Parametric and Semi-parametric Estimations of the Return to Schooling in South Africa*

Sonia Bhalotra† Claudia Sanhueza‡

Abstract

This paper estimates return to schooling for african and coloured women in South Africa. It compares parametric and semiparametric estimates of the sample selection model for the case of return to schooling. The parametric estimator is the one proposed by Heckman (1979) and the semiparametric estimator proposed by Newey (1991) and Klein and Spady (1993). It also attempts to correct endogeneity and mesurement error by using instruments of schooling. Following recent literature, the paper uses community variables primary and secondary school proximity and availability as instruments. Using instrumental variables increases the return to schooling substantially. Parametric corrections does not change the results but semiparametric corrections increases the return even more.

1 Introduction

This paper involves the application of parametric and semiparametric regressions of the sample selection model for the case of the estimation of the returns to schooling of African and coloured women in South Africa.

The reason to adopt a semi-parametric approach, in general, is that the often arbitrary functional form restrictions involved in parametric estimation can result in biases.

In addition, there appear to be no previous attempts to use semiparametric estimation specifically to estimate return to schooling. The

*We want to thank Dr. Simon Appleton, Dr. Geeta Kingdon, Prof. Jonh Knight and the participants at the Conference Understanding Poverty and Growth in Sub-Saharan Africa (CSAE, March 2002, Oxford University). Any mistake or omission is our responsability.

†Bristol University, UK. S.Bhalotra@bristol.ac.uk.

‡Universidad de Chile and Cambridge University. csanhueza@uchile.cl.

closest paper, in spirit, is Martins (2001). She estimates parametric and semiparametric wage and participation equations for married women in Portugal as we do. However, she does not discuss return to schooling and therefore does not consider the econometric issues behind the estimation of these.

Based on human capital theory the returns to schooling are usually estimated from a simple semi-log model of wages. In this model the distribution of labour outcomes (logarithm of wages) are assumed to be explained by human capital accumulated, which is usually measured as years of schooling and experience. Therefore the parameter of the returns to schooling is the one associated to the variable years of schooling.

But simple OLS regression to estimate these returns may be highly biased for basically three econometric problems. Firstly, the variable years of schooling is an endogenous variable. There are several explanatory variables of wages which may be correlated with education that are not observable, as ability for example. The omission of any relevant variable that explain wages and is correlated with education results on biased in the parameter of return to schooling. Secondly, the variable years of schooling is usually measured with error. And thirdly, wages are not observed for everyone. There are people who have decided not to participate in the labour market, whose characteristics are different from the people who are participating. This problem causes what is called a sample selection biased of the parameters in the wage equation and the problem is bigger always for women than for men.

Using a sample of African and coloured women of a fairly complete database for South Africa in this essay we attempt to correct these problems in the following way. We have used instrumental variables to face the problem of endogeneity and measurement error. Following recent literature (Kane and Rouse, 1993; Card, 1993; Conneely and Uusitalo, 1997 and Maluccio, 1997) we have used community variables primary and secondary school proximity as instruments. The sample selection problem bias is typically corrected for by using the Heckman parametric estimator (Heckman, 1979). Under this model a probit estimation of the participation into work equation is estimated in step 1. Using these results the inverse of Mills ratio is constructed and inserted into the wage equation in step 2. This methodology assumes that the error terms in the participation and wage equations are normally jointly distributed. Then we uses a semiparametric procedure which correct the sample selection bias but relaxes the joint normality assumption (Newey, 1991 and Klein and Spady, 1993).

The results indicate that once instrumental variables are used the return to schooling increases from 13% to 57% respect to the least squares estimate. Additionally, a parametric Heckman correction of the sample se-

lection bias give us a parameter on the returns to schooling of 57% and a semiparametric parameter of 60%.

The structure of the paper is the following. Next section 2 presents the econometric model. Section 3 presents the data. Section 4 analyses the results and section 5 are the conclusions.

2 The Econometric Model of Return to Schooling

In human capital framework, education is an investment of current resources in exchange for future returns. Thus, optimal investment decision implies that one would invest in the S^{th} year of schooling if the internal rate of return of that investment is higher than the market interest rate (Becker, 1967). Assuming that the costs of education are zero and the working period is large we can get the familiar functional form of the earnings equation (Mincer, 1974). According to this model the log of individual earnings ($y_i = \log w_i$) is explained by an additive function of a linear education term and a quadratic experience term (the wage equation):

$$\log w_i = \alpha + \beta S_i + \beta_1 E_i + \beta_2 E_i^2 + u_i \tag{1}$$

Where S_i is years of completed education and E_i is years that the individual has worked after completing education[1]. In this case, β is interpreted as the return to schooling[2].

The availability of microeconomic data has allowed the estimation of this equation for many countries in several different studies (Psacharopoulos, 1994). But, despite the overwhelming evidence of a positive correlation between schooling and labour market status, it is difficult to identify, in absence of experimental evidence, whether that correlation is indeed a causal effect of education on earnings.

The empirical application of the simple Mincerian specification involves the following econometric issues which have been much discussed in the literature (Card, 1999).

Endogeneity bias. Unobservables in the wage equation may be correlated with schooling. If this correlation is positive (as it would be if higher ability individuals were likely to acquire more schooling and, given schooling, likely to get paid higher wages) then this will create a positive bias in the return to schooling.

Measurement error bias. If educational level is measured with (random) error then the return to schooling will be biased downward. One might

[1]Experience is usually aproximated by Potential Experience. That is Age-Years of Schooling-6.

[2]This is the private return to schooling. It differs from social returns because the latter one should incorporate the externalities in the benefits of education.

imagine that this is problem is more likely in a developing country where, for example, the fact that school participation is intermittent may introduce some recall bias.

Sample selection bias. Wages are only observed for those in work. If the sample of workers is not a random sample, for example, because workers have higher ability or greater tastes for work than non-workers, then this will create a (positive) bias in the return to schooling.

Endogeneity and measurement error biases can be addressed if suitable instruments for schooling can be found. Sample selection bias is typically corrected for by using the Heckman estimator. In the following sections we will discuss all of them.

2.1 Instrumental Variables

If the unobservable factors can be measured and held constant in regression of equation 1, the endogeneity problem would be eliminated. However, in practice, neither economic theory nor the real word provides all the set of variables that should be held constant.

A standard solution to the problem of casual inference and also a solution to the problem of omitted variables[3] is instrumental variables. If we have a set of variables Z_i, which are correlated with schooling, but otherwise unrelated to earnings. That is, Z_i is uncorrelated with the omitted variables and the regression error in equation 1.

The econometric model is as follows. We have two equations. One is the main wage equation described before and the other is the schooling equation:

$$\log w_i = \alpha + \beta S_i + + X_i'\gamma + u_i \tag{2}$$

$$S_i = c + \Psi Z_i + X_i'\gamma + \xi_i \tag{3}$$

The procedure to estimate this model is 2SLS. In the first step equation Then an instrumental variable estimate of the returns to schooling is the sample analogue of $cov(y_i, Z_i)/cov(S_i, Z_i)$. And if also $cov(Z_i, measurement\ error) = 0$ then an unbiased and consistent instrumental variable estimator of the return to schooling can be found.

Then, the instrumental variable method allows us to consistently estimate the coefficient of interest free from bias from omitted variables, without actually having the data on the omitted variables. In addition, if

[3]Other solution to the omitted variable problem is randomly assigned the variable of interest. For example, social experiments are sometimes used to assign people to a social program. Random assignment assures that participation in the program is not correlated with omitted personal or social factors. But this is a very unlikely possibility in education.

the instruments are not measured with error they would also correct the measurement error bias.

A good instrument is correlated with the endogenous regressor (years of schooling) for the reasons that can be verified or explained, but uncorrelated with the outcome variables (wages) for reasons beyond its effect on the endogenous regressor.

2.1.1 Controling for Unobservable

Some studies that have been trying to control for unobservable factors are the following. Griliches and Mason (1972) studied a 1964 sample of US military veterans. They used several OLS specifications including three different measures of schooling (the total grades of school completed, grades of school completed before military service and the increment in years of schooling), measures of ability from an armed forces qualification test, personal-background variables such as father's schooling and occupational sector, place where they grew up, age and race. They also included some variables that measured "current location and success" such as current location zone, dummy for married and length of time in current job. However, they did not have measure of school quality. The coefficient of all schooling variables drop of around 15 percent when ability and personal background variables were included. The coefficient of return to schooling was around 3 percent-5 percent.

Blackburn and Neumark (1993) studied the increase in returns to schooling in the US over the 1980s. They used conventional wage regression in which they included years of education, years of high school and years of college, experience, age, union status, dummy for married and urban residence, they included ability variables such as academic, technical, computational and non academic test scores as other variables in the OLS regressions and they used them as instruments of schooling. They also have family background variables to as instruments of schooling. The coefficient of years of education laid between 2 and 5 percent. They do not corrected the selection bias.

