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

Jacopo Mazza*

University of Manchester

August 3, 2015

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 confirm that individuals self-select into education in order to minimize uncertainty and maximize returns irrespective of the degree chosen.

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

*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. This paper has also benefited from discussions with seminar and conference participant at Tinbergern Institute Amsterdam, LMU University Munich, CERGE-EI Prague, University of Manchester, IFS London, Royal Economic Society Congress, SOLE/EALE World Congress.

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 in education are taken under a considerable amount of uncertainty regarding performance in school, future labor market and macroeconomic conditions among many others. 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 perfectly anticipated by the individual, but for which some probabilistic assessment can be calculated. As it is plausible to assume that individuals possess some private information inaccessible to the researcher this discrepancy in the information set should be accounted for. In fact, 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, defined as future wage variance, 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 be-

tween 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 risk in 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 more standard parametric techniques have come under closer scrutiny and received growing criticism (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). Nevertheless, 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.
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, I classify similar individuals into cells. In this way, for each cell, I can obtain a distribution-free estimator for selection probabilities avoiding the undesirable independence of irrelevant alternative property intrinsic in other polychotomous choice models. In the second step these probabilities serve as basis for the construction of the correction function needed to consistently estimate the wage equation and retrieve risk estimates corrected for self-selection. Third, the various elements of wage variance are either point estimated or bounded within an admissible range of values. The results confirm the well known increase in transitory earnings volatility for the US in the past twenty years and show how graduates in Natural Science disciplines are better immunized against macroeconomic shocks compared to graduates in other subjects. The same type of educations also protect against total risk defined as the sum of transitory 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, non trivial, impact of risk on educational selection is confirmed for all educational groups.

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, uncer-

tainty 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] - Natural Science, Business and Economics, Humanities and Social Sciences and Health and Education, these probabilities serve as basis for constructing the selection adjustments terms that in the second stage are included in a wage equation. In this way, the zero mean condition on the error term is assured even in the presence of of self-selection allowing estimation by ordinary least squares of the parameters of interest. 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 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: Natural Sciences ($m_i = 1$); Business and Economics ($m_i = 2$); Humanities and 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 return of that specific investment. 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, 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 risk 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

[1] The choice of these four college major categories is fairly standard in the literature. Altonji (1993) and Arcidiacono (2004) use a very similar classification. Aggregation is necessary for statistical significance since many college major groups coded in the NLSY count little to no observations.

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 risk 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 $m = 1, 2, 3, 4$:

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

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

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 risk [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 5. 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.

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

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:

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

where I_{mi} is an indicator function assuming value 1 if that specific major is selected and 0 otherwise and $E[V_{mi}] = E[V_{mit_0}]$ since expectations are assumed to be age independent and time subscript t_0 is omitted in the remainder of the paper for ease of notation.
The system of equations in (1) and (2) cannot 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 only in 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.
In particular, 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_m, 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 therefore 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:

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

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). In case of multiple options, as for the model presented here, the approach depends on the structure of the outcomes that can either be ordered according to some natural or 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. 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 techniques 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. In a second stage these estimates form the basis for the construction of a 'single-index' correction function $g(.)$ which is then included in the outcome equation allowing consistent estimates of the parameters of interest. To be sure, research on semiparametric estimation methods for binary response models has received some attention in recent literature, however, 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 concerned with the estimation of migration decisions.
I exploit Dahl's work and adapt it to the different needs that my research question poses. The main difference between mine and Dahl's framework resides in the structure of the error term in the choice equation. In fact, to be able to separate risk from private information, the error term in the first stage is assumed to be normally distributed. In section 4.2 I show how this condition is necessary for deriving the so called 'truncation adjustment'.

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

Nonetheless, neither the derivation of consistent estimates for the probabilities of major choice in the first step, nor the unbiasedness of my second stage estimates hinge on normality as it is made clear in the follow up of the paper.
An additional advantage of the approach adopted in this paper consists in the faculty of leaving the exact subutility function unspecified which is otherwise one of the main challenges for the estimation of Roy models based on utility maximization. In my framework, a plethora of variables are potential candidates for inclusion in the correct utility function 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 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. Ahn and Powell exploit this property 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:

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

In this formulation $\varsigma_m(.) = E[u_{mit}|V_m - V_r, \ldots, V_M - V_r], \eta_{mit}$ 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}.
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 group 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).

