Pre-Analysis Plan: Using Revealed Choice Data with Subjective Expectations

June 18, 2019

Introduction

Below, we first present the baseline choice model which we have used in Angelov et al. (2019).[1] When writing this existing working paper, we only had access to the data collected as part of our 2014 survey, and Swedish administrative register data through the end of 2014. This pre-analysis plan lays out our intended follow-up research, to be conducted with administrative data from 2014-2018, which we have applied for, but not yet examined. With the baseline model as a starting point, we then lay out the hypotheses we plan on testing with the new data, along with the corresponding empirical specification. To this end, assume that each individual ($i = 1, ..., N$) chooses the most preferred field of study among $j = 1, ..., J$. The chosen field of study is the one maximizing the student's expected utility. The utility for individual i from choosing alternative j is specified as:

$$U_{ij} = \gamma_j + \boldsymbol{\theta}' \boldsymbol{x}_{ij} + \varepsilon_{ij} \tag{1}$$

where γ_j are 8 potential field of study indicators (with $\gamma_j = 0$ for one of the j for identification); $\boldsymbol{x}_{ij}$ represents the amenities which vary between individuals and alternatives; and the random component ε_{ij} is i.i.d. drawn from a certain distribution (for instance Gumbel, leading to the logit model). The amenities can be grouped as $\boldsymbol{x}_{ij} = (y_{ij}, \boldsymbol{a}_{ij})$ where $\boldsymbol{a}_{ij}$ denotes the students' expectations about amenities associated with varying potential field of study choices, such as perceived status and probability enjoying the job associated with each study choice, and y_{ij} denotes expectations about earnings associated with varying potential field of study choices. Table 1 reports γ_j and $\boldsymbol{\theta}'$ from the stated preference data, as estimated using conditional logit.

The existing working paper is focused solely on the students' stated expectations and preferences over fields as they reported in their last semester of high school. At the end of the survey we randomized the sample of students into an information experiment. After the students listed their expectations, and ranked the fields of study, they were randomly allocated into treatment ($T_i = 1$) or control ($T_i = 0$) . The treated students where asked if they wanted to receive information on the actual average earnings for individuals with the different types of education. The subset of treated students who agreed were then shown the earnings displayed in Table 2. As shown in the table, they were also told the no-college earnings, and the percent premium each type of college field represented relative to the no-college earnings.

[1] Angelov, Nickolay, Per Johansson, Mikael Lindahl, and Ariel Pihl. "Subjective Expectations, Educational Choice Heterogeneity, and Gender: Evidence from a Sample of Swedish High School Students." June 2019. *Mimeo.*

The data that was used to in our working paper, Angelov et. al. (2019), only goes through the end of 2014. Since it is very common for Swedish students to delay enrolling in university, this was insufficient to examine the impact of the information experiment, or to properly compare the surveyed preferences during high school to the students ultimate actions. We have cross-tabulated the top-ranked field of application with top ranked field from the survey for the subset of students who applied to college immediately after high school. We have not run any regressions using this data, and only 49% of our sample applied to a degree program in this period. This cross-tabulation is reported in Table 3.

We have requested data from Statistics Sweden covering the students' college applications and educational and labor market outcomes (from 2014-2018). We consider the students' college applications and college enrollments to reflect their revealed preferences, compared to stated preferences which they reported during the survey. The data was delivered to the administering organization, IFAU, while we were in the process of formulating this document. We therefore gave instructions to IFAU to delay uploading the data to the analysis server until after this document was registered. Hence, we will not be able to access the data until after the registration of this pre-analysis plan. With this data we would like to answer two broad questions:

1. Do students' revealed preferences differ from their stated preferences? In other words, do they place different weights on the $\boldsymbol{x_{ij}}$ amenities when they apply to colleges in period $g = 2$, than they had when stating their preferred fields of study in a survey format in period $g = 1$?

2. Did the experimental intervention have any impact on the students' choice of field. In other words, did it affect which programs they applied to in $g = 2$?

1 Simple estimation of "Question (1)"

We surveyed the students before they made their college applications. We will now be able to follow them over the subsequent four years to observe their revealed choices in applications to college programs, enrollment, and, in some cases, matriculation. This combination of survey and longitudinal register data provides an excellent environment to compare stated and revealed preferences. In equation (1) above, the $\boldsymbol{\theta}$ capture the students' stated preferences over $\boldsymbol{x_{ij}}$, and γ_j capture some population-wide preferences for particular fields that are not captured through $\boldsymbol{x_{ij}}$. We are interested in whether the relationship between $\boldsymbol{x_{ij}}$ and the choice they make in $g = 2$ differs from the corresponding relationship in $g = 1$.

