cemmap
centre for microdata methods and practice

Average and marginal returns to upper secondary schooling in Indonesia

Pedro Carneiro
Michael Lokshin
Cristobal Ridao-Cano
Nithin Umapathi

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP36/11

An ESRC Research Centre

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE **Tel:** 44 (0)20 7291 4800 **Email:** info@cemmap.ac.uk **www.cemmap.ac.uk**

Average and Marginal Returns to Upper Secondary Schooling in Indonesia

Pedro Carneiro[†]

Michael Lokshin

Cristobal Ridao-Cano

Nithin Umapathi

November 2011

Abstract:

This paper estimates average and marginal returns to schooling in Indonesia using a non-parametric selection model. Identification of the model is given by exogenous geographic variation in access to upper secondary schools. We find that the return to upper secondary schooling varies widely across individuals: it can be as high as 50 percent per year of schooling for those very likely to enroll in upper secondary schooling, or as low as -10 percent for those very unlikely to do so. Average returns for the student at the margin are well below those for the average student attending upper secondary schooling.

JEL Code: J31

Key words: Returns to Schooling, Marginal Return, Average Return, Marginal Treatment Effect

[†] Pedro Carneiro is at University College London, IFS, Cemmap and Georgetown University, Nithin Umapathi is Economist with Social Protection sector in East Asia and Pacific, Michael Lokshin is Advisor, DECRG, and Cristobal Ridao-Cano is a Senior Economist with Education sector in ECA. We thank Martin Ravallion for very useful comments. Carneiro and Umapathi gratefully acknowledge the financial support from the Economic and Social Research Council for the ESRC Centre for Microdata Methods and Practice (grant reference RES-589-28-0001) and the hospitality of the Development Economic Research Group of the World Bank. Carneiro gratefully acknowledges the support of ESRC-DFID (grant reference RES-167-25-0124), the European Research Council through ERC-2009-StG-240910-ROMETA and Orazio Attanasio's ERC-2009 Advanced Grant 249612 "Exiting Long Run Poverty: The Determinants of Asset Accumulation in Developing Countries". These are the views of the authors and do not reflect those of the World Bank, its Executive Directors, or the countries they represent. Correspondence by email to numapathi@worldbank.org

1. Introduction

The expansion of access to secondary schooling is at the center of development policy in most of the developing world. Analyzing the effects of such expansions requires knowledge of the impact of education on earnings for those affected by the expansions. In contrast with the standard model, much of the recent literature on the returns to schooling emphasizes that returns vary across individuals, and are correlated with the amount of schooling an individual takes (e.g., Card, 2001, Carneiro, Heckman and Vytlacil, 2011). In terms of the traditional Mincer equation, $Y = a + bS + u$ (where Y is log wage and S is years of schooling), b is a random coefficient potentially correlated with S. This has dramatic consequences for the way we conduct policy analysis.

In this model there is no single average return that summarizes the distribution of returns to schooling in the population. For example, the individual at the margin between two levels of schooling may have very different returns from all the infra-marginal individuals. Standard instrumental variables estimates of the returns to schooling estimate the Local Average Treatment Effect (or LATE; Imbens and Angrist, 1994), which does not in general correspond to the return to the marginal person (who is more likely to be affected by the expansion of secondary schooling than anyone else in the economy). Furthermore, different policies may affect different groups of individuals.

This paper studies the returns to upper secondary schooling in Indonesia in a setting where b varies across individuals and it is correlated with S (which in this paper is a dummy variable indicating whether an individual enrolls in upper secondary school or not). We find that the return to upper secondary schooling for the *marginal* person (who is indifferent between going to secondary schooling or not) is much lower than the returns for the *average* person enrolled in upper secondary schooling (14.2% vs. 26.9% per year of schooling).[1] Finally, we simulate what would happen if distance to upper secondary schooling was reduced by 10% for everyone in the sample, and we estimate that the return to upper secondary schooling for those induced to attend schooling by such an incentive is 14.2%.

[1] The estimated average and marginal returns to upper secondary schooling in Indonesia are 96% and 111% respectively. Average years of schooling for those who have and who have not enrolled in upper secondary schooling in Indonesia are 13.133 and 5.341, so the difference between the two is 7.79. We use this number to annualize the returns to schooling from the estimate of the total return.

When evaluating marginal expansions in access to school, the relevant quantities are the returns and costs for the marginal student, not the returns and costs for the average student. In spite of the importance of this topic, there are hardly any estimates of average and marginal returns to schooling in developing countries. Two exceptions using Chinese data are Heckman and Li (2004) and Wang, Fleisher, Li and Li (2011).

We estimate a semi-parametric selection model of upper secondary school attendance and wages using the method of local instrumental variables (Heckman and Vytlacil, 2005). Our data comes from the Indonesia Family Life Survey. Carneiro, Heckman and Vytlacil (2011) use a similar model to estimate the returns to college in the US. Although they examine a different country in a different time period, and a different level of schooling, they also find that the returns to college vary widely across individuals in the US, and that the return to college for the marginal student is well below the return to college for the average student (see also Carneiro and Lee, 2009, 2011).[2]

These papers document, across very different environments, how important it is to account for heterogeneity in the returns to schooling. They also show that it is possible to use exactly the same data which is used to produce an estimate of the return to education by instrumental variables methods (IV), and extract much more information from it (allowing us to characterize the heterogeneity in returns across individuals). This can be done using fairly standard parametric methods for estimating selection models, or using a more recent non-parametric approaches to the same problem.

Vytlacil (2002) shows that the monotonicity and independence assumptions supporting the interpretation of standard IV estimates of the effect of a particular program (such as attendance of upper secondary school) as local average treatment effects, are the same as the assumptions underlying a standard non-parametric selection model, and thus the two are equivalent. Heckman and Vytlacil (2001a, 2005) explain how to estimate such a model using the method of Local Instrumental Variables, which we apply in this paper, together with more parametric estimates of the same model. Both estimates show the importance of heterogeneity. The latter are more precise than the former, but the parametric model is more restrictive.

---

[2] There exist also papers which estimate returns for average and marginal student but which account only for selection and heterogeneity given by observable variables (ignoring selection on unobservables). One example is Dearden, McGranahan and Sianesi (2004).

This paper also proposes a methodological innovation. In the presence of multiple control variables, the construction various parameters (average returns for different groups of individuals) using the framework of Heckman and Vytlacil (2005) requires the estimation of conditional densities, where the conditioning set is of high dimensionality. These estimators are notoriously difficult to implement. We use instead a simulation method that avoids such a high dimensional non-parametric estimation problem (in contrast, Carneiro, Heckman and Vytlacil, 2010, 2011, impose restrictive assumptions to reduce the dimensionality of the problem).

Since schooling is endogenously chosen by individuals, we require an instrumental variable for schooling. We use as the instrument the distance (in kilometers) from the community of residence to the nearest secondary school (see also Card, 1995). Distance takes the value zero if there is a school in the community of residence. This variable is a strong determinant of enrolment in upper secondary school. One could be concerned that the forces driving the location of schools and parents are correlated with wages, implying that distance is an invalid instrument. Below we discuss this problem in detail.

We control for several family and village characteristics, namely father's and mother's education, an indicator of whether the community of residence was a village, religion, whether the location of residence is rural, province dummies, and distance from the village of residence to the nearest health post. Our assumption is that if we take two individuals with equally educated parents, with the same religion, living in a village which is located in an area that is equally rural, in the same province, and at the same distance of a health post, then distance to the nearest secondary school is uncorrelated with direct determinants of wages other than schooling. We present evidence that this assumption is likely to hold. In particular, we show that, once these variables are controlled for, there is no correlation between the distance to the nearest secondary school and whether the individual ever failed a grade in elementary school, how many times he repeated a grade in elementary school, and whether he had to work while attending elementary school. In addition, we show (using a different sample) that our distance variable is uncorrelated with test scores (Math, Bahasa, Science, and Social Studies) in elementary school. These are very important dimensions of the pre-secondary school experience which are measures of early ability and early home environments, and which we would expect to be correlated with distance to the nearest secondary school if this variable was endogenously determined.

Our instrumental variable estimates of the returns to schooling are higher than the returns to schooling for Indonesia estimated in Duflo (2000), with the qualification that the dataset, the instrumental variable, and the time period are not the same. Petterson (2010) finds similar rates of return using the same year and same data as us, but a different sample and a different instrument variable.

This paper proceeds as follows. Section 2 discusses the data. Section 3 reviews the econometric framework. Section 4 presents our empirical results. Section 5 concludes.

2. Data

We use data from the third wave of the Indonesia Family Life Survey (IFLS) fielded from June through November, 2000.[3] The IFLS is a household and community level panel survey that has been carried out in 1993, 1997 and 2000. The sample was drawn from 321 randomly selected villages, spread among 13 Indonesian provinces containing 83% of the country's population. The specific sample we use consists of males aged 25-60 who are employed in the labor market and who have reported non-missing wage and schooling information. We consider salaried workers, both in the government and in the private sector. We exclude females from the analysis because of low labor force participation, and we exclude self-employed workers because it is difficult to measure their earnings. The dependent variable in our analysis is the log of the hourly wage. Hourly wages are constructed from self-reported monthly wages and hours worked per week. The final sample contains 2608 working age males.

