UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

WORKING PAPERS IN ECONOMICS

No 638

Education and HIV incidence among young women: causation or selection?

Dick Durevall, Annika Lindskog, and Gavin George

November 2015

ISSN 1403-2473 (print)
ISSN 1403-2465 (online)

Department of Economics
School of Business, Economics and Law at University of Gothenburg
Vasagatan 1, PO Box 640, SE 405 30 Göteborg, Sweden
+46 31 786 0000, +46 31 786 1326 (fax)
www.handels.gu.se info@handels.gu.se

Education and HIV incidence among young women: causation or selection?*

Dick Durevall[+], Annika Lindskog[+], and Gavin George[++]

November 2015

Abstract

Several studies report that schooling protects against HIV infection in Sub-Saharan Africa. This study examines the effect of secondary school attendance on the probability of HIV incidence among young women aged 15-24, using panel data from rural KwaZulu-Natal in South Africa. Three approaches are used to distinguish causation from selection: instrumentation to identify the causal effect, a fixed effects model to control for constant unobserved factors and assessments of the bias from selection on unobserved variables. Although there is a strong negative association between secondary school attendance and HIV incidence, we are not able to find support for a causal effect. Thus, there is no evidence that interventions that increase secondary school attendance in KwaZulu-Natal would mechanically reduce HIV risk for young women. Our focus on school attendance, in contrast to studies that analyze school attainment, might explain the negative finding.

Key words: HIV/AIDS, Education, Schooling, South Africa

JEL: I12, I29, O12

[+] Department of Economics, School of Business, Economics and Law, University of Gothenburg, P.O. Box 640, SE 405 30, Gothenburg, Sweden, email: dick.durevall@economics.gu.se, annika.lindskog@economics.gu.se
[++] HEARD, University of KwaZulu-Natal, Westville Campus, Private Bag X54001, Durban 4000 South Africa, email: georgeg@ukzn.ac.za

*The authors gratefully acknowledge the Africa Centre for Population Health for providing us with data; Ralitza Dobreva for invaluable research assistance; Till Bärnighausen, Frank Tanser, and participants at seminars at HEARD, University of Gothenburg, and The Nordic Conference in Development Economics 2015, for helpful comments and suggestions. We would also like to thank the Swedish Research Council for financial assistance.

1. Introduction

In spite of recent achievements in the fight against HIV/AIDS, the disease remains a serious challenge in heavily affected sub-Saharan African countries. In 2014, there were roughly 1.4 million new HIV infections and 790 000 AIDS deaths, and as many as 26 million people, or one in 20 adults, were HIV positive (UNAIDS, 2015a). Recent achievements, such as increased access to antiretroviral treatment and reduced mother to child transmission of the virus, have decreased death and infection rates significantly, but new infections continue to outnumber death rates. In South Africa, the HIV prevalence is increasing by over 100,000 people per year, and in 2013 there were 6.8 million who were HIV infected, amounting to 18% of the world's HIV positive people (UNAIDS, 2015b).

Research has emphasized the key role of economic, social, political and environmental (i.e., structural) factors in explaining the spread of HIV and has discussed the need to address these factors (Gupta et al., 2008; Seely et al., 2012; Johnston et al., 2015). Using panel data from KwaZulu-Natal to analyze the role of schooling, Bärnighausen et al. (2007) conclude that increasing educational attainment in the general population is the most promising way to reduce HIV incidence in South Africa. Several studies have reached similar conclusions, affirming the protective effect of schooling (De Walque; 2007; Pettifor et al., 2012; Santelli et al., 2013; Hardee et al., 2014; Behrman, 2015; De Neve et al., 2015).

There are many potential reasons posited for the protective effect of schooling,[1] but there are also good reasons to suspect that self-selection into education matters. Educational and health investments are likely to be driven by similar factors, such as time preferences and ability (Fuchs, 1982), as highlighted in many studies that attempt to identify a causal impact of education on health outcomes other than HIV infection.[2] Still, there is a paucity of evidence on the impact of schooling on HIV infection. There are three previous attempts to estimate a causal effect: Alsan and Cutler (2013), Behrman (2015) and De Neve et al. (2015). All three studies find a protective effect of education, but Alsan and Cutler (2013) only analyze virginity because they lack information on HIV incidence and HIV prevalence.

[1] These are discussed in Section 2.
[2] Some examples are Berger and Leigh (1989), Lleras-Muney (2005) and Clark and Royer (2013). Amin et al. (2015) is a recent study that attributes the correlation between education and health to selection effects.

This study uses panel data from rural KwaZulu-Natal in South Africa to examine whether secondary school attendance has a causal effect on the probability of HIV incidence among young women aged 15-24. The data is population-based and includes annual HIV incidence for 2005-2012, so we can determine when infection occurred. We focus exclusively on current school attendance; many studies do not distinguish between school attendance and school attainment, and the impact of current attendance might differ from the effect of attainment among those who have already left school. Behrman (2015) and De Neve et al. (2015), on the other hand, study school attainment.

After having established a negative correlation between school attendance and HIV infection, three approaches are used to investigate whether this correlation is due to an underlying causal protective effect of school attendance or due to self-selection or other unobserved factors. Similar to Alsan and Cutler (2013), we use distance to secondary schools as an instrument of school attendance, both in a bivariate probit model and in a linearized instrumental variable model. Moreover, we estimate individual fixed effects models and apply approaches developed by Altonji et al. (2005) and Oster (2015) to evaluate the sensitivity of the estimated effect of school attendance on HIV incidence to unobserved factors.

Our results suggest that unobserved differences across the girls who attend school and those who drop out explain most of the association. Thus, we do not find support for the idea that policies that increase secondary school attendance automatically reduce HIV infections among young women. However, it is important to acknowledge that school environments, curriculums and social contexts differ, so the effect of current school attendance on HIV incidence is likely to be heterogeneous. It is also possible that long-run effects of school attainment, after having left school, are different than short-run effects of school attainment.

The following section elaborates on potential mechanisms and gives a brief description of the earlier studies on schooling and HIV. Section 3 describes the empirical strategy, Section 4 describes the study area and the data, and Section 5 presents the empirical analysis. Section 6 discusses the results and concludes the paper.

2. Education and HIV: Mechanisms and empirical evidence

There are several reasons why education might reduce HIV infection rates.[3] Some are more relevant for school attendance and others for school attainment or both. One explanation is that education increases human capital, and thereby raises the opportunity cost of infection and shortened life expectancy, which in turn leads to less risky sexual behaviors (Fortson, 2008; Oster, 2012). Education also makes people more able to adopt protective behaviors, and some studies focus on the ability to process and acquire new information and change behavior in accordance with that new information (De Walque, 2007; Hargreaves et al., 2008a). The fact that HIV prevention campaigns are often school-based should strengthen this effect. If a protective effect is due to increased opportunity costs or ability to process information, we should expect to see an effect while the individual is still investing in human capital, i.e. attending school, but also one after she has left school, since human capital will remain higher. The effect might, however, vary over time.

In a study using South African data, Hargreaves et al. (2007) argue that current school attendance reduces HIV risk because of its influence on social and sexual networks. However, the impact of school attendance on such networks is likely to vary depending both on the school and on the activities young women pursue if they do not attend school. A related hypothesis is that school attendance may have an incarceration effect, i.e., it reduces the time available for risky sex (Black et al., 2008; Alsan and Cutler, 2013).

