A Quantitative Theory of the HIV Epidemic: Education, Risky Sex and Asymmetric Learning*

Christian Alemán
ECARES
Université Libre de Bruxelles

Daniela Iorio
University of Bologna

Raül Santaeulàlia-Llopis
University of Pennsylvania
UAB, BSE and CEPR

November 3, 2022

Download the latest version HERE

Abstract

We show empirically that the relationship between education and the probability of HIV infection is U-shaped (positive-zero-positive) over the course of the epidemic. In contrast, the relationship between education and knowledge about the process of HIV infection follows an inverted U-shaped pattern. We develop a non-stationary quantitative macroeconomic theory with heterogeneous agents that is consistent with these facts. Our theory endogeneizes the entire course of the HIV epidemic across its different stages: a pre-HIV epidemic stage; a myopic HIV stage in which agents are not aware of the process of HIV infection; a learning stage in which agents heterogeneously—across education groups—learn about the process of infection; and an anti-retroviral (ARV) stage that modifies the effects of HIV infection on individuals. We show that asymmetric learning is key to reproduce both the micro patterns that we document and the aggregate evolution of the HIV epidemic. In further counterfactual experiments, we assess the effects of an early understanding of the virus and its mode of infection, improvements in the composition of education, the earlier (and universal) adoption of ARVs and the use of PrEP to prevent further spread.

*Raül Santaeulàlia-Llopis acknowledges financial support from the AGAUR 2020PANDE00036 "Pandemies" Grant 2020-2022, from the Spanish Ministry of Economy and Competitiveness, through the Proyectos I+D+i 2019 Retos Investigacion PID2019-110684RB-I00 Grant and the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S).

1 Introduction

In Sub-Saharan Africa (SSA) most HIV infections are due to heterosexual intercourse, and risky sexual behavior is one of the most relevant margins that policy intervention can affect [Behrman and Kohler, 2012, Greenwood et al., 2013, DePaula et al., 2014, Nyqvist et al., 2015]. If risky sexual behavior, and in turn HIV exposure, differs across education groups as the HIV epidemic evolves, then the timing of policy interventions targeted to specific educational groups is crucial for the effectiveness of these policies. Nowadays, however, the major international donors in the fight against HIV do not provide any guidance on specific targeting strategies across education groups.[1] This lack of policy advice could be explained by the fact that the current understanding of the sign and size of the relationship between education and HIV status lacks consensus [Beegle and de Walque, 2009]. That is, the knowledge of which education groups are at major risk of being infected with HIV remains unclear to scholarship, with a large body of mixed evidence that we review below. We provide a potential explanation that could reconcile the mixed evidence, departing from the observation that the existing works have, almost invariably, used data from different aggregate stages of the HIV epidemic, while the education gradient in HIV may vary over the evolution of the epidemic.

That the HIV epidemic in SSA evolves differently across countries and that these countries are at different stages of the HIV epidemic at any point in time is practically self-evident. In particular, we observe that the peak of HIV prevalence, the year of the HIV peak, the time it takes each country to reach its own peak, and the pace at which each country moves away from its peak differs greatly across SSA countries and over time. Based on these considerations, we propose an innovative unified macro framework that consists of a two-dimensional normalization of the HIV epidemic. Our definition of the stages of the HIV epidemic are analogous to the definition of stages of economic development [Lucas, 2004, Herrendorf et al., 2014] or the stages of the demographic transition [Galor and Weil, 2000, Lee, 2003, Greenwood et al., 2005]. In the context of HIV, our normalization adjusts for both the country-specific size of the epidemic (HIV prevalence rate) and the associated time paths of the epidemic (speed at which HIV epidemic

[1] According to Kates et al. [2011], $6.9 billion was given by donor governments to international AIDS assistance in 2010. The United States is the largest resource provider for the global fight against AIDS, and it channels its aid through the President's Emergency Plan for AIDS Relief (PEPFAR). Initiated by President George W. Bush for 2003-2008, PEPFAR has continued its activity under the mandate of President Barack Obama, who renewed the efforts for 5 years with few changes in policy implementation. However, they increased the amount of money—to about 50% in countries with a generalized epidemic— spent on preventing sexual transmission via abstinence, delay of age of first sexual intercourse, monogamy, fidelity, and reduction in the number of sex partners. More recently, new PEPFAR funds have been channeled to include "men who have sex with men, people who inject drugs, and sex workers", see the remarks of the Secretary of State John Kerry for the PEPFAR 10th annivarsary celebration at http://www.state.gov/secretary/remarks/2013/06/210770.htm. See also UNAIDS [2015].

reaches its peak and it moves away from it) in a comparable manner across countries. This way, our macro framework systematically defines aggregate stages of the HIV epidemic taking into account the large degree of cross-country heterogeneity in both the HIV prevalence rates and the speed at which the HIV epidemic evolves. We then use the heterogeneity in the stages of the HIV epidemic in 39 Demographic and Health Surveys (DHS) to document the stylized dynamic relationship between education and individual HIV status across the stages of the HIV epidemic.

Our main finding is that the education gradient in HIV follows a significant U-shaped (positive-zero-positive) pattern as the epidemic evolves. In particular, when individuals live in an economy that is at the early stages of the epidemic, the HIV-Education gradient is significantly positive and remarkably high: one additional year of education is associated with 1.12 percentage point increase in the probability of being HIV-positive. In other terms, completing five additional years of schooling doubles on average the likelihood of being infected (the HIV prevalence is 5.18% in our sample). Interestingly, the educational disparities in HIV gradually vanish as the epidemic progresses past these early stages, to then revert to a positive education gradient in HIV in the more advanced stages of the epidemic, where an additional five years of education result in a 2.40 percentage point increase in the probability of being HIV positive. This U-shaped pattern is also significant when controlling for country and year effects, although the size of the effects of education on HIV is smaller. We then explore whether there are heterogeneous effects by gender. While women and men share the U-shaped pattern of the education gradient in HIV, its magnitude is larger for women than for men. This gap is largest at the early stages of HIV development, and tends to disappear later on. Regarding women, completing five additional years of schooling is associated with a 7.40% rise in the probability of being HIV positive at the early stages of the epidemic, that is, an increase twice as large as that of men, 3.80% per five years of education. Thereafter, the education gradient in HIV substantially declines until it vanishes for women who live in the middle of the HIV epidemic. At this stage, the decline in the gradient is even sharper for men, and changes its sign, reaching a -0.65% per five years of schooling. Interestingly, the gradient reverts to positive in the more advanced stages of the epidemic for both women and men, respectively, 2.70% and 1.75% per five years of education.

To gain a better understanding of the dynamic relationship between education and the probability of infection, we explore educational disparities in the actual risky sexual behavior. Remarkably, the pattern of the education gradient in HIV closely resembles the pattern of the educational disparities in risky sexual behavior. While more education is associated with more extramarital partners at early stages of the epidemic (0.19 per five more years of schooling for women and 0.15 for men), this relationship rapidly and significantly declines in mid stages of the epidemic (0.06 and 0.05 extramarital partners per five more years of schooling for, respectively, women

and men). Interestingly, in the most advanced stages of the epidemic the relationship between education and the number of extramarital partners significantly increases (0.11 per five more years of schooling for women and 0.15 for men). The fact that the HIV-Education gradient closely follows education disparities in risky sexual behavior across stages of the epidemic points out the important role of educational disparities in determining the HIV incidence.[2]

The rest of our paper is organized as follows. In Section 1, we discuss the related literature. In Section 2.1, we describe our data. We document the heterogeneity of the epidemic across SSA countries and propose a unified macro framework to define the aggregate stages of the HIV epidemic in a comparable manner across time and space in Section 2.2. The estimates of the stylized evolution of the HIV-Education gradient across the aggregate stages of the epidemic are in Section 2.3. The potential mechanisms behind the evolution of the gradient are explored using a simple theoretical framework with risky sex choice that differs across education groups in Section **??**. Section 8 concludes.

Related literature We are not the first to investigate the HIV-Education gradient. A large number of epidemiological studies and small-scale studies examined socioeconomic disparities in HIV without reaching neither conclusive nor generalizable answers.While the current mixed evidence is likely to reflect differences in methodology, sampling strategy, and measures of socioeconomic indicators and HIV status, this may not entirely explain the differing conclusions reached by previous studies, which for instance overlooked the large differences in the evolution of the HIV epidemic across SSA countries. In a review of epidemiological studies Gregson et al. [2001] conclude that there could be temporal dynamics of the influence of socio-economic development on rates of HIV transmission, and in particular that the greater vulnerability of individual with a high socioeconomic status may be a transient feature of the early stages of epidemics. The findings of de Walque [2007] point in the same direction in his analysis of the HIV-education gradient in rural Uganda between 1989 and 2001. Importantly, these studies did not expect a rebound in the education gradient in HIV that we document. Recently, using nationally representative data from the Demographic and Health Surveys (DHS) for five SSA countries, Fortson [2008] finds education has a positive association with HIV status: adults with primary school are one half more likely to be infected than adults with no schooling conditioning on age, sex and area of residence (urban/rural). Using the DHS wealth index, de Walque [2009] also finds that wealth displays a positive association with HIV status. Mishra et al. [2007] find similar results for eight

[2]The prevalence of HIV is determined by the newly infected individuals as well as by the survival probabilities of the individual being infected in the past. Therefore, the educational disparities in the HIV prevalence could be the result of educational disparities in the incidence of HIV and/or educational disparities in the survival probabilities. The DHS data allow us only to focus on individual attitudes towards risky sexual behavior that might increase the probability of being HIV infected. Later on, we discuss the potential role that educational disparities in survival rates might play in shaping the HIV-Education patterns as the HIV epidemic evolves.

DHS countries in SSA.

We contribute to this literature in two major respects. First, we uncover the evolution of the HIV epidemic using a unified macro framework that systematically defines the aggregate stages of the HIV epidemic in a comparable manner across countries. Our approach addresses similar data challenges as those faced when defining the stages of economic development [Lucas, 2004, Herrendorf et al., 2014] or the stages of the demographic transition [Galor and Weil, 2000, Lee, 2003, Greenwood et al., 2005]. Second, in the context of this unified macro framework, and using repeated cross-sections of DHS surveys, our analysis exploits a rich variation in the aggregate stages of the HIV epidemic to document the stylized dynamic behavior of the HIV-Education gradient along the course of the epidemic, as well as the evolution of educational disparities in sexual responses. Our findings emphasize that while at early stages of the epidemic the HIV-Education gradient is large and positive (roughly three times larger than its stationary counterpart, à la Fortson [2008]), the gradient decreases to levels that are not significantly different from zero as the epidemic evolves; a macro cross-country decline that resembles the results by de Walque [2007] for several sites in rural Uganda. The time span of our data allows us to pick a rebound in the education gradient in HIV in mature epidemics, which we link to a positive change in risky sexual behavior among highly educated individuals. Interestingly, this U-shaped pattern is much more prominent for women than for men.

Due to the fact that heterosexual intercourse is the major mode of HIV infection [Behrman and Kohler, 2012], the relationship between HIV, risky sexual behavior, and HIV knowledge has been extensively studied. For example, information campaigns that improve the knowledge about HIV risk infection may induce people to adopt safer lifestyles. In this direction, de Walque [2007] documents substantial behavioral change in rural Uganda associated with the ABC campaign (Abstinence, Be faithful, and use Condoms).

Finally, our study is broadly related to the literature examining educational disparities in health outcomes. See Cutler and Lleras-Muney [2011] for a review of the studies examining this relationship in both developed and developing countries. Within this group of studies our work relates more closely to those that allow the relationship between education and health to be nonstationary. This is the case of the "fundamental cause" literature described in Cutler et al. [2006]in which the diffusion of information on technological improvements is an argument used to explain the changes in the education gradient in health.

2 Empirical Evidence

2.1 Data

The core of our exercise consists of examining the relationship between education and HIV education over the stages of the HIV epidemic. To address this question it would be ideal to use nationally representative long panel data for several SSA countries starting in the pre-HIV era. Unfortunately, available nationally representative data are neither long nor panel. However, from a macroeconomic perspective, we show it is possible to construct a normalized path of the patterns of HIV infection by education groups over the stages of the epidemic for several SSA countries to recover stylized patterns between education and HIV. To do so, we combine two sources of data: (i) cross-sectional data from the DHS, and (ii) aggregate data from the most recent World Population Prospects (WPP) provided by the United Nations.

The Demographic and Health Surveys. The DHS are based on nationally representative samples and are available for a large set of SSA countries. We consider the full sample of SSA DHS surveys for which individual HIV testing has been conducted (and available as of July 2014): Burkina Faso (2003, 2010), Burundi (2010), Cameroon (2004, 2011), Congo (2007), Côte d'Ivoire (2005, 2011), Democratic Republic of Congo (2007), Ethiopia (2005, 2011), Gabon (2012), Ghana (2003), Guinea (2005, 2012), Kenya (2003, 2008), Lesotho (2004, 2009), Liberia (2007), Malawi (2004, 2010), Mali (2006), Mozambique (2009), Niger (2006), Rwanda (2005, 2010), Senegal (2005, 2010), Sierra Leone (2008), Swaziland (2006), Tanzania (2003, 2007, 2011), Uganda (2011), Zambia (2007) and Zimbabwe (2005, 2010), for a total of 25 DHS countries. For a number of countries the survey was conducted in two consecutive years, so we can exploit variation across 51 country-year pairs, which provide sufficient observational heterogeneity on the stages of the HIV epidemic to obtain reliable estimates of the evolution of the education gradient over these stages.

While the DHS are primarily health interviews, they also contain cross-sectional information on individual socioeconomic characteristics, knowledge on HIV, several measures of risky sexual behavior (e.g., number of extramarital relationships and condom use) and most importantly, a large proportion of adult respondents have been tested for HIV.[3] We use this cross-referenced individual information harmonically collected across SSA countries. Our whole sample consists of

[3]The proportion of respondents who did not take the HIV test is .318 in the original whole sample (.098 and .432 among men and women, respectively). However, we find that the association between the likelihood of taking the HIV test and the educational attainment is virtually zero in the DHS sample. Further, our evidence suggests the DHS non-response bias for HIV testing is minimal. We find that statistics for age, schooling, and residence computed for the sample of HIV-tested adults resemble the analogous ones in the overall male sample. These results are available upon request. See also the discussions in Fortson [2008].