[4]To see how the equality is derived remember the selection criteria expressed by equation ([eq:selection]). That relation states that selectivity bias in y_{mit} 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.

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

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:

$$g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi}|V_m - V_r, \ldots, V_M - V_r) = \\ g_m(u_{mit},\ max_r(V_r - V_m + s_{ri} - s_{mi})|p_{im}, \ldots\ p_{iM}) \quad (16)$$

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

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

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:

$$g_m(u_{mit},\ max_r(V_r - V_m + w_{rit} - w_{mit}|V_m - V_r, \ldots, V_M - V_r) = \\ g_m(u_{mit},\ max_r(V_r - V_m + w_{rit} - w_{mit})|p_{im},\ \vec{q}) \quad (18)$$

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. Unfortunately there are no theoretical guidelines describing the choice of the probabilities to be included. It would be natural to include the first Nth best choices as additional terms. In my application to major choice the number of probabilities, other than the revealed one, candidate for inclusion is limited by construction. I decide to include the revealed choice probability plus the 'retention probability' which in this context is the probability attached to the decision of not pursuing a tertiary degree and obtain an high school degree only.
The choice of these probabilities implies the following distributional assumption:

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

and the following earning equation:

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

I refer to $\lambda_m(.)$ as the selection correction function which is an unknown function of two probabilities $p(m_i = I)$ and $p(m_i = 0)$ where with $P(m_i = 0)$ I indicate the probability of stopping at high school.

4 Empirical estimation

In the previous section I have outlined the general structure of a semi-parametric 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 anticipated 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 separate identification of this parameter. 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].
As explained in section 3.1 I will estimate the selection probabilities invoking the 'index sufficiency' assumption as proposed by Dahl and will use the observed probabilities of major selection augmented by the inclusion of the 'retention probability' as argument for the control function to be included in the second stage. As pointed out by Carneiro and Lee (2009) this procedure is in essence akin to estimating a classical two stage selection model in which the exclusion restriction is left implicit.
In this context, a selection probability is simply the fraction of individuals within a given cell who decide for one of the four possible major categories m. Individuals are assigned to different cells on the basis of a vector of individual characteristics discussed in section 5.2.
Two major advantages can be imputed to this methodology: the exact form for the subutility function does not need to be specified and distributional assumptions are not necessary for the consistent estimation of the selection probabilities.

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:

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

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

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

its analytical expression is given by:

$$\delta_{mi} = 1 - Var[\nu_i | p_{im}, \dots \ p_{iM}] = -\lambda(Z\varphi - \lambda) \tag{22}$$

Where $\lambda = E[\nu_i | p_{im}, \dots \ p_{iM}] = -\frac{\phi(z_i\varphi)}{\Phi(z_i\varphi)}$. The derivation[7] for the expression for the truncation adjustment in 22 only requires for the error term in the selection equation (ν_i in my case) to be distributed normally; no assumptions are made on the error term in the wage equation.
Remember that the probabilities for schooling selection are estimated via the within cell choice probabilities according to the method highlighted in section 4.2. 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:

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)\beta_m + (\kappa_{mit} - \bar{\kappa}_{mi}) \qquad \text{if } m_i = m, \tag{23}$$

$\overline{y}_i, \overline{x}_i$ 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 transitory component of wage variance ψ^2_{mt} is identified as the variance of the error term in equation 23.
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:

$$\overline{y}_i = \alpha_m + \overline{x}_i\beta_m + \sum [I_{im}\lambda_m(p_{im}, \dots \ p_{iM})] + \overline{\omega}_{mi} \tag{24}$$

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

[7] For a complete discussion see Maddala (1983).

expression for the permanent component corrected for truncation and self-selection:

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

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 24 $\overline{T} \equiv (\sum_i T_i^{-1}/N)^{-1}$ 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 prevents point identification for 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 thus 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 25 cannot be imputed completely to proper risk since 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 23. Identification of the permanent component σ^2_{mi} is more complicated. The complete specification of the permanent component given in equation 25 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 25 and rearrange it to obtain:

$$\sigma^2_{mi} = \frac{\widehat{Var}[\omega_{mi}|\overline{x}_i, m_i = m, z_i] - \sum_t \widehat{\psi}^2_{mt}/\overline{T}}{1 - \rho^2_m \hat{\delta}_{mi}} \tag{26}$$

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^2_m which can be bounded between 0 and 1. In case of no correlation between wages and the unobserved schooling factor (i.e.: $\rho_m = 0$) the permanent component is simply the variance of the error term in the between individual model of equation 25 minus the transitory shock; thus no private information is exploited for minimizing wage variance. The other extreme is given for perfect correlation (i.e.: $\rho_m = 1$). 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:

$$\lambda_m(p_{mi}, \dots\ p_{iM}) \simeq \sum_{q=1}^{Q} \kappa^q_m b^q_m(p_{mi}, \dots\ p_{iM}) \tag{27}$$

[8] See section 4.2.

