such beliefs data can also be used to identify the full distribution of the *ex ante* treatment effects. We characterize a consistent estimator of these distributions in Appendix A.5, and plot differences in estimates of the distributions of TT and TUT for each of the six occupations.

that it generally requires data on the actual occupational choices that may lie beyond the sample period. However, the two treatment on the treated parameters, TT_{jt} and TT^*_{jt}, coincide if certain conditions are met. Namely, taking d_{ij} as given, $TT_{jt} = TT^*_{jt}$ when:

$$E\left[\left(p_{ijt} - d_{ij}\right)\Delta\mathcal{Y}_{ijt}\right] = 0 \tag{4.7}$$

This condition is met when (A1) individuals are, on average, correct about the probability of working in occupation j, i.e. $E(p_{ijt} - d_{ij}) = 0$, and (A2) any forecast errors in their beliefs on the probabilities of working in occupation j, $p_{ijt} - d_{ij}$, are uncorrelated with their beliefs on earnings in occupation j relative to occupation 1, $\Delta\mathcal{Y}_{ijt}$. It follows that if (4.7) holds, $\widehat{TT}_{jt}$ in (4.5) is a consistent estimator of TT^*_{jt}.

Finally, each of these treatment on the treated measures tie to the conventional (*ex-post*) treatment effect on the treated parameter, $TT^P_j := E(Y_{ij} - Y_{i1})$ for each occupation j, under stronger assumptions. We provide a set of conditions under which the different treatment effect parameters coincide in Appendix A.4. For example, the *ex-ante* treatment on the treated is equal to the *ex-post* treatment on the treated when (A1) and (A2) are met and, in addition, (A3) for those who chose occupation j, the *ex-post* earnings gain from being in occupation j relative to occupation 1 are equal, in expectation, to their *ex-ante* counterpart, namely $E\left[d_{ij}\left(\Delta\mathcal{Y}_{ij} - \Delta Y_{ij}\right)\right] = 0$. A sufficient condition for Assumption (A3) is that students have rational expectations about their future earnings, and that there are no unanticipated aggregate earnings shocks.[28]

4.2 Treatment Effect Estimates and Evidence of Selection on *Ex Ante* Earnings

In Table 4 we present estimates of the various *ex ante* treatment effects defined in the previous section using elicited beliefs on expected earnings in particular occupations 10 years after graduation. Panel A shows estimates of the ex-ante ATE, TT, and TUT treatment effects using data from Phase 1. The ATE estimates in Column (1) of Panel A indicate clear differences in expected earnings gains across occupations. These ATE estimates for all occupations are significantly different from zero and indicate economically large *ex ante* gains to occupations relative to Education.[29] The largest mean expected earnings gains are found for Business occupations [$89,533, and gross earnings return[30] of 1.179] followed by Law [$88,750 and gross return of 1.169], while the lowest expected gains and gross returns are for Government [$25,875 or a 0.341 return] and Science [$22,320 or a 0.294 return].[31]

Corresponding ATE estimates from Phase 2 are displayed in Column (1) of Panel B. The expected earnings gains from being randomly assigned to the various occupations differ from those obtained in Phase 1, with larger expected gains for Business [$139,815] and Science [$51,968] but lower gains in

[28] Note that implicit here is the assumption that occupational choice is uncorrelated with forecast errors. This assumption is innocuous as long as choices are made before earnings are realized.

[29] The corresponding ATE estimates in Panel B that uses Phase 2 data and almost all of the estimates for the other treatment effects in Table 4 are significant as well.

[30] Gross returns, or premia, are defined as $E[(\mathcal{Y}_{ijt} - \mathcal{Y}_{i1t})/\mathcal{Y}_{i1t}]$.

[31] Table A.9 of Appendix A.6 presents estimates of the average *ex ante* treatment effects separately for lower-classmen and upper-classmen. While the estimates for all occupations are larger for upper-classmen compared to lower-classmen, none of them are significantly different at standard statistical levels.

Table 4: *Ex Ante* Treatment Effects by Occupation (Earnings in 2009 dollars)

Occupation	ATE	TT	TUT	TT^*	TUT^*
	(1)	(2)	(3)	(4)	(5)
	A. *Ph. 1 Elicited Earnings* ($\mathcal{Y}_{ij1}$) & *Occup. Choice Prob.* (p_{ij1}):			C. *Ph. 1 Elicited Earnings* ($\mathcal{Y}_{ij1}$) & *Occup. Choice* (d_{ij2})	
Science	22,320	29,820	20,674	34,808	19,290
	(3,121)	(4,786)	(3,246)	(7,612)	(3,269)
Health	68,065	117,700	57,808	122,570	52,358
	(8,575)	(18,802)	(6,879)	(14,222)	(7,323)
Business	89,533	104,224	84,201	90,726	84,021
	(8,480)	(14,664)	(8,052)	(11,445)	(10,086)
Government	25,875	26,733	25,753	37,111	23,758
	(3,970)	(7,162)	(3,918)	(16,673)	(3,979)
Law	88,750	110,423	84,343	88,667	89,474
	(11,280)	(20,033)	(10,595)	(30,611)	(9,616)
	B. *Ph. 2 Elicited Earnings* ($\mathcal{Y}_{ij2}$) & *Occup. Choice Prob.* (p_{ij2}):				
Science	51,968	61,879	49,942		
	(5,786)	(14,337)	(6,070)		
Health	54,224	119,588	35,131		
	(8,897)	(26,631)	(5,352)		
Business	139,815	220,938	82,518		
	(17,843)	(28,211)	(10,380)		
Government	10,046	18,008	9,524		
	(1,927)	(3,932)	(1,979)		
Law	65,995	54,175	66,990		
	(7,763)	(16,723)	(8,168)		

DATA: Sample who completed Phase 1 survey ($N = 173$) and the Phase 2 follow-up survey ($N = 112$), respectively.

NOTES:

[1] Standard errors are reported in parentheses.

[2] The base occupational category in the above tables is Education.

[3] In the Phase 1 sample (Panel A), TT is significantly different from TUT for Science (p-value = 0.051), and Health (p-value = 3.10^{-4}). For respondents that completed the Phase 2 survey (Panel B), TT is significantly different from TUT for Health (p-value = 0.001), Business (p-value = 0.000) and Government (p-value = 0.036). For the estimates based on actual choice of occupation (Panel C), TT^* is significantly different from TUT^* for Science (p-value = 0.061) and Health (p-value = 10^{-5}).

Law.[32] Nonetheless, the rank ordering of earnings gains across occupations is almost identical across time, with Business, Law and Health showing the largest ATE's, respectively, across the two waves of data.

We next consider our estimates of the *ex ante* TT in Column (2) of Table 4.[33] In all but one case (Law in Phase 2), the estimated *ex ante* TT parameters are larger than the estimated ATE's. These differences in *ex-ante* ATE and TT are consistent with positive sorting across occupations on *ex ante* earnings gains. Moreover, sorting on expected earnings is most pronounced for Heath, and this is true whether expectations were elicited before graduating from college (Phase 1) or 4-7 years after (Phase 2).

Comparing the estimates in Panels A and B, there are differences in expected gains depending on when the information was collected. Recall from columns (1) and (3) of Table 1 that our respondents in Phase 2 reported substantially lower probabilities for entering Law and Government and substantially higher probabilities of entering Business. As noted in Section 3.1, these differences in choice probabilities across the two survey waves may reflect, in part, the differential impact that the Great Recession had on entry into occupations, such as Law. The Great Recession also may have differentially affected the expected earnings and, thus, the sorting on earnings across occupations.

Estimates of *ex ante* treatment effects based on *ex post* occupations (TT^*) are presented in Panel C of Table 4, where we use data on earnings beliefs obtained in Phase 1 but chosen occupations as of Phase 2. The estimates in Columns (4) - (5) of this table for TT^* correspond to those in Columns (2) - (3) for the treatment effects on the treated. The estimates for TT^* and TUT^* in Panel C are fairly close to the corresponding estimates in Panel B, based on *ex ante* occupations. Furthermore, the former treatment effect estimates also provide evidence of sorting on *ex ante* earnings gains, with sorting again being most pronounced for Health occupations. In short, the findings in Panel C reinforce the conclusion that our treatment effects based on the *ex ante* earnings and occupational choice beliefs are informative about *ex ante* gains from earnings sorting.

Finally, it is interesting to compare our results with the related literatures. First, our results on earnings gains across occupations are consistent with those found in previous studies using *ex post* data on earnings and chosen occupations. For example, Lemieux (2014) found substantial returns to occupations using *ex post* earnings data.[34] Second, our results also complement prior evidence of sorting on *ex post* gains across occupations (see, e.g., Gibbons et al., 2005), by showing that individuals who report a higher probability of working in a particular occupation expect, on average, to earn significantly more in that occupation. Third, it is interesting to note that our estimates of treatment effects associated with different

[32]Since individuals in Phase 2 have already entered the labor market, some of the results could reflect ex-post rationalizations of their occupational choices. Note, however, that the fact that the gap between the TT and the TUT is smaller in Phase 2 than in Phase 1 for the case of law goes against this concern.

[33]Another way of assessing the role of selection is to construct the *ex ante* analogues of occupation-specific earnings, both unadjusted and adjusted for the selectivity of choosing a particular occupation. Unadjusted *ex post* earnings are just the observed earnings of individuals working in a particular occupation, as would be observed in national data sets such as the American Community Survey (ACS). Using our expectations data, we can produce *ex ante* analogues of both measures. Results under this alternative approach, which are shown in Appendix A.7, point to similar patterns of selection across occupations.

[34]See also Gibbons et al. (2005) who provide evidence of large and significant industry wage premia even after controlling for selection on (unobservable) comparative advantage.

occupations tend to be substantially larger than the already large returns to college majors that have been estimated in the literature (see Table 8 in Altonji et al., 2016, for a summary). For instance, using data from Norway, Kirkeboen et al. (2016) find that completing a four-year college degree in Business, relative to one in Education is associated with a 22% earnings premium. Our estimated premium for working in a Business occupation is more than 5 times larger.[35]

5 Occupational Choice and Sorting on Expected Earnings

The previous findings indicate a positive association between expected earnings and occupational choice. In this section we establish that non-pecuniary factors also play an important role in occupational choice, and then account for such factors to quantify the importance of sorting across occupations on expected gains. Specifically, Section 5.1 provides non-parametric evidence that subjective beliefs about earnings and occupational choice are inconsistent with a pure (expected) income-maximizing Roy model. This finding motivates our analysis in Section 5.2, where we estimate a variant of the generalized Roy model to calculate elasticities of occupational choice with respect to expected earnings.

