TI 2020-030/III
Tinbergen Institute Discussion Paper

Educational Choice, Initial Wage and Wage Growth

Jacopo Mazza[1]
Hans van Ophem[2]

[1] University of Essex
[2] Amsterdam School of Economics, University of Amsterdam and Tinbergen Institute

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and Vrije Universiteit Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at https://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Educational Choice, Initial Wage and Wage Growth

Jacopo Mazza

University of Essex

Hans van Ophem *

Amsterdam School of Economics, University of Amsterdam

and Tinbergen Institute

Spring 2020

Abstract

We investigate the major choice of college graduates where we make choice dependent on expected initial wages and expected wage growth per major. We build a model that allows us to estimate these factors semiparametrically and that corrects for selection bias. We estimate the model on the combined NLSY79 and NLSY97 samples. We find markedly different results in expected real wage growth and expected initial wages across majors. Furthermore, the differences in these expectations appear to be relevant for major choice.

Keywords: Wage inequality; Wage uncertainty; Unobserved heterogeneity; Selection bias; Decision-making under Risk and Uncertainty; Semiparametric estimation.

JEL classification: C14, C34, D81, J31

*Corresponding author at the Amsterdam School of Economics, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. Email: j.c.m.vanophem@uva.nl. All data and computer programs are available on request.

1 Introduction

The field of study in college is a key determinant of future earnings and earning differences among college graduates. Altonji et al. (2012) estimates that, in the US, the difference in earnings between male electrical engineers and male general education graduates is nearly as large as the difference in earnings between high-school and college students. Also the evolution of age earning profiles varies widely between majors. In a recent study, Deming and Noray (2018) document the different life-cycle returns of STEM, i.e. Science, Technology, Engineering and Mathematics, majors compared to non-STEM majors. They find that STEM graduates earn substantially more at the beginning of their career, but experience a slower wage growth in the first years of their working life.

The economic consequences of students' choices of a field of study in college are large, and the field of study also influences how these earnings are distributed throughout the life cycle. But do students take these factors into account when deciding on their future college career? More specifically, are expected initial earnings driving choices of college majors? And are expected age earning profiles important in the decisions to specialize in one major instead of another? In this paper we address these questions. In particular, we estimate how expected differences in earnings across fields of study affect major choice and we disentangle the effect of starting wages and wage growth for this choice using US data.

Economists have always been interested in understanding what drives investment decisions in human capital. This line of inquiry can be traced back directly to Adam Smith and in its modern conception to the seminal works of Mincer (1958, 1962), Becker (1975) and Willis and Rosen (1979). The attention is hardly surprising considering that for many individuals this investment is one of the largest that will ever be undertaken and its impact will profoundly shape their future life, career, and well-being.

The literature studying how expectations about future earnings affect decisions on investments in further years, or levels, of education (Willis and Rosen, 1979; Keane and Wolpin, 1997; Belzil and Hansen, 2002; Kaufmann, 2014) is fairly large and established. What drives selection into types of education, instead, is less examined. Studies considering the determinants of choices of types of education are recent (Arcidiacono, 2004; Arcidiacono et al., 2012; Beffy et al., 2012; Wiswall and Zafar, 2015; Altonji et al., 2016), and have not yet reached a consensus. Some argue for the primary importance of monetary considerations (Arcidiacono et al., 2012; Altonji et al., 2016), while others emphasize the role of taste for a particular field or other non-pecuniary factors (Arcidiacono, 2004; Beffy et al., 2012; Wiswall

and Zafar, 2015).

A common problem when studying the determinants of choices - of occupation, level of education, or type of education as in our case - is the lack of data. The econometrician can only observe the earnings of the chosen alternative, but the revealed choice needs to be compared to counterfactual outcomes for the other available options. To address this issue, the applied literature on schooling choices has resorted to two strategies: either directly elicit students' subjective expectations about future pay-offs from surveys (Arcidiacono et al., 2012; Stinebrickner and Stinebrickner, 2012; Zafar, 2013; Kaufmann, 2014; Wiswall and Zafar, 2015) or to assume rational agents who are utility maximizers and whose preference can be inferred from the choice data (Siow, 1984; Arcidiacono, 2004; Beffy et al., 2012). Both approaches impose assumptions and come with limitations.

A research design based on subjective expectations has two major drawbacks. First, these studies usually collect only information on expected wages at one particular future point in time, so that it is impossible to consider differences in the progression of wages in time [1]. Second, the timing of the collection of data is relevant for asking about expectations after choices have been made might bias the results. Some studies interview students still in high-school (Jensen, 2010; Zafar, 2013; Kaufmann, 2014); others interview former college students after graduation (Webber, 2014; Ruder and Van Noy, 2017), and some others a combination of these two groups (Arcidiacono et al., 2012; Wiswall and Zafar, 2015). But, especially when collected after the actual decision is made, subjective expectations can be endogenous (Bertrand and Mullainathan, 2001; Bound et al., 2001; Benitez-Silva et al., 2004; Zafar, 2013; Kaufmann, 2014) as individuals might try to rationalize their past or future choice. In this case, these studies are not eliciting students' expectations, but their rationalizations of a choice already made. This is the endogeneity that researchers should be careful about as it is likely to introduce a serious measurement error leading to biased estimation results.

Studies that adopt a more traditional revealed preference approach (Arcidiacono, 2004; Beffy et al., 2012; Webber, 2014), instead, try to determine the relevant factors of the choice process from the observed data. These models require assumptions on how students form their expectations, which are usually modeled as myopic or rational, both for the chosen and the counterfactual options. The disadvantage of this methodology is clear: as the

[1] A notable exception is Wiswall and Zafar (2015) who collected information on expected wages at three future moments.

econometrician only observes the revealed choice, selection bias needs to be addressed because of the endogeneity of educational choices. Or to put it differently, the counterfactuals are not observed and need to be created from the model and that is only possible under relatively strong assumptions. A clear advantage of using revealed preference information is that more information is available. For example, if panel data are available, the econometrician can observe the full evolution of age-wage profiles throughout the working life for the revealed choice.

In this paper we assess the elasticity of major choices to initial wages and wage growth rates starting from observed wage data. We exploit the panel structure of the NLSY to obtain consistent estimates for the two wage components and create meaningful counterfactual scenarios. A considerable advantage of this data is that it allows us to account for time-invariant unobserved heterogeneity that might drive both the choice of major and the personal life-cycle earning profile, without having to rely on exclusion restrictions.

We estimate the effect of differences in initial wages and wage growth with four procedures. The first two, logit and probit, are the workhorses for estimating multinomial choice models. These are parametric models that impose quite stringent and unattractive assumptions on the error structure. For these reasons, these techniques have come under closer scrutiny and are receiving growing criticism. In reaction to these mounting criticisms a series of semiparametric techniques have been proposed (Manski, 1975; Lee, 1995) in the theoretical literature, but their application has been very limited. The only exception that we are aware of is Dahl (2002) who proposes a two-step semiparametric method correcting for sample selection bias in the case of multiple possible outcomes, for the estimation of migration probabilities between the US states. This application though is very specific to the research question analyzed in that paper. In this paper we implement the semiparametric estimators of Manski (1975) and Lee (1995) that require only minimal distributional assumptions. The only distributional assumption on the error structure that we impose is the common one lying at the basis of any panel data estimation which serves us to estimate wage expectations corrected for selectivity for all possible choice alternatives. In a second step we use the corrected wage expectations as an explanatory factor in the major choice equation. Finally, we estimate this equation both parametrically and semiparametrically.

Our results show that initial wages and wage growth differ considerably across majors. The Health cluster is the one showing higher initial wages, whereas Education and Humanities the lowest. Major choice is determined by expected future wages as majors that offer

better wages in the future are preferred. Even if we find stronger evidence for an effect of initial wages on major selection, also wage growth influences major selection positively at least in the parametric models. To illustrate, increasing the mean initial wage and mean wage growth both by one standard deviation will increase the probability of choosing the Social Science major by 23% and as much as 64% for the Humanities major.

This paper makes three contributions. First, we introduce new evidence on the importance of financial pay-offs in the choice of field of study in college. We believe this to be important for at least two reasons. The first has to do with how economists think about, and model, individual decisions on human capital investments. A cornerstone of the enormous empirical literature on human capital initiated by Mincer (1958, 1962) and Becker (1975) is the maximizing agent. Rational people choose their education level, or their education type, by comparing the available alternatives and selecting the one that grants the highest expected value. One of our aims is to test this assumption. The second has to do with the implications of results for policies to address the shortages of skills - usually scientific ones - in the labor market that is often decried in the public debate. If students are insensitive to monetary returns of college majors, financial incentives as a solution to shortages will be ineffective.

The second contribution is to disentangle the separate effect of initial wages and wage growth on college major selection. This is a useful relaxation of the standard assumptions and it improves the understanding of the mechanics of educational choice formation.

The third contribution is to illustrate an application of semiparametric estimation methods for polychotomous choice models with panel data. Given the clear and well-understood limitations of standard parametric techniques, one would wish to see more applications of these class of estimators to unordered choice models. This has not been the case so far. We believe that one possible explanation for this lack of applications could reside in the heavy computational burden that these techniques bring about. In our application, we find the global optimum of semiparametrically estimated polychotomous choice model to be very hard to find.

2 Major choice

In this section we lay out a simple model for educational choice. The model takes a standard Roy model approach as a starting point for multiple and unordered educational choices in

which rational students are utility maximizers that need to choose among five major categories[2]. We focus on students who have completed college education only. After graduating from high school, these students have to decide on the major they want to pursue in college. We call this point in time $t = t_{<0}$. It lies before the individual starts to work ($t = 0$), but the exact timing is not specified. We distinguish five major categories: Natural Sciences ($m_i = 1$); Social Sciences ($m_i = 2$); Humanities ($m_i = 3$); Education ($m_i = 4$) and Health ($m_i = 5$), where the subscript i indicates a specific individual. After graduation from college, people start working on the labor market and a stream of income is expected for T periods.