As documented by Hardee et al. (2014), at least 15 studies find a negative association between education and the risk of HIV. Yet, during the early stages of the epidemic, HIV spread faster among the wealthy and well-educated (Jukes et al., 2008; Case and Paxson, 2013). This was most likely the combined effect of differences in lifestyle and a general lack of knowledge about HIV. Over time, the pattern of infections has changed in most countries, and several studies show that nowadays prevalence rates are lower among well-educated adults compared to others (de Walque, 2007; Hargreaves et al., 2008b; Hargreaves et al., 2010; Gummerson, 2013).

Previous studies on education and HIV risk in South Africa mostly use cross-sectional data. Pettifor et al. (2008) use nationally representative data on women aged 15-24 who report having

[3] See Jukes et al. (2008) for a review.

had only one life-time partner. They find that women who had not completed secondary school were three times as likely to be HIV infected, even when controlling for a range of mediating variables (including age of sexual debut, condom use, vaginal discharge, transactional sex, and pregnancy). Peltser et al. (2012) use survey data of persons aged 18-24 years from four South African provinces, and find a reduced HIV risk for those who have completed secondary school (grade 12 or more), but no difference between those with and without completed primary school (grade 7). Hargreaves et al. (2007) focus on school attendance, using survey data for persons aged 14-25 years from a rural area in Limpopo, South Africa. That study finds a negative association between attendance and HIV risk for men. For women, Hargreaves et al. (2007) do not find a significant association between school attendance and HIV infection, but there is a significant negative association between school attendance and risky sexual behaviors, such as earlier sexual debut, number of partners, and the age difference between partners. The study concludes that HIV risk is reduced because attending school affects sexual networks.

As far as we know, only two studies use panel data from sub-Saharan Africa: Bärnighausen et al. (2007) and Santelli et al. (2013). Bärnighausen et al. (2007) use two rounds of data from a rural area in KwaZulu-Natal – from 2003/2004 and 2005 – to determine whether socioeconomic factors affect HIV incidence for women aged 15-49 and men aged 15-54. The main finding is that one extra year of education reduces the risk of acquiring HIV by 7%. Santelli et al. (2013) analyze sexually experienced Ugandan youth (15–24 years-old) enrolled in the Rakai Community Cohort Study, 1999–2008. They find that school attendance reduces female HIV infection sharply, with an adjusted Incidence Risk Ratio of 0.22. Although these studies control for reverse causation and a number of observables that are correlated with HIV, they do not handle selection effects, i.e., that unobserved factors determine both schooling and HIV risk.

Alsan and Cutler (2013), Behrman (2015) and De Neve et al. (2015) are key studies because they employ strategies to identify a causal effect. Alsan and Cutler (2013) use distance to the nearest secondary school as an instrument. Because their data, the Uganda 1995 Demographic and Health Survey, does not include HIV testing, they use the age of sexual debut instead of HIV infection. The main finding is that education delays sexual debut. Using model simulations that link HIV to sexual debut, they suggest that the expansion of girls' secondary school enrollment in Uganda

accounted for between one-sixth and one-half of the decline in HIV prevalence between 1988 and 1995.

Behrman (2015) analyzes how the introduction of universal primary education policies in Malawi and Uganda in the mid-nineties affected HIV infection among adult women. She finds that the marginal effect of one additional year of primary schooling decreases the probability that a woman is HIV positive by 0.06 in Malawi and 0.03 in Uganda.

De Neve et al. (2015) exploit a change in the grade structure in Botswana in 1996 that increased educational attainment for cohorts born after 1980. Using two cross-sectional surveys and focusing on respondents over 17 years of age, they find that one more year of secondary schooling reduces the accumulated risk of infection by 12 percentage points for women and 5 percentage points for men.

A few randomized controlled trials provide indirect evidence on a potentially protective effect of education. Baird et al. (2010; 2012) find that a cash transfer program in Malawi increased school attendance and reduced HIV infections, but the effects could be due to the cash and/or the increased school attendance. Handa et al. (2014) find that an unconditional government cash transfer program in Kenya reduced the probability of sexual debut among those aged 15-25, though the impact on HIV infection was not significant. Kahn et al. (2015) assess the effect on HIV acquisition of a cash transfer conditioned on school attendance among young South African women. They find no effect of the transfers on HIV infection or school attendance, but young women who dropped out of school were three times more likely to become HIV infected than the others. However, it is possible that this finding is due to differences in personal traits between those who dropped out and the average young woman.

Finally, an impact evaluation of girls in Kenya by Duflo et al. (2015) finds that education subsidies reduced pregnancy but not STIs, while HIV information in school did not reduce STIs or pregnancy, but led to a shift from out-of-wedlock to within marriage pregnancies. However, when a combined program of education subsidies and HIV information was offered, both STI and pregnancy declined; the number of HIV infections was too small to be used as an outcome variable. The authors set up a theoretical model and show that their results are inconsistent with models where STI infection and pregnancy are the outcomes of a single common determinant:

unprotected sex. They propose a model with a distinction between committed relationships, which are likely to end in marriage in the event of pregnancy, and casual relationships, which are not likely to end in marriage. The likelihood of pregnancy is higher in committed relationships while the likelihood of STIs is higher in casual relationships.

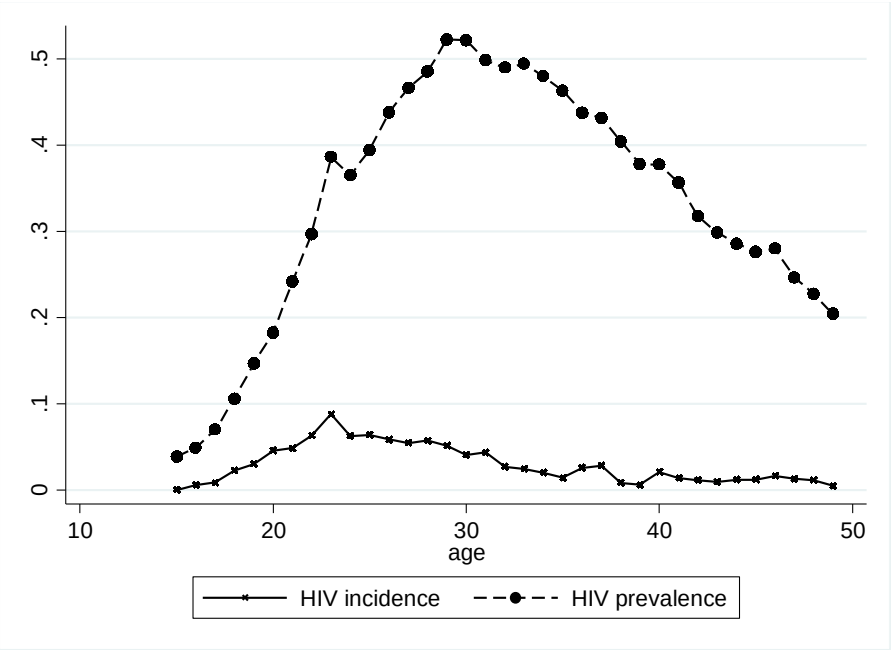
To conclude, there is some support for the claim that school attainment reduces HIV, but the evidence for school attendance is not strong (Baird et al., 2012; Kennedy et al., 2014). Most studies focus on associations, and those that evaluate cash transfer programs do not distinguish between the effect of schooling and the benefits of receiving cash, which, for example, could reduce transactional sex.