Table 1: The DHS Sample Characteristics (across Countries)

(A) Women	Mean	Median	Min.	Max.	Gini
HIV Prevalence (%)	6.1	4.1	0.5	31.2	0.54
Years of Schooling	3.2	2.8	0.6	4.8	0.26
Age	28.2	28.2	27.7	29.4	0.01
Urban (%)	32.7	33.6	10.6	88.3	0.25
Extramarital Partners, n	0.15	0.17	0.01	0.56	0.37
$n = 0$ (%)	87.4	84.9	57.9	99.2	0.07
$n = 1$ (%)	11.5	14.3	0.8	35.9	0.36
$n = 2$ (%)	1.0	1.2	0.0	6.3	0.47
$n \geq 3$ (%)	0.1	0.1	0.0	0.0	0.66
Frequency of Condom Use (%)	9.5	8.7	0.5	37.7	0.42
Not Sexually Active (%)	15.6	12.8	4.3	30.7	0.28

(B) Men	Mean	Median	Min.	Max.	Gini
HIV Prevalence (%)	4.1	2.2	0.4	19.7	0.56
Years of Schooling	3.8	3.2	1.3	5.3	0.18
Age	28.2	28.5	25.9	30.3	0.02
Urban (%)	33.7	35.6	15.4	87.6	0.24
Extramarital Partners, n	0.41	0.48	0.08	1.17	0.30
$n = 0$ (%)	73.7	65.7	36.8	93.5	0.14
$n = 1$ (%)	19.7	28.0	5.1	43.8	0.22
$n = 2$ (%)	4.5	5.2	0.8	14.4	0.36
$n \geq 3$ (%)	0.7	0.7	0.1	2.8	0.42
Frequency of Condom Use (%)	20.5	21.4	4.2	49.3	0.31
Not Sexually Active (%)	21.7	17.9	7.0	35.9	0.24

Notes: The computation of these statistics is performed by using individual HIV weights provided by the DHS to compute the mean. Then, country-specific population weights (i.e., the population size of each country provided by World Population Prospects) are used to compute the statistics across countries. The number of extramarital partners refers to the last 12 months. The frequency of condom use refers to the last sexual intercourse. Our sample is based on 39 DHS with 25 countries and a total of 227,935 women and 174,852 men.

a total of 402,670 individuals, of which 56.5% are women. We choose to explore HIV infection risk by education groups separately for women and men. The average HIV prevalence is 6.1% for women and 4.1% for men (respectively, panel A and B, Table 1). There is a substantial degree of heterogeneity in HIV prevalence across countries. The Gini index for the HIV prevalence across these 39 DHS (25 countries) is 0.54 for women and 0.56 for men, with a range from 0.5 to 31.2 for women and from 0.4 to 19.7 for men. Note that there is a substantial HIV gender gap of 2% that is roughly half of the total HIV prevalence for men. Interestingly, as we show below, this gender gap in HIV prevalence evolves across the stages of the HIV epidemic. We restrict our attention to HIV-tested adults men and women 15-49 years old who reported their schooling achievement, which is on average 3.2 for women and 3.8 for men. The urban population is roughly one third for both women and men.

Several aspects make DHS datasets appealing for our exercise. An important advantage of these data is that they provide unambiguous individual measures of individual HIV status, education, knowledge on HIV, and risky sexual behavior in a comparable manner across SSA countries. First, regarding individual HIV status, the DHS provides a direct measure as individuals have blood testing for HIV, so we do not rely on indirect proxy for HIV obtained from other health outcomes or biomarkers. Second, regarding education variables, DHS collect data on education (i.e., number of years of schooling and maximum degree attained) and also provide an asset-based wealth index.[4] Our preferred choice for measuring education is years of schooling—perhaps the most commonly used measure for education in the previous literature. The reasoning for our choice of years of schooling—rather than the DHS wealth index— is that while wealth is influenced by subsequent negative health conditions (such as HIV), or other shocks that will potentially determine one's health status in adulthood, educational attainment is not because, typically, education is completed before individuals in our sample—adults between 15 and 49 years of age—enter adulthood. However, we cannot entirely rule out the fact that investments in education might respond to changes in life expectancy. Indeed, Fortson [2011] suggests a significant negative effect of HIV on investment in children in a model where agents explicitly consider mortality risk when making human capital decisions.

Finally, regarding risky sexual behavior, we focus on (i) the number of sex partners (i.e., the extensive margin) in the past 12 months other than spouses and (ii) condom use in last intercourse (i.e., the intensive margin, quality). The number of extramarital partners is on average 0.15 for women and 0.41 for men (panel A and B, Table 1). This is consistent with the population of women not having extramarital partners being larger than men's, respectively 87.4% and 73.7%.

[4] Unfortunately, the DHS do not collect data on income, with very few exceptions that document wage earnings. This task is not easy as many SSA populations are mostly rural and a large proportion of these households' resources come from unsold agricultural production; see a discussion in De Magalhães and Santaeulàlia-Llopis [2015].

The frequency on condom use is 9.5% for women and 20.5% for men. That is, women not only report less extramarital partners than men, but also less condom use in the last sexual intercourse than men. These statements are consistent, as men's last sexual intercourse for casual sex is more common than women's. These figures are conditional on individuals being sexually active. Interestingly, the proportion of women not sexually active is smaller, 15.6%, than the one of men, 21.7%.[5]

The World Population Prospects. To uncover the evolution of the HIV epidemic we use the data from the 2015 revision of the World Population Prospects (WPP) constructed by the United Nations (Department of Economic and Social Affairs, Population Division). The WPP 2015 provides estimates of the HIV prevalence rates from 1980 until 2014 (at the country level), and their projections from 2014 onward for a large set of SSA countries. [6] The additional data on country-specific ART coverage used in our robustness exercise are also from the WPP.[7]

Finally, for all our SSA countries we use data on real output per capita from the Penn World Tables and data on agricultural share of output from the World Bank Development Indicators. We use these data in our empirical analysis to control for country-specific stages of aggregate economic development.

2.2 The Stages of the HIV Epidemic

This section describes the stages of the HIV epidemic. First, we discuss a set of challenges that we argue a useful definition of the stages of the HIV epidemic must address (Section 2.2.1). From our macroeconomic perspective these challenges arise from country differences in HIV prevalence across time (i.e., the evolution of the HIV epidemic within a country) and across space (i.e., heterogeneity of HIV prevalence across countries within a given period). Second, we provide an algorithm that circumvents those challenges by normalizing the HIV epidemic in both dimensions, time and space (Section 2.2.2). Our definition is provided in Section 2.2.3.

[5]The Gini index for the number of extramarital partners across these 39 DHS (25 countries) is 0.37 for women and 0.30 for men, with a range from 0.01 to 0.56 for women and from 0.08 to 1.17 for men. The Gini index for the frequency of condom use in last sexual intercourse is 0.42 for women and 0.31 for men, with a range from 0.5% to 37.7% for women and from 4.2% to 49.3% for men (Table 1).

[6]The WPP data represent the official 2014 estimates of UNAIDS. Until 2006 the UNAIDS estimates relied mostly on data aggregations collected from antenatal clinics that overestimated prevalence levels. Since 2008, the UNAIDS data belong to a downward revision largely originated by the appearance of nationally representative surveys such as the DHS, and do not suffer from overestimation problems. Indeed, the HIV prevalence levels computed from our DHS samples and the HIV prevalence levels from WPP are very similar. See also a detailed discussion of these data in Bongaarts et al. [2008].

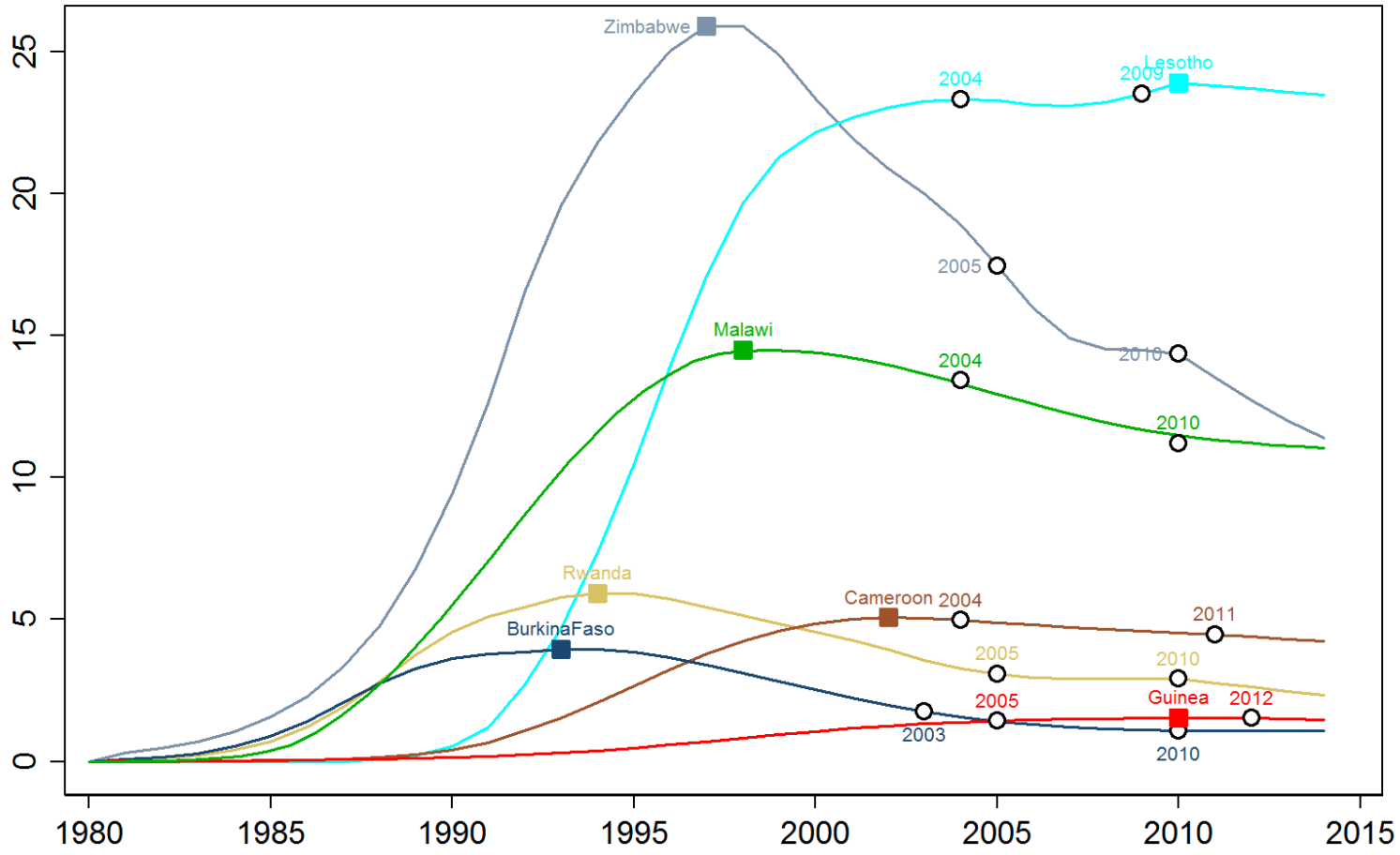
[7]Source: United Nations, Department of Economic and Social Affairs, Population Division: World Population Prospects. Unpublished Data - Special Tabulations. We thank Patrick Gerland for sharing these data.

Table 2: The Evolution of the HIV Epidemic across Sub-Saharan Countries: The DHS Sample

	DHS Obs.		Peak		DHS/Peak	
Country	t_i	HIV_i	$t_{i,*}$	$HIV_{i,*}$	$t_i - t_{i,*}$	$\frac{HIV_i}{HIV_{i,*}}$
Burkina Faso	2003	1.76	1993	3.96	-10	0.44
Burkina Faso	2010	1.10	1993	3.96	-17	0.27
Burundi	2010	1.39	1996	4.47	-14	0.31
Cameroon	2004	4.98	2002	5.06	-2	0.98
Cameroon	2011	4.47	2002	5.06	-9	0.88
Congo Brazaville	2009	2.74	1995	4.80	-14	0.57
Cote d'Ivoire	2005	4.79	1999	6.18	-6	0.77
Cote d'Ivoire	2011-12	3.41	1999	6.18	-12	0.55
D.R. Congo	2007	1.30	2001	1.43	-6	0.91
Ethiopia	2005	2.28	1999	3.23	-6	0.70
Ethiopia	2011	1.32	1999	3.23	-12	0.41
Gabon	2012	4.16	2003	5.53	-9	0.75
Ghana	2003	2.00	2000	2.15	-3	0.93
Guinea	2005	1.43	2010	1.54	5	0.93
Guinea	2012	1.52	2010	1.54	-2	0.98
Kenya	2003	7.08	1996	9.16	-7	0.77
Kenya	2008-09	5.84	1996	9.16	-12	0.64
Lesotho	2004-05	23.33	2010	23.91	6	0.98
Lesotho	2009-10	23.52	2010	23.91	1	0.98
Liberia	2006-07	1.34	2002	1.84	-5	0.73
Malawi	2004	13.41	1998	14.99	-6	0.89
Malawi	2010	11.21	1998	14.99	-12	0.75
Mali	2006	1.34	1999	1.71	-7	0.78
Mozambique	2009	9.98	2008	10.01	-1	1.00
Niger	2006	1.00	2002	1.20	-4	0.83
Niger	2012	0.60	2002	1.20	-10	0.50
Rwanda	2005	3.09	1994	5.93	-11	0.52
Rwanda	2010	2.90	1994	5.93	-16	0.49
Senegal	2005	0.90	2004	0.90	-1	1.00
Senegal	2010-11	0.70	2004	0.90	-6	0.78
Sierra Leone	2008	1.66	2008	1.66	0	1.00
Swaziland	2006-07	23.36	2010	24.46	4	0.95
Tanzania	2003-04	6.04	1996	7.36	-7	0.82
Tanzania	2007-08	5.30	1996	7.36	-11	0.72
Tanzania	2011-12	4.81	1996	7.36	-15	0.65
Uganda	2011	6.55	1991	12.62	-20	0.52
Zambia	2007	12.87	1998	14.77	-9	0.87
Zimbabwe	2005-06	17.46	1997	25.90	-8	0.67
Zimbabwe	2010-11	14.35	1997	25.90	-13	0.55

Notes: t_i is the calendar year of DHS data collection for country i; HIV_i is the prevalence rate for country i the year of DHS data collection; $t_{i,*}$ is the year country i reaches its HIV prevalence peak; $HIV_{i,*}$ is the peak prevalence rate for country i. Sources: United Nations, Department of Economic and Social Affairs, Population Division: World Population Prospects: The 2015 Revision, Medium-Variant Estimation and Projection.

Figure 1: Challenges for the Definition of Stages of the HIV: Epidemic. The Evolution of the HIV Epidemic for a DHS Subsample.