More recently, Glewwe (1996) estimated the returns to schooling using Ghanaian data. Glewwe had data on ability, school quality, dummy for gender, dummy for government job and current residence location. The measure of schooling was again years of completed education. He also corrected the estimates of selectivity bias using the Heckman two-step and full maximum likelihood estimation. The coefficient of returns to schooling was around 7 to 9 percent in OLS regressions, but once he corrected for selectivity they drop down to *zero*. He concluded that once the variable of school quality was included in the OLS estimation the schooling variable, measured as years of completed education, can overestimate the true re-

turn to education (if sample selection is ignored). But once he controlled for selectivity, years of schooling understates the rate of return to education. Hence, in countries were school quality varies widely across time and space, years of schooling may be a very imperfect indicator of education attainment and simple estimates of the return to schooling may be severely biased. Second, he also found that when data on cognitive skills and a measure of innate ability are used to assess the impact of education on wages, it appears that it is cognitive skills acquired, rather than years of schooling per se or innate ability, what determines wages in the private sector in Ghana. This is very important point in less developed countries where differences in school quality are dramatic.

2.1.2 Using Instruments

In the recent literature many instruments have been taken from institutional source of variation in schooling, such as minimum school leaving age, tuition costs or geographic proximity of schools[4].

Angrist and Krueger (1991) landmark study used quarter of birth interacted with year of birth in US 1970 and 1980. They show that men born earlier in the year have slightly less schooling than men born later in the year. Assuming that quarter of birth is independent of ability and other unobserved components. This phenomenon generates exogenous variation in education that can be used in an IV estimation scheme. Similar instruments were used by Staiger and Stock (1997).

Kane and Rouse (1993), Card (1993), Conneely and Uusitalo (1997) and Maluccio (1997) used geographic college proximity or similar variables as instruments. *The idea is that accessibility matters. Then it is more likely that individual chose higher level of education if there is colleges nearby.* Card found that accessibility is more important for individual on the margin of continuing their education. Since college proximity was found to have a bigger effect for children of less educated parents he used interactions of family background families with college proximity. Harmon and Walker (1995) used changes in compulsory school leaving age in England and Wales in 1947 and 1974 as instruments of education. Butcher and Case (1994) used a measure of sibling sex composition. They showed that sibling sex composition has an effect on women's schooling but does not appear to have effect on other relevant economic outcomes. Therefore it is useful as an instrument in estimating returns to education.

It is worth noting that all these studies have found that the IV estimator of the return to schooling is as big or bigger than the corresponding OLS

[4]It is worth noting that the literature using instrumental variables for schooling decisions has been mainly focused in developed countries. A fairly complete revision of the econometric issues and the literature is in Card (1999).

estimator. That means OLS estimates are understating the true value of the return to schooling. This difference can in principle be attributed to the non observation of true educational attainments (measurement error). But the differences between IV and OLS estimates are too large to be explicable by measurement error alone.

The most recent literature has suggested mainly two different interpretation of the results obtained with IV techniques. Card (1999) has shown that from a simple model of endogenous schooling choice, the return to education is not a single parameter in the population, but a random variable that may vary with other individual characteristics, such as family background, ability or level of schooling. *In other words, the differences between OLS and IV estimates of β might be caused by heterogeneity in returns to schooling.* The model implies that individuals choose an optimal level of schooling in a point where the marginal return equal marginal cost which is assumed to be the discount rate. Card (1999) suggested that the instrument is probably influencing the educational decision of individuals with high marginal returns and hence high discount rates. High discount rates are generally present in more disadvantage families due to liquidity constrains. *Then, if IV relies on "interventions" that affect the schooling choices of children from relatively disadvantaged family backgrounds (high discount rates) then their marginal return to schooling will be higher that the average return to schooling of the population as a whole.* Hence the IV estimator of β will be higher than the OLS estimator. Therefore, Card (1999) showed that β must be interpreted as the *average return to schooling* in the population and that not only OLS but also IV techniques can bias estimates of this parameter.

The second interpretation is based on the evaluation of "treatment effects" (Heckman, 1997). In this case, the treatment is defined as acquisition of additional education and the outcome as earnings. However, when treatment (education) effects are distinctive among people and participation into treatment is not random, the estimation of the effect of the treatment on a random person (the return to education in equation 1) is almost impossible. Heckman (1997) proposed, instead, the estimation of the "effect of the treatment on the treated" (the return to schooling for those who decided to acquire education) and he also showed that OLS and IV techniques require very restrictive assumption in order to estimate the return to schooling. Additionally, Angrist, Imbens and Rubin (1996) showed that the only treatment effect that IV can consistently estimate is the *Local Average Treatment Effect* (LATE) that is the average treatment effect (average return to education) for those who change treatment-status (educational choice) because they act in accordance with the assignment-to-treatment-mechanism (instruments). For example, IV estimates of the

return to schooling based on college proximity as an instrument should be interpreted as the average return to schooling for a person that acquires an additional year of education only because is close to college but would drop out of school if no college had been nearby. One consequence of this interpretation is that different instruments should estimate different returns to schooling associated with different subgroups in the population.

There have been some empirical evidence that effectively returns to schooling are heterogeneous. For example, Dearden (1993) using data on the UK argues that two group of individuals with same years of schooling but different "qualifications" have different return to education and therefore there is evidence of heterogeneity in the return to years of schooling. Ichino and Winter-Ebmer (1998) using data for Germany show that different instruments for education, father highly educated and father in war, result in different returns to schooling. They argued that these instruments affect different groups of the population and therefore the average return to schooling changes. However, Harmon and Walker (1999) found no evidence that different instrument affect different decision margins and hence no evidence of heterogeneity in the UK.

2.1.3 Using Panel Data of Twins

An alternative to the instrumental variables approach to the problem of endogeneity is to study education and earnings outcomes from a panel of twins. The key idea is that some of the unobserved differences that bias a cross-sectional relation between education and earnings are reduced or eliminated within families (fixed effect). So the parameter of the return to schooling is the within estimator for a panel of twins. A survey of the literature on twins can be found in Griliches (1979) and more recently Card (1999). Card found that the new studies contrast to the earlier ones in two features. First, the samples now are larger and include a broader range of age and family background. And second most of them fairly address the problem of measurement error. They almost always found that the within estimator of β is lower than the correspondent OLS.

Following recent innovations in the literature we use as instrument for schooling the community level variables, availabity and distance to primary and secondary school.

2.2 The Parametric Model of Sample Selection

This section in based on the work of Heckman (1979).

We only observe wages (or y_i) for people who are actually working. If the people we left out of the regressions have different characteristics from those in the sample then our estimates would be biased.

The typical sample selection model has the following form:

$$y_i^* = \alpha + \beta S_i + + X_i'\gamma + u_i \tag{4}$$
$$y_i = \log w_i = y_i^* P_i \tag{5}$$

where y is the log wage, S is schooling, X is a vector of exogenous control variables and P is an indicator variable for whether the individual participates in the work force and earns a wage. The participation (selection) condition is defined as:

$$P_i = 1 \quad if \quad P_i^* = W_i'\theta + \varepsilon_i > 0 \tag{6}$$
$$P_i = 0 \quad otherwise. \tag{7}$$

where W is a vector of exogenous variables. Identification may rely on functional form (the inverse Mills ratio is a non-linear function of ε_i but is assisted if W has at least one variable that is not contained in X. If the residuals u_i and ε_i are correlated the regression of equation (4) gives inconsistent estimates of α, β and γ.

The classical parametric correction of this problem is the Heckman estimator. In the first step a probit for work participation (6) is estimated:

$$\Pr(P = 1|W) = E(P|W) = \Phi(W'\theta/\sigma) \tag{8}$$

where Φ is the cumulative normal distribution and σ is the standard deviation of ε. Estimates of (5) are used to construct the inverse Mills ratio, λ. In the second step, the estimated λ is inserted into the wage equation (4) as an additional regressor. The probit model assumes normality of the residual. Even more, the maximun likelihood Heckman estimator assumes that (u_i,ε_i) have a bivariate normal distribution.

2.3 The Semiparametric Model of Sample Selection

The semi-parametric model is based on Newey (1991). It generalises the Heckman parametric procedure.

In the first step, we used a semiparametric estimator proposed by Klein and Spady (1993) to estimate the participation equation (6), which is consistent, asymptotically normal distributed and achieves the semiparametric efficiency bound.

$$\Pr(P = 1|W) = E(P|W) = G(W'\theta) \tag{9}$$

where G is an unknown continuous function. The estimation is semiparametric in the sense that it does not assume any distributional form on the disturbances. But it assumes that the choice probability function depends on the parametrically specified index function G.