5.1 Non-Parametric Evidence on the Role of Non-Pecuniary Factors in Occupational Choice

In this section we use our elicited data on future earnings beliefs and occupational choice probabilities to characterize the share of respondents whose expected occupation choices were not earnings maximizing. We also develop a measure of their *ex ante* willingness-to-pay for such choices, expressed in terms of foregone earnings. These *ex ante* measures of willingness-to-pay are directly identified from the data. Importantly, this measure does not require any distributional assumptions, nor does it require one to take a stand on what non-pecuniary factors affect the choice of occupation. Further, we do not need to impose the restriction that beliefs are rational, and we remain agnostic on the information set of the individuals.[36]

We define the occupation for which individual i gave their highest expected earnings given the information set at time t as j_{it}^{max}, and the expected earnings associated with it to be:

$$\mathcal{Y}_{it}^{max} := \max\{\mathcal{Y}_{i1t}, \mathcal{Y}_{i2t}, ..., \mathcal{Y}_{i6t}\} \tag{5.1}$$

Denote $\overline{\mathcal{Y}}_{it}$ as i's overall expected earnings, that is the weighted average of the occupation-specific elicited expected earnings where the elicited probabilities that the individual would work in each occupation are

[35] Differences also are stark when compared with the estimated returns to Business major (relative to Education) obtained by Arcidiacono (2004) on the NLS72 data, which range between 13% and 22% for males and females, respectively.

[36] Related work by D'Haultfœuille and Maurel (2013) investigates the relative importance of *ex ante* monetary returns versus non-pecuniary factors in the context of an extended Roy model applied to the decision to attend college. Their approach does not require measures of subjective expectations about future returns, but relies on stronger assumptions concerning the non-pecuniary factors. See also Eisenhauer et al. (2015), who use exclusion restrictions between monetary returns and non-pecuniary factors to separately identify these two components in the absence of subjective expectations.

the weights:

$$\overline{\mathcal{Y}}_{it} := \sum_{j=1}^{6} \mathcal{Y}_{ijt} p_{ijt} \tag{5.2}$$

Using $\mathcal{Y}_{it}^{max}$ and $\overline{\mathcal{Y}}_{it}$ one can characterize an individual's *ex ante* willingness-to-pay for not choosing their earnings maximizing occupation. Namely, denote $\mathcal{G}_{it}$ as:

$$\mathcal{G}_{it} := \mathcal{Y}_{it}^{max} - \overline{\mathcal{Y}}_{it} \tag{5.3}$$

If individual i reports that $p_{ijt} = 1$ for $j = j_{it}^{max}$, he is an earnings maximizer in an expected value sense, so that $\mathcal{G}_{it} = 0$. In contrast, if individual i is not an earnings maximizer and chooses occupations based, in part, on non-pecuniary factors, then $\mathcal{G}_{it} > 0$.

Our measures of expected earnings do not include any differences in the educational costs of pursuing particular occupations (for example, law or medicine). For this reason, we focus on Phase 2 beliefs which are elicited after most, if not all, of the relevant educational decisions have been made, and, as such, result in measures of *ex ante* willingness-to-pay which are not contaminated by tuition payments. In Table 5 we display estimates of the mean, median, first and third quartiles, and standard deviation of the distributions of $\mathcal{Y}_{i2}^{max}$, $\overline{\mathcal{Y}}_{i2}$, and $\mathcal{G}_{i2}$ in columns (1), (2) and (3), respectively. Panel A shows the distributions for the full sample of Phase 2 respondents. For this sample, almost 27% reported a choice probability equal to one for the occupation for which they reported the highest expected earnings.[37] Thus, the expected willingness-to-pay for non-pecuniary factors is $\mathcal{G}_{i2} = 0$ for all individuals in the bottom quartile of this measure. The mean willingness-to-pay is slightly less than \$30,000, which is about 14% of the maximum earnings individuals expect to receive.[38] Hence there is an alignment between the occupation that maximizes their earnings and their preferences for particular occupations, which tends to push the willingness-to-pay downward. NOT SURE

Panel B of Table 5 repeats Panel A for the 73% of respondents who were not certain of choosing a career that maximized their expected income, i.e., those for whom $\mathcal{G}_{i2} > 0$.[39] Note that these individuals tend to have lower maximum earnings than those who are certain of choosing their earnings-maximizing career; at all quartiles, the maximum earnings are lower (or equal) in Panel B than those in Panel A. Furthermore, as expected, this group has larger *ex ante* earnings losses. On average, this group expects to give up almost \$41,000 of annual earnings ten years after college, or a little over 21% of their maximum expected earnings, as a result of not choosing their (*ex ante*) earnings-maximizing occupation. However,

[37] An additional 10% of respondent are certain they will be working in a career where their income is *not* maximized. Overall, the respondents to our Phase 2 follow-up survey report a 57.6% chance of working in the occupation where their expected earnings are the highest.

[38] This estimate may seem low in light of the small earnings elasticities obtained in the education literature on major choices. As we will show in the next section, our estimates of occupational earnings elasticities are higher than the estimates in the literature on college major earnings elasticities. Further, occupations where one has an especially high premium may also be occupations that one would prefer for non-pecuniary reasons. For instance, economics majors may see higher pay in business but also prefer working in business.

[39] 29% of the sample members even expect to earn less than what they would under a random allocation across occupations, where we use as weights the aggregate shares in each occupation.

Table 5: Distribution of Maximum and Expected Earnings: Phase 2 Follow-Up Survey Data, 2009 dollars

	Max Earnings $[\mathcal{Y}_{i2}^{max}]$ (1)	Expected Earnings $[\overline{\mathcal{Y}}_{i2}]$ (2)	Difference $[\mathcal{G}_{i2}]$ (3)
Panel A: Full Sample			
Mean	212,946	183,020	29,926
1^{st} quartile	118,815	84,041	0
Median	158,419	143,370	14,258
3^{rd} quartile	237,629	210,405	35,124
Standard Dev.	165,133	148,179	48,427
Panel B: Conditional on $\mathcal{G}_{i2} > 0$			
Mean	193,000	152,126	40,874
1^{st} quartile	111,060	68,120	11,881
Median	158,419	128,478	23,961
3^{rd} quartile	198,024	187,727	47,526
Standard Dev.	154,956	120,947	52,543

DATA: Sample is 112 respondents to Phase 2 survey.
NOTE: 82 sample members (73.2%) were not certain of choosing the career that maximizes their expected earnings ten years out.

the distribution of earnings losses is skewed, with a smaller, but still substantial, median loss of about $24,000.

Taken together, the above results make clear that, in contrast with a pure (expected) earnings-maximizing model, non-pecuniary factors play an important role in explaining occupational choices. As we show in Appendix A.8, this remains true for two alternative specifications of the earnings component of the expected utility. In these specifications we replace the expected earnings ten years out by i) the discounted expected lifetime earnings, where we allow for occupation-specific growth rates (see Subsections A.8.1 and A.8.2), and ii) the expected lifetime utility associated with future earnings, where individuals are endowed with CRRA preferences (see Subsection A.8.3), and where we allow for heterogeneous earnings risks across occupations.[40] Notably, for both of these specifications, 76.8% of respondents were not certain of choosing the career that maximizes the expected lifetime utility associated with future earnings. This number turns out to be very similar to the share of respondents (73.2%) who were not certain of choosing a career that maximized their expected earnings ten years out. The results from the alternative specification i) further point to similar preferences for non-pecuniary factors to our baseline specification, with an associated mean expected lifetime earnings loss equal to 15.4% of the maximum expected lifetime earnings. Results from Specification ii) that also takes into account differences in risks across occupations indicate that the

[40] In the case that accounts for risk aversion, strong assumptions are needed regarding the predictability of future earnings shocks. See Cunha and Heckman (2016) for ways of measuring what individuals know about future earnings.

mean expected utility loss is equal to 42.8% of the maximum earnings component of the expected utility. In Subsection A.8.3 we show that this is equivalent to a sizable 11.4 log points permanent increase in mean prior (log-)beliefs about earnings.

5.2 Sorting on Expected Earnings: Evidence from a Generalization of the Roy Model

In this section we quantify the importance of sorting across occupations on expected monetary gains. To do so, we consider a variant of the Generalized Roy model that relates the choice of occupations to the elicited earnings beliefs and the (unobserved) non-pecuniary occupation attributes, which have been shown in the previous section to play an important role.

5.2.1 Model and Econometric Specifications

We model occupational choice using a simple static framework. Assume individual i at time t^* commits to an occupation j so as to maximize his expected utility. We assume that the expected utility of occupation j, U_{ijt^*}, is the sum of two components. The first is the pecuniary component that is a function of the earnings beliefs at time t^*, $\mathcal{Y}_{ijt^*}$. The second is the non-pecuniary component, that includes psychic and other costs associated with entering occupation j and other non-pecuniary factors of occupation j that individuals value.[41]

Our beliefs data on the probability of working in particular occupations capture the fact that some uncertainty will be resolved on the value of occupations before agents make their decision. Denote ϵ_{ijt} as the uncertainty to be resolved for occupation j when beliefs are captured at time t. Under the assumption that the ϵ_{ijt}'s are perceived to be distributed following a standard Type I extreme value distribution, we can invert the subjective probabilities, p_{ijt}, to obtain the the expected utility at time t for occupation j, U_{ijt}, relative to the baseline occupation:

$$\begin{aligned} \ln[p_{ijt}] - \ln[p_{i1t}] &= U_{ijt} - U_{i1t} \\ &= \Delta U_{ijt} \end{aligned} \tag{5.4}$$

for $j = 2, ..., 6$, where Δ denotes the differencing operator taken with respect to the baseline occupation $j = 1$.[42] In practice, we approximate the pecuniary component of U_{ijt} with the log of our elicited beliefs on earnings ten years out, $\mathcal{Y}_{ijt}$ where t refers to Phase 1 ($t = 1$) or Phase 2 ($t = 2$). We specify the non-pecuniary component as a function of observable characteristics, X_i, and unobserved characteristics, η_{ijt}, where the coefficient on X_i and η_{ijt} for the baseline occupation are normalized to zero.[43] Our initial

[41] Such costs may arise because of occupation-specific requirements, such as having a M.D. to enter a career in Health, passing the bar exam to enter Law, etc. Similar costs are included in versions of generalized Roy model of educational choice estimated in Carneiro et al. (2003), D'Haultfœuille and Maurel (2013), Eisenhauer et al. (2015), and Heckman et al. (2018).

[42] The expected utility of occupation j at time t and t^* are linked through the relationship $U_{ijt^*} = U_{ijt} + \epsilon_{ijt}$.

[43] This specification allows for preferences for particular occupations to vary across individuals based on observed as well as unobserved individual characteristics.

estimating equation is then:

$$\Delta \ln [p_{ijt}] = X_i \alpha_{jt} + \beta \Delta \ln (\mathcal{Y}_{ijt}) + \eta_{ijt} \tag{5.5}$$

We assume that, conditional on X_i, the earnings term is orthogonal to η_{ijt}. This assumption will be violated if, for instance, preferences for particular occupations not captured by X_i are higher for occupations where expected earnings are higher.

Importantly, our subjective expectations data allow us to deal with this potential issue in two ways. While up until now we have focused on beliefs conditional on the individual's chosen major, beliefs regarding earnings and probabilities of working in particular occupations were also collected for counterfactual majors. Hence, we have six measures of the probability of working in any given occupation, one for each of the six majors. This information allows to estimate (5.5) while including individual fixed effects for each occupation.

Our second approach is to use the panel data on respondents' elicited beliefs to difference out the permanent component of preferences for occupations. This ensures that occupation-major-individual fixed effects cancel out, so that the estimated sorting effects are robust to any preferences for occupation- and major-specific job attributes that may be correlated with expected earnings. Our estimating equation in this case is:

$$\Delta \ln [p_{ij2}] - \Delta \ln [p_{ij1}] = X_i(\alpha_{j2} - \alpha_{j1}) + \beta \left[\Delta \ln (\mathcal{Y}_{ij2}) - \Delta \ln (\mathcal{Y}_{ij1})\right] + \eta_{ij2} - \eta_{ij1} \tag{5.6}$$

where $\eta_{ij2} - \eta_{ij1}$ is the composite error term, having differenced out fixed effects.