4. The data

The data are from a population-based longitudinal surveillance of 85 000 people who are members of approximately 11 000 households (Tanser et al., 2008). It is conducted by the Africa Centre for Population Health in a predominantly rural community in KwaZulu-Natal. It is one of the poorest communities in South Africa. HIV is very common, as can be seen in Figure 1, which shows female HIV prevalence and incidence rates in 2005 by age. The HIV prevalence rate peaks at over 50%, for women aged 29-30, and the HIV incidence rate approaches 10% at age 23.

We use data for the period January 2005 to June 2012. Our estimation sample consists of young women aged 15 to 24, who have completed primary school but not secondary school (hence they should be attending secondary school), and who were HIV negative in the previous year. The share of the young women who have not completed primary school is small: 19.5% at age 15, decreasing to 5.8% at age 24. The share of the young women who have completed secondary school increases with age. At age 19 – when young women should have completed secondary school if they followed their age-appropriate grade throughout – 18.1% have completed secondary school, while 43.8% have completed secondary school by age 24.

Figure 1: Female HIV prevalence and HIV incidence in 2005 by age



The effects of education on health outcomes vary across levels of education (Clark and Royer, 2013). We analyze the educational level where we believe that school attendance ought to have the largest influence on HIV incidence, since young women attend secondary school during a period when most of them start having sexual relationships. Moreover, there is a substantial variation in secondary school attendance. In South Africa, secondary school spans from grade 8 to grade 12. Schooling is compulsory until age 15 or until grade 9 is completed; therefore, all the young women in the estimation sample could legally leave school if they wish.

The control variables, used in all models, are age dummies, year dummies, wealth quintile dummies, dummies for urban or peri-urban residence, distance to the primary road, distance to the secondary road and distance to the nearest health clinic. The wealth dummies are based on a wealth index constructed using principal component analysis. Table A1 in the appendix presents summary statistics of all variables.

3. The empirical strategy

The possibility of selection bias and causal effects can be illustrated using the potential outcomes framework of Rubin (Imbens and Rubin, 2008). Assume there is one potential HIV infection

outcome for each young woman if she attends school, $\mathrm{Y_i}(1)$, and another one if she does not, $\mathrm{Y_i}(0)$. The effect of school attendance for the young women *i* is $\mathrm{Y_i}(1) - \mathrm{Y_i}(0)$, but we observe only one of the potential outcomes. The observed difference in outcomes between young women who attend school and young women who do not, $\mathrm{E}[\mathrm{Y_1}(1)] - \mathrm{E}[\mathrm{Y_0}(0)]$, thus consists of two parts: the average treatment effect on the treated (ATT), $\mathrm{E}[\mathrm{Y_{1i}}(1)] - \mathrm{E}[\mathrm{Y_{1i}}(0)]$, and selection bias, $\mathrm{E}[\mathrm{Y_{1i}}(0)] - \mathrm{E}[\mathrm{Y_{0i}}(0)]$. The selection bias stems from the fact that school attendance is not random. Girls who attend school are likely to differ in various ways from girls who do not; for example they might have different time preferences and they might be more able in school. Girls who value the future more might be more careful not to contract HIV, independent of whether or not they go to school. And girls who are better academically might also be more able to process HIV and family planning information and to change their behavior in accordance with that information, whether or not they go to school. Hence, there are reasons to expect that $\mathrm{Y_{1i}}(0) - \mathrm{Y_{0i}}(0) \neq 0$.

We estimate the discrete time hazard function, $h(t) = \Pr[t \leq T < t + \Delta t \mid T \geq t]$, which is the probability of leaving a state, HIV negative in this instance, at time *t* conditional on not having done so before. The discrete, rather than continuous, time hazard function is estimated since data was collected at discrete points in time, and, importantly, since the discrete time hazard function can be estimated with any binary variable technique (it is a conditional probability). This makes it possible to use instrumental variables, techniques to evaluate robustness to selection on unobserved factors and models that control for time-constant individual effects. When estimating a discrete time hazard, it is essential to control for duration, i.e., how long young women have been at risk of HIV acquisition, as flexibly as possible. This is done with age dummies.

We first estimate the bivariate probit model (BPM), which can handle an endogenous regressor, either by the use of instruments or by evaluating robustness to selection. Let y_{it} be HIV incidence for individual *i* in year *t* (note that *i* was uninfected in *t-1*). s_{it} is secondary school attendance, x_{it} is a vector of control variables, and z_{it} is a vector of variables that predict school attendance but not HIV incidence conditional on s_{it} and x_{it}, , i.e., a vector of instruments. Then,

$$s_{it} = 1\left(x_{it}\delta + z_{it}\gamma + u_{it} > 0\right) \quad (1)$$

$$y_{it} = 1\left(x_{it}\pi + s_{it}\beta + \varepsilon_{it} > 0\right) \tag{2}$$

$$\begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{3}$$

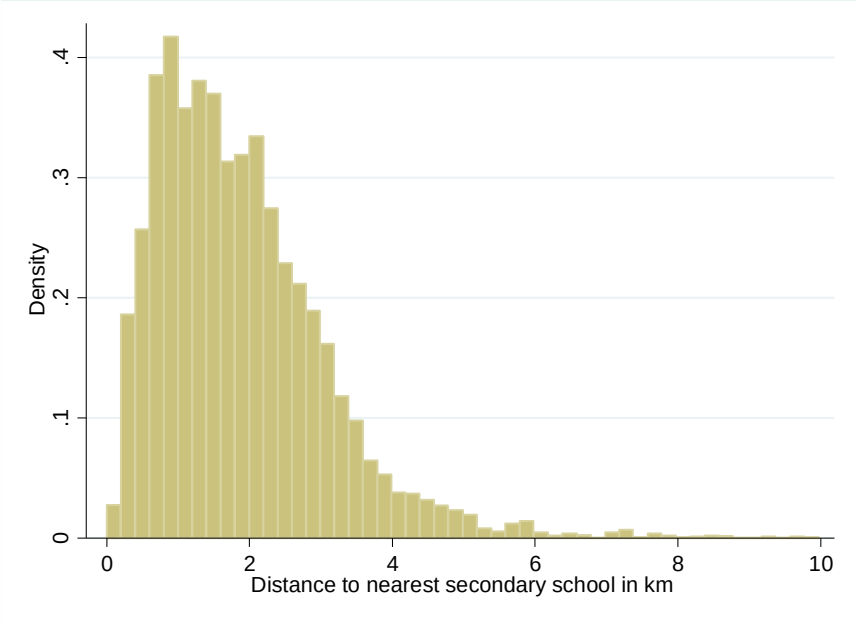
The unobserved determinants of school attendance, u, and HIV incidence, ε, are assumed to have a joint bivariate normal distribution, and ρ is the correlation between them. If $\rho = 0$, there is no selection on unobserved variables and the BPM will provide the same information as the (univariate) probit model. If ρ is significantly different from zero, we can reject the probit model in favor of the BPM.

The exclusion restriction is valid if z predicts school attendance, ($\gamma \neq 0$), but not HIV incidence conditional on x and s. Our instruments are based on distances to secondary schools. As seen in Figure 2, the distance to the nearest secondary school varies between 0 and 10 km, with most of the young women living within 2 km of the nearest secondary school.[4] The specific instruments are a dummy for living closer than 7 km to the nearest secondary school, and the difference (in km) between the nearest and second nearest secondary school. For girls living closer than 7 km to a secondary school, distance has little predictive power, while school attendance drops from 79% to 57% at the 7 km cut-off. The main drawback to this instrument is that few young women live 7 km or more from a secondary school. The second instrument reflects choice; more schools nearby makes it more likely that a young woman finds a school that suits her.