Source: United Nations, Department of Economic and Social Affairs, Population Division: World Population Prospects: The 2015 Revision, Medium-Variant Estimation. *Notes*: The solid square on each HIV time path displays the HIV prevalence at the peak year, and the open black circle on each HIV time path displays the HIV prevalence at the year that the DHS data were collected.

2.2.1 Challenges for a Definition of the Stages of the HIV Epidemic

The evolution of the HIV epidemic is largely heterogeneous. To illustrate this, we show in Figure 1 the country-specific time path of the epidemic for a selected subsample of countries.[8] In addition, we provide country-specific statistics of the HIV epidemic for our entire sample of 39 DHS country-year surveys (Table 2). The following patterns arise across time and space.

HIV Prevalence Differences Across Space While the HIV prevalence levels largely differs across countries (columns 1 and 2, Table 2),[9] two countries with the same HIV prevalence are not necessarily at the same epidemiological stage.

Remark 1. *The HIV prevalence alone is not sufficient to define the aggregate stage of the HIV epidemic.*

We show this argument with two straightforward counterexamples. Although Malawi in 1998 and Zimbabwe in 2010 both share the same HIV prevalence of 14.4, Malawi reaches this infection rate at its HIV peak while Zimbabwe reaches it only 13 years after its HIV peak (Figure 1). Indeed, in 2010 Zimbabwe's HIV prevalence is 55% of its HIV peak prevalence (column 6, Table 2). Another interesting counterexample arises from the comparison of the DHS observations of Zimbabwe and Lesotho. The DHS observation of Zimbabwe in 2005 delivers an HIV prevalence of 19.2%, lower than that of Lesotho in 2004, 23.4%. Looking at this statistic only, we would infer that Lesotho is at later stages of the epidemic than Zimbabwe. However, we actually know that Zimbabwe's HIV peak occurred at a higher level and earlier, 29.1% at 2009, than that of Lesotho, 23.8% at 2007, which suggests an opposite ordering over stages. That is, the ordering of DHS countries by HIV prevalence is a mere artifact of the years in which DHS were collected.

One step to address the problematic use of the absolute size of the HIV prevalence as a measure of the stages of the epidemic is to compute relative size of HIV prevalence dividing country-specific observations of HIV prevalence by their corresponding HIV peaks. However, this poses a new set of drawbacks because countries not only differ in the HIV peak level but also in the year of their HIV peak (columns 3 and 4, Table 2).

[8] This subsample of DHS countries consists of Burkina Faso, Cameroon, Guinea, Lesotho, Malawi, Rwanda and Zimbabwe. This subsample serves expositional purposes only as it is useful to highlight the heterogeneity of the evolution of the HIV epidemic across countries as we describe next. Many other subsample choices would be equally useful.

[9] For example, in year 2010, the HIV prevalence (in percentages) is 1.1 in Burkina Faso, 5.1 in Cameroon, 1.9 in Guinea, 11.7 in Malawi, 3.1 in Rwanda and 18.0 in Zimbabwe. Across all SSA countries, the inequality in HIV prevalence remains high across time with a Gini coefficient of 0.63 in 1990, 0.56 in 2000 and 0.57 in 2010. The SSA set consists of 44 countries. Similar figures are attained with our sample of 25 DHS countries with Gini's coefficients of 0.55 in 1990, 0.51 in 2000 and 0.55 in 2010.

Remark 2. *The relative HIV prevalence alone is not sufficient to define the aggregate stage of the HIV epidemic.*

To see this, note, for example, that while the DHS observations of Guinea in 2005 and Ghana in 2003 share the same relative HIV prevalence of .93 (column 6, Table 2), Guinea attains that relative size 5 years before reaching its peak, and Ghana does so 3 years after reaching its peak (column 6, Table 2). This observation suggests that two countries can be at different stages of the epidemic despite having the same relative HIV prevalence. Another interesting example is the one posed by the DHS observations of Rwanda 2005 and Uganda 2011. Both surpassed their respective peaks, and have the same relative prevalence level of .52 (column 6, Table 2). However, it took Rwanda 11 years to move from its peak to this relative prevalence, while it took Uganda almost twice as much time, 20 years, to reach the same relative prevalence (column 5, Table 2). The fact that the transition away from the peak is slower in Uganda than in Rwanda is, in itself, a phenomenon to which we would like our definition of the stages of the epidemic to be invariant. Constructively, the arguments posed here against the sole use of the absolute (or the relative) HIV prevalence to define stages of the epidemic also suggest what we need to add to our definition of the stages to resolve the exposed problems: some properties of the time path of the HIV epidemic.

HIV Prevalence Differences Across Time The time-path of HIV epidemic largely differs across countries. In particular, a large degree of heterogeneity exists for the HIV peak year across SSA countries (column 3, Table 2). In our DHS sample, the peak of the HIV year ranges from 1991 in Uganda to 2010 in Guinea, Lesotho and Swaziland. This leads to the following remark.

Remark 3. *Time (calendar year) alone is not sufficient to define the aggregate stage of the HIV epidemic.*

This remark states that two countries that suffer the HIV epidemic are not necessarily at the same epidemiological stage at the same calendar year. This is straightforward. Lesotho and Guinea reach the peak of their respective HIV epidemic in 2010, while Zimbabwe is at more advanced stage of its epidemic in 2010, precisely 13 years ahead of its HIV peak in 1997 (Figure 1).

One step to correct for the country-specific year of the HIV peak is to compute the relative time, i.e., calendar year minus year of HIV peak (column 5, Table 2). However, a large degree of heterogeneity exists for the speed by which SSA countries move to the respective HIV peaks and the speed by which SSA countries move away from their respective HIV peaks. This leads to the following remark.

Remark 4. *Relative time (calendar year minus year of HIV peak) alone is not sufficient to define*

the aggregate stage of the HIV epidemic.

To see this, note that the DHS observations of Ethiopia in 2011 and Malawi in 2010 both share the same time distance with respect to their own HIV; in both cases 12 years have passed between the peak and the DHS data collection. However, in those 12 years Ethiopia has managed to decrease its relative HIV prevalence to 0.41 (column 6, Table 2), while Malawi has only managed to decrease its relative HIV prevalence to 0.75. Again, as we noted for remark 2, the fact that the transition away from the peak of Malawi is slower than that of Ethiopia is, in itself, a phenomenon to which we would like our definition of the stages of the epidemic to be invariant. The relative time does not suffice to define stages of the HIV epidemic.

To address these four remarks at once, we propose a two-dimensional (2D) algorithm that normalizes both the HIV prevalence level and time.

2.2.2 A Two-Dimensional Normalization of the Evolution of the Epidemic Across Time and Across Space

This section builds a 2D algorithm that, for all countries, normalizes the country-specific level and time path of the epidemic, thereby making the evolution of the HIV epidemic comparable across countries. Once the evolution of the epidemic is normalized for all countries, the position of each DHS dataset on its associated epidemiological stage readily follows.

Algorithm 1. [A Two-Dimensional Normalization of the Evolution of the HIV Epidemic] *Given the time series of the level of HIV prevalence of each i, we follow three steps to conduct a 2D normalization of the level and time path of the HIV epidemic:*

1. *Interpolate the country-specific time path of prevalence for each country i, $\{\lambda_{i,t}\}_{t_0}^{t_p}$, for $p+1$ interpolation points (years), where p is a positive integer. Then, interpolate the aggregate (across countries) prevalence path as $\lambda_t = \frac{\sum_i^n \lambda_{i,t}\mu_{i,t}}{\sum_i^n \mu_{i,t}}$, where n is the total of number of countries and $\mu_{i,t}$ is the population level of country i at period t. Denote the country-specific interpoland function as $s_i : t \to [0, \max_t \lambda_{i,t}]$, where $\max_t \lambda_{i,t} \in [0,1]$ and $s_i \in \mathcal{S}$, where $\mathcal{S}$ is the collection of functions that can be written as a linear combination of a set of n-known linearly independent basis functions ψ_j, $j = 1, ..., n$,*

$$s_i(t) = \sum_{j=1}^{n} \theta_j \, \psi_j(t)$$

with n unknown θ_j coefficients. Denote the aggregate interpoland as $s(t)$ where $s(t)$

shares the same properties as the country-specific interpolands $s_i(t)$. *Importantly, note that* $\max_t s_i(t)$ *is not necessarily identical across countries or to the aggregate* $\max_t s(t)$.

2. *Level normalization*

   (a) *Compute the country-specific peak prevalence,*

   $$s_i(t_*^i) = \max_t s_i(t), \tag{1}$$

   where $t_*^i = \arg\max_t s_i(t)$ *is the period country* i *reaches its peak,* $s_i(t_*^i)$. *Redo equation (1) to obtain the aggregate peak* $s(t_*)$ *and aggregate peak period,* $t_* = \arg\max_t s(t)$.

   (b) *Normalize the country-specific and aggregate interpolands by their respective peak prevalence,*

   $$\widetilde{s}_i(t) = \frac{1}{s_i(t_*^i)}\, s_i(t) \quad \textit{and} \quad \widetilde{s}(t) = \frac{1}{s(t_*)}\, s(t),$$

   where $\widetilde{s}_i, \widetilde{s} : t \to \Lambda = [0,1]$ *and* $\arg\max_t s_i(t) = t_*^i = \arg\max_t \widetilde{s}_i(t)$. *Note now that* $\widetilde{s}_i(t_*^i) = \widetilde{s}(t_*) = 1\ \forall i$.

3. *Time normalization*

   (a) *For* $t_0^i < t^i \leq t_*^i$, *normalize the time interval between the initial period for which data are available,* $t_0^i = 1980$, *and the country-specific peak period,* t_*^i, *by the time interval between the aggregate initial period,* $t_0 = 1980$, *and the aggregate peak period,* t_*. *To do so, we compute the constant of time proportionality for the pre-peak era,*

   $$\alpha_i^L = \frac{t_* - t_0}{t_*^i - t_0^i}.$$

   For $t^i > t_*^i$, *normalize the time between the peak period,* t_*^i, *and the period* t_γ^i *in which country* i *reaches a given threshold* $\gamma \in [0,1]$, *that is,* $t_\gamma^i = \widetilde{s}_i^{-1}(\gamma)$, *over the analogous aggregate interval with* t_* *and* $t_\gamma = \widetilde{s}^{-1}(\gamma)$,

   $$\alpha_i^R(\gamma) = \frac{t_\gamma - t_*}{t_\gamma^i - t_*^i}.$$

   Here, note that t_γ^i *and* t_γ *may not occur at an interpolation node but elsewhere along their respective interpoland.*

(b) *Normalize the time input of the country-specific interpolands by* α_i^L *and* α_i^R*,*

$$\tau = \alpha_i^L (t - t_*^i) \quad \textit{for} \quad t \leq t_*^i \tag{2}$$

$$\tau = \alpha_i^R (t - t_*^i) \quad \textit{for} \quad t > t_*^i \tag{3}$$

where $\tau \in T$ *are the normalized units of time. Operations (2) and (3) compress/stretch the interpoland*[10] *to ensure that for* $\tau \leq \tau_*$ *(before the peak) the number of normalized periods* τ *that it takes each country to move from* τ_0 *to the peak are the same across countries,*

$$\widetilde{s}_i^{-1}(1) - \widetilde{s}_i^{-1}(0) = \widetilde{s}_j^{-1}(1) - \widetilde{s}_j^{-1}(0) = \widetilde{s}^{-1}(1) - \widetilde{s}^{-1}(0) \qquad \forall\, i, j\ ,$$

and for $\tau > \tau_*$ *(after the peak) the normalized periods* τ *that it takes each country to move from the peak to a threshold of prevalence* γ *is the the same across countries,*

$$\widetilde{s}_i^{-1}(\gamma) - \widetilde{s}_i^{-1}(1) = \widetilde{s}_j^{-1}(\gamma) - \widetilde{s}_j^{-1}(1) = \widetilde{s}^{-1}(\gamma) - \widetilde{s}^{-1}(1) \qquad \forall\, i, j.$$

This allows us to define the evolution of the epidemic for each country and the aggregate,

$$\widetilde{s}_i : \tau \to \Lambda \qquad \textit{and} \qquad \widetilde{s} : \tau \to \Lambda, \tag{4}$$

in the same—hence comparable—2D normalized space (T, Λ)*.*

To implement the algorithm we need to make two choices: the shape of the basis functions, $\psi(\tau)$, and the prevalence threshold γ for the time normalization after the peak. First, we specify $\widetilde{s}(\tau)$ as a B-spline with cubic pieces and solve for the θ_j coefficients accordingly. Our choice of splines as interpolands obeys our desired manageability of the interpoland given the size of the Lagrangian interpolation problem poised by 71 (1980-2050) interpolation data points. Second, our choice of γ responds to balance between minimizing the use of projection of U.N. data for our set of DHS countries and maximizing the number of countries that have already surpassed the threshold γ at the time of DHS data collection. Our search for this balance suggests a value of $\gamma = .8$. This choice of γ implies that more than half of the countries in our dataset have already passed the threshold. Our results are robust to alternative choices of γ. To see this, note that the value of γ does not alter the ranking of countries across stages of the epidemic.