5.2.2 Estimation Results

In Table 6 we present estimates for various versions of our occupational choice model, using elicited future earnings beliefs ($\mathcal{Y}_{ijt}$) and actual as well as subjective probabilistic data on occupational choices (d_{ij} and p_{ijt}, respectively). The measures of these variables used for particular specifications are noted in the column headings. For most of these specifications, we use elicited measures for a respondent's chosen major, although the specification for column (4) of the table uses measures for all of the possible majors.[44] We also estimate models with different sets of controls, which are indicated at the bottom of the table.

The estimates in column (1) of Table 6 are based on actual occupational choices in Phase 2 and earnings beliefs elicited in Phase 1. We use a conditional logit model to estimate β. The estimate of β equals 1.423, consistent with positive sorting on expected earnings across occupations, and is precisely estimated. Note that, consistent with the results discussed in Subsection 5.1, we strongly reject the hypothesis that occupation fixed-effects are all equal to zero (with a P-value $< 10^{-4}$), indicating that a pure (expected) earnings-maximizing version of the model is rejected by the data.[45]

[44] In Appendix A.9 we present Phase 1 estimates for a sample that excludes individuals who were seniors at Phase 1. Our main findings are robust to the use of this alternative estimation sample.

[45] Note that, strictly speaking, the terminology pure (expected) earnings-maximizing is an abuse of language since, even absent occupation fixed-effects, idiosyncratic shocks enter the occupation-specific utilities.

Table 6: Estimates of returns to (log of) expected earnings in occupational choice

Earnings Beliefs Used:	Phase 1 ($\mathcal{Y}_{ij1}$)				Phase 2 ($\mathcal{Y}_{ij2}$)			Phases 1 & 2 ($\mathcal{Y}_{ij1}$, $\mathcal{Y}_{ij2}$)	
Occup. Measured by:	Actual (d_{ij})	Phase 1 Probs. (p_{ij1})			Phase 2 Probs. (p_{ij2})			Phases 1 & 2 Probs. (p_{ij1}, p_{ij2})	
	(1)[1]	(2)	(3)	(4)[2]	(5)	(6)	(7)	(8)	(9)
Log Earnings	1.423	1.371	1.000	0.953	1.848	1.257	0.914		
	(0.294)	(0.271)	(0.332)	(0.148)	(0.202)	(0.182)	(0.159)		
Δ Log Earnings								1.274	1.020
								(0.239)	(0.235)
Controls:									
Occupation Only	Y	Y	N	N	Y	N	N	N	N
Major $\times$ Occupation	N	N	Y	Y	N	Y	Y	N	N
Individual $\times$ Occupation	N	N	N	Y	N	N	N	Y	Y
Occupation switching cost	N	N	N	N	N	N	Y	N	Y

DATA: The estimation sample includes 167 individuals that completed the Phase 1 survey and 112 individuals that also completed the Phase 2 follow-up survey.
NOTES:
[1] Estimates in column (1) are produced with a conditional logit model, using data on chosen occupations, d_{ij}. Estimates in all other columns use elicited data on occupational choice probabilities and are produced with a least absolute deviation (LAD) estimator.
[2] Column (4) uses Phase 1 data on respondents' elicitations of expected earnings and occupational choice probabilities for each possible major-occupation pair, providing 6 times the number of observations in the sample ($N = 1,002$). All other columns use elicited data only for the chosen college major of sample members.
[*] Standard errors in parentheses. The standard errors for estimates in column (4) are clustered at the individual $\times$ occupation level.

The estimates for the remaining columns in Table 6 use data on elicited choice probabilities, our *ex ante* measures of respondents' likelihood of choosing particular occupations. We first estimate Equation (5.5). To deal with the fact that some respondents report zero choice probabilities for some occupations, we replace them with an arbitrarily small number, as proposed by Blass et al. (2010), and then estimate the flow utility parameters using a least absolute deviation (LAD) estimator.[46]

Consider the LAD estimates in columns (2) and (3) based on expected earnings and occupational choice probabilities from Phase 1, and those in columns (5) through (7) based on Phase 2 data.[47] While the estimates of β vary, they are all positive and statistically significant at any standard levels.[48]

The estimators used to produce columns (1)-(3) and (5)-(7) of Table 6 maintained strong assumptions about the relationship between the unobserved utility components, η_{ijt}, and the elicited future earnings ($\mathcal{Y}_{ijt}$). Estimates for our more flexible specification that accounts for individual fixed effects for each occupation are presented in column (4) of Table 6. The estimate of β is 0.953, which remains economically as well as statistically significant.

Finally, we again use a LAD estimator to estimate our specification (5.6). Estimates for this specification that uses panel data on respondents' elicited beliefs are reported in columns (8) and (9). Estimates in column (9) include dummies for chosen occupation, as in the specification used to produce column (7) estimates. As with the estimates in the preceding columns, in these specifications the estimated earnings coefficients remain positive and significant, with magnitudes 1.27 (column 8) and 1.02 (column 9).

Taken together, these results provide robust evidence that individuals sort across occupations based on the earnings they would expect to receive in them. Importantly, and although non-pecuniary factors play an important role in occupational choice, our findings from Section 4 of a positive association between expected earnings and occupational choice still hold when we account for these non-pecuniary determinants.

To conclude this section, we quantify the responsiveness of occupational choice to expected earnings by evaluating the elasticities of the *ex ante* choice probabilities to expected earnings. Following Train (2003), for each individual i and occupation j, the earnings elasticity, denoted by e_{ij}, is given by:

$$e_{ij} = [1 - p_{ij1}]\beta, \tag{5.7}$$

which is estimated by replacing β with its estimate $\widehat{\beta}$ ($\widehat{\beta} = 1.020$ for our preferred specification from column (9) of Table 6).[49] The estimated elasticities $\widehat{e}_{ij}$ range from 0.04 to 1.01. The average occupation-specific elasticity estimates range from 0.70 (for Business) to 0.87 (for Education), with a mean elasticity

[46] This estimator is consistent under a zero conditional median restriction on the residuals.

[47] The specification reported in column (7) also includes as a control whether the future occupation is the same as the Phase 2 occupation in order to account for switching costs

[48] We also strongly reject again here the hypothesis that agents sort across occupations according to a pure (expected) income-maximizing choice model (P-value for the joint significance of occupation and major-specific non-earnings components $< 10^{-4}$).

[49] Note that this formula only applies for the intensive margin, that is for variation in the subjective probability p_{ij1} strictly between 0 and 1. It follows that we estimate these elasticities using only the subsample on individuals who provided non-zero choice probabilities.

across all occupations of 0.79.[50] That is, on average across individuals and occupations, a 10% increase in the expected earnings for a given occupation is associated with a 7.9% increase in the subjective probability of choosing that occupation. These elasticities are sizable, especially in comparison with those found in the related context of college major choices.[51]

6 Conclusion

This paper uses data on subjective beliefs about expected earnings in different occupations as well as on the probabilities of working in each of those occupations to provide new evidence on the determinants of occupational choice.

We use elicited beliefs data from a sample of individuals who were first interviewed as undergraduate students at Duke University and then followed up to seven years later, making it possible to compare subjective beliefs with realized labor market outcomes. Importantly, data on earnings beliefs and subjective probabilities of working in particular occupations are highly predictive of future actual earnings and occupational choices, illustrating the economic relevance of these elicited beliefs. We document large differences in the earnings individuals expect to earn in different occupations. We also find that individuals place higher probabilities of working in those occupations in which they have higher monetary returns, consistent with selection on expected earnings gains. Viewed through the lens of *ex ante* treatment effects, we quantify the importance of selection on expected gains from elicited beliefs data only.

We also provide direct evidence that individuals, however, are willing to forego sizable amounts of money when choosing their occupation, complementing prior evidence based on realized data that pure income-maximizing models fail to adequately capture individual choices (Heckman and Sedlacek, 1990). Using a variant of the generalized Roy model in which agents choose their occupation by taking expected earnings and occupation-specific non-pecuniary factors into account, we provide robust evidence of sorting on expected earnings, with large estimated earnings elasticities of occupational choices.

References

Altonji, Joseph, Peter Arcidiacono, and Arnaud Maurel, "The Analysis of Field Choice in College and Graduate School: Determinants and Wage Effects," in Eric Hanushek, Stephen Machin, and Ludger Wößmann, eds., *Handbook of the Economics of Education*, Vol. 5, Elsevier, 2016, pp. 305–396.

[50] The elasticity estimates for the other occupations are equal to 0.75 (Science), 0.77 (Health), 0.84 (Government) and 0.81 (Law).

[51] See, for example, the expected college major earnings elasticities in Beffy et al. (2012), Long et al. (2015), Wiswall and Zafar (2015), and Altonji et al. (2016) for a survey of the latter studies. Most of these studies use *ex post* earnings data to characterize *ex ante* earnings beliefs, rather than elicited ones, although Wiswall and Zafar (2015) do use elicited beliefs in their study of college major choices. That the estimated elasticities are larger than in earlier papers using observational data is consistent with our elicited beliefs being more accurate measurements of the true beliefs of the agents at the time of the choice. Differences in timing of choices, combined with major switching costs, may explain why our elasticities are also larger than in Wiswall and Zafar (2015).

Arcidiacono, Peter, "Ability Sorting and the Returns to College Major," *Journal of Econometrics*, 2004, *121* (1–2), 343–375.

_ , Esteban. Aucejo, Hanming Fang, and Kenneth Spenner, "Does Affirmative Action Lead to Mismatch? A New Test and Evidence," *Quantitative Economics*, 2011, *2* (3), 303–333.

_ , V. Joseph Hotz, and Songman Kang, "Modeling College Major Choices using Elicited Measures of Expectations and Counterfactuals," *Journal of Econometrics*, 2012, *166*, 3–16.

Attanasio, Orazio P. and Katja M. Kaufmann, "Education Choices and Returns to Schooling: Mothers' and Youth's Subjective Expectations and their Role by Gender," *Journal of Development Economics*, 2014, *109*, 208–216.

_ and _ , "Educational Choices and Returns on the Labor and Marriage Markets: Evidence from Data on Subjective Expectations," *Journal of Economic Behavior & Organization*, 2017, *140*, 35–55.

Barton, Benjamin H., *The Glass Half Full: The Decline and Rebirth of the Legal Profession*, Oxford University Press, 2015.

Beffy, Magali, Denis Fougere, and Arnaud Maurel, "Choosing the Field of Studies in Postsecondary Education: Do Expected Earnings Matter?," *Review of Economics and Statistics*, 2012, *94* (1), 334–347.

Blass, Asher A., Saul Lach, and Charles F. Manski, "Using Elicited Choice Probabilities to Estimate Random Utility Models: Preferences for Electricity Reliability," *International Economic Review*, 2010, *51* (2), 421–440.

Carneiro, Pedro M., Karsten T. Hansen, and James J. Heckman, "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 2003, *44* (2), 361–422.