We can imagine testing this in the following way:

$$U_{ijg} = \gamma_j + \delta_j \mathbb{1}(g = 2) + \boldsymbol{\theta}_1' \boldsymbol{x}_{ij} + \boldsymbol{\theta}_2' \boldsymbol{x}_{ij} \mathbb{1}(g = 2) + \varepsilon_{ij}, \tag{2}$$

where γ_j and $\boldsymbol{\theta}_1'$ are the stated preferences from equation (1). Thus, if $\delta_j = 0$ and $\boldsymbol{\theta}_2' = 0$ then stated preferences and revealed preferences are the same.

An added complication is the information experiment. Half of the individuals in our sample were offered the opportunity to receive information on average earnings for each field of study. If the experiment impacted the students' preferences, this needs to be controlled for in the estimation. The simplest way to avoid this is to estimate equation (2) only on the untreated subsample of students ($T_i = 0$, $N = 238$). If we find that the experiment has no effect (next section) we can include the treated individuals in this estimation.

2 Simple estimation of "Question (2)"

The experiment was designed to see if providing students with accurate wage information affects which fields they choose. One of the simplest ways we can see if the experiment had an impact is by focusing on changes in first-ranked field of study between the two periods, and testing if changing is more likely for the treated group. We can summarize this with the following equation:

$$DE_i = \alpha + \beta_1 T_i + \varepsilon_{ij}, \tag{3}$$

where the data has been reduced to one observation per person. Here DE_i is an indicator variable for whether the student's most highly ranked field in $g = 2$ differs from the one they stated in $g = 1$. Only 149 of the 260 treated students agreed to receive the additional information (we can denote them $Info_i = 1$). We will also estimate the treatment effect on the treated, by replacing T_i with $Info_i$ in equation (3), and using T_i as an instrument for it.

It is reasonable to expect $\beta_1 \geq 0$. Under the null hypothesis, the experiment has no impact ($\beta_1 = 0$), and the alternative is that people who received the additional information would be more likely to change their first choice than those who do not ($\beta_1 > 0$).

If we want to allow more nuance in the impact of the experiment, it is plausible that the wage information will motivate students to switch towards fields in which they under-estimated the returns, and away from fields in which they over-estimated the returns. The following equation can be used to investigate this:

$$\Delta E_{ij} = \alpha + \beta_2 (y_j - y_{ij}) T_i + \varepsilon_{ij}, \tag{4}$$

Here there are J observations per person, and ΔE_{ij} is the difference between the indicator for a field being ranked first in college applications, and an indicator for a field being ranked first in the survey.[2] y_j is the average earnings for field j that the student was shown in the experiment. This estimation model assumes that the impact of the experiment on students' choices will be proportional to how different the information we provided them is relative to the wage that they expected. Here again, we expect that if the experiment affected them, it will be in the direction of $\beta_2 > 0$.

If we want to mirror the conditional logit regressions summarized in equations (1) and (2) we can remove the differencing in the outcomes and return to a stacked data set:

$$\begin{aligned} U_{ijg} =& \mu_1 (y_j - y_{ij}) \mathbb{1}(g = 2) T_i \\ &+ \mu_2 (y_j - y_{ij}) \mathbb{1}(g = 2) + \varepsilon_{ijg}. \end{aligned} \tag{5}$$

This results in a more traditional difference-in-difference framework which can be estimated using conditional logit. μ_1 gives the impact of the difference in expectation of income vs. the information as a result of the experiment. The second term controls for any changes in decisions over time that happen also for the control group. We do not include a control for being in the

[2] So if the student ranked social sciences (ss first in $g = 2$) and education (ed first in $g = 1$, then $\Delta E_{i,ss} = 1$, $\Delta E_{i,ed} = -1$ and for the other fields $\Delta E_{i,j} = 0$. If the student chooses the same field as they said they would then the whole vector is 0.

treated group, because (a) it is randomly assigned, and (b) since it varies only at the individual level the conditional logit cannot identify it.

