The Impact of HIV Prevalence on Schooling Achievement in Sub-Saharan Africa: Evidence from the SACMEQ Assessments

1^{st} YEAR MASTER THESIS
Lund University, Department of Economics

Author: Oskar Johansson
Supervisor: Jan Bietenbeck

August, 2016

Abstract. *This paper investigates the impact of HIV prevalence on schooling achievement. By associating data of schooling test results, for seven countries in Sub-Saharan Africa, with regional measures of HIV prevalence for 70 different regions, I find evidence that a general increase in HIV prevalence of about 1%, lowers general test score results of about 1% of a standard deviation from the mean. Endogeneity in the HIV variable is addressed by instrumenting HIV prevalence with male circumcision rate.*

Keywords: SACMEQ, DHS, HIV Prevalence, Schooling Achievement

1 Introduction

Since the first occurrences of HIV in the 1960's, the virus has disseminated from more or less negligible levels, to a situation where a significant amount of people are actually infected. In some parts of Africa, regional HIV rates can be as high as 30 percent of the population (DHS). This scourge creates a whole new situation of challenges for these countries. Even if the study of the HIV virus in itself is mostly relevant in the context of natural sciences, its vast impact on infected people's lives and the societies in which they live, makes it relevant also within the scope of social sciences.

The societal and socioeconomic perspective of HIV has been a subject of research in many papers during the latest years and the fact that the virus is most widespread in relatively poor parts of the world makes it especially interesting when investigating the economic challenges of these countries. Knowledge regarding the impact of HIV on fields like schooling, labor market productivity, socioeconomic status etc. can be of great importance when designing policies aimed to decrease poverty and improve economic conditions. As an example, several papers has found that increased HIV rates leads to orphanage of children (since there parents are more likely to die) (McCannon and Rodriguez, 2016). This, in turn, decreases schooling achievement of these children which, by extension, decreases the productivity of the labor force as a whole (Evans and Miguel, 2007; Case and Ardington, 2006). Other papers focus more directly on the impact of HIV prevalence on schooling enrollment and find suggestions of a causal negative relationship (Alsan and Cutler, 2013; Fortson, 2011). Additionally, children of HIV/AIDS infected parents are more likely to be infected themselves (since HIV can be transferred from mother to child during pregnancy) which further creates additional challenges for these children (Lowenthal et al., 2014).

In this paper, I aim to investigate the relationship between HIV prevalence and schooling achievement within a Sub-Saharan Africa. Previously published research in this field focuses mainly on the fact that higher HIV prevalence typically means higher mortality rates, which in turn leads to a decrease in expected life length, affecting schooling decision of individuals. This is often reflected as a general decrease of schooling attainment for regions with high prevalence of HIV. I instead, choose a different perspective for the outset of this paper, focusing not on the impact of HIV prevalence on schooling attainment, but on its possible impacts on achievement while in school.

The empirical analysis will be using data of standardized schooling tests, provided by *'The Southern and Eastern Africa Consortium for Monitoring Educational Quality'*. Three different subjects of knowledge are tested; math, reading and health. By combining these tests with regional characteristics like HIV prevalence and circumcision rate, calculated from DHS, I am able to estimate the impact of HIV prevalence on the general test result performance. Exploiting the fact that HIV prevalence varies between regions. I control for unobserved characteristics by instrumenting regional HIV prevalence on male circumcision rates within the same region.

My findings are significant and suggest that an increase in HIV prevalence of 1%, corresponds to a decrease in test score results of about 1% of a standard deviation from the mean score.

The rest of this paper is structured as follows: Section 2 reviews previous research in this field, connecting what has been found earlier to the contributions I aim to make. Section 3 presents the theoretical framework on which I base my assumptions. I then describe my empirical strategy. Section 4 presents the data, immediately followed by section 5, which describes my empirical strategy in detail. Section 6 present the empirical results and section 7 discusses the validity of my findings. Finally, section 8 concludes.

2 Previous Research

This section will lift three different focuses of research regarding the relationship between HIV and socioeconomic conditions. I start out describing the general economic impacts of HIV/AIDS. Then follows a background of the broad research field regarding the connection between life expectancy and schooling. The last subject focuses on the difference between schooling length and schooling achievement. Finally, I will present how this paper contributes to the already existing research on this subject.

The economic impacts of HIV/AIDS work through two main channels, health and human capital. The first of these states that HIV decreases economic efficiency in a direct way; by decreasing productivity of infected workers and increasing the cost of health care. Research suggests that HIV-infected workers both are more likely to be absent from work due to sickness, as well as less productive while at work, compared to uninfected workers. (Fox et al., 2004; Larson et al., 2008; Habyarimana et al., 2010)

The second channel works at a longer run, decreasing the amount of human capital in societies. There is a variety of research focusing on this aspect, investigating how HIV prevalence affects the premises of schooling and human capital accumulation within societies. The reasoning in a number of these papers is that the prevalence of HIV leads to a general increase in mortality rates, thereby decreasing expected life length [1]. This, in turn, suggests that individuals adjust their schooling decisions accordingly, as economic theory[2] suggests that a decline in expected life length decreases the expected wage return to schooling.

Investigating the relationship between mortality rates and human capital investments is not straight forward, as there are a number of different ways in which lower mortality rates can be associated with higher levels of education, despite being causally related. (An example of this would be *poorness*, working as an underlying factor related both to lower education and higher mortality rates.) Therefore, conducted research using only cross-sectional data to analyze this subject, may suffer from bias.

Jayachandran and Lleras-Muney (2009) tries to address this problem by using variation in maternal mortality rates for Sri-Lanka over a period of seven years. By analyzing the same country over time, rather than looking at multiple countries at the same point of time, they argue that they are able to mitigate the amount of bias in their results. They conclude that a decrease in female literacy of about 1 percentage point increase schooling of about 0.2 years.

Fortson (2011) further investigates this question and targets the relationship between human capital, measured as length of schooling, and HIV/AIDS specifically. Using individual level data of citizens in 15 Sub-Saharan African countries, she investigates the general level of schooling for different birth cohorts. She utilizes the fact that the appearance and spreading of HIV was quite sudden and occurred over a relatively short period of time, and argues that people born before 1980 can be considered more or less unaffected by HIV. A diff-in-diff analysis is then used to compare birth

[1] See for example Soares (2005); Kalemli-Ozcan (2002); Kalemli-Ozcan et al. (2000)

[2] This conclusion is based on the Ben-Porath (1967)-model, which will be discussed more extensively in section 3.

cohorts born before and after 1980. Theoretically, such a methodological framework do not suffer from the type of bias that is described earlier. Her findings suggest that an increase in regional HIV prevalence is associated with a general decrease in length of schooling.

Oster et al. (2013) continues on the work of Jayachandran and Lleras-Muney (2009) an Fortson (2011). The authors estimate the impact of life expectancy on human capital investments using data on individuals at risk for Huntington Disease (HD). People with Huntington's have an expected life length of about 60 years. The risk of acquiring Huntington's is genetically inherited, but the first symptoms usually appears firsts at ages between 35 and 50. The researchers investigate whether children who learn that they carry the HD mutation differ in their human capital investments compared to those who learn that they do not. They find that children who learn that they carry the HD mutation are less likely to attend college while the level of high school completion is unaffected. Since the treatment and control groups are likely to share the same unobservable characteristics, the validity of their findings can be considered relatively robust.