Cunha, Flávio and James J. Heckman, "Identifying and Estimating the Distributions of *Ex Post* and *Ex Ante* Returns to Schooling," *Labour Economics*, 2007, *14* (6), 870–893.

_ and _ , "A New Framework for the Analysis of Inequality," *Macroeconomic Dynamics*, 2008, *12*, 315–354.

_ and _ , "Decomposing Trends in Inequality in Earnings into Forecastable and Uncertain Components," *Journal of Labor Economics*, 2016, *34* (2), S32–S65.

_ , _ , and Salvador Navarro, "Separating Uncertainty from Heterogeneity in Life Cycle Earnings," *Oxford Economic Papers*, April 2005, *57* (2), 191–261.

Delavande, Adeline, "Pill, Patch, or Shot? Subjective Expectations and Birth Control Choice," *International Economic Review*, 2008, *49* (3), 999–1042.

D'Haultfœuille, Xavier and Arnaud Maurel, "Inference on an Extended Roy Model, with an Application to Schooling Decisions in France," *Journal of Econometrics*, 2013, *174* (2), 95–106.

D'Haultfoeuille, Xavier, Christophe Gaillac, and Arnaud Maurel, "Rationalizing Rational Expectations? Tests and Deviations," 2018. NBER Working Paper No. 25274.

Dillon, Eleanor W., "Risk and Return Trade-Offs in Lifetime Earnings," *Journal of Labor Economics*, 2018, *36* (4), 981–1021.

Dominitz, Jeff and Charles F. Manski, "Eliciting Student Expectations of the Returns to Schooling," *Journal of Human Resources*, 1996, *31* (1), 1–26.

_ and _ , "Using Expectations Data to Study Subjective Income Expectations," *Journal of the American Statistical Association*, 1997, *92* (439), 855–867.

Eisenhauer, Philipp, James J. Heckman, and Edward Vytlacil, "The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs," *Journal of Political Economy*, 2015, *123* (2), 413–443.

Gibbons, Robert, Lawrence F. Katz, Thomas Lemieux, and Daniel Parent, "Comparative Advantage, Learning, and Sectoral Wage Determination," *Journal of Labor Economics*, 2005, *23* (4), 681–724.

Giustinelli, Pamela and Matthew D. Shapiro, "SeaTE: Subjective *Ex Ante* Treatment Effect of Health on Retirement," 2019. NBER Working Paper No. 26087.

Hai, Rong and James J. Heckman, "Inequality in Human Capital and Endogenous Credit Constraints," *Review of Economic Dynamics*, 2017, *25*, 4–36.

Heckman, James J. and Edward J. Vytlacil, "Econometric Evaluation of Social Programs: Causal Models, Structural Models and Econometric Policy Evaluations," in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Elsevier, 2007.

_ **and Guilherme L. Sedlacek**, "Self-selection and the Distribution of Hourly Wages," *Journal of Labor Economics*, 1990, *8* (1, Part 2), S329–S363.

_ **, John E. Humphries, and Gregory Veramendi**, "Returns to Education: The Causal Effects of Education on Earnings, Health and Smoking," *Journal of Political Economy*, 2018, *126* (S1), S197–S246.

Hurd, Michael, "Subjective Probabilities in Household Surveys," *Annual Review of Economics*, 2009, *1*, 543–562.

_ **and Kathleen McGarry**, "Evaluation of the Subjective Probabilities of Survival in the HRS," *Journal of Human Resources*, 1995, *30*, S268–S292.

_ **and** _ , "The Predictive Validity of Subjective Probabilities of Survival," *Economic Journal*, 2002, *112* (482), 966–985.

Kaufmann, Katja M., "Understanding the Income Gradient in College Attendance in Mexico: The Role of Heterogeneity in Expected Returns," *Quantitative Economics*, 2014, *5* (3), 583–630.

Keane, Michael and Kenneth I. Wolpin, "The Career Decisions of Young Men," *Journal of Political Economy*, 1997, *105* (3), 473–522.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad, "Field of Study, Earnings, and Self-Selection," *Quarterly Journal of Economics*, 2016, *131* (3), 1057–1111.

Lemieux, Thomas, "Occupations, fields of study and returns to education," *Canadian Journal of Economics*, 2014, *47* (4), 1047–1077.

Long, Mark C., Dan Goldhaber, and Nick Huntington-Klein, "Do Students' College Major Choices Reflect Changes in Wages?," *Economics of Education Review*, 2015, *49*, 1–14.

Manski, Charles F., "Adolescent Econometricians: How Do Youths Infer the Returns to Schooling?," in Charles T. Clotfelter and Michael Rothschild, eds., *Studies of Supply and Demand in Higher Education*, Chicago: University of Chicago Press, 1993, pp. 43–57.

_ , "Measuring Expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

Miller, Robert A., "Job Matching and Occupational Choice," *Journal of Political Economy*, 1984, *92* (6), 1086–1120.

Schennach, Suzanne, "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 2016, *8*, 341–377.

Siow, Alioysus, "Occupational Choice under Uncertainty," *Econometrica*, 1984, *52* (3), 631–645.

Stinebrickner, Todd and Ralph Stinebrickner, "A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout," *Review of Economic Studies*, 2014, *81* (1), 426–472.

Train, Kenneth, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2003.

van der Klaauw, Wilbert, "On the Use of Expectations Data in Estimating Structural Dynamic Choice Models," *Journal of Labor Economics*, 2012, *30* (3), 521–554.

_ **and Kenneth I. Wolpin**, "Social Security and the Retirement and Savings Behavior of Low-Income Households," *Journal of Econometrics*, 2008, *145* (1–2), 21–42.

Wiswall, Matthew and Basit Zafar, "Determinants of College Major Choice: Identification using an Information Experiment," *Review of Economic Studies*, 2015, *82* (2), 791–824.

_ **and** _ , "Human Capital Investments and Expectations about Career and Family," August 2018. Unpublished manuscript.

Zafar, Basit, "College Major Choice and the Gender Gap," *Journal of Human Resources*, 2013, *48* (3), 545–595.

A Appendix

A.1 Actual Majors at Duke and Major Groups

The following is the list of majors at Duke and the six groups we used to classify them:

Table A.1: Major Groups and Actual Majors Offered at Duke University

Science	***Engineering***
Biological Anthropology and Anatomy	Computer Science
Biology	Biomedical Engineering
Chemistry	Civil Engineering
Earth & Ocean Sciences	Electrical & Computer Engineering
Mathematics	Mechanical Engineering
Physics	
Humanities	***Social Sciences***
Art History	Cultural Anthropology
Asian and African Languages and Literature	History
Classical Civilization/Classical Languages	Linguistics
Dance	Psychology
English	Sociology
French Studies	Women's Studies
German	
International Comparative Studies	***Economics***
Italian Studies	Economics
Literature	
Medieval & Renaissance Studies	***Public Policy***
Music	Environmental Science and Policy
Philosophy	Political Science
Religion	Public Policy Studies
Spanish	
Theater Studies	
Visual Arts	

A.2 Descriptive Statistics of Samples

Table A.2 presents a descriptive overview of our sample. The composition of our sample corresponds fairly closely to the Duke male undergraduate student body. The sample includes slightly more Asians and fewer Hispanics and Blacks than in the Duke male student body, and it over-represents students in natural sciences majors while under-representing students in public policy. Finally, the sample is tilted towards upper-classmen.

Table A.3 reports the means of the expected incomes for the various major-occupation combinations collected in Phase 1 of the DuCMES.[52] Note that each cell contains averages of the responses by each of the 173 students. Expected incomes exhibit sizable variation both across majors and occupations. For instance, majoring in the natural sciences or engineering is perceived to lead to higher earnings in Science and Health careers, while expected earnings in Business are, on average, higher for economics majors. Differences across occupations are even starker. In particular, average expected incomes are lowest for a career in Education and generally highest for a career in Law, with the exception of natural sciences and economics majors, for which expected incomes are highest for Health and Business occupations, respectively.

Table A.4 presents the averages for the subjective probabilities of working in each occupation that were elicited from students in the Phase 1. The subjective probabilities of entering each occupation vary substantially across majors. At the same time, none of the majors are concentrated into one, or even two, occupations. For any given major, the average subjective probabilities are larger than 10% for at least three occupations.

Table A.5 reports the prevalence of students reporting that the probability they would choose a particular occupation was zero for each major-occupation combination.[53] While some combinations display a large share of zero subjective probabilities, the shares are all below 60%, suggesting that particular majors do not rule out certain occupations for all individuals.

Finally, Table A.6 compares the characteristics of our Phase 1 and Phase 2 samples, showing that, with the exception of one entry, the characteristics of the two samples are not statistically different at the five percent level.

[52] In our sample, only 1.6% of the expected earnings are missing. For these cases, expected earnings, for each major and occupation, are set equal to the predicted earnings computed from a linear regression of log-earnings on major and occupation indicators, interaction between major and occupation, individual-specific average log-earnings across all occupations and majors, and an indicator for whether the subjective probability of working in this occupation is equal to zero. One individual in our sample declared that he expected to earn $1,000 for some occupation-major combinations. We assume that this individual declared monthly rather than yearly incomes, and rescaled his expected income accordingly.

[53] The survey design was such that the default values of the subjective probabilities were set equal to zero for all occupation-major combinations. As a result, it might be that some of the zero probabilities observed in the data reflect missing probabilities rather than true zeros. However, in the former case, it seems likely that the latent (unobserved) probabilities are typically close to zero, so that aggregating these two types of zero probabilities should not be much of a concern.

Table A.2: Descriptive Statistics for Phase 1 Sample

	Sample	Duke Male Student Body
Current/Intended Major:		
Sciences	17.9%	14.8%
Humanities	9.3%	9.4%
Engineering	19.1%	20.7%
Social Sciences	17.9%	18.8%
Economics	19.7%	18.0%
Public Policy	16.2%	18.0%
Class/Year at Duke:		
Freshman	20.8%	
Sophomore	20.2%	
Junior	27.2%	
Senior	31.8%	
Characteristics of Students:		
White	66.5%	66.0%
Asian	20.2%	16.6%
Hispanic	4.6%*	8.3%
Black	4.0%	5.9%
Other	4.6%	3.0%
U.S. Citizen	94.8%	94.1%
Sample Size	173	

DATA SOURCES: Phase 1 of DuCMES for the sample characteristics and Campus Life and Learning (CLL) Project at Duke University for Duke Male Student Body. See Arcidiacono et al. (2011) for a detailed description of the CLL dataset.

NOTES: Current/Intended Major: Respondents were asked to choose one of the six choices (Sciences, humanities, engineering, social Science, economics, public policy) in response to the questions: "What is your current field of study?" "If you have not declared your major, what is your intended field of study?". * indicates significance at the 10% level of the difference in proportions between Phase 1 DuCMES sample and Duke male student body.

Table A.3: Mean of Phase 1 Expected Incomes for Different Major/Occupation Combinations 10 Years after Graduation (Annual Incomes, in dollars)

	Occupation:					
Major:	Science	Health	Business	Government	Education	Law
Natural Sciences	109,335	162,636	139,527	95,628	73,597	145,846
Humanities	82,897	126,891	131,254	92,024	71,925	149,058
Engineering	119,601	153,935	154,274	98,738	76,229	167,650
Social Sciences	86,686	126,614	145,856	96,632	71,996	151,323
Economics	96,004	131,822	198,665	103,085	79,303	160,526
Public Policy	90,319	126,521	157,341	110,517	72,928	166,211

DATA: Sample who completed Phase 1 survey ($N = 173$).
NOTES: Expected earnings were elicited for each possible major-occupation pair at Phase 1, regardless of the respondents' chosen or intended major.