3 Joint estimation

Above, we have described simple estimations that will separately tackle the two questions we are interested in. However, we can also estimate a joint model that accommodates both questions:

$$\begin{aligned} U_{ijg} =& \gamma_j + \delta_j \mathbb{1}(g=2) + \boldsymbol{\theta}_1' \boldsymbol{x}_{ij} + \boldsymbol{\theta}_2' \boldsymbol{x}_{ij} \mathbb{1}(g=2) \\ &+ \boldsymbol{\theta}_3' \boldsymbol{x}_{ij} \mathbb{1}(g=2) T_i + \mu_1 (y_j - y_{ij}) \mathbb{1}(g=2) T_i \\ &+ \mu_2 \mathbb{1}(group)(y_j - y_{ij}) T_i \mathbb{1}(g=2) + \varepsilon_{ijg}, \end{aligned} \tag{6}$$

where $T_i = 1$ if student i was randomized into the information treatment, and $T_i = 0$ otherwise. Here, γ_j and $\boldsymbol{\theta}_1'$ correspond to the stated choice parameters while δ_j and $\boldsymbol{\theta}_2'$ ideally identify preference changes between the survey and the actual choice. Although we acknowledge that students may update their expectations about amenities (i.e., $\boldsymbol{x}_{ij}$) between the time when the questionnaire was administered and the actual choice, we have no means of collecting new data on $\boldsymbol{x}_{ij}$ which suggests that it is better to use revealed preferences as close in time to the questionnaire as possible. This in turn implies that it would be better to rely on applications rather than actual enrollment. At this stage, we are not ready to decide exactly how we will choose to define revealed preferences in our most preferred specification, but will be transparent about the motives for our choice. In any case, we will present results using both application and enrollment.

Furthermore, $\boldsymbol{\theta}_3'$ captures potential preference changes as a consequence of being treated, regardless of the content of the information received in the treatment. The effect of the information content on the revealed schooling choice is instead measured by μ_1, which we expect to be positive. Note that the stated preferences by definition cannot be altered by the treatment, since the treatment was the final step in the survey, and there was no technical possibility to alter answers retroactively. The final term in equation (6) captures potential heterogeneity in the effect of new information on wages on the schooling choice (μ_2). From existing research we expect that low-SES students may be more impacted by the change, and that women and men may respond differently.

To this end, we plan on testing the following hypotheses:

$H_0^1 : \mu_1 = 0$, $H_1^1 : \mu_1 > 0$

H_1^1: In response to the information treatment, students increase (decrease) the probability of choosing a field at $g = 2$ proportionately to how much they under (over) estimated the earnings.

$H_0^2 : \boldsymbol{\theta}_2 = 0$ (joint test)

H_0^3 : Stated ($g = 1$) and revealed ($g = 2$) preferences ($\boldsymbol{\theta}$) over $\boldsymbol{x_{ij}}$ are the same.

$H_0^3 : \boldsymbol{\theta}_3 = 0$ (joint test)

H_0^4 : The treatment does not impact preferences (except through $(y_j - y_{ij})$).

$H_0^4 : \mu_2$

If *group* is low-SES: $H_0^4 : \mu_2 = 0$, $H_1^4 : \mu_2 > 0$. Low-SES students are more likely to change their major in response to the information.

If *group* is female: $H_0^4 : \mu_2 = 0$. Women and men respond to the information supplied in the same way.

Data

3.1 Students who do not apply to university

Of the students who participated in our initial survey, only 5% reported that they did not intend to apply to college. Although all the sample was asked to rank not going to college along side the choices of college major, they were only asked about their expectations on earnings (not the other amenities) because we judged "no-college" too broad a category. Therefore, in Angelov et. al. (2019) we remove the rank for no-college from all students' choice sets, and further drop the small sample who ranked it first.

There may be a larger share who we do not observe applying to university by the end of 2018.[3] If this choice is related to the x_{ij} or receiving the information on income, not including these individuals would introduce bias into our estimates. As a first check of this we can estimate the impact of x_{ij} and T_{ij} on the probability that the individual is not observed applying to college in $g = 2$. For T_{ij} this is as simple as replacing DE_i in equation (3) with an indicator for whether we observe the individual applying to college ($NoCollege_i$). For x_{ij} we can test if the γ_j and $\boldsymbol{\theta}$ are different between the $NoCollege_i = 0$ and the $NoCollege_i = 1$.