Distance to a secondary school is strongly correlated with urban residence. Because HIV incidence tends to be higher in urban than rural areas, it is important to control for urbanity/remoteness; we do so with dummies for urban and peri-urban residence and with distances (in km) to the primary road, to the secondary road and to the nearest health clinic. Conditional on these controls, we argue that distance to school is not related to the risk of HIV incidence through any other channel than secondary school attendance, and hence that our exclusion restriction is valid.

[4] Since secondary schools are relatively close by, students usually live with their family.

Figure 2: Distribution of distance to school (km) for young women age 15-24 who have not completed secondary school



Our next approach, by Altonji et al. (2005), does not depend on an exclusion restriction. Instead the BPM is treated as under-identified, where ρ is not estimated but imposed. Hence we estimate,

$$s_{it} = 1\left(x_{it}\delta + u_{it} > 0\right) \tag{4}$$

$$y_{it} = 1\left(x_{it}\pi + s_{it}\beta + \varepsilon_{it} > 0\right) \tag{5}$$

$$\begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho = \varphi \\ \rho = \varphi & 1 \end{bmatrix}\right) \tag{6}$$

under different assumptions about φ. This allows us to evaluate how sensitive the marginal effect of school attendance on HIV incidence is to different degrees of selection on unobserved factors. The non-trivial task is to select sensible values of ρ. Altonji et al. (2005) suggest using 0 as a lower bound and using the selection on observed control variables as an upper bound. More specifically, the upper bound on ρ is $\rho = Cov(x\delta, x\pi)/Var(x\pi)$. If all variables were observable and the researcher picked control variables randomly, the selection on unobserved (excluded) variables would asymptotically equal the selection on observed (included) variables.[5]

[5] In reality, researchers are likely to do better than random selection of included control variables.

Because we also estimate the BPM with an exclusion restriction, a convenient alternative could be to use ρ from this model as guidance on what a reasonable value is.

Next, we estimate linear probability models (LPMs), both because they are considered a robust alternative to probit and BPM models (Angrist and Pischke, 2008), and because they facilitate the use of instrumental variables and individual fixed effects. However, it is well known that the LPM is mis-specified (the linearity assumption is clearly not correct), although there is no agreement on the practical implications. Horrace and Oaxaca (2006) show that the LPM is biased when many predicted probabilities are outside of the unit range. HIV incidence is low, i.e., quite close to 0, so this is could potentially affect our estimates.

An advantage of the LPM compared to the BPM is that it provides a good framework to evaluate the strength of instruments. Yet, if the BPM is a good description of the data generating process, it is both more efficient and less biased than the linear probability instrumental variable model (LPIVM), although it is less clear which model, BPM or LPIVM, is most robust to deviations from the normality assumption (Bhattacharya et al., 2006; Chiburis et al., 2011). Another potential shortcoming of the LPIVM is that standard errors often are too large for meaningful hypothesis testing. On the other hand, BPM standard errors are too small in small samples – below 5,000 (Chiburis et al., 2011).

Fixed effects capture time-constant common determinants of secondary school education and HIV risk behavior. If there are no time-varying determinants of both school drop-out and HIV incidence, we will get the causal effect. However, because only within variation is used, individual fixed effects estimation does not use young women who attend secondary school every year until graduation or those who remain uninfected for identification.

To evaluate robustness to selection on unobserved variables in a linear framework, we use the approach developed by Oster (2015). It relies on the assumption that the selection on observed control variables is informative about potential selection on unobserved variables.[6] Specifically the selection on unobserved variables is assumed to be proportional to the selection on observed

[6] This does not imply correlation between observed controls and unobserved factors; in fact, if unobserved factors were highly correlated with the observed control variables, we would not have a selection problem. We therefore only need to consider the orthogonal parts of the unobserved factors and the direction of their impact relative to the observed (control) variables.

variables, where δ is the coefficient of proportionality. An assumption is also needed about the R-squared that would be obtained if all relevant variables, both observed and unobserved, were included in the model, denoted maximum R-squared. The main challenges are therefore to decide on the coefficient of proportionality and maximum R-squared. We follow Altonji et al. (2005) and Oster (2015) in assuming that the selection on unobserved variables is at most as large as the selection on observed variables, i.e., δ=1. The maximum R-squared is likely to be much lower than 1 in our case, both because our dependent variable is binary and because there is a great deal of idiosyncratic variation in HIV incidence. We follow Oster (2015) and use the R-squared from a fixed effects model. Individual fixed effects capture time-constant differences in, for example, time preferences and abilities.[7]

5. Empirical analysis

5.1 Probit results

Table 1 shows probit estimation results of the relationship between school attendance and HIV incidence. There is a strong negative association. Young women who attend secondary school are 1.3 percentage points less likely to become infected with HIV. This is a large effect given the average HIV incidence of 2.8% in the estimation sample. In particular, it is noteworthy that the association is robust to the inclusion of a full set of age dummies, since several previous studies use a linear age term or a smaller set of age dummies.[8] The complete results of Table 1 are available in the appendix (Table A2). Complete results from other regressions are available from the authors on request.

5.2 Accounting for selection - bivariate probit model

Next, to account for selection, we estimate HIV incidence and school attendance jointly with a BPM (Table 2). The marginal effect of school attendance switches the sign but is not statistically significant. Both instruments are strong predictors of school attendance and the marginal effects have the expected signs. Young women living closer than 7 km to a secondary school are 17.2 percentage points more likely to attend school than young women living farther away. Each extra

[7] If there are also time-varying factors that matter, maximum R-squared might be slightly larger, but we do not believe such time-varying factors to be of much importance.
[8] We also have information about mother's education in about 50% of the sample. It is negative and significant but does not affect the estimate of the marginal effect of schooling or the conclusions drawn in any of the following analyses. Thus, we report results without mother's education.

km between the nearest and second nearest school decreases the probability of school attendance by 1.4 percentage points. The correlation between the unexplained variation in school attendance and the unexplained variation in HIV incidence, ρ, is estimated to be -0.197, and is statistically significant. Hence, we can reject the probit model in favor of the BPM. A negative ρ means that unobserved factors that increase the probability of school attendance decrease the probability of HIV incidence, which is expected.

Table 1: The association between secondary school attendance and HIV incidence – probit marginal effect

School attendance	-0.013**
	(0.005)
Number of observations	6,502
Number of women	2,789

The model also includes a constant; age, year and wealth quintile dummies; peri-urban or urban residence; and distances to the primary road, the secondary road and the nearest health clinic. Standard errors, clustered at the individual level, in parentheses. * p<0.1; ** p<0.05; *** p<0.01

Table 2: The impact of secondary school attendance on HIV incidence – bivariate probit marginal effects

HIV incidence equation		
	School attendance	0.007
		(0.012)
School attendance equation		
	Nearest secondary school<7km away	0.179***
		(0.040)
	Distance between the nearest and second nearest secondary school	-0.015***
		(0.005)
ρ		-0.197
[p-value of test of ρ=0]		[0.039]
Number of observations		6,493
Number of women		2,785

Both equations also include a constant; age, year and wealth quintile dummies; peri-urban or urban residence; and distances to the primary road, the secondary road and the nearest health clinic. Standard errors, clustered at the individual level, in parentheses. * p<0.1; ** p<0.05; *** p<0.01

5.3 Evaluating the impact of selection

In this section, we use the approach developed by Altonji et al. (2005) to evaluate the sensitivity of the marginal effect from the probit model to potential selection on unobserved factors. Table 3 presents the marginal effects of school attendance on HIV incidence, with ρ equal to 0.00, -0.05,

-0.10, -0.15, -0.20, -0.25 and -0.30, which seems like a reasonable range since $\rho = -0.197$ in our estimate of the BPM.[9]

The estimated marginal effect of school attendance on HIV incidence does not appear to be robust to selection. It becomes statistically insignificant when $\rho = -0.05$, is already positive when $\rho = -0.15$, and is exactly zero when $\rho = -0.13$. Thus, far less correlation is required to remove the marginal effect of schooling on HIV incidence than 0.197, estimated with the BPM.