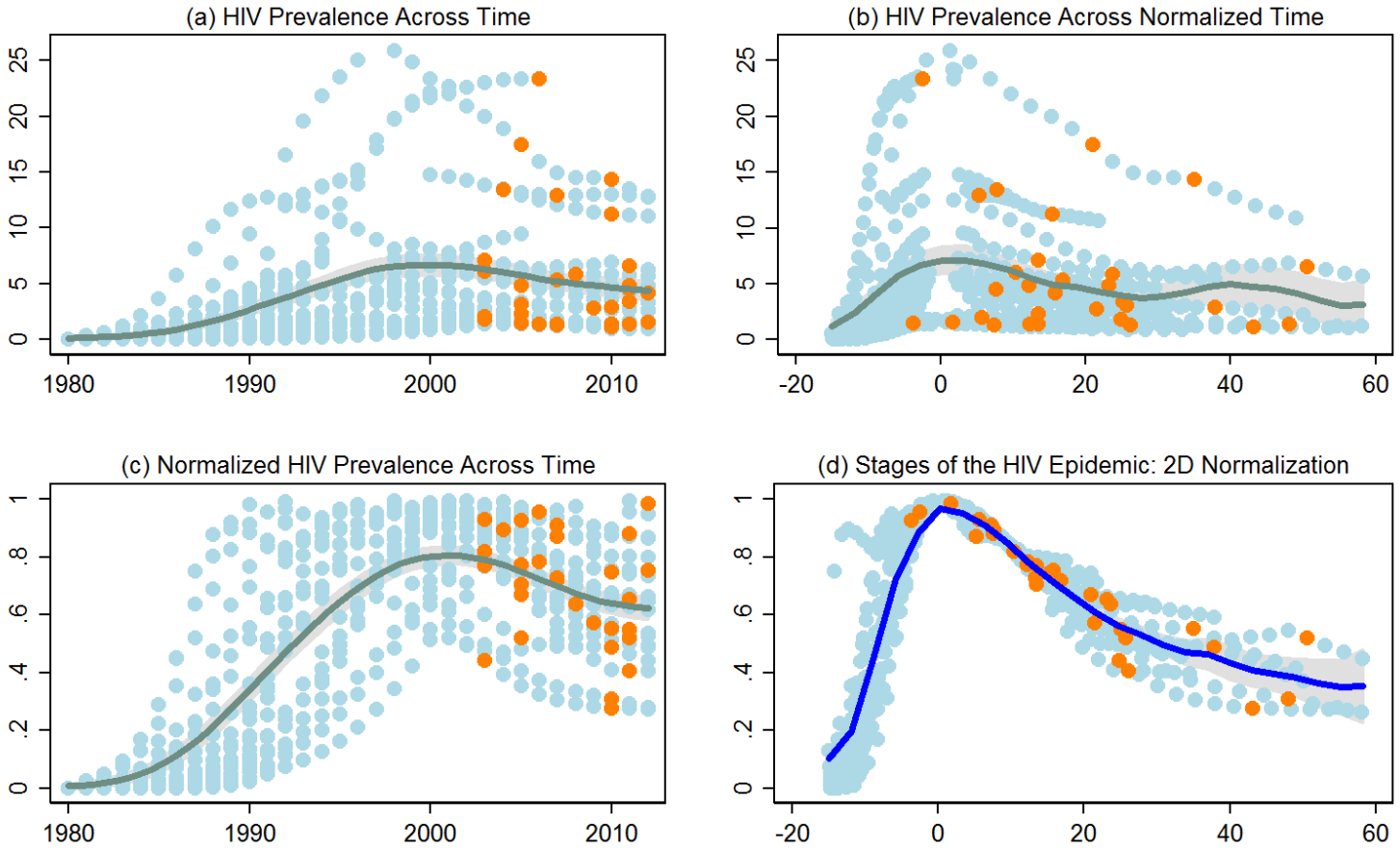
We next apply our algorithm to the SSA countries for which U.N. estimates and projections

[10] The interpoland s_i horizontally compresses when $\alpha_i^k > 1$ with $k = \{L, R\}$ and expands otherwise.

of the HIV prevalence time path are available.[11] The results are depicted in Figure 2. Panel (a) shows the HIV prevalence level across time for each and all countries. We highlight (in orange) the country-year observations for which DHS data with HIV testing results are available. This panel portrays all challenges discussed in Section 2.2.1. Panel (b) shows the result of the first normalization of our algorithm, the HIV prevalence level normalization, which tackles the issues associated with the absolute HIV prevalence, but does not correct for having HIV peaks at different calendar years for different countries. Panel (c) shows the result of the second normalization of our algorithm, the time normalization, which forces all countries to peak at the same calendar year but does not resolve the issues associated with the absolute HIV prevalence level. Finally, panel (d) closes our algorithm by jointly applying the level and the time normalizations. These 2D normalization resolves the full set of challenges posed in Section 2.2.1 as the relative HIV prevalence peaks in all countries at the same period. As it is obvious from panel (d), after the 2D normalization the level and time path of the epidemic are entirely comparable across countries.

[11] The interpolation Lagrangian points, that is, the country-specific prevalence time series, $\lambda_{i,t}$, are retrieved from the U.N. population division estimates and projections until 2050 (medium-variant).

Figure 2: Defintion of Stages of the HIV Epidemic



Source: Outcome of our 2D-normalization algorithm (Subsection 2.2.2) implemented using WPP, 2015, data for SSA. The vertical axis in the top panels is HIV prevalence. The vertical axis in the bottom panels is normalized HIV prevalence. The horizontal axis in the left panels is time. The horizontal axis in the right panels is normalized time. This way, the 2D-normalization is operative in panel (d) (Subsection 2.2.2). In each panel, the orange markers in the scatterplots represent a DHS dataset. The plotted trends are locally weighted polynomials with 95% confidence intervals.

2.2.3 A Definition of the Stages of the HIV Epidemic

Note that the position of each country i on its normalized HIV time path at the period t_{DHS} at which its respective DHS data were collected can be easily computed by solving for τ_i in

$$\widetilde{s}_i^{-1}\left(\frac{\lambda_{i,t_{DHS}}}{s_i(t_*^i)}\right) = \tau_i.$$

Then, the stage of the HIV epidemic is the continuous real variable,

$$\omega(\tau,\zeta) = \frac{\zeta}{\tau - \tau_*} \to \mathcal{R}^1, \tag{5}$$

with the pair (τ,ζ) belonging to the 2D normalized space, $T\times\Lambda$. Geometrically, $\omega(\tau,\zeta)$ represents the slope of the arrays from the origin in the (T,Λ) space with the following limiting properties:

$$\lim_{\tau\to\tau_*^-}\omega(\tau,\zeta) = -\infty,\ \lim_{\tau\to\tau_*^+}\omega(\tau,\zeta) = \infty,\ \text{ and }\ \lim_{\tau\to-\infty}\omega(\tau,\zeta) = \lim_{\tau\to+\infty}\omega(\tau,\zeta) = 0.$$

To conduct our empirical exercise, we discretize the continuous variable that defines the epidemiological stages in (5).[12]

Definition 1. [Stages of the HIV Epidemic] *Given a set of stage thresholds* $\{\zeta_0,...,\zeta_j,...,\zeta_n\}$ *with* $\zeta_j > \zeta_{j+1}$ *for all* j*, the stage* j *of the HIV epidemic consists of all pairs* $(\tau,\zeta)\in T\times\Lambda$ *such that* $\omega(\widetilde{s}^{-1}(\zeta_{j+1}),\zeta_{j+1}) \leq \omega(\tau,\zeta) \leq \omega(\widetilde{s}^{-1}(\zeta_j),\zeta_j)$*, where* $\widetilde{s}(\tau)$ *is the normalized (population-weighted) aggregate of the HIV epidemic defined in (4).*

Our choice for the stage thresholds $\{\zeta_0,...,\zeta_j,...,\zeta_n\}$ pursues the maximization of both countries per stage of the epidemic and number of stages. To do so, we set $\zeta_0 = 1$ and $\zeta_j = \zeta_1 - .05j\ \forall j$. The results of this exercise are shown in Figure 3 where each data point $(\tau_i,\widetilde{s}_i)$ represents a DHS dataset. This implies the following allocation of DHS datasets, $\omega(\tau_i,\widetilde{s}_i(\tau_i))$, over stages of the epidemic as follows:

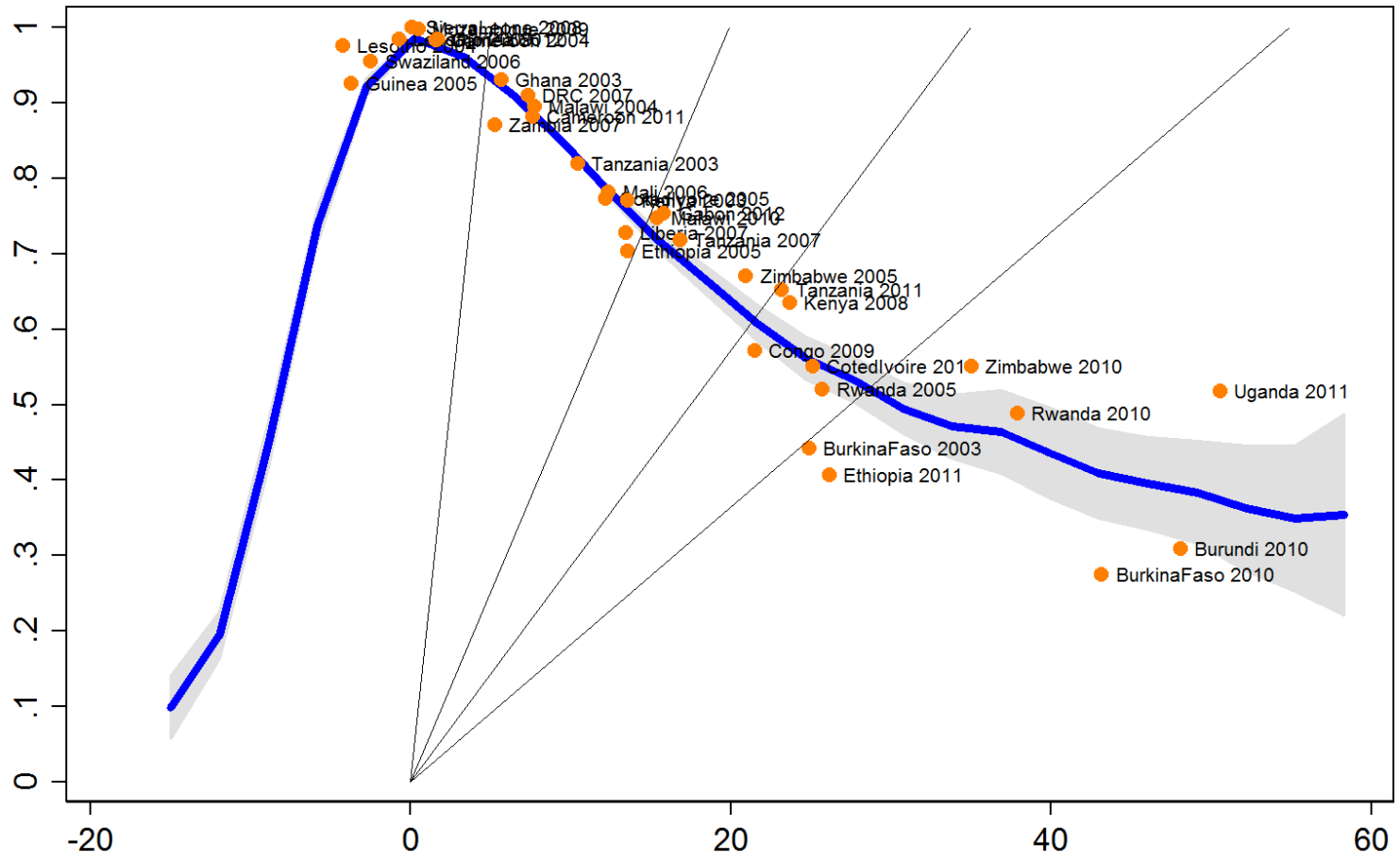
- Stage $\leq$0: Cameroon 2004, Guinea 2005, Guinea 2012, Lesotho 2004/05, Lesotho 2009/10, Mozambique 2009, Senegal 2005, Sierra Leone 2008, Swaziland 2006/07.
- Stage 1: Cameroon 2011, Cote d'Ivoire 2005, Democratic Republic Congo 2007, Ethiopia

[12] To capture the evolution of the HIV-Education gradient over the HIV epidemic we need to interact education with either some specific function (e.g., a quadratic or cubic polynomial) of $\omega(\tau,\zeta)$ or with a discretized version of $\omega(\tau,\zeta)$. We prefer to follow the latter approach and partition the continuous variable $\omega(\tau,\zeta)$ into the discrete pieces which delivers a cleaner interpretation of the estimates of the HIV-Education gradient across stages.

2005, Ghana 2003, Kenya 2003, Liberia 2006/07, Malawi 2004/05, Mali 2006, Niger 2006, Senegal 2010/11, Tanzania 2003/04, and Zambia 2007.

- Stage 2: Gabon 2012, Malawi 2010, Tanzania 2007/08, Zimbabwe 2005.
- Stage 3: Congo Brazaville 2009, Cote d'Ivoire 2011, Kenya 2008/09, Rwanda 2005, Tanzania 2011/12, Zimbabwe 2006.
- Stage $\geq$4: Burkina Faso 2003, Burkina Faso 2010, Burundi 2010, Cote d'Ivoire 2012, Ethiopia 2011, Niger 2012, Rwanda 2010/11, Uganda 2011 and Zimbabwe 2010/11.

Figure 3: Stages of the HIV Epidemic: DHS Sample



Source: The plot shows the location of SSA Countries (DHS Sample) on the 2D-normalized space at the time of DHS data collection. The vertical axis is the normalized HIV prevalence, and the horizontal axis is the normalized time (Subsection 2.2.2). The orange markers in the scatterplots represent a DHS dataset. The plotted trends are locally weighted polynomials with 95% confidence intervals.

Figure 3 depicts these allocations in two dimensions in the normalized space $T \times \Lambda$. It is identical to panel (d) in Figure 2 with the addition of the arrays initiating from the origin that define the breakdown of discrete stages of the HIV epidemic defined above. It is relevant to note that the DHS observations covers the entire evolution of the HIV epidemic except its initial rise, which is not surprising given that the first DHS surveys with HIV testing were conducted in 2003. In any case, the DHS observations provide a large degree of heterogeneity across country positions over the normalized HIV epidemic. The large heterogeneity across stages of the HIV epidemic is sufficient to provide reliable nonstationary estimates of the HIV-Education gradient.[13]

2.2.4 Discussion

The normalization procedure that we have proposed to define stages of the HIV epidemic is drawn upon ideas developed by macroeconomists to define the stages of the demographic transition and the stages of economic development.

As it is the case with the HIV epidemic, the demographic changes behind the demographic transition occur at different calendar years and at different speeds for different countries. To take this into account, the demographic model controls for calendar year and speed of stage completion with a time-normalization analogous to the one we propose, and in turn defines the demographic stages in a similar way to ours. Precisely, many countries have gone (or as still going) through a first demographic stage of high mortality and fertility, followed by a second demographic stage in which mortality drops and, with some period lag, fertility declines, and finally a third demographic stage of low mortality and low fertility [Lee, 2003]. This description of the stylized joint behavior of mortality and fertility exemplifies how the calendar year, the speed at which each country completes each demographic stage, and the country-specific levels of the variables of interest (mortality and fertility) needs to be controlled for to recover stylized facts across stages of the demographic transition.