To estimate G we use the method described by Klein and Spady (1993). It is assumed that G_{np}, the non-parametric estimator of G, is the following kernel regression:

$$G_{np}(\varphi_i) = \frac{\sum_j W_j K\left(\varphi_i - \varphi_j\right)/h_{np}}{\sum_j K\left(\varphi_i - \varphi_j\right)/h_{np}} \tag{10}$$

where $\varphi = W'\theta$ and θ is the vector of parameters to be estimated semi-parametrically. And the following two conditions are attained: the bandwidth h_{np} is non-stochastic and satisfies (i) $n^{-1/6} < h_{np} < n^{-1/8}$ (where n = sample size) and the kernel function have to be bias reducing (ii) $\int \varphi^2 K(\varphi)\, d\varphi = 0$.

The semi-parametric estimator θ_{sp} is obtained by maximising:

$$\log L_{np}(\theta) = n^{-1} \sum_{i=1} \{P_i \log G_{np}(W_i'\theta) + (1 - P_i) \log[1 - G_{np}(W_i'\theta)]\} \tag{11}$$

Assuming a normal distribution for ε independent of W reduces expression (to the probit specification. The parametric log-likelihood function is then obtained by replacing $G_{np}(W'\theta)$ with $\Phi(W'\theta/\sigma)$.

In the second step, a non-parametric selectivity-correction term is constructed using the first step estimates. Call this $\mu(W'\theta_{sp})$,this is the non-parametric analogue of the inverse Mills ratio, λ. The function μ is unknown since the distribution of the errors in (6) is unspecified.

Following Newey (1991), we approximate μ with the polynomial:

$$\mu(W'\theta_{sp}) = \sum_{k=1} \chi_k (W'\theta_{sp})^{k-1} \tag{12}$$

where K is increasing in sample size n. The larger the number of basis functions, k, the smoother is the relation, making K rather like the bandwidth parameter, h_{np}, used in the kernel estimation.

The wage equation is then:

$$y_i = \alpha + S_i'\beta + X_i'\gamma + \mu(W'\theta_{sp}) + u_i \tag{13}$$

which is estimated by ordinary least squares. Newey (1991) presents a consistent estimator of the asymptotic covariance matrix.

3 The Returns to Schooling in South Africa: Background and Data

3.1 Literature Review

This study uses a survey carried out just before the end of the apartheid government in South Africa. During that time residential and schooling

choices of African families were severely limited. They were segregated by law to live far from the cities centre where white families lived closed to their schools and jobs. In addition, funding decisions for most African schools were made centrally by White-controlled entities on which they had no representation. The way all the society worked marked great educational and labour outcomes disparities between different racial groups.

Case and Deaton (1999) presented a very detailed research on the educational system in South Africa at that time. Using the same survey, they studied the relationship between educational inputs and outcomes in South Africa. They showed that the disparities in educational outcomes were so important that they resulted in strong and significant effects on educational achievement, specially for African people.

Moll (1998) using the same survey investigate returns to education. However, instead of using years of schooling in equation ??, linear splines for primary, secondary and tertiary schooling were defined. He runs OLS, Least Absolute Deviations (LAD), Huber's M-estimator and Least Trimmed Squares (LTS) of equation ?? for African workers adding as explanatory variables urban dummy, female dummy, community dummies and test scores. The latter taken from the Literacy Assessment Module of the survey. He had a total of 133 observations. The OLS results are shown in table 1. Primary school had a return of about 3% but it was not significant. Secondary school had a return of 10% and tertiary education a return of 60%. These results contrast strongly with the ones presented by Psacharopoulus (1994) for Sub-Saharan Africa in which primary school had the largest return to schooling, followed by secondary and tertiary (see Table bellow).

Some Estimates of the Return to Schooling in South Africa (%)

	Primary	Secondary	Higher
Moll (1998)[a]: African workers	2.9	9.7*	60*
Psacharopoulos (1994)[b]: Sub-Saharan Africa	41.3	26.6	27.8
Appleton (2000)[a]: Sub-Saharan Africa	5	14	37
Mwabu and Schultz (2000)[a]:			
African Men	6.6*	13.5*	26.9*
African Women	5.4*	21.8*	39.4*
Coloured Men	3.0	17.2*	15.7*
Coloured Women	2.6	16.2*	31.5*
Indian Men	-6.7	21.1*	20.4*
Indian Women	-6.9	12.4*	29.5*
White Men	-0.6	8.4*	14.4*
White Women	-3.9*	6.2*	13.9*

[a]: Mincerian wage equations. [b]:Full Method. *:Significant at 5% level.

Also Appleton (2000) presented quite different estimates from that of

Psacharopoulos (1994) for Sub-Saharan Africa countries. He reported a survey of 28 studies from sub-Saharan African from 1980 onwards that produced a mean return to education of 5 percent for primary schooling, 14 percent for secondary schooling and 37 percent for tertiary education. Appleton pointed out that, firstly, the rates of return reported by Psacharopoulos are mainly "full" estimates[5] rather than the Mincerian returns. Second, the estimates reported by Psacharopoulos may be raised -especially primary education- by the inclusion of extremely high returns generated from studies with very poor data. Third, Psacharopoulos' estimates were mainly taken from studies in the 1960s and 1970s when education was more scarce in Africa and economic conditions lighter.

Mwabu and Schultz (2000) studied wage premiums (private returns) to primary, secondary and higher education in South Africa. Using the same survey, they run OLS regressions without taking to account the problem of endogeneity in schooling. They run the regressions across groups of South African workers aged 16 to 65 distinguished by race, sex and region. In general, the returns for tertiary education are higher than for secondary and primary education, and higher for women than for men. The group with the highest returns are for African women (see Table above).

3.2 The Data

The main source of data used in this study is the Project of Living Standards and Development (PLSD, South Africa, 1993).

The survey PSLD was carried out in South Africa in 1993 by the Southern Africa Labour and Development Unit (SALDRU) and The World Bank. The survey was carried out for 8,848 households and a total number of 43,974 individuals distributed among 360 clusters or communities.

The main part of the survey is a comprehensive Household Questionnaire (HHQ). The topics covered, among others, included demography, household services, household expenditures, educational status and expenditure, employment and income.

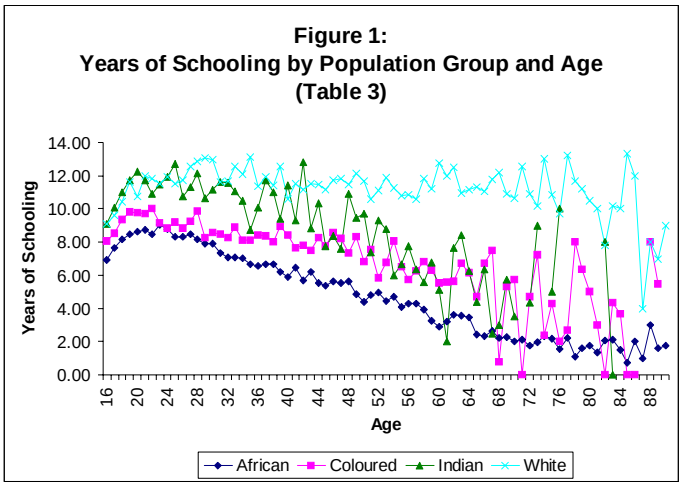
A Community Questionnaire (COMMQ) was also administered through the "communities" in the survey. The purpose was to collect information about the provision of facilities in the communities, including education, health and recreational facilities.

A third questionnaire was a Literacy Assessment Module (LAM). The aim of the LAM was to test proficiency in English and mother tongue to both reading comprehension and numerate. It was carried out over 2 members of one fifth of the households in each cluster, which is a total of 2,381 individuals.

[5]These include pecuniary costs and assume the opportunity cost of primary schooling is less than the adult uneducated wage.

The survey included 80% African people, 8% coloured, 3% Indian and 9% white people. 52% of them were women. 34% of them were children (younger than or up to 14). We take a sub-sample of 11,001 African and coloured women older than 15. We also carried out the analysis for african and coloured women who had taken the LAM in the survey. There were 919 of them.

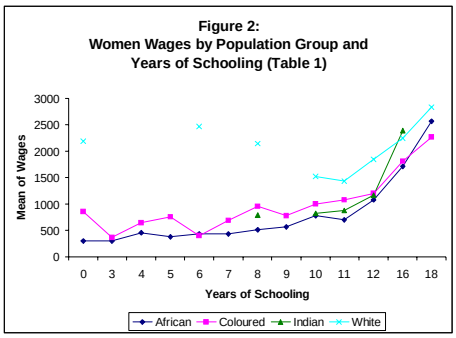
Figure 1[6] shows educational attainment by race and age (16 to 90). Younger people in South Africa show higher educational attainment than older for any population group, especially young Indians. Although it is white people who in average have more education.