Table A.4: Mean of Phase 1 Elicited Probabilities of Choosing Alternative Occupations, conditional on Majoring in Alternative Fields

	Occupation:					
Major:	Science	Health	Business	Government	Education	Law
Natural Sciences	0.352	0.319	0.120	0.070	0.068	0.070
Humanities	0.067	0.122	0.235	0.145	0.230	0.200
Engineering	0.411	0.194	0.190	0.072	0.065	0.068
Social Sciences	0.091	0.139	0.246	0.193	0.128	0.204
Economics	0.067	0.076	0.515	0.154	0.062	0.125
Public Policy	0.054	0.113	0.228	0.317	0.075	0.214

DATA: Sample who completed Phase 1 survey ($N = 173$).
NOTES: Probabilities were elicited for each possible major-occupation pair at Phase 1, regardless of the respondents' chosen or intended major.

Table A.5: Incidence of Elicited Zero Probabilities of Choosing Occupations in Phase 1, conditional on Majoring in Alternative Fields

	Occupation:					
Major:	Science	Health	Business	Government	Education	Law
Natural Sciences	4.62%	9.25%	30.06%	37.57%	41.04%	44.51%
Humanities	50.29%	35.84%	15.61%	20.81%	19.08%	17.92%
Engineering	8.09%	24.28%	22.54%	46.82%	48.55%	51.45%
Social Sciences	46.82%	32.95%	12.14%	15.03%	27.17%	18.50%
Economics	53.76%	50.87%	3.47%	19.65%	46.82%	30.64%
Public Policy	56.65%	38.15%	15.03%	5.78%	40.46%	12.72%

NOTES: Major can either be the chosen major or a counterfactual major ($N = 173$).

Table A.6: Comparison of Samples that completed the Phase 2 and Phase 1 Surveys

	Phase 2 Sample	Phase 1 Sample
Current/Intended Major:		
Sciences	17.9%	17.9%
Humanities	8.9%	9.3%
Engineering	21.4%	19.1%
Social Sciences	15.2%	17.9%
Economics	21.4%	19.7%
Public Policy	15.2%	16.2%
Class/Year at Duke:		
Freshman	21.4%	20.8%
Sophomore	18.8%	20.2%
Junior	26.8%	27.2%
Senior	33.0%	31.8%
Characteristics of Students:		
White	70.5%	66.5%
Asian	20.5%	20.2%
Hispanic	3.6%	4.6%
Black	1.8%	4.0%
Other	3.6%	4.6%
U.S. Citizen	96.4%	94.8%
Receives Financial Aid	41.1%	40.5%
Mean Subjective Probability (Phase 1):*		
Science	0.182	0.180
Health	0.181	0.171
Business	0.273	0.266
Government	0.136	0.124
Law	0.142	0.169
Education	0.086	0.095
Mean Expected Earnings (Phase 1):*		
Science	$92,598	$96,790
Health	$143,036	$142,540
Business	$160,420	$164,010
Government	$97,813	$100,350
Law	$150,214	$163,220
Education	$75,929	$74,470
Mean Realized Earnings (7 years later):**		
	$131,527	
Sample Sizes	112	173

DATA SOURCES: DuCMES Phase 1 and Phase 2 samples.
NOTES: Current/Intended Major: Respondents were asked to choose one of the six choices (natural sciences, humanities, engineering, social sciences, economics, public policy) in response to the questions "What is your current field of study? If you have not declared your major, what is your intended field of study?"'.
* Conditional on chosen/intended major.
** Earnings expressed in 2009 dollars, average over 81 individuals with non-missing earnings in Phase 2 follow-up survey.
The proportion of black students is the only characteristic that is significantly different between both samples (at the 5% level).

A.3 Dispersion of Earnings Beliefs about the Average Duke Student

Table A.7: Differences in Variances of the Log of Elicited Expected Incomes for the Average Duke Student between Upper- and Lower-Classmen

	Occupation:						
Major:	Science	Health	Business	Government	Education	Law	All
Natural Sciences	-0.16**	-0.05	-0.07	-0.15**	-0.14	-0.08	-0.11**
Humanities	-0.11	-0.01	-0.06	-0.19**	-0.27**	-0.09	-0.12**
Engineering	-0.14*	-0.17	-0.05	-0.13*	-0.14	-0.01	-0.11*
Social Sciences	-0.12	-0.05	0.00	-0.14	-0.18*	-0.09	-0.10*
Economics	-0.04	-0.01	-0.03	-0.10	0.37	-0.08	0.02
Public Policy	-0.07	-0.06	-0.08	-0.10*	-0.18*	-0.02	-0.09
All	-0.11*	-0.06	-0.05	-0.13**	-0.09*	-0.06	-0.08***

DATA: Data from Phase 1 ($N = 173$).
NOTES: "All" indicates average across majors (rows) and occupations (columns). *, ** and *** indicate statistical significance of differences at the 10%, 5%, and 1%, respectively.

Table A.8: Differences in Variances of the Log of Elicited Expected Incomes for the Average Duke Student between Chosen and Non-Chosen Majors

	Occupation:						
Major:	Science	Health	Business	Government	Education	Law	All
Natural Sciences	0.06	0.04	0.03	-0.01	-0.12**	0.09	0.05
Humanities	-0.07	-0.06	-0.13	-0.16***	-0.21***	-0.13	-0.12***
Engineering	-0.10**	-0.21***	-0.09	-0.08**	-0.12***	-0.17**	-0.14***
Social Sciences	-0.07	0.01	0.07	-0.05	0.06	0.01	0.01
Economics	-0.12***	-0.08	-0.06	-0.06	-0.48***	-0.10*	-0.13***
Public Policy	0.24***	0.14*	0.10	0.24***	0.31***	0.17**	0.18***
All	-0.01	-0.01	0.00	-0.01	-0.07***	-0.01	-0.02

DATA: Data from Phase 1 ($N = 173$).
NOTES: "All" indicates average across majors (rows) and occupations(columns). *, **, and *** indicate statistical significance of differences at the 10%, 5%, and 1%, respectively.

A.4 Robustness to measurement errors

A.4.1 Measurement errors on earnings beliefs

In the following we consider a situation where occupation-specific earnings beliefs are measured with error. Specifically, assume that for any given occupation j, the true earnings beliefs $\mathcal{Y}_{ijt}$ are unobserved. Instead, we only observe $\widetilde{\mathcal{Y}}_{ijt}$, which are affected by measurement errors. Namely:

$$\widetilde{\mathcal{Y}}_{ijt} = \mathcal{Y}_{ijt} + \xi^{\mathcal{Y}}_{ijt}, \tag{A.1}$$

where $\xi^{\mathcal{Y}}_{ijt}$ is the measurement error in earnings beliefs.

With respect to the average *ex ante* treatment effect, ATE_{jt}, in the presence of measurement errors affecting expected earnings, the following population parameter, $\widetilde{ATE}_{jt}$, is directly identified from the data (where Δ denotes the differencing operator with respect to the baseline occupation $j = 1$):

$$\begin{aligned} \widetilde{ATE}_{jt} &= E\left(\Delta\widetilde{\mathcal{Y}}_{ijt}\right) \\ &= ATE_{jt} + E\left(\Delta\xi^{\mathcal{Y}}_{ijt}\right) \end{aligned} \tag{A.2}$$

It follows that the parameter ATE_{jt} remains identified, provided that $E(\xi^{\mathcal{Y}}_{ijt}) = E(\xi^{\mathcal{Y}}_{i1t})$. More generally, the average *ex ante* treatment effects associated with all of the occupations $j \in \{2, .., 6\}$ remain identified from the elicited earnings beliefs, as long as measurement errors have the same mean across occupations.

We next consider the identification of the *ex ante* treatment effect on the treated, TT_{jt}. Provided that measurement errors on the earnings beliefs are uncorrelated with the subjective probabilities, this parameter is identified under the same condition as the average *ex ante* treatment effect. Indeed, note first that the following parameter is directly identified from the data:

$$\begin{aligned} \widetilde{TT}_{jt} &= E\left(\omega_{ijt}\Delta\widetilde{\mathcal{Y}}_{ijt}\right) \\ &= TT_{jt} + E\left(\omega_{ijt}\Delta\xi^{\mathcal{Y}}_{ijt}\right). \end{aligned}$$

where the weights are given by $\omega_{ijt} = p_{ijt}/E(p_{ijt})$. Assuming that measurement errors on the earnings beliefs are uncorrelated with the subjective probabilities yields:

$$\begin{aligned} \widetilde{TT}_{jt} &= TT_{jt} + E(\omega_{ijt})E\left(\xi^{\mathcal{Y}}_{ijt} - \xi^{\mathcal{Y}}_{i1t}\right) \\ &= TT_{jt} + E\left(\xi^{\mathcal{Y}}_{ijt} - \xi^{\mathcal{Y}}_{i1t}\right). \end{aligned} \tag{A.3}$$

It follows that the parameter TT_{jt} remains identified provided that $E(\xi^{\mathcal{Y}}_{ijt}) = E(\xi^{\mathcal{Y}}_{i1t})$. More generally, the average *ex ante* treatment effects on the treated associated with all of the occupations $j \in \{2, .., 6\}$ remain identified from the elicited earnings beliefs, as long as measurement errors have the same mean across occupations. The same arguments apply to the identification of the *ex ante* treatment effect on

the untreated.

A.4.2 Measurement errors on earnings beliefs and subjective probabilities

We now consider the case where both earnings beliefs ($\mathcal{Y}_{ijt}$) and subjective probabilities (p_{ijt}) of choosing particular occupations are measured with error. Namely, instead of the true beliefs $\mathcal{Y}_{ijt}$ and p_{ijt} we observe $\widetilde{\mathcal{Y}}_{ijt}$ and $\widetilde{p}_{ijt}$, which are given by:

$$\widetilde{\mathcal{Y}}_{ijt} = \mathcal{Y}_{ijt} + \xi^{\mathcal{Y}}_{ijt} \tag{A.4}$$

$$\widetilde{p}_{ijt} = p_{ijt} + \xi^{p}_{ijt} \tag{A.5}$$

where $\xi^{\mathcal{Y}}_{ijt}$ and ξ^{p}_{ijt} denote the measurement errors in earnings beliefs and subjective probabilities, respectively.

Clearly, the average *ex ante* treatment effect parameters are identified under the same conditions ($E(\xi^{\mathcal{Y}}_{ijt}) = E(\xi^{\mathcal{Y}}_{i1t})$) as in the previous subsection.