In our empirical model we collapse schooling into two categories: i) completed lower secondary or below, and ii) attendance of upper secondary or higher. While this division groups together several levels of schooling, it greatly simplifies the model and is standard in many studies of the returns to schooling (e.g., Willis and Rosen, 1979). The transition to upper secondary schooling is of substantial interest in the Indonesian context given its current effort to expand secondary education. We present both the return to upper

[3] For a description of the survey see Strauss, Beegle, Sikoki, Dwiyanto, Herawati and Witoelar (2004). In the appendix we list the main variables we use.

secondary schooling, as well as an annualized version of this parameter which we obtain by dividing the estimated return by the difference in average years of schooling completed by those with lower secondary or less and those with upper secondary or more. Upper secondary schooling corresponds to 10 or more years of completed education.[4] In order to compare our estimates with the rest of the literature (in particular, Duflo, 2000), in the appendix we also present ordinary least squares (OLS) and IV estimates of returns using a continuous education variable, corresponding to years of completed schooling.

The control variables in our models are indicator variables for age, indicators for the level of schooling completed by each of the parents (no education, elementary education, secondary education, and an indicator for unreported parental education), an indicator for whether the individual was living in a village at age 12, indicators for the province of residence, an indicator of rural residence, and distance (in kilometers) from the office of the head of the community of residence to the nearest community health post.

Our instrumental variable for schooling is the distance (in kilometers) from the office of the community head to the nearest secondary school. The distance is self-reported by the community head in the Service Availability Roster of the IFLS.[5]

Table 2 presents descriptive statistics for the main variables used in our analysis. It shows that individuals with upper secondary or higher levels of education have, on average, 108% higher wages than those with lower education. They have 7.778 more years of schooling. They are younger than those without and upper secondary education. They are more likely to have better-educated parents, to have lived in towns or cities at age 12, and to live closer to upper secondary schools, when compared to those with less than an upper secondary education.

[4] It is possible to estimate a non-parametric selection model with multiple levels of schooling but the data requirements to do it are very strong. In particular, one needs one instrumental variable for each transition. It is not feasible to pursue this with our dataset.

[5] We would have liked to use instead the distance between the community of residence in childhood and the nearest school in childhood. Our hope is that current residence and current school availability are good approximations (as in Card, 1995). We show below that this measure of distance to school is a good predictor of upper secondary school attendance.

3. Theoretical Framework

3.1 A Semi-Parametric Selection Model

This section of the model follows Heckman and Vytlacil (2005). We repeat part of the presentation in that paper because it lays out the empirical model we use, and provides the basis for discussing a new approach to estimating some of our parameters.

We consider a standard model of potential outcomes applied to schooling, as in Willis and Rosen (1979) or Carneiro, Heckman and Vytlacil (2010, 2011). Consider a model with two schooling levels:

$$\begin{aligned} Y_1 &= \alpha_1 + X\beta_1 + U_1 \\ Y_0 &= \alpha_0 + X\beta_0 + U_0 \end{aligned} \qquad (1)$$

$$S = 1 \text{ if } Z\gamma - U_s > 0 \qquad (2)$$

Y_1 are log wages of individuals if they have upper secondary education and above, Y_0 are log wages of individuals if they do not have upper secondary education, X is a vector of observable characteristics which affect wages, and U_1 and U_0 are the error terms. Z is a vector of characteristics affecting the schooling decision.

Equation (2) is a reduced form model of schooling. In theory, agents decide whether to enroll or not in upper secondary schooling based on the expected net present value of earnings with and without upper secondary schooling, and costs, which can be financial or not. There can be liquidity constraints. There is heterogeneity and we expect agents with the highest returns to upper secondary schooling ($Y_1 - Y_0$) to be more likely to enroll in higher levels of schooling. Costs and returns to schooling can be correlated. It is possible to summarize this decision process in the equation above. For a more detailed explanation see Willis and Rosen (1979) or Carneiro, Heckman and Vytlacil (2011).

It is convenient to rewrite the selection equation as:

$$S = 1 \text{ if } P(Z) > V \qquad (3)$$

$P(Z) = F_{U_S}(Z\gamma)$ and $V = F_{U_S}(U_S)$ and F_{U_S} is a cumulative distribution function of U_S . V is distributed uniformly by construction. This is an innocuous transformation given that U_S can have any density.

Finally, observed wages are:

$$Y = SY_1 + (1-S)Y_0 \qquad (4)$$

Notice that the return to schooling is

$$Y_1 - Y_0 = \alpha_1 - \alpha_0 + X(\beta_1 - \beta_0) + U_1 - U_0 \qquad (5)$$

The return to schooling varies across individuals with different X's and different U_1, U_0.

We require that Z is independent of (U_1, U_0) given X, and that Z is correlated with S (see Heckman and Vytlacil, 2005, for the full set of assumptions). These are the usual IV assumptions. In practice we use a stronger assumption: X, Z is independent of U_1, U_0, U_S. This stronger assumption is fairly standard in empirical applications of a selection model of the type described here. We discuss the advantages of using this stronger assumption in the empirical section (see also Carneiro, Heckman and Vytlacil, 2011).

The marginal treatment effect (MTE) is the central parameter of our analysis. In the notation of our paper it can be expressed as:

$$\begin{aligned} MTE(x,v) &= E(Y_1 - Y_0 \mid X = x, V = v) \\ &= \alpha_1 - \alpha_0 + x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid X = x, V = v) \end{aligned} \qquad (6)$$

The MTE measures the returns to schooling for individuals with different levels of observables (X) and unobservables (V), and therefore it provides a simple characterization of heterogeneity in returns. Heckman and Vytlacil (2005) show how to construct parameters of interest as weighted averages of the MTE. For example:

$$\begin{aligned} ATE(x) &= \int MTE(x,v) f_{V|x}(v \mid x) dv \\ ATT(x) &= \int MTE(x,v) f_{V|x}(v \mid x, S=1) dv \\ ATU(x) &= \int MTE(x,v) f_{V|x}(v \mid x, S=0) dv \end{aligned} \qquad (7)$$

where *ATE(x)* is the average treatment effect, *ATT(x)* is average treatment on the treated, *ATU(x)* is average treatment on the untreated (conditional on X=x), and $f_{V|X}(v|x)$ is the density of V conditional on X.[6]

A less standard parameter but equally (if not more) important is the policy relevant treatment effect (PRTE), introduced in the literature by Heckman and Vytlacil (2001b). It

[6] Notice that $f_{V|x}(v|x) = 1$. Heckman and Vytlacil (2005) do not use exactly this representation of the parameters. For example, they write: $ATT(x) = \int MTE(x,v) h_{TT}(v|x) dv$, where $h_{TT}(v|x)$is a parameter weight (in this case, the parameter for TT). Our representation is equivalent since $h_{TT}(v|x)$ in their paper can be shown to be equal to $f_{V|x}(v|x, S = 1)$. The only reason we make this slight change is because it is helpful for explaining our new procedure for constructing the parameter weights.

measures the average return to schooling for those induced to change their enrolment status in response to a specific policy. Obviously, it depends on the policy being considered. Consider a determinant of enrolment Z, which does not enter directly in the wage equation. The policy shifts Z from Z=z to Z=z'. The weights for the corresponding PRTE are:

$$PRTE(x) = \int MTE(x,v) f_{V|x}(v|x, S(z)=0, S(z')=1)dv$$

3.2 Estimating the *MTE*

Assuming that the unobservables in the wage (1) and selection (2) equations are jointly normally distributed the MTE could be estimated using a standard (parametric) switching regression model (see Heckman, Tobias and Vytlacil, 2001). Assume:

$$U_0, U_1, U_s \sim N(0,\Omega) \tag{8}$$

where Ω represents the variance and covariance matrix. Under this assumption:

$$MTE(x,v) = E(Y_1 - Y_0 \mid X = x, V = v) = (\alpha_1 - \alpha_0) + x(\beta_1 - \beta_0) + (\frac{\sigma_{U_S,1}}{\sigma_{U_S}} - \frac{\sigma_{U_S,0}}{\sigma_{U_S}})\Phi^{-1}(P(Z))$$

where $\sigma^2_{U_S}$ denotes variance of U_s, σ^2_i variance of U_i with i = 0,1, $\sigma^2_{U_S,i}$ covariance between U_s and U_i, $\sigma^2_{i,j}$ the covariance between U_i and U_j and Φ is the c.d.f. of the standard normal. Therefore MTE can be constructed by estimating parameters $\alpha_1, \alpha_0, \beta_1, \beta_0, \rho_1, \rho_2$.

This model relies on strong assumptions about the distribution of the error terms in equations (1-2). To relax these restrictions, we use the method of local instrumental variables that imposes no distributional assumptions on the unobservables of the model (Heckman and Vytlacil, 2000). In particular, Heckman and Vytlacil (2000) show that:

$$MTE(x,v) = \frac{\partial E(Y \mid X, P)}{\partial P}\Big|_{X=x, P=v} \tag{9}$$

where,

$$\begin{aligned} E(Y \mid X, P) &= E[\alpha_0 + X\beta_0 + S(\alpha_1 - \alpha_0) + SX(\beta_1 - \beta_0) + U_0 + S(U_1 - U_0) \mid X, P] \\ &= \alpha_0 + X\beta_0 + P(\alpha_1 - \alpha_0) + PX(\beta_1 - \beta_0) + E(U_1 - U_0 \mid S = 1, X, P)P \\ &= \alpha_0 + X\beta_0 + P(\alpha_1 - \alpha_0) + PX(\beta_1 - \beta_0) + K(P) \end{aligned} \tag{10}$$

($K(P)$ is a function of P, which can be estimated non-parametrically). Therefore, taking the derivative of (10) with respect to P:

$$MTE(x,v)=\frac{\partial E\left(Y\mid X,P\right)}{\partial P}\Big|_{X=x,P=v}=X(\beta_1-\beta_0)+K'(P) \qquad (11)$$

V can take values from 0 to 1. However, in practice it is only possible to estimate the MTE over the observed support of P. In our data the support of P is almost the full unit interval, so we are able to estimate the MTE close to its full support.

