Does Risk Matter? A Semiparametric Model for Educational Choices in the Presence of Uncertainty

Jacopo Mazza [*]
School of Social Sciences - Economics
University of Manchester
M13 9PL Oxford rd. Manchester, UK

December 15, 2014

Abstract

Standard human capital theory suggests that individuals select into education in order to maximize their utility. If agents are risk averse, they select the educational level that minimizes future uncertainty. The possibility of self-selection complicates the identification of the causal contribution of education to uncertainty in future payoffs. In this paper the assumption of endogenous school choices due to concerns about future risk is tested and the importance of uncertainty in shaping schooling choices is assessed. Relying on a flexible semiparametric procedure allowing for self selection, bounds for the effect of field of study in college on uncertainty are estimated and, in a second stage, exploited for modeling schooling choices. The results of the empirical investigation do confirm that individuals self-select into education in order to minimize uncertainty and maximize returns. Only selection of Humanities type of majors is unaffected by risk or expected returns.

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

JEL classification: C14; C34; D81; J31

[*] Email: jacopo.mazza@manchester.ac.uk. All data and computer programs are available on request. I am grateful to Hans van Ophem and Joop Hartog for helpful discussions and suggestions. I am in sole charge of any error or omission.

1 Introduction

The enormous empirical literature on human capital and earnings stemming from the seminal works of Mincer (1958; 1962) and Becker (1975) often assumes utility maximizing agents selecting their educational level as a consequence of their expected present value of education. This successful approach postulates agents possessing an adequate knowledge on future payoffs of different types of educations and on their ability to successfully complete the educational path chosen. Obviously, investment decisions on education are taken under a considerable amount of uncertainty. Uncertainty regarding ones performance in school, uncertainty about the future labor market conditions and uncertainty about future macroeconomic environment, just to name a few. Incorporating these elements into the usual framework of schooling and career choices would be a natural relaxation of standard assumptions and would greatly improve the understanding of the mechanics of educational choice formation.

Surprisingly enough, empirical evidence on schooling choices under uncertainty is scarce at best (Altonji, 1993; Cunha et al., 2005; Zafar, 2011). Even scarcer is the body of literature assessing the role that concerns about non predictable future returns play in the selection of education. This seems at odds with recent literature on risk in education (Cunha et al., 2005; Lemieux, 2006; Chen and Khan, 2007; Chen, 2008; Mazza and van Ophem, 2010) treating self-selection into education, motivated by risk concerns on the part of choice makers, as given. In this framework, self-selection might arise as a consequence of risk aversion. The possibility of self-selecting into education complicates the identification of the specific parameters of interest. Proper risk, in fact, should be defined as that part of labor market performance which can not be anticipated by the individual, but each individual possesses some private information inaccessible to the researcher. If the private information is acted upon and, consequently, education is selected in order to minimize uncertainty, simple metrics such as the variance of error terms of a wage equation would confuse risk and private information.

In this article, I test the existence of self-selection into type of education triggered by distaste for risk and the role that uncertainty plays in shaping educational decisions. Before identifying the effects of risk on individuals preferences for field of study two hurdles must be cleared. First, potential self-selection needs to be accounted for. Second, wage variance corrected for self-selection has to be separated between risk and private information. Building up on recent developments of the literature on semiparametric estimators, this paper proposes a model for educational choices correcting for self-selection when uncertainty of future payoffs is accounted for and able to disentangle the separate contribution of uncertainty and unobserved heterogeneity.

The empirical strategy adopted falls into the growing literature on semiparametric estimation. As the common parametric techniques have come under closer scrutiny and received growing criticism (see, for example, Goldberger, 1983), a series of new semiparametric estimators for dichotomous choice models have been developed in the literature (Lee, 1983; Robinson, 1988; Cosslett, 1991; Ahn and Powell, 1993; Newey, 2009). On the other hand, polychotomous choice models have received considerably less attention. Dahl (2002) proposes a two-step semiparametric method correcting for sample selection bias in the case of multiple possible outcomes. I combine this semiparametric estimation method for unordered outcomes with a parametric method in the first stage. Ideally, I would like to avoid any distributional assumption for both error terms in the choice and outcome equation. In my case, as I need to decompose the variance of the wage equation in its different elements, some structure for the error terms is necessary. The estimation strategy adopted in the present work assumes normality only for the distribution of the disturbance term for the choice equation without imposing joint normality of the error terms. Furthermore, I extend the original model by introducing uncertainty of future payoffs in the choice formation routine.

Next to the obvious advantages of producing consistent parameter estimates in a fairly general set of data generating processes, the particular method adopted here presents some additional attractive features that could be easily extended to other polychotomous settings. In particular, consistency of the second stage estimation does not require an exclusion restriction as most other parametric or semiparametric estimators (Robinson, 1988; Cosslett, 1991; Newey, 2009) do. Since valid exclusion restrictions are, in practice, hard to come across (Bound et al., 1995) and exogeneity is often hard to justify and test, not having to depend on a valid instrumental variable can only increase the estimates reliability. Endogenous instruments, in fact, can lead to an amplification in parameters' estimates bias compared to simple OLS (Cameron and Trivedi, 2005).

To my knowledge, this is the first paper adopting a semiparametric strategy, able to assess the separate impact of risk and unobserved heterogeneity on unordered choices for type of education. The only other paper semiparametrically correcting for self selection and separately identifying risk and unobserved heterogeneity is Mazza and van Ophem (2010), while Chen (2008) accomplishes the same result, but strictly parametrically. Both works are only interested in gauging the causal effect of education on risk and not the effect of uncertainty on schooling choices. Additionally, this is the first paper that disentangles the various components of wage variance via a semiparametric estimator in a context for which a clear order of choices is not *a-priori* determined.

Theoretical advancement is not the only motivation behind the present research. Understanding the extent of the influence that uncertainty exerts on individuals choices is of direct interest for policy makers and sound empirical evidence on this matter is severely lacking. Consider, for example, an economy in which some particular occupation can not meet enough supply in the labor market due to excessive risk in the required education for accessing it. A government willing to propel a more efficient labor supply structure might consider the public provision of insurance coverage for those individual ready to undertake that particular educational path. Furthermore, if riskier human capital investments are leading to higher returns to education, and if poorer individuals avoid them due to the absence of the intrinsic financial buffer that family income offers, intergenerational and social mobility might be severely reduced.

The analysis, which exploits data from the National Longitudinal Survey of Youth (NLSY), proceeds in four steps. First, probabilities for major filed selection are estimated with a multinomial probit model. In the second step these probabilities serve as basis for the construction of the correction function to get consistent estimates for the wage equation. The correction functions enter the wage equation significantly, suggesting that self-selection exists. The results for the wage equation show that self-selection into education leads to a significant *downward* bias of OLS estimation for returns to education up to more than 100%. Third, the various elements of wage variance are either point estimated or bounded within some admissible range of values. Results confirm the well known increase in transitory earnings volatility for the US in the past twenty years and show how graduates in Science and Social Sciences disciplines are better immunized against macroeconomic shocks compared to graduates in Humanities and Health and Education. At the same time, those same type of educations protect against total uncertainty defined as the sum of transitory volatility and individual specific permanent volatility. In the final step, the responsiveness of educational choices to differences in risk associated with the distinct major type is tested. I find that the theoretical prediction of a negative impact of risk and a positive impact of expected returns on educational selection is confirmed for three out of four educational groups, the only exception being the Humanities group.

2 Theoretical model

I present here a four steps model for the estimation of the impact of future wage uncertainty on educational choices. The model builds on Dahl (2002) who proposes a semiparametric estimation method for polychotomous choice models. The original model concerns internal migration choices in the US where self-selection raises from differentials in returns for education in the 51 US states. In my framework choices are limited to four educational categories and self-selection occurs as a consequence of individual specific tastes for education. Additionally, the focus of my research is not centered on means returns to education, but on the dispersion of returns, thus, uncertainty is added to the original model.

The first steps of a four stages procedure consist in estimating the probability of selection into one of the four educational groups[1] - Humanities, Science, Social Science and Health and Education, these probabilities serve as basis for constructing four selection adjustments terms that in the second stage are included in a wage equation reestablishing the condition of zero mean on the error term allowing estimation by ordinary least squares. In the third step the real magnitude of risk is assessed and disentangled from private information. Finally, the assumption of individuals self-selecting into education as a consequence of comparative advantages is tested and the impact of uncertainty concerns on type of education selection is estimated.

2.1 A model for school choice and wages in the presence of uncertainty

In this section, I present a Roy (1951) model for multiple educational choice that builds on Dahl (2002) in its general structure, adapting the analysis to educational choices and introducing uncertainty on future payoffs.

Consider N individuals facing four possible choices for major type in college m: Humanities ($m_i = 1$); Sciences ($m_i = 2$); Social Sciences ($m_i = 3$); and Health and Education ($m_i = 4$). In this stylized world there are two periods. In the first period, after high school and conditional to wanting to acquire a college education, the individual selects the type of major that he wants to pursue according to his inclinations and the expected income that that specific type of education allows him to earn. In the second period, once a college degree has been attained, he enters the labor market and a stream of income is earned for T periods. Observing all relevant variables for schooling choice,

[1] The choice of these four college major categories is fairly standard in the literature. Additionally, many of the college major groups coded in the NLSY count little to no observations, thus some aggregation is necessary for the statistical analysis. See appendix for detailed major classification.

each individual (i) compares the benefits obtainable in each of the m categories and opts for the utility maximizing one, with utility being a function of expected earnings, earnings uncertainty and tastes affecting choices. Tastes affecting educational choice are potentially infinite. Among others they include tastes and inclination for a specific type of education, private information including individuals' own assessment on the riskiness of major m and individual specific risk attitude. A common feature of these factors is that they are all unobservable to the econometrician. How these personal characteristics translate in the labor market is not completely revealed to the choice maker even though private information allows him to form a more precise estimate for both the profitability and the uncertainty of incomes associated with each of the m categories compared to the econometrician who is unable to use the same information.

Formally, my model comprises two inter-related equations: an additively separable utility function (1) and a potential wage equation (2) for each major $m = 1, 2, 3, 4$:

$$E[V_{mit_0}|\nu_i] = \vartheta_1 E[y_{mit}|\ x_{it_0},\ \nu_i] + \vartheta_2 E[\tau^2_{mit}|\ x_{it_0},\ \nu_i] + \nu_i, \tag{1}$$

$$y_{mit} = \alpha_m + x_{it}\beta_m + \sigma_{mi}e_{mi} + \psi_{mt}\epsilon_{it},\ with\ (m = 1,\ 2,\ 3,\ 4). \tag{2}$$

In equation (1) the dependent variable $E[V_{mit_0}]$ is the expected utility that individual i attaches to major type m at time t_0, where the subscript 0 denotes the beginning of the first period. Utility is a function of expected wages ($E[y_{mit}|\ x_{it_0},\ \nu_i]$), expected uncertainty[2] ($E[\tau^2_{mit}|\ x_{it_0},\ \nu_i]$) and private information (ν_i). ϑ_1 and ϑ_2 are the coefficients associated with expected wages and uncertainty. Parameter ϑ_2 is the key parameter in this paper, its estimates are reported in table 8. Expectations are formed conditioning on individual observed (x_{it_0}) and unobserved (ν_i) characteristics evaluated at time t_0.