Figures 2 and 3 show average net monthly wage for women and men older than 15 by population group and educational attainment. This wage correspond to people who have regular employment only[7]. I do not show the plot when in the subgroup there are 10 or less observations. Wages for white people are almost always higher, expect for Indian women with higher education who have the highest wage amongst women with more than 16 years of education. Generally, in all racial group people with more education have also higher wages. But there is an strange hump around the 7 years of schooling, which corresponds to finishing primary school, for white people. The reasons for than may be that white people with only primary school can still get better jobs because their relative privilege position under the apartheid.

[6] All the Figures are based on Tables that are reported in the Appendix.

[7] There are some people who apart from their regular employment have also some casual jobs. Therefore their wages are the sum of both regular income and casual income.



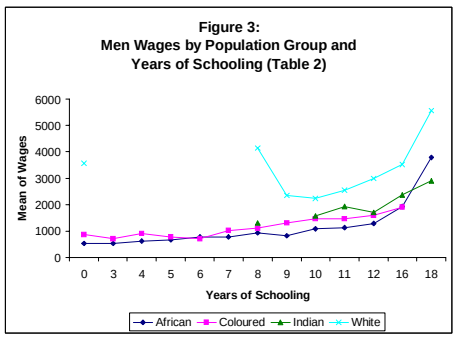
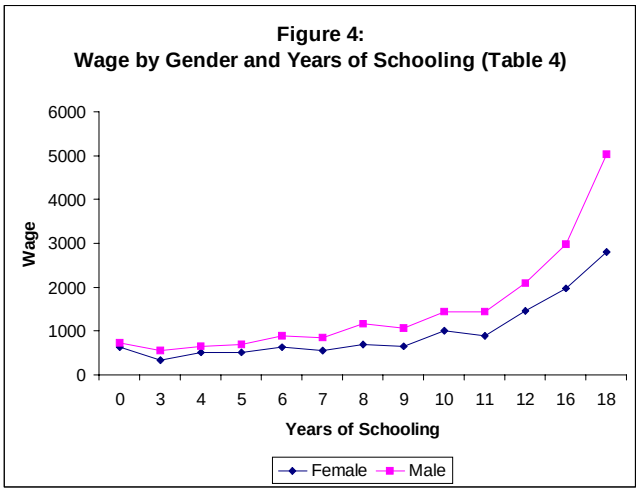
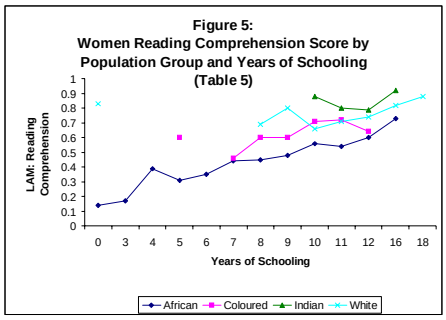


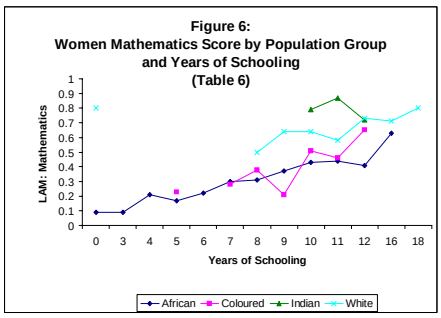
Figure 4 shows wages by gender and years of schooling. Men have higher wages at all level of schooling. The gender gap is increasing in years of schooling.



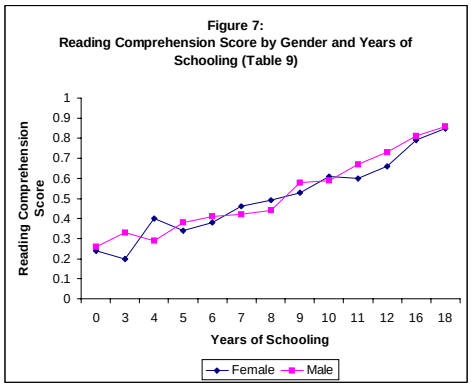
Figures 5 and 6 reports scores of women on the Reading Comprehension and Mathematics test in the LAM by population group and years of schooling. The reading comprehension test consisted in two paragraph in which there were 3 questions corresponding to each of them. They could have answered right, wrong or omitted each question. Similarly in the test of practical mathematical problems there were 2 sections, computational exercises and practical mathematical problems[8]. The score was calculated as the proportion of right questions answered. I just plotted the group with 5 or more observations. The figures show that in reading comprehension and mathematics indian women have better results. African women can improve their scores if they attain more years of schooling.

[8]The questions taken are shown in the Appendix.





Figures 7 and 8 shows the scores by gender and years of schooling. Both test, reading comprehension and mathematics, show that women and men do similarly for each year of schooling. However, men did generally better amongst people with higher level of education.



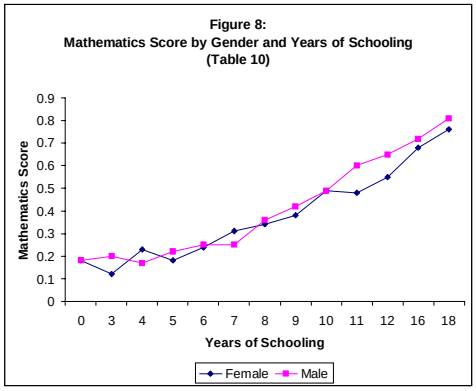


Table 7 reports the set of variables of the sample used in the econometric analysis of this work. Recalling that this sample correponds to african and coloured women aged 15 or more. Table 8 in the Appendix includes people who has the LAM. The Tables show mean, standard deviation, minimum and maximum values and the meaning of the variables.

The sample with no LAM has 11,001 observations. Only 2,178 or 20% african and coloured women are working in a regular employment. They get an in average 709 ZAR per month (45 UKP). The average age of the sample is 35. 16% of them are head of the household. 88% of them are married. The household size has in average 7 members. From which approximately 3 are children. 5% of the children are aged 2 or less, 11% are aged 3 to 7 and 16% are aged 8 to 14. There are families from 360 different clusters or communities. And the average unemployment rate per community is 68%. The average proportion of casual workers per community is 15%. There are 1749 african women whose husband who got a wage from regular employment, with a mean wage of 1049 ZAR (67 UKP). To run the regressions is necessary to replace non husband wage by a very small number to get some logarithm from it.

Table 7: Descriptive Statistics Sample African and Coloured Women older than 15

Variable	Obs	Mean	Std. Dev.	Min	Max	Meaning
clustnum	11001	210.42	101.39	1.00	360.00	community number
wage	2178	709.18	648.22	2.00	4616.67	Monthly Net Wage Income plus Cash Paid, . If non wage
age	11001	35.21	16.57	15.00	110.00	Age in years
lwage	2178	6.13	1.00	0.69	8.44	log(wage)
age2	11001	1514.16	1459.60	225.00	12100.00	age^2
head	11001	0.16	0.36	0.00	1.00	1 if head, 0 if not
married	11001	0.88	0.33	0.00	1.00	1 if married, 0 if not
hhsize	11001	7.35	3.92	1.00	31.00	number of household members
lhhsize	11001	1.84	0.58	0.00	3.43	log(hhsize)
children	11001	2.58	2.13	0.00	18.00	number of people less than 14 yo in the household
pk02	11001	0.05	0.08	0.00	0.50	prop. of children less that 2 yo
pk37	11001	0.11	0.12	0.00	0.67	prop. of children between 3 and 7 yo
pk814	11001	0.16	0.14	0.00	0.80	prop. of children between 8 and 14 yo
pcas	11001	0.15	0.16	0.00	1.00	prop. of casual workers in the community
punemp	11001	0.68	0.18	0.05	0.97	prop. of unemployment in the community
educa	11001	6.81	4.10	0.00	18.00	years of schooling
swage	11001	166.73	522.46	0.00	14620.00	Spouse Wage, 10^(-21) if non wage
lswage	11001	-39.62	20.10	-48.35	9.59	log(swage)
dps	11001	0.80	3.15	0.01	55.00	distance to primary school
ldps	11001	-3.24	2.25	-4.61	4.01	log(dps)
dss	11001	3.00	9.01	0.01	120.00	distance to secondary school
ldss	11001	-2.07	2.94	-4.61	4.79	log(dss)
dummyps	11001	0.74	0.44	0.00	1.00	1 if there is a primary school in the community, 0 if not
dummyss	11001	0.55	0.50	0.00	1.00	1 if there is a secondary school in the community, 0 if not
pexp	11001	22.39	18.87	-5.00	104.00	potential experience=age-educa-6
pexp2	11001	857.57	1240.71	0.00	10816.00	pexp^2
pexpchd	11001	56.03	76.50	-12.00	990.00	pexp*children
pexpchd2	11001	2094.19	4296.67	0.00	55112.00	pexp2*children
durban	11001	0.45	0.50	0.00	1.00	1 if community is urban or periurban, 0 if rural
coloured	11001	0.10	0.30	0.00	1.00	1 coloured population group
oswage	1749	1048.71	890.09	1.00	14620.00	Original Spouse Wage, . if non wage
e	11001	0.00	3.03	-9.62	14.00	Residuals Step 1 IV
part	11001	0.20	0.40	0.00	1.00	participation binary variable: 1 if work, 0 if not

The average level of years of schooling attained are 6.8. The data on years of schooling was not taken directly from the survey but was constructed based on the educational question. This latter is shown in the Appendix.