Table 3: Robustness of the impact of secondary school attendance on HIV incidence to selection on unobserved factors

Assumed ρ	0.00	-0.05	-0.1	-0.15	-0.2	-0.25	-0.3
Marginal effect	-0.013**	-0.008	-0.003	0.002	0.007	0.012**	0.018***
Standard error	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)

Based on constrained bivariate probit estimations. Both the school attendance and the HIV incidence equations include a constant, age, year and wealth quintile dummies, peri-urban or urban residence, and distances to the primary road, the secondary road and the nearest health clinic.

5.4 A linearized model

In this section, we use LPMs to analyze the data (Table 4). When the LPM is estimated with OLS, the marginal effect is slightly larger than in the probit model: HIV incidence is 1.7 percentage points lower among young women attending secondary school. The LPM is more robust and produces less bias if predicted probabilities are within the unit range (Horrace and Oaxaca, 2006). Predicted probabilities from the OLS regression in column 1, Table 4, are below 0 for 15.09% of the observations.

When the instruments are used and the model is estimated with 2SLS and LIML (IV-2SLS and IV-LIML), the marginal effect is even more negative (in both models), and thus differs greatly from the one obtained with BPM, which is positive. However, the estimates are statistically insignificant and the confidence intervals are very wide, including extreme values on both the negative and positive sides. This is in line with Chiburis et al. (2012), who show that LPIVM estimations sometimes result in extremely broad confidence intervals.

Nonetheless, the instruments seem to be reasonably strong. The first stage regression results (in the appendix, Table A3) show that the instrument coefficients are statistically significant and of

[9] If we let ρ be such that selection on unobserved factors equals selection on observed control variables, which is -0.37, the BPM does not converge.

the expected signs. The commonly used rule-of-thumb is that the F-statistic of excluded instruments should be at least 10; our F-statistic is 11.47 (Table 4). We also report critical values based on the maximum IV estimator bias, developed by Stock and Yogo (2005).[10] The critical values depend on both the specific IV estimator and the number of instruments used. Because the limited information maximum likelihood (LIML) estimator is more robust to weak instruments than the more commonly used two-stage least square (2SLS) estimator, it also has lower critical values. Our instruments are always strong enough for the LIML estimator (i.e., the F-statistic is larger than the critical value for the 10% maximal IV size). Whether or not they are strong enough for the 2SLS estimator depends on how tolerant we are (in terms of maximum IV estimator bias). Thus, we conclude that, although the instruments are reasonably strong, there is no evidence of a causal effect. However, the large negative value of the estimated coefficient makes this conclusion highly tentative, so further analysis is required.

The last column of Table 4 reports estimation with individual fixed effects, which controls for the influence of unobserved constant factors. The estimated marginal effect is very small and statistically insignificant, supporting earlier findings.

Next, we use the approach developed by Oster (2015) to evaluate robustness to selection on unobserved factors. We set the maximum R-squared to the R-squared obtained with the fixed effects estimator, which is 0.070. Oster's approach is used in two ways. First, we compute bounds on the coefficient, given the coefficient of proportionality δ. We get the lower bound (our effect is negative) by assuming no selection on unobserved factors, i.e., $\delta = 0$, which is the OLS coefficient. To get the upper bound, we assume that the contribution of the unobserved factors is equal to that of the observed control variables, i.e., $\delta = 1$. The second approach is to compute the value of δ that makes the estimated coefficient equal to zero. This tells us how much selection on unobserved factors, in comparison to the selection on observed control variables, is needed to explain away the estimated impact of school attendance on HIV incidence. Outcomes of both approaches are reported in Table 5. In the second column, all control variables are treated as potentially informative about the selection on unobserved factors, while in the third column we

[10] The IV estimator will always contain a bias if the endogenous regressor is indeed endogenous. The size of this bias depends on the strength of the instrument. The maximum IV estimator bias is expressed in relation to the OLS bias. For example, a 15% maximum IV estimator bias means that the IV estimator bias is at most 15% of the OLS estimator bias.

include only control variables that are likely to affect the probability of HIV incidence and the probability of school attendance in the opposite direction of each other, i.e., age and wealth quintile dummies (the other control variables are in both the uncontrolled and the controlled model).

Table 4: Linear estimations of the impact of secondary school attendance on HIV incidence

	OLS	IV – 2SLS	IV - LIML	Individual fixed effects
School attendance	-0.017**	-0.027	-0.028	-0.002
	(0.008)	(0.067)	(0.074)	(0.010)
R^2	0.03	0.03	0.03	0.07
Number of observations	6,502	6,493	6,493	6,502
Number of women	2,789	2,785	2,785	2,789
F – test of excluded instruments		11.47	11.47	
Stock-Yogo weak ID test critical values:				
10% maximal IV size		19.93	8.68	
15% maximal IV size		11.59	5.33	
20% maximal IV size		8.75	4.42	
25% maximal IV size		7.25	3.92	

All models also include a constant; age, year and wealth quintile dummies; peri-urban or urban residence; and distances to the primary road, the secondary road and the nearest health clinic. Standard errors, clustered at the individual level, in parentheses. * $p<0.1$; ** $p<0.05$; *** $p<0.01$. The instruments are Nearest secondary school<7km away and Distance between the nearest and second nearest secondary school.

When all control variables are viewed as potentially informative about the selection on unobserved variables, the upper bound is 0.034 and the lower one is -0.017, and the marginal effect of school attendance switches sign when $\delta=0.412$. When we view as informative only variables that are more likely to share covariance properties with unobserved factors, the upper bound increases to 0.056 (the lower bound does not change), and the coefficient switches sign when $\delta=0.306$. Because the upper bound is large in both cases, and relatively little selection bias is needed to remove the school attendance effect, there is no support for a causal effect of schooling on HIV.

Table 5: Robustness of the impact of secondary school attendance on HIV incidence to selection on unobserved factors

	Use all controls[a]	Only use controls expected to share covariance properties[b]
Assumed maximum R^2	0.070	0.070
Uncontrolled R^2	0.012	0.017
Controlled R^2	0.032	0.032
Coefficient bounds		
Upper bound (δ=1)	0.034	0.056
Lower bound (δ=0)	-0.017	-0.017
How much selection to explain away the estimated OLS effect		
δ making the estimated effect =0	0.412	0.306

a) Age, year and wealth quintile dummies, peri-urban or urban residence, and distances to the primary road, the secondary road and the nearest health clinic.
b) Year dummies, peri-urban or urban residence and distances to the primary road, the secondary road and the nearest health clinic are included both in the controlled and the uncontrolled regressions

5.5 Sexual activity and pregnancy

Our results seem to be at odds with those in Alsan and Cutler (2013), Behrman (2015) and De Neve et al. (2015). While Behrman (2015) and De Neve et al. (2015) study long-term effects of school attendance, Alsan and Cutler (2013) do not observe HIV infection and therefore use the effect of schooling on age at sexual debut as their outcome. However, HIV infection is the outcome of various proximate determinants, not only sexual debut; most prominent are the frequency of intercourse, number of partners, type of partners, and condom use. Schooling might affect these proximate determinants differently. In particular, as suggested by Duflo et al. (2015), it might induce some young women to delay their sexual debut, while encouraging others to shift from stable relationships (with a higher likelihood of pregnancy) to casual relationships (with a higher likelihood of infection with HIV and other STIs). Hence, in this section, we analyze sexual activity, pregnancy and schooling to shed some further light on the association between school attendance and HIV.