Equation (2) specifies individual log earnings (y_{mit}) in each of the four major types m as a function of a major type specific constant (α_m), a vector of individual characteristics (x_{it}), an individual fixed effect component ($\sigma_{mi}e_{mi}$) and an idiosyncratic transitory shock capturing macroeconomics or institutional changes and affecting individuals earnings ($\psi_{mt}\epsilon_{it}$). e_{mi} and ϵ_{it} are random unit root variables uncorrelated with each other. Note also that the loading factor σ in front of the individual fixed effect component is allowed to vary with type of education. In this way, considerations of comparative advantages enter individuals' decision mechanism. If the loading factor is equal across major types, the individual fixed effect is rewarded equally at all levels. For the scope of this paper the identification of the variance of potential wages ($\sigma^2_{mi} + \psi^2_{mt}$) plays a key role since this variance serves as basis for the construction of the risk coefficient whose effect on choices I want to estimate. It is important to note that while the shock term does not correlate either with observed or unobserved characteristics, the individual fixed effect does with both.

Selection of the preferred type of education is determined by considerations of comparative advantages depicted in equation (1). Formally, individuals choose the educational levels for which:

$$\begin{aligned} I_{mi} \quad &= 1 \qquad \text{if and only if } E[V_{mi}] = max(E[V_{1i}], ..., E[V_{4i}]), \\ &= 0 \qquad \qquad \qquad \qquad \text{otherwise} \end{aligned} \tag{3}$$

[2] The exact specification of $\tau^2_{mit_0}$ is provided in equation (6).

where I_{mi} is an indicator function assuming value 1 if that specific major is selected and 0 otherwise. Where $E[V_{mi}] = E[V_{mit_0}]$ since expectations are assumed to be age independent and therefore time subscript t_0 is omitted in the remainder of the paper for ease of notation.

The system of equations in (1) and (2) can not be directly estimated for three reasons: first, all the relevant variables for major choice are unobserved; second, private information affects both the choice of major type and the realization of wages introducing a selection bias in the estimation of the wage equation; third, in the data individuals are observed in only one of the four possible states thus the estimation of the determinants of major choice requires generating counterfactual earnings and uncertainty, accounting for self-selection, for the other three options. Self-selection is treated in section 3.1, counterfactual imputation is treated in section 6 while for the identification of the unknown parameters σ^2_{mi}, τ^2_{mi} and ν_i some additional assumption regarding the functional form are necessary.

More specifically, I need to specify how unobserved heterogeneity (ν_i) relates to the individual specific permanent component ($\sigma_{mi}e_{mi}$). I indicate the correlation term between the two with (ρ_m) and in equation (4), following Mazza and van Ophem (2010), I define a linear relation for the conditional expectations of the two:

$$\sigma_{mi}e_{mi} = \gamma_m\nu_i + \xi_{mi}, \tag{4}$$

where I assume that: $Var[e_{mi}|x_{it}] = \sigma^2_{mi}$, $Var[\nu_i] = \sigma^2_\nu$, $Cov[e_{mi}, \nu_i] = \gamma_m = \rho_m\sigma_m\sigma_\nu$, $E[\xi_{mi}|\nu_i] = 0$ and $Var[\xi_{mi}] = \sigma^2_\xi$. As in Willis and Rosen (1979), the correlation coefficient is not restricted to assume positive values allowing either positive or negative selection into type of education. In the presence of positive selection (i.e.: $\rho_m > 0$) a high predisposition for a specific type of education translates into higher wages in the labor market, the opposite occurs in case of negative selection (i.e.: $\rho_m < 0$). The linear assumption is needed for the separate identification of wage uncertainty and unobserved heterogeneity.

Using these distributional assumptions, an equation for expected wages and expected uncertainty from the individual standpoint can be derived:

$$E[y_{mi}|x_i, \nu_i] = \alpha_m + x_i\beta_m + \gamma_m\nu_i, \tag{5}$$

$$\tau^2_{mit} = Var[\sigma_{mi}e_{mi} + \psi_{mt}\epsilon_{it}|\ x_{it},\ \nu_i] = \sigma^2_{mi}(1 - \rho^2_m\sigma^2_\nu) + \psi^2_{mt}. \tag{6}$$

This formulation illustrates the contribution of the parameter ν_i to wage expectations and, through the correlation coefficient ρ_s, to personal uncertainty. Regarding the first relationship, we can easily see from equation (5) that in the presence of positive selection individuals with a high degree of predisposition for a specific type of education are rewarded in the labor market while the opposite occurs in the case of negative selection. On the other hand, expression (6) illustrates the channel through which the unobserved schooling factor relates to the uncertainty components. In fact, if the correlation between unobserved schooling factor (ν_i) and the fixed individual effect $\sigma_{mi}e_{mi}$ is perfect (i.e.: $\rho_m = 1$) individuals can predict perfectly how their own inclinations translate in the labor market and uncertainty is only caused by variance in transitory shocks (ψ^2_{mt}). On the other hand, when correlation is absent (i.e.: $\rho_m = 0$) the individual does not posses any additional information compared to the econometrician on how his unobserved abilities affect his wages in the future and uncertainty equates observed wage variance.

Using the relation expressed in (5) I define an equation for the deviation of individuals' expected wages from population average earnings, obtaining:

$$E[y_{mit}|\ x_{it}, \nu_{mi}] - E[y_{mit}|\ x_{it}] = \gamma_m \nu_i. \tag{7}$$

Equation (7) simply states that the deviation of individual expected earnings from the average students in category m given his observable characteristics and unobservable tastes for schooling is the individual specific error term $\gamma_m \nu_i$ in equation (5). The transitory shock component in equation (2) is differenced out since it is supposed to be uncorrelated with individual characteristics and thus it affects all individuals with $m_i = m$ equally. The equality makes clear that deviations from the population mean are a function of the specific schooling tastes expressed by ν_i and how these tastes correlate with individual specific component.

I define a similar equation for the deviation of individuals taste for education from the population average:

$$\nu_i - E[\nu_i|\ x_i] = w_{mi}, \tag{8}$$

w_{mit} is an error term for individual deviations from mean tastes. Tastes for type of education m include a number of possible variables such as the inclination for a specific subject, anticipated likelihood of obtaining a degree for major m, or the anticipated individual wage risk associated with that type of education.

I can now rewrite expression (1) in terms of population means and individual specific error component:

$$E[V_{mit}] = E[V_{mt}] + s_{mi} \tag{9}$$

where $E[V_{mt}] = E[y_{mit}|x_{it}, m_i = m] + E[\nu_i|x_i]$ and $s_{mi} = w_{mi} + \gamma_m \nu_i$. In the selection literature V_{mt} is referred to as the subutility function. I assume the error term s_{mit} to be multivariate normally distributed with mean zero and covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & \dots & \sigma_{14} \\ \vdots & \sigma_2^2 & & \\ \vdots & & \sigma_3^2 & \\ \sigma_{41} & \dots & \dots & \sigma_4^2 \end{bmatrix} \tag{10}$$

The selection rule expressed in equation (3) can now be rewritten as:

$$\begin{aligned} I_{mi} \quad &= 1 \qquad \text{if and only if } V_m + s_{mi} \geq V_r + s_{ri}\ \forall r \neq m, \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned} \tag{11}$$

Thus, earnings are observed only for the utility maximizing choice and if the selection equations outlined in (11) are satisfied simultaneously. Equations (1)-(11) describe a Roy model of schooling and earnings with multiple choices and in the presence of uncertainty. For this paper the main equation of interest is equation (1) which, after the necessary transformation, is estimated in section 6.

3 Semiparametric estimation of a Roy model with multiple sectors

The most common procedure for estimation of models with self-selection and binary outcomes is the Heckman selection model (Heckman, 1974; 1976; 1979). The model presented here allows for four possible choices. In case of multiple options, the approach depends on the structure of the outcomes that can either be ordered according to some natural and evident structure, or unordered, in case this ordering is not apparent. In the first case, the selection correction term is usually derived from an ordered probit regression in the first stage which, after some transformation, is then included in the outcome equation (Vella, 1998) obtaining consistent estimates of the β's. In the second case, when no ordering of choices is possible, the first stage can be estimated via a conditional logit model or its extension the nested logit model (McFadden, 1984; Trost and Lee, 1984; Falaris, 1987).

All these methods rely on heavy assumptions on the distribution of the error terms in the choice and selection equations. If the true joint distribution is not correctly specified and it is different from the designated one, the estimated parameters in the outcome equation are severely biased (Goldberger, 1983) with the level of bias increasing as the self-selected sample size increases(Dahl, 2002). These criticisms generated a fertile line of research proposing alternative methods imposing limited distributional assumptions (Cosslett, 1983; Gallant and Nychka, 1987; Robinson, 1988; Ahn and Powell, 1993; Powell, 1994; Newey, 2009).

All these methods address binary choice models and, similarly to their parametric counterparts, imply estimation in two steps[3]. In the first step, some nonparametric or semiparametric estimator of the parameters in the choice equation, for which the distribution of the error term remain unspecified, is used. These estimates form the basis for the construction of a 'single-index' correction function $g(.)$ which is then included in the second stage allowing consistent estimates of the parameters in the outcome equation.

If research on semiparametric estimation methods for binary response models has received some attention in recent literature, very little effort has been dedicated to the semiparametric estimation of polychotomous choice models. One of the few exceptions is Dahl (2002) who proposes a model for unordered choices regarding migration decisions.

I exploit Dahl's work and adapt it to the different needs that my research question poses. The main methodological difference between my and Dahl's framework resides in the structure of the error term in the choice equation. In fact, in order to be able to separate risk from private information, the error term in the first stage is assumed to be normally distributed.

Additionally, Roy models based on utility maximization, such as the present one, present a specific challenge: the correct specification of the subutility function V_m and the choice of variables to include in it. In my framework, a plethora of variables are potential candidates for inclusion

[3] For a textbook discussion of parametric and semiparametric selection models see Cameron and Trivedi (2005).

and many of these variables are either unobservable or non perfectly measurable. The model that I present here sidesteps the estimation of underlying parameters of the subutility function and thus, does not require the correct specifications of tastes.