All of the above papers have one thing in common, they use *schooling lenght* to measure the level of human capital. Assuming that length of schooling is a good determinant of labor market performance, it is consistent to argue that, since HIV/AIDS decreases the average length of schooling, HIV/AIDS by extension also decreases labor market efficiency in the long run. However, it is not clear that such an assumption is perfectly valid.

In addition to what is suggested by the human capital theory, which supports the assumption above, another theory regarding the connection between schooling and labor market performance exists. This theory, often referred to as the 'signaling theory' argues that the level of schooling is only an effect of the individuals innate ability, which is determined at birth and impossible to change. The level of education depends on this ability since it defines how long a certain individual manages to stay in school. Hence, the level of education is not valuable for performance in itself. It simply serves as a *signal* for *ability* which, in turn, determines labor market performance.

Viewed in the light of signaling theory, it becomes clear that the length of schooling measure may not be perfectly suitable if one wants to investigate the effects of HIV/AIDS on labor market efficiency in the long run. If a general increase in HIV prevalence decreases length of schooling, but leaves innate individual ability unchanged, the conclusion that HIV/AIDS has a bad impact on productivity in the long run may be false.

The credibility of this suspicion increases when seen in the light of research by Young (2005), who finds that a general increase in HIV prevalence is associated with a higher amount of GDP per capita, partly due to the fact that increased mortality rates means fewer people. Provided that aggregated GDP is unchanged, fewer people means higher GDP per capita. A high level of aggregated GDP, despite a rise in HIV prevalence, could be explained by the fact that the labor market manages to stay productive.

On this background, I contribute by investigating whether a high prevalence of HIV is associated not only with a decline in the length of schooling, but if it also decreases schooling *achievement* for those who are still in school.

3 Theory

The theoretical argumentation behind the hypothesis of this paper, that an increase in HIV/AIDS prevalence lowers schooling achievement at a regional level, is based on the implications of the Ben-Porath (1967)-model of human capital and the life cycle of earnings. Essentially, this model argues that every individual is born with at least some amount of human capital, which in turn can be used as an input to produce two things, earnings or further human capital. Depending on how efficient the individual is in the production of human capital (e.g. different individuals require different amounts of initial human capital to produce the same amount of further human capital) she chooses her level of schooling in a way that maximizes total earnings over the life cycle. Hence, there is a trade of between production of earnings and production of human capital. More human capital certainly yields a higher wage, but the production of more human capital is costly in terms of forgone earnings (e.g. time spent in school could alternatively be used as time producing earnings). Ultimately, an individual choosing to invest all human capital in education, deciding never to quit school, would end up with a zero amount of life time earnings, despite the fact that she would be very productive if ever entering the labor market. Alternatively, an individual choosing to invest all her initial amount of human capital in the production of earnings, would end up with a relatively low wage stream accumulating earnings over a long period of time. The model can be formalized into the following objective function, which individuals will try to maximize:

$$max \int_0^T e^{-r_i t} Y_{it} \, \mathrm{d}t \tag{1}$$

Meaning that an individual living between point 0 and point T, with earnings Y_{it} and a discount rate r_i, will chose the amount of schooling, t^*, that maximizes total earnings over the period 0 to T.

Earnings, Y_t, is defined as:

$$Y_{it} = R(H_{it} - K_{it}) \tag{2}$$

where R is the rental rate of one unit of human capital, H_{it} is the amount of capital used to produce earnings and K_{it} is the amount of human capital used as input to produce further human capital.

In the context of this model, it is likely that an increase in HIV rate impact the schooling decision through two channels. Firstly, HIV rates affect expected life length, thereby affecting the level of T in equation 1. Secondly, it is possible that the prevalence of HIV changes the general capability to assimilate human capital. Assume that an increase in regional HIV prevalence would decrease the general quality of schooling in these regions, for example by increasing the amount of sick days for infected teachers. This would be reflected in the model as a decrease in the individual efficiency to produce human capital, H_{it}, from existing human capital, K_{it}, which would lower the value of actual earnings, Y_t, Thereby affecting the decision of schooling.

Connecting this reasoning to previous findings, and to the outset of my own contributions, it is clear that most of the previously conducted research focuses on the first channel. I, instead, aim to investigate the validity of the second channel, arguing that a general decrease in the ability to produce human capital should be reflected as a general decrease in schooling achievement. This relies on the assumption that the underlying ability to produce human capital should, on average, be equal across groups.

4 Data and Descriptive Statistics

The empirical analysis draws on data from The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and the Demographics and Health Surveys (DHS). This chapter begins with a description of each of these sources separately. Then follows a detailed account of how I exploit overlap between these data sets in order to associate individual level data in SACMEQ with regional characteristics calculated from DHS.

4.1 The SACMEQ Assessments

The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) is a network of 15 Sub-Saharan African Ministries of Education and the UNESCO International Institute for Educational Planning (IIEP). The main activity of SACMEQ is to carry out regular assessments of the math and reading knowledge of 6th-grade primary school students and their teachers. As part of these assessments, information on student, teacher and school characteristics is gathered via questionnaires.

Three waves of SACMEQ assessments have been conducted to date. The most recent of these was carried out during 2007 and apart from tests in math and reading, this wave also contains additional tests measuring knowledge of health, focusing on HIV. Since this paper investigates the effects of HIV on schooling, I restrict the analysis using data only from this last wave, which contains nationally representative samples of 6th-grade students from 15 countries. In total, the dataset contains 56,671 students representing 4,773 classes from 2,528 schools.

The main outcome variables of interest for this thesis are students' test scores from the math, reading and HIV tests. The math and reading tests measure students' knowledge of the 6th-grade math and reading curricula, and are described in detail in Bietenbeck et al. (2015). The HIV knowledge tests measures students' knowledge about the ways of prevention, spread of and risk factors/consequences of the disease. All tests are graded centrally by SACMEQ and test scores are standardized into a mean of 500 and standard deviation of 100. Here, I transform these scores into having a mean of 0 and a standard deviation of 1. The reason for this is ease of interpretation, as estimated coefficients then can be interpreted in terms of percentages of a standard deviation from the mean.

In addition to the individual test scores, SACMEQ contains information about the regional location of all schools. Regions are defined according to the ISO 3166-2 standard and consists of the principal subdivision (e.g., provinces or states) for each country. This is essential for the forthcoming empirical analysis, as it allows me to associate regions in SACMEQ with corresponding regional variable in the DHS data.

4.2 The Demographics and Health Surveys

The Demographic and Health Surveys (DHS) is a series of surveys carried out periodically in less developed countries of the world. The standard DHS surveys provides data monitoring indicators of population, health and nutrition. Apart from these standard surveys, the DHS Program supports a variety of additional data collection options monitoring indications regarding for example AIDS, Malaria and more. Data from DHS forms a representative sample of individuals at a regional level, containing information of individual characteristics like religious view, etc. Since 2001, the DHS program has been carrying out HIV tests as part of the standard survey. Before this, reliable sources of HIV prevalence in developing countries were rare.