When choosing the favorite major category, each individual compares the benefits obtainable in the five educational categories and opts for the utility maximizing one, with utility being a function of the expected lifetime earnings as perceived by the individual at $t = t_{<0}$, $h(E(Y_{mi0}), E(Y_{mi1}), .., E(Y_{miT}))$, where Y_{mit} is the income of individual i at time t if opted for major m. By adding an error term, ξ^*_{mi} for each major $m = 1, 2, 3, 4, 5$ and for each individual $i = 1, .., N$ we can specify the following utility function:

$$U_{mi} = h(E(Y_{mi0}), E(Y_{mi1}), .., E(Y_{miT})) + \xi^*_{mi}, \tag{1}$$

To assess individually expected wages we need to rely on observed wages during working life. We are now faced with three problems:

- How do individual expectations relate to economic reality i.e. wage observations?
- All the wages we observe are conditional on the optimal choice made on $t = t_{<0}$. This will give rise to a selectivity bias and therefore corrections need to be made.
- We observe only the wage of the optimal choice and not the counterfactual wages.

We now discuss how we tackle each issue.

2.1 The wage equation

After the educational choice has been made and after graduation, the individual starts working and wages are observed for several periods. The starting wage is the wage observed in the first and we model it as follows:

[2]The choice of these five college major categories is fairly standard in the literature. Many of the college major groups coded in the NLSY count little to no observations, thus some aggregation is necessary for the statistical analysis. How these major categories were precisely created from the NLSY classification is available on request. Also see the appendix.

$$log(y_{mi0}) = \beta'_m x_{i0} + \varepsilon_{mi0} = log(\tilde{y}_{mi0}) + \varepsilon_{mi0} \tag{2}$$

Note that in this specification there is no time dimension. Obviously, only $t = 0$ is relevant here. All wages earned after the starting period contribute to form the age-wage profile. We model the later period individual wages as follows:

$$y_{mit} = y_{mi0} e^{\rho_{mi}(t)} \qquad t > 0 \tag{3}$$

where $\rho_{mi}(t)$ is a time varying growth rate of wages specific for individual i and major m. This growth rate is approximated by a Kth order polynomial of time:[3]

$$\rho_{mi}(t) = \rho_{mi0} \left(\sum_{j=1}^{K} \alpha_{mj} t^j \right) + \varepsilon^*_{mit} \qquad t > 0 \tag{4}$$

This functional form of individual wage growth, allows for the empirical observation of a concave function of wages in time, initially increasing but at a diminishing rate and potentially decreasing for large t. In our specification such a functional form is only possible when $K > 1$. By using a Kth order polynomial a large number of functional forms can be approximated. Although we expect that $\rho_{mi0} > 0$, indicating that the initial growth rate of wages ($t = 0$) is positive for every individual and major choice, positivity needs not be the case for all growth rates. Substituting (4) in (3) we obtain:

$$y_{mit} = y_{mi0} e^{\rho_{mi0}\left(\sum_{j=1}^{K} \alpha_{mj} t^j\right) + \epsilon^*_{mit}} = \tilde{y}_{mi0} e^{\rho_{mi0}\left(\sum_{j=1}^{K} \alpha_{mj} t^j\right) + \varepsilon_{mit}} \qquad t = 1, 2, ..., T_i \tag{5}$$

where y_{mit} is the individual wage received at moment t if major choice m is made and $\varepsilon_{mit} = \varepsilon^*_{mit} + \varepsilon_{mi0}$, a zero mean error term. Taking logarithms this can be written as:

$$log(y_{mit}) = \beta'_m x_{i0} + \rho_{mi0} \left(\sum_{j=1}^{K} \alpha_{mj} t^j \right) + \varepsilon_{mit}. \tag{6}$$

The wage equation (6) is different for each major as reflected by major specific initial wage, growth rates and error structure. In the panel data literature it is common to specify the

[3]Note that we do not add a constant to the polynomial. The reason for this is that we need $\rho_{mi}(0) = 0$ so that $y_{mit} = y_{mi0}$ if $t = 0$.

error structure as follows:[4]

$$\varepsilon_{mit} = \varepsilon^*_{mit} + \varepsilon_{mi0} = e_{mi} + \varsigma_{mit} + \varepsilon_{mi0}. \tag{7}$$

This error structure consists of individual fixed effect e_{mi}, and an idiosyncratic term ς_{mit}. For the error term of (2) we make an equivalent assumption:

$$\varepsilon_{mi0} = \tilde{e}_{mi} + \varsigma_{mi0}. \tag{8}$$

Note that we distinguish two individual fixed effects. The reason for this is that ε_{mi0} contains ε^*_{mit}, and both have a different origin: ε_{mi0} stems from the starting wage equation whereas ε^*_{mit} relates to the wage growth equation. Effectively, we allow the polynomial approximation of the growth rate to have an individual fixed effect of its own, although we will impose a direct relation later.

An important problem is that wages differ across major and are only observed for the utility-maximizing major choice. As a result, the error terms of the major choice equation (ξ^*_{mi}) and the wage equation (ε_{mit}) are likely to be correlated due to self-selection and as a result, estimating the wage equations in (6) with OLS will result in biased estimates.[5] As in Chen (2008) and Mazza and van Ophem (2018), we will assume that there is no statistical relation between ξ^*_{mi} and ς_{mit}. The expected value of future wages given that the individual has chosen $m_i = m, m = 1, 2, 3, 4, 5$, is given by:

$$E(log(y_{mit})|m_i = m) = \beta'_m x_{i0} + \rho_{mi0}\left(\sum_{j=1}^{K} \alpha_{mj} t^j\right) + E(e_{mi} + \tilde{e}_{mi}|m_i = m) \quad t = 1, ..., T_i \tag{9}$$

$$E(log(y_{mi0})|m_i = m) = \beta'_m x_{i0} + E(\tilde{e}_{mi}|m_i = m)$$

We now relate the private information the individual possesses to the individual fixed effects in the following way:

[4] We do not include time specific fixed effects to avoid multicollinearity. As we already allow for a flexible time pattern of wage growth using a high degree polynomial, adding year dummies to the specification will pick up a considerable part of the time pattern.

[5] Another reason for a bias is that the fixed effects might correlate with the regressors of the wage equation.

$$e_{mi} = \gamma_m \nu_i. \tag{10}$$

The scalar ν_i is not observed and due to the presence of γ_m, we can assume that ν_i has unit variance. Furthermore we assume that:

$$\tilde{e}_{mi} = \tau_m e_{mi} = \tau_m \gamma_m \nu_i. \tag{11}$$

What this means is that we assume that the unobserved abilities, interests, motivation as combined in ν_i can be important for wages, both for the initial wages and the growth rates, and that we allow for differences across majors.

2.2 The estimation of wages

Our aim is to estimate the major choice as faced by college students. Students maximize utility and this utility, as reflected in (1), depends on expected future wages and unobserved personal characteristics. As specified in the previous subsection, wages in time are characterized by an initial wage (y_{mi0}) and a growth rate ($\rho_{mi}(t)$). The individual has to form expectations on these factors using the individual information available. Part of this information is observable, but another part is not observed but known to the individual (ν_i). In this subsection we show how the parameters of the model described previously can be estimated.

Wages are only observed given the educational choice made by the individual and, as a result, we need to correct for this potential selectivity. Under the assumptions made, the element in the wage equation (6) causing the problem are the fixed effects e_{mi} and $\tilde{e}_{mi}$. Disregarding for the moment the initial wage, the fixed effect e_{mi}, and consequently the selectivity problem, can be removed from the equation by taking difference across the mean in time, i.e. the usual within transformation in panel data models, cf. Hsiao (1986) or Baltagi (2013). Alternatively, a first difference estimator can be used, but it is somewhat less efficient. Due to the growth rate dependence on time, it is more convenient in the present case to subtract the previous individual observation:[6]

[6]We avoid using the term first differences here because we use the preceding (in time) observation of each individual observation. There are two reasons why the preceding observation is not always last years observation: (i) the NLSY cohorts were created annually in the first couple of years and after that biannually;

$$\triangle log(y_{mit}) = log(y_{mit}) - log(y_{mit\_}) = \rho_{mi0} \sum_{j=1}^{K} \alpha_{mj}(t^j - t_-^j) + (\varsigma_{mit} - \varsigma_{mit\_}) \qquad t = 2, ..., T_i \tag{12}$$

where the time invariant component of the error term in (7) cancels and t_- indicates the previous observation in time. As a result, selectivity is removed. The baseline growth rate is specified as:

$$\rho_{mi0} = e^{\delta_{m0} + \delta'_m z_{i0}}, \tag{13}$$

where z_{i0} is a vector of individual characteristics observed at $t = 0$ which also (potentially) includes a constant.[7] Given this specification we can rewrite (12) as follows:

$$\triangle log(y_{mit}) = \left(\alpha_{m1} e^{\delta_{m0}}\right) e^{\delta'_m z_{i0}} \left((t - t_-) + \sum_{j=2}^{K} \left(\frac{\alpha_{mj}}{\alpha_{m1}} \right) (t^j - t_-^j) \right) + \zeta_{mit} - \zeta_{mit\_} \quad t = 2, ..., T_i \tag{14}$$

From this it is clear that α_{m1} and δ_{m0} are not identified separately, but that the sign of α_{m1} is identified. By applying NLS on the subsample having opted for major m, consistent estimates of the parameters of $\rho_{mi}(t)$, i.e. $\alpha_{m1} e^{\delta_{m0}}$, δ_m, and α_{mj}/α_{m1} $(j = 2, .., K)$ can be found. Since

$$log(y_{mit}) - \rho_{mi0} \left(\sum_{j=1}^{K} \alpha_{mj} t^j \right) = \beta'_m x_{i0} + e_{mi} + \varsigma_{mit} + \varepsilon_{mi0} \tag{15}$$

$$log(y_{mi0}) = \beta'_m x_{i0} + \varepsilon_{mi0} \tag{16}$$

where the left hand sides are either observed or can be calculated given the estimates obtained thus far, the difference between eqs (15) and (16) for a given t $(t = 1, ...T)$ equals:

(ii) for some individual the observation per year of two-years is interrupted for some years.

[7] Note that we assume that the initial wage growth is positive. From the view point of economic theory this appears to be a natural assumption. However, it can be relaxed, e.g. by assuming $\rho_{mi0} = \delta_{m0} + \delta'_m z_{i0}$ but the resulting model will be harder to estimate.

$$log(y_{mit}) - \left[\rho_{mi0}\left(\widehat{\sum_{j=1}^{K}\alpha_{mj}t^j}\right)\right] - log(y_{mi0}) = \gamma_m\nu_i + \varsigma_{mit} \tag{17}$$

Since $E(\varsigma_{mit}) = 0$, from this a consistent estimate of $\gamma_m\nu_i$ can be obtained by averaging the left hand side across time for every individual but only for the major the individual opted for.