3.1 Schooling probabilities as sufficient statistics in single and multiple-index models

The estimation method that I present here for schooling choices is building on previous works by Dahl (2002), Lee (1983) and Ahn and Powell (1993) on semiparametric estimation methods.

As already noted by Heckman and Robb (1985) and Ahn and Powell (1993) in single-index selection models the selectivity bias can be expressed as the probability of selection given covariates. This follows from the fact that in latent index models, the mean of the error term in the outcome equation for the selected sample is an invertible function of the selection probability (Dahl, 2002). Ahn and Powell exploit this fact in order to avoid estimation of an unknown distribution function for the selection errors. Dahl extends this idea to multiple-index models providing a relatively simple semiparametric correction for polychotomous selection models. In this section I first show the formulation of Ahn and Powell (1993) for single-index models and then the extension that Dahl provides to multiple-index.

Considering the theoretical model presented in section 2.1 I rewrite the earnings equation as:

$$y_{mit} = \alpha_m + x_{it}\beta_m + \sum_{m=1}^{M} [I_{mi}\varsigma_m(V_m - V_r, \ldots,\ V_M - V_r)] + \eta_{mit}. \tag{12}$$

In this formulation $\varsigma_m(.) = E[u_{mit}|V_m - V_r, \ldots, V_M - V_r]$, η_{mit} is a zero mean error term in the selected sample and I_{mi} is the usual indicator function assuming value 1 if $m_i = m$. This is a partially-linear, multiple-index model since the control functions ς_m are unknown functions of the multiple index $V_m - V_r, \ldots, V_M - V_r$.

Let's now define the joint density function of the error term in equation (2) and in equation (11) describing the selection criteria, as: $f_m(u_{mit},\ s_{mi} - s_{ri}, \ldots, s_{Mi} - s_{ri})$. Lee (1983) shows that $f_m(u_{mit},\ s_{mi} - s_{ri}, \ldots, s_{Mi} - s_{ri}|V_m - V_r, \ldots, V_M - V_r) = g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi}|V_m - V_r, \ldots, V_M - V_r)$[4]. Dahl takes advantage of Lee's results and imposes the following index-sufficiency assumption:

$$\begin{gathered} g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi}|V_m - V_r, \ldots, V_M - V_r) = \\ g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi}|p_{mi}) \end{gathered} \tag{13}$$

where p_{mi} is the probability that individual i selects major type m given the vector of subutilities differences $V_m - V_r, \ldots, V_M - V_r$. Equation (13) assumes that $p_{mi} = p_{mi}(V_m - V_r, \ldots, V_M - V_r)$ exhausts all the information about how the differences in subutility functions influence the joint distribution of the error term in the outcome equation and $max_r(V_r - V_m + s_{ri} - s_{mi})$ contained in the sample, which is equivalent from stating that the conditional distribution of u_{mit} and $max_r(V_r - V_m + s_{ir} - s_{mi})$ can depend on the conditioning variables only through the single index p_{mi}.

[4] To see how the equality is derived remember the selection criteria expressed by equation (11). That relation states that selectivity bias in y_{sit} is driven by the event that the maximum of the collection of random variables $V_r - V_m + t_{ri} - t_{mi}, \ldots,\ V_M - V_m + t_{Mi} - t_{mi}$ is less than or equal to zero.

The single index p_{mi} is the probability of each individual first best education choice; in other words it is the major choice observed in the data and can be rewritten as:

$$p_{mi} = Pr(I_{mi} = 1|V_m - V_r, \ldots, V_M - V_r). \tag{14}$$

The differences in subutility functions determine the choice for type of education, thus they need to be accounted for when estimating p_{mi}. Using equation (13) the earnings equation expressed in (12) can be rewritten as:

$$y_{mit} = \alpha_m + x_{it}\beta_m + \sum_{m=1}^{M}[I_{mi}\lambda_m(p_{mi})] + \omega_{mit}, \tag{15}$$

where for each supergroup m, $\lambda_m(.)$ is an unknown function of the single index $p_{mi}$and $E[\omega_{mit}|x_{it}, p_{mi}, I_{mi} = 1] = 0$ by construction[5].

All the results reported until this point were already obtained by Lee (1983). The specific contribution of Dahl (2002) is extending the single index correction function in equation (15) to multiple index framework.

Dahl's intuition is that, subject to the invertibility condition:

$$\begin{array}{c} g_m(u_{mit},\, max_r(V_r - V_m + w_{rit} - w_{mit}|V_m - V_r, \ldots, V_M - V_r) = \\ g_m(u_{mit},\, max_r(V_r - V_m + w_{rit} - w_{mit}|p_{im}, \ldots,\, p_{iM}) \end{array} , \tag{16}$$

which simply implies that multiple education type choice probabilities contain the same information as the difference in subutilities functions, the earnings equations can be rewritten as multiple-index, partially linear models that depend on all M schooling probabilities:

$$y_{mit} = \alpha_m + x_{it}\beta_m + \sum[I_{mi}\mu_m(p_{im}, \ldots,\, p_{iM})] + \eta_{mit} \tag{17}$$

where $\mu_m(.) = E[u_{mit}|p_{mi}, \ldots,\, p_{Mi}] = E[u_{mit}|V_m - V_r, \ldots, V_M - V_r]$. The assumption contained in equation (13) reduces this equivalence by imposing that only the probability of the utility maximizing choice matters. The assumption can be relaxed allowing for other probabilities beside the first-best choice to influence the distribution of g_m. Indicating with $\vec{q}$ the subset, or full set, of schooling probabilities $\{p_{im}, \ldots,\, p_{Mi}\}$, a less restrictive assumption can be written as:

$$\begin{array}{c} g_m(u_{mit},\, max_r(V_r - V_m + w_{rit} - w_{mit}|V_m - V_r, \ldots, V_M - V_r) = \\ g_m(u_{mit},\, max_r(V_r - V_m + w_{rit} - w_{mit})|p_{im},\, \vec{q}) \end{array} . \tag{18}$$

From this expression the earnings equation can be rewritten as a multiple-index, partially linear model, where the bias correction is an unknown function of the revealed first-best choice plus a few other chosen probabilities.

In my application of this model to type of major choice the number of probabilities, other than the revealed choice, candidate for inclusion is necessarily limited. I can then estimate a very rich model with the inclusion of all major type selection probability and compare it with the most parsimonious model possible. This is the way I proceed and describe in Section 5.3. The choice of

[5] See Dahl (2002) for analytical proof of this result.

these probabilities implies the following distributional assumption:

$$\begin{array}{c} g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi}|V_m - V_r, \ldots, V_M - V_r) = \\ g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi})|p_{im}, \ldots\ p_{iM}) \end{array}, \tag{19}$$

and the following earning equation:

$$y_{mit} = \alpha_m + x_{it}\beta_m + \sum_{m=1}^{M}[I_{im}\lambda_m(p_{im}, \ldots\ p_{iM})] + \omega_{mit} \tag{20}$$

I refer to$\lambda_m(.)$ as the selection correction function which is an unknown function of four probabilities $p(m_i = 1)$, $p(m_i = 2)$, $p(m_i = 3)$ or $p(m_i = 4)$.

4 Empirical estimation

In the previous section I have outlined the general structure of a semiparametric model in a polychotomous choice framework in the presence of self-selection as presented by Dahl (2002) and my adaptation to the present application for college major choice. OLS estimates of equation (20) produce consistent estimates for the parameters of interests.

The focus of this paper is first obtaining consistent estimates for the level of unanticipated wage dispersion that each schooling level entails and then, in a second step, assessing how heavily individuals weigh the risk factor when taking schooling decisions. Both steps need to account for individuals' private information and thus, intrinsic to risk estimation, is the identification of private information. In the following section I illustrate the empirical implementation choices and the necessary steps for identification of the transitory component of wage variance (ψ^2_{mt}), the permanent component of wage variance (σ^2_{mi}), risk (τ^2_{mit}) and private information (ν_i) starting from the wage equation corrected for self-selection presented in (20).

4.1 Estimation for the selection probabilities

The model presented hinges on the assumption that the researcher can consistently estimate the probabilities associated with each schooling choice for each individual. The most common procedures adopted in the literature for estimation of selection probabilities are the conditional logit model and the ordered probit model in case of unordered or ordered outcomes respectively. The main drawbacks of these two methods are their dependence on heavy distributional assumptions[6].

Ideally, I would like to semiparametrically estimate both stages. The literature on semiparametric estimators in the presence of unordered choice structure is very scarce (Matzkin, 1993; Dahl, 2002; Bayer et al., 2011) and for none of these estimators the full asymptotic properties are derived. As evident from the expressions for σ_{mi}, τ_{mit} and δ_{mi} estimates for the conditional and unconditional variance of the error term in the choice equation are needed if the variance of wages has to be decomposed between our parameters of interest. Therefore I estimate the first stage and the probabilities of schooling selection via a multinomial probit model, assuming normality for

[6] An additional and unattractive property of the conditional logit model is the independence of irrelevant alternatives .

the distribution of the disturbance term in the secondary equation, but avoiding to impose joint normality on the error terms for the selection and outcome equation.

Compared to the conditional logit model the multinomial probit has the considerable advantage of allowing for the error terms for the different options to be correlated eluding the independence of irrelevant alternatives (IIA) assumption imposed by multinomial or conditional logit models.

4.2 Identifying the two components of wage variance

Intra-educational wage variance can result from observed heterogeneity expressed by β_m in equation (2) or unobserved heterogeneity which is captured by the error term in the same equation.

In this model the error term in equation (2) is composed by an individual specific fixed term ($\sigma_{mi}e_{mi}$) and an idiosyncratic shock ($\psi_{mt}\epsilon_{it}$); the variance of these two elements ($\sigma^2_{mi} + \psi^2_{mt}$) captures the unobserved part of wage variance which, in turns, includes both risk and private information. This part of wage variance is my target of identification in the first step.

Starting from the same premises Chen (2008), in a parametric setting, and Mazza and van Ophem (2010), semiparametrically, derive an expression for variance of wages. Adapting their results to the present framework with utility maximization I obtain:

$$Var[\sigma_m e_{mi} + \psi_{mt}\epsilon_{it}|p_{im}, \dots\ p_{iM}] = \sigma^2_m(1 - \rho^2_m\delta_{mi}) + \psi^2_{mt}. \tag{21}$$

δ_{mi} is referred to as the truncation adjustment needed in order to retrieve the untruncated distribution of wage variance. Following Lee (1982; 1983) and Maddala (1983) and given the distributional assumptions in (10) its analytical expression is given by:

$$\delta_{mi} = 1 - Var[\nu_i|p_{im}, \dots\ p_{iM}] = \lambda^2 - \frac{-\phi(z_i\varphi)}{\Phi(z_i\varphi)}$$

Where $\lambda = E[\nu_i|p_{im}, \dots\ p_{iM}] = \frac{\phi(z_i\varphi)}{\Phi(z_i\varphi)}$. The probabilities for schooling selection are estimated with a multinomial probit model given the distributional assumptions in (10)[7]. δ_{mi} determines whether observed wage inequality overstates or understates potential wage inequality. If $\delta_{mi} > 0$ observed wage inequality overstates potential inequality and vice versa in case $\delta_{mi} < 0$.