For this thesis, the variables of interest are, in general, such variables that may determine schooling outcomes at a regional level. The purpose of the thesis naturally causes HIV prevalence to be especially interesting. Consequently, I exploit information from DHS in order to calculate mean values of regional HIV prevalence. DHS reports information on several regional subdivisions, one corresponding to the same province or state that is reported in the SACMEQ data. Consequently, regional characteristics from DHS are calculated at the same regional level that is used in SACMEQ. As I will address possible endogeneity in the HIV rate variable by using male circumcision rate as an instrument, this variable is also of certain interest. Additionally, mean values of education, access to electricity and share of population being christian-protestant or non-religous are included in some of the regressions as controls.

4.3 Preparation and Sample Selection

Exploiting the fact that both SACMEQ and DHS associates each observation with a sub divisional region, I am able to measure regional characteristics at a relatively fine level, linking each student in SACMEQ with regional characteristics of the area in which he or she attends school. This in turn, allows me to investigate whether higher prevalence of HIV is in general associated with lower levels of schooling. The availability of reliable data regarding HIV rates in Africa has historically been poor, and the ability to associate such data with a reliable measure of schooling achievement has not existed until recently. To my knowledge, this is the first attempt to estimate the effect of HIV-prevalence on schooling achievement using this type of detailed data.

Even though the process of combining the observations in SACMEQ with the regional characteristics obtained from DHS may seem straightforward at first, several aspects has to be considered in order to obtain a matched data set.

First of all, there is no perfect overlap between countries included in SACMEQ and countries included in DHS. As DHS data about HIV prevalence is not available for all 15 countries investigated within SACMEQ, I am limited to use only data from countries which are available in both data sets.

Additionally, it is necessary that that both data sets corresponds to the same time period, at least to some extent. SACMEQ surveys for the third wave were all carried out during 2007. DHS Surveys are carried out repeatedly at intervals of around five years. The years vary between countries and I choose to restrict my sample to use only DHS data collected at a maximum of plus minus three years from 2007.

Consequently, out of the 15 countries initially surveyed withing SACMEQ, I restrict my sample to include these seven countries: Kenya, Lesotho, Malawi, Mozambique, Tanzania, Zambia and Zimbabwe. [3] A detailed description regarding included countries, and which wave of DHS that is used, is presented in table 7. Together, they make out 70 different regions which are used to calculate different levels of HIV prevalence.

The propensity of the sample forms two main threats to my estimation strategy. Firstly, the relatively small sample of different regions heavily decreases variation in my sample. Even though the number of observations in SACMEQ is large, the fact that all these observations are associated with one of these 85 regions may mitigate the importance of each individual observation. This could mean that the sample may be too small in order achieve statistically significant results. However, it will not bias the estimates of my regressions. Secondly, the fact that the measure of HIV prevalence does not exactly correspond to the same time point as the SACMEQ assessments

[3] The following eight countries are excluded from the sample: Botswana, Mauritius, Namibia, Seychelles, South Africa, Swaziland and Uganda.

could potentially bias the results. Assuming that the average HIV prevalence increases for all regions over time, this could bias the estimated results. If so, I would overestimate the true HIV prevalence for countries surveyed after 2007 and underestimate true HIV prevalence in countries surveyed before 2007. However, as HIV prevalence varies more heavily between regions than over time (UNAIDS, 2014) this problem is assumed to be of negligible importance.

After imposing the sample restrictions described above, and dropping variables that contains missing values for at least one of the relevant variables, the dataset consists 21,819 students of which 10,622 are boys and 11,197 are girls. As mentioned earlier, the test scores for the tests of maths, reading and health knowledge are the main outcome variables of interest in the forthcoming analysis. Table 1 presents descriptive statistics for all relevant variables, and shows that the scores on these test varies at a rate of around ±5, after transformed into having mean 0 and standard deviation 1. Furthermore, I present summary statistics for five other variables in SACMEQ, determined either at school or individual level. This includes characteristics such as sex, indicated as a dummy taking value one if the student is a girl and zero otherwise ('Girl'); age in years ('Age'); Paternal education in years ('Fathers education') and total number of possessions owned by the student ('№of possessions') [4]. Furthermore, the ratio between number of students and number of teachers in included as a proxy quality in education for schools ('Pupil-teacher ratio').

All of the above variables from SACMEQ, that are not test scores, are used as additional controls in the regressions presented in section 6. Number of possessions is included mainly as a proxy of individual socioeconomic conditions.

Additionally, table 1 presents summary statistics for for variables that are calculated from DHS. HIV prevalence serves as the main variable of interest, and the fact that it varies at a standard deviation of 6.5% suggests that there is enough regional variance in my sample to estimate the impact of HIV prevalence on schooling achievement. Furthermore, male circumcision rates are calculated for each region. This variable will serve as an instrument for HIV prevalence and the econometric logic behind this will be presented in

[4] This variable is measured in levels, but takes a maximum value of one. Consequently, students with 32 possessions are reported as owning only 31.

section 5. Measurements of the average regional share of homes with access to electricity together with the regional share of people that are christian-protestant or non-religious ('Religion') are presented, as they will serve as additional control variables for some of the regressions presented in section 6.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.
Math test score	0	1	-5.482	6.5
Reading test score	0	1	-4.523	4.776
Health test score	0	1	-4.674	5.155
Girl	0.513	0.5	0	1
Age	13.919	1.808	9.75	26.917
Fathers education	5.605	3.085	0	12
№ of possessions	8.040	4.731	0	31
Pupil-teacher ratio	54.519	36.908	1.039	732.5
HIV prevalence	0.125	0.065	0.023	0.248
Circumcision rate	0.513	0.325	0.02	1
Religion	0.682	0.281	0.145	1
Has electricity	0.201	0.194	0.005	0.99
N		21,819		

5 Methodology

To explain the relationship between regional HIV prevalence and schooling achievement in primary school, I assume that the relationship can be formalized by the following linear equation:

$$T_{isj} = \alpha + \beta HIV_j + X_i'\gamma + R_j'\delta + S_s'\rho + \epsilon_{isj} \tag{3}$$

Were T_{isj} denotes the schooling achievement variable of interest for a pupil i at school s, situated in region j. α is the intercept. X_i is a vector of observable individual characteristics, such as the parental level of education. S_s is a vector of school specific characteristics, such as pupil-teacher ratio. ϵ_{isj} denotes the error term. Since available data measures HIV rate at a regional level, β is the coefficient of interest.

Estimates of the coefficients in equation 3 above can theoretically be achieved by conducting a regular OLS regression. However, it is practically impossible to identify all relevant characteristics that are part of vectors X_i, R_j and S_s. As a consequence, such estimates are likely biased compared to the true values (e.g. omitted variable bias). An example of such bias, in the context of this subject, arise if one fails to control for example for *poverty*. Assume that a higher levels of poverty are associated with a general decrease of schooling achievement, and that regions with high prevalence of HIV are in general poorer. Carrying out an OLS regression estimating the effect of HIV prevalence on schooling achievement without controlling for the underlying aspect of poverty would then suggest that the impact of HIV on schooling achievement is bigger than it actually is. On the other hand, other factors may work in the other direction. An example of such a factor could be the general level of education in a region. Children of well educated parents may be more likely to perform better in school, but there is evidence that highly educated people are also more likely to be infected with HIV (Fortson, 2008). Hence, it is not unlikely that failing to control for regional education would mitigate the results of an OLS regression aiming to estimate the impact of HIV prevalence on schooling achievement.