Regarding to the community question on geographical proximity and availability of primary and secondary schools, 74% of our sample had a primary school in the community. And only 55% of them had a secondary school. We include in the regressions the logarithm of the distance to a primary and secondary schools. The actual question was formulated as follows:

> Q1. Is there a primary school in this community? If No go to Q3
>
> Q2. If yes, how many primary schools are there in the community?
>
> Q3. If no, how far away is the nearest primary school to this community?...Km.

Then in communities where there was at least a primary school the distance to this one was not shown in the survey. We fix this problem letting be this distance 0.01 Km, given that the nearest primary school out

of the community was 0.05 km away. Thus, the average distance to primary school (dps) is 0.8 km and the average distance to a secondary school (dss) is 3 km.

4 The Returns to Schooling in South Africa: Econometric Analysis

We turn now to report the results of the parametric and semiparametric estimations described in section 2.

We proceed first to estimate equation 2 by OLS for women who participate in the labour market. The vector of exogenous variables in the wage equation includes: pexp (potential experience), pexp2 (potential experience squared), pexpchd (interaction of potential experience and number of children), pexpchd2 (interaction of potential experience squared and number of children), pcas (proportion of casual workers by community), punemp (unemployment rate by community) and coloured (dummy for coloured population group).

Column 1 in Table 11A and 11B reports these results. The returns to schooling are 0.134 for the whole sample and drops to 0.071 if we included as regressors the test scores in mathematics and reading comprehension from the LAM. Although these test scores are not statistically significant they do have and effect on the return of years of schooling. These two measures are measure of human capital, they are a measure of the acquisition of cognitive skills. Therefore this drop in the returns to schooling has to be interpreted as if in some sense they matters to the distribution of wages. Not only the number of years in school but also what people have learnt during those years.

Recalling that this OLS regression presents the following econometric problems. Any omitted variable which explain wages and is correlated with years of schooling will result in bias of the returns to schooling. For example, we do not have a measure of ability in our survey, or a measure of the quality of the schools people went to. And these two variables may clearly explain wages. People with more ability and went to better schools are able to get better jobs and have higher wages. But also people with higher ability and in better schools may decide to spend more years at school. This double correlation causes a bias on β, being the OLS parameter higher than it should be.

Also, in this estimation we do not take into account that we have left out of the relevant sample 80% women because they either have a casual job or have decided not to work. If the sample we left out have different characteristics, or the sample we have taken is not random sample of the population, the OLS estimates of equation 2 are biased. This biased can be positive or negative.

4.1 Instrumental Variables

As we said before an standard solution to the problem of omitted variables is instrumental variables. We have chosen then a set of variables Z_i, which are correlated with schooling, but otherwise unrelated to earnings. That is, Z_i is uncorrelated with the omitted variables and the regression error in equation 2. If also $cov(Z_i, measurement\ error) = 0$ then we have found an instrumental variable estimator of the returns to schooling which is free from endogeneity and measurement error bias.

Following the recent literature we have chosen as instruments availability and geographic proximity of primary and secondary schools. The idea behind this is that accessibility matters. Then it is more likely that individual chose higher levels of education if there is school nearby. In particular, the instruments chosen were the following four:

- Dummy variable which indicates whether there is a primary school in the community.
- Dummy variable which indicates whether there is a secondary school in the community.
- Logarithm of the distance to primary school.
- Logarithm of the distance to secondary school.

Column 2 in Table 11A and 11B reports the IV results for the sample of women who participates in the labour market. We can see that the parameter of return to schooling has increased from 0.134 to 0.572. In the case of the sample with the LAM it has increased to 0.27.

Why IV is higher than OLS? Although this result coincide with the literature it is still valid to ask us what is happening in our particular case. One possible explanation is that the return to schooling is not a single parameter in the population but there are heterogenous return to schooling. In that case OLS estimator is the average return to schooling in the population. And IV is just estimating the return to schooling of those individuals whose education choices are affected really by our instruments. In our case, we are estimating the return to schooling of african and coloured women in South Africa. This is the poorest population group and therefore it is possible that our IV is really affecting their educational decison. In the sence that it is more likely that availability and geographical proximity of primary and secondary schools affect more the educational decisions of those poor individuals than of the richest populatiuon group in South Africa.

Statistical Validity of the Instruments. The first stage equation of the two-stage least squares procedure shows a F test of 463.68. And the F

test for the significance of the four instruments is 3.43 (p value is 0.0093). But t tests show that only two of the instruments are significant. The two related to secondary school. One possible cause for this is that availability of secondary is relatively more scarce than primary schools. Recalling that 74% of our sample have a primary school in the community but only 55% of them a secondary school. However, for identification purpose it is still useful leave all the instruments[9]. It was laso tested whether the instruments were exogenous by a test of everidentifying restrictions. The J-statistics is 0.87. We cannot reject with 99% of confidence the nule hyphothesis of exogeneity of the instruments.

In the next sections we estimate the sample selection model using the parametric and the semiparametric methods. All estimations are also including instruments to correct endogeneity bias.

4.2 Parametric Results

So we are now assuming that there is an underlying participation equation which model the decision to participate in the labour market and therefore have a positive wage in the wage equation.

We run two different regressions. We first assume that there is no correlation between these two equations (rho=0) and therefore is the same to regress OLS of wage equation and probit for the participation equation separately. And then we run a Heckman maximum likelihood regression which assumes this correlation is not 0.

The participation equation includes as explanatory variables: ldps-ldss-dummyps-dummyss (instruments of years of schooling), age, age2 (age-squared), married (dummy for married women), lhhsize (logarithm of the household size), pk02 (proportion of children aged between 0 and 2), pk37 (proportion of children aged between 3 and 7), pk814 (proportion of childen between 8 and 14), head (dummy if the woman is head of the household), pcas, durban (dummy for urban community), punemp, coloured, lswage (logarithm of monthly husband wage).

The most important explanatory variables in the participation equation are age, age2, married pk02, p714, head, punemp and coloured. Surprisingly husband's wage is not significant. And married affects positively the probability of being observed working.

We can see that the parameter on return to schooling does not change significantly between the two parametric approaches. And indeed the same happens to the rest of the parameters.

We check whether there is or not sample selectivity bias. To do that we

[9] An estimation with only the two instruments related to secondary school gives us a return to schooling of 51%. With an F test of 4.83 for the two instruments in the first stage equation.

test the null hypothesis of no correlation (rho=0) between the disturbance in the wage equation and the disturbance in the participation equation, given the alternative hypothesis that there is correlation. Given the Heckman ML results, Wald test of independent equations (rho = 0): chi2(1) = 3.18 and Prob > chi2 = 0.07, the hypothesis can be rejected at a 5% level of significance. For the sample with the LAM the hypothesis can not be rejected, Wald test of indep. eqns. (rho = 0): chi2(1) = 0.42 and Prob > chi2 = 0.52. However, we know that Heckman ML procedure relies heavily on the assumption that both disturbance terms from the wage equation and the participation equation are jointly normal distributed.

4.3 Semiparametric Results

We are estimating now the sample selection model under the semiparametric model. As we know under this specification it is not necessary to assume that the disturbances in the participation and wage equation are jointly normal distributed.

In the first step, the participation equation was estimated following Klein and Spady. Under the latter we need to attain the conditions (i) and (ii) described in section 2.3. That is the bandwidth h_{np} is non-stochastic and satisfy (i) $n^{-1/6} < h_{np} < n^{-1/8}$ (where n is sample size) and the kernel function have to be bias reducing (ii) $\int_{\mathbb{R}} \varphi^2 K(\varphi)\, d\varphi = 0$.

To satisfy these conditions h_{np} has to lay in the interval [0.21;0.31] for the whole sample and in [0.32;0.42] for the sample with the LAM. And G_{np} was calculated using fourth-order kernel $K(\varphi) = \varphi(\mu)(1.5 - 0.5\mu^2)$.