In order to be able to disentangle the transitory shock component from the permanent component a panel data structure is essential. In fact, an individual fixed-effect model differences out the time invariant permanent component $\sigma_{mi}e_{mi}$ so that the unexplained part of wage variance in the model can be attributed to external and unanticipated idiosyncratic shocks which is one part of wage risk properly defined.

In the present framework a fixed-effect model for individual earnings takes the form:

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)\beta_m + (\kappa_{mit} - \bar{\kappa}_{mi}) \qquad \text{if } m_i = m, \tag{22}$$

$\overline{y}_i$, $\overline{x}_i$ and $\bar{\kappa}_{mi}$ denote the average of individual earnings, time varying covariates and error term, respectively, over the time period taken into consideration and $\bar{\kappa}_{mi} \equiv \psi_{mt}\epsilon_{it}$. Consequently, the

[7] For derivation see Maddala (1983).

transitory component of wage variance ψ^2_{mt} is identified as the variance of the error term in equation (22).

The next step is identifying the permanent component of wage variance σ^2_{mi}. The parameter is identified with a between-individual model based on equation (20):

$$\overline{y}_i = \alpha_m + \overline{x}_i\beta_m + \sum[I_{im}\lambda_m(p_{im}, \dots\ p_{iM})] + \overline{\omega}_{mi} \tag{23}$$

With the inclusion of the correction term, the between-individual model can be consistently estimated by OLS since $E[\overline{\omega}_{mi}|x_i, \gamma_m] = 0$. Mazza and van Ophem (2010) show that with only the assumption of linearity on the error terms discussed in section 2.1, it is possible to obtain an analytical expression for the permanent component corrected for truncation and self-selection:

$$\widehat{\sigma}^2_{mi} = \widehat{Var}[\omega_{mi}|\overline{x}_i, m_i = m, z_i] + \gamma_m\hat{\delta}_{mi} - \sum_t \widehat{\psi}^2_{mt}/\overline{T}. \tag{24}$$

As in Chen (2008) and Mazza and van Ophem (2010) $\widehat{Var}[\omega_{mi}|\overline{x}_i, m_i = m, z_i]$ is estimated as the mean squared error of the between individual model in equation (23), $\overline{T} \equiv (\sum_i T_i^{-1}/N)^{-1}$ and $\hat{\delta}_{mi}$ is the truncation adjustment. The only parameter that remains unidentified is γ_m. The very flexible structure of the error terms and of the correction function selected in this application hampers point identification of this parameter. In section 4.3 I show how this parameter can be bounded within a given interval of admissible values. As I show in the last section of the present work, these bounds are informative enough for determining the contribution of the permanent component to education selection.

I have now point identified or bounded both elements of wage variance. Remember that since individuals posses private information, the permanent component $\widehat{\sigma}^2_{im}$ bounded in (24) cannot be imputed completely to proper risk as the individual can foresee part of it. The proper expression for risk, defined as the unforeseeable part of wage variance from the individual standpoint, is $\tau^2_{mit} = Var[u_{mit}|z_i; x_{it}, \nu_i] = \sigma^2_{mi}(1 - \rho^2_m\delta_{mi}) + \psi^2_{mt}$. Remembering that ρ_m expresses a correlation and can thus vary only between -1 and 1, I can conclude that all elements for bounding the risk parameter τ^2_{mit} are at hand.

4.3 Separate identification for risk and unobserved heterogeneity

For the purpose of this paper it is essential to separately identify the risk coefficient τ^2_{mit} from the unobserved heterogeneity component ν_i and further split τ^2_{mit} into transitory shock ψ^2_{mt} and permanent component of wage variance σ^2_{mi}.

Transitory shocks are easily identified as the variance of the error term in equation (22). Identification of the permanent component σ^2_{mi} is more complicated. The complete specification of the permanent component given in equation (24) includes the coefficient for the selectivity adjustments differentiated by schooling type γ_m. Therefore, point identification of σ^2_{mi} presupposes the possibility of separately identify one selectivity adjustment per schooling level. This is not possible in the context of this paper where the correction function is a series of polynomial expansions.

Instead of pursuing point identification for the permanent component of wage variance, I derive informative lower and upper bounds for the range of possible values that this component can assume. I decide to trade off precision of identification, that would be possible if stricter assumptions on the structure of the error terms were imposed, with generality of results that in my case do not rely on the specific distributional form chosen. I believe that these bounds are still informative since they allow for estimation of schooling choices based on comparative advantages which is the final purpose of the present work. To see how the permanent component can be bounded consider equation (24) and rearrange it to obtain:

$$\sigma_{mi}^2 = \frac{\widehat{Var}[\omega_{mi}|\overline{x}_i, m_i = m, z_i] - \sum_t \widehat{\psi}_{mt}^2/\overline{T}}{1 - \rho_m^2 \hat{\delta}_{mi}} \tag{25}$$

the numerator of this fraction is easily identified[8], following Mazza and van Ophem (2010) I can also identify δ_{mi} as $1 - Var[\nu_i|z_i,\ I_{mi} = 1]$ where $Var[\nu_i|z_i,\ I_{mi} = 1] = E[\nu_i^2|z_i,\ I_{mi} = 1] - E[\nu_i|z_i,\ I_{mi} = 1]^2$. The only unknown in this equation is the squared correlation coefficient ρ_m^2 which can be bounded between 0 and 1. In case of no correlation between wages and the unobserved schooling factor (i.e.: ρ_m=0) the permanent component is simply the variance of the error term in the between individual model of equation (24) minus the transitory shock; thus no private information is exploited for minimizing wage variance. The other extreme is given for perfect correlation (i.e.: ρ_m=1). In this case, the width of the bounds depends on the magnitude of $\hat{\delta}_{mi}$.

4.4 Estimating the correction function

In a semiparametric framework the correction function is left unspecified. Different methods exist for estimation of an unknown function. In this paper I employ a series expansions for estimation of the unknown function. The method was first introduced by Newey (1997). The approximation for individuals in major category m is:

$$\lambda_m(p_{mi}, \dots\ p_{iM}) \backsimeq \sum_{q=1}^{Q} \kappa_m^q b_m^q(p_{mi}, \dots\ p_{iM}) \tag{26}$$

where the functions $b_m^q(.)$ are referred to as the basis functions. Common choices for basis functions are the terms of a polynomial or Fourier series. In my estimation I chose the polynomial expansion so that Q denotes the number of terms in the approximating series. I now have a model that is linear in parameters and thus estimable by ordinary least squares. The number of series expansions should increase as the sample size increases, in practice, there is no standard procedure that the researcher can follow for choosing the correct number. Additionally, consistency for the parameters estimation in the outcome equation requires the number of probabilities entering the basis function to be sufficiently large. The probabilities for each individual and for each schooling categories are calculated with a multinomial probit in the first stage.

[8] See section 4.2.

5 The causal impact of risk on education

My empirical estimation for the importance of concerns on risk on the choice of education proceeds in four steps. In the first, the probability of major type selection is estimated following the procedure explained in section 4.1; these probabilities are then used for calculating the basis functions, and thus the selectivity correction terms, in equation (26) in the second step. The correction functions are included in the wage equation obtaining estimation corrected for selectivity, these estimations serve as basis for identification of permanent component σ_{mi}^2, transitory component ψ_{mt}^2, private information ν_i and risk τ_{mit}^2 as described in section 4.2. In the last step the responsiveness of major type selection probabilities to differences in risk level, corrected returns to education and other amenities, are estimated.

5.1 Data

For my purpose I use the National Longitudinal Survey of Youth 1979 (NLSY79). The NLSY is a longitudinal study of a representative sample of U.S. citizens who were 14 to 22 years old in 1979 when the survey first started. The sample size is 12,686 strong and it includes a wide variety of economic, sociological and psychological measures. Particularly important for my study, the survey includes information about the major selected in college for those individuals who proceed to tertiary education. The survey begun in 1979 and it is still ongoing as the last available wave dates back to 2008. The cohort was interviewed annually until 1994 and biennially thereafter.

Since my analysis regards major choice in college, I restrict the sample analyzed to males and females who attended college, this reduces my sample to 6,325 individuals. The first wave considered in my analysis is that of 1990 so that all individuals in the sample have already terminated their studies and are entering the work force. Observations are organized in 11 subsequent waves until the last available survey of 2008.

My model counts two dependent variables: major choice for the selection probabilities and earnings for the wage equation. Major in college is recorded as a four digit code distinguishing among the various fields of study[9] (e.g.: Biological Sciences, Engineering, Business and Management, etc.) and sub fields within the bigger field (e.g.: Microbiology, Chemical Engineering, Banking and Finance etc.). Earnings are expressed as the logarithm of hourly earnings in the period considered translated in 2008 dollars. The historical series for the Consumer Price Index (CPI) in the US for the period considered is obtained from the Bureau of Labor Statistics[10].

The information contained in the NLSY allows me to control for gender, ethnic background, family income when the respondent was 17 years old or as close to 17 as possible (in 2008 dollars), parents' levels of education, ability measured by the Armed Forces Qualification Test (AFQT) and dummies for geographical characteristics for the area of origin at age 17[11]. The AFQT is a series of four tests in mathematics, science, vocabulary and automotive knowledge. The test was administered in 1980 to all subjects regardless their age and schooling level. For this reason it can

[9] For a detailed description of the NLSY79 major classification see the appendix.

[10] source: ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt (accessed 11/07/2011)

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

include age and schooling effects in the ability index that the test is meant to construct. To correct for these undesired effects, I follow Kane and Rouse (1995) and Neal and Johnson (1996). First I 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.

The choice variable deserves some further discussion. The multinomial probit estimation procedure becomes intractable with the standard statistical package used[12], when the possible outcomes exceed four. Therefore, I grouped the different major as defined in the NLSY in four big categories: Humanities, Sciences, Social Sciences and Health and Education[13]. In this way I obtain four unordered categories that can be estimated via a multinomial probit procedure that allows for correlation of errors.

All the control variables used in the first stage[14] are also added to the between individual model in equation (23). In addition to these common variables, work experience is added as a time varying control in both the between individuals and fixed-effect estimation. In case information for any of the control variables is lacking the observation is dropped. For this reason I delete 319 individuals lacking information about the AFQT test score, 748 about parents education, 947 without information for family income and 647 whose information for earnings in the labor market is lacking. The final balanced panel counts 3,664 individuals observed in 11 waves generating 40,304 individual-year pair observations[15].