If we had assumed that Z is independent of (U_1,U_0) given X, instead of full independence between (Z,X) and (U_1,U_0), it would be difficult to estimate the MTE over a large support. In that case, for each value of X it is only possible to estimate the MTE over the support of P conditional on X, which usually will be much smaller than the unconditional support of P (for a detailed discussion see Carneiro, Heckman and Vytlacil, 2011). The assumption of full independence of (Z,X) and (U_1,U_0) is common in empirical applications of selection models and it allows us to use the full support of P.

Equations (10) and (11) can be estimated using standard methods. In particular, we use the partially linear regression estimator of Robinson (1988) to estimate (β_1,β_0). Then we compute $R=Y-\left[\alpha_0+X\beta_0+PX\left(\beta_1-\beta_0\right)\right]$. (α_1,α_0) cannot be identified separately from $K(P)$. $K(P)$ (and $K'(P)$) is estimated using a locally quadratic regression (Fan and Gijbels, 1996) of R on P. A simple test of heterogeneity and selection on unobserved characteristics is a test of whether $K'(P)$ is flat (or of whether $E(Y\mid X, P)$ is nonlinear in P). If $K'(P)$ is flat then heterogeneity is not important, or individuals do not select on it.

3.3 Average Marginal Returns to Education

Economic decisions involve comparisons of marginal benefits and marginal costs. Therefore it is important to estimate the average returns to schooling for individuals at the margin between enrolling or not. They would be those who are the most likely to change their upper secondary schooling decision in response to a change in education policy.

The definition of who is marginal depends on the policy being considered. This is made clear in Carneiro, Heckman and Vytlacil (2010, 2011), who focus on three particular definitions of individuals at the margin:

$$i)\left|P-V\right|<\varepsilon,\;\; ii)\left|Z\gamma-U_s\right|<\varepsilon,\;\; iii)\left|\frac{P}{U}-1\right|<\varepsilon.$$

These correspond to three different marginal policy changes.[7]

In this paper we estimate the average marginal returns to upper secondary schooling in Indonesia according to the definition of marginal in ii) above, although we could have chosen a different one. The MTE provides a general characterization of heterogeneity in returns and from it we can construct various other parameters.

Carneiro, Heckman and Vytlacil (2010) show how it is possible to write the average marginal treatment effect (or AMTE, the return for the marginal person) as a weighted average of the MTE:

$$AMTE(x) = \int MTE(x,v) f_{X,V}(v \| Z\gamma - Us| < \varepsilon, x) dv \qquad (12)$$

3.4 Estimating vs. Simulating the Weights: A New Procedure

So far this section has shown how to recover the MTE from the data, and how to construct economically interesting parameters as weighted averages of the MTE. Heckman and Vytlacil (2005) and Carneiro, Heckman and Vytlacil (2010) provide formulas for the necessary weights in equations 7 and 12, conditional on *X*:

$$\begin{aligned}
&f_{V|X}(v)=1\\
&f_{V|X}(v \mid X, S=1)=\frac{1-F_{P|X}(v \mid X)}{E(P \mid X)}\\
&f_{V|X}(v \mid X, S=0)=\frac{F_{P|X}(v \mid X)}{E(P \mid X)}\\
&f_{V|X}(v \mid X, S(z)=0, S(z')=1)=\frac{F_{P|X}(v \mid X)-F_{P'|X}(v \mid X)}{\int\left[F_{P|X}(v \mid X)-F_{P'|X}(v \mid X)\right]dv}\\
&f_{V|X}(v \mid X, |Z\gamma-V|<\varepsilon)=\frac{f_{P|X}(v \mid X) f_{U_S|X}\left[F_{U_S|X}^{-1}(v \mid X)\right]}{E\left[f_{U_S|X}(Z\gamma \mid X)\right]}
\end{aligned} \qquad (13)$$

[7] The three policy changes considered are (i) a policy that increases the probability of attending college (P) by an amount α, so that $P_\alpha = P_0 + \alpha$; (ii) a policy that changes each person's probability of attending college by the proportion (1+ α), so that $P_\alpha = (1+\alpha)P_0$; and (iii) a policy intervention that has an effect similar to a shift in one of the components of Z, say Z^k, so that $Z_\alpha^k = Z^k + \alpha$ and $Z_\alpha^j = Z^j$ for $j \neq k$.

where $f_{P|X}(p \mid X)$ and $F_{P|X}(p \mid X)$ are respectively the p.d.f and the c.d.f. of *P* conditional on *X*, $f_{U_S|X}(u_S \mid X)$ and $F_{U_S|X}(u_S \mid X)$ are respectively the p.d.f and the c.d.f. of U_S conditional on *X*, and $F_{P'|X}(p \mid X)$ is the c.d.f. of *P* conditional on *X* when *Z=z'*.

In practice it is difficult to implement these formulas since they involve estimation of conditional density and distribution functions, and *X* is generally a high dimensional vector (there are 28 variables in *X* in our empirical work). Therefore, Carneiro, Heckman and Vytlacil (2010, 2011) have aggregated *X* into an index, namely $I = X(\beta_1 - \beta_0)$, and proceeded by estimating conditional densities and distributions of *P* with respect to *I*.

There is no theoretical basis for this aggregation which makes it quite unattractive. In this paper we use an alternative procedure, which avoids making this aggregation, and sidesteps the problem of estimating a multidimensional conditional density function.

Notice that the selection equation relates *S*, *X*, *Z*, and *V* (which is uniform by construction). Using the estimated parameters, we can simulate the following objects:

$$f_{V|X}(v|x), f_{V|X}(v|S=1,x), f_{V|X}(v|S=0,x), f_{V|X}(v||Z\gamma - Us| < \varepsilon, x)$$

Once we construct these objects, we just need apply them to equations (7) and (12). This simulation procedure is simple, and its steps are described in detail in the appendix.

4. Empirical Results

4.1 Is Distance to School a Valid Instrument?

To account for the potential endogeneity of the schooling decision we instrumented schooling with the distance to the nearest secondary school.[8] In order for it to be a valid instrument distance to school needs to satisfy two conditions: i) it should affect the probability of school enrolment and ii) it should have no direct effect on adult wages.

We show that condition i) is satisfied. Condition ii) is controversial if families and schools do not randomly locate across locations in Indonesia. For example, Carneiro and Heckman (2002) and Cameron and Taber (2004) show that individuals living closer to universities in the US have higher levels of cognitive ability and come from better family backgrounds. In Indonesia, those who have better educated parents are also located closer

[8] Distance to the nearest school has been used by Card (1995), Kane and Rouse (1995), Kling (2001), Currie and Moretti (2003), Cameron and Taber (2004) and Carneiro, Heckman and Vytlacil (2011).

to secondary schools. However, it is possible that school location is exogenous after we account for a very detailed set of individual and regional characteristics, namely: age (or cohort), parental education, religion, an indicator for whether the individual was living in a city or in village at age 12, an indicator for whether the individual lived in a rural area at age 12, dummies for the province of residence, and distance to the nearest health post.

One way to investigate the plausibility of such a story is to check whether distance to the nearest secondary school is correlated with pre-secondary educational outcomes of each individual (grade repetition, work in school, test scores). If there was non-random sorting of families and schools across locations in such a way that distance to secondary school was correlated with adult wages, it would surely appear in these variables.

Table 3 examines whether distance to upper secondary school is correlated with whether an individual ever repeated a grade in elementary school, the number of repetitions in elementary school (both of which are measure of early school success), and whether the individual worked while in primary school. If our instrument is valid it should not be correlated with such early characteristics of educational experience. Our results show no apparent correlation between distance to school and these variables.

In addition, Table 4 examines comprehensive exam scores in math, science, social studies and Bahasa. The sample used in this table is not exactly the sample used in our regressions, because it is only possible to gather elementary school test scores for a very small proportion of individuals in our original sample. Therefore, in the regression showed in this table, we placed no age or gender restrictions in the sample. Again, we find no correlation between the distance to school and test scores in four different subjects.[9] This evidence is suggestive that our empirical strategy is valid.

There is another important reason why condition ii) might be violated. If regions where schools are abundant are also regions where other infrastructure is also abundant, then we may be confounding the impact of school availability on wages with the impact of other infrastructure on wages (see the argument in Jalan and Ravallion, 2002). This will be true unless labor is perfectly mobile, which is unlikely to be the case in Indonesia.

[9] Considering a more restricted sample results in a small number of observations. Our main conclusions are unchanged, but results are fairly imprecise.

Our model includes a large set of regional controls which should absorb much of this variation. The argument we use is that our assumption is valid conditional on all the included controls. In addition, we show that removing these detailed regional controls hardly affects our results, indicating that this problem is unlikely to be important in our setting. As argued in Duflo (2004), perhaps the response of other (private or public) infrastructure to school construction and to a better skilled workforce is very slow.

Table 5 shows that distance to the nearest secondary school is a strong predictor of enrolment in secondary school. We run a logit regression where the dependent variable is an indicator taking value 1 if an individual ever attended upper secondary school and the regressors include distance to the nearest secondary school and all the control variables mentioned above. The table displays marginal effects of each variable on the probability of enrolling in upper secondary education. We include as a control the distance to the nearest health post as a proxy for location characteristics and, unlike distance to school, distance to health post does not predict school enrollment. Children of highly educated parents are more likely to attend upper secondary school than children of parents with low levels of education. Catholics and Protestants are much more likely to attend secondary school than Muslims (the omitted category). Children living in small villages and in rural areas are less likely to attend upper secondary school than those living in large cities and urban areas.