With regard to the aggregate stages of economic development (or structural transformation out from agriculture), the heterogeneity across time and across space depends on the fact that countries take off at different calendar years and move across stages of economic development at different speeds [Hansen and Prescott, 2002, Gollin et al., 2002, 2007]. For example, China's income per capita raised by a factor of 6 between 1989 and 2009, while it took the U.K. from 1820 to 1970 to generate such growth. That is, China has grown at roughly 7.5 times the speed of the first industrial revolution [Bolt and van Zanden, 2013]. With a definition of the stages of

[13] For the country of Niger and Senegal, we lack of projected paths, and therefore we cannot implement the time-normalization. We allocate the DHS surveys of these countries across stages by visual inspection. We find reassuring that the exclusion of Niger and Senegal from our analysis does not alter our results.

development that nets out calendar years and speed, we can study stylized economic relationships along the process of economic development such as the skill wage premium [Buera et al., 2015].

In sum, the spirit of our exercise is similar to what has been done to describe the stages of demographic transitions and the stages of economic development. First, we provide an algorithm that defines aggregate stages of the HIV epidemic netting out the calendar year effects, the country-specific speeds of the HIV epidemic, and the HIV prevalence levels. Second, we study the stylized dynamic relationship between HIV status and education across stages of the HIV epidemic.

2.3 The HIV-Education Gradient

Our empirical analysis consists of posing a simple econometric specification suitable for documenting the potentially nonstationary behavior of the HIV-Education gradient.

2.3.1 Econometric Specification

We consider a linear probability model (LPM) where the HIV-Education gradient is allowed to change over the stages of the HIV epidemic (j) defined in Section 2.2.3. Since for a large set of countries there are at least two cross-sections in our sample, and individuals in different periods are not the same people for each country (g), we index the variables by a double subscript. Namely, $i(t) \in \{ 1, ..., N_t\}$ denotes the individuals in cross-section t. Let $s_{i(t),t,j}$ denote the educational attainment of an individual $i(t)$ that lives in stage j of the HIV epidemic at time t. Also, let $y_{i(t),t,j}$ be the individual's HIV status, a dummy variable equal to one if the individual HIV testing result is positive and zero otherwise. We estimate a linear projection of the type,

$$y_{i(t),t,j} = \alpha_0 + \sum_{j>0} \alpha_j \mathbf{1}_j + \left(\gamma_0 + \sum_{j>0} \gamma_j \mathbf{1}_j \right) s_{i(t),t,j} + \beta\, x_{i(t),t} + \psi\, m_{g,t} + \theta_t \mathbf{1}_t + \theta_g \mathbf{1}_g + \varepsilon_{i(t),t}, \quad (6)$$

where $\mathbf{1}_j$ is an indicator function that is equal to one when the stage of the HIV epidemic is j and zero otherwise. That is, if the stage of the epidemic is $j = 0$ then the intercept is α_0 and the slope is γ_0. However, for each the stage of the epidemic is $j > 0$, the associated intercept is $(\alpha_0 + \alpha_j)$ and the slope is $(\gamma_0 + \gamma_j)$. Namely, γ_j is the difference in the HIV-Education gradient between individuals that are in stage j and stage 0 of the epidemic. This implies that the HIV-Education gradient is $\gamma_0 + \gamma_j$ for each epidemiological stage j. We cluster the individual observations at the country level to account for any unobserved shock that correlates observations within a country. Given that the number of countries is 25, we use the wild cluster bootstrap from Cameron et al.

[2008] to get better approximations to asymptotically valid standard errors.[14]

The vector $x_{i(t),t}$ corresponds to a set of individual characteristics that are likely to be correlated with both education and HIV status. Hence, controlling for these characteristics reduces the impact of omitted variables bias. Precisely, we find it is important to control for the type of area in which agents live because the HIV prevalence is, on average, higher in urban than in rural areas (respectively, 6.73% and 4.41% in our whole sample), and it is in urban areas where adult education levels are also higher (the average number of years of schooling in rural area is 3.23, while in urban areas reaches 3.83). Thus, a positive association between education and HIV may be driven by the fact that people living in urban areas are both more likely to be HIV-positive and more educated. Similarly, we also control for age because HIV prevalence is increasing with age (the DHS age sample is 15-49), and education is negatively correlated with age as younger cohorts are more educated than older cohorts.

We also control for time varying country-specific economic variables, $m_{g,t}$, which correct for the stage of economic development in which each country is. To do so, we use measures of output per capita and share of agricultural output following the literature on structural transformation. We also include year dummies (θ_t) and country dummies (θ_g) to pick up any spurious correlation between the regressors and the dependent variable. To the extent that such contextual effects affect all individuals in a country in a similar manner, the country dummies will sweep them up. All our specifications are weighted least squares regressions, where the weights are proportional to the relative population size of each country. By doing so, when we pool a number of countries in the same stage of the HIV epidemic, the relative DHS sample size of a given country corresponds to the relative population size of the country. We then combine these weights with the individual weights provided by the DHS surveys.

2.3.2 Results

We first consider a linear probability model where the dependent variable is the individual HIV status. The estimates of the HIV-Education gradient using the whole sample are reported in Table 3. We then re-conduct our analysis separately for women and men in Table **??**.

[14] The results are robust to clustering at the country-year level. While we believe it would be interesting to explore also the within-country variation (e.g., across regions), it is not feasible to recover the epidemiological stages using our algorithm proposed in section 2.2 due to data limitations about the evolution of the HIV epidemic at the regional level; recall that the algorithm would require complete time-series of HIV prevalence for each region within a country and these estimates are generally provided at the national level [UNAIDS, 2015].

Stationary Specification To study the stationary gradient we restrict the econometric model (6) with $\alpha_j = \gamma_j = 0$ for all $j > 0$. We find that the stationary HIV-Education gradient is highly significant and positive (column 1, Table 3). The probability of being HIV infected increases by 0.43% per year of schooling. This suggests that completing five additional years of schooling increases the probability of being HIV positive by 2.15%, which is not small if we consider that the HIV prevalence is 5.18% in our sample.[15] Further, the probability of being HIV positive is higher for women (by 2.24%), for urban areas (by 2.12%), and it increases significantly with age (0.25% per year of age).[16] Aggregate variables denoting the stage of development such as the agricultural share of output and output per capita are negatively related with the probability of being infected. Next we explore how much this gradient changes as the HIV epidemic evolves.

Non-Stationary Specification Our non-stationary specification follows the econometric model in (6). Our key finding is that the HIV-Education gradient is significantly nonstationary and displays a positive-zero-positive U-shaped pattern over the stages of the HIV epidemic.

Focusing on our benchmark specification (column 2, Table 3), we find that at Stage 0 an additional schooling year raises the probability of being infected by $\gamma_0 = 1.12\%$. That is, for individuals in an economy that is at early stages of the epidemic the HIV-Education gradient is significantly positive and remarkably high (roughly three times larger than that of the stationary specification). Interestingly, as the HIV epidemic evolves, the HIV-Education gradient rapidly declines. At Stage 1 the rise in the probability of being infected associated with one additional year of schooling is $\gamma_0 + \gamma_1 = 0.51\%$, i.e., less than one-half of its value at Stage 0, and it is significantly different from zero at 1% level (column 2, panel A of Table 4). The educational disparities in HIV then vanish as the epidemic reaches Stage 2, where we cannot reject the null that $\gamma_0 + \gamma_2 =$ -.04% is different from 0 (column 2, panel A in Table 4). As we move away from Stage 2, the HIV-Education gradient becomes increasingly positive as the epidemic evolves with $\gamma_0 + \gamma_3 = 0.19\%$ and $\gamma_0 + \gamma_4 = 0.48\%$ in Stages 3 and 4, respectively. This way, the HIV-Education gradient bounces back reaching a significant gradient in Stage 4 that is almost half the size of the gradient in Stage 0. Note that both the initial decline of the HIV-Education gradient from Stage 0 to Stage 2 and its posterior rebound from Stage 2 to Stage 4 are both significant. The size of the rebound from Stage 2 to Stage 4 is a significant 0.52% (column 2, panel B in Table 4). We conclude that the HIV-Education gradient exhibits a positive-zero-positive U-shape pattern over stages of the HIV epidemic. To illustrate this pattern, Figure 4 shows the isomorphic representation of the estimated HIV-Educ gradient across stages of the epidemic, $\gamma_0 + \sum_{j>0} \gamma_j \mathbf{1}_j$

[15] These findings are consistent with those obtained by Fortson [2008], who specifies a similar stationary econometric model for five DHS countries.

[16] While we introduce age linearly, we do find that the estimated coefficients for the HIV-Education gradient are robust when age enters non-linearly.

Table 3: The HIV-Education Gradient

(A) *HIV Status*	(1)	(2)	(3)	(4)	(5)
Education	0.0043***	0.0112***	0.0098***	0.0040***	0.0037***
	(0.0006)	(0.0007)	(0.0010)	(0.0003)	(0.0003)
Education * Stage1		-0.0059***	-0.0046***	-0.0010***	-0.0008**
		(0.0008)	(0.0010)	(0.0003)	(0.0003)
Education * Stage2		-0.0116***	-0.0103***	-0.0027***	-0.0024***
		(0.0007)	(0.0011)	(0.0002)	(0.0003)
Education * Stage3		-0.0093***	-0.0076***	-0.0020*	-0.0018
		(0.0015)	(0.0011)	(0.0011)	(0.0011)
Education * Stage4		-0.0064***	-0.0051***	-0.0015***	-0.0012***
		(0.0008)	(0.0012)	(0.0003)	(0.0003)
Male	-0.0224***	-0.0229***	-0.0228***	-0.0223***	-0.0224***
	(0.0027)	(0.0027)	(0.0021)	(0.0020)	(0.0019)
Age	0.0025***	0.0025***	0.0025***	0.0025***	0.0025***
	(0.0004)	(0.0004)	(0.0003)	(0.0003)	(0.0002)
Urban Area	0.0212***	0.0197***	0.0227***	0.0280***	0.0285***
	(0.0044)	(0.0045)	(0.0040)	(0.0027)	(0.0027)
Stage 1	-0.0023	0.0124***	0.0111***	-0.0055***	0.0088***
	(0.0048)	(0.0036)	(0.0041)	(0.0008)	(0.0019)
Stage 2	0.0103	0.0498***	0.0598***	-0.0012	0.0200***
	(0.0078)	(0.0089)	(0.0100)	(0.0025)	(0.0032)
Stage 3	-0.0094	0.0197	0.0314***	-0.0102**	0.0110**
	(0.0128)	(0.0122)	(0.0096)	(0.0052)	(0.0043)
Stage 4	-0.0032	0.0131***	0.0394***	-0.0147***	0.0032
	(0.0042)	(0.0030)	(0.0072)	(0.0019)	(0.0029)
Agricultural Share	-0.0029***	-0.0029***	-0.0031***	0.0021***	-0.0008***
	(0.0002)	(0.0002)	(0.0003)	(0.0004)	(0.0003)
Output per Capita	-0.0000***	-0.0000***	-0.0000***	0.0000**	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Constant	0.0806***	0.0615***	0.0307***	-0.1363***	-0.1832***
	(0.0089)	(0.0072)	(0.0089)	(0.0344)	(0.0175)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes
Sample Size	402,766	402,766	402,766	402,766	402,766

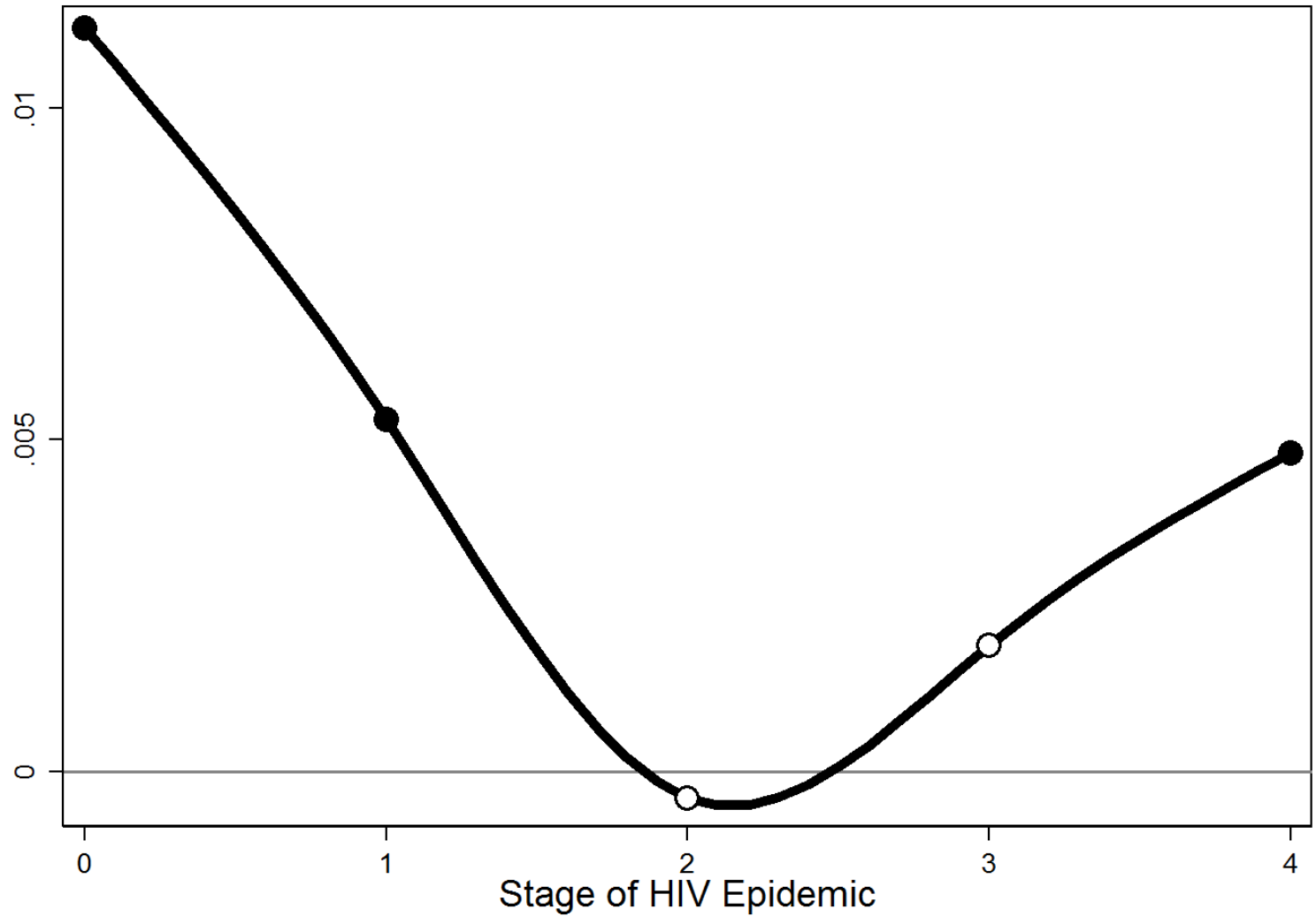
(B) *HIV Status*	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
Education	0.0038***	0.0022***	0.0014***	0.0020*	0.0025***
	(0.000)	(0.000)	(0.000)	(0.0011)	(0.000)
Year-Country Dum.	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes
Sample Size	66,322	119,700	48,615	50,535	118,425

Notes: All specifications use the "Full Sample" described in Section 2.1 and the same set of controls. In Panel (A), Column (1) reports the results for the stationary specification, and columns (2) to (5) report the results for the non-stationary specification. We add year dummies in column (3), country dummies in column (4) and year-country dummies in column (5). Panel (B) reports the estimates of the HIV-Education gradient for each stage separately. Standard errors are clustered at the country level using the wild cluster bootstrap from Cameron et al. [2008], and reported in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

for each j (panel A, Table 4).[17] Similar results are attained if we consider only the sexually active subsample (Appendix Table A-1), with an HIV-Education gradient of 1.23% at Stage 0, 0.59% at Stage 1, -0.06% at Stage 3, 0.22% at Stage 3 and 0.56% at Stage 4.