Table 6 reports the results from probit and linear regressions and an individual fixed effects model (the instruments are too weak to be of any use), where initiation of sexual activity is the dependent variable. We include the same control variables as in the previous models. The effect of school attendance varies between -0.342 and -0.197 across the models; it is much stronger and

more significant than in the models with HIV infection. Furthermore it is clearly significant in the fixed effects model.

Table 6: Estimations of the impact of secondary school attendance on sexual debut

	Probit (marginal effects)	OLS	Individual fixed effects
School attendance	-0.284***	-0.342***	-0.197***
	(0.017)	(0.036)	(0.021)
R^2		0.29	0.35
Number of observations	5,401	5,401	5,429
Number of women	3,489	3,489	3,507

Sexual debut is a binary variable, which equals 0 if the respondent reports never having had sex and equals 1 the first time the respondent reports having had sex. All models also include a constant, age, year and wealth quintile dummies, peri-urban or urban residence dummies, and distances to the primary road, the secondary road and the nearest health clinic. Standard errors, clustered at the individual level, in parentheses. * p<0.1; ** p<0.05; *** p<0.01

.

To evaluate the stability of the coefficient in the linear model (OLS), we repeat Oster's (2015) analysis of the role of unobserved factors. (The approach of Altonji et al. (2005) gives similar results). We use an R-squared of 0.35, which is obtained with a fixed effects model (Table 7). The results seem reasonably robust: the upper and lower bounds are – 0.211 and –0.342. There is only a marginal difference between the results when we treat all control variables as potentially informative or when only those that are likely to share covariance properties with the unobserved variables are treated as informative. The effect of school attendance switches sign when $\delta = 1.43$, which is higher than the assumed maximum of 1. Thus, school attendance is associated with, and possibly causes, delayed sexual debut. This is a key finding of Alsan and Cutler (2013).

Table 8 reports probit model regression results with HIV incidence as the dependent variable and sexually active (ever had sex) and pregnancy (ever been pregnant) as explanatory variables (column 1). As expected, sexual activity increases the risk of being infected; the marginal effect is 0.05. However, young women who have been pregnant are clearly less likely to be infected than are other sexually active young women. Adding school attendance to the model (column 2) does not affect the coefficients of either pregnancy or ever having had sex. Because school attendance is insignificant and the marginal effect is close to zero, sexual activity is clearly the

main mediator, and school attendance does not seem to affect HIV risk among those who are sexually active.

Table 7: Robustness of the impact of secondary school attendance on sexual debut to selection on unobserved factors

	Use all controls[a]	Only use controls expected to share covariance properties
Assumed maximum R^2	0.35	0.35
Uncontrolled R^2	0.193	0.203
Controlled R^2	0.292	0.292
Coefficient bounds		
Upper bound if δ=1	-0.211	0.211
Lower bound if δ=0	-0.342	--0.342
How much selection to explain away the estimated OLS effect		
δ making the estimated effect =0	1.43	1.41

a) Age, year and wealth quintile dummies, peri-urban or urban residence, and distances to the primary road, the secondary road and the nearest health clinic.
b) Year dummies, peri-urban or urban residence and distances to the primary road, the secondary road and the nearest health clinic are included in both the controlled and the uncontrolled regressions

Table 8: The impact of pregnancy and sexual activity on HIV incidence, probit marginal effects

	(1)	(2)
Ever been pregnant	-0.016**	-0.017***
	(0.006)	(0.007)
Ever had sex	0.052***	0.051***
	(0.008)	(0.008)
Attend school		-0.007
		(0.006)
Number of observations	5,163	5,144
Number of women	2,567	2,561

All models also include a constant, age, year and wealth quintile dummies, peri-urban or urban residence, and distances to the primary road, the secondary road and the nearest health clinic. Standard errors, clustered at the individual level, in parentheses. * $p<0.1$; ** $p<0.05$; *** $p<0.0$

To conclude, schooling seems to affect the initiation of sexual activity but not HIV infection among those who are sexually active. Moreover, among sexually active young women, those who have been pregnant are less likely than other young women to be infected. This is probably because women who become pregnant often have stable relationships, while other sexually active

young women are more likely to have casual relationships, and are therefore more likely to be HIV infected (Duflo et al., 2015).

6. Conclusion

This study utilizes a longitudinal dataset on HIV incidence from rural KwaZulu-Natal in South Africa, an area with very high HIV rates, to estimate the causal effect of school attendance on HIV incidence among young women. Although we employ a variety of approaches, each with its virtues and vices, none suggests a causal effect; selection seems to be the main reason behind the negative correlation between school attendance and HIV infection. Although it is not possible to completely rule out a causal effect, if there is one, it is probably small.

Our findings have implications for the interpretation of the multitude of studies that only analyze associations. Young women who attend school differ from those who drop out, so selection is likely to be of key importance in all of these studies (whether cross-sectional or panel-data studies). This means that claims about the protective effects of schooling should be interpreted with caution.

At first glance, our results might appear to be in conflict with those of Behrman (2015) and De Neve et al. (2015), who find a protective effect of schooling. However, these studies analyze the impact of increased school attainment on cumulative risk of HIV infection until the age at which their data was collected, while we analyze effects of current school attendance. The protective effect found can thus be due both to the direct impact of increased human capital on risk taking and to an indirect effect via labor and marriage market outcomes. Network effects are likely to be important in our study, but of little importance for their findings, and these might not be protective. Women are likely to meet sexual partners in school, and whether they meet more or riskier partners than young women who have dropped out depends on the specific school and community.

Our results might also appear to be at odds with those of Alsan and Cutler (2013), who rely on the impact of schooling on delayed sexual debut (virginity). However, delayed sexual debut is just one proximate determinant: school attendance could make some girls delay their sexual

debut, while others have more premarital casual relationships. The study by Duflo et al. (2015) on Kenyan data suggests that education might lead to such a shift, i.e., from committed relationships (with a higher likelihood of pregnancy) to casual ones (with a higher likelihood of STIs, including HIV). There is some support for delayed sexual debut as a consequence of school attendance in our data too, but many of the sexually active girls remain in school, while those who become pregnant, who often drop out, are less likely than other sexually active young women to be HIV infected.

Again, the specific setting might matter. Secondary education might have had a more beneficial impact in Uganda due to the simultaneous massive "No Grazing" and ABC campaigns in Uganda. This would also be consistent with the experimental study from Kenya by Duflo et al. (2015). In South Africa, the Life Orientation Program, which includes HIV education, was introduced in the curriculum in 2000, but the quality of implementation has been questioned (Jacobs, 2011; Adams-Tucker et al., 2015). Young women in South Africa are also surrounded by other types of information, which likewise shape their beliefs.