The results are shown in Tables 13A and 13B. It is also shown the results for the Probit Model for effects of comparison. To identify the model we are assuming that the coefficient of the variable age is equal to 1.

We notice that there are some

In the second step, we used the semiparametric above results $Z'\theta_{sp}$ for $h_{np} =$ and $h_{np} = 0.37$ for the whole sample and the sample with the LAM. For the whole sample $\mu(Z'\theta_{sp})$ has $K = 4$ basis functions and for the sample with the LAM $\mu(Z'\theta_{sp})$ has $K = 2$. And this is added as explanatory "function" in the wage equation.

Results are shown in Tables 14A and 14B. We can see that ... for the subsample which has got LAM the significance of the varaibles has not changed. The specific coefficient correspondening to the return to schooling has increased from the parametric procedure from 0.259 to 0.355. So the introduction of the variables of the LAM, mathematics and reading scores, implies finally that wages of african and coloured women in South Africa are only explained by their population group.

5 Conclusions

References

[1] Angrist, J., Imbens, G. and Rubin, D. (1996). "Identification of causal effects using instrumental variables", Journal of American Statistical Association 91: 444-472.

[2] Angrist J. and Krueger A. (2001) "Instrumental Variables and the Search for Identification: from Supply and Demand to Natural Experiments", Working Paper 8456, National Bureau of Economic Research.

[3] Angrist, J. and Krueger, A. (1991). "Does compulsory school attendance affect schooling and earnings", Journal of Business and Economics Statistic 13: 225-235.

[4] Appleton, S. (2000). "Education and Health at the Household Level in Sub-Saharan Africa", Center for International Development Working Paper No. 33, Harvard University.

[5] Becker (1967). Human capital and the personal distribution of income (University of Michigan Press, Ann Arbor, MI).

[6] Blackburn, M. and Neumark, D. (1993). "Omitted-Ability Bias and the Increase in the Return to Schooling", Journal of Labor Economics, Vol. 11, No. 3: 521-544.

[7] Butcher, K. and Case, A. (1994). "The effect of sibling composition on women's education and earnings", Quarterly Journal of Economics 109: 531-563.

[8] Case, A. and Deaton, A. (1999). "School Inputs and Educational Outcomes in South Africa", Quarterly Journal of Economics, 1047-1084.

[9] Card, D. (1993). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling", Working Paper No. 4483, National Bureau of Economic Research.

[10] Card, D. (1999). "The Causal Effect of Education on Earnings", in Ashenfelter, O. and Card, D. (eds), Handbook of Labor Economics, Vol 3A, North Holland, Amsterdam.

[11] Conneely, K. and Uusitalo, R. (1997). "Estimating Heterogeneous treatment effects in the Becker schooling model", Discussion Paper, Industrial relations Section, Princeton University.

[12] Dearden L. (1993). "Qualifications and Earnings in Britain: How reliable are conventional OLS estimates of the returns to education?, Institute of Fiscal Studies, London.

[13] Glewwe, P. (1996). "The relevance of standard estimates of rates of return to schooling for education policy: A critical assessment", Journal of Development Economics, Vol 51: 267-290.

[14] Griliches, Z. and Mason, (1972). "Education, Income, and Ability",

Journal of Political Economy, Vol 80: S74-S101.
[15] Griliches, Z. (1979). "Sibling models and data in economics: beginnings of a survey", Journal of Political Economy 87: S37-S65.
[16] Harmon, C. and Walker, I. (1995). "Estimates of the Economic Return to Schooling for the United Kingdom", American Economic Review 85: 1278-1286.
[17] Harmon, C. and Walker, I. (1999). "The marginal and average return to schooling in the UK", European Economic Review 43: 879-887.
[18] Heckman, J. (1979). "Sample Selection Model as Specification Error". Econometrica, 47:153-161.
[19] Heckman, J. (1997). "Instrumental Variables: A study of Implicit Behavioral Assumptions Used in Making Program Valuations", Journal of Human Resources, Vol 32, No. 3: 441-462.
[20] Ichino, A. and Winter-Ebmer, R. (1998). "Lower and Upper Bounds of Returns to Schooling", Unpublished Working Paper, CEPR.
[21] Kane, T. and Rouse, C. (1993). "Labor market return to two and four year colleges: is a credit a credit and do degrees matter?", Working Paper No. 4268, National Bureau of Economic Research.
[22] Klein R.L. and Spady, R.H. (1993). "An Efficient Semiparametric Estimator for Binary Response Models". Econometrica, 61: 387-421.
[23] Maluccio, J. (1997). "Endogeneity of Schooling in the Wage Function", Unpublished Manuscript, Yale University, Department of Economics.
[24] Martins, M.F.O. (2001). "Parametric and Semiparametric Estimation of Sample Selection Models: An empirical Application to the Female Labour Force in Portugal", Journal of Applied Econometrics, 16: 23-39.
[25] Moll, P. (1998). "Primary Schooling, Coginitive Skills and Wages in South Africa", Economica, 65: 263-284.
[26] Mwabu, G. and Schultz, T.P. (2000). "Wage Premiums for Education and Location of South African Workers, by Gender and Race", Economic Development and Cultural Change, 307-334.
[27] Newey W.K. (1991). "Two Step Estimation of Sample Selection Models. MIT Working Paper.
[28] Newey, W. Powell, J. and Walker, J. (1990). "Semiparametric Estimation of Selection Models: Some Empirical Results", American Economic Review Papers and Proceedings, Vol. 80, No. 2: 324-328.
[29] Mincer, J. (1974). "Schooling Experience and Earnings (Columbia University Press, New York).
[30] Psacharopoulos, G. (1994). "Returns to investment in education: a global update", World Development 22: 1325-1343.
[31] Staiger, D. and Stock, J. (1997). "Instrumental Variables Regression with Weak Instruments", Econometrica 65: 557-586.

6 Appendix

6.1 Literature Assessment Module: Reading Comprehension

Please read the following passages. Then, answer the questions that follow the passage. (15 minutes)

Passage 1:

"When Mbaya was a child, he got very excited when his mother, Corfu, asked if he would like to go to the meat market with her. As they walked into the centre of town, the wonderful odours of meat - both fresh and spoiled - could be smelled up to one kilometre away. The hundreds of market stalls formed a row of almost 1 and $\frac{1}{2}$ kilometres long. It took almost one hour to walk slowly from one end of the meat market to the other.

"Sometimes Corks would let Mbaya choose what meat they would buy that morning. The smell of fresh beef was Mbaya's favorite. But sometimes Mbaya would accidentally choose the beef that was not fresh. Corfu would go up dose to the big piece of meat hanging from the rack and smell it. Once she was close to it, Corks could tell immediately that the beef was not fresh. Then, she would laugh at Mbaya and tease him for picking spoiled meat. But the meat seller would be angry, as Corks let on to other shoppers that his beef was not fresh. Mbaya would then start looking around for beef that seemed more fresh, no longer trusting that his nose is the best instrument for finding fresh meat."

Q1. How long was the row of meat stalls, from one end to the other end? (Tick one ..)

a. 1 and 1/2 kilometers long.

b. 1 kilometre long.

c. It was very dose from one end to the other end.

d. Hundreds of stalls were lined up.

Q2. What did Mbaya most like to do inside the meat market?(Tick one..)

a. Try, to find spoiled meat.

b. Walk from one end to the other end of the market.

c. Have his mother tease him when he found spoiled meat.

d. Find fresh beef for his mother to buy.

Q3. What did not happen when Mbaya would choose a spoiled piece of meat? (Tick one..)

a. The meat seller would get angry.
b. Mbaya and his mother would leave the market.
c. Corfu would tease Mbaya.
d. They would keep shopping for fresh meat.

Passage 2:
"Zenariah was riding to work in her usual combi. The driver and the woman sifting next to him, named Roseline, were arguing over whether it was any use for the woman's son to stay in school. The son, named Philemon, was 16. His secondary school had been closed for many days over the past 6 months. Teachers often did not show-up for work. But the woman felt that if he could pass matric, Philemon could eventually find a good job, perhaps as a clerk or office worker. The driver, however, claimed that even university graduates were having difficulty finding jobs as clerks. Zenariah had graduated from the University of the Western Cape, and it had taken her 3 months to find her job as an assistant accountant. She was sympathetic to the woman's position, but also had to agree that until the economy improved, education would not guarantee a good job."

Q4. When Zenariah goes to work in the morning.. (Tick one.)
a. She usually takes the same combi.
b. She always sits next to Roseline.
c. She tries to find different drivers and combis.
d. She usually talks about her son, Phileon, in the combi.

Q5. What kind of job did Philemon's mother hope he would find? Tick one)
a. Assistant accountant
b. Conti driver
c. Office worker
d. Teacher

Q6. What was the combi driver's position? (Tick one..)
a. Schools are of high quality.
b. Philemon should go to University of the Western Cape.
c. Completing school will lead to a good job.
d. Schooling does not guarantee a good job.