One way of including these individuals in the estimation is by thinking of not applying as the outside option. In equation (3), since neither the x_{ij} nor y_j are included, we can include the non-college ranks from the survey. Those who stated they would apply to any university field, but did not, would have $DE_i = 1$. The same follows for (4) and (5), where we can think of not going to college as being among the fields j.

Equations (2), (5) and (6) are formulated as utility models. In practice, these are estimated using conditional logit regressions. These estimations use the information both for the students' chosen fields of study (in each period), and the ones they do not. The non-chosen fields are essentially treated as being tied among each other. The outcome we use to proxy for utility is a binary indicator for having chosen the given field in the given period, over the alternative options. We can call this outcome C_{ijg}.

$$C_{ijg} = \mathbb{1}(1stChoice_{ig} = j) \quad \forall g \in \{1, 2\} \tag{7}$$

for each of the eight j field of college study choice. Equations (2) and (6) include the x_{ij}, most of which were not collected for the no-college option. Thus we must remove individuals whose first choice in the survey is no-college and not include no-college as a j. However, we can still think of not applying to college as changing one's mind relative to $g = 1$. Thus, if the student doesn't have a first choice of field of study in $g = 2$ then $C_{ij2} = 0$ for all j. This binary variable (there will be 16 observations per individual) is used as the outcome when estimating the conditional logit regressions.

[3] In 2016, roughly 56% of Swedish 24 year olds had not begun higher education studies. In our data, this group will be a combination of students who will eventually apply to college, those who went abroad for college, and those who will never go to college.

3.2 Outcomes of Interest

Our eight categories of field of study are defined by the Swedish classification of education, so it is straightforward to map all possible majors within Sweden to the same categories. In the four years since they graduated high school, we expect most of our surveyed sample to have applied to university, and enrolled. Typical bachelors programs are three years long, so many will also have graduated.

In Sweden, students submit a ranked list of major-university choices. They must meet basic qualifications (courses taken), and then competitive programs are offered based on their high school performance. Students can also apply to courses, for example if they want to try something out.

Although they can in theory rank many fields of study, in practice most students focus on one or two fields of study and apply to multiple narrow majors and universities within that field. Thus, it makes sense to focus on first-choice field only, and use conditional logit or another binary-outcome method. The following two choices of the student will be used to create C_{ij2}.

Field of first-ranked degree program in college application: The application to programs is the first revealed choice the students make. We will focus on applications to degree granting programs. We will ignore rankings of courses.

Field of first enrolled degree program: The students may be rejected from their top-ranked programs of study, or choose not enroll in a program that they were admitted to. This means that the top ranked application program and where students ultimately enroll can differ. Enrollment is interesting because it combines what the student wants with what it feasible given their grades, coursework, and commitment.

4 Possible Additions

This document describes hypotheses that we will test and report in the resulting paper. However, if our hypotheses are shown to be incorrect, we could possibly extend the analysis to understand why. In particular, we think that information from the whole college application (which include information on field-of-study and university choice ranked lower than first), may help us examine mechanisms for how and why our main hypotheses are violated.

The rest of the application rankings will show how committed a student is to a particular field of study, versus to a particular university or region of Sweden. This may be especially helpful if hypothesis H_0^2 is rejected.

5 Tables

Table 1: Coefficients from stated preference model of college major choice.