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.

5 The causal impact of risk on education

My empirical estimation for the importance of concerns about 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 27 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. Participants were interviewed annually until 1994 and biennially thereafter.

Since the focus of my analysis is on college major choice, I restrict the sample analyzed to males and females who attended and graduated from college, this reduces my sample to 3,529 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 12 subsequent waves until the survey of 2010. 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 (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[9].

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[10]. The AFQT is a weighted sum of four different ASVAB components in mathematics, science, vocabulary and automotive knowledge. I present here only the score on the mathematical section of ASVAB since I use only this information for assigning individuals to mutually exclusive cells while the (corrected) AFQT score is used in the wage regressions. I explain this decision in detail in section 5.2 when I discuss the cell creation procedure. The test was administered in 1980 to all subjects regardless their age and schooling level. For this reason it can include age and schooling effects in the ability index that the test is meant to construct. To correct for these undesired effects, 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.

Given the the sample size and the very detailed classification of majors available in the NLSY79 some aggregation of major categories is needed for statistical power. For the grouping of the different major I follow Arcidiacono (2004) and define four categories: Natural Sciences (including math and engineering), Business and Economics, Humanities and Social Sciences and Health and Education.

In addition to these common variables, work experience is added as a time varying control in both the between individuals and fixed-effect estimation.

[9]Source: ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt (accessed 11/07/2014)

[10]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.

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 3,592 individuals observed in 11 waves generating 30,366 individual-year pair observations[11]. Summary statistics for the variables used in the analysis are presented in table 1. Panels A to D show summary statistics of static variables for a series of personal and family characteristics, geographical area of residence at age 17 and test score for the AFQT test and for the mathematical component of it.

Panel E summarizes the available information on high school curricula The NLSY79 provides information on the type of subjects that pupils attended during their last year of high school. I group the possible options into three mutually exclusive categories, as shown in the table, and this information enters the vector of individual characteristics used to create the cells in the first stage. The numbers in panel E refer to the percentage of individuals for whom the majority of subjects taken in high school fell into one of the three possible options. This information plays a crucial role for my identification strategy since it will be used for cell assignment, but excluded from the wage regressions. Evidently subject in humanities constituted the biggest share of high school curricula for the majority of students.
Finally, Panel F shows means and standard deviations for annually or biennially varying hourly wages at most recent job used for the second stage wage analysis. Potential work experience is calculated as age minus schooling minus six.

5.2 Step 1: Schooling choice first stage estimates

The first step to correct for self-selection is the estimation of the choice probabilities. Individual's probabilities serve as basis for the construction of the correction functions $b_m^q(.)$ in equation 27. Similar individuals are grouped into cells and this assignment is then used to estimate individual choice probability as the fraction of cell members selecting one of the four possible college majors or selecting to stop with an high school degree. In fact, as

[11] 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: Descriptive Statistics

Variable	Mean	Standard Deviation	Sample Size
	A. Personal characteristics		
Female	0.540	0.499	3,529
Black	0.257	0.437	3,592
Hispanic	0.158	0.365	3,592
	B. Family characteristics		
Family Income *(in 2008 $)*	47,151.8	33,316.3	3,592
Number of Siblings	3.226	2.298	3,592
Highest Grade Mother	11.03	4.080	3,592
Highest Grade Father	10.55	5.565	3,592
	C. Geographic Controls (*Measured at 17)*		
North-East	0.172	0.377	3,952
North-Central	0.270	0.444	3,952
South	0.367	0.482	3,952
West	0.185	0.388	3,952
Urban	0.794	0.395	3,952
	D. Test Scores		
ASVAB Math Score	14.64	6.113	3,952
AFQT (*Adjusted)*	70.425	22.847	3,952
	E. Favored High School Subject		
Natural Science	0.102	0.303	3,592
Humanities	0.755	0.430	3,592
Social Sciences	0.045	0.207	3,592
	F. Time-Varying Labor Market Variables		
Log Hourly Wage	2.704	.802	30,366
Potential Work Experience (years)	16.623	7.039	30,366

