Subjective Expectations and the Gender Gap in STEM Majors: Evidence from a Sample of Swedish Students

Nikolay Angelov[1], Per Johansson[2], Mikael Lindahl[3] and Ariel Pihl[4]

Key words: Fields of study, counterfactual choices, subjective expectations, incentives
JEL: J24, I23

Abstract

How important are economic incentives and non-pecuniary expectations in students' choice of field of university study? How does their relative importance vary across genders? We have collected high quality survey data on expectations for counterfactual education choices for a broad sample of Swedish students to shed light on these questions. We use a rank-ordered logistic model to explain students' rankings of the potential choices, and find that expectations about both pecuniary and non-pecuniary outcomes of an educational choice strongly predict which field students chose for both genders, but with non-pecuniary expectations being relatively more important for female students. The existing STEM gender gap can be fully explained by our elicited subjective expectations variables. Most important are the non-pecuniary factors, especially the two taste variables capturing enjoyment of course work and of expected occupation. We find similar results when we look at the gender gaps across all fields of study.

1. INTRODUCTION

The choice of field of study has long lasting consequences for an individual in terms of future earnings (Kirkeboen, Leuven and Mogstad, 2016; Hastings, Neilson and Zimmerman, 2013), as well as for the society as a whole in terms of economic growth (Murphy, Shleifer and Vishny, 1993). Field of study is also important as a source of inequality between groups, such as the gender wage differences due to the underrepresentation of women in certain fields. Understanding why individuals choose to acquire different types of education is therefore of paramount importance if we want to design policies effective in reducing these inequities.

An issue of particular concern is the under-representation of women in some well-paid fields like Science, Technology, Engineering and Mathematics (STEM); both in university studies and in the labor market. This contributes to the gender pay gap as well as the potentially below-optimal participation of women in STEM fields: perhaps resulting in a "female STEM talent reserve" of high ability women not being utilized. Although Sweden is seen as one of the most gender equalized countries in the world, it is very close to the OECD average when it comes to fraction of women in STEM fields (Figure 1), and

1 Swedish National Audit Office
2 Uppsala University, IFAU, IZA and UCLS
3 University of Gothenburg, CESifo, IFAU, IZA and UCLS
4 University of Gothenburg. Corresponding Author: ariel.pihl@gu.se

the low share of women has been similarly steady over time, with only a slight increase (Figure 2). The gender earnings gap in Sweden also appears to be partly attributable to fields of study, with around 10% due to STEM fields and 22% due to general fields of study choices (see Appendix Table A1). These estimates are similar to those for the US and Canada (see Card and Payne, 2017).

In this paper we investigate gender differences in fields of study by utilizing data on subjective expectations to understand how beliefs and expectations about future pecuniary and non-pecuniary outcomes determine choice of fields of study at universities. More specifically, we attempt to provide answers to questions such as: How do expectations about future pecuniary and non-pecuniary outcomes differ between those preferring STEM majors versus those preferring non-STEM majors and how does their relative importance vary between men and women? How much of the STEM gender gap can be explained by our subjective expectations measures and which of these measures are most important?

In this study we follow a small but growing literature that has analyzed the determinants of educational choice by using survey data on students' subjective expectations associated with both their preferred and non-preferred choices (see e.g. Zafar, 2011; Arcidiacono et al., 2012; Stinebrickner and Stinebrickner, 2012; Attanasio and Kaufmann, 2014; Wiswall and Zafar, 2016; Reuben et al., 2015 and the pioneering study by Dominitz and Manski, 1996). The advantages of such a survey design is that it provides information about the students' actual expected outcomes rather than assumed rational expectations. It is also an opportunity to elicit preferences about a number of outcomes typically not available in data sets, such as perceived ability and parents' approval in choice of education. This earlier literature, to which Altonji et al. (2015) provides an excellent introduction, has begun answering some of the questions posited above. Zafar (2013), in research that is most closely related to our study, finds that the gender gap in fields of study is mainly due to gender differences in tastes and preferences, but not due to perceived ability or expected earnings. However, the studied samples are often small and unrepresentative (frequently currently enrolled college students from a single university) so imprecise estimates and low generalizability are common limitations.

For the purpose of this project we have collected survey data on a sample of 505 high school students in Stockholm. The students were asked to provide their expectations about various hypothetical educational choices (in eight fields of study, and also if they preferred not to go to college). To study the relative importance of factors such as status, perceived ability, consumption value/enjoyment of study and work, possibility of combining family with work, and future income streams in choice of higher education, the students were also asked to rank the fields of study based on how likely they were to choose them. We have matched the survey data to administrative data from Swedish registers on high school grades, gender, if immigrated, and parents' education and income.

An important advantage of our study compared to earlier research which has elicited subjective expectations in education, is in regard to the high quality of our data. This is evidenced by: a) the relatively large sample, which makes it possible to analyze all possible determinants jointly as well as to estimate the models separately for various groups; b) the wide range of determinants of educational choice available; c) a better survey design than most previous studies: mostly administered through home visits and a more representative sample of students, and; d) that we have merged the survey data with data from administrative registers.

A suitable estimation framework should utilize all information that we have available on expected outcomes for all field-of-study choices, as well as the individual's rank of each field of study. Since we have counterfactual outcomes for each field of study for all individuals, our data has a panel structure. At the same time linear models are not, in general, well suited for this data. Our preferred estimates use rank-ordered logistic regressions (as in Zafar, 2013, and Arcidiacono et al., 2012). This incorporates all the information contained in the rank of possible majors and the associated expectations for each. However, we also show that our conclusions are robust to using conditional logit regressions, comparing the top ranked major to the other options, and to the use of linear probability models. The underlying decision to choose a college major can be modelled as in Zafar (2013): an individual student maximizes their expected utility over the set of educational paths available. Their expected utility is based on their beliefs about what will happen in college for the given field and what it will mean for their career path and lives after college. Namely, expected utility is a function of the students' expectations of various future outcomes both during and after college.

We provide several interesting findings. First, the pecuniary and non-pecuniary expectations are important for fields of study choice for both genders, but their relative importance differ among men and women with non-pecuniary expectations being relatively more important for females. This result is similar to Zafar (2013). However, the effects of our elicited survey expectations measures are estimated precisely enough to always generate estimates that are statistically significant, regardless of whether this is done for the sample of males or females. This includes perceived ability and expected earnings, similarly to the result in Arcidiacono et al. (2012) for men. However, expected enjoyment of coursework and future employment, explain more of the choice than the pecuniary variables do. Second, the existing STEM gender gap can be fully explained by the elicited subjective expectations variables. Most important are the non-pecuniary factors, especially the two taste variables, enjoying coursework and job. We find similar results when we look at the fields of study gender gap across all fields of study. More than 80% explained by the elicited subjective expectations. Humanities and arts is the only field of study where these variables do not explain a major part of the observed gender gap in preferred field choice. We also note that other factors, like earlier math test scores and family background, do not explain the gender gap.

2. BACKGROUND AND DESCRIPTIVE STATISTICS

2.1 Related Literature

As summarized by Altonji et al. (2015), without data on subjective expectations, models of choice must make strict and often inaccurate assumptions about individuals' beliefs. For example, assuming that expectations of the returns to education are unbiased. Several foundational papers have illustrated how survey data on students' subjective expectations can be used to estimate more informative choice models (Zafar, 2011; Arcidiacono et al., 2012; Stinebrickner and Stinebrickner, 2012). In particular, Zafar (2011) shows that students do not appear to exhibit cognitive dissonance when reporting their beliefs.

Building upon these foundations, economists have used subjective expectations data to examine how students chose college majors and occupations (Arcidiacono et al., 2012; Zafar 2013, Stinebrickner and Stinebrickner 2014; Wiswall and Zafar 2015; Hastings et al. 2015; Reuben Wiswall and Zafar, 2017; Wiswall and Zafar 2018). These papers broaden the possible types of expectations which may affect major or occupation choice. A central focus is whether expected earnings play an important role in student's choices, as simple economic models would typically predict. While a subset of papers find that expected earnings are an important predictor of choice (e.g. Arcidiacono et al. (2012); Arcidiacono et al. (2017)), some do not (e.g. Zafar, 2013). Wiswall & Zafar (2015) show that estimates of the role of expected earnings are too high when the correlation between expected earnings and tastes are not accounted for.

Differences both by gender and science/STEM majors have also been a common focus in this literature. Zafar (2013) and Wiswall and Zafar (2018) finds that women value different features of majors and occupations, which contribute to their lower rates of pursuing STEM and business fields. Almås et al., (2016) and Reuben, Wiswall and Zafar (2015) incorporate lab measures of overconfidence and risk preferences and show that the differences between genders in these measures drive some of their educational beliefs and behaviors. Kugler et al. (2017) find that women are more likely to change majors in response to a bad grade in STEM then men are. Rapoport and Thibout (2018) use confidence measures to simulate expectations and find that boys put more weight on their previous math test scores in choosing high school track than girls do. Wiswall and Zafar (2016) add marriage markets in to their analysis and find that women believe there is a marriage penalty to STEM and business degrees.

2.2 Institutional details

All children in Sweden should attend nine years of compulsory school. During their 9th year (aged 15-16), pupils apply for a high school program within their municipality of residence. All students that have passing grades in a sufficient number of subjects at the end of compulsory school are qualified to attend a high school program. In 2011, about 91 percent of girls and 89 percent of boys leaving compulsory school in Stockholm were qualified to attend a high school program (in Sweden as a whole the figures are slightly lower). A student applies for the combination of a program and a high school (public or independent) and provides a preference rank of such choices. If there is excess demand, selection is made based on GPA at the end compulsory school.

There are a total of 18 national high school programs: 6 academic (university preparation) programs and 12 occupational-oriented programs, where the academic content varies depending on the program and at the discretion of the student's choice. Exams from an academic program always provide basic qualification to attend university education, whereas in the occupational oriented programs there is always a possibility to obtain basic qualifications to attend university, as long as the student chooses enough academic courses. Many university fields of study also have special requirements regarding high school academic qualifications. For instance, if a student chooses any of the natural science-oriented university programs, a high school exam from the natural science program is sufficient, whereas a high school exam from a social science high school program is only sufficient if the student actively chooses enough natural science courses. In addition to the 18 national high school programs, there are also some specialized programs that only exist in some locations, including some academic programs like the International Baccalaureate program. High school programs in Sweden are three years long, regardless of type of program.