	Rank	First Choice		
	(1) All	(2) All	(3) Male	(4) Female
Mean expected earnings 30-40	0.181**	0.136	0.270	0.0240
	(0.0456)	(0.109)	(0.173)	(0.138)
Expected hrs/wk (age 30)	0.0201	0.0448	0.0116	0.0388
	(0.0438)	(0.103)	(0.173)	(0.122)
Prob find a job	0.110**	0.158	0.280*	-0.00194
	(0.0398)	(0.109)	(0.136)	(0.176)
Perceived status for degree	0.255**	0.179+	0.0422	0.271+
	(0.0460)	(0.106)	(0.152)	(0.155)
Prob enjoy job (age 30)	0.580**	0.910**	1.232**	0.625**
	(0.0527)	(0.148)	(0.216)	(0.198)
Prob work-life balance (age 30)	0.0766*	0.172	0.282*	0.0656
	(0.0382)	(0.107)	(0.142)	(0.173)
Prob of passing the degree	0.291**	0.387*	0.593**	0.159
	(0.0555)	(0.165)	(0.230)	(0.226)
Expected study hrs/wk	-0.00738	0.107	0.201	0.00569
	(0.0578)	(0.135)	(0.190)	(0.204)
Parental approval	0.270**	0.433**	0.303	0.577**
	(0.0557)	(0.148)	(0.221)	(0.195)
Prob of enjoying coursework	0.589**	0.921**	0.760**	1.142**
	(0.0570)	(0.165)	(0.244)	(0.221)
Pedagogy	-0.382**	-0.616*	-0.622	-0.656
	(0.0946)	(0.270)	(0.381)	(0.407)
Humanities and Art	-0.624**	-0.360	-0.315	-0.391
	(0.109)	(0.232)	(0.341)	(0.303)
Social Science	(.)	(.)	(.)	(.)
Science and Math	-0.289**	-0.0749	-0.176	0.0370
	(0.0917)	(0.173)	(0.243)	(0.249)
Tech and Engineering	-0.309**	0.119	0.0967	0.0941
	(0.0880)	(0.176)	(0.239)	(0.297)
Agro and Animal	-0.666**	-1.182**	-1.463**	-0.951*
	(0.0990)	(0.299)	(0.467)	(0.416)
Healthcare	-0.450**	-0.663**	-0.798**	-0.469+
	(0.0917)	(0.188)	(0.305)	(0.242)
Services	-0.453**	-0.598*	-0.300	-0.953*
	(0.0958)	(0.271)	(0.393)	(0.401)
Pseudo R^2	0.244	0.425	0.442	0.427
N	3808	3808	1928	1880

Notes: Coefficients and field specific intercepts using only the stated preference (survey) information. Columns 1 uses the full rankings of fields, columns 2-3 are estimated using conditional logit comparing the first-ranked field to the remaining seven. The amenities have been standardized to mean 0, standard deviation 1. The two variables measuring hours (work hours and study hours) have been reversed such that higher values correspond to lower hours spent. Standard errors in parentheses (+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table 2: Intervention table with translations: Average earnings by field of study

English Translations:	Education	Average salary /mo before tax	%-difference relative to high school
	Utbildning	**Genomsnittslön /mån före skatt**	**%-skillnad jämfört med gymnasium**
High School Only	Gymnasium	26,361 kr	
College education (bachelors degree)	*Högskoleutbildning (grundexamen)*		
Teaching and Education	Pedagogik och lärarutbildning	27,942 kr	6.0%
Humanities and Art	Humaniora och konst	28,955 kr	9.8%
Social Science, Law and Business	Samhällsvetenskap, juridik, handel, administration	37,642 kr	42.8%
Science, Math and Data	Naturvetenskap, matematik och data	34,515 kr	30.9%
Engineering and Technology	Teknik och tillverkning	39,924 kr	51.5%
Agriculture, Forestry and Vet	Lant- och skogsbruk samt djursjukvård	33,726 kr	27.9%
Health and Social work	Hälso- och sjukvård samt social omsorg	32,407 kr	22.9%
Services	Tjänster	33,564 kr	27.3%

Källa: Lönestrukturstatistiken och LISA, SCB (2010). Lön uttryckt i 2013 års priser. *Translation: Source: Salary Tables and LISA, SCB (2010). Salaries in 2013 terms.*

Notes: Students were only shown the Swedish.

Table 3: Cross-tabulation of surveyed 1st choice and 1st choice of college application in 2014

	First ranked field in 2014 survey:								
	Ed	Hum	Soc	Sci	Tech	Agro	Health	Serv	Total
First Ranked Program in College Application:									
STEM prep year	0	1	0	4	2	0	2	0	9
Education/Teaching	3	0	0	1	0	0	0	0	4
Humanities and Art	0	7	1	0	0	0	0	0	8
Social Science	4	3	53	9	3	1	8	2	83
Science and Math	0	0	4	7	0	1	0	0	12
Tech and Engineering	0	1	5	27	40	1	1	1	76
Agro and Animal	0	0	0	0	0	2	0	0	2
Healthcare	1	2	3	6	2	0	22	0	36
Services	0	0	1	0	0	0	0	0	1
Total	8	14	67	54	47	5	33	3	231

Notes: Columns show the first ranked field in the survey responses, rows show the top-ranked program among those students who applied to college programs for Fall 2014. "STEM prep year" is a short college program that allows students (for example those who completed non Science-track high school programs) to fulfill the entry requirements for STEM college majors.