Given all the estimates retrieved thus far, we can obtain consistent estimates of β_m and for each major m, after having substituted $\hat{e}_{mi} = \widehat{[\gamma_m\nu_i]}$[8] and using assumption (11) using OLS on:

$$\begin{aligned} log(y_{mit}) - \left[\rho_{mi0}\left(\widehat{\sum_{j=1}^{K}\alpha_{mj}t^j}\right)\right] - \widehat{[\gamma_m\nu_i]} &= \beta'_m x_{i0} + \tau_m\widehat{[\gamma_m\nu_i]} + \varsigma_{mit} + \varsigma_{mi0} \qquad \text{for } t = 2,..,T_i \\ log(y_{mi0}) &= \beta'_m x_{i0} + \tau_m\widehat{[\gamma_m\nu_i]} + \varsigma_{mi0} \qquad \text{for } t = 0 \end{aligned} \tag{18}$$

An estimate of τ_m is also found. Eq. (18) represents the initial wages. The first equation corrects post initial wages such that at the right-hand side the initial wage remains although with additional random error. The resulting serial correlation and heteroskedasticity will not introduce a bias, but the standard errors of the estimates need to be corrected.[9] For this reason and apart from the first step estimation, all standard errors presented in the result section are bootstrapped.

We can now set out to estimate the expected wages, or more precisely the major-specific initial wage and wage growth. We first start from the observed major for each individual. The relevant expectations are:

$$E(log(y_{mi0})|\nu_i) = \beta'_m x_{i0} + \tau_m\gamma_m\nu_i \tag{19}$$

$$E(\rho_{mi}(t)|\nu_i) = \rho_{mi0}\sum_{j=1}^{K}\alpha_{mj}t^j + \gamma_m\nu_i \tag{20}$$

[8] We use square brackets to indicate that the complete term within the square brackets is estimated. The parameters building the expression within the square brackets are not (yet) identified.

[9] As we use estimates from previous estimations, standard errors need to be corrected anyway.

Next, we need to estimate these expectations for the counterfactuals. We only observe earnings for the major the individual actually chooses. For a large part we can calculate the expectations in eqs (19) and (20) since we estimated β_m, τ_m, ρ_{mi0} and α_{mj}. The problem is that γ_m and ν_i are not identified separately: we only have an estimate $\widehat{[\gamma_m \nu_i]}$ for the major actually chosen. Note that this is only a scaling problem: the order of $\gamma_m \nu_i$ and therefore ν_i is fixed, the absolute level of ν_i is unknown. We solve this identification problem in three steps:

1. We calculate the Mahalanobis-distance of the observations using all the explanatory variables.

2. Given m, we match $\widehat{[\gamma_m \nu_i]}$ using kernel matching for each alternative major j ($j = 1, .., 5, j \neq m$). We do this for all m. This gives us: $\widetilde{[\gamma_{jm} \nu_i]}$

3. In order to maintain the ordering, for each individual that opted for major m and for each counterfactual major j, $j \neq m$, we regress the matched $\widetilde{[\gamma_{jm} \nu_i]}$ on the estimated $\widehat{[\gamma_m \nu_i]}$ and use the predicted value as the counterfactual estimate of $\gamma_j \nu_i$ if we combine for all m.

The counterfactuals are based on the ordering of our estimate of $\widehat{[\gamma_m \nu_i]}$. The unknown scaling component γ_j for the majors that the individual did not choose, is determined by kernel matching. Regression ensures that the estimated order is not violated. Note that we do not apply full scale matching. We only need matching to make the scales of γ_j comparable and as a result we are able to create two the explanatory variables in the major choice equation using eqs 19 and 20.

2.3 The estimation of the major choice equation

We characterize expected future wages by the expected initial wages and growth rates. The procedure described in the previous subsection yields T_i different expected growth rates for each individual and for each major. To reduce the number of, quite likely highly correlated, explanatory variables in the major choice equation and to solve the problem of an unequal number of growth rates per individual, we will reduce the T_i growth rates to dimension 1 by averaging across time. This average is denoted by $\bar{\rho}_{mi}$. Moreover, we will also introduce major specific constants: κ_{0m}. The utility of choosing major m as perceived by individual i is specified as:

$$U_{mi} = \theta_1 E(log(y_{mi0})|\nu_i) + \theta_2 \bar{\rho}_{mi} + \kappa_{0m} + \xi^*_{mi}. \tag{21}$$

The error term ξ^*_{mi} may be correlated to the unobserved heterogeneity ν_i introduced earlier. We make this explicit by assuming: [10]

$$\xi^*_{it} = \kappa_{1m}\gamma_m\nu_i + \xi_{it}. \tag{22}$$

Note that we do not have an estimate of ν_i, but of $\gamma_m\nu_i$, this is what determines our particular error structure, but this is not particularly restrictive since the inclusion of alternative constant regressors allow the inclusion of alternative specific coefficients and κ_{1m} automatically correct the scaling. Substituting in (21) yields:

$$U_{mi} = \theta_1 E(log(y_{mi0})|\nu_i) + \theta_2 \bar{\rho}_{mi} + \kappa_{0m} + \kappa_{1m}\gamma_m\nu_i + \xi_{mi}, \tag{23}$$

There are two determinants that are alternative and individual specific ($E(log(y_{mi0})|\nu_i)$ and $\bar{\rho}_{mi}$) and one individual specific regressor (ν_i) plus a constant (κ_{0m}).

We estimate major choice both parametrically by multinomial logit and probit and semiparametrically implementing the estimation method proposed by Manski (1975) and Li (2011). Our preferred methods are the semiparametric ones as they allow us to make no distributional assumptions on the error term ξ_{mi}.

The theoretical literature on semiparametric estimation of choice models has mostly concentrated on the binary case (Lee, 1982; Cosslett, 1983; Robinson, 1988; Newey, 2009). Semiparametric estimators for multiple and unordered choice models are harder to find. Two theoretical examples of this class of estimators are Manski (1975) and Lee (1995). However, in the empirical literature very few applications of these methods can be found. This paper presents a practical application for these two methodologies that have, so far, rarely been used in the applied literature. We apply the maximum score estimator of Manski (1975) and use the smoothing idea of Horowitz (1992) to make the objective function continuous and differentiable. The idea is to maximize the number of correct major predictions, where the predicted major is the major with the largest probability. Because of the yes (correct prediction) or no (incorrect prediction) character of the objective function it is hard to maximize using standard techniques. Horowitz (1992) suggests smoothing the objective

[10] A more general factor describing unobserved individual tastes and characteristics, say $\sigma_{\nu m}\nu_{mi}$, can be added as well, but it can not be distinguished from the error term in (1).

function by using a continuous function that closely approximates the 0-1 situation. This can be done e.g. by $\Phi((V_{ij} - V_{ik})/h)$ where V_{ij} is the deterministic part of utility, j indicates the chosen major and k represents the other majors. h is a bandwidth parameter chosen. The smaller it is, the closer it resembles the 0-1 situation. Although, the resulting objective function is now continuous and differentiable, the global optimum is still hard to find due to the many local optima. We first use simulated annealing (10 million iteration steps) to find good starting values and then optimized our routine. The optimum we found is the best result we encounter in numerous attempts. The simplicity of the objective function makes the procedure tractable. Note that, Manski (1975, 1985) is actually on a conditional logit model setting (only including explanatory variables that differ across individuals and alternatives), but this can be generalized to include individual-specific regressors as is discussed in Maddala (1983, p. 42, footnote 4). [11]

The second semiparametric estimator that we consider is based on a different idea. The idea, in this case, is to use a multinomial logit model and thereby assuming independent type I extreme value distributed error terms, but to estimate the systematic component semiparametrically. Such methods are discussed in e.g. Briesch, Chintagunta, and Matzkin (2002) and Li (2011). We follow the suggestion of Li (2011) and use splines to approximate the systematic part of the utility in (23).

3 Data

For our purpose, we use the 1979 (NLSY79) and 1997 (NLSY97) waves of the National Longitudinal Survey of Youth. These are two widely used longitudinal surveys representative of the U.S. population. The NLSY79 started in 1979 surveying 12,686 individuals who were 14 to 22 years old at the time. The NLSY97 contains information on 8,984 young individuals who were between 12 and 18 years of age in 1997. Both surveys are still ongoing. The last waves we use, are collected in 2014 for the NLSY79 and 2015 for the NLSY97. Respondents were interviewed annually until 1994 for the NLSY79 and 2011 for the NLSY97 and biannually thereafter.

Both surveys include a wide variety of economic, sociological, and psychological mea-

[11] We also tried to estimate the method proposed by Lee (1995). Although, convergence was achieved, using different starting values resulted in finding different local optima. Given the slow speed of convergence, it is impractical to engage in some kind of grid search so that we were forced to abandon the idea of using this method.

sures. In particular, both surveys include information on the major selected in college for those individuals who proceed to tertiary education.

Since our analysis regards major choice in college, we restrict the sample to males and females who completed college and for whom the major choice is known. This reduces our sample to 5,205 individuals.

Our model has two dependent variables: major choice for the selection probabilities and earnings for the wage equation. In both NLSYs the major in college is recorded as a four-digit code distinguishing among the various fields of study (e.g.: Biological Sciences, Engineering, Business and Management, etc.) and subfields within the bigger field (e.g.: Microbiology, Chemical Engineering, Banking and Finance, etc.). We combine this information into five major categories: Natural Sciences, Social Sciences, Humanities, Education and Health[12]. Earnings are expressed as the logarithm of hourly earnings in the period considered translated in 2010 constant dollars. The historical series for the Consumer Price Index (CPI) in the US for the period considered is taken from the Bureau of Labor Statistics.[13] We are interested only in wages earned after graduation, therefore our initial wage is the first wage earned thereafter. In the NLSY79 the first 'graduate' wage observed is in 1990 while for NLSY97 in 1999. We use 2014 (NLSY79) and 2015 (NLSY97) as final observation years.