[17] Our results also hold under a Probit specification. The partial effects (and p-values) are as follows: $\gamma_0 =$ 0.691% (0.000), $\gamma_0 + \gamma_1 =$0.465% (0.000), $\gamma_0 + \gamma_2 =$0.018% (0.674), $\gamma_0 + \gamma_3 =$0.135% (0.133), and $\gamma_0 + \gamma_4 =$0.339% (0.004). Regarding the rebound, we also reject the null that $\gamma_4 - \gamma_2 =$0 (0.018).

Figure 4: The HIV-Education Gradient Across Stages of the Epidemic



Notes: This graph plots the benchmark estimates of the HIV-Education gradient using the full sample (with year controls). For each stage j we plot $\left(\gamma_0 + \sum_{j>0} \gamma_j \mathbf{1}_j\right)$. We construct this estimates from column 3 of Table 3 (also reported in column 3, panel A, Table 4). Significance at 10%, 5%, and 1% is represented by, respectively, markers with open circles, markers with medium transparency fill, and markers with solid fill. We use a cubic spline for interpolation across stages.

The U-shape pattern of the HIV-Education gradient across stages of the HIV epidemic is robust to the addition of year dummies, country dummies, and year and country dummies (columns 3 to 5 in Table 3). When we control for year dummies, the HIV-Education gradient is significantly different from zero for all stages except Stage 2 showing the same positive-zero-positive behavior as our benchmark (column 3 in Table 4). When we control for country dummies the magnitude of the HIV-Education gradient is smaller (columns 4 and 5, panel A, Table 4) but the U-shape pattern is preserved and remains significant, that is, γ_2 and γ_4 are significantly different from each other (column 4 and 5, panel B, Table 4). While this specification is quite demanding, as it exploits only variation witin each country, the direction of the results discussed above remains unchanged. Finally, the same U-shape pattern is displayed in panel B of Table 3 where we re-estimate the model for each stage separately.

To summarize, the HIV-education gradient shows a U-shape positive-zero-positive pattern across stages of the epidemic. This stylized fact is twice more sizable for women than for men at early stages of the epidemic, the gradients across genders tend to equalize in later stages.

2.4 Further Evidence: Risky Sex and ARVs

Our previous theoretical interpretation suggests a parallel evolution between the HIV-Education gradient and education disparities in risky sexual behavior. We provide an empirical investigation of this phenomenon in this section.

The Risky Sex-Education Gradient The margins of risky sexual behavior that we study are: (i) the number of sex partners other than spouses (i.e., extramarital partners) during past 12 months, i.e., the extensive margin of sexual behavior,[18] and (ii) a dummy variable equal to 1 if the respondent used a condom during the last intercourse.We label the first Risky Sex-Education gradient as the Partners-Education gradient and the second as the Condom-Education gradient.

The results for the Partners-Education gradient are in panel A of Table 5. We report the results separately for women and for men, and we follow the same econometric specifications described for the HIV-Education gradient in Table **??**. The more educated have significantly more sexual partners that the less educated. An additional year of schooling increases the chances of having an extramarital partners by 1.77% for women and by 2.85% for men (column (1) and (6), panel A, Table **??**).[19] The non-stationary specification uncovers an interesting inverted U-

[18] For the extensive margin of risky sexual behavior we use the number of sex partners other than spouses (i.e., extramarital partners) in the past 12 months and note that for individuals who are single or do not cohabit, all sex partners are extramarital.

[19] Note that the females' Partners-Education gradient is smaller (about two thirds) than that of males but it shows a similar pattern over the stages of the epidemic. One potential caveat of this analysis is that women

Table 4: Additional Inference

(A) *HIV-Education Gradient*	(1)	(2)	(3)	(4)	(5)
γ_0	0.0043***	0.0112***	0.0098***	0.0040***	0.0037***
	(0.0006)	(0.0007)	(0.000)	(0.0003)	(0.0003)
$\gamma_0 + \gamma_1$		0.0053***	0.0052***	0.0029***	0.0029***
		(0.0005)	(0.000)	(0.0003)	(0.0003)
$\gamma_0 + \gamma_2$		-0.0004	-0.0005	0.0013***	0.0013***
		(0.0003)	(0.228)	(0.0001)	(0.0001)
$\gamma_0 + \gamma_3$		0.0019	0.0022***	0.0020*	0.0019*
		(0.0014)	(0.005)	(0.0011)	(0.0011)
$\gamma_0 + \gamma_4$		0.0048***	0.0047***	0.0025***	0.0025***
		(0.0005)	(0.000)	(0.0002)	(0.0001)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes

(B) *Rebound*	(1)	(2)	(3)	(4)	(5)
$\gamma_4 - \gamma_2$		0.0052***	0.0051***	0.0011***	0.0012***
		(0.0006)	(0.0005)	(0.0001)	(0.0001)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes

Notes: The underlying econometric models are as specified in the columns of Table 3. Column (1) reports the tests results for the stationary specification. Columns (2) to (5) report the tests results for the non-stationary specification. Standard errors are clustered at the country level using the wild cluster bootstrap from Cameron et al. [2008], and reported in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

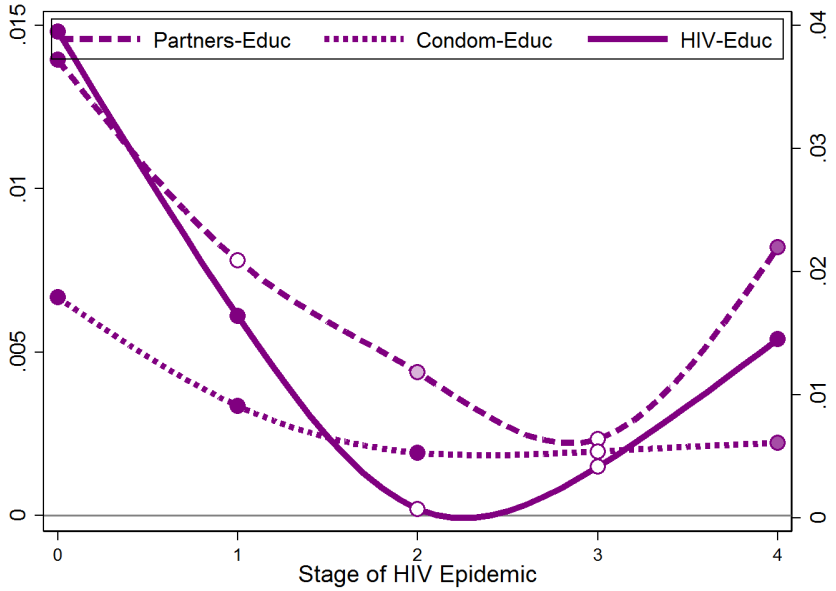
shaped pattern of the Partners-Education gradient that represents our main finding in this section. The Partners-Education gradient first decreases (between aggregate stages 0 and 2) and then increases (between aggregate stages 2 and 4) for both women and men, and this dynamics are significant. The dynamics across stages of the epidemic show that the evolution of the Partners-Education gradient is remarkably consistent with the pattern of the HIV-Education gradient, as it is predicted by our theory in the previous section. To see this, Figure 5 shows separately for women (panel A) and for men (panel B) the isomorphic representation of the HIV-Education gradient and the Partners-education gradient (as in Figure 4, the significance of the gradients is denoted by the color of the marker). The sizes of the gradients are different, being larger for the Partners-Education gradient, which implies a positive elasticity of less than one from risky sex to HIV. In contrast, the Condom-Education gradient is reported in panel B of Table 5. In the stationary specification (columns 1 and 4) we find that the more educated women and men use, on average, more condoms than the less educated ones. Interestingly, we do not find a significant pattern across stages of development in the Condom-Education gradient. After Stage 0, the Condom-Education gradient remains positive but relatively constant across stages of the epidemic for both women and men. Finally, knowledge about the transmission mechanisms of HIV might affect sexual behavior [Dinkelman et al., 2006, Dupas, 2011a, Duflo et al., 2015a]. We document that more-educated individuals acquire more information about HIV transmission than less-educated individuals at earlier stages of the epidemic, but these educational differences in knowledge remain constant as the epidemic evolves. This way, while knowledge might affect the HIV-Education gradient at early stages of the epidemic, this effect should rapidly vanish after the first stages.[20]

To sum up, the evolution of the Partners-Education gradient is consistent with the evolution of the HIV-Education gradient, as predicted by the theory. The evolution of educational disparities in the number of extramarital partners help explain both the decline and rebound of the HIV-Education gradient across stages of the epidemic.

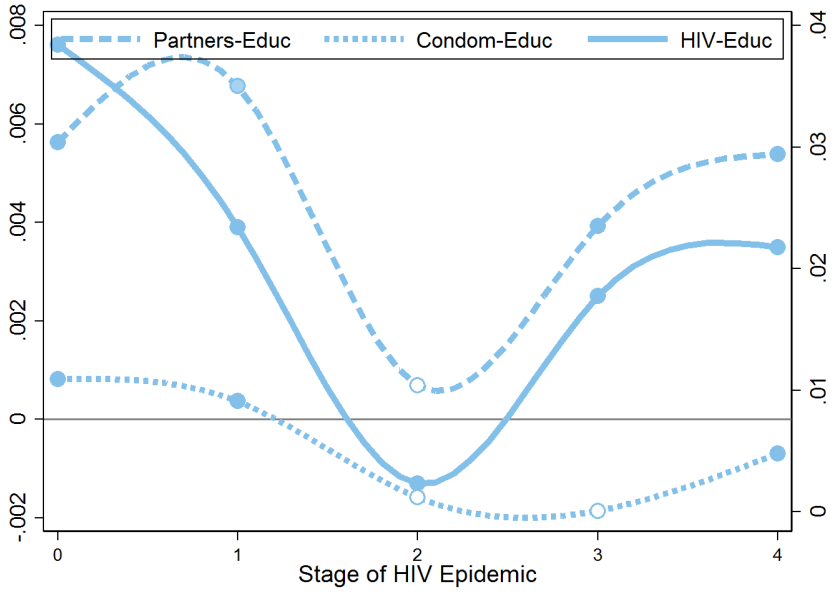
might under-report risky sexual behavior more than men, as it was pointed out by Smith [1992] and Gersovitz et al. [1998]. However, more recently, Helleringer et al. [2009] find that men and women are equally likely to under-report risky sexual behavior when using sexual network data from Likoma Island, Malawi. In our analysis, as long as misreporting occurs systematically across all stages of the epidemic, the shape of the Partners-Education gradient for women over stages will not be biased.

[20] To this end, we consider two DHS questions regarding ways to avoid HIV infection that are directly related to the risky sex margins studied in the previous subsection. Specifically, respondents answer two questions: (i) *"Can you (the respondent) reduce the chances of getting HIV by having one sex partner who has no other partners?"* and (ii) *"Can you (the respondent) reduce the chances of getting HIV by always wearing a condom?"*. We estimate the education gradients in these knowledge variables. Our results are in appendix Table A-2.

Figure 5: The Risky Sex-Education Gradient: Evolution Across Stages of the HIV Epidemic



(a) Women



(b) Men

Notes: The HIV-Education gradient is plotted on the left vertical axis. The Partners-Education gradient and the Condoms-Eduction gradient are plotted on the right vertical axis. For each stage j we plot $\left(\gamma_0 + \sum_{j>0} \gamma_j \mathbf{1}_j\right)$. The specification we plot is with year controls. That is, for the HIV-Education gradient we use column 3 (8) in Table **??** for respectively women (top panel) and men (bottom panel). For the Partners-Education and Condoms-Education gradient we use column 3 (8) in Table 5 for respectively women (top panel) and men (bottom panel). Significance at 10%, 5%, and 1% is represented by, respectively, markers with open circles, markers with medium transparency fill, and markers with solid fill. We use a cubic spline for interpolation across stages.