To address this problem of possible endogeneity in the HIV rate variable, I use an instrumental variables approach. For this strategy to be successful, the instrument needs to fulfill two conditions; it has to be correlated with the endogenous regressors which I aim to treat, but uncorrelated with the outcome variable. As mentioned earlier, I instrument HIV prevalence with male circumcision rates [5], both calculated from DHS in terms of prevalence at a regional level.

Male circumcision has been shown to significantly decrease the risk of infection and transformation of HIV (Auvert et al., 2005). Hence, relevance condition seems theoretically valid. The exclusion restriction, implies that there should be no independent relationship between schooling achievement of individual pupils, and the circumcision rate in the area in which he or she lives. Such a threat to my strategy arise if religious faith impacts schooling achievement. As people of certain religions may be more likely to be circumcised, and also differ in their general schooling achievement propensity. Robustness checks will be conducted and presented in the following section, in order to control for this possibility.

[5] This instrument has previously been used by Werker et al. (2007) and Marinescu (2014).

It should be mentioned that other possible instruments of HIV prevalence may exist. Oster (2012) instruments regional HIV prevalence with the average distance to the origin of the HIV virus. She finds that HIV prevalence is indeed correlated with this distance, and argues that the spreading of HIV is slow enough to validate such conclusions. However, since Osters' analysis uses data over other African countries that those that are part of Sub-Saharan Africa, the variance in the distance to HIV-origin is much greater than corresponding numbers in this paper. Early attempts of instrumenting HIV prevalence with this distance, were shown not to be successful. Therefore abandoning this instrument in favor of circumcision rates.

The setup and compounds of the IV strategy follows the functional form of equations 4 to 6.

The first stage of the IV strategy estimates values of the endogenous HIV rate variable by regressing HIV rate on the same regressors that are part of equation 3. In addition to this, the current instrumental variable is also included as a regressor in this first stage. This first stage setup can be described by the following equation:

$$HIV_j = \phi i_j + X_i'\gamma + R_j'\delta + S_s'\rho + \epsilon_{isj} \tag{4}$$

The notation in this equation is almost analogous to the one in equation 3. Apart from the fact that the coefficients, despite being notated with the same Greek letters, now reflect other true values.

The instrument varies at a regional level and is denoted by i_j. ϵ_{isj} denotes the error term.

Carrying out an estimation on the setup in equation 4 leads to the following estimated model:

$$H\hat{I}V_j = pi_j + X_i'G + R_j'D + S_s'R \tag{5}$$

The second stage of the IV strategy follows the same argumentation as the one described in equation 3. However, instead of using the endogenous HIV rate variable HIV_j directly, HIV_j is substituted with the corresponding prediction from the first stage regression, $H\hat{I}V_j$. Hence, the final IV estimation is given by:

$$\hat{T_{isj}} = a + bH\hat{I}V_j + X_i'G + R_j'D + S_s'R \tag{6}$$

As long as the predicted value $H\hat{I}V_j$ is exogenous with respect to the outcome variable, this estimation strategy should reduce bias that arise from omitted variables. However, this exclusion restriction is only satisfied as long as the instrument variable i_j significantly explains variation in HIV_j, while being uncorrelated with the outcome variable T_{isj} (apart from the indirect link through HIV_j). Equation 6 explains why the two conditions of the instrumental variable, the relevance condition and the exclusion restriction, has to be fulfilled.

6 Results

This section is divided into four subsections. The first of these presents the main estimates of carrying ot the methodology described in section 5. The second one investigates whether the estimated effects are heterogeneous between boys and girls. Thereafter follows a section of additional robustness test. Finally, I investigate one possible channel through which HIV prevalence may affect schooling achievement. The results of this investigation is presented in the last section.

6.1 Main Results

Table 2 presents the result of the first stage in the two stage least square strategy described in section 5. The estimated coefficients from this regressions are used in order to predict values of HIV prevalence that are exogenous with respect to schooling achievement.

Circumcision-rate is negative and significant for all specifications. Suggesting that the first instrumental requirement holds; circumcision-rate is correlated with HIV rate. Furthermore, HIV prevalence seems uncorrelated with all included controls apart from the instrument. One exception is *age* which is negatively related to HIV rates. The reason behind this is intuitive; a comparably higher HIV-rate means higher mortality rates. Since the likelihood of dying from HIV increases with age, high prevalence of HIV creates a skewed age distribution, with a relatively higher share of younger people.

Table 3 summarizes the estimated results from the fully specified model including all relevant, available controls described in section 4. The table includes both OLS and IV estimations for the standardized test scores in mathematics, reading and health respectively. [6]

The estimated impact of HIV on schooling achievement is negative and significant for all three tests. IV estimates are consistently more severe than corresponding OLS estimates, suggesting that there is an upwards bias of some unknown variable, which attenuates the direct impact of HIV on schooling. The interpretation of the IV estimates suggests that an increase in HIV prevalence of 1% decreases test scores of about 1.1%, 1.2% and 0.56% of a standard deviation for math, reading and health respectively. Additionally, all included controls have the expected sign.

An interesting aspect of the results is the fact that the estimated effect of HIV prevalence is substantially lower for health test score compared to the other two tests. One explanation of this could be that students achieve knowledge of HIV from other sources apart from school, decreasing the effect of HIV prevalence on schooling quality. Another possible explanation could be that regions with high prevalence of HIV are more concerned about their students HIV knowledge, thereby putting extra effort into the subject of HIV knowledge. If so, my findings would suggest that such efforts seem to be successful.

The main conclusion from these findings is that there is indeed evidence that HIV prevalence decreases schooling achievement, consequent with the suggestions of the Ben-Porath (1967) model. Furthermore, it is possible to argue that these findings are in line with the findings of Fortson (2011), even though they are not directly related to each other. In order to strengthen the validity of these findings and investigate this question in a more extensive way, I will now present some additional estimates from the same dataset.

6.2 Hetrogeneity and Robustness Checks

Table 3 shows that the estimated impact of being female is negative and significant for all three tests. In the light if this, it is interesting to know if the investigated determinants of schooling achievement works different between boys and girls. Consequently, I have also carried out regressions separately for boys and girls. Table 4 shows the estimated coefficients for determinants of schooling achievement separately for boys and girls. All these estimations are IV-estimates. The results of these regressions are

[6] More detailed tables showing estimations with successively included controls, for each subject separately, are provided in tables 8 to 13 of the appendix.

presented in table 4.

This table reveals the fact that the impact of HIV prevalence on schooling achievement seems to be a bit more severe for boys compared to girls. However, the differences in magnitudes are very small. In reality, both genders seem to suffer as much from the impacts of HIV prevalence.

One interesting side effect of these regressions is that they show significant results of the same magnitudes, even though sample size is divided in half. Strengthen the validity of my findings. However, since the main limitation of my data is the limited number of regions rather than the amount of individual observations, this aspect of robustness should not be overrated.

To further investigate the level of robustness in my estimates, I perform a number of additional regressions that include more control variables. Table 5 presents estimates of the same type as the main results, but with the addition of two additional controls determined at the regional level. The first of these, titled 'Has electricity', describes the regional share of people having access to electricity at home, serving as a proxy for poverty level. The second variable, titled 'Religion', reflect the share of people that identifies themselves as either christian-protestants or non-religious. The reason behind adding this variable is that I assume that people of these religions are relatively less likely to be circumcised than others. As discussed in the methodology [7] it could threaten my design if people of certain religions have different propensities regarding schooling achievement. In the light of this, it is interesting to see if the estimates are still robust after controlling for religious characteristics.