During the last year of high school, individuals can choose to continue to university programs, which are typically between 2 and 4 years of duration. If students are not qualified for their university program of choice at the end of high school (by coursework or grades), they can top up their education with an additional year of high school. University programs in Sweden can be divided into the following eight broad categories:

- Healthcare and social care (Medical training; Social work and guidance)
- Humanities and Art (Media production; History and archeology)
- Services (Tourism and travel; Police training)
- Pedagogy and Teacher education (Subject teacher training; Pedagogy and didactics)
- Social science, Law, Business, etc. (Psychology; Business administration)
- Agricultural and Forestry and animal health care (Veterinary care; Agriculture and forestry)
- Natural science, Mathematics and Data (Computer science; Mathematics and statistics)

- Technology and Manufacturing (Civic and building engineering; Technical industry)

Within these eight categories there are more specific university tracks, some of the most common ones which are listed in parentheses. In the analysis below, we use a division of STEM versus non-STEM fields of study. We define “Natural science, Mathematics and Data” and “Technology and Manufacturing” as STEM fields of studies and the rest as non-STEM fields of studies.

2.3 Survey design, Data collection, Variables and Sample restrictions

For the purpose of this project we have collected survey data on a sample of high school students in the municipality of Stockholm. To be part of our population, the students must have attended the third year of a municipality high school in 2014 and lived in the municipality of Stockholm. Although the fraction of independent high school students is high in Stockholm, the majority of the students in academic programs attend municipality schools. The municipality of Stockholm includes many suburbs, some well-off and some much less so.

In order to elicit reliable subjective expectations and beliefs on counterfactual outcomes, we wanted to have an expert interviewer present during the interview, something that is quite costly. At the same time, it was important to get a large enough sample that was also as representative as possible of the student population. This is especially important since the subjective expectations data sets used in most previous papers are based on small and non-representative samples. We hired a survey company (SKOP) to carry out the sampling and survey for us.

A prerequisite for SKOP to even attempt to book time for an interview with a student was that they could be contacted by phone. The municipality of Stockholm provided us with a list of the 3368 3rd year students enrolled in their public schools. Of this original population, SKOP managed to identify correct (either mobile or stationary) phone numbers to 1682 persons.[5] All these students were sent an introductory letter with information about the survey. Of these, 68 persons were dropped because they were found to have died, moved away etc., leaving 1614 individuals. Of these, SKOP was unable to contact 789 individuals by phone (perhaps because of wrong phone numbers; a refusal to answer unknown phone numbers; etc.)

Of the 825 students that SKOP managed to establish some contact with, 505 students choose to participate in the survey; 258 persons choose to not participate in the survey and 62 persons dropped out

[5] A gross file of the names and addresses of the whole student population was provided by the municipality of Stockholm authorities, and sent to SKOP. The figure 56% is for the 1682 of the original 3368 individuals.

during the process. Hence, we consider this study to have a response rate of 61% (505/825).[6] Students being difficult to contact and choosing not to participate in the survey could give us a sample of students which is not representative of the starting population. The original list provided gymnasium program, and a comparison of the frequency of each program for the 3368 originals and 479 in our sample is reassuringly similar, although there appears to be overrepresentation of science students. The comparison in provided in the Appendix as Table A2.

The questionnaires were answered using CAPI (computer-assisted personal interviewing), i.e., in person. Hence, SKOP mostly administered the survey through home visits. It was conducted during the time period February-April of 2014, which was chosen to be before the students had made their actual choice (applied or accepted college programs), but late enough that they had likely put considerable thought into their college path.[7]

Our study differs from most other studies of subjective expectations data, in that it surveys students at many schools and before they apply to college instead of at a single university. Hence, we have a more representative sample of students. Since the sampling is done at the high school level, we allow for students to end up at any university of their choice, including technical universities and business schools.

The students were asked to provide their expectations and beliefs about various hypothetical educational choices in the eight fields of study (listed above). The students were also asked to rank the alternatives based on how likely they were to choose them, which also included an option of no additional education. This is the choice set of the participants in the estimations below.[8]

For each of these eight fields of study, we elicited students' beliefs and expectations regarding:

- Probability of passing a degree
- Probability enjoying course work
- Study time (hours per week) required
- Family approval
- Probability of getting a job
- Enjoying your job (age 30)
- Combining work and family life

[6] This might be called an upper response rate bound since some of the 789 may have received the letter and chosen not to respond. We might regard 31% (505/1614) to be a lower response rate bound.
[7] Hence, we avoid the issue of cognitive dissonance/ex-post rationalization (discussed in Zafar, 2011) when students provide subjective expectations and a field of study selection is already made.
[8] In Appendix Table 1 we show details of the fields of study categories, as the students also were asked to specify a more specific choice within each of these eight categories.

- Work hours per week required
- Monthly earnings at age 30 and at age 40
- Status (not related to pay)

Hence, they are asked about labor market outcomes (earnings, hours of work and job probabilities) as well as about their expectations for how educational choice will affect their experience in college (study hours, probability of passing, enjoying coursework, parental approval), as well as after college (enjoying their work, balancing family life, status).

For all questions, except for the earnings and hours questions, students are asked to respond with a number between 0 and 100. So, for instance, for each hypothetical choice we asked “How high is the probability that your parents and other family members would approve of your choice of major?” The average response to this question for males was 72.3, meaning that on average they expected that there was a 72.3% chance that their parents would approve the choice. Note that we carefully explained what is meant by a probability. A translated version of the general instructions and the specific questions in the survey is included section 1 of the Data Appendix**.**

To the survey data we then matched administrative data from Swedish registers on high school track, grades and test scores, gender, if immigrated and the parents’ education and income.

Since this paper focuses on university field of study choice, we drop the 29 individuals that rank not continuing on to university as their top ranked choice. We further ignore the rank of not choosing a university program among the remaining individuals. For this option, stopping at high school education, we do not have the counterfactual responses anyway. Hence, our final sample consists of 476 individuals: 241 males and 235 females.

2.4 Descriptive Statistics

Background characteristics of the sample are found in Table 1. Roughly half of the sample is female, and 18% are either immigrants or have two parents who are foreign born. Among their parents, roughly half of both mothers and fathers have completed a college degree. The median of the sum of both parents’ gross employment and business income (including parental leave and unemployment benefits) is 866,000 SEK (around €88,000). Median earnings seem high, but this is because this is a broad measure of income and because most households in Sweden are two-income households, where the women often work full time. We combine the education and income measures of the parents into one socioeconomic status (SES) measure using Principal Component Analysis (see the data appendix for details). This SES measure has mean zero and standard deviation one by construction. We also use two measures of student achievement in high school: English test score and Math test score, both based on averages of several tests (and then standardized in the whole sample). We only use courses completed in the first or second year of high school, i.e. *before* the survey was conducted. The sources and creation of all variables are described in the data appendix.

43% of students in the sample were attending STEM oriented high school tracks; 47% were attending academic non-STEM specialized high school tracks; and 10% of students were attending one of the 12 occupational programs. In our main analysis we have kept all HS students originally surveyed, regardless of HS track. The reason is that, as explained above, all HS programs in Sweden can be individually designed so as to qualify for university programs, including STEM programs, and that later complementing high school education with courses from a post high school education program (at KOMVUX) is quite common in Sweden.[9] In the main estimations we include an (interacted) indicator for whether the individual attended a STEM high school program (or not). When we estimate models where we limit the sample to those attending academic high school programs or STEM-oriented high school programs, we get very similar results. This is also not surprising since our estimations look at within individual choices for individuals that rank all university fields of study available. Also, regardless of whether they attend a STEM high school program or not, they still rank both STEM and non-STEM university programs highly. This can be seen in Figure 3 where we show how individuals rank STEM fields of study at the university (1 is highest rank and 8 is lowest rank).[10] Of those students who rank a STEM field as the most preferable, about 27 percent attended a non-STEM high school track.

As we explained in previous subsections, the sample we use might or might not be representative of the population. In the next version of this paper, column 4 of Table 1 will report means for some of the variables using population data of municipality high- school students living in Stockholm.[11]

Next, we show some results regarding gender differences in preference for fields of study at the university. We have already seen large differences in the population regarding graduates in various fields and that gender shares have been fairly stable over time and across countries. In Figure 4 we show gender differences in the ranking of STEM fields of study in our sample. It shows that there is clear STEM gender gap in preference rank, with men ranking STEM fields much higher than women.

[9] We will later try to provide statistics on how many people from non-natural science program actually attended social science university programs, and vice versa, in the population.

[10] Note that the sample sizes in the filled and dashed bars differ, since there are more students in non-STEM focused high school programs.

[11] This is in progress, but not currently available.

Next, we summarize the students' expectations from the survey in Table 2. We show summary statistics for all fields, and divided by STEM and non-STEM fields. We also show gender differences in these means. All means are calculated over the set of eight fields of study options.

Over the set of educational options, the surveyed men estimate that they will earn an average of 37,150 SEK per month at age 40. Women expect lower earnings: 35,620 SEK. These are not unreasonable figures if they are compared with average earnings in Sweden. Expected hours of work per week are high: 45.8 for men and 47.1 for women. The mode is 40 hours (28% of the respondents), but over 40% of the respondents provide figures between 41 and 60 hours. They are also asked about their expectations for how educational choice will affect their experience in college (study hours, probability of passing, enjoying coursework, parental approval), as well as after college (finding a job, enjoying their work, balancing family life).

Note that the numbers look mostly reasonable, although the expected hours of work per week is high on average, especially for women. However, it should be noted that these are averages over all options, including choices that individuals rank low and do not expect to experience. Overall, the stated expectations are similar for men and women. The differences that are statistically significant (column 3) are the predicted earnings, which are lower for women, and the predicted study hours, which are higher for females.[12]

Next, we look at the means shown separately for STEM and non-STEM university fields of study. We saw above that expected earnings are higher for males than females, but that the means in general were similar for males and females. But it is possible that a gender pattern across fields of study is hidden in these numbers. If we compare the gender differences in columns 6 (STEM) and 9 (non-STEM) there are some interesting patterns present. For instance, expected earnings is always higher for males, regardless of if it follows a STEM or non-STEM education. Whether this is because males expect to have an absolute advantage in labor market productivity over females or because of other factors such as expected discrimination in pay in all fields, we do not know.

We also see that for some expectations, the pattern in STEM fields reverses in non-STEM fields. Men expect to find a job more easily, achieve a degree at higher rates and enjoying course work more than women in STEM fields, whereas the reverse gender pattern is found for non-STEM fields.

[12] Note that these gender differences might to some degree be driven by differences in general preference for continue to college as men rank staying at HS higher as a choice (half a rank) than women.

We next look at whether students' expectations and beliefs differ depending on the rank of the fields of study. In Table 3, we compare expectations for individuals' first ranked education compared to the other seven fields. There are highly significant differences in expectations for all variable and both genders. All of the differences are in the (normatively) positive direction, except for expected hours of work and expected study hours.[13] This is early evidence that students chose education fields for which they have high expected future outcomes. Arcidiacono et al. (2012) find that students expect to earn higher wages in their major relative to most majors they didn't chose. They argue that this is evidence of comparative advantage: students selecting fields for which they are relatively best suited. We also find evidence of this for our sample in Table 3. The higher expected wages and probability of graduating that we see for chosen field is also consistent with students incorporating comparative advantage into their educational choice. They chose educational fields for which they expect themselves to have success in college and in the labor market.