I explained in section 3.1 I include two different probabilities in the basis function: the selection probability and the 'retention' probability. The 'retention' probability is the within cell probability of opting out of education after high school. In section 5.1 I explained that the final sample of analysis is restricted to college graduates for whom a retention probability would be impossible to estimate. Therefore, the sample over which these probabilities are estimated differs from the final one as it includes individuals who had the option to proceed to tertiary education since they obtained an high school degree, but decided otherwise. The augmented sample includes the 3,592 individuals of the main sample plus additional 4,428 high school graduates for a total of 8,020 subjects.
The covariates determining cell assignment are: gender; ethnic background (e.g.: white, African-American or Hispanic); quartile of ASVAB mathematical score and whether the majority of classes taken in the last high school year was in subjects pertaining to humanities, social sciences or natural sciences (including mathematics). This assignment rule produces 96 mutually exclusive cells.

Table 2: Major Choice by Personal Characteristics *(%)*

	Natural Sciences	Business & Economics	Social Sciences & Humanities	Education & Health
Gender				
Males	34.5	23.6	31.9	10.1
Females	14.3	30.3	27.0	28.4
Ethnic Background				
African-American	21.8	30.9	26.6	20.7
Hispanic	21.4	27.5	30.6	20.5
White	23.3	27.3	29.2	20.2
Quartile of Mathematic Test Score				
1^{st}	19.2	26.7	30.5	23.6
2^{nd}	17.2	30.4	30.8	21.7
3^{rd}	19.1	27.5	31.4	22.0
4^{th}	32.1	26.1	25.7	16.2
High School Curricula Pred. Choice				
Sciences	37.1	25.6	17.9	19.5
Humanities	21.6	27.0	31.5	20.0
Social Sciences	15.2	35.0	26.9	22.8

In table 2 the probabilities for the best choice by personal characteristics are reported for the main sample only. Looking at choices by gender first, there is a clear difference in Natural Science and Health and Education choice. The former is male dominated while the opposite occurs for the

latter.
On the other hand, few differences exists between the three ethnic groups. As expected, the most mathematically gifted students tend to select a degree in the Natural Sciences, whilst students within the first three quartiles of the mathematical ability distribution shy away from them and more or less evenly distribute between the remaining three categories.
Lastly, it should be noted that the predominant choice in the last year of high school curricula is a good predictor of student major choice in college.

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, potential work experience, ethnic origin (white being the omitted category), mother's and father's years of education, family of origin income, four dummies for geographic area of origin (with people grown up in the North East as 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 24 for identification and run separate regression for each of the four college categories. It should also be noted that the regression constant includes both any major specific constant and the intercept for the correction functions. Therefore, these two elements cannot be separately identified.
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. Other than the revealed choice probability I include the 'terention probability as argument for the basis function and the second order expansion for the both.
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[12].

[12]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.

Results of the wage equation estimation are presented in Table 3. The four columns in Table 3 report estimations of the wage equations for the four major categories.
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 3 indicates that the correction function enters significantly at the one percent confidence level for the Natural Science regression and at the 5As for the other coefficients the only covariates that systematically affect earnings for each of the four majors are the ability index as measured by the AFQT score, gender and, for three out of four groups, family income. As expected, ability positively influences earnings while being a female has the opposite effects. The gender disparity is particularly accentuated for Natural Sciences and Business graduates. Unsurprisingly students coming from rich families enjoy higher payoffs for their degree with the only exception of those studying Business.
The effect of all other covariates on earnings are statistically indistinguishable from zero.