Table 5: The Risky Sex-Education Gradient: Women and Men Separately

(A) Number of Extramarital Partners

	Women					Men				
Risky Sex	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Education	0.0177**	0.0372***	0.0348***	0.0294***	0.0296***	0.0285***	0.0304***	0.0269***	0.0290***	0.0295***
	(0.029)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.002)	(0.000)	(0.000)
Education * Stage1		-0.0163**	-0.0222***	-0.0204***	-0.0199***		0.0046	-0.0071	-0.0167**	-0.0167**
		(0.025)	(0.000)	(0.006)	(0.008)		(0.742)	(0.515)	(0.050)	(0.029)
Education * Stage2		-0.0254***	-0.0267***	-0.0243***	-0.0249***		-0.0200	-0.0238***	-0.0290***	-0.0294***
		(0.006)	(0.000)	(0.003)	(0.003)		(0.125)	(0.007)	(0.000)	(0.000)
Education * Stage3		-0.0308***	-0.0288***	-0.0201**	-0.0213**		-0.0069	-0.0047	0.0010	0.0003
		(0.002)	(0.000)	(0.019)	(0.018)		(0.550)	(0.588)	(0.932)	(0.980)
Education * Stage4		-0.0152	-0.0134	-0.0134*	-0.0138*		-0.0010	0.0019	0.0021	0.0018
		(0.107)	(0.185)	(0.081)	(0.087)		(0.928)	(0.845)	(0.840)	(0.866)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes	No-No	No-No	Yes-No	No-Yes	Yes-Yes
Sample Size	227,935	227,935	227,9358	227,935	227,935	174,831	174,831	174,831	174,831	174,831

(B) Condom Use in Last Intercourse

	Women					Men				
Risky Sex	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Education	0.0079***	0.0192***	0.0193***	0.0125***	0.0126***	0.0066***	0.0129***	0.0141***	0.0162***	0.0166***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Education * Stage1		-0.0102***	-0.0114***	-0.0054**	-0.0055***		-0.0038	-0.0054	-0.0064*	-0.0066*
		(0.001)	(0.142)	(0.011)	(0.002)		(0.0406)	(0.122)	(0.084)	(0.065)
Education * Stage2		-0.0135***	-0.0132***	-0.0075***	-0.0075***		-0.0109***	-0.0111**	-0.0149***	-0.0156***
		(0.000)	(0.000)	(0.000)	(0.000)		(0.010)	(0.028)	(0.000)	(0.000)
Education * Stage3		-0.0138	-0.0133	-0.0068	-0.0068		-0.0101	-0.0108	-0.0131	-0.0136
		(0.164)	(0.000)	(0.482)	(0.501)		(0.611)	(0.640)	(0.592)	(0.585)
Education * Stage4		-0.0134**	-0.0133***	-0.0073**	-0.0080**		-0.0078**	-0.0088*	-0.0066***	-0.0074***
		(0.000)	(0.000)	(0.024)	(0.007)		(0.017)	(0.062)	(0.004)	(0.001)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes	No-No	No-No	Yes-No	No-Yes	Yes-Yes
Sample Size	163,883	163,883	163,883	163,883	163,883	120,840	120,840	120,840	120,840	120,840

Notes: In panel (A) we report the marginal effects of the associated Tobit model where the endogenous variable is the number of extramarital partners in the past 12 months. In panel (B) we report the coefficients of a linear model where the endogenous variable is binary and refers to use of condom in last sexual intercourse. In both panels we include the same set of controls and fixed effects as in our benchmark specifications in Table **??**. Standard errors are clustered at the country level using the wild cluster bootstrap from Cameron et al. [2008], and reported in parenthesis.* significant at 10%; ** significant at 5%; *** significant at 1%.

Antiretroviral Therapy (ART) The effects of ART on the HIV-Education gradient are potentially ambiguous. On the one hand, ART increases survival probabilities [Greenwood et al., 2013] and decreases the degree of infectiousness of the HIV+ population that takes ART [Apondi et al., 2011]. These effects, respectively, an increase in γ^+ and a reduction in $\lambda(x)$, unambiguously increase sexual behavior by reducing the future marginal cost of today's risky sex (panel (b) and (c) in Appendix Figure **??**). On the other hand, if agents have to pay for this treatment they will reduce today's risky sex because of the higher future marginal cost.[21] As long as the monetary cost of ART does not offset the reductions in the marginal cost of sex through γ and λ, ART will increase risky sex (see panel (b) and (c) in Appendix Figure **??**).

Treatment is indeed costly. While prices of the most common first-line ART regiments have declined over time, they remain relatively high for a vast majority of the population, with median prices in low and medium income countries of USD115 per patient per year (ppy) in 2013. The prices of USD330 (ppy) in 2013 for the second-line treatment, and more than USD1,500 for the third-line treatment, [WHO, 2014]. This is prohibitive for a vast majority of SSA households.[22] In this context, if the more educated have more access to ART , then ART might help explain the rebound in the most advanced stages. The findings in Table 6 point in this direction. We estimate the HIV-Education gradient by epidemiological stage with and without ART controls in panel A and B, respectively. Note that when we include ART, the estimate of HIV-Education gradient in the most advanced stages of the epidemic decreases with respect to the counterpart in panel A. This suggests that the provision of ART partially accounts for the rebound in the gradient. This result should be taken with a grain of salt though, since we cannot directly test the impact of the education gradient on HIV because the DHS does not provide information of ART at the individual level.

[21] In our model, a monetary cost of ART would imply a change in the future marginal cost of having risky sex today, which becomes $\beta\lambda_x(\gamma^- u(y_1(s)) - \gamma^+ u(y_1(s) - cost))$.

[22] For example, income per capita in Malawi is on average USD250 in 2014, and the average income per capita in SSA is USD1638. We also expect more-educated individuals to have greater access to ART treatments for several obvious reasons—they (i) are more likely to live in the city (e.g., in Malawi, anyone who has a university degree is likely to live in the two largest cities, Lilongwe or Blantyre, where the ART drugs are available), (ii) have better transportation (do not have to walk several miles to refill prescriptions), or (iii) have access to someone in a hospital who can help them gain priority status when necessary to obtain ART .

Table 6: The HIV-Education Gradient: Specification by Stage

(A) *HIV Status*	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
Education	0.0044***	0.0022***	0.0014***	0.0020*	0.0025***
	(0.000)	(0.000)	(0.000)	(0.081)	(0.000)
Year-Country Dum.	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes
Sample Size	58,560	112,024	48,615	50,535	118,425

(B) *HIV Status with ART*	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
Education	0.0043***	0.0029***	0.0016***	0.0011	0.0015***
	(0.000)	(0.0003)	(0.0002)	(0.0021)	(0.0002)
ART Coverage	0.0003*	-0.0025***	0.0015***	-0.0012***	-0.0003***
	(0.0002)	(0.0005)	(0.0000)	(0.0001)	(0.0000)
Education * ART Coverage	-0.0000	-0.0000	-0.0000***	0.0000***	0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Year-Country Dum.	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes
Sample Size	58,560	112,024	48,615	50,535	118,425

Notes: We apply our stationary specification of the HIV-Education gradient separately for each stage of the epidemic. We include year and country fixed effects in all columns. In both panels we include the same set of controls as in our benchmark specifications in Table 3. We exclude Senegal (in Stage 0) and Niger (in Stage 1) since WPP does not provide information about ART coverage for these countries. Standard errors are clustered at the country level using the wild cluster bootstrap from Cameron et al. [2008], and reported in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

3 The Model

This is a perpetual youth economy with heterogeneous agents where the mortality rate is a function of the endogenously determined individual HIV status. There is an exogenous fertility rate that feeds newborns into the economy at a constant rate f. Agents differ in education level $e \in \mathcal{E}$, income shock $s \in \mathcal{S}$, sex type $i \in \mathcal{I}$, HIV status $h \in \mathcal{H}$ and whether the agent takes antiretroviral drugs(ARVs) against HIV $d \in \mathcal{D}$. The education level is permanent and exogenously given at the beginning of life. The sex type states whether agents are sex consumers or sex producers and is permanent and exogenously determined.[23]

[23] That is, we focus on endogenezing the intensive margin of sex and leave out from the model the endogenous decision on who becomes a sex consumer or a sex producer. This is clearly a caveat of this exercise which is

The economy transits across aggregate HIV stages, $g \in \mathcal{G}$ ={-1, 0, 1-2, 3-4}, through a sequence of unexpected shocks that reflect the evolution of the HIV epidemic. At stage -1, the economy lives in a stationary Pre-HIV epidemic era; at stage 0, the HIV epidemics starts, but agents are unaware of its workings; at stage 1-2, agents start to learn the sexual nature of the process of HIV infection and the speed of learning differs by education group; finally, at stage 3-4, ARVs are introduced.

Let us cast the household problem recursively within and across aggregate HIV stages and then explain it.

[Stage -1] The Pre-HIV Era

At this stage there are no individuals infected with HIV. At any given period t, agents with education level, $e \in \mathcal{E}$ and income shock $s \in \mathcal{S}$, solve the following problem depending on their sex type, $i \in \mathcal{I}$:

Risky sex-consumer household problem. Sex consumers choose consumption c, and non-marital risky sex x to solve the following dynamic problem:

$$V(e,i,s,\Phi) = \max_{c\geq 0, x\geq 0} \chi u(c,x) + \beta\gamma \sum_{s'|s} \pi(s'|s) V(e,i,s',\Phi') \tag{7}$$

s.t

$$c + p(\Phi)x = zy(e)s \tag{8}$$

$$s = (s_n) \text{ with } s_n \in \mathcal{S} \text{ and stochastic matrix } \pi \tag{9}$$

That is, sex consumers derive utility from c and x with preference parameter χ that depends on the individual' HIV status, and discount future at factor β times a survival probability γ. The preference parameter χ controls the level of utility derived from c and x. Moreover, $u(c,x)$ is concave, continuous and twice diferenciable.[24].We use consumption as numeraire and the relative price of sex is denoted by p. Labor income is the product of a permanent component $y(e)$ that depends on the level of education, a transitory component s that follows a Markov process with

overcome, for example, by **?**. However, the fact that we introduce a substantial degree of endogenous heterogeneity within each of the two groups implies that part of our sex consumers will consume a positive but negligible amount of sex, and part of our sex producers will produce a negligible amount of sex. The size of this population, for which sex transactions are small, which is endogenous, determines the size of the population that consumes/sells risky sex. Moreover, the joint distribution of education groups $e \in \mathcal{E}$ and sex types $i \in \mathcal{I}$ (with respective proportions $\vartheta_{e=1,i}, \vartheta_{e=1,-i}, \vartheta_{e=0,i}, \vartheta_{e=0,-i}$) is endogenous throughout the model. However, we are still required to provide an exogenous measure for $t = 0$, see Section 4 for more detail.

[24]We choose an additive separable CRRA functional form with common parameter ξ, at this stage the preference parameter χ is set to one for all agents.

given transition matrix π, and a permanent component z that depends on the individual's health status. Since there is no HIV, z is set equal to one for everyone. We assume that $y(e)s > y(e)\widetilde{s}$ for $s > \widetilde{s}$ and $\forall\, e \in \mathcal{E}$ [25]

Risky sex-producer household problem. Sex-producer households choose consumption c and the fraction of time devoted to sex production l to solve the following dynamic problem:

$$V(e,-i,s,\Phi) = \max_{c\geq 0, 1\geq l\geq 0} \chi u(c) + \beta\gamma \sum_{s'|s} \pi(s'|s) V(e,-i,s',\Phi') \tag{10}$$

s.t

$$c = z[p(\Phi)l^{\alpha} + y(e)s(1-l)], \tag{11}$$

Where s follows the income shock process (9). Sex-producer households derive utility from the consumption good, but not from risky sex. The production of sex follows a technology $x = l^{\alpha}$ using time l with decreasing returns to scale, $\alpha \in (0,1)$. This technology puts an upper bound to the amount of sex produced. Notice that labor is inelastically supplied, this way, l denotes the fraction of labor allocated to the production of risky sex and the remaining labor $(1-l)$ is allocated to sex production. The permanent component z affects both the ability to generate labor market income and sex production. This idea pursues the notion that HIV+ individuals do not necessarily increase the production of sex as a response to their lower labor income.

At any point in time in the pre-HIV stage -1, the economy is summarized by the joint distribution Φ of individual states (e,i,s). The aggregate state variable evolves according to:

$$\Phi' = H(\Phi) \tag{12}$$

Where the function $H : \mathcal{M} \to \mathcal{M}$ is the aggregate law of motion, mapping distributions to distributions. H summarizes the distribution of type and education evolves from one period to the next, however this is exactly what a transition function tell us. Define the transition function $\mathcal{Q} : \mathcal{Z} \times \mathcal{B}(\mathcal{Z}) \to [0,1]$ by:

$$Q((e,i,s)(\mathcal{E},\mathcal{I},\mathcal{S})) = \gamma \quad \forall \ (e,i,s) \in \mathcal{Z} \ \text{ and } \ (\mathcal{E},\mathcal{I},\mathcal{S}) \in \mathcal{B}(\mathcal{Z})$$

Where $\mathcal{Z}$ consists of all n-tuples of $E \times I \times S$. [26]

We now describe the recursive competitive equilibrium (RCE) for this Stage -1 economy. We

[25] This means that anyone at $s > \widetilde{s}$, will have a higher labour income regardless of his/her education

[26] Define $\mathcal{B}(\mathcal{Z})$ as the set of Borel sets on $\mathcal{Z}$, in particular $\mathcal{E},\mathcal{G},\mathcal{S} \in \mathcal{B}(\mathcal{Z})$ where $\mathcal{E},\mathcal{G},\mathcal{S}$ are projections of $\mathcal{Z}$ over the spaces E, G and S respectively. Let $\mathcal{P}$ be a probability measure on $\mathcal{B}(\mathcal{Z})$, then $\mathcal{P} : \mathcal{B}(\mathcal{Z}) \to [0,1]$.

focus on stationary solutions and its associated equilibrium.

Definition of the Stage -1 (Stationary) Recursive Competitive Equilibrium

A Stage-1 stationary RCE is a value function $V : \mathcal{Z} \to R$, policy functions $c : \mathcal{Z} \to R$, $x : \mathcal{Z} \to R$, and $l : \mathcal{Z} \to R$, price p, and a measure $\Phi \in \mathcal{M}$ such that:

1. Given p the policy functions $c(e,i,s)$, $x(e,i,s)$ and $l(e,i,s)$ solve the sex-consumer household problem (7)-(8) and sex-producer households problem (10)-(11).

2. All markets clear.

$$\sum_{e,i,s} x(e,i,s) = \sum_{e,-i,s} x(e,-i,s),$$

   The sex markets clear and the consumption market clears by Walras law.

3. The stationary probability distribution,

$$\Phi = H(\Phi)$$

   is induced by the equilibrium policy functions.

Notice that value function, policy functions, and price are not any longer indexed by measures Φ because all conditions must be satisfied for the equilibrium stationary measure Φ. The last requirement states that the measure Φ reproduces itself: starting with a measure of education, sex type, and income shocks today generates the same measure tomorrow.