6.2 Literature Assessment Module: Mathematics

- Computational Problems. (15 minutes)

Please solve the following maths problems.

Q1. 103 kg - 37 kg = ...kg

Q2. R35.50 x 7 = R...

Q3. 25% of R225 = R...

Q4. R22.25 - R7.88 = R...

- Practical Mathematical Problems. (10 minutes)

Q5. "According to the doctor, the mother must buy 0,30 litres of cough mixture for her two sick children. She can either buy three bottles, each containing 0,10 litres. for R 9.50 per bottle or she can buy four bottles, each containing 0.08 litres, for R 7.00 per bottle. What is the least amount of Rand she needs to spend to get the 0,30 litres required by the doctor?
Answer...

Q6. "Namane was trying to figure-out her transport costs from the township to the city to get to her job. The combi cost R 2.00 each day. If she took a combi, then a taxi for part of the trip, she would have to spend R 3.50 each day. How much more would the taxi plus the combi cost for the week, than just taking a taxi, if she went to work five days during the week?'
Answer...

6.3 Years of Schooling

Q: What is the highest educational qualification attained by ...?

00=None $\Longrightarrow$ 0 years of schooling.
01=Sub A - Std 1 (Class 1/Grade 1 - Std 1) $\Longrightarrow$ 3 years of schooling.
02=Std 2$\Longrightarrow$ 4 years of schooling.
03=Std 3$\Longrightarrow$ 5 years of schooling.
04=Std 4$\Longrightarrow$ 6 years of schooling.
05=Std 5$\Longrightarrow$ 7 years of schooling. (End Primary School)
06=Std 6 (Form 1)$\Longrightarrow$ 8 years of schooling.
07=Std 7 (Form 2)$\Longrightarrow$ 9 years of schooling.
08=Std 8 (Form 3/Junior Certificate)$\Longrightarrow$ 10 years of schooling.
09=Std 9 (Form 4)$\Longrightarrow$11 years of schooling.
10=Std 10 (Matric/Form 5/Senior Certificate)$\Longrightarrow$ 12 years of schooling. (End Secondary School)

11=Std 7,8, or 9+diploma$\Longrightarrow$ 12 years of schooling.

12=Std 10 + teacher training$\Longrightarrow$ 16 years of schooling.

13=Std 10+Nursing$\Longrightarrow$ 16 years of schooling.

14=Std 10 + diploma at techmnikon or other technical institution$\Longrightarrow$ 16 years of schooling.

15=Std 10 + some university courses$\Longrightarrow$ 16 years of schooling.

16=Completed university degree$\Longrightarrow$ 18 years of schooling.

17=Creche/daycare$\Longrightarrow$ 0 years of schooling.

18=Pre-primary$\Longrightarrow$ 0 years of schooling.

19=Other (Specify)$\Longrightarrow$ missing value for years of schooling.

7 Tables

Table 1: Women Wages by Population Group and Years of Schooling

Years of schooling	Population Group											
	African			Coloured			Indian			White		
	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.
0	297.83	311.96	300	857.28	783.89	19	1360.1	619.15	5	2189.75	2396.45	60
3	302.36	348.16	107	363.94	308.22	15	1947.83		1	883.33	824.96	2
4	456.94	600.3	113	648.26	696.76	15		0	0	2108	1230.36	3
5	375.64	387.23	110	761.05	728.45	25	908.06	115.68	3	2459.38	1189.18	4
6	431.94	405.48	130	403.46	412.45	33	1576.67	1305.79	2	2472.1	1610.24	17
7	432.55	349.81	223	686.59	523.8	43	1921.07	1717.43	6	2415.1	1089.36	7
8	516.55	355.64	261	952.15	659.18	54	793.98	232.89	10	2145.03	1558.56	20
9	570.36	405.67	128	781.11	436.18	49	1014.47	428.11	5	1141	694.38	2
10	773.26	437.29	222	1003.92	556.8	72	819.63	216.53	16	1522.5	949.2	111
11	695.36	513.11	88	1081.59	703.33	34	878.85	385.65	10	1433.7	516.92	19
12	1072.32	683.45	195	1201.05	607.24	63	1169.27	606.79	49	1842.53	1032.63	270
16	1711.15	652.6	155	1810.09	921.69	20	2384.98	1024.33	13	2241.2	1184.06	145
18	2571.82	882.59	11	2265.49	1589.91	3	2925.79	1930.6	6	2836.54	1434.43	75
Total	641.49	613.32	2,043	923.03	687.22	445	1324.77	954.56	126	2020.22	1329.85	735

Table 2: Men Wages by Population Group and Years of Schooling

Years of schooling	Population Group											
	African			Coloured			Indian			White		
0	523.76	449.78	511	854.72	766.56	24	3099.32	1997.78	7	3554.33	2401.5	31
3	535.43	438.25	191	717.45	588.93	15		0	0	1630	1065.37	2
4	621.59	490.38	195	906.15	652.99	12	908.56	125.28	3	1333.3	1044.98	3
5	674.26	419.79	202	782.16	434.04	20	2580		1		0	0
6	778.06	534.22	257	713.17	511.94	26	847.1		1	4685.21	1412	9
7	769.94	533.06	346	1027.45	739.84	45	869.27	607.69	5	6747.22	3404.86	3
8	920.58	649.85	352	1110.34	725.2	97	1298.23	794.11	21	4137.58	4156.18	31
9	822.99	467.55	179	1304.34	1809.21	64	1872.41	1080.67	10	2349.9	1596.47	15
10	1074.91	775.44	275	1460.13	846.21	90	1576.37	795.23	36	2243.14	946.15	120
11	1120.59	650.84	147	1466.04	751.86	40	1932.26	1324.24	17	2543.53	2207.81	34
12	1294.12	772.79	314	1592.99	1049.7	77	1713.03	895.06	79	2993.9	1921.52	348
16	1929.73	839.35	98	1907.32	977.4	18	2370.7	1341.87	21	3522.29	2919.95	253
18	3786.72	2226.85	18	2418.35	1369	5	2893.13	1002.8	21	5553.76	8518.11	153
Total	871.64	719.36	3,085	1262.45	1028.04	533	1863.28	1118.56	222	3474.81	4065.88	1,002

Table 4: Wages by Gender and Years of Schooling

years of schooling	Gender					
	Female			Male		
	Mean	SD	Obs	Mean	SD	Obs
0	636.66	1215.54	382	733.86	1044.81	571
3	332.21	382.98	125	559.08	467.59	208
4	517.55	673.01	130	650.49	515.35	212
5	513.44	610.69	142	688.74	441.06	219
6	629.92	867.19	182	891.93	887.48	292
7	552.82	604.24	275	845.17	800.03	399
8	687.18	674.63	345	1173.58	1432.98	500
9	644.75	431.31	184	1062.56	1123.05	268
10	1012.01	700.78	421	1446.34	953.18	520
11	887.38	608.7	151	1439.9	1185.98	238
12	1455.02	927.18	577	2085.84	1613.82	818
16	1975.57	982.92	332	2985.56	2525.02	390
18	2793.49	1404.26	95	5029.1	7606.57	197
Total	1008.13	1013.5	3,341	1500.29	2236.4	4,832

Table 5: Women Reading Comprehension Score by Population Group and Years of Schooling

Years of	Population Group											
schooling	African			Coloured			Indian			White		
	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.
0	0.14	0.26	70	0.5	0.58	4	0.75	0.12	2	0.83	0.17	9
3	0.17	0.22	31	0.58	0.59	2		0	0		0	0
4	0.39	0.29	41	0.33		1		0	0	0.83		1
5	0.31	0.27	42	0.6	0.22	5		0	0		0	0
6	0.35	0.23	52	0.42	0.17	4	0.17		1	0.89	0.19	3
7	0.44	0.24	80	0.46	0.29	9	0.58	0.12	2	1	0	2
8	0.45	0.24	122	0.6	0.25	24	0.54	0.21	4	0.69	0.34	8
9	0.48	0.22	102	0.6	0.26	18	0.71	0.21	4	0.8	0.16	11
10	0.56	0.21	124	0.71	0.21	22	0.88	0.13	7	0.66	0.23	37
11	0.54	0.23	77	0.72	0.19	9	0.8	0.27	5	0.71	0.26	19
12	0.6	0.22	76	0.64	0.24	12	0.79	0.19	13	0.74	0.23	40
16	0.73	0.21	18	0.92	0.1	4	0.92	0.12	2	0.82	0.12	14
18	0.67	0	2		0	0		0	0	0.88	0.14	10
Total	0.44	0.27	837	0.62	0.27	114	0.75	0.22	40	0.75	0.22	154