In spite of the many studies on education and HIV incidence, those who go beyond associations and attempt to separate causation from selection are surprisingly few. It is therefore a challenge to draw general conclusions without more research. However, our study suggests that secondary school attendance does not have a large causal protective effect on HIV incidence. Therefore, if we are concerned about HIV infections among young women, we cannot be satisfied with increased schooling; more targeted measures are needed. The papers by Behrman (2015) and De Neve et al. (2015) show that there can be a protective effect of school attainment, at least given the right circumstances, but this affect occurs after many years, that is, in the long run. Our results suggest the need for short-run measures during the secondary school years.

References

Adams-Tucker, L., George, G., Reardon, C. and Panday, S. (2015). "Learning the Basics": Young peoples' engagement with sexuality education at secondary schools. *Sex Education* (in Press)

Alsan, M. M. and D. M. Cutler, (2013) Girls' education and HIV risk: Evidence from Uganda. *Journal of Health Economics*, 32(5): 863-872.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 113(1): 151-184.

Amin, V., Behrman, J. R., and Kohler, H. P. (2015). Schooling has smaller or insignificant effects on adult health in the US than suggested by cross-sectional associations: New estimates using relatively large samples of identical twins. *Social Science and Medicine*, 127: 181-189.

Angrist, J. D., and Pischke, J. S. (2008) *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Baird, Sarah, Chirwa, Ephraim, McIntosh, Craig, and Özler, Berk (2010) The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. *Health Economics*, 19(S1): 55-68.

Behrman, J. A. (2015). The effect of increased primary schooling on adult women's HIV status in Malawi and Uganda: Universal primary education as a natural experiment. *Social Science and Medicine*, 127: 108-115.

Berger, M. C., and Leigh, J. P. (1989). Schooling, self-selection, and health. *Journal of Human Resources*, 24(3): 433-455.

Bhattacharya, J., Goldman, D., and McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in medicine*, *25*(3): 389-413.

Black, Sandra E. and Devereux, Paul J. and Salvanes, Kjell G (2008). Staying in the Classroom and out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births, *The Economic Journal* 118(530): 1025-1054.

Branson, N., Hofmeyr, C., and Lam, D., (2013) Progress through school and the determinants of school dropout in South Africa. A Southern Africa Labour and Development Research Unit Working Paper Number 100. Cape Town: SALDRU, University of Cape Town

Bärnighausen, T., Hosegood, V., Timaeus, I. M. and Newell, ML. (2007). The socioeconomic determinants of HIV incidence: evidence from a longitudinal, population-based study in rural South Africa. *AIDS* 21(Suppl. 7): S29–S38.

Case, A., and Paxson, C. (2013). HIV Risk and Adolescent Behaviors in Africa. *The American Economic Review Papers and Proceedings*, 103(3): 433-438.

Chiburis, R. C., Das, J., and Lokshin, M. (2011). *A Practical Comparison of the Bivariate Probit and Linear IV Estimators.*

Clark, D., and Royer, H. (2010). The effect of education on adult health and mortality: Evidence from Britain (No. w16013). National Bureau of Economic Research.

De Neve, J. W., Fink, G., Subramanian, S. V., Moyo, S., and Bor, J. (2015). Length of secondary schooling and risk of HIV infection in Botswana: evidence from a natural experiment. *The Lancet Global Health*, 3(8): e470-e477

De Walque, D. (2007). How does the impact of an HIV/AIDS information campaign vary with educational attainment? Evidence from rural Uganda. *Journal of Development Economics*, 84(2): 686-714.

Duflo, E., Dupas, P., and Kremer, M. (2015). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economic Review* 105(9): 2257-97.

Fortson, J. (2008). The Gradient in Sub-Saharan Africa: Socioeconomic Status and HIV/AIDS" *Demography,* 45(2):303 -322

Grapsa E., Zaidi J., Tanser F., Newell ML., and Bärnighausen, T. (2013) The effects of school attendance on HIV acquisition: Evidence from a full population cohort study in rural South Africa. Mimeo, Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa.

Gummerson, E. (2013) Have the educated changed HIV risk behaviours more in Africa? *African Journal of AIDS Research*, 12(3): 161 -172.

Gupta, G. R., J. O. Parkhurst, J. A. Ogden, P. Aggleton, A. Mahal, (2008) Structural approaches to HIV prevention", *The Lancet*, 372(9640): 764-775.

Handa S., Halpern C.T., Pettifor A., Thirumurthy H. (2014) The Government of Kenya's Cash Transfer Program Reduces the Risk of Sexual Debut among Young People Age 15-25. *PLoS ONE* 9(1): e85473.

Hardee, K., J. Gay, M. Croce-Galis and A. Peltz (2014) Strengthening the enabling environment for women and girls: what is the evidence in social and structural approaches in the HIV response?" *Journal of the International AIDS Society*, 17:18619.

Hargreaves, J. R., C. P. Bonell, L. A. Morison, J. C. Kim, G. Phetla, J.D.H. Porter, C. Watts, P.M Pronyk (2007) Explaining continued high HIV prevalence in South Africa: socioeconomic factors, HIV incidence and sexual behaviour change among a rural cohort, 2001-2004. *AIDS* 21(Suppl 7):S39 -S48.

Hargreaves, J.R., L.A. Morison, J.C. Kim, C. P. Bonell, J. D. H. Porter, C. Watts, J. Busza, G. Phetla, P. M. Pronyk (2008a) "Evidence-based public health policy and practice: The association between school attendance, HIV infection and sexual behaviour among young people in rural South Africa, *Journal of Epidemiological Community Health,* 62(2): 113-119.

Hargreaves, J. R., C.P. Bonell, T. Boler, D. Boccia, I. Birdthistle, A. Fletcher, P.M. Pronyk, J.R. Glynn (2008b) Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa. *AIDS* 22(3): 403-414.

Hargreaves, J. R., L. D. Howe (2010) Changes in HIV prevalence among differently educated groups in Tanzania between 2003 and 2007. *AIDS*, 24(5):755 -761.

Horrace, W. C., and R. L. Oaxaca (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, *90*(3): 321-327.

Imbens G. W. and D. B. Rubin (2008) "Rubin causal model" in *The New Palgrave Dictionary of Economics*, Eds. Steven N. Durlauf and Lawrence E. Blume, Second Edition.

Jacobs, A. 2011. "Life Orientation as experienced by learners: A qualitative study in North-West Province*." South African Journal of Education* 31 (2): 212-223.

Johnston, D., K. Deane and M. Rizzo (2015) The political economy of HIV. *Review of African Political Economy*, 42(145): 335-341.

Jukes, M., S. Simmons, D. Bundy (2008). Education and vulnerability: the role of schools in protecting young women and girls from HIV in southern Africa. *AIDS* 22(Suppl 4): 41-56.

Kahn, K. et al. (2015). Cash transfers conditional on schooling does not prevent HIV infection among young south African women" HPTN068 phase III study, presented at 8th IAS Conference on HIV Pathogenesis, Treatment and Prevention in Vancouver, Canada. Accessed http://www.wits.ac.za/newsroom/newsitems/201507/26918/news_item_26918.html

Kennedy, C. E., H. Brahmbhatt, S. Likindikoki, S. W. Beckham, J. K. Mbwambo, D. Kerrigan (2014). Exploring the potential of a conditional cash transfer intervention to reduce HIV risk among young women in Iringa, Tanzania. *AIDS Care* 26(3): 275-281.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *The Review of Economic Studies*, *72*(1): 189-221.

Oster, E. (2012) "HIV and sexual behavior change: Why not Africa?" *Journal of Health Economics* 31(1): 35-49.