This model is fairly flexible in the sense that the impact of distance on secondary school attendance varies with X. In particular, we interact distance to school with age (which, for a fixed year, also captures cohort), religion, parental education, and rural residence. It is useful to estimate such a rich model for two related reasons. First, because of its flexibility. Second, by allowing the impact of the instrument to vary will the variables in X we are able to use extra variation in the instrument. As a result, the standard errors in the IV estimates and in the selection model are smaller than if we just used a simpler model without these interactions. Therefore, the basic estimates in this paper will come from this model, while estimates of a simpler model without interactions are presented in the appendix (we discuss them below). All average derivatives are computed at the mean value of the X variables.

Table 5 also displays p-values for the test of the null hypothesis that distance to school does not affect upper secondary school attendance. We perform a joint test on all coefficients involving distance. We reject that distance to school does not determine upper secondary school attendance.

4.2 Standard Estimates of the Returns to Schooling

In order to more easily make a comparison between our data and estimates and those in the literature we start by presenting standard OLS and IV estimates of the returns to schooling. Throughout the paper schooling takes two values: 0 for less than upper secondary, and 1 for upper secondary or above. We use the log hourly wage in 2000 as our dependent variable. The full set of controls consists of: age (or cohort), parental education, religion, an indicator for whether the individual was living in a city or in a village at age 12, an indicator for whether the individual lived in a rural area at age 12, dummies for the province of residence, and distance to the nearest health post.

We present ordinary least squares (OLS) and IV results. This is shown in Table 6. Recall from table 2 that individuals with upper secondary schooling or above have on average 13.133 years of schooling, while those with less than upper secondary have on average 5.341 years of schooling. The difference between the two groups is 7.792 years of schooling. Using this figure to annualize the returns to upper secondary education we have an OLS estimate of 9% and an IV estimate of 12.9% (without annualizing returns we have OLS and IV estimates of 70.5% and 100% respectively).

These estimates are higher than (but of comparable magnitude to) those in Duflo (2001), although we use more recent data. Petterson (2010) finds a return of 14% using the same data as we do, but a different sub-sample and instrument.

As in most of the literature, our IV estimates of the return to education are larger than OLS estimates. Card (2001) suggests that such a finding indicates that returns to schooling are heterogeneous and the marginal individual induced to enroll in school by the change in the instrument has a higher return than the average individual. Carneiro and Heckman (2002) and Carneiro, Heckman and Vytlacil (2011) show that, in the case of college attendance in the US, IV estimates can be above OLS estimates even if the marginal individual has a lower return than the average. Another reason why IV can exceed OLS is measurement error in schooling. Although schooling is relatively well measured in the US (Card, 1999), this is not necessarily the case in Indonesia.

Appendix table A1 presents OLS and IV estimates where we use years of schooling as the main explanatory variable (as opposed to upper secondary schooling). The first column in this table shows coefficients of an OLS regression of log wages on years of schooling and several controls. The estimated return to a year of schooling is 9.6%. The second column shows the first stage of the two stage least squares estimator, i.e., a regression of years of schooling on the instrument and the control variables. It shows that distance to school is negatively related to schooling attainment. Finally, column 3 shows the IV estimate of the return to schooling, which is 15.7%. In appendix table A2 we also present IV estimates of returns for models where we do not interact the instrument with the variables in X. The point estimate is smaller than the one in Table A1, and the standard error is larger, but the main pattern remains: the IV estimate is much higher than the OLS estimate. In a model with heterogeneous returns, it is not surprising that the instrumental variable is sensitive to the choice of instrument. For the remaining of the paper, we present a parallel set of results in the appendix in which we do not interact the instrument with X in the selection equation.[10] Finally, in appendix table A3 we present results were we omit regional dummies from the model. Our IV estimate is very similar to the ones in tables A1 and A2. This indicates that regional variation in infrastructure, which is correlated with the availability of schooling, is unlikely to be driving our results.

OLS and IV estimates hide considerable heterogeneity in returns and, as emphasized in Heckman and Vytlacil (2005), Heckman, Urzua and Vytlacil (2006), and Carneiro, Heckman and Vytlacil (2011), it is not clear which question is answered by the IV estimate. In order to further investigate this issue we use the framework of section 3. We estimate parametric (assuming joint normality of (U_1, U_0, U_S)) and semi-parametric versions of the model (relaxing assumptions on the joint distribution of (U_1, U_0, U_S)).

[10] We do this for two reasons. First, to show that the main patterns in our results are not driven by choosing the specific way the instrument enters the model. Second, because the first stage F-statistic is higher in the case where we use a single IV (F=11.34) than when we use multiple IVs (F=3.62) consisting of distance interacted with different components of X. We will see throughout the paper that using the expanded set of instruments allows us to get similar results and lower standard errors than we use a single (but apparently stronger) instrument.

4.3 Average and Marginal Treatment Effect Estimates

We start with the semi-parametric model. We construct P as a predicted probability of ever attending upper secondary school from a logit regression of upper secondary school attendance on the X and Z variables of section 3. Table 5, discussed above, reports the coefficients of the logit model. All variables work as expected.

It is only possible to identify the MTE over the support of P. Therefore, we need to examine the density of P for individuals who attend upper secondary school or above, and those who do not. This is done in Figure 1, which shows the distributions of the predicted propensity score (P) for these two groups. The supports for these two distributions overlap almost everywhere, although the support at the tails is thin for low values of P among those with upper secondary school or above. We construct the MTE as described in Section 2. In order to estimate $K(P)$ we run a local quadratic regression of R on P, using a Gaussian kernel and a bandwidth of 0.2. The implied $MTE(x,v)$ is computed by calculating the slope on the linear term of the local quadratic regression.[11]

Figure 2 displays the estimated MTE (which we evaluate at the mean values of the components of X). The MTE is monotonically decreasing for all values of V. Returns are very high for individuals with low values of V (individuals who are more likely to enroll in upper secondary school or facing high costs). The figure demonstrates substantial heterogeneity in the return to schooling, which ranges from 34% for individuals with V around 0.1 to 13% for those with V close to 0.5, and becomes negative for those with values of V close to 1. The fact that returns are the lowest for individuals who are least likely to go to school is consistent with a simple economic model where agents sort into different levels of schooling based on their comparative advantage.

Unfortunately the standard errors on our estimated MTE are quite wide (standard errors are estimated using the bootstrap). However, it is still possible to reject that the MTE is flat. Table 7 tests whether adjacent segments of the MTE are equal (see Carneiro, Heckman and Vytlacil, 2011). Take, for example, the first column of the table. In the first line we show the average value the MTE takes when X is fixed at its mean and V takes values between 0 and 0.1, while the second line corresponds to values of V between 0.1 and 0.2. The third line shows the difference between the first two lines, and the fourth

[11] The coefficients on X in the outcome equations are presented in table A4 in the appendix.

line reports the p-value of a test of whether this difference is equal to zero. We reject equality in almost all columns of the table at the 5% significance level. Therefore, we are able to reject that the MTE is flat, even with the large standard errors shown in figure 2.

Figure 3 shows that the standard errors improve when we estimate the MTE assuming joint normality of (U_1, U_0, U_S). The shape of the MTE is declining as before, although the normality assumption does not allow the MTE to have a flat section as in Figure 2, so the MTE is declining everywhere, again taking negative values for very high values of V.

Table 8 presents average returns to upper secondary schooling for different groups of individuals. The return to upper secondary school for a random person is 12.3%. The return for those individuals who were enrolled in upper secondary schooling is considerably higher, at 26.9%. The return that individuals who did not go to upper secondary school would experience had they gone there is 1.7%. Average parameters are estimated with the assumption of full support (although figure 1 shows a very small lack of support in the left tail of the distribution of P). Estimates of the return to the marginal student (AMTE) are robust to the lack of full support (Carneiro, Heckman and Vytlacil, 2010, 2011). The return to the marginal student is 14.2%, well below the return to the average student in upper secondary school (26.9%).

Finally, the last line of Table 8 reports the average return for those induced to attend upper secondary school by a particular policy shift: a 10% reduction in distance to an upper secondary school. This is the parameter needed to understand the impacts of such an education expansion. By coincidence, it is remarkably similar to the MPRTE.

In the appendix we show that results are similar but more imprecise when we do not interact Z and X in the selection equation. This is reassuring, and shows the usefulness of accounting for a more flexible model for the precision of our estimates.[12]

5. Conclusion

Indonesia has an impressive record of educational expansion since the 1970s. The enrollment rates are nearly universal for elementary schooling and are around 75% for secondary education. There is an ongoing effort to extend universal education attainment to the secondary level. And although enrollment in secondary education continues to rise

[12] See tables A2 , A5 and A6, and figures A1, A2 and A3.

we find striking inequality in returns to education. Individuals who are more likely to be attracted by educational expansions at the upper secondary level (marginal) have lower average returns than those already attending upper secondary schooling. In this paper we document a large degree of heterogeneity in the returns to upper secondary schooling in Indonesia. We estimate the return to upper secondary education to be 12 percentage points higher (per year of schooling) for the average than for the marginal student.

Therefore, efforts aimed at educational expansion will attract students with lower levels of returns. However, returns are still fairly high for the marginal person, and therefore further expansions are probably justified. Our estimates also show that it is probably not optimal to bring everybody into upper secondary education.

What is behind such a large inequality in the returns to schooling? There is a growing body of literature that argues that human capital outcomes later in life (including the ability to learn) are largely influenced by what happens early in life (e.g., Carneiro and Heckman, 2003). It is therefore important for the design of schooling policy to determine whether the inequality in secondary schooling outcomes can be remedied at earlier stages, for example during early childhood, or during the elementary school years.

References

Bjorklund, A. and R. Moffitt (1987) , "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics* , 69:42-49.