The information contained in the NLSY allows us to control for gender, ethnic background and geographical characteristics for the area of origin at age 17[14]. Following other studies (Neal and Johnson, 1996; Altonji et al., 2012; Deming and Noray, 2018) we use respondents' standardized scores on the Armed Forces Qualifying Test (AFQT) for the NLSY79 and the Armed Service Vocational Aptitude Battery (ASVAB) for the NLSY97 to proxy for ability. Both tests are a series of tests in mathematics, science, vocabulary, and automotive knowledge. The AFQT was administered in 1980 to all subjects regardless of their age and schooling level. For this reason it can include age and schooling effects in the ability index that the test is meant to construct. To correct for these undesired effects, we follow Kane and Rouse (1995) and Neal and Johnson (1996). First we regress the original test score on age dummies and quarter of birth, then we replace the original test score with the residuals obtained from this regression. For comparability between the two tests, we re-scale both

[12]For a detailed description of the NLSY major classifications and our mapping into five categories see the appendix.

[13]Source: fttp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt.

[14]The geographical controls include a dummy indicating whether the respondent grew up in an urban area and four dummies for the area of origin: North Central, North East, South, and West.

scores to a maximum of 100. Figure 1 provides information on the differences in the distribution of the two test scores. The distribution is fairly similar but AFQT-scores are more concentrated at the high end of the scale. The average of the AFQT score is somewhat higher but with a smaller variance. In the econometric analysis we will include an indicator variable for the two sub-samples and its interaction with test scores where necessary, to account for possible differences in the two tests.

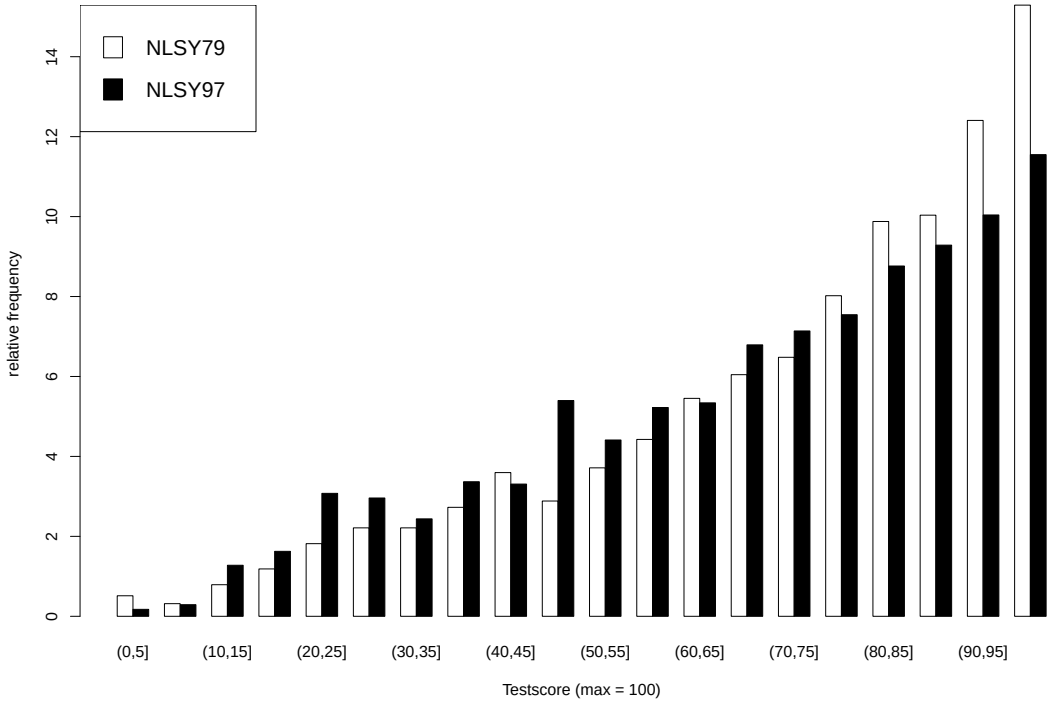

Figure 1: Distribution of the re-scaled test scores for NLSY79 (AFQT) and NLSY97 (ASVAB).

After having removed unknown and unrealistic hourly wages, (i.e. wages smaller than $e^1 = \$2.71$), wages observed before college graduation and individuals with majors that could not be assigned to any of the five groups, we are left with 27,982 observations for 4,519 individuals. We observe 6.2 wages per individual on average with a maximum of 15 wages. The number of observed wages are summarized in Table 1. For 824 individuals we observe at least one wage whereas for 3,695 individuals we observe more than once.

We report both starting wages and mean wages observed throughout the survey period in Table 2. For both these measures Humanities is the lowest paying field, while Health

At least	1	2	3	4	5	6	7	8
NLSY79	2531	1946	1803	1665	1524	1387	1296	1166
NLSY97	1988	1749	1457	1260	1046	804	561	363
Combined	4519	3695	3260	2925	2570	2191	1857	1529

At least	9	10	11	12	13	14	15
NLSY79	1043	932	847	740	638	526	324
NLSY97	208	116	50	7	4	1	0
Combined	1251	1048	897	747	642	527	324

Table 1: Count of the observed number of wages

pays the highest wages at the start and throughout the career. The gap between the highest and lowest paying fields is around 32% for starting wages and 45% for average wages. Test scores are highest for Natural Science graduates and lowest for Humanities ones, but the spread is substantial. As expected, Education and Health are female dominated fields and overall, women are more numerous than men in our sample. About 20% of our sample is black, 13% Hispanic and 78% grew up in a city and 56% is taken from the 1979 survey. In Education, respondents start working at the late age of 32, whereas in Humanities the first working experience after college completion is acquired at the age of 28. We also observe that almost half of the sampled individuals graduated in a Social Science discipline, 1,212 in Natural Science, and only 231 in one of the Humanities.

	Natural Sciences	Social Sciences	Humanities	Education	Health	Total
Initial log hourly wage	2.822	2.766	2.605	2.688	2.884	2.775
	(0.714)	(0.731)	(0.782)	(0.604)	(0.802)	(0.725)
Mean log hourly wages	3.176	3.139	2.885	2.920	3.256	3.120
	(0.685)	(0.723)	(0.722)	(0.592)	(0.750)	(0.709)
Test score	68.889	65.902	64.960	66.119	66.646	66.753
	(28.646)	(28.235)	(33.392)	(25.185)	(26.057)	(28.107)
Age started working	29.403	29.885	27.952	32.224	31.345	30.076
	(6,244)	(7.078)	(6,178)	(7.784)	(7.412)	(7.018)
Dummy variables						
Female	0.442	0.550	0.571	0.767	0.790	0.571
Black	0.172	0.219	0.121	0.177	0.192	0.194
Hispanic	0.138	0.117	0.130	0.144	0.146	0.129
Urban	0.768	0.791	0.801	0.761	0.776	0.780
North East	0.196	0.219	0.260	0.192	0.178	0.208
West	0.185	0.160	0.221	0.172	0.183	0.173
North-Central	0.274	0.259	0.277	0.241	0.283	0.264
NLSY97	0.572	0.405	0.654	0.257	0.352	0.440
N	1,212	2,102	231	536	438	4,519

Standard deviation in parentheses. Wages in 2010 real dollars.

Table 2: Descriptive Statistics

4 Estimation results

Before the estimation of the major choice equation (23), we first have to estimate wage growth rates and the determinants of the initial wage. As discussed in section 2.2 the estimation entails three steps. The wage growth equation is specified in (14) and is estimated with non-linear least squares. The determinants of the wage results from an ordinary least-squares estimation of eq. (2). These estimation results are combined to retrieve the expected initial wage and expected annual growth rate as specified in eqs (19) and (20) and these are used in the estimation of the major choice equation (23). To obtain the correct standard errors of the estimates in the second (section 4.2) and third step (section 4.3), we employ a non-parametric bootstrap with 200 replications.[15]

[15]According to Efron and Tibsharani (1993, p. 52), using 200 replication in the bootstrap almost always suffices.

4.1 Estimation of the wage growth rates

To start with, let us stress that we analyze real wages. The wage growth considered here is real wage growth. The non-linear least squares estimates of the growth rate equation (14) are presented in Table 3. To acquire reasonable significance, we have to limit the number of explanatory variables. None of the deleted explanatory variables, as discussed in section 3, have a significant effect. To illustrate the problem, even the straightforward addition of the dummy NLSY97 reduces the significance as presented in Table 3 considerably. The loss of significance is not observed for the time variables presented in the lower panel of the table. However, in the upper panel 12 estimates are significantly different from 0, whereas this reduces to 4 significant effects if the dummy NLSY97 is included.

Equation (14) consists of two parts: a time pattern, involving the α-parameters, and an individual scale factor, involving the δ-parameters, that renders wage growth observation specific. The time pattern is generic, although different across majors. Individual variation is reflected in the scale factors.

The order of the polynomial of the time pattern of wage growth is chosen according to the AIC criterium.[16] The lowest AIC values were found for a third-order polynomial, apart from the Humanities major where a polynomial of order 4 needs to be preferred according to AIC. The patterns are hard to evaluate using only the presented estimates. To get a better insight, consider Figure 2. In this plot we show the real wage growth rate for a male with an average test score from the NLSY79 subsample. All curves show the expected curvature apart from Humanities. The growth paths of Natural and Social sciences are very similar. The highest real wage growth is 40% or a little less than 3% a year for the first 15 years of working life. Individuals who chose one of these majors experience wage growth for the first 15 years of their working life; after that wages remain roughly constant. Health graduates show a similar evolution, but the growth rate is considerably larger, approximately twice the growth rate of the Sciences majors, and their wages plateau earlier. Education graduates experience the lowest wage growth. Compared to the initial wage, their wages increase only about 18%, i.e. less than 2% real wage growth a year. On top of that, the leveling off starts earlier. The estimated pattern for Humanities graduates is somewhere between the previous cases. Hardly any wage growth is experienced in the first 10 years of working life, and an accelerated wage growth is experienced after that. If we restrict the polynomial of Humanities to order 3, only the second-order coefficient is significant at 5% and the pattern

[16] We estimated polynomials up to order 6.

becomes very similar to the one of the Education major.