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 educational 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 are those better protected

Table 3: Estimated Wage Regressions

	Natural Sciences	Business & Economics	Social Sciences & Humanities	Education & Health
Female	-0.305***	-0.314***	-0.241***	-0.164**
	(0.06)	(0.04)	(0.03)	(0.07)
Potential Experience	-0.037	-0.017	0.006	-0.001
	(0.03)	(0.03)	(0.03)	(0.04)
Potential Experience^2	0.002*	0.001	0.000	0.000
	(0.00)	(0.00)	(0.00)	(0.00)
Adj. AFQT Score	0.004***	0.007***	0.006***	0.006***
	(0.00)	(0.00)	(0.00)	(0.00)
Mother's Highest Qualification	0.002	0.006	0.004	-0.008
	(0.01)	(0.00)	(0.00)	(0.01)
Father's Highest Qualification	0.006	0.002	0.007**	0.006
	(0.00)	(0.00)	(0.00)	(0.00)
Siblings	-0.009	-0.005	-0.005	-0.005
	(0.01)	(0.01)	(0.01)	(0.01)
Quartile of Family Income				
1st	-0.163*	-0.074	-0.131*	-0.117
	(0.09)	(0.09)	(0.07)	(0.08)
2nd	-0.045	0.049	0.053	0.022
	(0.09)	(0.08)	(0.07)	(0.10)
3rd	0.048	0.054	0.028	0.144*
	(0.08)	(0.08)	(0.06)	(0.09)
4th	0.167**	0.023	0.111*	0.210**
	(0.08)	(0.08)	(0.06)	(0.09)
African-American	-0.012	-0.069	0.080	0.099
	(0.05)	(0.06)	(0.05)	(0.07)
Hispanic	0.012	0.063	0.179***	0.057
	(0.06)	(0.06)	(0.06)	(0.08)
Wald test for λ	19.62***	12.37**	6.69	10.95**
	[0.001]	[0.015]	[0.153]	[0.027]
Geographic Controls Added	Yes	Yes	Yes	Yes
Demographic Controls Added	Yes	Yes	Yes	Yes
R^2	0.243	0.229	0.165	0.189
N	7,084	8,336	8,496	5,841

$*p < .10$, $**p < .05$, $***p < 0.01$. Bootstrapped standard errors based on 200 replications in parentheses. *p-values* in brackets. Geographic controls include a dummy for urban residence and residence in one of the four Census regions (North-east, South, North-Central and West) at 14. Cohort controls include dummies for cohort of birth.

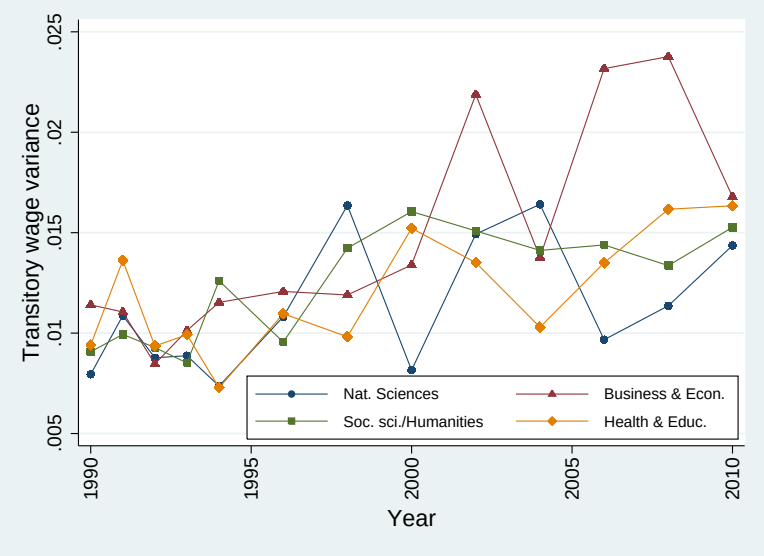


Figure 1: Transitory component of wage variance by college major, 1990 to 2010

from macroeconomic and institutional shocks. No significant difference is apparent between the other three groups, although the hike in transitory volatility for Business and Economics graduates during the 2007-2009 recession is noteworthy. 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 degree type. 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 4: Estimates of Variance of Potential Wages

	Natural Sciences		Business & Economics		Social Sciences & Humanities		Education & Health	
A. Transitory Component ($\hat{\psi}^2_{mt}$)	.012		.014		.013		.013	
	LB	UB	LB	UB	LB	UB	LB	UB
B. Permanent Component ($\hat{\sigma}^2_{mi}$)	.252	.680	.223	.760	.220	.721	.223	.809
Potential Wage Variance (A+B)	.264	.733	.237	.771	.233	.780	.236	.804
C. Wage Uncertainty ($\hat{\tau}^2_{mi}$)	.090	.258	.104	.244	.047	.219	.124	.240
D. Unobserved Heterogeneity ($\hat{\nu}_i$)	0	.679	0	.760	0	.721	0	.809