As seen in the table (5), estimates of HIV prevalence are almost unchanged compared to the estimates of the main results in table 3. OLS estimates have slightly larger absolute values, suggesting that the amount of bias compared to the IV estimates, has decreased. The IV estimates of the impact of HIV prevalence of schooling achievement stays stable at rates of around 1.1% for maths and reading, and 0.5% for health test score results. Even though this additional test of robustness is no guarantee for an unbiased estimate, close to the true impact of HIV prevalence, it does increase the strength of my results and their internal validity within 6th grade student of Sub-Saharan Africa.

[7] See section 5

6.3 Possible Channels between HIV and Schooling Achievement

Even if my results propose a causal relationship between HIV prevalence and schooling, they provide no information regarding the mechanisms through which this relationship works. One example of such a chain of events, would be if children living in areas of relatively high rates of HIV are in general more likely to take care of sick family members. If so, one can argue that children living in these areas have less time to focus on school work, since the process of care taking is costly in terms of time. SACMEQ contains information of such data in the form of av dummy variable indicating whether or not children are engaged in the caretaking of sick family members. This allows me to regress the prevalence of HIV on the likelihood that an individual does engage in caretaking. Even if this is far from the only channel in which the relationship between schooling achievement and HIV prevalence works, I am able to accept or reject the relevance of this hypothesis.

Table 6 presents OLS estimates of a logit model, regressing the probability of being engaged in caretaking on regional HIV prevalence. The estimated effect is positive but insignificant. Hence, these findings show no evidence that a decrease in schooling achievement due to increases in HIV prevalence can be explained by the fact that students are more likely to take care of their sick parents. However, these estimates are quite trivial, and to be able to reject the validity of this channel with certainty, further research is required.

Table 2: First Stage Regression - Dependent Variable: HIV-rate (in logs)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES							
Circumcision rate (in logs)	-0.219	-0.213	-0.213	-0.213	-0.216	-0.221	-0.221
	(0.0501)	(0.0490)	(0.0490)	(0.0484)	(0.0485)	(0.0480)	(0.0480)
Age		-0.0314	-0.0308	-0.0264	-0.0243	-0.0212	-0.0212
		(0.0107)	(0.0111)	(0.0104)	(0.00995)	(0.00971)	(0.00971)
Girl			0.0106	0.0103	0.0121	0.0119	0.0119
			(0.0145)	(0.0145)	(0.0141)	(0.0139)	(0.0139)
Fathers education				0.0144	0.0131	0.0115	0.0115
				(0.00425)	(0.00390)	(0.00355)	(0.00355)
№ of possessions (in logs)					0.0348	0.0271	0.0271
					(0.0257)	(0.0238)	(0.0238)
Pupil-teacher ratio						-0.00143	-0.00143
						(0.00126)	(0.00126)
Constant	-2.469	-2.025	-2.038	-2.180	-2.272	-2.219	-2.219
	(0.106)	(0.178)	(0.182)	(0.169)	(0.160)	(0.168)	(0.168)
Observations	21,819	21,819	21,819	21,819	21,819	21,819	21,819
R-squared	0.151	0.161	0.161	0.166	0.167	0.175	0.175

Robust standard errors in parentheses

Table 3: Main Results - Summary

	Mathematics		Reading		Health	
VARIABLES	OLS	IV	OLS	IV	OLS	IV
HIV prevalence (in logs)	-0.641	-1.105	-0.767	-1.193	-0.514	-0.557
	(0.0605)	(0.235)	(0.0636)	(0.252)	(0.0665)	(0.204)
Age	-0.0721	-0.0898	-0.0831	-0.0994	-0.0119	-0.0135
	(0.0104)	(0.0106)	(0.0108)	(0.0110)	(0.0105)	(0.0115)
Girl	-0.180	-0.178	-0.0670	-0.0653	-0.0859	-0.0857
	(0.0266)	(0.0287)	(0.0192)	(0.0217)	(0.0211)	(0.0211)
Fathers education	0.0292	0.0352	0.0343	0.0399	0.0204	0.0210
	(0.00525)	(0.00546)	(0.00511)	(0.00568)	(0.00505)	(0.00602)
№ of possessions (in logs)	0.212	0.216	0.234	0.237	0.138	0.138
	(0.0299)	(0.0336)	(0.0339)	(0.0373)	(0.0273)	(0.0273)
Pupil-teacher ratio	-0.00354	-0.00388	-0.00353	-0.00384	0.000186	0.000154
	(0.000845)	(0.00104)	(0.000876)	(0.00106)	(0.000979)	(0.00102)
Constant	-0.708	-1.523	-0.973	-1.721	-1.328	-1.403
	(0.207)	(0.533)	(0.220)	(0.536)	(0.208)	(0.415)
Observations	21,819	21,819	21,819	21,819	21,819	21,819
R-squared	0.204	0.131	0.270	0.208	0.100	0.099
First stage F-stat		16.71		16.71		16.71

Robust standard errors in parentheses

Table 4: Heterogeneity - Separate IV Estimates for Boys and Girls

	Mathematics		Reading		Health	
VARIABLES	Boys	Girls	Boys	Girls	Boys	Girls
HIV prevalence (in logs)	-1.234	-0.980	-1.236	-1.155	-0.506	-0.611
	(0.228)	(0.244)	(0.231)	(0.274)	(0.205)	(0.214)
Age	-0.0765	-0.106	-0.0812	-0.123	0.000339	-0.0324
	(0.0112)	(0.0127)	(0.0101)	(0.0144)	(0.0117)	(0.0134)
Fathers education	0.0361	0.0332	0.0383	0.0404	0.0246	0.0170
	(0.00568)	(0.00603)	(0.00556)	(0.00664)	(0.00671)	(0.00641)
№ of possessions (in logs)	0.208	0.227	0.220	0.256	0.116	0.160
	(0.0300)	(0.0404)	(0.0350)	(0.0428)	(0.0324)	(0.0272)
Pupil-teacher ratio	-0.00323	-0.00456	-0.00334	-0.00437	0.000622	-0.000361
	(0.000918)	(0.00124)	(0.000948)	(0.00122)	(0.00120)	(0.000867)
Constant	-2.029	-1.172	-2.061	-1.392	-1.489	-1.344
	(0.563)	(0.504)	(0.534)	(0.538)	(0.465)	(0.381)
Observations	10,622	11,197	10,622	11,197	10,622	11,197
R-squared	0.146	0.119	0.231	0.192	0.102	0.092
First stage F-stat	18.11	15.29	18.11	15.29	18.11	15.29