		Nat Sciences	Soc Sciences	Humanities	Education	Health
Constant	$(\alpha_{m1}e^{\delta_{m0}})$	0.141***	0.138***	0.082	0.056	0.036
		(0.038)	(0.027)	(0.095)	(0.046)	(0.036)
Test score	(δ_{m1})	-0.009**	-0.008**	-0.017	-0.006	0.022**
		(0.004)	(0.003)	(0.026)	(0.010)	(0.011)
Female	(δ_{m2})	-0.346*	-0.399***	-0.480	0.170	-1.031***
		(0.200)	(0.136)	(0.489)	(0.542)	(0.343)
Test score x NLSY97	(δ_{m3})	0.011***	0.013***	0.036*	0.013*	0.001
		(0.003)	(0.002)	(0.021)	(0.007)	(0.004)
Δt^2	$(\alpha_{m2}/\alpha_{m1})$	-0.056***	-0.058***	-0.183***	-0.062*	-0.063***
		(0.010)	(0.006)	(0.029)	(0.018)	(0.008)
Δt^3	$(\alpha_{m3}/\alpha_{m1})$	0.001**	0.001***	0.014***	0.001	0.001***
		$(3 \cdot 10^{-4})$	$(2 \cdot 10^{-4})$	(0.004)	(0.001)	$(3 \cdot 10^{-4})$
Δt^4	$(\alpha_{m4}/\alpha_{m1})$	-	-	$-3 \cdot 10^{-4}$***	-	-
		-	-	$(1 \cdot 10^{-4})$	-	-

Standard errors in parentheses. ***/**/* = significant at 1%/5%/10%.

Table 3: Estimated of growth rates equation (14) across majors

Regarding the scaling factor in the wage growth specification, i.e. $\alpha_{m1}e^{\delta_{m0}}e^{\delta' m z_{i0}}$ in eq. (14), we find significant effects for all variables for Natural and Social Science majors. The scale factor is lower for women than for men, indicating that females experience slower wage growth than men. The differences are quite substantial: in Natural Sciences female experience a 29%, in Social Sciences the difference is -33% and in Health it is -64%. For Humanities and Education the effects are not significant and in Education females are estimated to experience a larger wage growth than men. The test score has a different effect on the two NLSY subsamples. The effect of the test score is negative for the older subsample whereas it is positive in the more recent one. The negative effect is unexpected and might be explained by the AFQT-test score not being a very reliable indicator of ability (Schofield, 2014). For Humanities and Education majors only the cross term of test score and the NLSY97-dummy is significant. A one point increase in the test score, which ranges between 0 and 100, is associated with a 1% slower wage growth for the Sciences for the NLSY79-participants and about 2% faster wage growth for Health for the NLSY79-participants. For the NLSY97 samples an additional point on the test, gives extra wage growth ranging from

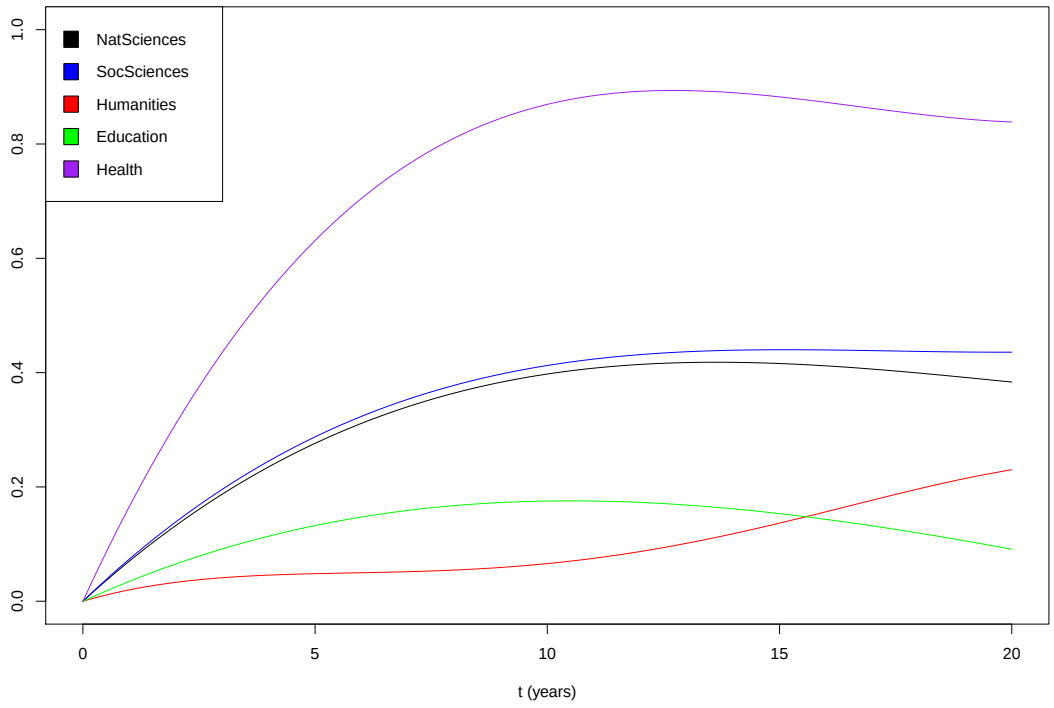

Figure 2: Estimated growth rates across majors.

0.2% (Natural Sciences) to 2.3% (Health).

To get an idea about what the scale factors in the wage growth specification look like, consider Figure 3. In this figure, we plotted the kernel density of the scale factor per major relative to the same reference as used in Figure 2, i.e., each of the curves displays the estimated distribution of $e^{\delta'_m(z_{i0}-z_{reference})}$ for a specific major. The relative scale factor starts from a minimum of about 0.5 in our sample. Values below 0.5 are artificial and due to the smoothing process used to plot kernel densities. The curves indicate that Figure 2 exaggerates the true wage growth differences across majors. The major experiencing the highest wage growth, Health, tends to have a smaller relative scale factor whereas, for Education, the wage growth appears to be small, but there is some compensation due to a higher scale factor. The other majors are between these extremes and again Natural and Social science majors look similar.

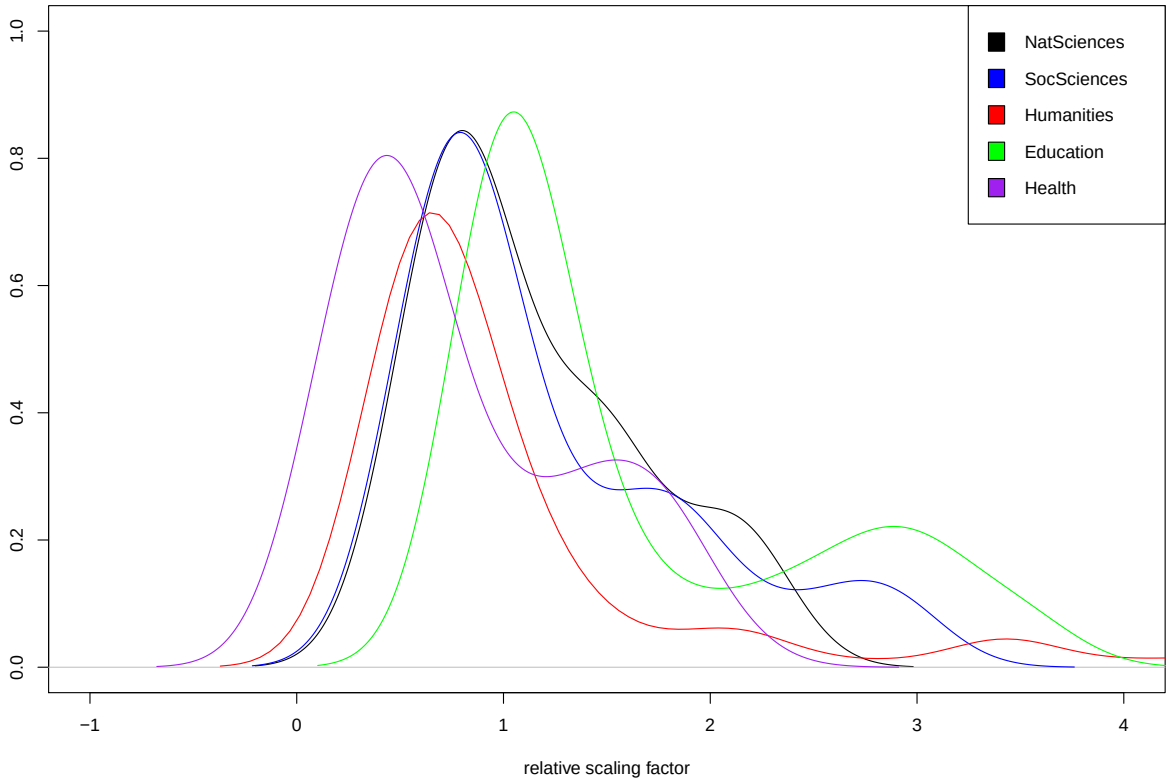

Figure 3: Kernel density plots of the estimated scale factors in the wage growth equation across majors.

4.2 Estimation of initial wages

We present the results of the ordinary least squares estimation of eq. (18) in Table 4. Initial wages are strongly and positively affected by the individual's test score, except for Health. For three out of the five major groups interactions of the test score with the subsample used are strongly significant. Note the reversal in sign of the estimates compared to the wage growth rate equation: now the test score itself has a positive sign whereas the interaction has a negative sign, again except for Health. For Social Sciences and Education the effect of the test score vanishes for the NLSY97 subsample. For the NLSY79 subsample, a 1 point better score will increase the starting wage by 0.5% (Education) and up to 1.3% (Humanities), apart from the Health major for which we find no effect.

The effect of gender is statistically significant only for Social Sciences majors where

women starting wage is 8% lower than that of comparable men. Some differences are estimated for the NLSY97 subsample; not only the interaction with test score has a negative impact, but also the dummy indicating the NLSY97 subsample is positive although significant only for Social Sciences. Our estimates for the ethnic background are not statistically significant for most major groups, except for a negative and large penalty for Black Health graduates of about 14%, a positive premium for Black Education graduates of 11% and a positive premium of 7% for Hispanic graduates in one of the Social Sciences. Living in an urban area in 1979 or 1997, has a positive effect on wages but it is only significant for the majors Social Sciences and Education. Living in the North-East or West of the U.S. at the age of 18 increases initial wages compared to living in the South. The effect of the age at which the college graduates started working is positive for three out of five majors. Starting one year later will increase the initial wage with 2% for Natural and Social Sciences and even almost 4% for Health. Finally, the effect of unobserved heterogeneity or fixed effect, as measured by ν_i, is strongly negative and significant. The significance is not surprising and indicates that time-invariant unobserved heterogeneity is present. As for the negativity, remember that since $1 + \tau_m > 0$, cf. eq. (9)-(11) an estimated negative unobserved heterogeneity term posits a negative effect on initial wages only if the coefficient is estimated to be between 0 and -1, in our case, the estimated coefficient is larger than -1, implying a *positive* impact of unobserved heterogeneity on wages earned at a later date. Note that due to the inclusion of the fixed effect term, measures of goodness of fit are unusually high. An R^2 of 75% (Health) is rare in the estimation of individual wages. The lowest goodness of fit is found for Education, but it is still 58.5%.