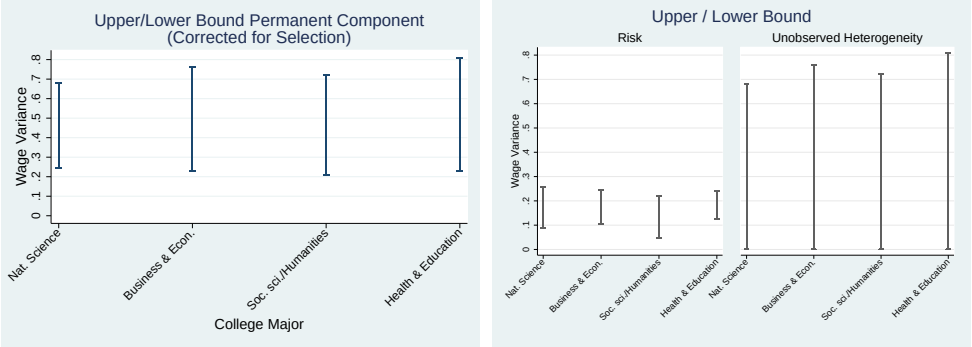
table 4. Row A describes the mean over time and by major category for the transitory component of wage inequality visually described in figure 1. Clearly, wage uncertainty due to idiosyncratic shocks is minimal for the Science group and at its maximum for the Humanities graduates, but differences are minor.

Row B shows lower and upper bounds for the permanent component corrected for selection and truncation as described in equation 26. Remember that the lower bound is set for $\rho^2_m = 1$ while we have an upper bound when $\rho^2_m = 0$.

The width of the bounds is fairly similar between the four groups. The only category slightly deviating from the norm is natural scientists experiencing the highest lower bound and the lowest upper bound of all. Overall, lower bounds are approximately 30% of upper ones. Looking at the intra-category differences in estimated upper bounds, the smallest parameter estimated for natural scientists which is 10% under the largest one for the Health and Education category. Differences in lower bounds are of comparable magnitudes.

Undeniably, the permanent component is the biggest contributor to total wage variance. The ratio of the permanent to transitory component even considering the lower bound is estimated at 20 to 1. This is in line with previous parametric (Mazza et al., 2013) and semiparametric (Mazza and van Ophem, 2010) estimates.

The key parameter in this study is 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 the Social Science and Humanities category while bounds on values for health and Education are the narrowest. Estimated risk somehow reflects the same pattern of the



(a) Permanent component of wage variance

(b) Risk and Observed Heterogeneity

Figure 2: Potential wage variance, risk and unobserved heterogeneity by major category: bounds

permanent component. The highest lower bound on risk is associated with Business graduates, while the larger upper bound is for Natural Scientists. Overall, Social Scientists are those for whom risk is lower at both ends of the interval.

The last parameter of interest is unobserved heterogeneity. By construction the interval on this parameter are larger than for any other. In fact, remembering that the expression for ν_i is given by: $\nu_i = \sigma^2_{mi}\rho^2_m$ it should be evident that when correlation is 0 so will be unobserved heterogeneity (i.e.: the individual and the econometrician share the same informations) while, at the other extreme, in the case of perfect foresight on the individual's part (i.e.: $\rho = 1$) the entirety of permanent component is attributable to unobserved heterogeneity.

The highest possible unobserved heterogeneity is found for the Health and Education category, probably reflecting the heterogeneity of this specific category classification, while Natural Scientists are at the other extreme.

Figures 2a and 2b graphically display the estimated intervals for the permanent component and for risk and unobserved heterogeneity respectively. The results discussed above are effectively summarized in this graphical representation. The difference in level between risk and unobserved heterogeneity is obvious as evident 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 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 Business and Health and Education at the high end of the spectrum, while, at the other we see that the parameter for Natural Scientists is the largest. Third, risk is highest for students of Sciences, Business and related disciplines. Fourth, the estimated bounds for the risk parameter are not as wide as one could have feared.