Cameron, S. and C. Taber (2004), "Estimation of Educational Borrowing Constraints Using Returns to Schooling", *Journal of Political Economy* , part 1, 112(1): 132-82.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling ", *Aspects of Labour Economics: Essays in Honour of John Vanderkamp* , edited by Louis Christofides, E. Kenneth Grant and Robert Swindinsky. University of Toronto Press.

Card, D. (1999), "The Causal Effect of Education on Earnings," Orley Ashenfelter and David Card, (editors), Vol. 3A, *Handbook of Labor Economics*, Amsterdam: North-Holland.

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* , 69(5): 1127-60.

Card, D. and T. Lemieux (2001), "Can Falling Supply Explain the Rising Return to College For Younger Men? A Cohort Based Analysis," *Quarterly Journal of Economics* 116: 705-46.

Carneiro, P. and J. Heckman (2002), "The Evidence on Credit Constraints in Post-secondary Schooling," *Economic Journal* 112(482): 705-34.

Carneiro, P. and J. Heckman (2003), "Human Capital Policy," in *Inequality in America: What Role for Human Capital* Policy, J. Heckman, A. Kruger and B. Friedman eds, MIT Press.

Carneiro, P., J. Heckman and E. Vytlacil (2010), "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin", *Econometrica*.

Carneiro, P., J. Heckman and E. Vytlacil (2011), "Estimating Marginal Returns to Education ", *American Economic Review*, 101(6).

Carneiro, P. and S. Lee (2009), "Estimating Distributions of Potential Outcomes using Local Instrumental Variables with an Application to Changes in College Enrolment and Wage Inequality", Journal of Econometrics.

Carneiro, P. and S. Lee (2011), “Trends in Quality Adjusted Skill Premia in the US: 1960 to 2000”, *American Economic Review*, 101(6).

Currie, J. and E. Moretti (2003), Mother's Education and the Intergenerational Transimission of Human Capital: Evidence from College Openings, *Quartely Journal of Economics* , 118:4.

Dearden, L., L. McGranahan, B. Sianesi (2004), “Returns to Education for the ‘Marginal Learner’: Evidence from the BCS70”, CEEDP 45, Center for the Economics of Education, London School of Economics and Political Science.

Duflo, E. (2001), “Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment”, *American Economic Review*, 91(4), 795-813.

Duflo, E. (2004), “The medium run effects of educational expansion: evidence from a large school construction program in Indonesia”, Journal of Development Economics, 74, 163-197.

Fan, J. and I. Gijbels (1996), Local Polynomial Modelling and its Applications, New York, Chapman and Hall.

Griliches, Z. (1977), “Estimating the Return to Schooling: Some Persistent Econometric Problems”, Econometrica.

Petterson, G. "Do supply-side education programs targeted at under-served areas work? The impact of increased school supply on education and wages of the poor and women in Indonesia." PhD dissertation (Draft), Department of Economics, University of Sussex

Heckman, J. and X. Li (2004), “Selection Bias, Comparative Advantage, and Heterogeneous Returns to Education: Evidence from China in 2000”, *Pacific Economic Review*.

Heckman, J., S. Urzua and E. Vytlacil (2006), “Understanding What Instrumental Variables Really Estimate in a Model with Essential Heterogeneity”, *Review of Economics and Statistics*.

Heckman, J. and E. Vytlacil (1999), Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects, *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

Heckman, J. and E. Vytlacil (2001a), "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powells, (eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honorof Takeshi Amemiya*, (Cambridge: Cambridge University Press, 2000), 1-46.

Heckman, J. and E. Vytlacil (2001b), "Policy Relevant Treatment Effects", *American Economic Review Papers and Proceedings*.

Heckman, J. and E. Vytlacil (2005), "Structural Equations, Treatment, Effects and Econometric Policy Evaluation," *Econometrica*, 73(3):669-738.

Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects,' *Econometrica*, 62(2):467-475.

Jalan, J. and M. Ravallion (2002), "Geographic poverty traps? A micro model of consumption growth in rural China", *Journal of Applied Econometrics*, 17, 329-346.

Kane, T. and C. Rouse (1995), "Labor-Market Returns to Two- and Four-Year College", *American Economic Review*, 85(3):600-614.

Kling, J. (2001), "Interpreting Instrumental Variables Estimates of the Returns to Schooling", *Journal of Business and Economic Statistics*, 19(3), 358-364.

Robinson, P. (1988), "Root-N-Consistent Semiparametric Regression", Econometrica, 56(4), 931-954.

Strauss, J., K. Beegle, B. Sikoki, A. Dwiyanto, Y. Herawati and F. Witoelar (2004), "The Third Wave of the Indonesia Family Life Survey (IFLS): Overview and Field Report", March 2004. WR-144/1-NIA/NICHD.

Vytlacil, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result", Econometrica, 70(1), 331-341.

Wang, X., B. Fleisher, H. Li and S. Li, "Access to Higher Education and Inequality: the Chinese Experiment", working paper, Ohio State University.

Willis, R. and S. Rosen (1979), "Education and Self-Selection,"Journal of Political Economy, 87(5):Pt2:S7-36.

Table 1: Definitions of variables used in the empirical analysis

Variable	Definition
Y	Log hourly earnings for salaried males
S = 1	Ever enrolled in upper secondary school; zero otherwise
X	Age, age squared, respondent's religion – protestant, catholic and other, mother's and father's education – elementary, secondary or higher, distance to the nearest health post in km from the community, rural residence, province of residence – West Sumatra, South Sumatra, Lampung, Jakarta, Central Java, Yogyakarta, East Java, Bali, West Nussa Tengara, South Kalimanthan, South Sulawesi
Z	Distance in km from the community heads office to nearest secondary school, interactions of distance with age, age squared, religion, parental education and rural residence

Table 2: Sample statistics for the treatment groups

	Upper secondary or higher	*Less than upper secondary*
	N = 1085	*N = 1523*
Log hourly wages	8.198	7.481
Years of education	13.133	5.341
Distance to school in km	1.053	1.564
Distance to health post in km	0.889	1.079
Age	37.058	38.675
Religion Protestant	0.050	0.022
Catholic	0.029	0.009
Other	0.062	0.043
Muslim	0.860	0.927
Father uneducated	0.130	0.383
…elementary	0.503	0.507
...secondary and higher	0.330	0.061
...missing	0.020	0.037
Mother uneducated	0.201	0.425
…elementary	0.484	0.406
...secondary and higher	0.204	0.022
...missing	0.098	0.133
Rural household	0.240	0.476
North Sumatra	0.057	0.063
West Sumatra	0.047	0.058
South Sumatra	0.048	0.032
Lampung	0.016	0.027
Jakarta	0.181	0.095
Central Java	0.085	0.163
Yogyakarta	0.092	0.054
East Java	0.121	0.180
Bali	0.056	0.038
West Nussa Tengara	0.050	0.048
South Kalimanthan	0.040	0.020
South Sulawesi	0.035	0.035

Source: Data from IFLS3. Sample restricted to males aged 25-60 employed in salaried jobs in government and private sectors. Hourly wages constructed based on self-reported monthly wages and hours.

Table 3: Regression of elementary education experiences on distance to school

	Failed grade	Number of repeats	Worked
Dist. to nearest secondary school in km	0.007	0.011	-0.001
	(0.007)	*(0.008)*	*(0.005)*
Number of observations	2,248	2,244	2,250
R2	0.041	0.043	0.043

Note: Sample restricted to males with the repeated grade information non-missing. The individual and family controls include age, age squared, religion, fathers and mother's schooling levels completed, distance to local health outpost, rural and province dummies. All regressions include individual and family controls, and location fixed effects. Standard errors (in parenthesis) are robust to clustering at the community level, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table 4: Regression of comprehensive exam test scores from elementary school on distance to school

	Math	Bahasa	Science	Social Studies
Distance to nearest secondary school	0.001	-0.002	-0.004	-0.005
	(0.005)	*(0.005)*	*(0.005)*	*(0.005)*
Number of observations	1,652	1,668	1,621	1,605
R2	0.134	0.187	0.124	0.115

Note: Sample includes everyone with non-missing test scores. Test scores recorded from score cards. The individual and family controls include age, age squared, religion, fathers and mother's schooling levels completed, distance to local health outpost, rural and province dummies. All regressions include individual and family controls, and location fixed effects. Standard errors (in parenthesis) are robust to clustering at the community level, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table 5: Upper school decision model – Average Marginal Derivatives

	Coef	Average Derivative
Dist. to secondary school in km	-0.123***	-0.0300**
	(0.040)	(0.0127)
Age	0.077*	0.0130
	(0.044)	(0.0090)
Age Squared	-0.096*	-0.0162
	(0.055)	(0.0111)
Protestant	0.730***	0.1382***
	(0.264)	(0.0484)
Catholic	1.211***	0.2123**
	(0.395)	(0.0890)
Other religions	0.245	0.0552
	(0.363)	(0.0878)
Fathers education elementary	0.766***	0.1342***
	(0.127)	(0.0217)
Father higher education	1.835***	0.3769***
	(0.178)	(0.0320)
Mother education elementary	0.443***	0.0852***
	(0.123)	(0.0230)
Mother higher education	1.851***	0.3730***
	(0.237)	(0.0418)
Rural	-0.593***	-0.1143***
	(0.110)	(0.0276)
Distance to health post in km	-0.017	0.0000
	(0.040)	(0.0083)
Location fixed effect	Yes	
Test for joint significance of instruments: Chi-square/p-value	9.42/0.0021	