Table 5 and Figure 4 provide information on the estimated initial wages and the counterfactuals. The expected initial wages for the alternative majors are estimated using eq. (19). From Table 5 we see that the observed wages correspond closely to the calculated wages for the relevant major.[17] Larger deviations are found for the counterfactuals. For instance, if a Natural Science graduate had chosen one of the Social Sciences instead, her initial wage would have been about 7.5% lower. For Humanities and Education the counterfactual penalty would have been 15.4% and 10.4% respectively. Choosing Health, on the

[17]The observant reader will note that the reported actual wages in Table 5 deviate from those reported in Table 2. This is due to the number of observations used: in Table 2 we use all (4,519) observations whereas in Table 5 we use the observations for which wages are observed at least twice (3,695 observations). As is clear from the estimation procedure, the unobserved heterogeneity component can only be estimated for the individuals who reported at least two wages.

other hand, would have increased the wage by about 9.2%. The expected starting wage for Health majors is the highest regardless of the real choice, whereas the lowest wages are estimated for Humanities or Education. Note however, that individuals that have chosen the Education major, are predicted to do very well if they would have chosen one of the other majors. The reason for this is the average age of this group: they are substantially older when they start working than individuals that did not choose the Education major. Figure 4 provides a kernel plot of the estimated counterfactual initial wages. The estimated wage is most concentrated for Health whereas the largest spread is found for Humanities.

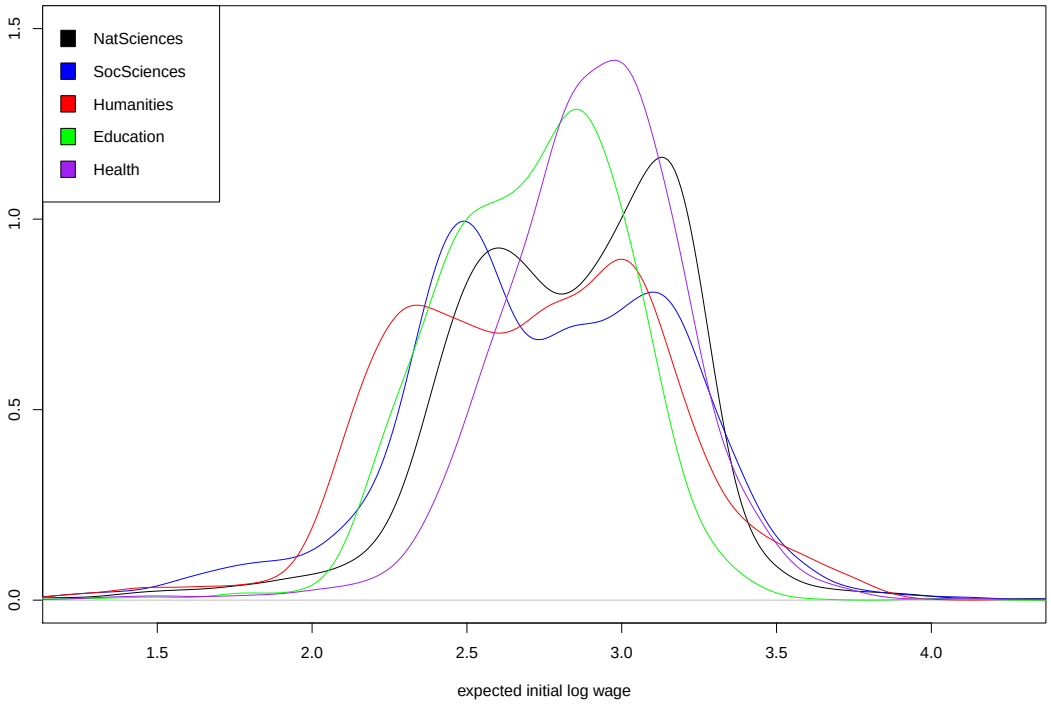

Figure 4: Kernel density plots of the expected initial log wages across majors

4.3 Estimation of major choice

Table 6 presents the major choice estimates for four different estimation methods.[18] Two fully parametric ones: multinominal logit and multinomial probit, a parametric, but more flexible parametric method (Li, 2011) and one semiparametric estimation method, (Manski, 1975).

In this table we present our estimates for our two key parameters: the effect of starting wages and wage growth rates on major choice. Remember that we have calculated major-specific initial wages and wage growth rates, therefore in the multinomial estimation, these two parameters are alternative specific, implying that only one coefficient is estimated per major and per variable. The estimated unobserved heterogeneity is also added as a regressor, but this variable is individual-specific, so we obtain one estimate per major. We set the Natural Sciences category as the reference group. Apart from these variables, we also include a major specific constant, but these estimates are not presented. All standard errors are bootstrapped.

Both parametric estimation methods find significant effects of expected initial wages and expected wage growth. Both effects are positive indicating that individuals prefer higher starting wages and higher wage growth. The estimates of the probit specification are somewhat larger but the scaling of logit and probit models differ so that it is better to look at the ratios. These are indeed quite similar. For the less restrictive estimation methods, we find statistically significant effects only for the initial wage for the Manski-estimation method. In that case, we estimate a negative, but insignificant, effect of wage growth. For the flexible parametric method of Li, no significance is found at all. The estimated effects of the semiparametric estimation are relatively large compared to the other effects. Furthermore, only in this case, we find significant effects of unobserved heterogeneity. The relative large absolute size of the estimates is not due to a different scaling. The Manski (1975) method requires to impose additional identification restriction and we have set the constant of the Social Sciences major equal to its corresponding multinomial logit estimate.

In the type of research we employ, i.e. using counterfactuals to create two explanatory

[18] There are some additional parameters estimated for the Multinomial Probit model and the Li (2011) method. Multinomial Probit: Only the variance of ΔU is identified and additionally one variance has to be put to 0. Of the remaining 9 parameters, three variances and six covariances, two variances are significant at 1%, one covariance is significant at 5% and one other covariance is significant at 10%. Li (2011): We present here estimations with only 3 breaks. None of the additional parameters is significant. Increasing the number of breaks does not improve significance.

variables based on the model specification and correcting for potential selectivity, it is always questionable whether this is a valid procedure. Finding estimates with signs that correspond to economic theory assuming rational individuals and significant estimates, is an indication that we, at least to some extent, are able to do what we intended. If the variables created in this way are without meaning, quite likely, a lack of significance or unexpected signs would have been found. But, we do not estimate significant effects for the less restrictive estimation methods as for both Li's and Manski's estimation methods the expected growth rates of wages are not significant. We believe, this is likely to be caused by the overall reduction in significance when more flexible estimation methods are used.

In our experience the practical applicability of the multinomial probit model is rather limited. We applied simulated maximum likelihood and convergence depends on the seed and number of replications per simulated probability used.[19] While bootstrapping the standard errors, we found strongly fluctuating estimated variances and constants per major and variances and covariances. This also the case for some of the estimated variances and their standard errors. For instance, the estimated variance of the error term of the Education major is 126.9 with and standard error of 363). In contrast, the estimated effects of the initial wages and annual wage growth were more stable as represented by the bootstrapped standard errors in Table 6.

The estimation of the semiparametric Manski (1975) is also quite problematic. There appear to be numerous local optima and it is hard to find the global one. We employed simulated annealing in combination with optimizing the objective function, to find the global optimum. On top of that we used many different starting values. Obtaining convergence or the time needed to estimate is not a real problem, however.

To illustrate the impact of the main variables, consider Table 7. It gives the estimated multinomial logit probabilities calculated in the average values of the explanatory variables across each alternative and the probability differences as a result of adding one standard deviation to the expected log initial wage and/or the expected annual wage growth. Note that the estimates of the Manski estimation method can not be used here, because this method does not provide estimates of the probabilities. The results indicate that the effect of the explanatory variables is considerable. If both the initial wage and wage growth rate is increased with one standard deviation for a specific major, the probability of choosing this

[19] To give an idea: we present results using R = 35 simulations. R = 30, 40 or 100 do not yield convergence, R = 5, 10, 20, 35, 50, or 60 do yield convergence.

major will increase by 45.2% (Natural Sciences), 23.2% (Social Sciences), 64.0% (Humanities), 57.1% (Education) or 53.0% (Health). Only increasing the initial wage or the wage growth rate with one standard deviation, shows that the effect of the change of the initial wage is about 30-40% larger than the effect of the change in wage growth.

4.4 Robustness checks

In this section we check if our results are robust within the two subsamples, to the inclusion of additional covariates, and to different methods of imputation for initial wages and wage growth rates.

As the four methods produce comparable results, for the robustness checks we concentrate on the multinomial logit specification only.

4.4.1 NLSY subsamples

So far we have combined, the 1979 and 1997 NLSY samples and added an indicator variable to distinguish between the two. In this section we test whether the models are the same in both subsamples. Due to space restrictions, we do not present the estimation results here. The estimation results of the NLSY79 sample are very similar to the overall results presented earlier, although the significance is reduced to some extent. The effects of expected log initial wages and expected annual wage growth are estimated to be quite similar. For instance, the expected wage growth curves as presented in Figure 2 are hardly distinguishable from the ones resulting from estimating on the 1979 subsample. Also, the estimate of the variable initial wage in the major choice equation remains positive buy is no longer significant (estimate 0.089 with standard error 0.185). The estimate of the effect of wage growth remains positive and is significant at 5%: 0.146, with standard error 0.060.

In the 1997 subsample the conclusions are different. The estimation results are more different and far less significant than in the combined estimation. Especially the wage growth curves exhibit different behavior. Until 5 years after having started working the curves are very similar but after that wage growth is estimated to explode. Quite likely this is due to the more limited observation period available which might cause overfitting. The estimates of initial wages and the wage growth parameters remain positive: the effect of wage growth is estimated to equal 0.340 (s.e. 0.173, significant at 5%) and the estimated effect of wage growth equals 0.001 (s.e. 0.416).