Table 6: Women Mathematics Score by Population Group and Years of Schooling

Years of	Population Group											
schooling	African			Coloured			Indian			White		
	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.
0	0.09	0.2	70	0.25	0.4	4	0.33	0.47	2	0.8	0.26	9
3	0.09	0.22	31	0.58	0.59	2		0	0		0	0
4	0.21	0.26	41	0.67		1		0	0	0.67		1
5	0.17	0.24	42	0.23	0.15	5		0	0		0	0
6	0.22	0.22	52	0.13	0.25	4	0.5		1	0.72	0.19	3
7	0.3	0.26	80	0.28	0.24	9	0.17	0.24	2	0.67	0.24	2
8	0.31	0.26	122	0.38	0.3	24	0.54	0.28	4	0.5	0.32	8
9	0.37	0.25	102	0.21	0.23	18	0.71	0.16	4	0.64	0.29	11
10	0.43	0.25	124	0.51	0.29	22	0.79	0.19	7	0.64	0.29	37
11	0.44	0.25	77	0.46	0.27	9	0.87	0.14	5	0.58	0.24	19
12	0.41	0.24	76	0.65	0.29	12	0.72	0.25	13	0.73	0.3	40
16	0.63	0.16	18	0.67	0.14	4	0.92	0.12	2	0.71	0.21	14
18	0.58	0.12	2		0	0		0	0	0.8	0.19	10
Total	0.32	0.27	837	0.4	0.31	114	0.69	0.27	40	0.68	0.27	154

Table 8: Descriptive Statistics Sample African and Coloured Women older than 15 with LAM.

Variable	Obs	Mean	Std. Dev.	Min	Max	Meaning
clustnum	919	201.50	101.76	2.00	360.00	community number
wage	146	805.58	666.07	50.00	3081.67	Monthly Net Wage Income plus Cash Paid, . If non wage
age	919	27.28	11.96	15.00	78.00	Age in years
lwage	146	6.30	0.97	3.91	8.03	log(wage)
age2	919	887.21	842.45	225.00	6084.00	age^2
head	919	0.10	0.30	0.00	1.00	1 if head, 0 if not
married	919	0.95	0.23	0.00	1.00	1 if married, 0 if not
hhsize	919	7.36	3.16	1.00	25.00	number of household members
lhhsize	919	1.91	0.43	0.00	3.22	log(hhsize)
children	919	2.71	1.82	0.00	14.00	number of people less than 14 yo in the household
pk02	919	0.05	0.08	0.00	0.40	prop. of children less that 2 yo
pk37	919	0.11	0.11	0.00	0.60	prop. of children between 3 and 7 yo
pk814	919	0.19	0.14	0.00	0.80	prop. of children between 8 and 14 yo
pcas	919	0.15	0.16	0.00	1.00	prop. of casual workers in the community
punemp	919	0.66	0.18	0.15	0.97	prop. of unemployment in the community
educa	919	7.99	3.40	0.00	18.00	years of schooling
swage	919	128.65	412.71	0.00	3433.33	Spouse Wage, 10^(-21) if non wage
lswage	919	-41.18	18.53	-48.35	8.14	log(swage)
dps	919	0.91	2.48	0.01	25.00	distance to primary school
ldps	919	-2.81	2.41	-4.61	3.22	log(dps)
dss	919	2.59	5.95	0.01	60.00	distance to secondary school
ldss	919	-1.77	2.86	-4.61	4.09	log(dss)
dummyps	919	0.64	0.48	0.00	1.00	1 if there is a primary school in the community, 0 if not
dummyss	919	0.47	0.50	0.00	1.00	1 if there is a secondary school in the community, 0 if not
pexp	919	13.30	13.40	-4.00	72.00	potential experience=age-educa-6
pexp2	919	356.18	637.66	0.00	5184.00	pexp^2
pexpchd	919	36.97	49.26	-8.00	371.00	pexp*children
pexpchd2	919	976.49	2168.72	0.00	22445.00	pexp2*children
durban	919	0.52	0.50	0.00	1.00	1 if community is urban or periurban, 0 if rural
coloured	919	0.13	0.33	0.00	1.00	1 coloured population group
math	919	0.33	0.28	0.00	1.00	Mathematics Test Score
read	919	0.46	0.28	0.00	1.00	Reading Comprehension Test Score
oswage	120	985.26	680.35	100.00	3433.33	Original Spouse Wage, . if non wage
e1	919	0.00	2.51	-10.82	8.78	Residuals Step 1 IV
parta	919	0.16	0.37	0.00	1.00	participation binary variable: 1 if work, 0 if not

Table 9: Reading Comprehension Score by Gender and Years of Schooling

	Gender					
years of	Female			Male		
schooling	Mean	SD	Obs	Mean	SD	Obs
0	0.24	0.35	85	0.26	0.33	44
3	0.2	0.26	33	0.33	0.28	23
4	0.4	0.29	43	0.29	0.24	30
5	0.34	0.28	47	0.38	0.27	24
6	0.38	0.25	60	0.41	0.27	42
7	0.46	0.25	93	0.42	0.23	44
8	0.49	0.26	158	0.44	0.24	83
9	0.53	0.24	135	0.58	0.25	78
10	0.61	0.22	190	0.59	0.25	103
11	0.6	0.25	110	0.67	0.24	67
12	0.66	0.23	141	0.73	0.22	95
16	0.79	0.18	38	0.81	0.21	44
18	0.85	0.15	12	0.86	0.17	16
Total	0.51	0.29	1,145	0.55	0.3	693

Table 10: Mathematics Score by Gender and Years of Schooling

	Gender					
years of	Female			Male		
schooling	Mean	SD	Obs	Mean	SD	Obs
0	0.18	0.31	85	0.18	0.28	44
3	0.12	0.26	33	0.2	0.26	23
4	0.23	0.27	43	0.17	0.23	30
5	0.18	0.23	47	0.22	0.24	24
6	0.24	0.25	60	0.25	0.25	42
7	0.31	0.26	93	0.25	0.23	44
8	0.34	0.28	158	0.36	0.27	83
9	0.38	0.27	135	0.42	0.26	78
10	0.49	0.28	190	0.49	0.28	103
11	0.48	0.26	110	0.6	0.25	67
12	0.55	0.3	141	0.65	0.26	95
16	0.68	0.18	38	0.72	0.24	44
18	0.76	0.19	12	0.81	0.26	16
Total	0.39	0.31	1,145	0.44	0.32	693

Table 11A
Simple OLS and Instrumental Variables Estimates

Wage Equation		
	OLS	**IV**
educa	0.134	0.572
	(0.006)**	(0.089)**
pexp	0.034	0.102
	(0.007)**	(0.015)**
pexp2	-0.000	-0.001
	(0.000)**	(0.000)**
pexpchd	-0.002	0.003
	(0.001)	(0.002)
pexpchd2	0.000	-0.000
	(0.000)	(0.000)
pcas	-0.498	-0.299
	(0.243)*	(0.255)
durban	0.440	-0.231
	(0.071)**	(0.158)
punemp	0.612	0.224
	(0.182)**	(0.194)
coloured	0.287	0.068
	(0.081)**	(0.097)
e		-0.442
		(0.088)**
Constant	4.060	0.124
	(0.157)**	(0.787)
Observations	2178	2178
R-squared	0.41	0.43
Robust standard errors in parentheses		
* significant at 5%; ** significant at 1%		
Instruments		
ldps		0.009
		(0.063)
ldss		-0.294
		(0.089)**
dummyps		0.199
		(0.370)
dummyss		-1.879
		(0.571)**

Table 11B (Sample with LAM)
Simple OLS and Instrumental Variables Estimates

Wage Equation		
	OLS	**IV**
educa	0.071	0.270
	(0.025)**	(0.523)
math	0.587	0.127
	(0.307)	(1.213)
read	0.180	-0.365
	(0.291)	(1.468)
pexp	0.039	0.047
	(0.023)	(0.032)
pexp2	-0.001	-0.001
	(0.001)	(0.001)
pexpchd	-0.001	0.003
	(0.005)	(0.010)
pexpchd2	0.000	0.000
	(0.000)	(0.000)
pcas	-0.657	-0.399
	(0.873)	(1.173)
durban	0.600	0.354
	(0.158)**	(0.644)
punemp	0.730	0.466
	(0.499)	(0.781)
coloured	0.507	0.578
	(0.141)**	(0.239)*
e1		-0.200
		(0.530)
Constant	4.102	2.920
	(0.338)**	(3.194)
Observations	146	146
R-squared	0.47	0.47
Robust standard errors in parentheses		
* significant at 5%; ** significant at 1%		
Instruments		
ldps		-0.030
		(0.071)
ldss		-0.029
		(0.118)
dummyps		-0.264
		(0.445)
dummyss		0.066
		(0.731)