Robust standard errors in parentheses

Table 5: Robustness Check

	Mathematics		Reading		Health	
VARIABLES	OLS	IV	OLS	IV	OLS	IV
HIV prevalence (in logs)	-0.682	-1.093	-0.835	-1.168	-0.558	-0.536
	(0.0687)	(0.189)	(0.0637)	(0.180)	(0.0613)	(0.176)
Age	-0.0662	-0.0787	-0.0779	-0.0880	-0.0124	-0.0117
	(0.0106)	(0.0109)	(0.0113)	(0.0116)	(0.0111)	(0.0116)
Girl	-0.182	-0.181	-0.0747	-0.0743	-0.0949	-0.0949
	(0.0260)	(0.0272)	(0.0187)	(0.0201)	(0.0216)	(0.0214)
Fathers education	0.0290	0.0341	0.0357	0.0399	0.0228	0.0225
	(0.00530)	(0.00509)	(0.00498)	(0.00497)	(0.00471)	(0.00557)
№ of possessions (in logs)	0.175	0.161	0.187	0.176	0.120	0.120
	(0.0231)	(0.0263)	(0.0261)	(0.0282)	(0.0224)	(0.0229)
Pupil-teacher ratio	-0.00356	-0.00385	-0.00362	-0.00386	6.53e-05	8.12e-05
	(0.000794)	(0.000934)	(0.000896)	(0.000979)	(0.000828)	(0.000853)
Has electricity	0.592	0.855	0.802	1.015	0.371	0.357
	(0.266)	(0.341)	(0.234)	(0.288)	(0.168)	(0.225)
Religion	0.207	0.311	0.516	0.600	0.484	0.479
	(0.150)	(0.164)	(0.125)	(0.147)	(0.125)	(0.130)
Constant	-1.064	-1.918	-1.612	-2.305	-1.787	-1.741
	(0.287)	(0.528)	(0.272)	(0.467)	(0.207)	(0.430)
Observations	21,819	21,819	21,819	21,819	21,819	21,819
R-squared	0.216	0.161	0.305	0.269	0.120	0.120
First stage F-stat		17.27		17.27		17.27

Robust standard errors in parentheses

Table 6: The Channel of Caretaking

VARIABLES	(1) OLS
HIV prevalence (in logs)	0.164
	(0.112)
Age	0.0827
	(0.0142)
Girl	-0.0294
	(0.0371)
Fathers education	-0.0116
	(0.00631)
№ of possessions (in logs)	0.191
	(0.0415)
Pupil-teacher ratio	0.000395
	(0.000818)
Constant	-1.358
	(0.304)
Observations	21,819

Robust standard errors in parentheses

7 Discussion

I suggest that the empirical strategy carried out above may fail mainly due to two alternate reasons. First of all, the strategy requires that there is a direct and significant relationship between the endogenous regressor, in this case the HIV-rate variable, and the instrument, here being male circumcision rate in the same region. It has already been stated, in connection to table 2, that there is indeed a significant relationship between these two variables. However, it can be the case that this relationship is not strong enough. If so, the standard errors of the estimated coefficients cannot be trusted, as it makes estimators inconsistent. They will however, in theory, still be unbiased. This means that using significance tests such as the t-test as proof of a causal relation is not a very good idea. Estimated standard error can be either bigger or smaller than what the results suggest.

Since all estimated coefficients for the impact of HIV prevalence on schooling are robust, in the sense that estimated standard errors are small.

The standard errors are allowed to increase substantially, without making estimated results insignificant. However, the bias of the estimated covariance matrix can always be big enough to imply an untrue significance, provided that the instrument is weak enough. Since their is no econometric solution to this problem other than adding additional data, which I do not have, I have to rely on the assumption that the instrument that I use is a good enough predictor for the endogenous regressor that I try to address in the first place. I base the validity of this assumption on the fact that this instrument has been used in previous published papers that has been subject to a thorough review process.

Furthermore, the exclusion restriction of the instrument can, and will, always be a big subject of discussion when addressing endogeneity with instrumental variables. Since it is impossible to formally test whether the exclusion restriction holds or not, the only way to persuade oneself that it does hold is by using pure argumentation. In the context if this paper, the question that has to be asked can be formulated as follows: “Is there any way that schooling test results can be affected by the circumcision rate in the region in which the school is situated, apart from the fact that circumcision decreases HIV-rates, which in turn affects schooling achievement?” For my strategy to work, the answer to this question should be “no”.

At first sight, It may seem counter intuitive that the prevalence of circumcised men in your area should affect your achievement in school. However, taking a more detailed look, it is possible to find at least some explanations to why this could actually be the case.

Looking at the underlying causes for why men are circumcised or not. I identify three possible causes: Hygienic/health-related, aesthetic, and/or religious reasons.

People choosing to circumcise themselves or their sons due to hygienic reasons may be different from people choosing not to. It could be the case that, in areas with relatively high HIV-prevalence, well informed, highly educated people are more likely to get circumcised than others because they want to decrease their risk of being infected with HIV. If so, the children of these people may perform better at the SACMEQ-tests compared to others, because of the fact that they are helped by their well-educated parents while studying. Thereby providing a threat to my instrumentation strategy. However, I do control for this to some extent by including parental education as an additional control variable in my regressions. Additionally, there are multiple other health related reasons behind circumcision. For example, one of the most common treatments for *Phimosis* is circumcision. It seems unlikely however, that men suffering from Phimosis differ in schooling achievement compared to those who do not.

For those who chose to circumcise themselves or their children due to aesthetic reasons, I cannot identify any obvious reasons for why these persons would differ in ability compared to others choosing not to. In this case it is simply a matter of aesthetic taste, which I believe to be uncorrelated with schooling outcomes.

The third and final cause, religion, is the one that I find most likely to threaten the validity of my instrumentation strategy. People of certain religious views may have a general preference towards circumcision, and the moral implications in this religious view may provide a preference towards or against the importance of schooling and achievement in school. The severity of this factor and its impact on my findings is impossible to estimate. However, the fact that I include the regional share of individuals that identifies themselves as Christian-Protestant or Non-Religious, as one of the regional characteristics controls for this at least partially.

In conclusion, the above discussion regarding the exclusion of my instrument can be developed to a far greater extent. And it is possible to find reasons and mechanisms that opposes the use of any instrument. My prospect is that the reasoning and argumentation provided above, together with the fact that this instrument has been used earlier in published research, is enough evidence to persuade the reader of the relevance of this instrument.

Provided that my instrumentation strategy is successful, the level of external validity in my results can still be discussed. It is clear from the nature of the data, that my findings are of most relevance to children attending school in Sub-Saharan Africa. The presence of HIV, for example in the US, is more or less negligible. Hence, it is unlikely that my findings would be relevant for a potential increase of HIV rates within the US. However, Sub-Saharan Africa is a large, dense region, and even though the level of external validity may be small, the findings of this paper can be applicable for a huge amount of people. Hopefully shedding some light on the specific conditions and challenges that these people, and their societies, meet.

8 Conclusion

The main conclusion of this paper is that HIV prevalence decreases schooling achievement, The effects are small but significant and assuming that the instrumentation strategy is successful, I am able to proof a causal relationship stating that a regional increase in HIV prevalence of 1% decreases general schooling achievement for children in Sub-Saharan Africa of more than 1% of a standard deviation form mean. However, the underlying assumptions of the estimation strategy can always be discussed. Consequently, my findings should be interpenetrated carefully. As is always the case with research, findings should be considered in terms of implications rather than in terms of proofs.

My findings are in line with those of Fortson (2011) and others, who concludes that HIV prevalence has a negative impact on level of education within societies. The fact that I am able to strengthen the validity of this conclusion, investigating another channel in which these mechanisms work, illustrate the complexity of the ways in which HIV affects societies.