Note: This table reports the coefficients and average marginal derivatives from a logit regression of upper secondary school attendance (a dummy variable that is equal to 1 if an individual has ever attended upper secondary school and equal to 0 if he has never attended upper secondary school but graduated from lower secondary school) on several variables. Type of location is controlled for using province dummy variables. A dummy variable for missing parental education is included in the regressions but not reported in the table. The first column presents coefficients of logit where only distance to school is used an IV. In the second column average derivatives (computed at the average values of X) are presented and instruments include distance to secondary school and interactions with all the Xs. Reference categories are Muslim, not educated. Standard errors (in parenthesis) are robust to clustering at the community level, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table 6: Annualized OLS and IV estimates of the return to upper secondary schooling

	OLS	**IV**
Upper secondary (annualized)	0.090***	0.129***
	(0.005)	*(0.048)*
Age	0.052***	0.048**
	(0.019)	*(0.020)*
Age Squared	-0.042*	-0.037
	(0.023)	*(0.025)*
Protestant	0.182**	0.142
	(0.084)	*(0.104)*
Catholic	0.059	0.001
	(0.189)	*(0.202)*
Other religions	0.109	0.097
	(0.126)	*(0.125)*
Fathers education elementary	0.135***	0.091
	(0.048)	*(0.070)*
Fathers education secondary or higher	0.215***	0.101
	(0.067)	*(0.153)*
Mother's education elementary	-0.052	-0.080
	(0.048)	*(0.060)*
Mother's education secondary or higher	-0.031	-0.128
	(0.078)	*(0.136)*
Rural household	0.111**	0.152**
	(0.045)	*(0.068)*
Distance to health post in km	-0.023	-0.020
	(0.018)	*(0.017)*
Location controls	YES	YES
Number of observations	2,608	2,608
Test for joint significance of instruments: F-stat/p-value	2.22/0.00	
R2	0.210	0.190

Note: This table reports the coefficients for OLS and 2SLS IV for regression of log of hourly wages on upper school attendance (a dummy variable that is equal to 1 if an individual has ever attended upper secondary school and equal to 0 if he has never attended upper secondary school but graduated from lower secondary school), controlling for parental education, religion and location. Excluded instruments are distance to secondary school and interactions with parental education, religion and age. Type of location is controlled using province dummies. A dummy variable for missing parental education is included in the regressions but not reported in the table. Reference categories are Muslim for religion, and not educated for education.Standard errors (in parenthesis) are robust to clustering at the community level with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table 7: Test for heterogeneity in returns: compare adjacent sections of the semi-parametric MTE

Ranges of U_S for *LATE*^{j}	(0,0.1)	(0.1,0.2)	(0.2,0.3)	(0.3,0.4)	(0.4,0.5)	(0.5,0.6)	(0.6,0.7)	(0.7,0.8)	(0.8,0.9)
Ranges of U_S for *LATE*$^{j+1}$	(0.1,0.2)	(0.2,0.3)	(0.3,0.4)	(0.4,0.5)	(0.5,0.6)	(0.6,0.7)	(0.7,0.8)	(0.8,0.9)	(0.9,1)
Difference in LATEs	-0.078	-0.039	-0.013	-0.012	0.00	0.005	-0.014	-0.024	-0.04
p-value	0.00	0.00	0.00	0.00	0.597	0.759	0.005	0.00	0.00

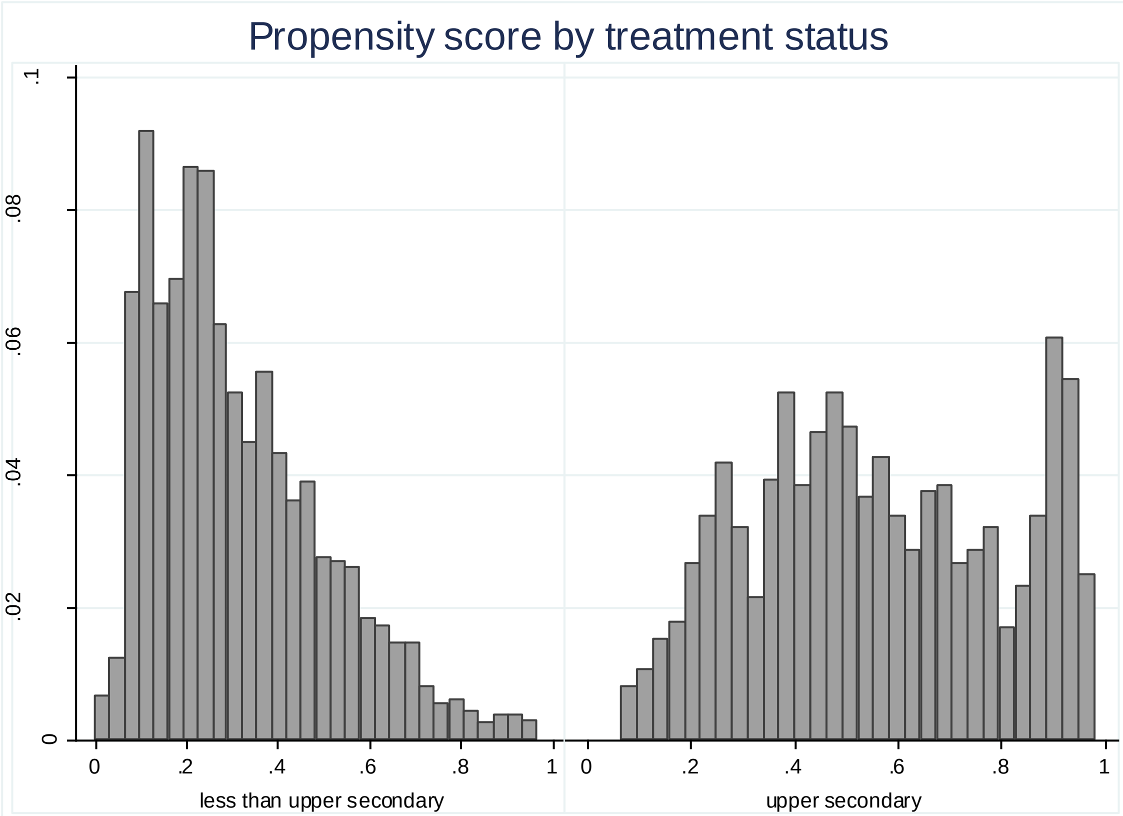
Note: In order to compute the numbers in this table we construct groups of values of *Us* and average the MTE within these groups, where U_S^{Lj} and U_S^{Hj} are the lowest and highest values of *Us* defined for interval *j*. Then we compare the average MTE across adjacent groups and test whether the difference is equal to zero using the bootstrap with 250 replications.

Table 8: Estimates of Average Returns to Upper Secondary Schooling with 95% confidence interval

Parameter	*Non parametric Estimate*	*Normal selection model*
ATT	0.269***	0.201***
	(.069, 0.47)	(0.05,0.35)
ATE	0.123*	0.066
	(-0.019, 0.266)	(-0.029,0.163)
ATU	0.017	-0.029
	(-0.236, 0.27)	(-0.175,0.116)
MPRTE	0.142***	
	(.038, 0.246)	
PRTE	0.142***	
	(.038, 0.247)	

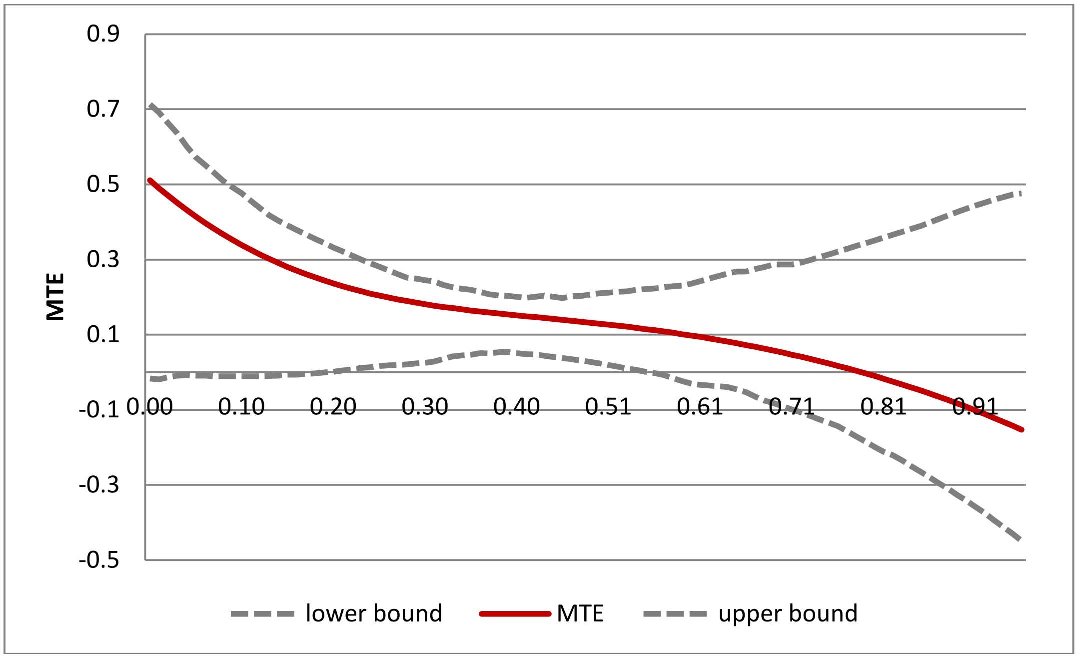
Note: This table presents estimates of various returns to upper secondary school attendance for the semi-parametric and normal selection models: average treatment on the treated (ATT), average treatment effect (ATE), treatment on the untreated (ATU), marginal policy relevant treatment effect (MPRTE), and the policy relevant treatment effect (PRTE) corresponding to a 10% reduction in distance to upper secondary school. Returns to upper school are annualized to show returns for each additional year. Bootstrapped 95% confidence interval are reported in parentheses, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Figure 1: Propensity score (*P*) support for each schooling group *S* = 0 and *S* = 1