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 28 describes a multinomial logit model for the selection of major m instead of type r in terms of earnings, risk and individual specific taste for education:

$$V_{im} = \vartheta_1 + \vartheta_2 \hat{y}_{im} + \vartheta_3 \hat{\tau}_{imt} + \vartheta_4 \hat{\nu}_i + x_{it}\beta_m + \varrho_{mi}. \tag{28}$$

I model the probability of selecting major type m as a function of wage and risk associated with that specific category estimated in the previous step plus the same vector of covariates included in the wage regression. $\hat{y}_{im}$ 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"[13].
Remember that I do not posses point estimation neither for the risk parameter nor for the unobserved schooling factor. For this reason, I decide to estimate the effect of risk on educational choice at different values of the bounded parameter ρ.
Estimation of equation 28 via multinomial logit 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 5 are obtained through bootstrapping.
Table 5 lists the estimation for the coefficients and marginal effects at mean for equation 28 by college major category for the responsiveness to a risk increase at different values of correlation parameter.

The key parameter in this study is the effect that differences in personal risk have on educational choices. From the estimated coefficient displayed in table 5 it is immediately evident how educational decisions are significantly and negatively influenced by comparative differences in risk levels. As the theory would suggest, risk discourages selection of the specific category. This is true for any simulated level of the correlation coefficient and the magnitude of this effect increases with the amount of uncertainty on the individuals' side.
Marginal effects reported in the second section of the table highlight the economic effect of increased risk. Irrespective of level of uncertainty and major group, all coefficients are negative at least at 10% level, with most at 5%. As expectable, with as the ability to make reasonable predictions on own future payoffs diminishes, the effect of a hike in uncertainty causes a stronger movement out of the elected college major. This effect is particularly strong for the Business and Economics students and less pronounced for students in the Health and Education category.

[13] 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 5: Responsiveness to risk at different values of ρ

	(1) $\rho = .90$	(2) $\rho = .75$	(3) $\rho = .50$	(4) $\rho = .25$	(5) $\rho = .10$
Coefficient					
Risk	-2.449**	-3.221*	-8.037***	-34.68**	-221.4**
	(-2.25)	(-1.95)	(-2.75)	(-2.36)	(-2.16)
Marginal Effects					
Natural Sciences	-.412**	-.542*	-1.353**	-5.838**	-37.273**
	(.184)	(.279)	(.503)	(2.492)	(17.703)
Business & Economics	-.496**	-.652*	-1.627**	-7.021**	-44.826**
	(.218)	(.336)	(.594)	(2.952)	(20.361)
Humanities & Soc. Sciences	-.532**	-.699*	-1.744**	-7.528**	-48.061**
	(.239)	(.362)	(.635)	(3.183)	(22.406)
Education & Health	-.369**	-.486**	-1.212**	-5.233**	-33.401**
	(.160)	(.240)	(.432)	(2.256)	(15.427)
Controls added	Yes	Yes	Yes	Yes	Yes
N	15,376	15,376	15,376	15,376	15,376

$*p < .10$, $**p < .05$, $***p < 0.01$. Bootstrapped standard errors based on 200 replications in parentheses

A useful guideline for interpreting the distinct parameters shown here can be to keep in mind the different estimates in the literature for the ability of high school students to forecast the variance of their future earnings. Parametric estimates produced by Chen (2008) place this number somewhere around 0.5. Semiparametric estimates in Mazza and van Ophem (2010) find a remarkably similar result at 0.48 while Cunha et al. (2005) estimate that 60% of wage variability is foreseen by student at the time of college enrollment. In any case, previous literature has produced fairly consistent estimates which can be of great help in reading the results presented here.
Therefore, in interpreting the marginal effects displayed in table 5 I concentrate on column 3, both because this is the midpoint of possible values for ρ and because this is the closest value to previous estimates produced in the risk and education literature. The estimated marginal effects reveal that a doubling of the risk coefficient would decrease the probability of selecting a scientific degree by 27%, a Business degree by 32%, a Humanities and Social Science degree by 35% and an Education and Health degree by 24%. It is manifest that the effects are not only statistically, but also economically significant.
In conclusion, the results presented here support a Roy model for selection of education driven by comparative advantages. As expected, higher risk discourages selection of a particular type of education and the effects of

uncertainty for selection into education are far from trivial.

7 Conclusions

Exploiting recent advancements in the literature for semiparametric estimators for polychotomous choice models, this paper tests the 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 all college major types. Additional results, contributing to the growing literature of causal effects of schooling on risk, show how the well known long-running rise of earning transitory volatility for college graduates' earnings in the US for the past twenty years is confirmed here, but considerable variation exists among different major types with Scientific degrees more apt at providing some protection against macroeconomic fluctuations. I also offer some evidence for the relatively more severe consequence that the 2007-2009 economic downturn had on Business related professions compared to other categories of college graduates.