Descriptive statistics for the entire sample as well as for the four major categories appear in table 1 and table 2. The tables reveal sufficient variation in individuals own characteristics and background. Graduates from Social Sciences constitute the largest group in my sample and Humanities graduates the smallest, the other two groups of Sciences and Health and Education are quite balanced.

It is evident that people graduating from Humanities belong to families with a more favorable economic and educational background. Both mother's and father's education, as well as family income, are at their highest for this category. Additionally, AFQT score is also higher for them, while the share of ethnic minorities is the lowest among the four categories. The opposite occurs in the case of Health and Education group which is at the bottom for parents education, family income and ability measure.

It is also worth noting how ethnic minorities are overrepresented and that the majority of individuals in my sample were brought up in an urban environment.

5.2 Step 1: Schooling choice first stage estimates

The individuals probabilities to chose one of the four fields of study in college serve as basis for the construction of the correction functions $b_m^q(.)$ in equation (26). The first stage estimates for the multinomial probit model described in section 4.1 from which the choice probabilities are derived

[12] Stata version 10.0.

[13] See appendix for the exact definition of these categories.

[14] See section 4.1.

[15] A simple probit analysis for the probability of dropping out of my sample due to lack of information shows how females and ethnic minorities are less prone to attrition than white males while family income and AFQT score are very precisely estimated to have a 0 effect. All coefficients for the other observable characteristics are not significant. Estimation results available on request.

Table 1: Summary statistics: time invariant variables

Variable	Total sample	Humanities	Sciences	Social Sciences	Health & Education
Percentage of total sample		9.7	27.8	37.1	25.4
		(.29)	(.45)	(.48)	(.43)
Background and ability					
Female	.55	.50	.36	.55	.77
	(.49)	(.50)	(.48)	(.50)	(.42)
African American	.23	.18	.23	.25	.23
	(.42)	(.39)	(.42)	(.43)	(.42)
Hispanic	.15	.17	.16	.13	.17
	(.36)	(.38)	(.36)	(.34)	(.37)
AFQT score (adjusted)	54.20	57.00	55.55	54.91	50.59
	(27.29)	(28.23)	(27.38)	(26.97)	(27.01)
Mother's years of schooling	11.75	12.06	11.81	11.76	11.54
	(3.01)	(3.18)	(2.90)	(3.01)	(3.07)
Father's years of schooling	11.94	12.51	11.94	12.03	11.59
	(3.82)	(4.12)	(3.83)	(3.67)	(3.85)
Family income (in 2008 dollars)	38,443.4	39,782.29	36,696.87	40,472.96	36,878.01
	(29,403.64)	(32,440.4)	(26,449.62)	(31,555.33)	(27,788.65)
Geographic region grew up in:					
Urban	.81	.81	.81	.81	.79
	(.39)	(.39)	(.39)	(.39)	(.40)
Northeast	.19	.24	.18	.21	.16
	(.39)	(.43)	(.38)	(.41)	(.37)
North Central	.26	.22	.27	.27	.26
	(.44)	(.41)	(.45)	(.44)	(.44)
South	.35	.33	.34	.34	.37
	(.48)	(.47)	(.47)	(.47)	(.48)
West	.18	.20	.18	.17	.19
	(.38)	(.40)	(.39)	(.37)	(.39)
Observations	3,664	355	1,019	1,360	930

Note: Standard deviations in parentheses.

Table 2: Summary statistics: time variant variables

Year	1990	1995	2000	2005	2008
Log hourly wage	4.81	5.19	4.91	4.80	5.10
	(2.72)	(2.95)	(3.45)	(3.59)	(3.64)
Work experience	8.64	10.97	15.47	18.24	19.71
	(3.01)	(3.56)	(5.05)	(6.22)	(6.85)

Note: Standard deviations in parentheses.

Table 3: First stage estimates

	Full sample	Science	Social science	Health and education
High school curriculum	.090***			
	(.018)			
Mother education		.011	-.341	-.045
		(.052)	(.320)	(.139)
Father education		-.080*	.155	-.149
		(.044)	(.222)	(.115)
Female		-1.089***	.711*	3.569***
		(.081)	(.383)	(.217)
Afro-American		.180*	1.212***	-.605**
		(.118)	(.503)	(.316)
Hispanic		-.097	-1.069	-.178
		(.122)	(.805)	(.301)
AFQT		.002	.015**	-.011***
		(.001)	(.008)	(.005)
Urban		.058	.454	-.089
		(.101)	(.488)	(.267)
Wald χ^2			425.86	

Note: */**/*** indicate confidence levels of 10/5/1 percent respectively. Reference category: humanities. White ethnic background is the reference category.

are presented in table 3. The omitted category is Humanities, therefore all coefficients should be interpreted in comparison to this category.

The covariate high school curriculum records the number of hours per week that each respondent dedicates to subjects belonging to one of the four educational supercategories in the last year of high school[16]. The coefficient shown is not group dependent (i.e.: it holds for each of the four categories). As expected, there is a positive relation between this variable and college major selection. The only other factors consistently affecting the choice for field of study are gender, being Afro-American and ability. Not surprisingly, girls are significantly more likely to select Humanities than Sciences, but even more likely to choose Health and Education. African-American college students are particularly attracted by Social Science subjects and little by Health and Education. No discernible pattern is evident for Hispanic students. Last, students with high AFQT score appear to select Social Science category.

Table 4 reports the estimated variance covariance matrix. The variance for Sciences is fixed at 2. The interpretation of the estimated covariance matrix coefficient is quite difficult and of limited practical interest since it only describes the differences in errors relative to alternative Humanities (i.e.: $(m_2 - m_1)$).

In table 5 the probabilities for the best choice by personal characteristics are reported. Looking at gender first, there is a clear dominance of Social Science and Health and Education choice for females, while males are, predictably, dominant in scientific majors. Humanities is clearly the least favored selection among all three ethnic groups, with Hispanic students showing the highest

[16] For example a students whose curricula include 4 hours of history and 4 hours of English literature per week in the last year of high school is recorded to have a 8 hour experience for humanities in high school.

Table 4: Estimated variance-covariance matrix

	Sciences	Social sciences	Health and educ.
Sciences	2		
Social sciences	-.265	84.112	
Health & educ.	-4.906	-8.840	14.907

Note: covariances are for alternatives differenced with Humanities. Variance for Hard sciences fixed at 2 for identification purposes.

Table 5: Mean best probability

	Humanities	Sciences	Social sciences	Health and education
White	.100 (.033)	.278 (.110)	.372 (.021)	.250 (.118)
Afro-American	.075 (.026)	.278 (.106)	.402 (.020)	.244 (.109)
Hispanic	.112 (.033)	.280 (.114)	.323 (.020)	.284 (.128)
Female	.085 (.030)	.181 (.020)	.375 (.033)	.357 (.037)
Male	.109 (.034)	.397 (.028)	.367 (.029)	.127 (.021)
Urban	.095 (.034)	.280 (.109)	.374 (.032)	.249 (.118)
N	355	1,019	1,360	930

Note: standard deviations in parentheses.

probability of selection for this category. Social science is the most likely choice for both African-American and Caucasian respondent, while Hispanic prefer Health and Education disciplines.

5.3 Step 2: Corrected estimates for the returns on education

The estimation of field of study choice probabilities illustrated in the previous section is propaedeutical to the identification of unbiased college major coefficients in the earning equation. In this section I report estimates of the earnings equation according to the implementation choices outlined in section 4. The dependent variable of the earnings equation is here the log of hourly wages. The independent variables are gender, work experience, ethnic origin (Caucasian is the omitted category), mother's and father's years of education, family of origin income, three dummies for field of study category (I exclude the dummy for Health and Education), four dummies for geographic area of origin (with people grown up in the North East the excluded category) and a control for personal ability measured by the AFQT (adjusted) score. As most of these variables, particularly the four major categories, are time invariant I use the between-individual model in equation (23) for identification.

In the specification of the basis function $b_m^q(.)$ the choice of the number of probability to be included is essential. The first natural choice is that of including only the best (revealed) choice. As mentioned in section 4.4 consistency of the methodology adopted in this paper requires the number of probabilities included in the basis function to be sufficiently large. Since no standard procedure exists for guiding the researcher in the correct choices of probabilities to include, in a second specification of the earning equation I augment the most parsimonious specification possible by the inclusion of the probabilities for all schooling types. A likelihood ratio test for the two models will provide me with some insight for the choice of the best specification. The test shows that the more extensive model outperforms the other, therefore I show, and base my estimates for wage variabilities on, only the favorite specification[17]. As for the choice of the order of polynomial expansions used for the creation of the correction functions, after the appropriate likelihood ratio test for model selection, I decide to use a third degree expansion[18].

Since I substitute estimates of the real schooling probabilities in the earning equation, in the second stage naive standard errors would probably be downward biased (Dahl, 2002; Cameron and Trivedi, 2005). I correct for the extra sample variability by bootstrapping [19].

Results of the wage equation estimation are presented in Table 6[20]. The two columns in Table 6 report estimations of the uncorrected model and the corrected model respectively.

When reading the results, particular attention should be paid to the major field coefficients. The difference between the uncorrected and the corrected college major coefficients is evident. Major coefficients increase by a factor of ten in all three cases and change sign.

A test for presence of self-selection is given by the Wald test statistic testing the significance of the correction term in my wage equation. The test statistic reported in Table 6 indicates that the correction function enters significantly at the one percent confidence level granting some confidence on the ability of my correction function to detect selection bias.

The other coefficients are very similar between the corrected and uncorrected specification: they all show the expected sign and, with the exception of parents schooling, are very precisely estimated. The only exceptions are the two coefficients for ethnic minorities that show a counterintuitive positive sign. This somehow surprisingly result was encountered also in previous estimates of wage equations by Cameron and Taber (2004) and Chen (2008) on the same sample of American young men.

5.4 Step 3: Point identification and bounds on wage variance parameters

In this section I provide estimates and bounds for the four crucial parameters for assessing the impact of risk on college major choice: the transitory component of wage variance(ψ^2_{mt}) and bounds for the permanent component (σ^2_{mi}), the risk parameter (τ^2_{mit}) and unobserved heterogeneity (ν_i).

As explained in section 4.3 point identification for the permanent component σ^2_{mi} is not possible given the flexible structure of the error term and of the correction functions adopted in this paper. Since the risk parameter τ^2_{mit} is a function of the permanent component also this parameter can only be bounded within a given interval.

The only parameter which can be point estimated, given the methodology adopted in this work, is (ψ_{mt}). Details for its derivation are given in section 4.2. Figure 1 plots the time series of estimated transitory component of wage variance by field of study (ψ^2_{mt}). At lest two important pieces of evidence can be extrapolated from this figure, the first regarding the coverage that different edu-

[17] Results of the test are available on request.