Later in this paper we estimate unconditional models as well as models where we include all expectations and beliefs simultaneously. Since these measures are correlated, estimating the importance of one measure (e.g. expected earnings) for the rank order choice should be interpreted as the effect of earnings expectations and those other beliefs and expectations which are correlated with expected earnings. We therefore report correlation matrices for the full set of variables in Appendix Table A3. As correlations are of similar size for men and women, we only show results for the full sample. The upper panel provides raw correlation coefficients and the bottom panel correlations, where we first have regressed out individual fixed effects. The largest correlation for each variable (going row-by-row) is in bold. We find fairly similar results in both panels. For instance, status is most correlated with parental approval, but also highly correlated with wage, and the two enjoyment variables (for coursework and job) are highly correlated. However, notably, most correlations are around 0.2-0.3 (and never above 0.7). Hence, there is independent variation in all these variables.

3. ECONOMETRIC MODEL AND ESTIMATION STRATEGY

An ideal estimation framework should utilize all information that we have available on expected outcomes for all field-of-study choices, as well as the individual's rank of each field of study. Since we have counterfactual outcomes for each field of study for all individuals, our data has a panel structure. Since rank is ordinal, linear models are not well suited to capture this aspect of the data. Our preferred

[13] Expecting to work/study more in one's preferred major may be due to the student incorporating the effect of the other variables, or that more time-consuming fields are correlated with other things that they value. An example of the first would be if students expect to study more in their chosen fields because they enjoy the coursework. An example of the second is if higher earnings fields have higher hours requirements.

estimates use rank-ordered logistic regressions. This incorporates all the information contained in the rank of possible majors and the associated expectations for each. The rank-ordered logit model is a generalization of the conditional logit regression model (McFadden, 1974; Beggs et al., 1981; Hausman and Ruud, 1987). Allison and Christakis, 1994 provide an accessible presentation of the features of this model (which we follow below).

The underlying decision to choose a college major can be modelled as in Zafar (2013): an individual student maximizes their expected utility over the set of educational paths available. Their expected utility is based on their beliefs about what will happen in college for the given field, and what it will mean for their career path and lives after college. Namely, expected utility is a function of the students' expectations of various future outcomes both during college and after completed college.

More specifically, we specify that each individual i=1, …., N rank field of study j=1, …, J from most to least preferred choice (the survey did not allow ties). It is assumed that the rank order represents the ordering of their expected utility $\boldsymbol{U_{ij}}$ for each of the alternatives:

$$\boldsymbol{U_{ij}} = \boldsymbol{\mu_{ij}} + \boldsymbol{\varepsilon_{ij}}, \tag{1}$$

where $\boldsymbol{\varepsilon_{ij}}$ is i.i.d. with an extreme value (or Gumbel) distribution, meaning that the difference between $\boldsymbol{\varepsilon_{ij'}}$ and $\boldsymbol{\varepsilon_{ij''}}$, where $\boldsymbol{j'}$ and $\boldsymbol{j''}$ are two alternatives, has a logistic distribution.

We decompose $\boldsymbol{\mu_{ij}}$ into linear parts as:

$$\boldsymbol{\mu_{ij}} = \boldsymbol{\beta_j x_i} + \boldsymbol{\gamma_j} + \boldsymbol{\theta w_{ij}} \tag{2}$$

Where $\boldsymbol{x_i}$ represent variables that vary between individuals, $\boldsymbol{\gamma_j}$ are 7 potential field of study indicators, and $\boldsymbol{w_{ij}}$ represent variables that vary between individuals and alternatives. In the present study $\boldsymbol{x_i}$ is an indicator for student being a female, and $\boldsymbol{w_{ij}}$ are individuals' stated preferences/expectations about future outcomes associated with varying potential field of study choices, such as expected earnings and probability enjoying the job associated with each study choice.

Three things are worth noting with equations (1) and (2). First, the rank ordered logistic regression model conditions on the choice set for the individual.[14] This means that we condition on unobserved individual characteristics, and that main effects of individual specific variables are not identified. In fact,

[14] This is analogous to OLS with individual fixed effects. In fact, if we treat rank as a linear variable and estimate a linear model with individual fixed effects, we get remarkably similar results.

since all individuals rank all fields of study choices, there is no variation in the outcome variable (the rank order) if only the between-individual variation is utilized in the data. Hence, coefficient estimates stemming entirely from between-individual variation in the data cannot be estimated. However, interaction effects of individual specific characteristics with fields of study indicators are identified. Suppose that we group the fields of study into STEM and non-STEM fields. This means that the coefficient for an interaction between *STEM* and *female* captures the female-male gender gap in preferring STEM versus non-STEM fields of study.

Second, assuming $\boldsymbol{\varepsilon_{ij}}$ is i.i.d. rules out violations of the Independence of Irrelevant Alternatives (IIA) assumption. In our setting means assuming that introducing or eliminating a field of study should not alter the preference rank for the remaining field of studies. However, a benefit with rank ordered data is that it can be tested by dividing the ranked fields of study into groups.[15] Third, the underlying behavioral model assumes that individuals first choose a preferred field of study among the choice set of *J*=8 fields of study, then choose a preferred field of study among the remaining choice set of *J-1*=7 fields of study, and so forth until all choices are made. In reality one might worry that only the first few choices are well thought out by an individual. In order to gauge the sensitivity to this issue we also estimate models with fewer ranks utilized, including the extreme case where the first ranked alternative is contrasted with all the remaining alternatives.

The estimation is conducted by maximum likelihood using the *rologit* command in STATA. We group on the individual, meaning that we only utilize within-individual between-fields of study variation.

4. RESULTS

4.1 The importance of expectations and beliefs for preferred fields of study

In Table 4 we show estimates from estimating a rank ordered logistic model. To facilitate comparison between the size of the coefficients across variables, we have standardized all subjective expectations/beliefs variables to have mean zero and standard deviation one in the full sample. We do not include the field of study fixed effects, nor any of the individual characteristics interacted with fields of study (so in equation 2 we set $\gamma_j = 0$ and $\beta_j = 0$ for all j's). The first three columns show results for models including a single subjective expectations variable (w_{ij}), for instance log predicted wage at age 40 in the first row. The last three columns report results from models which include all 10 subjective expectations variables simultaneously. We show results from both models since (as we saw in Appendix Table A3) the variables are more or less correlated. The two sets of results allow a better understanding

[15] To be added to the next version of the paper.

of the importance of each separate variable. Columns 3 and 6 show differences in the estimates and associated standard errors) between men and women.

In estimation of the separate models we see that all variables are statistically significantly associated with rank of field of study for both genders. This strongly suggests that all these measures manage to represent reliable information of importance for the choice. This likely depends on both the relatively large sample and the careful survey design we employed. The variables of most importance for the study ranks are job and coursework enjoyment. We refer to these two variables as "taste" variables since they should capture a general taste for work or studies in the field.[16] The effect sizes are similar across gender. The variable of least important is how easy it is to balance work and family.

In column 3, we report estimates and standard errors for the gender differences in the importance of the subjective expectations. We see that the expected wage, probability of finding a job and the number of study hours are of more importance for the study choice for men than for women. However, all these variables, including expected wage, are sizable for both genders.

In the combined models, coefficient estimates decrease. This is unsurprising given the correlation of these variables within individuals. A striking example of this is the positive and significant coefficient on hours of work in columns 1 and 2. This coefficient suggests that individuals rank fields that require more hours per week of work, higher. This counter-intuitive result is because there is a strong correlation between expected hours of work and things that are more generally positive like earnings and social status. When we control for all expectations simultaneously in columns 4-6, there is no longer a relationship between hours and rank. The taste variables are still of most importance. The differences show that probability of finding a job and the number of study hours are of more importance for the study choice for men than for women, even conditional on all the variables. However, expected wage is no longer significantly different. Hence, the gender difference in expected wage decreases conditional on the other expectations and beliefs, which is mostly due to conditioning on the probability of finding a job, expected study hours and status, which are the variables that are statistically significant across genders and correlated with expected earnings within individuals (see Appendix Table A3).

Next, similar in spirit to Zafar (2013), we attempt to decompose the relative contributions of the different expectations and beliefs. We follow Zafar and divide the variables into pecuniary (wage, work hours, probability of finding a job, probability of graduating and hours of study required) and non-pecuniary

[16] However, it is important to differentiate these "taste variables" from the coefficient estimates on the expectations variables. The latter can be thought of as preference parameters, i.e. how heavily the variable is weighted in the utility function and therefore the choice of field.

(social status, enjoying work, combining work and family, family approval and enjoying coursework) variables.[17]

We calculate the contribution of the pecuniary variables to the choice as:

$$\boldsymbol{M_{pec}} = \sqrt{\left(\frac{1}{8}\sum_j (\widehat{y_{J,full}} - \widehat{y_{J,non\ pec}})^2\right)} \tag{3}$$

where $\widehat{y_{J,full}}$ is the predicted rank for each field (averaged over individuals) using all the expectations variables and $\widehat{y_{J,non\ pec}}$) is the predicted rank using only the non-pecuniary variables. Intuitively, this tells us how far away from the full model we would be if we ignored the pecuniary expectations.

We report the relative contribution of the subset of variables, calculated as:

$$\boldsymbol{R_{pec}} = \frac{M_{pec}}{M_{pec}+M_{non\ pec}} \tag{4}$$

Where $M_{non\ pec}$ is calculated analogously. R_{pec} and $R_{non\ pec}$ will then sum to one and each tell us the fraction of the model's prediction we can attribute to the given subset of variables.

The relative contributions are reported in the final rows of Table 4. For men 37.3% of the prediction comes from pecuniary variables, and the remaining 63.7% from non-pecuniary variables. Women have a lower R_{pec}, 18.4%, suggesting that they care relatively less about the pecuniary aspects of educational choice. The R_{pec} for both genders combined is 24.6%, which is remarkably close to the 24.95% found in Zafar (2013), despite the fact that our survey was administered on high school students in Sweden instead of college students in the US.

4.2 Calculating willingness to pay for amenity expectations

Since we have precise estimates of the coefficient on expected wages, we can translate our estimates into willingness to pay for the various expected amenities. To align with existing literature, we use the unstandardized measures of the variables and calculate, for each amenity:

$$\boldsymbol{WTP_{amenity}} = \boldsymbol{\hat{\theta}_{amenity}/\hat{\theta}_{log\_earnings}}$$

[17] Zafar (2013) includes status in the category of pecuniary variables, however our survey specifically asked students to consider status while holding earnings constant. We therefore think it's more appropriate to think of it as a non-pecuniary outcome in our context.

And then estimate standard errors using the delta method. Table 5 reports these separately for men and women, along with the difference. We estimate that male students are willing to forgo 0.43% of their wages for a 1 percentage point increase in the probability of finding a job immediately after graduation. Women on the other hand have effectively no willingness to pay for this amenity.[18]

Still, we see that both genders seem willing to pay for increasing their probability of passing, higher parental approval in their choice as well as both measures of enjoyment.