Oster, E. (2015). "Unobserved selection and coefficient stability: Theory and validation" unpublished, Brown University. Available at http://www.brown.edu/research/projects/oster/

Peltzer, K., S. Ramlagan, W. Chirinda, G. Mlambo, and G. Mchunu (2012) "A community-based study to examine the effect of a youth HIV prevention programme in South Africa" *International Journal of STD and AIDS* 23: 653-658.

Pettifor, A. E., B.A. Levandowski, C. MacPhail, N.S. Padian, M.S. Cohen, H.V. Rees (2008) "Keep them in school: the importance of education as a protective factor against HIV infection among young South African women" *International Journal of Epidemiology,* 37(6): 1266-1273.

Pettifor, A., S. I. McCoy, N. Padian, (2012) "Paying to prevent HIV infection in young women?" *The Lancet*, 379(9823): 1280-1282.

Santelli, J. S., Z. R. Edelstein, S. Mathur, Y. Wei, W. Zhang, M.G. Orr, ... & D.M. Serwadda (2013). Behavioral, Biological, and Demographic Risk and Protective Factors for New HIV Infections among Youth, Rakai, Uganda. *Journal of Acquired Immune Deficiency Syndromes* (1999), 63(3): 393.

Seeley, J., C. H. Watts, S. Kippax, S. Russell, L. Heise, A. Whiteside (2012). Addressing the structural drivers of HIV: a luxury or necessity for programmes? *Journal of the International AIDS Society,* 15(Suppl 1):17397.

Stock, J. H., and M. Yogo (2005). Testing for weak instruments in linear IV regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, *1*.

Tanser, F., V. Hosegood, T. Bärnighausen, K. Herbst, M. Nyirenda, W. Muhwava,... M. L. Newell (2008). Cohort Profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. *International Journal of Epidemiology*, 37(5): 956-962.

UNAIDS (2015a) "Fact Sheet: 2014 Global Statistics". accessed 2015-08-04 http://www.unaids.org/en/resources/documents/2015/20150714_factsheet

UNAIDS (2015b) "South Africa, Epidemiological Fact Sheet on HIV and AIDS". accessed 2015-08-04 http://www.unaids.org/en/regionscountries/countries/southafrica/

Appendix: Additional Tables

Table A1: Estimation sample summary statistics

		Mean	Std. dev.	Min	Max
HIV incidence	6,502	0.028	0.164	0.000	1.000
School attendance	6,502	0.789	0.408	0.000	1.000
Urban	6,502	0.068	0.252	0.000	1.000
Peri-urban	6,502	0.189	0.391	0.000	1.000
Distance to the primary road	6,502	0.188	0.391	0.000	1.000
Distance to the secondary road	6,502	0.153	0.360	0.000	1.000
Distance to nearest health clinic	6,502	0.118	0.322	0.000	1.000
First wealth quintile	6,502	0.088	0.283	0.000	1.000
Second wealth quantile	6,502	0.066	0.247	0.000	1.000
Third wealth quintile	6,502	0.051	0.220	0.000	1.000
Fourth wealth quintile	6,502	0.042	0.202	0.000	1.000
Fifth wealth quintile	6,502	0.038	0.191	0.000	1.000
Age 15	6,502	0.093	0.290	0.000	1.000
Age 16	6,502	0.123	0.329	0.000	1.000
Age17	6,502	0.117	0.322	0.000	1.000
Age18	6,502	0.133	0.340	0.000	1.000
Age19	6,502	0.100	0.300	0.000	1.000
Age20	6,502	0.136	0.342	0.000	1.000
Age21	6,502	0.127	0.333	0.000	1.000
Age22	6,502	0.170	0.376	0.000	1.000
Age23	6,502	0.029	0.168	0.000	1.000
Age24	6,502	0.315	0.465	0.000	1.000
Year 2005	6,502	0.227	0.419	0.000	1.000
Year 2006	6,502	0.226	0.418	0.000	1.000
Year 2007	6,502	0.218	0.413	0.000	1.000
Year 2008	6,502	0.192	0.394	0.000	1.000
Year 2009	6,502	0.137	0.344	0.000	1.000
Year 2010	6,502	7.306	6.786	0.027	23.889
Year 2011	6,502	1.449	1.241	0.006	6.624
Year 2012	6,502	3.076	1.867	0.123	9.941
Nearest school closer than 7 km	6,502	0.990	0.099	0.000	1.000
Distance between the nearest and second nearest secondary school	6,493	1.616	1.143	0.003	7.534
Ever been pregnant	5,511	0.351	0.477	0.000	1.000
Sexual debut	5,612	0.347	0.476	0.000	1.000
Ever had sex	5,443	0.496	0.500	0.000	1.000

Table A2: The association between secondary school attendance and HIV incidence – probit marginal effects

School attendance	-0.013**
	(0.005)
Age 16	0.189***
	(0.015)
Age17	0.197***
	(0.015)
Age 18	0.219***
	(0.015)
Age 19	0.226***
	(0.015)
Age 20	0.237***
	(0.016)
Age 21	0.238***
	(0.016)
Age 22	0.244***
	(0.017)
Age 23	0.254***
	(0.017)
Age 24	0.247***
	(0.017)
Year 2006	0.019*
	(0.011)
Year 2007	0.020*
	(0.011)
Year 2008	0.013
	(0.011)
Year 2009	0.020*
	(0.011)
Year 2010	0.024**
	(0.011)
Year 2011	0.033***
	(0.011)
Year 2012	0.036***
	(0.010)
Urban	0.009
	(0.014)
Peri-urban	0.012**
	(0.006)
Second wealth quantile	-0.008
	(0.006)
Third wealth quintile	-0.007
	(0.006)
Fourth wealth quintile	-0.012*
	(0.007)
Fifth wealth quintile	-0.017**
	(0.008)
Distance to the primary road	-0.000
	(0.000)
Distance to the secondary road	-0.000
	(0.002)
Distance to nearest health	-0.001

clinic kmclinic2012	
	(0.001)
N	6,502

Standard errors, clustered at the individual level, in parenthesis. * $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table A3: First stage regression results

Nearest school closer than 7 km	0.213***
	(0.050)
Distance between the nearest and	-0.015***
second nearest secondary school	(0.005)
Age 16	-0.022**
	(0.010)
Age17	-0.043***
	(0.011)
Age 18	-0.090***
	(0.013)
Age 19	-0.189***
	(0.017)
Age 20	-0.329***
	(0.021)
Age 21	-0.418***
	(0.026)
Age 22	-0.523***
	(0.029)
Age 23	-0.660***
	(0.029)
Age 24	-0.753***
	(0.029)
Year 2006	0.021
	(0.015)
Year 2007	0.043**
	(0.017)
Year 2008	0.028
	(0.018)
Year 2009	0.028
	(0.020)
Year 2010	-0.001
	(0.018)
Year 2011	0.036*
	(0.018)
Year 2012	0.033*
	(0.018)
Second wealth quantile	0.025*
	(0.015)
Third wealth quintile	0.053***
	(0.017)
Fourth wealth quintile	0.065***
	(0.018)
Fifth wealth quintile	0.067***
	(0.019)
Urban	-0.081**

	(0.036)
Peri-urban	-0.013
	(0.016)
Distance to the primary road	-0.002
	(0.001)
Distance to the secondary road	0.003
	(0.005)
Distance to nearest health clinic	0.004
kmclinic2012	(0.004)
Constant	0.729***
	(0.063)
R^2	0.29
N	6,693