Turning to the *ex ante* treatment effect on the treated, and letting $\widetilde{\omega}_{ijt} = \widetilde{p}_{ijt}/E(\widetilde{p}_{ijt})$, the following parameter is directly identified from the data:

$$\begin{aligned}\widetilde{TT}_{jt} &= E\left(\widetilde{\omega}_{ijt}\Delta\widetilde{\mathcal{Y}}_{ijt}\right) \\ &= TT_{jt} + E\left(\omega_{ijt}\Delta\xi^{\mathcal{Y}}_{ijt}\right) + E\left(\left(\widetilde{\omega}_{ijt} - \omega_{ijt}\right)\Delta\mathcal{Y}_{ijt}\right) + E\left(\left(\widetilde{\omega}_{ijt} - \omega_{ijt}\right)\Delta\xi^{\mathcal{Y}}_{ijt}\right)\end{aligned}$$

Assuming that the measurement errors on the subjective probabilities are mean zero ($E(\xi^{p}_{ijt}) = 0$), it follows that:

$$\widetilde{TT}_{jt} = TT_{jt} + E\left(\omega_{ijt}\Delta\xi^{\mathcal{Y}}_{ijt}\right) + \frac{1}{E(p_{ijt})}\left(E\left(\xi^{p}_{ijt}\Delta\mathcal{Y}_{ijt}\right) + E\left(\xi^{p}_{ijt}\Delta\xi^{\mathcal{Y}}_{ijt}\right)\right) \tag{A.6}$$

Under the assumption that (i) the measurement errors on the earnings beliefs ($\xi^{\mathcal{Y}}_{ijt}$) are uncorrelated with the subjective probabilities (p_{ijt}), (ii) measurement errors on the subjective probabilities (ξ^{p}_{ijt}) are uncorrelated with the earnings beliefs ($\mathcal{Y}_{ijt}$), and that (iii) both types of measurement errors are mutually uncorrelated, it follows that:

$$\widetilde{TT}_{jt} = TT_{jt} \tag{A.7}$$

which implies that the *ex ante* treatment effect on the treated, TT_{jt}, is identified. The same arguments apply to the identification of the *ex ante* treatment effect on the untreated in the presence of measurement errors on earnings beliefs and choice probabilities.

A.5 Distributions of *ex ante* treatment effects

Our elicited expectations data not only allow us to estimate the means of the *ex ante* treatment effects defined in the previous section, but also estimate their distributions. We first consider the estimation of the unconditional distribution of the *ex ante* treatment effects and then turn to the conditional distributions of the *ex ante* treatment effects on the treated and untreated subpopulations. All of the *ex ante* treatment effects are computed for students' chosen college majors, using data from Phase 1. In order to simplify the exposition, we first assume that earnings beliefs and subjective probabilities are measured without error, and then provide at the end of the section conditions under which distribution of *ex ante* treatment effects remain identified if subjective beliefs are measured with error.

The density of the *unconditional* distribution of the *ex ante* treatment effects for occupation j, i.e., $\Delta\mathcal{Y}_{ijt}$, in the overall population is directly identified from the data and can be simply estimated with a kernel density estimator, using the fact that we have direct measures of the *ex ante* treatment effects for each occupation j, $j \in \{2, .., 6\}$, for each student in our sample. We denote the resulting density by $f_{TE,j}(\cdot)$ and its estimator by $\widehat{f}_{TE,j}(\cdot)$.

Next, consider the distributional counterpart of the *ex ante* treatment effects on the treated, which is characterized by a weighted version of $f_{TE,j}(\cdot)$, where the weights are functions of the elicited probabilities of choosing the various occupations. This density function is defined as:

$$f_{TE,j}^{Treated}(u) = \omega_{ijt}(u) \cdot f_{TE,j}(u), \tag{A.8}$$

where $\omega_{ijt}(u) := g(u)/E(p_{ijt})$ and $g(u) = E(p_{ijt}|\Delta\mathcal{Y}_{ijt} = u)$.[54] $f_{TE,j}^{Treated}(.)$ is identified from the subjective beliefs data, and the following plug-in estimator:

$$\widehat{f}_{TE,j}^{Treated}(u) = \widehat{\omega}_{ijt}(u) \cdot \widehat{f}_{TE,k}(u), \tag{A.9}$$

is a consistent estimator of $f_{TE,j}^{Treated}(u)$, where $\widehat{\omega}_{ijt}(u) = \widehat{g}(u)/(N^{-1}\sum_i p_{ijt})$ and $\widehat{g}(u)$ is the Nadaraya-Watson estimator of the nonparametric regression $g(u)$. In the following we will simply refer to $f_{TE,j}^{Treated}(.)$ as the density of the *ex ante* treatment effects on the treated for occupation j. The distribution of the *ex ante* treatment effects on the untreated can be estimated in a similar fashion by replacing p_{ijt} with $1 - p_{ijt}$ in Equation (A.9).

Figure A.1 plots the densities of the *ex ante* treatment on the treated and untreated for Science, Health, Business, Government, and Law occupations, respectively.[55] Each figure shows a different pattern

[54] If individuals form rational expectations over their future occupational choices, it follows from Bayes' rule that $f_{TE,j}^{Treated}(.)$ coincides with the density of the distribution of the *ex ante* treatment effects on the (*ex post*) treated subpopulation. This remains true in the presence of unanticipated aggregate shocks, provided that these shocks affect the shares of workers in each occupation in a multiplicative fashion. While for most occupations we are underpowered to provide informative estimates of the distribution of *ex ante* treatment effects on the treated based on Phase 2 data, estimation results for the two most frequently chosen occupations (Health and Business) point to similar patterns of selection to the ones obtained using subjective probabilities collected in Phase 1.

[55] All densities were estimated using 100 grid points over the support, and a Gaussian kernel with optimal default bandwidth

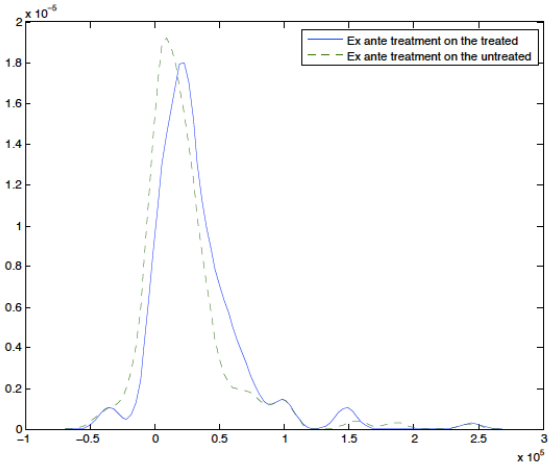
of selection. For Government, the distributions for the treated and the untreated are essentially the same: there is little role for selection into Government jobs, at least relative to Education. For Health, the treated distribution is to the right of the untreated distribution, suggesting substantial selection on expected returns throughout the distribution. For Business careers, while there appears to be significant selection at the bottom end of the distribution, the discrepancy between the two distributions is attenuated in the top end.[56] This latter pattern suggests that there is a group of individuals who would do quite well in Business – essentially as well as the highest returns individuals from the treated group – but whose preferences, or expected earnings in other occupations, lead them away from Business.

The conditions under which the distribution of *ex ante* treatment effects remain identified in the presence of measurement errors on the earnings beliefs and subjective probabilities are, as expected since we are working with distributions instead of mean parameters, stronger than the identifying assumptions for the *ex ante* treatment effect parameters (see Section 4). In particular, identification in this context requires imposing the conditions that measurement errors on earnings beliefs are classical, and drawn from a known distribution with non-vanishing characteristic function. Under these conditions, identification of the distribution of the unconditional distribution of the *ex ante* treatment effects and of the distribution of the *ex ante* treatment effects on the treated follows from standard parametric deconvolution arguments (see Subsection 3.1 in Schennach, 2016).

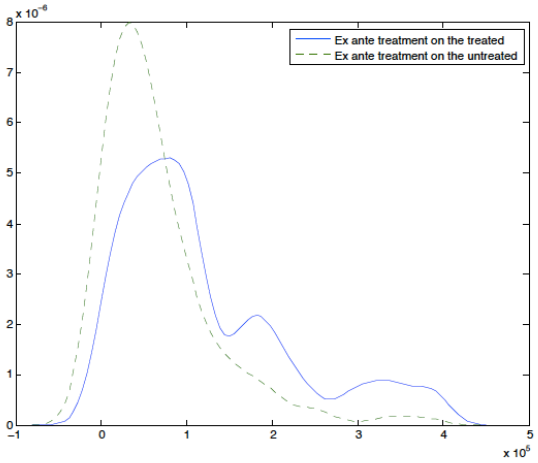
returned by the procedure `ksdensity` in Matlab.

[56] While, for Business, the average *ex ante* treatment on the treated is not significantly different from the average *ex ante* treatment on the untreated, one can indeed reject at 5% the equality of the first quartiles of these two distributions (p-value of 0.015).

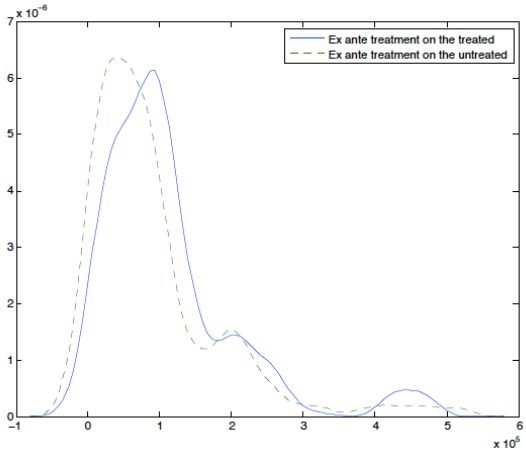
Figure A.1: Densities of *Ex Ante* Treatment Effects of Treated & Untreated Occupations on Expected Earnings (Education is base category)



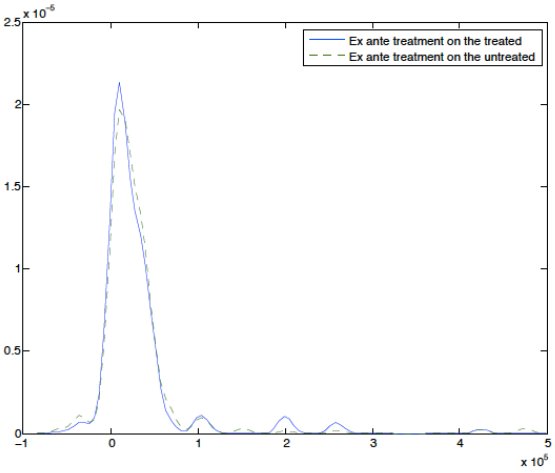
(a) Science



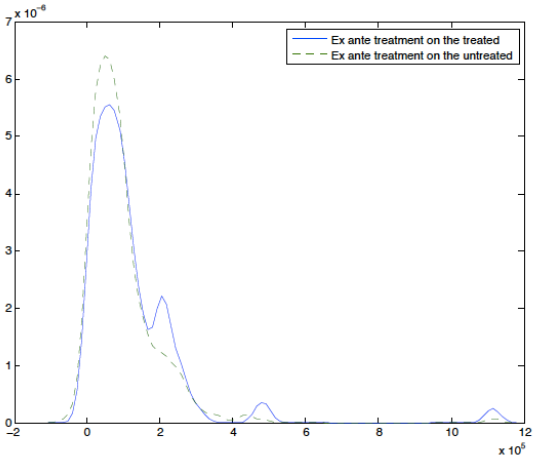
(b) Health



(c) Business



(d) Government



(e) Law

A.6 *Ex Ante* Treatment Effects by Students' Class

Table A.9: Average *Ex Ante* Treatment Effects (ATE) of Occupations: Lower-Classmen versus Upper-Classmen (Annual Earnings, in dollars) using Data elicited in Phase 1

Occupation	Lower-classmen	Upper-classmen	P-value
Science	20,796	23,424	0.66
	(4,652)	(3,733)	
Health	61,657	72,492	0.51
	(13,911)	(8,448)	
Business	75,981	98,961	0.48
	(30,760)	(10,406)	
Government	24,803	26,608	0.83
	(6,333)	(5,627)	
Law	74,450	98,608	0.33
	(19,873)	(15,011)	

NOTES: Standard errors are reported in parentheses. Reported P-values correspond to a t-test of equality of the average *ex ante* treatment effects between lower-classmen and upper-classmen.