If one wants to investigate how these effects may affect socioeconomic outcomes later in life, It depends on ones beliefs regarding signaling compared to human capital theory. Following the implications of the Ben-Porath (1967) model, the estimated decrease in schooling achievement will in the long term affect individual earnings not only by decreasing time spent in school, but also by decreasing the amount of human capital gained from this time. This in turn could provide a threat for the economy as a whole, as it decreases the efficiency and true potential of its residents. This new channel through which HIV plagues the Sub-Saharan countries, shows that the problem of HIV can be even more complex to solve for concerned policymakers. On the other hand, those relying more on signaling theory should not be as worried by the results of my analysis, since the results do not say anything about the effect of HIV on actually getting a degree.

Finally, my suggestion for further research within this field, would be to investigate the mechanisms behind the relationship between HIV prevalence and schooling achievement. As is briefly discussed in this paper, the SACMEQ dataset may contains some of the answers to this question. However, I leave it to others to look for and discover these answers.

References

Marcella M. Alsan and David M. Cutler. Girls' education and hiv risk: Evidence from uganda. *Journal of Health Economics*, 32(5):863–872, 2013.

Bertran Auvert, Dirk Taljaard, Emmanuel Lagarde, Joelle Sobngwi-Tambekou, Rémi Sitta, and Adrian Puren. Randomized, controlled intervention trial of male circumcision for reduction of hiv infection risk: the anrs 1265 trial. *PLos med*, 2(11):e298, 2005.

Yoram Ben-Porath. The production of human capital and the life cycle of earnings. *The Journal of Political Economy*, pages 352–365, 1967.

Jan Bietenbeck, Marc Piopiunik, and Simon Wiederhold. Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance? 2015.

Anne Case and Cally Ardington. The impact of parental death on school outcomes: Longitudinal evidence from south africa. *Demography*, 43(3): 401–420, 2006.

David K Evans and Edward Miguel. Orphans and schooling in africa: A longitudinal analysis. *Demography*, 44(1):35–57, 2007.

Jane G Fortson. The gradient in sub-saharan africa: socioeconomic status and hiv/aids. *Demography*, 45(2):303–322, 2008.

Jane G Fortson. Mortality risk and human capital investment: The impact of hiv/aids in sub-saharan africa. *The Review of Economics and Statistics*, 93(1):1–15, 2011.

Matthew P Fox, Sydney Rosen, William B MacLeod, Monique Wasunna, Margaret Bii, Ginamarie Foglia, and Jonathon L Simon. The impact of hiv/aids on labour productivity in kenya. *Tropical Medicine & International Health*, 9(3):318–324, 2004.

James Habyarimana, Bekezela Mbakile, and Cristian Pop-Eleches. The impact of hiv/aids and arv treatment on worker absenteeism implications for african firms. *Journal of Human Resources*, 45(4):809–839, 2010.

Seema Jayachandran and Adriana Lleras-Muney. Life expectancy and human capital investments: Evidence from maternal mortality declines. *The Quarterly Journal of Economics*, 124(1):349–397, 2009.

Sebnem Kalemli-Ozcan. Does the mortality decline promote economic growth? *Journal of Economic Growth*, 7(4):411–439, 2002.

Sebnem Kalemli-Ozcan, Harl E Ryder, and David N Weil. Mortality decline, human capital investment, and economic growth. *Journal of Development Economics*, 62(1):1–23, 2000.

Bruce A Larson, Matthew P Fox, Sydney Rosen, Margaret Bii, Carolyne Sigei, Douglas Shaffer, Fredrick Sawe, Monique Wasunna, and Jonathon L Simon. Early effects of antiretroviral therapy on work performance: preliminary results from a cohort study of kenyan agricultural workers. *Aids*, 22(3):421–425, 2008.

Elizabeth D Lowenthal, Sabrina Bakeera-Kitaka, Tafireyi Marukutira, Jennifer Chapman, Kathryn Goldrath, and Rashida A Ferrand. Perinatally acquired hiv infection in adolescents from sub-saharan africa: a review of emerging challenges. *The Lancet infectious diseases*, 14(7):627–639, 2014.

Ioana Marinescu. Hiv, wages, and the skill premium. *Journal of health economics*, 37:181–197, 2014.

Bryan McCannon and Zachary Rodriguez. A lasting effect of the hiv/aids pandemic: Orphans and pro-social behavior. Working Papers 16-10, Department of Economics, West Virginia University, 2016.

Emily Oster. Hiv and sexual behavior change: Why not africa? *Journal of Health Economics*, 31(1):35–49, 2012.

Emily Oster, Ira Shoulson, and E Dorsey. Limited life expectancy, human capital and health investments. *The American Economic Review*, 103(5): 1977–2002, 2013.

Rodrigo R Soares. Mortality reductions, educational attainment, and fertility choice. *The American Economic Review*, 95(3):580–601, 2005.

UNAIDS. The gap report, 2014. available at: http://www.refworld.org/docid/53f1e1604.html [accessed 17 August 2016].

Eric Werker, Amrita Ahuja, and Brian Wendell. *Male Circumcision and AIDS: The Macroeconomic Impact of a Health Crisis.* 2007.

Alwyn Young. The gift of the dying: The tragedy of aids and the welfare of future african generations. *The Quarterly Journal of Economics*, pages 423–466, 2005.

Appendices

Table 7: Details of Data Sources

Country	DHS Wave	DHS Year	School Survey
Kenya	V	2008-09	2007
Lesotho	VI	2009	2007
Malawi	VI	2010	2007
Mozambique	V	2009	2007
Tanzania	VI	2010	2007
Zambia	V	2007	2007
Zimbabwe	VI	2005-06	2007

Table 8: Mathematics Test Score - OLS Regression

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIV prevalence (in logs)	-0.573	-0.613	-0.613	-0.629	-0.631	-0.641	-0.641
	(0.0748)	(0.0716)	(0.0714)	(0.0690)	(0.0661)	(0.0605)	(0.0605)
Age		-0.0951	-0.104	-0.0921	-0.0786	-0.0721	-0.0721
		(0.0142)	(0.0145)	(0.0125)	(0.0109)	(0.0104)	(0.0104)
Girl			-0.188	-0.188	-0.179	-0.180	-0.180
			(0.0267)	(0.0265)	(0.0269)	(0.0266)	(0.0266)
Fathers education				0.0405	0.0329	0.0292	0.0292
				(0.00626)	(0.00550)	(0.00525)	(0.00525)
№ of possessions (in logs)					0.233	0.212	0.212
					(0.0329)	(0.0299)	(0.0299)
Pupil-teacher ratio						-0.00354	-0.00354
						(0.000845)	(0.000845)
Constant	-1.278	-0.0438	0.177	-0.250	-0.851	-0.708	-0.708
	(0.193)	(0.294)	(0.296)	(0.262)	(0.229)	(0.207)	(0.207)
Observations	22,118	22,118	22,118	22,118	21,819	21,819	21,819
R-squared	0.114	0.144	0.153	0.168	0.187	0.204	0.204