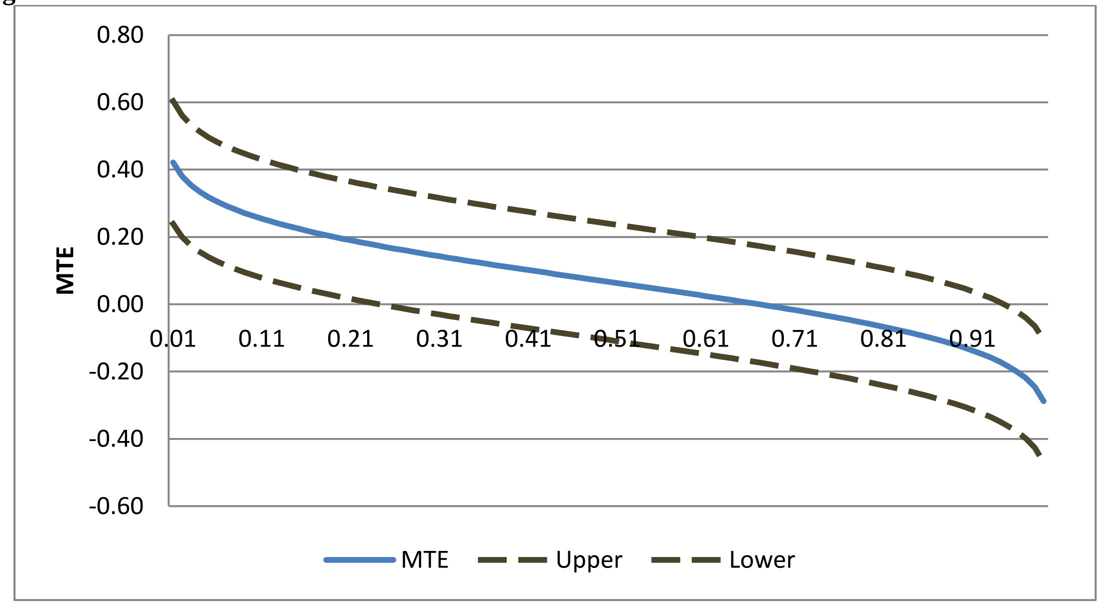
Note: *P* is estimated probability of going to upper secondary school. It is estimated from a logit regression of upper school attendance on Xs, distance to school, interactions of X and distance to school (Table 5).

Figure 2: Marginal treatment effect with 90% Confidence Interval – Semi-parametric regression estimates



Note: To estimate the E(Y_1-Y_0|X, U_s) function we used a partial linear regression of log wages on X and K(P) ,with a bandwidth of 0.2. X includes age, age squared, religion, parental education, rural and province dummy variables. 90% confidence interval constructed using 250 boostrap repetitions. Values of V on the x-axis.

Figure 3: MTE with 90% Confidence Interval – Parametric normal selection model estimates



Note: Parametric MTE estimated using a switching regression model with normally distributed errors.

Appendix

<u>Simulation-based approach for estimating average treatment effects in equations 7 and 12.</u>

Step 1: Estimate MTE(x, v) as described in section 3.

Step 2: For each individual in the sample construct the corresponding *P(Z)* and take *n* draws from $V \sim Unif\,[0,1]$ (recall that we assumed that *V* was independent of *X* and *Z*). Since there are 2608 individuals in the sample this creates a simulated dataset of size $2608^{*}n$ (we use *n*=1000). Evaluate the *MTE(x,v)* for each value of *X* and each value of simulated *V*.

Step 3: In this simulated dataset both *X* and *V* are observed for all 2608^{*}n observations. In addition, we have estimates of *MTE(x,v)* for each of them. Therefore it is trivial to construct the following quantities:

$$ATE = \iint MTE(x,v) f_{X,V}(x,v)\,dxdv$$

$$ATT = \iint MTE(x,v) f_{X,V}(x,v|S=1)\,dxdv = \iint MTE(x,v) f_{X,V}(x,v|P>V)\,dxdv$$

$$ATU = \iint MTE(x,v) f_{X,V}(x,v|S=0)\,dxdv = \iint MTE(x,v) f_{X,V}(x,v|P \leq V)\,dxdv$$

by respectively averaging the *MTE* for everyone in the simulated sample, for those who have *P*>*V*, and for those with *P*≤*V*.

Step 4: There is one parameter that remains to be estimated: the AMTE. The version of the AMTE we use in this paper defines marginal individuals as those for whom:

$$|Z\gamma - Us| < \varepsilon$$

Carneiro, Heckman and Vytlacil (2010) show that this is equivalent to estimating the average return to schooling for those induced to enroll in upper secondary schooling when one of the components of *Z*, say the intercept, changes my a marginal amount. This is exactly what we do in our simulations: we change the intercept of the selection equation marginally and we see which members of our simulated dataset change their schooling decision. Finally, we average the MTE for this group.

Table A1: OLS and IV estimates of the return to a year of schooling

	OLS		First stage		IV	
	Coef	**se**	**Average Marginal Derivative**	**se**	**Coef**	**se**
Years of education	0.096***	*0.005*			0.157***	*0.037*
Age	0.058***	*0.017*	0.027	*0.078*	0.055***	*0.018*
Age Squared	-0.047**	*0.022*	-0.062	*0.098*	-0.042*	*0.022*
Muslim						
Protestant	0.084	*0.082*	2.033	*0.381*	-0.037	*0.118*
Catholic	0.003	*0.152*	2.196	*0.856*	-0.117	*0.149*
Other religions	0.055	*0.121*	0.987	*0.754*	0.002	*0.128*
Father uneducated						
… elementary	0.062	*0.048*	1.759	*0.228*	-0.049	*0.080*
… secondary or higher	0.135**	*0.067*	3.627	*0.312*	-0.083	*0.144*
Mother uneducated						
… elementa	-0.086*	*0.046*	1.000	*0.216*	-0.147**	*0.063*
… secondary or higher	-0.119	*0.078*	3.173	*0.344*	-0.316**	*0.145*
Rural household	0.149***	*0.044*	-1.146	*0.301*	0.234***	*0.073*
Distance to health post in km	-0.020	*0.015*	0.037	*0.084*	-0.015	*0.013*
Location controls			Yes			
Dist to nearest sec school			-0.298***	*0.102*		
Number of observations	2,608				2,608	
Test for joint significance of instruments: F-Stat/p-value			3.62/0.000			
R2	0.260				0.204	

Note: This table reports the coefficients for OLS and 2SLS IV for regression of log of hourly wages on years of schooling controlling for parental education, religion and location. We report average marginal derivatives for the first stage equation. Excluded instruments are distance to secondary school and interactions with parental education, religion, age and distance to health center. Type of location is controlled using province dummies. A dummy variable for missing parental education is included in the regressions but not reported in the table. Standard errors (in parenthesis) are robust to clustering at the community level, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table A2: IV estimates of the return to a year of schooling without distance and X interactions

	IV		First stage	
	coef	**se**	**coef**	**se**
Years of education	0.144***	0.053		
Age	0.056***	0.017	0.036	0.077
Age Squared	-0.043*	0.022	-0.072	0.096
Muslim				
Protestant	-0.011	0.141	2.050***	0.380
Catholic	-0.091	0.164	2.229**	0.906
Other religions	0.014	0.128	0.839	0.778
Father uneducated				
… elementary	-0.025	0.102	1.800***	0.231
… secondary or higher	-0.036	0.198	3.525***	0.316
… education missing	-0.034	0.109	0.353	0.444
Mother uneducated				
… elementary	-0.134*	0.073	0.973***	0.215
… secondary or higher	-0.274	0.185	3.180***	0.331
… education missing	-0.183***	0.063	0.367	0.301
Rural household	0.215**	0.091	-1.144***	0.302
Distance to health post in km	-0.016	0.013	0.007	0.082
W Java				
N Sumatra	0.114	0.088	-0.615	0.500
W Sumatra	0.282**	0.112	-0.704	0.476
S Sumatra	0.137	0.125	0.667	0.476
Lampung	-0.044	0.108	0.149	0.477
Jakarta	-0.077	0.078	0.752*	0.421
C Java	0.051	0.091	-0.937*	0.498
Yogyakarta	-0.303***	0.100	1.128**	0.570
E Java	-0.007	0.066	-0.300	0.411
Bali	-0.197	0.159	1.027	0.946
W Nusa Tenggara	-0.176	0.107	0.715	0.839
S Kalimantan	0.298***	0.114	1.726***	0.540
S Sulawesi	0.032	0.097	0.226	0.702
Dist to nearest sec school			-0.244***	0.072
Number of observations	2,608			
Test for joing significance of instruments: F-stat/p-value			11.34/0.00	
R2	0.206			

Note: This table reports the coefficients for 2SLS IV for regression of log of hourly wages years of schooling, controlling for parental education, religion and location. Excluded instruments are distance to secondary school. Type of location is controlled using province dummies. Dummy variable for missing parental education is included in the regressions but not reported in the table. Reference categories are Muslim, and not educated. Standard errors (in parenthesis) are robust to clustering at the community level, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table A3: IV estimates of the return to a year of schooling without regional dummies

	IV	
	coef	**Se**
Years of education	0.135***	0.034
Age	0.059***	0.018
Age Squared	-0.046**	0.022
Muslim		
Protestant	-0.032	0.100
Catholic	-0.153	0.154
Other religions	-0.109	0.091
Father uneducated		
… elementary	-0.006	0.077
… secondary or higher	-0.004	0.141
… education missing	-0.002	0.107
Mother uneducated		
… elementary	-0.074	0.057
… secondary or higher	-0.190	0.131
… education missing	-0.156***	0.060
Rural household	0.227***	0.072
Distance to health post in km	-0.008	0.014
Number of observations	2,608	
Test for joing significance of instruments: F-stat/p-value	4.08/0.00	
R2	0.22	