A.7 Alternative measures of selection on expected gains

Another way of assessing the role of selection with our data is to construct the *ex ante* analogues of occupation-specific earnings, both unadjusted and adjusted for the selectivity of choosing a particular occupation. Unadjusted *ex post* earnings are just the observed earnings of individuals working in a particular occupation, as would be observed in national data sets such as the American Community Survey (ACS). Using our expectations data, we can produce *ex ante* analogues of both measures. Namely, define the selected *ex ante* earnings for occupation j at time $t \in \{1, 2\}$ (i.e., elicited in Phase 1 or Phase 2) to be $SE_t(j) := E(\omega_{ijt}\mathcal{Y}_{ijt})$, which is consistently estimated using its sample analog ($\widehat{SE}_t(j)$). As with the *ex ante* treatment effect on the treated, this parameter upweights the expected earnings by individuals' subjective probabilities of being in occupation j, thereby mimicking the realized earnings of those who chose to work in occupation j. The corresponding estimator for the selected *ex ante* earnings difference between occupation j and Education ($j = 1$) is given by:

$$\Delta\widehat{SE}_t(j) \quad = \quad \widehat{SE}_t(j) - \widehat{SE}_t(1) \tag{A.10}$$

This is the *ex ante* equivalent of the raw earnings premium of occupation j relative to Education. We can then compare these estimates with the average *ex ante* treatment effects to quantify how much of the selected *ex ante* earnings premium $\Delta\widehat{SE_t(j)}$ is due to selection.

Panel A of Table A.10 performs this decomposition using Phase 1 data. Row (1) displays estimates for the selected earnings $SE_1(j)$ and, as a point of reference, the simple unweighted means of the *ex ante* earnings for occupation j, $\overline{\mathcal{Y}_{j1}} := N^{-1}\sum_i \mathcal{Y}_{ij1}$, are displayed in Row (2). The fact that the means of selected *ex ante* earnings are all greater than $\overline{\mathcal{Y}_{j1}}$ is indicative of positive sorting on expected earnings.

In Rows (3) and (4), we display the occupation-specific estimates of the selected *ex ante* earnings differentials, $\Delta\widehat{SE}_1(j)$, and the average *ex ante* treatment effects, $\widehat{ATE}_{j1}$. The nature of selection into occupations based on *ex ante* returns is illustrated by the relationship between $\Delta\widehat{SE}_1(j)$ and $\widehat{ATE}_{j1}$. In Rows (5) and (6), we show the simple difference between the two (Selection amount), and the share of that difference with respect to the selected *ex ante* earnings differentials $\Delta\widehat{SE}_1(j)$ (Selection share). The selection amounts for all occupations are positive, which is consistent with positive sorting on *ex ante* earnings. Furthermore, the selection share estimates show that selection is much stronger for the Health occupation, least important for the Business occupation, with the other occupations somewhere in between.

Panel B of Table A.10 presents the same statistics as in Panel A, using data from Phase 2. Using these more recent elicited expectations allows us to assess the role selection plays after educational decisions are essentially finalized. Comparing Row (1) across the two panels, there is a sizable increase in respondents' selected *ex ante* earnings for Business occupations ($176,393 vs $294,728), a noticeable decline for careers in Government ($109,419 vs $76,514) and Law ($190,072 vs $138,062), and almost no change for those in Education and Health. Comparing the selection amounts and shares in Rows (5) and (6) across the two Panels, we see a very large increase in selection for Business careers, smaller increases in Health and

Science careers and actual declines in selection for careers in Government and Law. These changes may reflect our respondents learning more about their prospects in these careers over the 7 years between the two surveys, as well as changes that may have occurred to the relative demands and wages across different occupations.

Table A.10: Relative Importance of Selection in *Ex Ante* Earnings Returns: Phases 1 and 2 (Earnings in 2009 dollars)

	Occupation:					
	Science	Health	Business	Government	Law	Education
Panel A: Phase 1						
(1) $\widehat{SE}_1(j)$ (Selected *ex ante* earnings)	102,699	188,189	176,393	109,419	190,072	72,725
(2) $\overline{\mathcal{Y}_{j1}}$ (Aver. *ex ante* earnings)	96,793	142,538	164,006	100,348	163,223	74,473
(3) $\Delta\widehat{SE}_1(j)$ (Selected earnings difference from Educ.)	29,974	115,463	103,668	36,694	117,346	
(4) $\widehat{ATE}_{j1}$ (Aver. *ex ante* effect)	22,320	68,065	89,533	25,875	88,750	
(5) $\Delta\widehat{SE}_1(j) - \widehat{ATE}_{j1}$ (Selection amount)	7,654	47,399	14,135	10,819	28,596	
(6) $\frac{\Delta\widehat{SE}_1(j)-\widehat{ATE}_{j1}}{\Delta\widehat{SE}_1(j)}$ (Selection share)	25.5%	41.1%	13.6%	29.5%	24.4%	
Panel B: Phase 2						
(1) $\widehat{SE}_2(j)$ (Selected *ex ante* earnings)	137,631	183,852	294,728	76,514	138,062	64,380
(2) $\overline{\mathcal{Y}_{j2}}$ (Aver. *ex ante* earnings)	123,301	125,557	211,147	81,379	137,328	71,333
(3) $\Delta\widehat{SE}_2(j)$ (Selected earnings difference from Educ.)	73,251	119,472	230,347	12,134	73,681	
(4) $\widehat{ATE}_{j2}$ (Aver. *ex ante* effect)	51,968	54,224	139,814	10,046	65,995	
(5) $\Delta\widehat{SE}_2(j) - \widehat{ATE}_{j2}$ (Selection amount)	21,283	65,248	90,534	2,087	7,687	
(6) $\frac{\Delta\widehat{SE}_2(j)-\widehat{ATE}_{j2}}{\Delta\widehat{SE}_2(j)}$ (Selection share)	29.1%	54.6%	39.3%	17.2%	10.4%	

DATA: Data from Phase 1 ($N = 173$) and Phase 2 ($N = 112$).

A.8 Alternative Measures of *Ex Ante* Earnings Losses

We examine in this section the robustness of the findings discussed in Section 5.1 to alternative specifications of the monetary component of the expected utility associated with each occupation.

A.8.1 Measuring Present Values of Future Expected Earnings

In the following, we define and compute the *ex ante* earnings associated with any given occupation j $\left(\mathcal{Y}^s_{ij2}\right)$ as the subjective expectation of the discounted stream of earnings associated with occupation j, where, as is the case in our baseline specification, subjective beliefs are elicited in our Phase 2 follow-up survey. Recall that respondents were asked to give their expected earnings 10 years after graduating from college. Let t_{10} denote this age. Then the present value of expected future earnings over the next T years of a individual's work career is given by:

$$\mathcal{Y}^s_{ij2} := \sum_{\tau=0}^{T} \delta^\tau \mathcal{Y}_{ij2,t_{10}+\tau} \tag{A.11}$$

where $\mathcal{Y}_{ij2,t_{10}+\tau}$ denotes the expected earnings $10+\tau$ years after graduation and δ is the annual discount factor. We set $T = 30$ (assuming individuals retire 40 years after graduation) and $\delta = 0.9$. To compute $\mathcal{Y}_{ij2,t_{10}+\tau}$ beyond t_{10} (that was elicited in Phase 2), we assume that an individual's earnings grow at a rate ρ that we allow to vary by college major (m), occupation (j), and over time (τ). That is:

$$\begin{aligned}
\mathcal{Y}_{ij2,t_{10}+\tau} &= (1+\rho^1_{mj})^\tau \mathcal{Y}_{ij2} \text{ for } 0 < \tau \leq t_1 \\
&= (1+\rho^2_{mj})^\tau \mathcal{Y}_{ij2,t_{10}+t_1} \text{ for } t_1+1 \leq \tau \leq t_2 \\
&= (1+\rho^3_{mj})^\tau \mathcal{Y}_{ij2,t_{10}+t_2} \text{ for } t_2+1 \leq \tau \leq t_3 \\
&= (1+\rho^4_{mj})^\tau \mathcal{Y}_{ij2,t_{10}+t_3} \text{ for } t_3+1 \leq \tau \leq T
\end{aligned}$$

where $(\rho^1_{mj}, \rho^2_{mj}, \rho^3_{mj}, \rho^4_{mj})$ denotes the annual earnings growth rates associated with occupation j and major m, over the intervals $[0, t_1]$, $[t_1+1, t_2]$, $[t_2+1, t_3]$ and $[t_3+1, T]$, respectively. The growth rates are estimated using data from the American Community Survey, over the period 2009 to 2017 and conditional on being male, working at least 48 weeks a year and 40 or more hours per week. Specifically, the growth rates are calculated across five age bins (25 to 29; 30 to 35; 36 to 41; 42 to 47; 48 to 53) by computing the difference in the average log wages for each major-occupation pair across two adjacent bins, and then dividing by the difference in the average age of each of these bins.

A.8.2 *Ex Ante* Losses based on Present Values of Expected Earnings

Using the measures of the occupation-specific present values of expected earnings defined in (A.11), we define the discounted expected lifetime earnings for that occupation in which individual i expected to earn the most as:

$$\mathcal{Y}^{s^{max}}_{i2} := \max\{\mathcal{Y}^s_{i12}, \ldots, \mathcal{Y}^s_{i62}\}. \tag{A.12}$$

which we then compare with the weighted average of discounted expected lifetime earnings, using as weights the elicited probabilities that the individual would work in each of these occupations, i.e.

$$\overline{\mathcal{Y}_{i2}^{s}} := \sum_{j=1}^{6} \mathcal{Y}_{ij2}^{s} p_{ij2} \tag{A.13}$$

We report in Table A.11 features of the distribution of the differences between both quantities, namely $\mathcal{G}_{i2}^{s} := \mathcal{Y}_{i2}^{s^{max}} - \overline{\mathcal{Y}_{i2}^{s}}$. For any given individual i, $\mathcal{G}_{i2}^{s}$ is our estimate of the lifetime *ex ante* earnings loss associated with not choosing the highest paying occupation. Overall, 76.8% of respondents were not certain of choosing the career that maximizes their lifetime expected income.