4.3 The importance of expectations and beliefs for the STEM gender gap in preferred fields of study

We have established that males rank STEM fields of study much higher than females (see Figure 1) and that there exist apparent gender differences in what are the beliefs and expectations that matter most in ranking of fields of study (Table 4). This suggests that it is a worthwhile exercise to see if the gender gap in STEM choice can be explained by differences in beliefs and expectations.

We investigate this issue in Table 6, where we report estimates from rank ordered logistic regression models. The dependent variable is the rank of the eight fields of study, from most to least preferred. The main variable of interest is an interaction term of a STEM indicator (=1 for the rank of a STEM field of study; 0 otherwise) and an indicator for student being female. If an estimate of this variable is negative, there is evidence of a STEM gender gap where females are less likely to prefer STEM fields of study at the university. We always control for fields of study fixed effects. All models effectively condition on unobserved individual characteristics, so that only effects of individual characteristics that vary with field of study choice can be estimated. The first three columns of Table 6 report the main results using the full sample, and the next three columns use a subsample where we also observe background variables of the individuals.

In the model in column 1, we simply include the interaction term between STEM and being female. The resulting coefficient estimate is estimated negative, meaning that females rank STEM fields lower than males. If we convert the estimate to an odds ratio (*exp(-0.688)*) we find that the odds of ranking a STEM field higher than a non-STEM field is 50% lower for females. In column 2, we also control for an interaction term of STEM and indicator for if the student attended a high school program that was more specialized in STEM subjects. This could potentially be important for our results since preferring STEM

[18] Note that the change in magnitudes of the coefficients relative to each other is because we are now using the units directly from the survey instead of standardized (mean 0, standard deviation 1) variables.

university fields of study is higher for those attending STEM-specialized high-school programs. However, the STEM gender gap only decreases moderately (about 20%) when this control is added.

In column 3 we add the full set of belief and expectations variables. Strikingly, the whole STEM gender gap is now eliminated.[19] This suggests that if we want to understand why male and female students vary in their preference for STEM fields we should look among these predictors, where the most important determinants of field of study ranks are how much students expect to enjoy their studies and work. Below, we will investigate the sources for eliminating the STEM gender gap further.

Although the estimated models condition on the choice set for an individual (effectively conditioning on individual characteristics), factors such as ability and family background can still be important in explaining the STEM gender gap. We therefore also estimate models on a smaller subsample where data on such characteristics is available. In column 4 we report the STEM gender gap for this subsample, finding it to be similar to the gap for the full sample. In column 5, we then add ability and family background variables interacted with STEM, and find, as expected, that students with high Math score and low English score are more likely to rank STEM fields higher. However, the STEM gender gap is unaffected. In Column 6 we add the full set of beliefs and expectations variables and find, again, that the STEM gender gap is eliminated.

Even though we control for attending a STEM specialized high school program for the models estimated in Table 6, the sample includes all students regardless of whether they attend a STEM-, non-STEM academic or occupational oriented high school program. We therefore present results in Table 7 where we estimate models on two subsamples: those students attending academic high-school programs (columns 1-3) and those students attending STEM specialized high-school programs (columns 4-5). As can be seen, results and conclusions from previous estimations in Table 6 are virtually unchanged, even though the standard errors increase a bit for the smallest STEM-specialized subsample.

While Table 6 shows us that the expectations elicited by the survey can explain the gender gap in STEM preferences, it is also interesting to ask which variables play the largest roles. Similarly to the previous section, in Table 8 we split the expectation variables into pecuniary and non-pecuniary factors. Columns 1 and 5 are repeated from Table 6, but in columns 2 and 3 we add first the pecuniary and then the non-pecuniary variables separately.

[19] If coefficients on beliefs and expectations are allowed to vary by gender in column 3, the coefficient on STEM*female barely changes (to -0.066, with se 0.126).

Following the addition of the pecuniary variables, the STEM-gender gap drops from 0.554 to 0.273 (50.7%). While this seems large, when we add only the non-pecuniary variables in column 3, the drop is 77.6% and the gender gap is no longer statistically different from zero. Strikingly, if we include only two of the non-pecuniary variables, the probabilities of enjoying coursework and job, the gap is even smaller. The smaller coefficient in column 4 relative to columns 3 and 5 tells us that conditioning on at least one of the expectation variables actually increases the gender gap.[20] While the gender gap gets very close to zero with only these two variables, it's not correct to conclude that the remaining eight don't matter. As seen earlier in Appendix Table A3, the enjoyment variables are significantly correlated with all of the other variables.

Thus, when no other variables are included, the coefficients on the enjoyment variables incorporate the importance of these other variables. It appears true that both pecuniary and non-pecuniary variables matter, but the pecuniary variables matter relatively more. And among the non-pecuniary variables, the two measures of enjoyment in college and during the career are the most important.

Finally, while Table 6 showed that subjective expectations explain the gender gap in STEM, we may also ask how they explain gender-gaps in all fields. To explore this, we repeat the estimation in the first three columns of Table 6 replacing the "STEM" dummy with indicators for the specific fields of study. The results are shown in Table 9, with only the gender-gap in ranking and expectations coefficients printed for simplicity. Column 2 shows the gender gaps in the rank of each of the fields without accounting for expectations. Most of the coefficients are meaningful in size and five are significantly different from zero. It is important to note that the coefficients tell us the gap in relation to the omitted category. Here, the omitted field is Social Sciences, which is the most popular field and which has a below average gender gap in average ranking.[21] Still, we can see that most of the gaps are meaningful in size and significantly different from zero.

In column 3 we add the expectations variables, and overall the coefficients shrink in size. The two categories that make up STEM both get smaller by more than an order of magnitude. The only category that remains large and significant is humanities, which is more preferred by women after controlling for expectations. One way of quantifying how small the gender gaps are, is by calculating the mean sum of squares of the coefficients.

$$\text{MSS} = \frac{1}{N}\sum_j(\widehat{\gamma}_{j,female})^2 \tag{5}$$

[20] As an example, it could be that given their expectation about how easy it will be to balance work and family, men actually rank STEM fields higher than we would expect.

[21] If we were to redo Table 6 replacing the "STEM" indicator with a "Social Science" indicator, the coefficient in column 1 would be 0.252, significant at the 10% level.

Where $\hat{\gamma}_{j,female}$ is the estimated gender gap for field j. This is analogous to the variance around 0, and tells us how close the coefficients are getting to zero regardless of sign. The results for columns 2 and 3 are in the end of Table 9. The MSS of the gaps without expectations is 0.153, but with expectations it drops to 0.029 (a fall of 80%). This is simply another way of saying that expectations explain much of the differences in choice of field of study between men and women, not only the choice of STEM/non-STEM.

5. INTERPRETING TASTES AND ENJOYMENT

In the model described in equation (2), we think of individual's utility, and thus their preference rankings, varying by the expectations $\boldsymbol{w_{ij}}$, and the weights on these expectations $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ translates the inputs into individual utility, we think of this vector as measuring individual's preferences. Thus far the results have indicated that the gender gap in major choice is more due to differences in expectations ($\boldsymbol{w_{ij}}$'s) than it is in preferences ($\boldsymbol{\theta}$'s).

However, the two enjoyment measures seem intuitively different from the remainder of the $\boldsymbol{w_{ij}}$'s. We can think of them in two ways. First, it is possible that enjoyment is a non-pecuniary amenity similar to job flexibility. The students may intrinsically find the coursework and job tasks in some fields to be more appealing than others. If a student is afraid of blood, it makes sense that they would not expect to enjoy being in the medical field. The second way to think of enjoyment is as a measure of utility. We may expect that asking students how much they will enjoy college or their job, is a bit like asking them what utility they expect from the choice. To put this concretely, we can think of enjoyment as being itself a function of the other $\boldsymbol{w_{ij}}$'s and their associated preference parameters $\boldsymbol{\theta}$. This is different, and more problematic than saying that the different expectations are correlated, because it also incorporates $\boldsymbol{\theta}$. As the true interpretation of enjoyment is likely a combination of the two described above, we may also model that expected enjoyment is also a function of an intrinsic taste for the content of the field: $\boldsymbol{taste_{ij}}$.

$$\boldsymbol{enjoy_{ij}} = \boldsymbol{f(\theta, w_{ij,\neg enjoy})} + \boldsymbol{taste_{ij}}. \tag{6}$$

Where $\boldsymbol{w_{ij,\neg enjoy}}$ is the vector of expected amenities excluding the enjoyment measures. While one could make an argument that any stated expectation also captures preferences, the argument is by far the strongest for the two variables related to enjoyment. Indeed, our finding that the enjoyment variables alone can explain the entire gender gap is at least suggestive that they are capturing overall utility.

Zafar (2013) calls the enjoyment variables measures of "tastes" for the field of study, and the intuition for thinking of them as tastes is clear: they ask how much the individual likes the job. Wiswall & Zafar (2015) manipulate expected wage using an experiment and find that unobserved tastes for programs must be positively correlated with expected wage. As they point out, this means that if individual field-specific preferences are ignored, the importance of wage is likely upwardly biased. For this reason, we expect that we get better estimates of the preference parameters on the other amenities when we control for enjoyment.

6. DISCUSSION

College major choice is an important determinant of outcomes for individuals (e.g. in determining future wages), labor markets (in supplying trained workers in growing fields), and societies (in reducing or perpetuating sources of inequality). Many of these outcomes are of great interest to policy makers, but without an accurate idea of which incentives matter in choice, it's difficult to design appropriate policy instruments.

The focus of this paper has been a particular setting and source of inequality: the gender gap in STEM major choice. We use this example to show that in our diverse sample of Swedish high school students, expectations substantially explain the large differences in probability of choosing a STEM field by gender. We estimated a choice model relating subjective expectations to students' ranking of fields of study. The results suggest that young men and women value most of the expected college and work amenities related to field of study quite equally. Instead, what drives the gender gap is differences by gender in expected amenities by field of study, and differences in both observed and unobserved tastes for certain fields.

Our findings are very similar to those found previously for a very different sample of US college students (Zafar, 2013). This builds upon the existing literature to emphasize that while expected wages do seem to matter when students choose majors, they predict a fairly small amount of the choice. This means that increasing wages in a field, a common policy suggestion, would be a costly way of changing enrollment levels.

REFERENCES:

Allison, Paul D., and Nicholas A. Christakis. 1994. "Logit Models for Sets of Ranked Items." *Sociological Methodology*, 199–228.

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø Sørensen, and Bertil Tungodden. 2016. "What Explains the Gender Gap in College Track Dropout? Experimental and Administrative Evidence." *American Economic Review* 106 (5): 296–302. https://doi.org/10.1257/aer.p20161075.