Note: This table reports the coefficients for 2SLS IV for regression of log of hourly wages years of schooling, controlling for parental education, religion and location. Excluded instruments are distance to secondary school. Type of location is controlled using province dummies. Dummy variable for missing parental education is included in the regressions but not reported in the table. Reference categories are Muslim, and not educated. Standard errors (in parenthesis) are robust to clustering at the community level, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Table A4: Outcome equation: Partial linear regression estimates

	Coeffients	Standard Errors
Age	0.070*	*0.042*
Age Squared	-0.076	*0.051*
Protestant	-0.022	*0.368*
Catholic	-0.816	*0.634*
Other religions	0.786*	*0.406*
Father with elementary education	0.042	*0.192*
… secondary or higher	0.103	*0.675*
… education missing	0.425	*0.292*
Mother with elementary education	-0.144	*0.156*
… secondary or higher	-1.570*	*0.938*
… education missing	-0.173	*0.170*
Rural household	0.288*	*0.161*
Distance to health post in km	-0.016	*0.030*
N Sumatra	0.333	*0.214*
W Sumatra	0.177	*0.218*
S Sumatra	0.233	*0.309*
Lampung	0.253	*0.294*
Jakarta	-0.248	*0.233*
C Java	0.071	*0.153*
Yogyakarta	-0.127	*0.301*
E Java	-0.071	*0.149*
Bali	-1.022**	*0.478*
W Nusa Tenggara	-0.267	*0.325*
S Kalimantan	0.013	*0.451*
S Sulawesi	-0.434	*0.274*
N Sumatra	-0.550	*0.465*
S Sumatra	-0.134	*0.595*
C Java	-0.197	*0.415*
Yogyakarta	-0.127	*0.602*
E Java	0.326	*0.357*
Bali	1.660*	*0.898*
W Nusa Tenggara	0.192	*0.711*
S Kalimantan	0.367	*0.860*
W Sumatra*P	0.465	*0.535*
Lampung*P	-0.993	*0.839*
Jakarta*P	0.394	*0.452*
S Sulawesi*P	0.979	*0.598*
Age*P	-0.069	*0.097*
Age Squared*P	0.124	*0.121*
Protestant*P	0.130	*0.639*
Catholic*P	1.171	*0.931*
Other religions*P	-1.261*	*0.703*
Father with elementary*P	0.053	*0.605*
Father with secondary/higher*P	0.002	*1.280*
Father education missing*P	-1.322	*0.942*
Mother with elementary*P	0.187	*0.393*
Mother with secondary/higher*P	1.977	*1.433*
Mother education missing*P	0.109	*0.458*
Rural *P	-0.275	*0.362*
Distance to health post*P	0.037	*0.082*
Number of observations	2,608	
R2	0.080	

note: *** p<0.01, ** p<0.05, * p<0.1 The table presents the coefficients on X and P^*X from the Robinson's (1988) double residual semi-parametric regression estimator. The logit estimated pscore (P) enters the equation nonlinearly according to a non-binding function and estimated using a gaussian kernel regression with bandwidth equal to 0.2.

Table A5: Testing for equality of LATEs over different Intervals of MTE

$$\left(\mathbf{H_0}: \mathbf{LATE^j}\left(\mathbf{U_S^{L_j}}, \mathbf{U_S^{H_j}}\right) - \mathbf{LATE^{j+1}}\left(\mathbf{U_S^{L_{j+1}}}, \mathbf{U_S^{H_{j+1}}}\right) = \mathbf{0}\right)$$

Ranges of U_S for $LATE^j$	(0,0.1)	(0.1. 0.2)	(0.2,0.3)	(0.3,0.4)	(0.4,0.5)	(0.5,0.6)	(0.6,0.7)	(0.7,0.8)	(0.8,0.9)
Ranges of U_S for $LATE^{j+1}$	(0.1. 0.2)	(0.2,0.3)	(0.3,0.4)	(0.4,0.5)	(0.5,0.6)	(0.6,0.7)	(0.7,0.8)	(0.8,0.9)	(0.9,1)
Difference in LATEs	-0.078	-0.04	-0.014	-0.012	-0.010	-0.011	-0.012	-0.014	-0.014
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

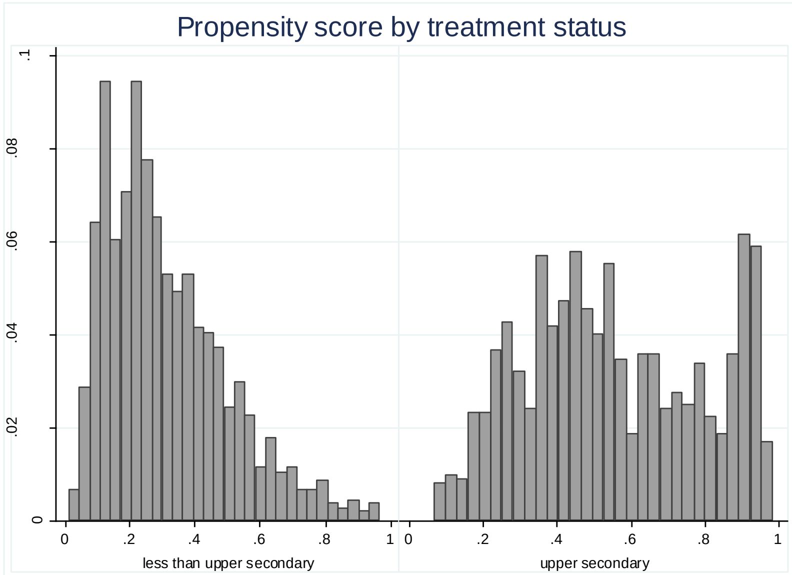
Note: In order to compute the numbers in this table we construct groups of values of *Us* and average the MTE within these groups, where $U_S^{L_j}$ and $U_S^{H_j}$ are the lowest and highest values of *Us* defined for interval *j*. Then we compare the average MTE across adjacent groups and test whether the difference is equal to zero using the bootstrap with 250 replications.

Table A6: Estimates of Average Returns to Upper Secondary Schooling with 95% confidence interval

Parameter	*Non parametric Estimate*	*Normal selection model*
ATT	0.217 (-.1, 0.525)	0.198** (-0.041,0.438)
ATE	0.13 (-0.06, 0.32)	0.065 (-0.099, 0.231)
ATU	0.07 (-0.227, 0.365)	-0.028 (-0.217, 0.160)

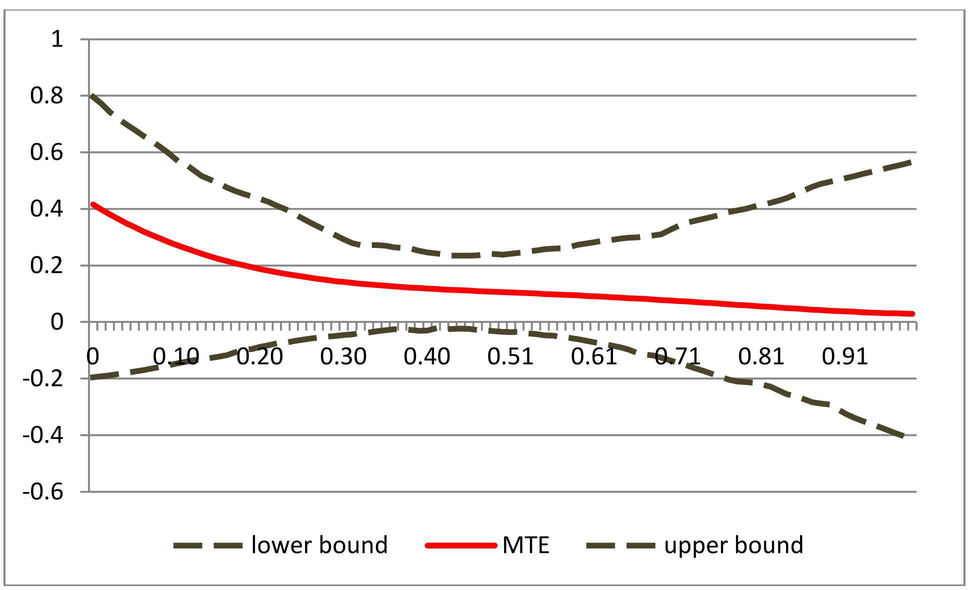
Note: This table presents estimates of various returns to upper secondary school attendance for the semi-parametric and normal selection models: average treatment on the treated (ATT), average treatment effect (ATE), treatment on the untreated (ATU), and marginal policy relevant treatment effect (MPRTE). Returns to upper school are annualized to show returns for each additional year. Bootstrapped 95% confidence interval are reported in parentheses, with significance at *** p<0.001, ** p< 0.05, * p<0.1 indicated.

Figure A1: Propensity score (*P*) support for each schooling group *S* = 0 and *S* = 1



Note: *P* is estimated probability of going to upper secondary school. It is estimated from a logit regression of upper school attendance on Xs, distance to school, interactions of X and distance to school (Table 5).

Figure A2: Marginal treatment effect with 90% Confidence Interval – Semi-parametric regression estimates (without distance and Xs interactions)



Note: To estimate the $E(Y_1\text{-}Y_0|X, U_s)$ function we used a partial linear regression of log wages on X and K(P) ,with a bandwidth of 0.2. X includes age, age squared, religion, parental education, rural and province dummy variables. 90% confidence interval constructed using 250 boostrap repetitions. Values of V on the x-axis.

Figure A3: MTE with 90% Confidence Interval – Parametric normal selection model estimates