4.4.2 Adding more alternative specific regressors to the major choice equation.

We now consider the effect of adding extra individual-specific variables as explanatory variables. Table 8, panel A, contains the estimation results. Specification 1 is the same as the one presented in Table 6 and is added for comparison. This specification only has the constant and the unobserved heterogeneity term (ν_i) as individual-specific regressors in the major choice equation. Specification 2 adds gender, whereas specification 3 also adds ethnic background (black and Hispanic). Specification 4 also adds the test score.

Concentrating on the main parameters of the expected log initial wage and expected annual wage growth, the differences across the specifications listed are limited. In all cases except one, the estimates remain positive and significant. The coefficient for initial wage drops when more variables are added to the choice equation, whereas this is not true for the expected wage growth. Gender has a strong and significant effect and this is true for all majors. For ethnic background and the test score very few significant estimates are found and the significance if always at the lower level of 5% or 10%.

4.4.3 Alternative estimates of the expected initial wage and the expected wage growth.

Thus far, we estimated the expected initial log wage and the expected annual wage growth using regression. In particular, eqs (14) and (18) are used. In Table 8 panel B we present simpler measures for the two key parameters. In specifications 1 and 2 we use averages across majors and gender (specification 1) and ethnic groups (specification 2). Specifications 3 and 4 are based on regression, but in specification 3 we do not correct for self-selection, while in specification 4 we do not include the fixed effect term and therefore ignore unobserved heterogeneity.

Specifications 1 and 2 give unexpected negative, in absolute value large and significant, effects of initial wages on major choice. The effect of wage growth is estimated to be positive and is also significant in one case. In specifications 3 and 4 both coefficients are estimated to be positive. The significance without using a correction for selectivity is reduced considerably (specification 3), although the effect of wage growth remains significant at a 10% level. Ignoring the fixed effects, as we do in specification 4, reduces the significance to an extent that no estimate is significant anymore.

These robustness checks suggest that regression methods seem to lead to expected

outcomes from an economic point of view. They also highlight the importance to correct for selectivity in the estimations, and suggest some role for unobserved heterogeneity.

4.4.4 Conclusions robustness checks.

The overall conclusion from our robustness checks is that more elaborate specifications do not really change the conclusions. Simplification of the estimation of expected initial log wages and expected annual wage growth is not a good idea. Splitting up the sample might indeed be advisable, but the information in the most recent subsample is still too limited to get trustworthy results. Perhaps, it would have been better to restrict the sample to NLSY 1979. On the other hand the estimation results of the combined samples are quite similar but stronger statistical effects are found. In our main estimations, we combined male and female respondents. Although we corrected for gender to some extent in our estimations, our results indicate that major choice behavior differs across the sexes.

		Nat Sciences	Soc Sciences	Humanities	Education	Health
Constant	(β_{m0})	1.193***	0.995***	0.263	1.794***	1.653**
		(0.450)	(0.349)	(1.230)	(0.554)	(0.681)
Test score	(β_{m1})	0.006***	0.009***	0.013***	0.005***	-0.002
		(0.002)	(0.001)	(0.004)	(0.001)	(0.002)
Female	(β_{m2})	-0.072	-0.080**	0.037	-0.095	0.131
		(0.059)	(0.034)	(0.164)	(0.086)	(0.145)
Test score x NLSY97	(β_{m3})	-0.001	-0.010***	-0.018***	-0.005*	0.001
		(0.002)	(0.001)	(0.005)	(0.003)	(0.003)
NLSY97	(β_{m4})	0.010	0.390**	0.976	0.041	0.296
		(0.244)	(0.170)	(0.647)	(0.277)	(0.322)
Black	(β_{m5})	-0.016	0.016	0.159	0.111*	-0.144*
		(0.043)	(0.035)	(0.130)	(0.060)	(0.081)
Hispanic	(β_{m6})	0.014	0.069*	-0.010	0.047	-0.062
		(0.054)	(0.037)	(0.175)	(0.060)	(0.074)
Urban	(β_{m7})	0.065	0.075**	0.023	0.077*	0.038
		(0.041)	(0.031)	(0.122)	(0.045)	(0.060)
NorthEast	(β_{m8})	0.053	0.083**	0.091	0.173***	0.124
		(0.049)	(0.036)	(0.132)	(0.064)	(0.078)
West	(β_{m9})	-0.003	0.107***	0.023	0.138***	0.139*
		(0.051)	(0.037)	(0.126)	(0.053)	(0.081)
NorthCentral	(β_{m10})	-0.027	-0.001	-0.091	-0.063	-0.092
		(0.039)	(0.031)	(0.116)	(0.061)	(0.072)
Age started working	(β_{m11})	0.020**	0.020***	0.027	0.012	0.037***
		(0.010)	(0.007)	(0.021)	(0.010)	(0.012)
Fixed effect ($\gamma_m \nu_i$)	(τ_m)	-0.767***	-0.645***	-0.790***	-0.717***	-0.804***
		(0.031)	(0.025)	(0.117)	(0.046)	(0.072)
Adjusted R^2		0.636	0.596	0.678	0.585	0.750

Bootstrapped standard errors in parentheses. ***/**/* = significant at 1%/5%/10%. The specification also includes a dummy for missing test score and dummies for year of observation.

Table 4: Estimated initial wage equations (18) across majors.

Major chosen		Actual logwage	Nat Sciences Expected logwage	Soc Sciences Expected logwage	Humanities Expected logwage	Education Expected logwage	Health Expected logwage
Nat Sciences	mean	2.778	2.787	2.715	2.644	2.688	2.875
	(st dev)	(0.723)	(0.586)	(0.337)	(0.436)	(0.269)	(0.277)
Soc Sciences	mean	2.745	2.842	2.743	2.693	2.709	2.869
	(st dev)	(0.728)	(0.305)	(0.564)	(0.422)	(0.273)	(0.246)
Humanities	mean	2.564	2.754	2.645	2.571	2.626	2.885
	(st dev)	(0.769)	(0.291)	(0.349)	(0.676)	(0.283)	(0.275)
Education	mean	2.683	2.872	2.824	2.786	2.686	2.982
	(st dev)	(0.609)	(0.278)	(0.330)	(0.450)	(0.478)	(0.257)
Health	mean	2.853	2.844	2.770	2.723	2.698	2.870
	(st dev)	(0.810)	(0.287)	(0.347)	(0.449)	(0.273)	(0.721)

Table 5: Mean and standard deviation of actual and expected initial wages across majors

		MNLogit	MNProbit	Li	Manski
Expected log initial wage	(θ_1)	0.431**	0.564**	0.032	1.471**
		(0.170)	(0.246)	(2.060)	(0.699)
Expected annual wage growth	(θ_2)	0.066**	0.098*	0.262	-0.125
		(0.028)	(0.041)	(0.193)	(0.147)
Constant Soc Sciences	(κ_{02})	0.846***	0.444**	0.620	
		(0.101)	(0.210)	(2.726)	
Constant Humanities	(κ_{03})	-1.625***	-6.466	-0.229	-3.465***
		(0.090)	(21.866)	(6.069)	(1.104)
Constant Education	(κ_{04})	-0.718***	-15.426	0.181	-1.110
		(0.078)	(84.325)	(5.202)	(0.680)
Constant Health	(κ_{05})	-0.586***	-3.688	-2.137	-1.920**
		(0.199)	(20.325)	(5.181)	(0.857)
Fixed effect Soc Sciences $(\gamma_2\nu_i)$	(κ_{12})	-0.074	-0.041	-0.991	1.351***
		(0.115)	(0.114)	(0.949)	(0.373)
Fixed effect Humanities $(\gamma_3\nu_i)$	(κ_{13})	-0.110	0.100	0.062	-3.490***
		(0.278)	(3.821)	(1.979)	(0.781)
Fixed effect Education $(\gamma_4\nu_i)$	(κ_{14})	0.040	6.454	0.105	-0.989***
		(0.155)	(40.851)	(1.741)	(0.348)
Fixed effect Health $(\gamma_5\nu_i)$	(κ_{15})	0.011	0.183	0.312	-2.107***
		(0.222)	(2.559)	(2.198)	(0.577)

Bootstrapped standard errors in parentheses. ***/**/* = significant at 1%/5%/10%.
Reference category (fixed effect, constant): Natural Sciences. In the estimated model using the Manski (1975) method, an additional restriction has to be added because one of the variances of the error terms can not be restricted to a constant. We opted to restrict the constant of Soc Sciences to be equal to the corresponding multinomial logit estimate.

Table 6: Estimated major choice equation (23).

	Nat. Sciences	Soc. Sciences	Humanties	Education	Health
Estimated Probability	0.221	0.531	0.025	0.091	0.132
Probability differences by adding 1 standard deviation to the intial wage and growth rate of:					
Nat. Sciences	+0.100	-0.068	-0.003	-0.012	-0.017
Soc. Sciences	-0.058	+0.123	-0.007	-0.024	-0.035
Humanties	-0.004	-0.009	+0.016	-0.001	-0.002
Education	-0.013	-0.030	-0.001	+0.052	-0.008
Health	-0.018	-0.043	-0.002	-0.007	+0.070
Probability differences by adding 1 standard deviation to the intial wage of:					
Nat. Sciences	+0.058	-0.040	-0.002	-0.007	-0.010
Soc. Sciences	-0.036	+0.076	-0.004	-0.014	-0.021
Humanties	-0.002	-0.005	+0.009	-0.001	-0.001
Education	-0.007	-0.017	-0.001	+0.029	-0.004
Health	-0.010	-0.024	-0.001	-0.004	+0.040
Probability differences by adding 1 standard deviation to the wage growth of:					
Nat. Sciences	+0.037	-0.025	-0.001	-0.004	-0.006
Soc. Sciences	-0.023	+0.050	-0.003	-0.010	-0.014
Humanties	-0.001	-0.003	+0.005	-0.000	-0.001
Education	-0.004	-0.011	-0.000	+0.018	-0.003
Health	-0.006	-0.015	-0.001	-0.003	+0.025

Estimates based on the MNLogit estimated presented in Table 6.

Table 7: Estimated probabilities and probability difference of major choice.