Table A.11: Distribution of Maximum and Expected Lifetime Earnings ($\times 10^6$ 2009 USD)

	Max Earnings $[\mathcal{Y}_{i2}^{s^{max}}]$ (1)	Expected Earnings $[\overline{\mathcal{Y}_{i2}^{s}}]$ (2)	Difference $[\mathcal{G}_{i2}^{s}]$ (3)
Panel A: Full Sample			
Mean	3.022	2.557	0.466
1^{st} quartile	1.620	1.114	0.040
Median	2.487	1.821	0.282
3^{rd} quartile	3.362	3.061	0.603
Standard Dev.	2.228	2.052	0.653
Panel B: Conditional on $\mathcal{G}_{i2}^{s} > 0$			
Mean	2.707	2.100	0.607
1^{st} quartile	1.475	1.019	0.217
Median	2.177	1.726	0.392
3^{rd} quartile	3.086	2.688	0.750
Standard Dev.	1.986	1.576	0.686

DATA: Sample is 112 respondents to Phase 2 follow-up survey.
NOTE: 86 sample members (76.8%) were not certain of choosing the career that maximizes their expected lifetime earnings.

A.8.3 Accounting for Occupation-Specific Earnings Risk

We now discuss how one can extend the previous framework to account for differences across occupations in earnings risk. Specifically, we assume that individuals are endowed with CRRA preferences, with a risk aversion parameter θ, so that the component of the expected utility associated with the earnings ten years after graduation is given by:

$$\mathcal{U}_{ij2} := \mathcal{E}\left[\frac{Y_{ij}^{1-\theta}}{1-\theta}\middle|\mathcal{I}_{i2}\right]$$

where, as before, Y_{ij} denotes individual i's potential earnings ten years after graduation in occupation j, $\mathcal{I}_{i2}$ denotes individual i's information set at the time of the Phase 2 survey, and $\mathcal{E}\left[\cdot|\mathcal{I}_{i2}\right]$ the subjective expectation operator.

Assuming that individuals' prior beliefs about future earnings are log-normally distributed with parameters m_{ij} and σ_j^2, $\mathcal{U}_{ij2}$ can be rewritten as:

$$\mathcal{U}_{ij2} = \frac{1}{1-\theta}\exp\left((1-\theta)m_{ij} + \frac{(1-\theta)^2\sigma_j^2}{2}\right) \tag{A.14}$$

More generally, assuming that prior beliefs about future earnings $\tau + 10$ years after graduation are log-normally distributed, with parameters $m_{ij\tau}$ and $\sigma_{j\tau}^2$, the expected utility associated with the earnings $\tau + 10$ years after graduation is given by:

$$\mathcal{U}_{ij2\tau} = \frac{1}{1-\theta}\exp\left((1-\theta)m_{ij\tau} + \frac{(1-\theta)^2\sigma_{j\tau}^2}{2}\right) \tag{A.15}$$

We evaluate the expected utilities for each occupation as follows. First, we set the risk aversion parameter, θ, equal to 2 (as in Hai and Heckman, 2017). Second, we compute the occupation-specific variance parameters $\sigma_{j\tau}^2$ from the set of estimates that were obtained by Dillon (2018) using data from the PSID (1976-2011). Specifically, using the notations from Section II.A in Dillon (2018) (pp. 985-986), it follows from the earnings process assumed in that paper that the occupation-specific variance parameters are given by, for $\tau \leq 10$:[57]

$$\sigma_{j\tau}^2 = \sigma_{je}^2\frac{1-\rho_j^{2(3+\tau)}}{1-\rho_j^2} + (3+\tau)\sigma_{ju}^2 + \sigma_{j\varepsilon}^2 \tag{A.16}$$

and, for $\tau > 10$:

$$\sigma_{j\tau}^2 = \sigma_{je}^2\frac{1-\rho_j^{26}}{1-\rho_j^2} + \widetilde{\sigma}_{je}^2\frac{1-\widetilde{\rho}_j^{2(\tau-10)}}{1-\widetilde{\rho}_j^2} + 13\sigma_{ju}^2 + (\tau-10)\widetilde{\sigma}_{ju}^2 + \widetilde{\sigma}_{j\varepsilon}^2 \tag{A.17}$$

where the stochastic earnings components $(\sigma_{je}^2, \sigma_{ju}^2, \sigma_{j\varepsilon}^2, \rho_j)$ and $(\widetilde{\sigma}_{je}^2, \widetilde{\sigma}_{ju}^2, \widetilde{\sigma}_{j\varepsilon}^2, \widetilde{\rho}_j)$ are set equal to the estimates obtained by Dillon (2018) for workers below and above the age of 40, respectively (see Tables A8 and A9 pp.17-18 in the Online Appendix of Dillon, 2018).[58]

Third, using the earnings growth rates $(\rho_{mj}^1, \rho_{mj}^2, \rho_{mj}^3, \rho_{mj}^4)$ introduced in Subsection A.8.1, we assume

[57] This expression is derived under the simplifying assumption that beliefs in the Phase 2 survey are all elicited seven years after graduation.

[58] In practice, we use the following mapping between our occupations and the occupations used in Dillon (2018): Science includes Sciences and Engineering; Business includes Financial, Management, Sales, and Entertainment; Government includes Community. We use the same classification as in Dillon (2018) for Health, Education and Law (Legal occupations).

that the deterministic component $m_{ij\tau}$ is given by:

$$\begin{aligned} m_{ij\tau} &= m_{ij} + \tau \times \log(1+\rho^1_{mj}) \text{ for } \tau \leq t_1 \\ &= m_{ijt_1} + (\tau - t_1) \times \log(1+\rho^2_{mj}) \text{ for } t_1+1 \leq \tau \leq t_2 \\ &= m_{ijt_2} + (\tau - t_2) \times \log(1+\rho^3_{mj}) \text{ for } t_2+1 \leq \tau \leq t_3 \\ &= m_{ijt_3} + (\tau - t_3) \times \log(1+\rho^4_{mj}) \text{ for } t_3+1 \leq \tau \leq T \end{aligned}$$

where, given the parameter values for σ^2_{j0} computed using the expression in (A.16), we estimate the parameters m_{ij} from the elicited earnings beliefs using the expression:

$$\mathcal{Y}_{ij2} = \exp\left(m_{ij} + \frac{\sigma^2_{j0}}{2}\right) \tag{A.18}$$

Finally, we proceed as in our baseline set-up, replacing the expected earnings, $\mathcal{Y}_{ij2}$, with expected lifetime utilities, which are given by:

$$\mathcal{V}_{ij2} = \frac{1}{1-\theta}\sum_{\tau=0}^{T}\delta^\tau \exp\left((1-\theta)m_{ij\tau} + \frac{(1-\theta)^2\sigma^2_{j\tau}}{2}\right) \tag{A.19}$$

where, as before, we set the discount parameter δ equal to 0.9.

The maximum occupation-specific discounted expected lifetime utility associated with future earnings is then given by:

$$\mathcal{V}^{max}_{i2} := \max\{\mathcal{V}_{i12}, \ldots, \mathcal{V}_{i62}\}. \tag{A.20}$$

which we compare with the weighted average of discounted expected lifetime utilities, using as weights the elicited probabilities that the individual would work in each of these occupations, i.e.

$$\overline{\mathcal{V}_{i2}} := \sum_{j=1}^{6}\mathcal{V}_{ij2}p_{ij2} \tag{A.21}$$

The resulting estimates, presented in Table A.12, document features of the distribution of the *ex ante* utility losses associated with not choosing the occupation that maximizes the pecuniary component of utility (denoted by $\mathcal{G}^{\mathcal{V}}_{i2} := \mathcal{V}^{max}_{i2} - \overline{\mathcal{V}_{i2}}$).

One can alternatively express the expected utility losses in terms of the equivalent permanent increase in mean prior (log) beliefs about earnings. Specifically, consider a permanent increase $m_{2\tau} - m_{1\tau} = \Delta_m$ in mean log beliefs, such that the following equality holds:

$$\mathcal{G}^{\mathcal{V}} = \frac{1}{1-\theta}\sum_{\tau=0}^{T}\delta^\tau\left(\exp\left((1-\theta)m_{2\tau} + \frac{(1-\theta)^2\sigma^2}{2}\right) - \exp\left((1-\theta)m_{1\tau} + \frac{(1-\theta)^2\sigma^2}{2}\right)\right)$$

which, assuming a constant earnings growth rate ρ, yields the following closed-form expression for

Table A.12: Distribution of Maximum and Expected Lifetime Utilities ($\times 10^{-3}$)

	Max Utilities $[\mathcal{V}_{i2}^{max}]$ (1)	Expected Utilities $[\overline{\mathcal{V}_{i2}}]$ (2)	Difference $[\mathcal{G}_{i2}^{\mathcal{V}}]$ (3)
Panel A: Full Sample			
Mean	-0.067	-0.096	0.029
1^{st} quartile	-0.089	-0.131	0.001
Median	-0.060	-0.079	0.015
3^{rd} quartile	-0.043	-0.049	0.039
Standard Dev.	0.036	0.064	0.042
Panel B: Conditional on $\mathcal{G}_{i2}^{\mathcal{V}} > 0$			
Mean	-0.072	-0.109	0.037
1^{st} quartile	-0.094	-0.144	0.009
Median	-0.066	-0.094	0.023
3^{rd} quartile	-0.047	-0.059	0.048
Standard Dev.	0.036	0.066	0.044

DATA: Sample is 112 respondents to Phase 2 follow-up survey.
NOTE: 86 sample members (76.8%) were not certain of choosing the career that maximizes their expected lifetime utility associated with future earnings.

Δ_m (where m_0 denotes the initial mean prior beliefs):

$$\Delta_m = \frac{1}{1-\theta} \log \left(\frac{(1-\theta)\mathcal{G}^{\mathcal{V}}}{\exp\left((1-\theta)m_0 + (1-\theta)^2\sigma^2/2\right) \frac{1-(\delta(1+\rho))^{T+1})}{1-\delta(1+\rho)}} + 1 \right)$$

Evaluating this expression for $\theta = 2$, $\mathcal{G}^{\mathcal{V}} = 0.029 \times 10^{-3}$ (mean absolute *ex ante* utility loss), $T = 30$, $\rho = 0.04$, $m_0 = 11$ and $\sigma^2 = .34$ yields $\Delta_m = 0.114$ (11.4 log points increase).[59]

[59]The latter two parameters (m_0 and σ^2) are set to match the mean values across occupations and majors of the mean and variance of prior earnings (log)-beliefs.

A.9 Occupational Choice Model: Additional Estimation Results

Table A.13: Estimates of returns to (log of) expected earnings in occupational choice (excluding seniors)

Earnings Beliefs Used:	Phase 1 ($\mathcal{Y}_{ij1}$)			
Occup. Measured by:	Actual (d_{ij})	Phase 1 Probs. (p_{ij1})		
	(1)[1]	(2)	(3)	(4)[2]
Log Earnings	1.519	1.337	0.688	1.014
	(0.346)	(0.310)	(0.333)	(0.177)
Controls				
Occupation Only	Y	Y	N	N
Major × Occupation	N	N	Y	Y
Individual × Occupation	N	N	N	Y

DATA: The *Excluding Seniors* sample contains 113 individuals.

NOTES:

[1] Estimates in column (1) are produced with a conditional logit model, using data on chosen occupations, d_{ij}. Estimates in all other columns use elicited data on occupational choice probabilities and are produced with a least absolute deviation (LAD) estimator.

[2] Column (4) uses Phase 1 data on respondents' elicitations of expected earnings and occupational choice probabilities for each possible major-occupation pair, providing 6 times the number of observations in the sample. All other columns use elicited data only for the chosen college major of sample members.

[*] Standard errors in parentheses. The standard errors for estimates in column (4) are clustered at the individual × occupation level.