[18] For model choice I used a likelihood test ratio. The null hypothesis is strongly rejected at a 1% confidence level. Results of the test are available on request.

[19] Bootstrapping, with 400 repetitions, increases the standard errors by around 2% for most of the imputed regressors, but has no impact on standard errors for the other regressors.

[20] My estimation results are based on a more parsimonious specification for the wage equation assuming $\beta_m = \beta$ for all m in equation (2). The identification method and results, are not affected by allowing β to vary with major type. Results available on request.

Table 6: Estimated wage equations

	Uncorrected	Corrected
Humanities	-.130***(.035)	2.050***(.322)
Science	-.072***(.025)	1.725***(.286)
Social sciences	-.166***(.022)	1.479***(.282)
Work experience	.401***(.002)	.401***(.002)
Female	-.154***(.018)	-.108(.029)
Mother education	.001(.004)	.001(.004)
Father education	.001(.003)	-.002(.002)
Black	.936***(.024)	.931***(.025)
Hispanic	.557***(.027)	.555***(.026)
AFQT score	.002***(.000)	.002***(.000)
Geographic controls	yes	yes
Demographic controls	yes	yes
Wald test for λ		226.30 [.000]

Note: Bootstrapped standard errors based on 400 replications in parentheses. p-values in brackets. */**/*** indicate confidence levels of 10/5/1 percent respectively. Geographic controls include three dummies for residence at 14 (South is the excluded category). Demographic controls for year and quarter of birth.

cational paths offer to macroeconomic and institutional shocks, the second concerning the evolution of wage volatility for American college graduates throughout the past twenty years.

From this plot we can easily see how graduates of scientific disciplines, in particular, and Social Sciences are those better protected from macroeconomic and institutional shocks. At the opposite, Humanities graduates are those more exposed to macroeconomic fluctuations in almost every survey year staring from the early 1990s onwards. The last category of Health and Education behaves quite similarly to Humanities.

As for the time trend, the well known long-running rise of earning transitory volatility (Dynan et al., 2007) is confirmed here and it is irrespective of schooling level. In accordance with previous literature (Haider, 2001; Shin and Solon, 2011), I find a consistent increase in earning volatility starting with the last years of the past century and accentuating in the past decade.

It is worth noting that in my model on the job training is absent by construction. In fact, remember that the model envisages only two periods: the first when individuals invest in education and the second when individuals enter the labor market and collect their wages. It is evident that if on the job training investments are undertaken after completion of the selected course of study, these investments are overlooked in my estimation and their effects would be confounded with macroeconomic shocks in the transitory component.

Estimates for all the parameters of interest are concisely reported in table 7. Row A describes the mean over time and by schooling level for the transitory component of wage inequality visually described in figure 1. Clearly, wage uncertainty due to idiosyncratic shocks is minimal for the Science group and at its maximum for the Humanities graduates.

Row B shows lower and upper bounds for the permanent component corrected for selection and truncation as described in equation (25). Remember that the lower bound is set for $\rho_m^2 = 1$ while we have an upper bound when $\rho_m^2 = 0$. The situation changes for this parameter. In fact, whilst for the

Figure 1: Transitory component of wage variance by college major, 1990 to 2008

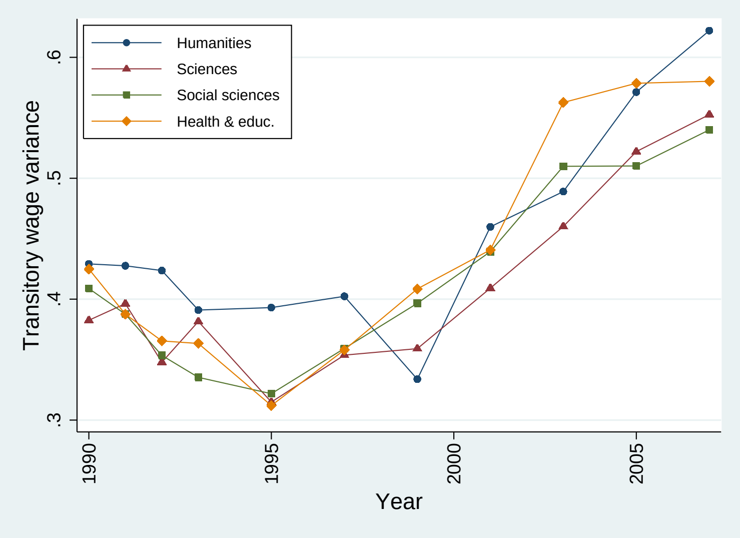


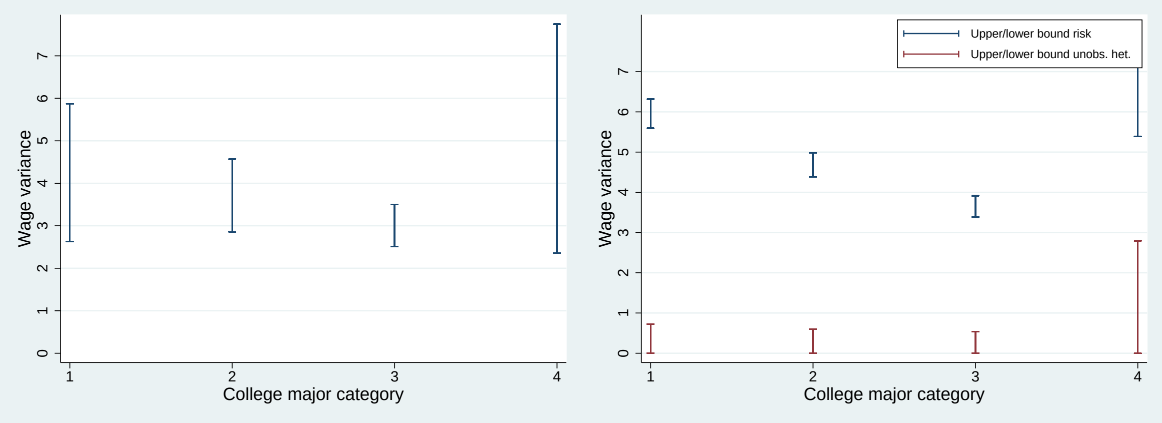
Table 7: Estimates of variance of potential wages

	Humanities		Sciences		Social sciences		Health and education	
A. Transitory component(ψ^2_{mt})	.449		.407		.415		.435	
	LB	UB	LB	UB	LB	UB	LB	UB
B. Permanent component(σ^2_{mi})	2.629	5.867	2.854	4.570	2.512	3.501	2.357	7.747
Potential inequality (A+B)	**3.078**	**6.316**	**2.946**	**4.977**	**2.927**	**3.916**	**2.792**	**8.182**
C. Degree of wage uncertainty(τ^2_{mi})	5.593	6.316	4.379	4.977	3.381	3.916	5.388	8.182
D. Unobserved heterogeneity(ν_i)	0	.723	0	.598	0	.535	0	2.794

transitory component we have that scientists experience the lowest variation, in terms of permanent component, the lowest level is encountered, here, for Social scientists. The width of the bounds is largest for Health and Education disciplines graduates and lowest for Social scientists. Compared to the least varying category of Social scientists, people belonging to the Health and Education category experience a 100% permanent component induced higher wage variance if we consider the high end of the admissible values. Interestingly enough, the lower bound for the latter category is slightly smaller than the former. Undeniably, the permanent component is the biggest contributor to total wage variance. The ratio of the permanent to transitory component (considering the upper bound) varies between 10 to 1, for the Science and Social Science group, and 20 to 1, in the case of the Health and Education group.

The key parameter in this study is pure risk and its effect on educational choices. The estimated intervals of admissible values for wage uncertainty are reported in Row C. Reassuringly for my methodology, the width of bounds is greatly diminished compared to the estimated bounds on the permanent component. The widest bound is on risk for Health and Education while bounds on values for Social Science almost collapse to point identification. Estimated risk reflects the same pattern of the permanent component. The highest risk, both in terms of upper and lower bounds, is associated with Humanities and Health and Education, while risk is minimized for Social Scientists.

Figure 2: Potential wage variance, risk and unobserved heterogeneity by schooling level: bounds



Note: College major category classification: 1 = Humanities; 2 = Sciences; 3 = Social Sciences; 4 = Health and Education

The last parameter of interest is unobserved heterogeneity. Compared to all other estimated parameters, its magnitude is minimal. Also in this case, with the exception of Health and Education, the bounds are comfortably narrow and the most precisely estimated bounds are those for the Social Science category. The highest unobserved heterogeneity is found in the Health and Education category, probably reflecting the heterogeneity of this specific category classification, but even in this case and considering the largest possible impact, its contribution to total wage variance is a mere 34%. The much higher contribution of risk compared to unobserved heterogeneity to wage variance was also encountered by other estimations concerned with level of education (Chen, 2008).

Figure 2 graphically and concisely displays the estimated intervals for the permanent component on the left panel and for risk and unobserved heterogeneity on the right panel. The results discussed above are effectively summarized in this graphical representation. The difference in level between risk and unobserved heterogeneity is evident as is the narrowness of the estimated bounds for these two key parameters.

The key empirical results reported in this section are four. First, some types of education, namely Scientific and Social Sciences disciplines, offer better immunization than others to idiosyncratic shocks. Second, the permanent component of wage variance, as well as the transitory component, is highest for graduates of Humanities and Health and Education in particular and lowest for the other two groups. Third, risk is highest for students of Health and Education disciplines. Fourth, pure risk accounts for the vast majority of potential wage variance.

6 The effect of risk on educational choices

In this section I estimate the responsiveness of educational choices to differences in individual specific risk level across the four college major categories. If personal risk differs across the four categories and if individuals are informed and act on this information, behaving according to what the theory of comparative advantage suggests, the probability of selecting one of the four possible choices should respond to these differences.

Equation (27) describes a multinomial probit model for the selection of major m instead of type r in terms of earnings, risk and individual specific taste for education:

$$V_{im} = \vartheta_1 + \vartheta_2 \hat{y}_{im} + \vartheta_3 \hat{\tau}_{imt} + \vartheta_4 \hat{\nu}_i + \varrho_{mi}. \tag{27}$$

The two main explanatory variable for the probability of selecting major type m are the wage and risk associated with that specific category estimated in the previous step. $\hat{y}_{im}$ is the estimated log individual earnings for major type m, $\hat{\tau}_{im}$ the log of the risk component, ν_i the log of the estimated taste for schooling parameter and ϱ an error term. The subscripts m indicate the different college majors. I only observe earnings and associated risk in the case that $m_i = m$, while earnings and risk for the counterfactual are not observed. What I can observe in the data is the outcome for individuals for whom observable characteristics x_{it} closely match those of the individual of interest. Matching the two type of individuals and imputing the revealed outcome for the "treated" as counterfactual for the "untreated" individual is a viable methodology given that I can control for a rich set of variables and given that selection is driven only by observables (Cameron and Trivedi, 2005). The assumption is strong and most likely not respected in my framework. Equation (1), in fact, describes the mechanism governing schooling selection and makes evident that individual select into education according to two criteria: expected income and the unobserved schooling parameter ν_i. Nevertheless, in section 5.4 I provide estimates for the admissible range of values of the unobserved heterogeneity parameter. I can then include this parameter in the matching algorithm and match on both observable characteristics and unobservable schooling factor rendering the selection mechanism only dependent on observable characteristics. As for the implementation of the matching procedure, I apply the propensity score matching method originally proposed in Rosenbaum and Rubin (1983) with "caliper matching"[21].