Altonji, Joseph G., Peter Arcidiacono, and Arnaud Maurel. 2015. "The Analysis of Field Choice in College and Graduate School: Determinants and Wage Effects." Working Paper 21655. National Bureau of Economic Research. https://doi.org/10.3386/w21655.

Arcidiacono, Peter, V. Joseph Hotz, and Songman Kang. 2012. "Modeling College Major Choices Using Elicited Measures of Expectations and Counterfactuals." *Journal of Econometrics* 166 (1): 3–16.

Arcidiacono, Peter, V. Joseph Hotz, Arnaud Maurel, and Teresa Romano. 2017. "Ex Ante Returns and Occupational Choice."

Attanasio, Orazio P., and Katja M. Kaufmann. 2014. "Education Choices and Returns to Schooling: Mothers' and Youths' Subjective Expectations and Their Role by Gender." *Journal of Development Economics* 109 (July): 203–16. https://doi.org/10.1016/j.jdeveco.2014.04.003.

Beggs, S, S Cardell, and J Hausman. 1981. "Assessing the Potential Demand for Electric Cars." *Journal of Econometrics* 17 (1): 1–19. https://doi.org/10.1016/0304-4076(81)90056-7.

Card, David, and A. Abigail Payne. 2017. "High School Choices and the Gender Gap in STEM." Working Paper 23769. National Bureau of Economic Research. https://doi.org/10.3386/w23769.

Chapman, Randall G., and Richard Staelin. 1982. "Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model." *Journal of Marketing Research* 19 (3): 288–301. https://doi.org/10.2307/3151563.

Dominitz, Jeff, and Charles F. Manski. 1996. "Eliciting Student Expectations of the Returns to Schooling." *The Journal of Human Resources* 31 (1): 1–26. https://doi.org/10.2307/146041.

Hastings, Justine, Christopher A. Neilson, and Seth D. Zimmerman. 2015. "The Effects of Earnings Disclosure on College Enrollment Decisions." National Bureau of Economic Research.

Hastings, Justine S., Christopher A. Neilson, and Seth D. Zimmerman. 2013. "Are Some Degrees Worth More than Others? Evidence from College Admission Cutoffs in Chile." National Bureau of Economic Research.

Hausman, Jerry A., and Paul A. Ruud. 1987. "Specifying and Testing Econometric Models for Rank-Ordered Data." *Journal of Econometrics* 34 (1): 83–104. https://doi.org/10.1016/0304-4076(87)90068-6.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection." *The Quarterly Journal of Economics* 131 (3): 1057–1111. https://doi.org/10.1093/qje/qjw019.

Kugler, Adriana D., Catherine H. Tinsley, and Olga Ukhaneva. 2017. "Choice of Majors: Are Women Really Different from Men?" National Bureau of Economic Research.

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior."

———. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3 (4):

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. 1993. "Why Is Rent-Seeking so Costly to Growth?" *The American Economic Review* 83 (2): 409–414.

Rapoport, Benoît, and Claire Thibout. 2018. "Why Do Boys and Girls Make Different Educational Choices? The Influence of Expected Earnings and Test Scores." *Economics of Education Review* 62 (February): 205–29. https://doi.org/10.1016/j.econedurev.2017.09.006.

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. 2017. "Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender." *The Economic Journal* 127 (604): 2153–86. https://doi.org/10.1111/ecoj.12350.

Stinebrickner, Ralph, and Todd R. Stinebrickner. 2014. "A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout." *The Review of Economic Studies* 81 (1): 426–72. https://doi.org/10.1093/restud/rdt025.

Stinebrickner, Todd, and Ralph Stinebrickner. 2012. "Learning about Academic Ability and the College Dropout Decision." *Journal of Labor Economics* 30 (4): 707–748.

Wiswall, Matthew, and Basit Zafar. 2015. "Determinants of College Major Choice: Identification Using an Information Experiment." *The Review of Economic Studies* 82 (2): 791–824. https://doi.org/10.1093/restud/rdu044.

———. 2016. "Human Capital Investments and Expectations about Career and Family." National Bureau of Economic Research.

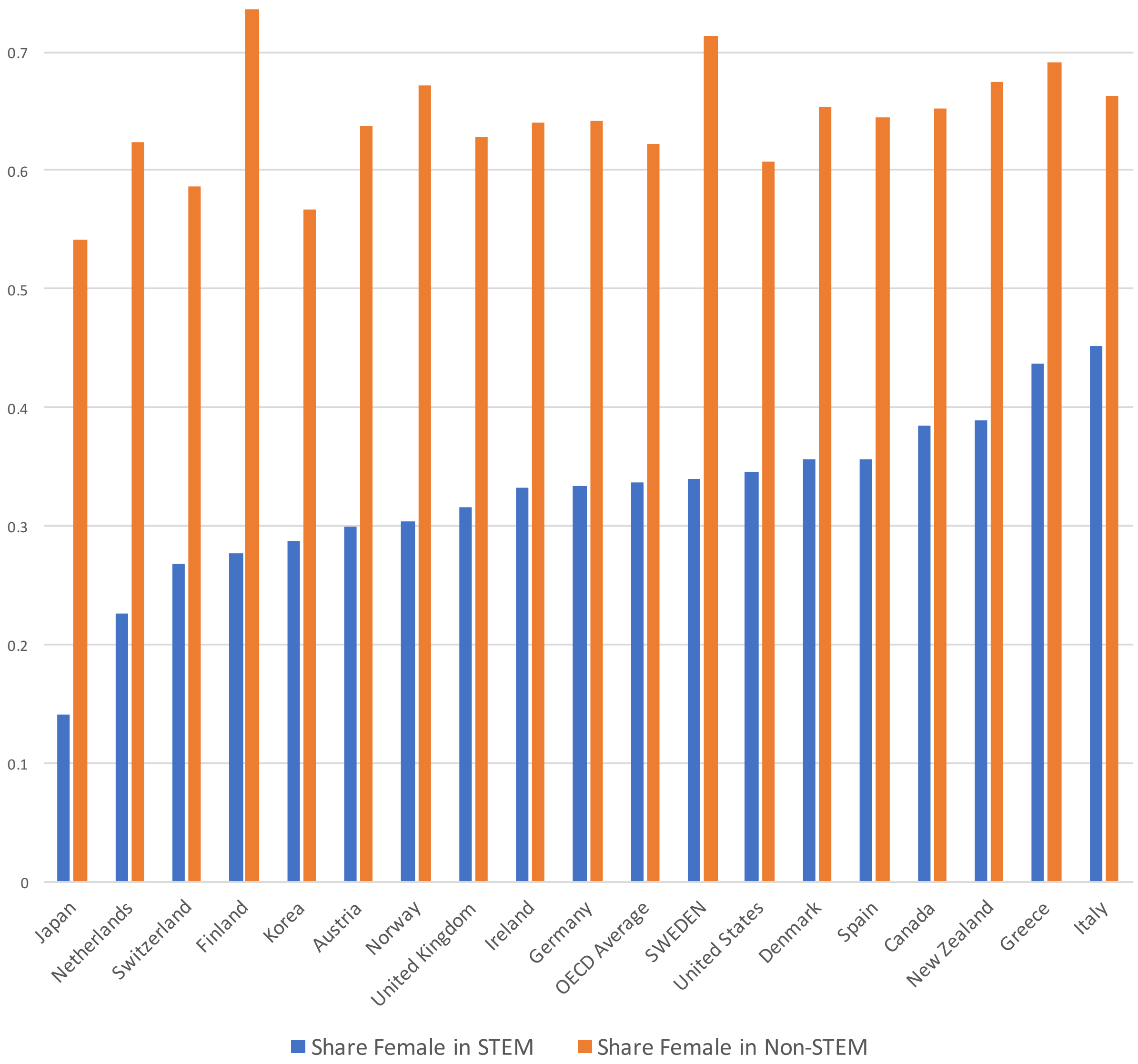
———. 2018. "Preference for the Workplace, Investment in Human Capital, and Gender." *The Quarterly Journal of Economics* 133 (1): 457–507. https://doi.org/10.1093/qje/qjx035.

Zafar, Basit. 2011. "Can Subjective Expectations Data Be Used in Choice Models? Evidence on Cognitive Biases." *Journal of Applied Econometrics* 26 (3): 520–544.

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

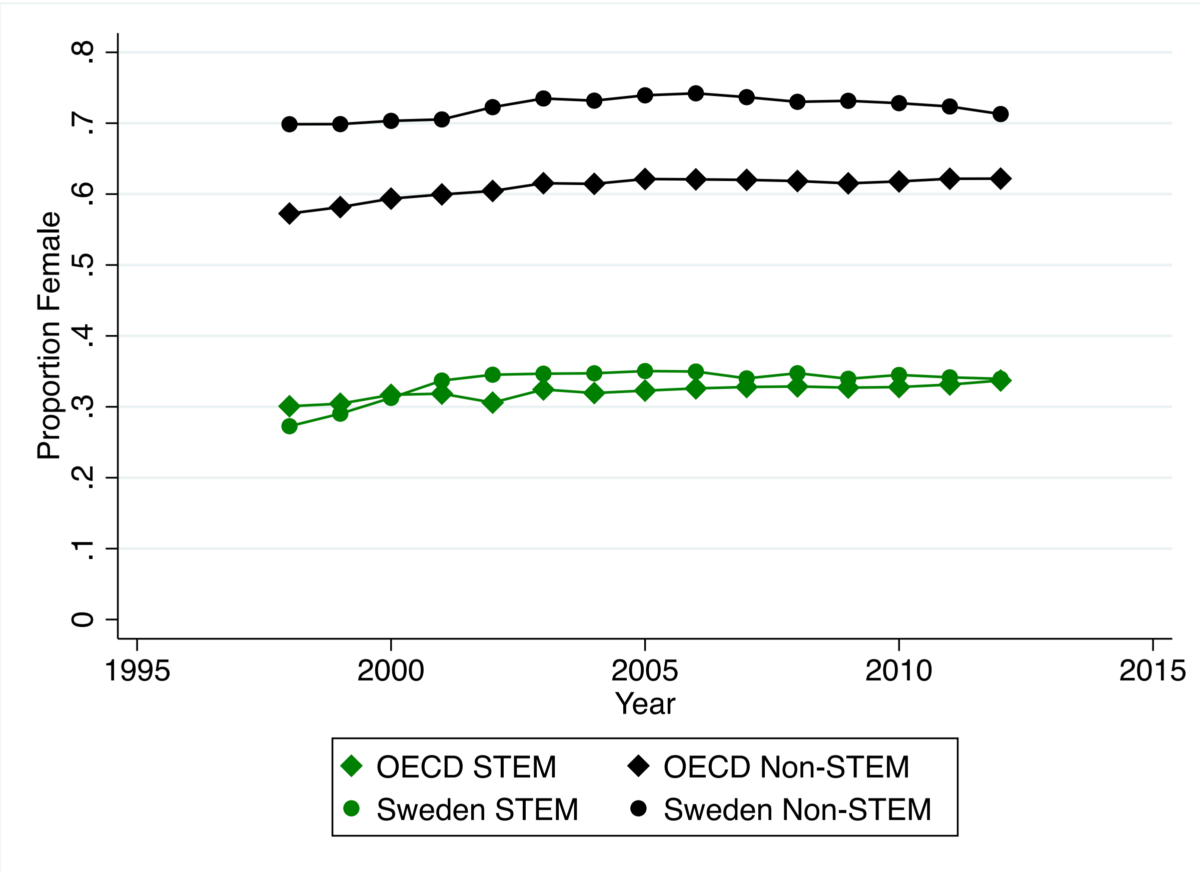
Figures

Figure 1: Gender composition of college graduates, STEM and Non-STEM selected OECD countries, 2012