Panel A		Spec. A1	Spec. A2	Spec. A3	Spec. A4
Expected log initial wage	(θ_1)	0.431**	0.316**	0.279*	0.260*
		(0.170)	(0.162)	(0.165)	(0.157)
Expected annual wage growth	(θ_2)	0.066**	0.051*	0.047	0.053*
		(0.028)	(0.030)	(0.029)	(0.031)
Panel B		Spec. B1	Spec. B2	Spec. B3	Spec. B4
Expected log initial wage	(θ_1)	-9.060*	-7.528***	0.066	0.006
		(4.660)	(0.646)	(0.237)	(0.208)
Expected annual wage growth	(θ_2)	13.403*	0.944	0.080*	0.090
		(7.765)	(10.081)	(0.044)	(0.064)

Bootstrapped standard errors in parentheses. ***/**/* = significant at 1%/5%/10%.

Panel A: indicated variables constant across alternatives added.

Spec. A1: specification as in Table 6.

Spec. A2: as spec. A1 + female.

Spec. A3: as spec. A1 + female + ethnic background.

Spec. A4: as spec. A1 + female + ethnic background + test score.

Panel B: alternative measure for initial wages and wage growth.

Spec. B1: average initial wage and wage growth across majors and NLSY subsamples.

Spec. B2: average initial wage and wage growth across majors, gender and ethnic background.

Spec. B3: initial wage and wage growth based on regression without selectivity correction.

Spec. B4: initial wage and wage growth based on regression without unobserved heterogeneity (ν_i) in the major choice equation.

Table 8: Robustness checks using MNL.

5 Conclusion

In this investigation we considered whether considerations of future pay-offs drive the major choice of college graduates. We proposed a model relating future earnings to individual choices. We distinguished short-run wage effects, i.e. the initial wage, and longer run wage effects, i.e. wage growth. We estimated initial wages and wage growth using the NLSY data after having corrected for sample selection. Our correction of selectivity is dependent on a strong assumption: the individual fixed effect causes the problem (cf. eqs (10) and (11)). Semiparametric estimation methods were employed to estimate the model.

We found differences in starting wages and growth rates across majors. Test scores, often used as some measure of abilities, proved to have very different effects on wage growth and initial wages depending on the subsample considered. Stronger wage growth is experienced for higher test scores for the NLSY97 subsample, whereas test scores for the NLSY79 can have a negative impact. For initial wages, the 1979 respondents do better. Females experience less wage growth but there is no gender effect on initial wages.

Real wage growth rates behave differently. We find high real hourly wage growth for Health graduates and low for Education and Humanities graduates. All majors display the expected pattern of wage growth apart from Humanities. Wages tend to grow strongly at the beginning of the working career with the effect wearing off with time. For Humanities a different pattern was found. An explanation for this result is that the major Humanities has the smallest number of graduates in our sample and is the only major in which a fourth-order polynomial provided a significantly better fit than the third-order polynomial employed for the other majors. Initial wage also display differences across majors. The largest variation in the initial wages is found for Humanities whereas the starting wages in Health are much more concentrated.

In the estimation of the major choice equation, we used counterfactual initial wages and wage growth based on the modeling of wages. The fact that we find significant effects of short-run and long-run wage characteristics, reassures us of our ability to characterize counterfactual wages to a reasonable degree. We find the probability of selecting one major to be increasing with expected starting wages. For expected wage growth, significant positive effects are found for the parametric specifications only. Therefore, we tentatively conclude that expected financial payoffs do play a role in the major choice of college students.

Our results are robust to alternative specifications that allow for individual-specific regressors and alternative specification of initial wage and wage growth.

Finally, regarding our estimation procedures, we find that both multinomial probit and the Manski semiparametric estimation do not guarantee to provide a global optimum. Multinomial probit often tends to unrealistically large variance and covariance estimates. Manski's maximum score estimator requires extensive computational time to find the global optimum even with few explanatory variables.

Appendix: Major Grouping

Assigned category	NLSY major
Natural Sciences	Mathematics
	Physics
	All Other Engineering
	Mechanical Engineering
	Electrical Engineering
	Chemistry
	Computer & Info Tech.
	Civil Engineering
	Chemical Engineering
	Engineering Tech.
	Earth and Other Physical Sci.
	Computer Programming
	Biological Sciences
	Multidisciplinary or General Sci.
	Agriculture and Agr. Science
Social Sciences	Economics
	Accounting
	Architecture
	Business Management and Admin.
	Family and Consumer Science
	Psychology
	Communications
	Other Social Science
	Area, Ethnic, and Civ. Studies
	Political Science
	History
	Art History and Fine Arts
	Public Administration and Law
	Social Work and Human Resources
	Journalism
Humanities	Foreign Language
	Music and Speech/Drama
	Letters: Lit, Writing, Other
	Philosophy and Religion
Education	Secondary Education
	Library Science and Education (Other)
Health	Misc. Business and Med. Support
	Other Med/Health Services
	Public Health (Physical and Mental)
	Nursing

References

Altonji, J. G., P. Bharadwaj, and F. Lange (2012). Changes in the Characteristics of American Youth: Implications for Adult Outcomes. *Journal of Labor Economics 30*(4), 783–828.

Altonji, J. G., E. Blom, and C. Meghir (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annual Review of Economics 4*(1), 185–223.

Altonji, J. G., L. B. Kahn, and J. D. Speer (2016). Cashier or consultant? entry labor market conditions, field of study, and career success. *Journal of Labor Economics 34*(S1), S361–S401.

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

Arcidiacono, P., J. Hotz, and S. Kang (2012). Modeling college major choices using elicited measures of expectations and counterfactuals. *Journal of Econometrics 166*, 3–16.

Baltagi, B. (2013). *Econometric Analysis of Panel Data* (5th ed.). John Wiley & Sons, Chichester.

Becker, G. (1975). *Human Capital: A Theoretical and Empirical Analysis*. National Bureau of Economic Research.

Beffy, M., D. FougÃ¨re, and A. Maurel (2012). Choosing the field of study in postsecondary education: do expected earnings matter? *The Review of Economics and Statistics 94*, 334–347.

Belzil, C. and J. Hansen (2002). Unobserved ability and the return to schooling. *Econometrica 70*(5), 2075–2091.

Benitez-Silva, H., M. Buchinsky, H. M. Chan, S. Cheidvasser, and J. Rust (2004). How large is the bias in self-reported disability? *Journal of Applied Econometrics 19*, 649–670.

Bertrand, M. and S. Mullainathan (2001). Do people mean what they say? implications for subjective survey data. *American Economic Review Papers and Proceedings 91*, 67–72.

Bound, J., C. Brown, and N. Mathiowetz (2001). *Measurement error in survey data*, Chapter of the Handbook of Econometrics, vol. 5, E. Leamer and J. Heckman (eds.), pp. 3705–3843. Elsevier Science, Amsterdam.

Briesch, R., P. Chintagunta, and R. Matzkin (2002). Semiparametric estimation of brand choice behavior. *Journal of the American Statistical Association 97*, 973–982.

Chen, S. (2008). Estimating the variance of wages in the presence of selectivity and unobserved heterogeneity. *The Review of Economics and Statistics 90*, 275–289.

Cosslett, S. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica 51*, 765–782.

Dahl, G. B. (2002). Mobility and the return to education: testing a Roy model with mulitple markets. *Econometrica 70*, 2367–2420.

Deming, D. J. and K. L. Noray (2018, September). STEM careers and the changing skill requirements of work. Working Paper 25065, National Bureau of Economic Research.

Efron, B. and J. Tibsharani (1993). *An introduction to the bootstrap*. Chapman and Hall.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica 60*, 505–531.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge University Press, Cambridge (Mass.).

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics 125*, 515–548.

Kane, T. J. and C. E. Rouse (1995). Labor-market returns to two and four years college. *American Economic Review 85*(3), 600–614.

Kaufmann, K. M. (2014). Understanding the income gradient in college attendance in mexico: the role of heterogeneity in expected returns. *Quantitative Economics 5*, 583–630.

Keane, M. P. and K. I. Wolpin (1997). The career decisions of young men. *Journal of Political Economy 105*(3), 473–522.

Lee, L. (1982). Some approaches to the correction of selectivity bias. *Review of Economic Studies XLIX*, 355–372.

Lee, L. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics 65*, 381–428.

Li, B. (2011). The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis. *Transportation Research Part B 45*, 461–473.

Maddala, G. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.

Manski, C. F. (1975). Maximum score estimation ot the stochastic utility model of choice. *Journal of Econometrics 3*, 205 – 228.

Mazza, J. and H. van Ophem (2018). Separating risk from uncertainty in education: a semiparametric approach. *Journal of the Royal Statistical Society, Series A 181*, 249–275.

Mincer, J. (1958). Investments in human capital and personal income distribution. *Journal of Political Economy 66*, 281–302.

Mincer, J. (1962). On-the-job-training: Costs, returns, and some implications. *Journal of Political Economy 70*, 50–79.

Neal, D. A. and W. R. Johnson (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy 104*(5), 869–895.

Newey, W. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal 12*, S217–S229.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society 56*(4), 931–954.

Ruder, A. I. and M. Van Noy (2017). Knowledge of earnings risk and major choice: Evidence from an information experiment. *Economics of Education Review 57*, 80–90.

Schofield, L. (2014). Measurement eror in the afqt in the nlsy79. *Economics Letters 123*, 262–265.

Siow, A. (1984). Occupational choice under uncertainty. *Econometrica 52*(3), 631–645.

Smith, A. (1938). *An Inquiry into the Nature and Causes of the Wealth of Nations* (5 ed.). New York: Random House.

Stinebrickner, T. and R. Stinebrickner (2012). Learning about academic ability and the college dropout decision. *Journal of Labor Economics 30*(4), 707–748.

Webber, D. A. (2014). The lifetime earnings permia of different majors: correcting for selection based on cognitive, noncognitive, and unobserved factors. *Labour Economics 28*, 14–23.

Willis, R. and S. Rosen (1979). Education and self-selection. *Journal of Political Economy 87*, S7–S36.

Wiswall, M. and B. Zafar (2015). Deterninants of college major choice: identification using an information experiment. *Review of Economic Studies 82*, 791–824.

Zafar, B. (2013). College major choice and the gender gap. *The Journal of Human Resources 48*, 545–595.