Remember that I do not posses point estimation neither for the risk parameter nor for the unobserved schooling factor. I decide to estimate the effect of risk on educational choice when both unobserved heterogeneity and risk are at their maximum possible values. Therefore, estimates shown in table 8 should be interpreted as the upper bound of the effect of risk on education decisions.

Estimation of equation (27) via multinomial probit would produce consistent estimates, but since I substitute estimates for the schooling coefficient ν_i and for wage and risk, the extra sampling variability needs to be accounted for. Therefore, standard errors shown in table 8 are obtained through bootstrapping.

Table 8 lists the estimation for the coefficients and marginal effects at mean for equation (27) by college major category. In order to make the coefficient more readily interpretable, I take logarithms for all covariates. Therefore, the β's reported should be interpreted as the percentage change in the probability of selecting that particular group of majors, that a 1% change in the covariate causes.

The key parameters are the effect that differences in personal risk have on educational choices. From the estimated coefficient it is immediately evident how educational decisions are significantly and negatively influenced by comparative differences in risk levels. As the theory would suggest,

[21] Caliper matching matches individuals within a predefined radius around the estimated propensity score to the untreated observation. For a textbook discussion of matching procedures see Cameron and Trivedi (2005). Matching procedure and results available on request.

Table 8: Responsiveness of education selection to risk and returns to education

Dependent variable: major choice	Coefficients	Marginal effect at mean			
		Humanities	Sciences	Social sciences	Health and education
Log Return to education	.797***	.009*	.064***	.183***	.222***
	(.095)	(.005)	(.010)	(.010)	(.012)
Log Risk	-.851***	-.010	-.069***	-.196***	-.237***
	(.105)	(.013)	(.010)	(.017)	(.020)
Log unobserved heterogeneity	.001***	-.000	-.000	-.003***	.003***
	(.000)	(.000)	(.000)	(.001)	(.001)
Wald χ^2	79.73				
N	3,664	355	1,019	1,360	930

Note: */**/*** indicate confidence levels of 10/5/1 percent respectively. Bootstrapped standard errors based on 400 replications in parentheses.

risk discourages selection of the specific supergroup. This holds for three out of four categories, Humanities being the only exception. The distaste for risk is particularly strong for the Health and Education group, in fact, doubling the risk associated with this category would cause a 23.7% decrease in the likelihood of selection for this particular category. The absence of any effect for the Humanities group might signal some different inclination towards risk for individuals of this category as consequence of their particular socio-economic status. In fact, if we look back at Table 1 we see how this group is formed by individuals with the best possible family background in terms of parents education and family income. This piece of evidence might support the intuition that good family background, serving as a buffer in case of failure, encourages people to select riskier educations.

The effect of returns to education is also of the expected sign and also particularly strong for Health and Education group. As for the risk coefficient, the faintest effect of wages on educational selection is detected for Humanities graduates.

Table 9 provides a further scrutiny of the correlation patterns and thus individuals' personal advantages in terms of both expected wages and wages uncertainty in selecting one of the four major categories. The terms presented here refer to the correlation across the four possible major categories of expected wages and risk used in equation (27) for the estimation of the effect of returns and risk on choices. The three counterfactuals are imputed by matching and the risk coefficients are correlated under the hypothesis of $\rho = 1$. From the Table we can detect a weak negative correlation for wages and a weak positive correlation for uncertainty for each possible pair. The highest correlation exists between Social Sciences and Sciences for wages and Social Sciences and Health and Education for wage uncertainty. Both the direction and the size of correlation signal the existence of comparative advantages. In fact, the direction suggests that individual do select their most advantageous options as a high income in the selected category is always associated with lower income in the alternative category and the opposite happens in the case of wage uncertainty. On the other hand, the low correlation detected points towards a the existence of real outcome differences as a consequence of major type selection.

In conclusion, the results presented here support a Roy model for selection of education driven

Table 9: Correlation matrices for wages and wage uncertainty

	Expected Wages			
	Humanities	Sciences	Social Sciences	Health&Education
Humanities	1			
Sciences	-.087	1		
Social Sciences	-.131	-.260	1	
Health&Education	-.137	-.197	-.244	1
	Wage uncertainty			
Humanities	1			
Sciences	.031	1		
Social Sciences	.078	.126	1	
Health&Education	.185	.102	.259	1

by comparative advantages. As expected, higher risk discourages selection of a particular type of education, while higher returns have the opposite effect. The parameter estimates for the Humanities group, in light of the particularly favorable environment that these people enjoyed during their upbringing, offers some ground for theories suggesting that socio-economic status bears consequences on risk taking behaviors of individuals.

7 Conclusions

Exploiting recent advancements in the literature for semiparametric estimators in the case of polychotomous choice models, this paper tests the often made assumption of endogenous schooling choices when future outcomes are uncertain and estimates the effect that differences in personal risk level have for those choices.

My main finding is that concerns about risk significantly bias observed wages and observed wage variances for every College major category. The test of the Roy model for educational choices supports the role of comparative advantages in schooling decisions for three out of four categories. I advance the hypothesis that the almost complete absence of risk aversion for the Humanities graduates group could be a resultant of the particularly favorable family background that this group enjoys. I also find that OLS estimation severely *underestimates* returns to education up to more than 100%. Additional results, contributing to the growing literature of causal effects of schooling on risk, show how some types of education protect against macroeconomics shocks better than others, how Scientific and Social Sciences type of degrees entail less risk compared to Humanities and Health and Education type of educations and confirms the long-running growth of transitory volatility for college graduates earnings in the US for the past twenty years.

The semiparametric approach employed in the present work can be easily extended to other unordered or ordered choice settings with just few modifications. In the context of schooling choices the most relevant case would probably regard choices between vocational or academic educations at high school level. Future research could tackle this aspect and could also try to explain how and through which channels, socio-economic status influences educational investments under uncertainty as the results in this paper appears to suggest.

References

Ahn, H. and J. Powell: 1993, 'Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism'. *Journal of Econometrics* **58**(1/2), 3–29.

Altonji, J.: 1993, 'The Demand for and Return to Education When Education Outcomes are Uncertain'. *Journal of Labor Economics* **11**(1), 48–83.

Bayer, P., S. Kahn, and C. Timmins: 2011, 'Nonparametric Identification and Estimation in a Roy Model With Common Nonpecuniary Returns'. *Journal of Business and Economics Statistics* **29**(2), 201–215.

Becker, G. S.: 1975, *Human Capital: A Theoretical and Empirical Analysis.* New York: National Bureau of Economic Research.

Bound, J., D. A. Jaeger, and R. M. Baker: 1995, 'Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak'. *Journal of the American Statistical Association* **90**(430), 443–450.

Cameron, C. and P. Trivedi: 2005, *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Cameron, S. V. and C. Taber: 2004, 'Estimation of Educational Borrowing Constraints Using Returns to Schooling'. *Journal of Political Economy* **112**(1), 132–182.

Chen, S.: 2008, 'Estimating the Variance of Wages in the Presence of Selection and Unobserved Heterogeneity'. *The Review of Economics and Statistics* **90**(2), 275–289.

Chen, S. and S. Khan: 2007, 'Estimating the Casual Effect of Education on Wage Inequality Using IV Methods and Sample Selection Models'. Working paper, University at Albany-SUNY.

Cosslett, S.: 1983, 'Distribution Free Maximum Likelihood Estimator of the Binary Choice Model'. *Econometrica* **51**(3), 765–782.

Cosslett, S.: 1991, 'Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics'. In: W. Barnett, J. Powell, and G. Tauchen (eds.): *Semiparametric Estimation of a Regression Model with Sample Selectivity.* Cambridge, UK: Cambridge University Press, pp. 175–197.

Cunha, F., J. Heckman, and S. Navarro: 2005, 'Separating Uncertainty from Heterogeneity in Life Cycle Earnings'. *Oxford Economic Papers* **57**(2), 191–261.

Dahl, G. B.: 2002, 'Mobility and the Return to Education: Testing a Roy Model with Multiple Markets'. *Econometrica* **70**(6), 2367–2420.

Dynan, K., D. W. Elmendorf, and D. E. Sichel: 2007, 'The Evolution of Household Income Volatility'. Finance and economics discussion series: 2007/61, Board of Governors of the Federal Reserve System.

Falaris, E. M.: 1987, 'A Nested Logit Migration Model with Selectivity'. *International Economic Review* **28**(2), 429–443.

Gallant, R. A. and D. W. Nychka: 1987, 'Semi-Nonparametric Maximum Likelihood Estimation'. *Econometrica* **55**(2), 363–390.

Goldberger, A.: 1983, 'Abnormal Selection Bias'. In: S. Karlin, T. Amemiya, and L. Goodman (eds.): *Studies in Econometrics, Time Series, and Multivariate Statistics*. New York: Academic Press.

Haider, S. J.: 2001, 'Earnings Instability and Earnings Inequality of Males in the United States: 1967-1991'. *Journal of Labor Economics* **19**(4), 799–836.

Heckman, J.: 1974, 'Shadow Prices, Market Wages and Labour Supply'. *Econometrica* **42**(4), 679–694.

Heckman, J.: 1976, 'The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models'. *Annals of Economics and Social Measurement* **5**(4), 475–492.

Heckman, J.: 1979, 'Sample Selection Bias as a Specification Error'. *Econometrica* **1**(74), 153–162.

Heckman, J. J. and R. Robb: 1985, 'Alternative Methods for Evaluating the Impact of Interventions: An Overview'. *Journal of Econometrics* **30**(1-2), 239 – 267.

Kane, T. J. and C. E. Rouse: 1995, 'Labor-Market Returns to Two and Four Years College'. *American Economic Review* **85**(3), 600–614.

Lee, L.-F.: 1982, 'Some Approaches to the Correction of Selectivity Bias'. *The Review of Economic Studies* **49**(3), 355–372.

Lee, L.-F.: 1983, 'Generalized Econometric Models with Selectivity'. *Econometrica* **51**(2), 507–512.

Lemieux, T.: 2006, 'Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill'. *American Economic Review* **96**(3), 1–64.