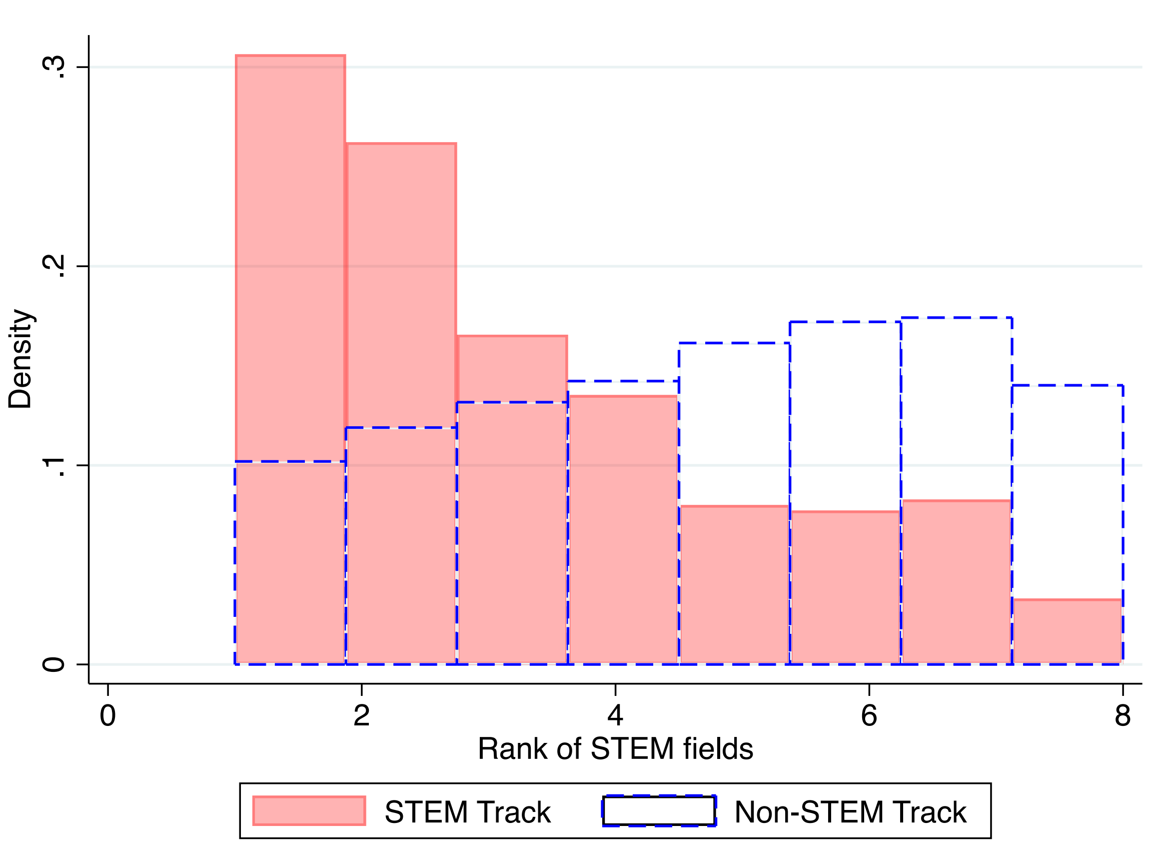
Note: Data from OECD (http://stats.oecd.org/viewhtml.aspx?datasetcode=RGRADSTY). Accessed January 2017.

Figure 2: Trends in gender composition of STEM and non-STEM graduates, Sweden vs OECD average



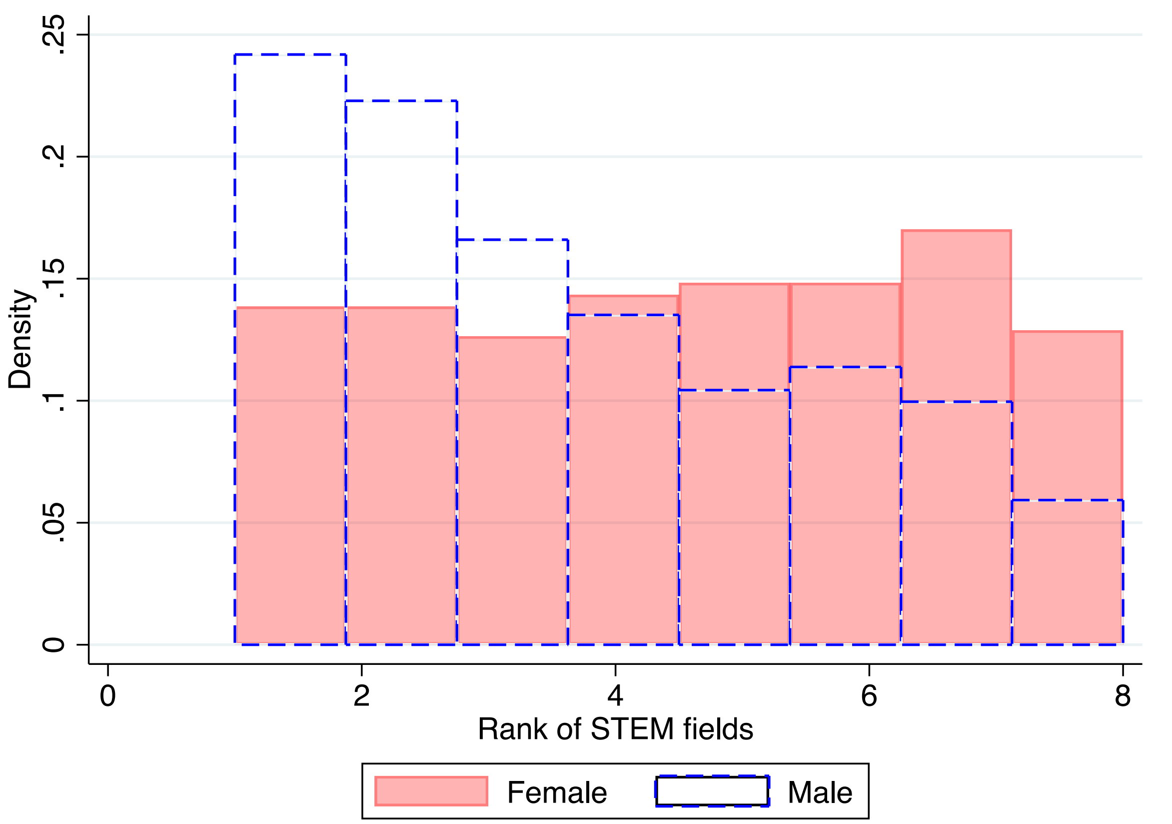
Note: Data from OECD (http://stats.oecd.org/viewhtml.aspx?datasetcode=RGRADSTY). Accessed January 2017.

Figure 3: Rank of fields divided by High School STEM or non-STEM programs



Note: The highest (most preferred) rank is 1, lowest is 8. STEM Track means that the student was enrolled in a high school program that was designed to prepare for STEM university studies. Each individual in the sample is included twice, once for their ranking of the science and mathematics field, and once for their ranking of the technology and manufacturing field. The filled bars sum to 207*2=414 and the dashed bars sum to 269*2=538.

Figure 4: Rank of STEM Field by Gender



Note: The highest rank is 1, lowest is 8. Each individual in the sample is included twice, once for their ranking of the science and mathematics field, and once for their ranking of the technology and manufacturing field. The female bars sum to 235*2=470 and the male bars sum to 241*2=482.

Tables

Table 1: Summary statistics on family and high school variables

	Surveyed Sample			
	(1)	(2)	(3)	(4)
	Male	Female	Total	Population
Background Variables:				
Foreign background	0.163	0.207	0.185	
	(0.371)	(0.406)	(0.389)	
Mom went to university	0.506	0.498	0.502	
	(0.501)	(0.501)	(0.501)	
Father went to university	0.496	0.484	0.490	
	(0.501)	(0.501)	(0.500)	
Median parental income	883.6	844.7	868.8	
(in 1,000s SEK)	(800.3)	(541.3)	(687.5)	
SES Index using PCA	0.0835	-0.0857	0.000	
	(1.656)	(1.455)	(1.561)	
School Variables:				
Avg. English Score (/20)	16.26	15.98	16.13	
	(5.169)	(5.086)	(5.124)	
Avg. Math Score (/20)	12.99	12.60	12.80	
	(5.169)	(5.086)	(5.124)	
College Prep Program	0.884	0.915	0.899	
	(0.321)	(0.280)	(0.301)	
STEM Prep Program	0.506	0.362	0.435	
	(0.501)	(0.482)	(0.496)	
Total Observations	241	235	476	
N for Math Scores	190	191	381	
N with all Vars	163	167	330	

Note: Students are regarded as being of foreign background is if either the they or both their parents are immigrants to Sweden. Parent income is annual and is the sum of income from parents who can be matched to the child. SES index combines parent income and education levels, see Data Appendix for further detail. Test scores and foreign background are not available for the entire sample, with math scores having the lowest coverage. “N with all Vars” on the bottom row provides the sample for which we have all the information in this table. *Population column to be filled in in future version.*