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.

All told, this research strongly supports the recent effort made in the risk and education literature for accounting for self-selection mechanisms when modeling causal impacts of educational paths on individual risk. This finding bears consequences also for the design of public policies aimed at efficiently allocating talents within the educational market. If individuals do care about uncertainty then insurance mechanism would allow for choices less guided by insurance motives and more by personal talent and inclinations. Determining which mechanisms are more efficients, under which conditions and which would be the consequence of a different allocation of talents for the economy at large is a fascinating research program that I leave for future investigation.

References

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

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

Arcidiacono, Peter, "Ability sorting and the returns to college major," *Journal of Econometrics*, 2004, *121* (1-2), 343–375.

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

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

Carneiro, Pedro and Sokbae Lee, "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 2009, *149* (2), 191 – 208.

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

_ **and Shakeeb Khan**, "Estimating the Casual Effect of Education on Wage Inequality Using IV Methods and Sample Selection Models," Working paper, University at Albany-SUNY February 2007.

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

_ , "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, 1991, pp. 175–197.

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

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

Dynan, Karen, Douglas W. Elmendorf, and Daniel E. Sichel, "The Evolution of Household Income Volatility," Finance and Economics Discussion Series: 2007/61, Board of Governors of the Federal Reserve System 2007.

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

Gallant, Ronald A. and Douglas W. Nychka, "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 1987, *55* (2), 363–390.

Goldberger, Arthur, "Abnormal Selection Bias," in S. Karlin, T. Amemiya, and L. Goodman, eds., *Studies in Econometrics, Time Series, and Multivariate Statistics*, New York: Academic Press, 1983.

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

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

_ , "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*, 1976, *5* (4), 475–492.

_ , "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *1* (74), 153–162.

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

Kane, Thomas J. and Cecilia E. Rouse, "Labor-Market Returns to Two and Four Years College," *American Economic Review*, 1995, *85* (3), 600–614.

Lee, Lung-Fei, "Some Approaches to the Correction of Selectivity Bias," *The Review of Economic Studies*, 1982, *49* (3), 355–372.

_ , "Generalized Econometric Models with Selectivity," *Econometrica*, 1983, *51* (2), 507–512.

Lemieux, Thomas, "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill," *American Economic Review*, 2006, *96* (3), 1–64.

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

Mazza, Jacopo and Hans van Ophem, "Separating Risk in Education from Heterogeneity: a Semiparametric Approach," UvA Econometrics Discussion Paper: 2010/07, Amsterdam School of Economics December 2010.

_ , _ , and Joop Hartog, "Unobserved heterogeneity and risk in wage variance: Does more schooling reduce earnings risk?," *Labour Economics*, 2013, *24* (C), 323–338.

McFadden, Daniel L., "Econometric Ananlysis of Qualitative Response Models," in Griliches D. and M. D. Intriligator, eds., *Griliches D. and M. D. Intriligator, eds.*, 1984.

Mincer, Jacob, "Investments in Human Capital and Personal Income Distribution," *Journal of Political Economy*, August 1958, *66*, 281–302.

_ , "On-the-Job Training: Costs, Returns, and Some Implications," *Journal of Political Economy*, 1962, *70* (5), 50–79.

Neal, Derek A. and William R. Johnson, "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 1996, *104* (5), 869–895.

Newey, Whitney K., "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 1997, *79* (1), 147 – 168.

_ , "Two-Step Series Estimation of Sample Selection Models," *Econometrics Journal*, 2009, *12* (1).

Powell, James, "Estimation of Semiparametric Models," in Robert F. Engle and Daniel L. McFadden, eds., *Handbook of Econometrics*, Vol. 4, Amsterdam: North Holland, 1994, pp. 2444–2523.

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

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

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

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

Trost, Robert P. and Lung-Fei Lee, "Technical Training and Earnings: A Polychotomous Choice Model With Selectivity," *The Review of Economics and Statistics*, 1984, *66* (1), 151–156.

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

Willis, Robert J and Sherwin Rosen, "Education and Self-Selection," *Journal of Political Economy*, October 1979, *87* (5), S7–36.

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