Robust standard errors in parentheses

Table 9: Reading Test Score - OLS Regression

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIV prevalence (in logs)	-0.688	-0.736	-0.736	-0.754	-0.757	-0.767	-0.767
	(0.0726)	(0.0718)	(0.0718)	(0.0704)	(0.0671)	(0.0636)	(0.0636)
Age		-0.114	-0.118	-0.104	-0.0895	-0.0831	-0.0831
		(0.0146)	(0.0147)	(0.0127)	(0.0110)	(0.0108)	(0.0108)
Girl			-0.0758	-0.0766	-0.0660	-0.0670	-0.0670
			(0.0197)	(0.0190)	(0.0191)	(0.0192)	(0.0192)
Fathers education				0.0468	0.0380	0.0343	0.0343
				(0.00607)	(0.00527)	(0.00511)	(0.00511)
№ of possessions (in logs)					0.255	0.234	0.234
					(0.0367)	(0.0339)	(0.0339)
Pupil-teacher ratio						-0.00353	-0.00353
						(0.000876)	(0.000876)
Constant	-1.542	-0.0618	0.0273	-0.465	-1.115	-0.973	-0.973
	(0.196)	(0.305)	(0.303)	(0.273)	(0.241)	(0.220)	(0.220)
Observations	22,118	22,118	22,118	22,118	21,819	21,819	21,819
R-squared	0.166	0.208	0.210	0.230	0.253	0.270	0.270

Robust standard errors in parentheses

Table 10: Health Test Score - OLS Regression

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIV prevalence (in logs)	-0.492	-0.501	-0.501	-0.511	-0.515	-0.514	-0.514
	(0.0622)	(0.0632)	(0.0631)	(0.0644)	(0.0655)	(0.0665)	(0.0665)
Age		-0.0219	-0.0263	-0.0190	-0.0115	-0.0119	-0.0119
		(0.0117)	(0.0114)	(0.0112)	(0.0106)	(0.0105)	(0.0105)
Girl			-0.0930	-0.0934	-0.0859	-0.0859	-0.0859
			(0.0206)	(0.0204)	(0.0212)	(0.0211)	(0.0211)
Fathers education				0.0251	0.0202	0.0204	0.0204
				(0.00573)	(0.00526)	(0.00505)	(0.00505)
№ of possessions (in logs)					0.137	0.138	0.138
					(0.0286)	(0.0273)	(0.0273)
Pupil-teacher ratio						0.000186	0.000186
						(0.000979)	(0.000979)
Constant	-1.102	-0.817	-0.708	-0.972	-1.320	-1.328	-1.328
	(0.132)	(0.197)	(0.203)	(0.214)	(0.221)	(0.208)	(0.208)
Observations	22,118	22,118	22,118	22,118	21,819	21,819	21,819
R-squared	0.083	0.085	0.087	0.092	0.100	0.100	0.100

Robust standard errors in parentheses

Table 11: Mathematics Test Score - IV Regression

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIV prevalence (in logs)	-1.119***	-1.224***	-1.230***	-1.229***	-1.176***	-1.105***	-1.105***
	(0.315)	(0.292)	(0.292)	(0.276)	(0.261)	(0.235)	(0.235)
Age		-0.122***	-0.131***	-0.116***	-0.100***	-0.0898***	-0.0898***
		(0.0159)	(0.0160)	(0.0137)	(0.0119)	(0.0106)	(0.0106)
Girl			-0.185***	-0.186***	-0.177***	-0.178***	-0.178***
			(0.0299)	(0.0294)	(0.0296)	(0.0287)	(0.0287)
Fathers education				0.0490***	0.0404***	0.0352***	0.0352***
				(0.00644)	(0.00576)	(0.00546)	(0.00546)
№ of possessions (in logs)					0.239***	0.216***	0.216***
					(0.0366)	(0.0336)	(0.0336)
Pupil-teacher ratio						-0.00388***	-0.00388***
						(0.00104)	(0.00104)
Constant	-2.500***	-1.030	-0.824	-1.306**	-1.826***	-1.523***	-1.523***
	(0.732)	(0.682)	(0.691)	(0.654)	(0.597)	(0.533)	(0.533)
Observations	22,118	22,118	22,118	22,118	21,819	21,819	21,819
R-squared	0.011	0.016	0.022	0.045	0.086	0.131	0.131
First stage F-stat	16.29	15.58	15.60	15.56	15.76	16.71	16.71

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 12: Reading Test Score - IV Regression

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIV prevalence (in logs)	-1.203***	-1.323***	-1.326***	-1.324***	-1.263***	-1.193***	-1.193***
	(0.332)	(0.313)	(0.313)	(0.297)	(0.277)	(0.252)	(0.252)
Age		-0.140***	-0.144***	-0.127***	-0.110***	-0.0994***	-0.0994***
		(0.0169)	(0.0168)	(0.0143)	(0.0123)	(0.0110)	(0.0110)
Girl			-0.0738***	-0.0748***	-0.0639***	-0.0653***	-0.0653***
			(0.0237)	(0.0228)	(0.0226)	(0.0217)	(0.0217)
Fathers education				0.0548***	0.0451***	0.0399***	0.0399***
				(0.00681)	(0.00606)	(0.00568)	(0.00568)
№ of possessions (in logs)					0.260***	0.237***	0.237***
					(0.0402)	(0.0373)	(0.0373)
Pupil-teacher ratio						-0.00384***	-0.00384***
						(0.00106)	(0.00106)
Constant	-2.696***	-1.011	-0.929	-1.468**	-2.021***	-1.721***	-1.721***
	(0.765)	(0.688)	(0.693)	(0.663)	(0.604)	(0.536)	(0.536)
Observations	22,118	22,118	22,118	22,118	21,819	21,819	21,819
R-squared	0.073	0.090	0.090	0.119	0.166	0.208	0.208
First stage F-stat	16.29	15.58	15.60	15.56	15.76	16.71	16.71

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 13: Health Test Score - IV Regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES							
HIV prevalence (in logs)	-0.559***	-0.581***	-0.584***	-0.583***	-0.554***	-0.557***	-0.557***
	(0.213)	(0.221)	(0.221)	(0.218)	(0.214)	(0.204)	(0.204)
Age		-0.0255*	-0.0300**	-0.0219*	-0.0131	-0.0135	-0.0135
		(0.0146)	(0.0142)	(0.0131)	(0.0122)	(0.0115)	(0.0115)
Girl			-0.0927***	-0.0932***	-0.0858***	-0.0857***	-0.0857***
			(0.0207)	(0.0205)	(0.0213)	(0.0211)	(0.0211)
Fathers education				0.0261***	0.0208***	0.0210***	0.0210***
				(0.00705)	(0.00651)	(0.00602)	(0.00602)
№ of possessions (in logs)					0.137***	0.138***	0.138***
					(0.0291)	(0.0273)	(0.0273)
Pupil-teacher ratio						0.000154	0.000154
						(0.00102)	(0.00102)
Constant	-1.252***	-0.946**	-0.843**	-1.099**	-1.391***	-1.403***	-1.403***
	(0.473)	(0.402)	(0.408)	(0.442)	(0.453)	(0.415)	(0.415)
Observations	22,118	22,118	22,118	22,118	21,819	21,819	21,819
R-squared	0.082	0.083	0.084	0.091	0.099	0.099	0.099
First stage F-stat	16.29	15.58	15.60	15.56	15.76	16.71	16.71

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1