Table 2: Expectations: Means by Gender and STEM Field

	Overall			STEM Fields			Non-STEM Fields		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Male	Female	Difference	Male	Female	Difference	Male	Female	Difference
Predicted earnings at 40	37.15	35.62	1.483*	42.09	39.87	2.219*	35.48	34.24	1.238
	(13.04)	-11.18	(0.746)	(12.81)	(11.56)	(1.024)	(12.64)	-10.69	(0.715)
Predicted hrs/wk (age 30)	45.75	47.1	-1.417	46.44	46.53	-0.0878	45.37	47.23	-1.860*
	(12.16)	-13.14	(0.825)	(10.35)	(12.28)	(0.944)	(12.79)	-13.43	(0.851)
Prob find a job	61.76	62.17	-0.233	70.16	63.38	6.785***	59.15	61.72	-2.572
	(25.41)	-24.93	(1.381)	(20.51)	(24.60)	(1.840)	(26.24)	-25	(1.447)
Prob of passing the degree	69.65	71	-1.128	70.33	60.39	9.943***	69.59	74.41	-4.819**
	(25.41)	-27.77	(1.740)	(24.07)	(29.53)	(2.284)	(25.80)	-26.24	(1.794)
Predicted study hrs/wk	36.66	41.92	-5.320***	41.54	48.13	-6.595***	34.81	39.7	-4.895***
	(18.88)	-19.75	(1.450)	(18.87)	(20.66)	(1.714)	(18.64)	-18.95	(1.420)
Perceived status for degree	58.04	58.52	-0.430	68.82	69.14	-0.327	54.64	55.1	-0.465
	(23.98)	-24.56	(0.928)	(17.33)	(19.80)	(1.355)	(24.88)	-24.95	(1.054)
Prob enjoy job (age 30)	57.26	55.51	1.803	63.72	50.29	13.43***	55.16	57.23	-2.072
	(25.54)	-27.61	(1.363)	(23.31)	(27.75)	(2.106)	(25.90)	-27.26	(1.454)
Prob work-life balance (age 30)	62.35	64.03	-1.612	63.89	62.41	1.473	62.00	64.64	-2.641
	(23.25)	-22.68	(1.432)	(20.92)	(22.29)	(1.826)	(23.98)	-22.71	(1.438)
Parental approval	72.29	72.62	-0.0647	81.90	78.14	3.760	69.38	70.72	-1.340
	(27.75)	-28.61	(1.745)	(20.60)	(26.34)	(1.925)	(29.08)	-29.04	(1.897)
Prob of enjoying coursework	56.47	55.62	0.965	63.54	47.76	15.78***	54.38	58.35	-3.972**
	(27.07)	-28.56	(1.397)	(26.08)	(28.93)	(2.242)	(27.05)	-27.94	(1.499)
Individuals	241	235	476	241	235	476	241	235	476
N	1928	1880	3808	482	470	952	1446	1410	2856

Note: Columns labeled "Male" or "Female" report means with standard deviation in parentheses. Columns labeled "Difference" report the male-female gap with standard error in parentheses (clustered at the individual level). STEM fields includes two categories: Natural Sciences/Math, and Technology and Engineering. Non-STEM fields include six categories: Social Sciences, Health, Humanities/Art, Pedagogy, Services and Agriculture/Animals. Stars below differences represent the significance level (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Table 3: Means of expectations by first ranked or not first ranked

	Male		Female	
	(1)	(2)	(3)	(4)
	First Rank	Later Ranks	First Rank	Later Ranks
Expected earnings at 40	42.71	36.33	40.86	34.90
	(13.81)	(12.69)	(12.27)	(10.81)
Expected hrs/wk (age 30)	48.15	45.28	49.94	46.64
	(11.11)	(12.34)	(13.14)	(13.11)
Prob find a job	74.23	60.14	70.20	60.99
	(21.01)	(25.47)	(23.92)	(24.84)
Prob of passing the degree	82.90	67.90	83.93	69.05
	(17.39)	(25.79)	(20.57)	(28.16)
Expected study hrs/wk	40.69	35.89	46.13	41.19
	(19.76)	(18.72)	(19.87)	(19.64)
Perceived status for degree	71.52	56.28	75.32	56.23
	(17.90)	(24.18)	(20.00)	(24.19)
Prob enjoy job (age 30)	78.44	54.28	78.70	52.18
	(15.34)	(25.28)	(19.79)	(26.89)
Prob work-life balance (age 30)	68.38	61.63	67.94	63.54
	(20.80)	(23.48)	(21.03)	(22.79)
Parental approval	87.70	70.34	88.86	70.25
	(17.35)	(28.26)	(16.33)	(29.17)
Prob of enjoying coursework	79.25	53.44	81.45	52.03
	(19.28)	(26.52)	(20.41)	(27.65)
Individuals	241	241	235	235
Observations	241	1687	235	1645

Note: Standard deviation in parentheses. Later ranks are the mean of the 7 fields that were not ranked as first choice.

Table 4: The impact of expectations on the ranking of fields, by gender

	Separate			Combined		
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Difference	Male	Female	Difference
Log Expected earnings at 40	0.750**	0.540**	0.210*	0.332**	0.227**	0.105
	(0.0630)	(0.0567)	(0.0846)	(0.0680)	(0.0579)	(0.0892)
Expected hrs/wk (age 30)	0.341**	0.305**	0.0352	-0.0268	0.0599	-0.0867
	(0.0545)	(0.0414)	(0.0684)	(0.0605)	(0.0563)	(0.0826)
Prob find a job	0.616**	0.388**	0.228**	0.241**	-0.0221	0.263**
	(0.0503)	(0.0517)	(0.0721)	(0.0444)	(0.0543)	(0.0701)
Prob of passing the degree	0.668**	0.757**	-0.0888	0.288**	0.323**	-0.0348
	(0.0627)	(0.0625)	(0.0884)	(0.0749)	(0.0800)	(0.110)
Expected study hrs/wk	0.615**	0.381**	0.233*	0.174*	-0.138+	0.312**
	(0.0818)	(0.0692)	(0.107)	(0.0804)	(0.0801)	(0.113)
Perceived status for degree	0.666**	0.582**	0.0841	0.205**	0.384**	-0.180*
	(0.0549)	(0.0462)	(0.0717)	(0.0661)	(0.0575)	(0.0875)
Prob enjoy job (age 30)	1.181**	1.124**	0.0570	0.579**	0.521**	0.0582
	(0.0743)	(0.0765)	(0.107)	(0.0716)	(0.0758)	(0.104)
Prob work-life balance (age 30)	0.167**	0.239**	-0.0716	0.0641	0.0935+	-0.0295
	(0.0516)	(0.0519)	(0.0731)	(0.0507)	(0.0553)	(0.0749)
Parental approval	0.970**	0.873**	0.0975	0.251**	0.323**	-0.0719
	(0.0784)	(0.0828)	(0.114)	(0.0802)	(0.0719)	(0.108)
Prob of enjoying coursework	1.168**	1.102**	0.0654	0.569**	0.619**	-0.0499
	(0.0744)	(0.0670)	(0.100)	(0.0757)	(0.0805)	(0.110)
$R_{pecuniary}$				0.373	0.184	
$R_{nonpecuniary}$				0.627	0.816	
Log Likelihood				-1920.47	-1876.44	-3796.91
Pseudo R^2				0.249	0.247	0.248
Individuals	241	235	476	241	235	476
N	1928	1880	3808	1928	1880	3808

Note: First three columns report coefficients from 10 separate regressions, each only including a single expectation variable to predict rank. The final three columns include all the expectation measures in one regression. All columns estimate rank ordered logistic models. Standard errors in parentheses (+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table 5: Willingness to Pay for non-wage expecations

	(1) Male	(2) Female	(3) Difference
Expected hrs/wk (age 30)	0.151	-0.359	0.510
	(0.282)	(0.387)	(0.479)
Prob find a job	-0.429**	0.0109	-0.440+
	(0.136)	(0.194)	(0.237)
Prob of passing the degree	-0.506*	-0.961*	0.455
	(0.222)	(0.403)	(0.460)
Expected study hrs/wk	-0.394	0.682	-1.076*
	(0.241)	(0.372)	(0.443)
Perceived status for degree	-0.405	-0.993*	0.588
	(0.214)	(0.451)	(0.499)
Prob work-life balance (age 30)	-0.130	-0.344	0.214
	(0.122)	(0.210)	(0.243)
Parental approval	-0.421*	-0.895*	0.474
	(0.206)	(0.366)	(0.420)
Prob enjoy job (age 30)	-1.199***	-1.597*	0.398
	(0.355)	(0.621)	(0.715)
Prob of enjoying coursework	-1.103***	-1.684**	0.581
	(0.332)	(0.633)	(0.715)
Individuals	241	235	
N	1928	1880	

Willingness to pay estimates derived as the quotient of the estimated coefficient on the given variable and the one on earnings. Unlike in other tables, these variables have not been standardized, thus they can be interpreted as the willingness to pay in % of earnings for a unit increase in the expectation. Standard errors calculated via the delta method in parentheses (+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table 6: The impact of expectations on the STEM gender gap in preferred fields of study

	Main Sample			Controls Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
STEM field=1 × Female	-0.688**	-0.554**	-0.0743	-0.605**	-0.604**	-0.0849
	(0.118)	(0.113)	(0.121)	(0.139)	(0.137)	(0.138)
STEM field=1 × STEM Spec. HS Program=1		1.098**	0.453**	1.170**	1.065**	0.477**
		(0.119)	(0.122)	(0.142)	(0.142)	(0.154)
STEM field=1 × Math Test Score					0.199**	0.0101
					(0.0771)	(0.0819)
STEM field=1 × English Test Score					-0.237*	-0.223+
					(0.101)	(0.126)
STEM field=1 × SES Index					0.113	0.186
					(0.0788)	(0.115)
STEM field=1 × Foreign background=1					0.199	0.249
					(0.196)	(0.211)
Expectations (Standardized):						
Log Expected earnings at 40			0.202**			0.244**
			(0.0454)			(0.0580)
Expected hrs/wk (age 30)			0.0248			0.0620
			(0.0449)			(0.0607)
Prob find a job			0.101*			0.0881+
			(0.0396)			(0.0480)
Prob of passing the degree			0.271**			0.172*
			(0.0558)			(0.0698)
Expected study hrs/wk			0.00401			0.0447
			(0.0575)			(0.0640)
Perceived status for degree			0.254**			0.241**
			(0.0466)			(0.0589)
Prob enjoy job (age 30)			0.568**			0.588**
			(0.0522)			(0.0623)
Prob work-life balance (age 30)			0.0806*			0.0536
			(0.0385)			(0.0451)
Parental approval			0.255**			0.231**
			(0.0556)			(0.0723)
Prob of enjoying coursework			0.578**			0.678**
			(0.0568)			(0.0711)
Field Fixed Effects	X	X	X	X	X	X
Log Likelihood	-4755.75	-4687.47	-3806.56	-3188.35	-3181.66	-2546.06
Pseudo R^2	0.058	0.071	0.246	0.089	0.091	0.272
Individuals	476	476	476	330	330	330
N	3808	3808	3808	2640	2640	2640

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Note: All columns estimate rank ordered logistic regressions on the ranking of a field. Columns 1-3 use the full sample, while columns 4-5 use only the sample for which all demographic controls are available. All continuous variables have been standardized to mean zero, standard deviation one.