Maddala, G. S.: 1983, *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

Matzkin, R. L.: 1993, 'Nonparametric Identification and Estimation of Polychotomous Choice Models'. *Journal of Econometrics* **58**(1-2).

Mazza, J. and H. van Ophem: 2010, 'Separating Risk in Education from Heterogeneity: a Semiparametric Approach'. Uva econometrics discussion paper: 2010/07, Amsterdam School of Economics.

McFadden, D. L.: 1984, 'Econometric Ananlysis of Qualitative Response Models'.

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**(5), 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. K.: 1997, 'Convergence rates and asymptotic normality for series estimators'. *Journal of Econometrics* **79**(1), 147 – 168.

Newey, W. K.: 2009, 'Two-Step Series Estimation of Sample Selection Models'. *Econometrics Journal* **12**(1).

Powell, J.: 1994, 'Estimation of Semiparametric Models'. In: R. F. Engle and D. L. McFadden (eds.): *Handbook of Econometrics*, Vol. 4. Amsterdam: North Holland, pp. 2444–2523.

Robinson, P. M.: 1988, 'Root-N-Consistent Semiparametric Regression'. *Econometrica* **56**(4), 931–954.

Rosenbaum, P. and D. B. Rubin: 1983, 'The Central Role of Propensity Score in Observational Studies for Causal Effect'. *Biometrika* (70), 41–55.

Roy, A. D.: 1951, 'Some Thoughts on the Distribution of Earnings'. *Oxford Economic Papers* **3**(2), 135–146.

Shin, D. and G. Solon: 2011, 'Trends in men's earnings volatility: What does the Panel Study of Income Dynamics show?'. *Journal of Public Economics* **95**(7-8), 973 – 982.

Trost, R. P. and L.-F. Lee: 1984, 'Technical Training and Earnings: A Polychotomous Choice Model With Selectivity'. *The Review of Economics and Statistics* **66**(1), 151–156.

Vella, F.: 1998, 'Estimating Models with Sample Selection Bias: A Survey'. *The Journal of Human Resources* **33**(1), 127–169.

Willis, R. J. and S. Rosen: 1979, 'Education and Self-Selection'. *Journal of Political Economy* **87**(5), S7–36.

Zafar, B.: 2011, 'How Do College Students Form Expectations?'. *Journal of Labor Economics* **29**(2), 301–348.

A College majors classification

In the following scheme I report the different college majors as coded in the NLSY and the educational group that I classify them into.

Humanities	**Sciences**	**Social Sciences**	**Health and Education**

Fine and Applied Arts

-Fine Arts, General

-Art

-Art History and Appreciation

-Music (Performing, Composition, Theory)

-Music (Liberal Arts Program)

-Music History and Appreciation

-Dramatic Arts

-Dance

-Applied Des. & Graphic Des. & Fashion Des.

-Cinematography

-Photography

-Applied Music

-Studio Arts

-Commercial Art

-History of Architecture

-Other

Foreign Languages

-Foreign Languages, General

-French

-German

-Italian

Agriculture and Natural Resources

-Agriculture, General

-Agronomy

-Soils Science

-Animal Science

-Dairy Science

-Poultry Science

-Fish, Game, and Wildlife Management

-Horticulture

-Ornamental Horticulture

-Agricultural and Farm Management

-Agricultural Economics

-Agricultural Business

-Food Science and Technology

-Forestry

-Natural Resources Management

-Agriculture and Forestry Technologies

-Range Management

-Pest Control and Crop Protection

-Other

Architecture and Environmental Design

-Environmental Design, General

Area Studies

-Asian Studies, General

-East Asian Studies

-South Asian (India, etc.) Studies

-Southeast Asian Studies

-African Studies

-Islamic Studies

-Russian and Slavic Studies

-Latin American Studies

-Middle Eastern Studies

-European Studies, General

-Eastern European Studies

-West European Studies

-American Studies

-Pacific Area Studies

-French Studies

-Other

Business and Management

-Business and Commerce, General

-Accounting

-Business Statistics

-Banking and Finance

Education

-Education, General

-Elementary Education, General

-Secondary Education, General

-Junior High School Education

-Higher Education, General

-Junior and Community College Education

-Adult and Continuing Education

-Special Education, General

-Administration of Special Education

-Education of the Mentally Retarded

-Education of the Gifted

-Education of the Deaf

-Education of the Culturally Disadvantaged

-Education of the Visually Handicapped

-Speech Correction and Communicative Disord.

-Education of the Emotionally Disturbed

-Remedial Education

-Special Learning Disabilities

-Education of the Physically Handicapped

-Education of the Multiple Handicapped

-Social Foundations

-Educational Psychology

-Spanish

-Russian

-Chinese

-Japanese

-Latin

-Greek, Classical

-Hebrew

-Arabic

-Indian (Asiatic)

-Scandinavian Languages

-Slavic Languages (Other than Russian)

-African Languages (Non-Semitic)

-Portuguese

-Other

Law

-Law, General

-Pre-law

-Other

Letters

-English, General

-Literature, English

-Comparative Literature

-Classics

-Linguistics

-Architecture

-Interior Design

-Landscape Architecture

-Urban Architecture

-City, Community, and Regional Planning

-Other

Biological Sciences

-Biology, General

-Botany, General

-Bacteriology

-Plant Pathology

-Plant Pharmacology

-Plant Physiology

-Zoology, General

-Pathology, Human and Animal

-Pharmacology, Human and Animal

-Physiology, Human and Animal

-Microbiology

-Anatomy

-Histology

-Biochemistry

-Biophysics

-Molecular Biology

-Cell Biology

-Marine Biology

-Investments and Securities

-Business Management and Administration

-Operations Research

-Hotel and Restaurant Management

-Marketing and Purchasing

-Transportation and Public Utilities

-Real Estate

-Insurance

-International Business

-Secretarial Studies

-Personnel Management

-Labor and Industrial Relations

-Business Economics

-Organizational Behavior

-Other

Communications

-Communications, General

-Journalism

-Radio - Television

-Advertising

-Communication Media

-Mass Communications

-Public Relations

-Group Communications

-Other

-Pre-Elementary Education

-Educational Statistics and Research

-Educational Testing, Evaluation and Measur.

-Student Personnel

-Educational Administration

-Educational Supervision

-Curriculum & Instruct. & Educational Media Nurs. Educ.

-Reading Education

-Art Education

-Music Education

-Mathematics Education

-Science Education

-Physical Education

-Driver and Safety Education

-Health Education

-Business, Commerce, and Distributive Educ.

-Industrial Arts, Vocational & Technical Educ.

-Guidance and Counseling

-English Education

-Foreign Languages Education

-Social Studies Education

-School Management

-Speech and Drama Education

-School Librarianship

-Urban Education

-Bilingual Education

-Speech, Debate, and Forensic Science

-Creative Writing

-Teaching of English as a Foreign Language

-Philosophy

-Religious Studies

-Literature, General (except English)

-Other

Library Science

-Library Science, General

-Other

History

-Archaeology

-History

Theology

-Theological Professions, General

-Religious Music

-Biblical Languages

-Religious Education

-Other

-Biometrics and Biostatistics

-Ecology

-Entomology

-Genetics

-Radiobiology

-Nutrition, Scientific

-Neurosciences

-Toxicology

-Embryology

-Pre-med

-Pre-vet

-Pre-dentistry

-Immunology

-Other

Computer and Information Sciences

-Computer and Information Sciences, General

-Information Sciences and Systems

-Data Processing

-Computer Programming

-Systems Analysis

-Other

Engineering

-Engineering, General

Home Economics

-Institutional Management and

-Cafeteria Management

Psychology

-Psychology, General

-Experimental Psychology

-Clinical Psychology

-Psychology for Counseling

-Social Psychology

-Psychometrics

-Statistics in Psychology

-Industrial Psychology

-Developmental Psychology

-Physiological Psychology

-Behavioral Science

-Comparative Psychology

-Rehabilitation Counseling

-Animal Behavior

-Other

Public Affairs and Services

-Community Services, General

-Public Administration

-Parks and Recreation Management

-Multicultural Education

-Community Education

-Agricultural Education

-Education of Exceptional Children,
Not Classified Above

-Home Economics Education

-Other

Health Professions

-Health Professions, General

-Hospital and Health Care Administration

-Nursing

-Dental Specialties

-Medical Specialties

-Occupational Therapy

-Optometry

-Pharmacy

-Physical Therapy

-Dental Hygiene

-Public Health

-Medical Record Librarianship

-Podiatry or Podiatric Medicine

-Biomedical Communication

-Veterinary Medicine Specialties

-Speech Pathology and Audiology

-Chiropractic

-Aerospace, Aeronautical, Astronautical Eng.

-Agricultural Engineering

-Architectural Engineering

-Bioengineering and Biomedical Engineering

-Chemical Engineering

-Petroleum Engineering

-Civil, Construction & Transportation Eng.

-Electrical, Electronics, Communications Eng.

-Mechanical Engineering

-Geological Engineering

-Geophysical Engineering

-Industrial and Management Engineering

-Metallurgical Engineering

-Materials Engineering

-Ceramic Engineering

-Textile Engineering

-Mining and Mineral Engineering

-Engineering Physics

-Nuclear Engineering

-Engineering Mechanics

-Environmental and Sanitary Engineering

-Naval Architecture and Marine Engineering

-Ocean Engineering

-Engineering Technologies

-Other

Mathematics

-Mathematics, General

-Statistics, Mathematical and Theoretical

-Applied Mathematics

-Social Work and Helping Services

-Law Enf. & Correct. & Criminol. & Crim. Just.

-International Public Service

-Administration of Justice

-Other

Social Sciences

-Social Sciences, General

-Anthropology

-Economics

-Geography

-Political Science and Government

-Sociology

-Criminology

-International Relations

-Afro-American (Black Culture) Studies

-American Indian Cultural Studies

-Mexican-American Cultural Studies

-Urban Studies

-Demography

-Group Studies

-Other

-Clinical Social Work

-Medical Laboratory Technologies

-Dental Technologies

-Radiologic Technologies

-Rehabilitation

-Expressive Therapy(ies)

-Allied Health

-Other

	-Other **Physical Sciences** -Physical Sciences, General -Physics, General -Molecular Physics -Nuclear Physics -Chemistry, General -Inorganic Chemistry -Organic Chemistry -Physical Chemistry -Analytical Chemistry -Pharmaceutical Chemistry -Astronomy -Astrophysics -Atmospheric Sciences and Meteorology -Geology -Geochemistry -Geophysics and Seismology -Earth Sciences, General -Paleontology -Oceanography -Metallurgy -Industrial Chemistry -Other Earth Sciences -Other Physical Sciences **Interdisciplinary Studies** -Biological and Physical Sciences -Engineering and Other Disciplines		