Table 7: Sensitivity to dropping students in vocational high schools and non-STEM focused programs

	College Prep			STEM Program	
	(1)	(2)	(3)	(4)	(5)
STEM field=1 × Female	-0.683**	-0.507**	-0.0315	-0.517**	-0.0836
	(0.127)	(0.121)	(0.127)	(0.192)	(0.195)
STEM field=1 × STEM Spec. HS Program=1		1.203**	0.529**		
		(0.124)	(0.129)		
Expectations (Standardized):					
Log Expected earnings at 40			0.217**		0.228**
			(0.0485)		(0.0720)
Expected hrs/wk (age 30)			0.0323		-0.0352
			(0.0477)		(0.0759)
Prob find a job			0.0776^{+}		0.0850
			(0.0410)		(0.0594)
Prob of passing the degree			0.229**		0.185*
			(0.0604)		(0.0887)
Expected study hrs/wk			0.00285		0.151
			(0.0605)		(0.0966)
Perceived status for degree			0.247**		0.339**
			(0.0496)		(0.0790)
Prob enjoy job (age 30)			0.574**		0.651**
			(0.0545)		(0.0829)
Prob work-life balance (age 30)			0.0729^{+}		0.0862
			(0.0392)		(0.0668)
Parental approval			0.278**		0.0760
			(0.0587)		(0.0967)
Prob of enjoying coursework			0.604**		0.596**
			(0.0610)		(0.0900)
Field Fixed Effects	X	X	X	X	X
Log Likelihood	-4253.88	-4179.12	-3379.75	-1980.91	-1592.27
Pseudo R^2	0.063	0.079	0.255	0.098	0.275
Individuals					
N	3424	3424	3424	1656	1656

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Note: All columns estimate rank ordered logistic regressions on the ranking of a field. Columns 1-3 use all students who are in a college preparatory gymnasium program, while columns 4-5 use only those who attend a STEM-focused gymnasium program.

Table 8: Decomposing the impact of expectations on the STEM gender gap in preferred fields of study

	(1) No Expl	(2) Pecuniary	(3) Non-Pec	(4) Enjoyment	(5) All
STEM field=1 × Female	-0.554**	-0.273**	-0.124	-0.0116	-0.0743
	(0.113)	(0.116)	(0.120)	(0.117)	(0.121)
STEM field=1 × STEM Spec. HS Program= 1	1.098**	0.714**	0.524**	0.529**	0.453**
	(0.119)	(0.118)	(0.120)	(0.120)	(0.122)
Expectations (Standardized):					
Log Expected earnings at 40		0.420**			0.202**
		(0.0426)			(0.0454)
Expected hrs/wk (age 30)		0.0799+			0.0248
		(0.0437)			(0.0449)
Prob find a job		0.287**			0.101*
		(0.0410)			(0.0396)
Prob of passing the degree		0.727**			0.271**
		(0.0528)			(0.0558)
Expected study hrs/wk		0.197**			0.00401
		(0.0540)			(0.0575)
Perceived status for degree			0.300**		0.254**
			(0.0425)		(0.0466)
Prob enjoy job (age 30)			0.609**	0.739**	0.568**
			(0.0520)	(0.0528)	(0.0522)
Prob work-life balance (age 30)			0.0848*		0.0806*
			(0.0374)		(0.0385)
Parental approval			0.284**		0.255**
			(0.0558)		(0.0556)
Prob of enjoying coursework			0.689**	0.723**	0.578**
			(0.0556)	(0.0541)	(0.0568)
Field Fixed Effects	X	X	X	X	X
Log Likelihood	-4687.47	-4283.51	-3844.89	-3927.08	-3806.56
Pseudo R^2	0.071	0.151	0.238	0.222	0.246
Individuals	476	476	476	476	476
N	3808	3808	3808	3808	3808

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Note: All columns estimate rank ordered logistic regressions on the ranking of a field.

Table 9: Expectations role in explaining gender gaps in all fields

	(1)	(2)	(3)
Pedagogy × Female		-0.192	-0.123
		(0.161)	(0.169)
Humanities and Art × Female		0.340^{+}	0.430^{*}
		(0.188)	(0.198)
Social Science × Female		0	0
		(.)	(.)
Science and Math × Female		-0.469^{**}	0.00210
		(0.175)	(0.186)
Tech and Engineering × Female		-0.814^{**}	-0.0292
		(0.179)	(0.176)
Agro and Animal × Female		-0.336^{+}	-0.0983
		(0.190)	(0.188)
Healthcare × Female		0.0184	0.119
		(0.162)	(0.179)
Services × Female		-0.279^{+}	0.101
		(0.161)	(0.180)
Expectations (Standardized):			
Log Expected earnings at 40	0.196^{**}		0.200^{**}
	(0.0456)		(0.0458)
Expected hrs/wk (age 30)	0.0249		0.0235
	(0.0454)		(0.0454)
Prob find a job	0.0996^{*}		0.0985^{*}
	(0.0396)		(0.0398)
Prob of passing the degree	0.274^{**}		0.271^{**}
	(0.0563)		(0.0565)
Expected study hrs/wk	0.00205		-0.00309
	(0.0573)		(0.0568)
Perceived status for degree	0.248^{**}		0.247^{**}
	(0.0463)		(0.0478)
Prob enjoy job (age 30)	0.569^{**}		0.567^{**}
	(0.0524)		(0.0525)
Prob work-life balance (age 30)	0.0852^{*}		0.0882^{*}
	(0.0388)		(0.0392)
Parental approval	0.252^{**}		0.257^{**}
	(0.0555)		(0.0558)
Prob of enjoying coursework	0.583^{**}		0.579^{**}
	(0.0571)		(0.0568)
Track FE	X	X	X
Track FE × STEM Spec. HS Program	X	X	X
Mean Sum of Squares		0.153	0.029
Pseudo R^2	0.247	0.079	0.248
N	3808	3808	3808

Note: Standard errors in parentheses (+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$). All columns estimate rank ordered logistic regressions on the ranking of a field. Mean sum of squares is calculated over the eight field-gender gaps in the top panel.

1 Appendix Tables

Table A1: Effect of Field of study on the gender earnings gap

	(1)	(2)	(3)
Female	-0.309***	-0.279***	-0.240***
	(0.004)	(0.004)	(0.004)
STEM		0.170***	
		(0.006)	
Field of Study FE			X
R^2	0.09	0.10	0.13
N	140,822	140,822	140,822

Note: Standard errors in parentheses, (*** $p < 0.001$). Sample is from Swedish register data and includes individuals born 1950-1967 with at least short (2 years) of university education. Dependent variable is log of average yearly earnings, using earnings 2003-2005. Earnings is defined as total income from work, retirement benefits, social benefits and other labor market related benefits. All models control for birth year indicators and three educational level indicators. STEM is defined as having a degree in science, technology, engineering and mathematics fields.

Table A2: Selection into being surveyed, by HS program

Program	(1) College Prep?	(2) Original List	(3) Percent	(4) Our Sample	(5) Percent
Natural Science	yes	919	27.3%	182	38.2%
Social Science	yes	734	21.8%	116	24.4%
Business	yes	436	12.9%	46	9.7%
Introduction		337	10.0%	9	1.9%
Arts	yes	220	6.5%	36	7.6%
Technology	yes	148	4.4%	22	4.6%
Humanities	yes	126	3.7%	6	1.3%
Electricity and Energy		90	2.7%	11	2.3%
International Baccalaureate	yes	47	1.4%	4	0.8%
Crafts		41	1.2%	8	1.7%
Restaurant		41	1.2%	2	0.4%
Trade and Administration		40	1.2%	1	0.2%
Individual Program		38	1.1%		0.0%
Treatment and Care		34	1.0%	5	1.1%
Hotel and Tourism		19	0.6%		0.0%
Vehicles and Transport		18	0.5%	2	0.4%
SM		18	0.5%		0.0%
Childcare		16	0.5%	1	0.2%
Natural Resource Use		16	0.5%	1	0.2%
Total		3,367	100	476	100

Note: Column 2 reports frequencies of gymnasium programs from the full original list we began with, column 4 includes only the individuals who were successfully contacted and surveyed. Programs in Column 2 are the ones the students were enrolled in in year 20XX. In Column 4-5 we use the application program. Note, a handful of rare programs are not included in this figure.

Table A3: Correlation between expectations

	(1) Log(Earn) Age 40	(2) Hrs/Wk Age 30	(3) Prob Find Job	(4) Prob Graduate	(5) Study Hrs/Wk	(6) Status	(7) Enjoy Job Age 30	(8) Work-Life Balance	(9) Parental Approval	(10) Enjoy Courses
Log expected earnings at 40		0.229	0.257	0.0645	0.183	**0.438**	0.222	-0.002	0.248	0.172
Expected hrs/wk (age 30)	0.229		0.146	0.0644	**0.352**	0.242	0.173	-0.034	0.195	0.121
Prob find a job	0.257	0.146		0.299	0.127	0.325	**0.393**	0.203	0.371	0.356
Prob of passing the degree	0.065	0.0644	0.299		-0.0171	0.107	0.383	0.263	0.219	**0.567**
Expected study hrs/wk	0.183	**0.352**	0.127	-0.017		0.326	0.099	0.052	0.213	0.095
Perceived status for degree	0.438	0.242	0.325	0.107	0.326		0.415	0.118	**0.476**	0.339
Prob enjoy job (age 30)	0.222	0.173	0.393	0.383	0.099	0.415		0.312	0.445	**0.660**
Prob work-life balance (age 30)	-0.002	-0.034	0.203	**0.263**	0.052	0.118	0.312		0.156	0.243
Parental approval	0.248	0.195	0.371	0.219	0.213	**0.476**	0.445	0.156		0.407
Prob of enjoying coursework	0.172	0.121	0.356	0.567	0.095	0.339	**0.660**	0.243	0.407	
Panel B: Correlations after removing individual fixed effects from all expectations										
Log expected earnings at 40		0.301	0.327	0.046	0.441	**0.610**	0.304	-0.041	0.444	0.248
Expected hrs/wk (age 30)	0.301		0.231	0.125	**0.312**	0.299	0.203	-0.141	0.286	0.171
Prob find a job	0.327	0.231		0.256	0.217	0.303	0.355	0.089	**0.402**	0.341
Prob of passing the degree	0.046	0.125	0.256		-0.036	0.084	0.446	0.199	0.225	**0.584**
Expected study hrs/wk	0.441	0.312	0.217	-0.036		**0.511**	0.188	-0.070	0.400	0.165
Perceived status for degree	**0.61**	0.299	0.303	0.084	0.511		0.415	0.010	0.561	0.346
Prob enjoy job (age 30)	0.304	0.203	0.355	0.446	0.188	0.415		0.249	0.497	**0.694**
Prob work-life balance (age 30)	-0.041	-0.141	0.089	0.199	-0.070	0.010	**0.249**		0.069	0.198
Parental approval	0.444	0.286	0.402	0.225	0.400	**0.561**	0.497	0.069		0.471
Prob of enjoying coursework	0.248	0.171	0.341	0.584	0.165	0.346	**0.694**	0.198	0.471	

Note: Table reports correlation between the standardized expectation variables. The correlations are duplicated in the upper and lower triangles of the matrix for ease of comparison. Bold denotes the highest correlation for each variable, going row-by-row. This differs from column-by-column. The second panel is similar, but using variables for which individual fixed effects (means) have been removed.