[Stage 0] The Myopic Onset of the HIV Epidemic

The HIV epidemics starts in this stage, but agents are neither aware of its presence nor its workings. Agents can be healthy or HIV infected $h \in \mathcal{H} = \{-,+\}$. Specifically, agents live with HIV myopia in two dimensions. First, agents are unaware of the fact that HIV infection

Then the evolution of the population distribution is,

$$\Phi'(\mathcal{E},\mathcal{I},\mathcal{S}) = F(\Phi)(\mathcal{E},\mathcal{I},\mathcal{S}) = \sum_{e,i,s} Q((e,i,s)(\mathcal{E},\mathcal{I},\mathcal{S})) + f\Phi((e,i,s')(\mathcal{E},\mathcal{I},\mathcal{S})), \tag{13}$$

which is the fraction of people with education $\mathcal{E}$, type $\mathcal{G}$ and states in $\mathcal{S}$ as measured by Φ, that transit to $(\mathcal{E},\mathcal{G},\mathcal{S})$ as measured by $\mathcal{Q}$. The last term accounts for the new born. Population of each group increases according to respective fertility rate f.

is occurring and that its spread depends on the amount of risky sex transactioned x, (either consumed or produced) and the current prevalence rate, $\phi^+ \in [0,1]$. Specifically, HIV infection occurs at an endogenous rate:

$$\phi^+ \lambda(x;\rho) = \phi^+ \frac{e^x}{e^x + \rho e^{-x}}, \tag{14}$$

where $\rho \in [0,\infty)$ is a parameter that governs the mapping from the amount of sex transactioned to the probability of HIV infection. The lower is ρ the higher is the probability of HIV infection per amount of sex transactioned.[27][28] Second, although agents are unaware of the nature of HIV infection (14), at every period, agents observe higher average mortality rates (γ), lower average labor market and sex production productivity (z) and a lower overall satisfaction (χ). Since agents do no know that these changes in γ, z and χ are due to HIV, at every period, the economy makes the mistake of taking these observations as unexpected one-time aggregate permanent shocks in mortality rates, productivity and preferences . We model this as a sequence of permanent unexpected aggregate shocks in γ, z and χ. That is, at every period, agents notice a change between γ_t and γ_{t-1}, between z_t and z_{t-1}, between χ_t and χ_{t-1} and agents assume that $\widetilde{\gamma}_\tau = \gamma_t$, $\widetilde{z}_\tau = z_t$, and $\widetilde{\chi}_\tau = \chi_t$ for all $\tau \geq t$.

In reality, however, this is not the case because the average survival rates, labor productivity and felicity depend on the distribution of HIV status across the population which is endogenous to risky sex. In particular, the true HIV distribution of the population evolves according to

$$\begin{bmatrix} \phi^-_{t+1} \\ \phi^+_{t+1} \end{bmatrix} = \begin{bmatrix} \gamma_- & 0 \\ 0 & \gamma_+ \end{bmatrix} \begin{bmatrix} 1-\phi^+_t \lambda_\rho(x_t) & 0 \\ \phi^+_t \lambda_\rho(x_t) & 1 \end{bmatrix} \begin{bmatrix} 1+f & f \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \phi^-_t \\ \phi^+_t \end{bmatrix} \tag{15}$$

where ϕ^+_t and ϕ^-_t are the measures of HIV infected and HIV non-infected populations, γ_+ and γ_- are survival rates for HIV infected and HIV non-infected populations, respectively, the odds of HIV infection $\phi^+_t \lambda_\rho(x)$ are defined by (14), and f is the fertility rate for the aggregate economy,

[27] Notice that in order for the virus to start spreading we need to provide an initial HIV prevalence rate $\phi^+_{t=0} > 0$. The value for $\phi^+_{t=0}$ is exogenous, meaning that it is not directly linked to the risky sex practices between humans, which is the main mechanism of HIV spread in this model. The calibration of the initial prevalence is described in Section 4. We conduct some robustness on this assumption by setting the probability of infection equal to $\lambda_\rho(x)$, that is, the aggregate rate of HIV infection in the economy does not explicitly affect the infection probability. Our main results remained unchanged.

[28] Medical studies show that the conditional probability of infection depends on the stage of infection (primary, asymptomatic, late) Hollingsworth et al. [2008], Leynaert et al. [1998]. Our model abstracts from these differences.

which is independent of HIV status.[29] Then, the evolution of the aggregate population is:

$$\phi_{t+1} = \phi^-_{t+1} + \phi^+_{t+1} \tag{16}$$

with:

$$\phi^-_{t+1} = \gamma_-(1 - \phi^+_t \lambda_\rho(x_t))[(1+f)\phi^-_t + f\phi^+_t]$$
$$\phi^+_{t+1} = \gamma_+ \left(\phi^+_t \lambda_\rho(x_t)[(1+f)\phi^-_t + f\phi^+_t] + \phi^+_t\right)$$

Since our agents are myopic in HIV they only see the current population ϕ_{t+1}, the fertility rate, and the previous population ϕ_t so as to infer an aggregate survival rate, $\widetilde{\gamma}_t$,

$$\phi_{t+1} = \widetilde{\gamma}_t(1+f)\phi_t. \tag{17}$$

Notice that we can equate (16) and (17) to find the survival rate $\widetilde{\gamma}$ observed by myopic agents. We can proceed analogously to find the average labor productivity observed by agents,

$$\widetilde{z}_t = \frac{z_-\phi^-_t + z_+\phi^+_t}{\phi^-_t + \phi^+_t}. \tag{18}$$

And analogously for the preference parameter $\widetilde{\chi}_t$. We calibrate the preference parameter χ_+ such that, conditional on infection and the state, the following inequality holds :

$$u(\cdot)\chi_+ < u(\cdot)\chi_- \tag{19}$$

Given the myopic updating formulas for γ and z in (17), (18) and the preference parameter calibration (19), we are now ready to formulate the risky-sex consumer and producer problems. Importantly, notice that $\widetilde{\gamma}_t$ and $\widetilde{z}_t$ get updated every period and, hence, the formulation of the households problem and the definition of equilibrium for the Stage 0 of the HIV epidemic needs to reflect this phenomenon.

[29] Note that we can develop (15) as:

$$\begin{bmatrix}\phi^-_t \\ \phi^+_t\end{bmatrix} = \begin{bmatrix}\gamma_- & 0 \\ 0 & \gamma_+\end{bmatrix}\begin{bmatrix}1-\phi^+_t\lambda_\rho(x_t) & 0 \\ \phi^+_t\lambda_\rho(x_t) & 1\end{bmatrix}\begin{bmatrix}(1+f)\phi^-_t + f\phi^+_t \\ \phi^+_t\end{bmatrix} = \begin{bmatrix}\gamma_- & 0 \\ 0 & \gamma_+\end{bmatrix}\begin{bmatrix}(1-\phi^+_t\lambda_\rho(x_t))[(1+f)\phi^-_t + f\phi^+_t] \\ \phi^+_t\lambda_\rho(x_t)[(1+f)\phi^-_t + f\phi^+_t] + \phi^+_t\end{bmatrix}$$

and hence,

$$\phi^-_{t+1} = \gamma_-(1-\phi^+_t\lambda_\rho(x_t))[(1+f)\phi^-_t + f\phi^+_t] \text{ and } \phi^+_{t+1} = \gamma_+\left(\phi^+_t\lambda_\rho(x_t)[(1+f)\phi^-_t + f\phi^+_t] + \phi^+_t\right).$$

Risky-sex consumer household problem. Risky-sex consumers choose c, and x to solve:

$$V_t(e,i,s,\Phi) = \max_{c_t \geq 0, x_t \geq 0} \widetilde{\chi} u(c_t, x_t) + \beta \widetilde{\gamma} \sum_{s'|s} \pi(s'|s) V_{t+1}(e,i,s',\Phi') \tag{20}$$

s.t

$$c_t + p_t(\Phi) x_t = \widetilde{z} y(e) s \tag{21}$$

$$\widetilde{\gamma} = \widetilde{\gamma}_{t-1} = \frac{\phi_t}{(1+f)\phi_{t-1}} \tag{22}$$

an income shock process that follows (9)

Risky sex-producer household problem. Sex producers choose c and l, to solve:

$$V_t(e,-i,s,\Phi) = \max_{c_t \geq 0, 1 \geq l_t \geq 0} \widetilde{\chi} u(c_t) + \beta \widetilde{\gamma} \sum_{s'|s} \pi(s'|s) V_{t+1}(e,-i,s',\Phi') \tag{23}$$

s.t

$$c_t = \widetilde{z} \left[(p_t(\Phi) l_t^{\alpha} + y(e) s (1 - l_t)) \right] \tag{24}$$

$$\tag{25}$$

with income shock process (9), survival rates $\tilde{\gamma}$, productivity $\tilde{z}$, and preference parameter $\tilde{\chi}$.

At any point in time in the HIV Stage 0, the economy is summarized by the joint distribution Φ of individual states (e,i,s). Importantly, notice that HIV status is not part of the individual states, as agents in our economy are unaware of HIV. The aggregate state variable of the economy evolves following:

$$\Phi_{t+1} = H_t(\Phi_t) \tag{26}$$

Notice that our objective functions and prices are indexed by time which captures the non-stationarity nature of the Stage 0 problem. This sequential formulation of the recursive problem is required to address the unexpected changes in γ, z and χ. On the top of that, notice that the myopia makes agents assume that last period's mortality, productivity and preferences will be permanent, that is, $\widetilde{\gamma}_{\tau} = \widetilde{\gamma}\ \forall \tau \geq t$, $\widetilde{z}_{\tau} = \widetilde{z}\ \forall \tau \geq t$ and $\widetilde{\chi}_{\tau} = \widetilde{\chi}\ \forall \tau \geq t$. Because of this myopia, the changes in average mortality, productivity and felicity are not only unexpected, but also occur at every period.

Definition of the Stage 0 (Nonstationary) RCE

Given a Stage-1 stationary joint distribution of (e, i, s), $\Phi_{t=0}(g = -1)$, and a sequence of myopically unexpected and permanent changes in mortality rates $\{\widetilde{\gamma}_t\}_{t=0}^{\infty}$, labor productivity $\{\widetilde{z}_t\}_{t=0}^{\infty}$ and preference shocks $\{\widetilde{\chi}_t\}_{t=0}^{\infty}$, constructed from (17), (18) and (19), respectively, a competitive equilibrium is a sequence of individual household functions $\{V_t, c_t, x_t, l_t : Z \times M \to M\}_{t=0}^{\infty}$, sequence of prices $\{p_t\}_{t=0}^{\infty}$, and a sequence of measures $\{\Phi_t\}_{t=0}^{\infty}$ such that, $\forall t$:

1. The policy functions $c_t(e, i, s)$, $x_t(e, i, s)$ and $l_t(e, i, s)$ solve the sex-consumer household problem (20) and sex-producer households problem (23).

2. All markets clear.

$$\sum_{e,i,s} x_t(e, i, s) = \sum_{e,-i,s} x_t(e, -i, s),$$

   The sex markets clear and the consumption market clears by Walras law.

3. The aggregate law of motion is,

$$\Phi_{t+1} = H_t(\Phi_t)$$

   where Φ is the joint distribution of (e, i, s) is induced by the equilibrium policy functions.

4. The true distribution of the HIV population, which is used to construct the sequences $\{\widetilde{\gamma}_t\}_{t=0}^{\infty}$, $\{\widetilde{z}_t\}_{t=0}^{\infty}$ and $\{\widetilde{\chi}_t\}_{t=0}^{\infty}$, endogenously evolves according to (15).

The main characteristic of this Stage 0 is that agents do not know that the HIV epidemic is unravelling. That is, agents in this Stage 0 do not internalize the evolution of the HIV (15) because they are unaware that their sexual behavior affects their chances of survival, labor productivity and felicity. We model this through a myopic updating of γ, z and χ. Where agents perceive the updates in γ, z and χ as permanent changes. This implies that after one of these permanent changes, agents rationalize their current behavior by looking forward and solving the entire transition from today to a new steady state associated with the new triple $(\widetilde{\gamma}, \widetilde{z}, \widetilde{\chi})$. That is, we need to solve for a transition every time there is a perceived permanent change. Because this permanent changes occur every period, then at every period we need to compute the entire transition. The equilibrium value functions and policy functions are the sequence of first-period

solutions to the sequence of transitional problems. The sequence of myopic permanent changes do not go *ad infinitum* because Stage 1-2 (Maturity) arrives after the economy has been a finite amount of periods T_0in Stage 0.

[Stage 1-2] Learning the HIV Epidemic

At this stage, agents are aware of the HIV epidemic and its consequences: higher mortality rates, lower productivity and lower felicity for HIV infected individuals. Agents are aware of their own HIV status and that of the rest of the population.[30] However, at the beginning of Stage 1-2, agents are not fully aware of the sexual nature of HIV infection in so far they do not accurately know ρ in (14). Although agents do not know the odds of infection as function of risky sexual activity, agents learn about it through Bayesian updates on ρ with some noise.

The speed in which agents learn about the actual odds of infection differs across education groups, this being faster for the more-educated agents. The learning speed is the only source of heterogeneity across education groups introduced in this stage. The initial degree of accuracy in which individuals know the odds of infection is the same across education groups, this follows from the fact that both groups were completely unaware of ρ in the previous Stage.
More precisely, each educational group $e \in \mathcal{E}$ has a prior belief about the distribution of $\lambda(x;\rho)$, denoted by the p.d.f $\mathcal{P}_e(\lambda(x;\rho))$. Furthermore, at the beginning $\mathcal{P}_e(\lambda(x;\tilde{\rho}_o)) \sim N(\lambda(x;\tilde{\rho}_o),\sigma^2_\varepsilon)$ and $\mathcal{P}_{e=1}(\lambda(x;\tilde{\rho}_o)) = \mathcal{P}_{e=0}(\lambda(x;\tilde{\rho}_o))$. Afterwards, agents receive a signal $\widetilde{\lambda}_\rho(x)$ per period. This signal contains information about the actual probability of infection plus some noise ε_t that is normally distributed with zero mean and variance $\sigma^2_\varepsilon(e)$. Explicitly:

$$\widetilde{\lambda}_\rho(x) = \lambda_\rho(x) + \varepsilon_t, \tag{27}$$

where the signal follows the following covariance stationary process:

$$\varepsilon_t = v_t + \mathbf{1}_{e=0} u_t \tag{28}$$

with $v \sim N(0,\sigma^2_v)$ and $u \sim N(0,\sigma^2_u)$. The dummy $\mathbf{1}_{e=0}$ equals one if an agent belongs to the less educated group, and zero otherwise. That is, the signal is noisier for the less educated individuals than for the more educated individuals. In particular, every period t agents update their beliefs $\mathcal{P}_e(\lambda_\rho(x))$ given the information up to $t-1$ according to Bayes rule:

$$\mathcal{P}_e(\lambda_\rho(x)) = \mathcal{P}_e(\lambda_\rho(x)|\widetilde{\lambda}_\rho(x)) = \frac{\mathcal{P}_e(\widetilde{\lambda}_\rho(x)|\lambda_\rho(x))\mathcal{P}_e(\lambda_\rho(x))}{\mathcal{P}_e(\widetilde{\lambda}_\rho(x))} \tag{29}$$

[30] See Carli and Santaeulalia-Llopis (2019)**?** for an economy in which individuals can hide their HIV status.

where the Bayesian updates will transit faster to the actual odds of HIV infection for the more educated individuals due to (28).[31]

Let us now write the nonstationary recursive problem and then explain it

Risky-sex consumer household problem. Risky-sex consumers choose c, and x to solve:

$$V_t(e,i,s,h,\Phi) = \max_{c_t\geq 0, x_t\geq 0} \chi(h)u(c_t,x_t) + \beta \sum_{h'|h,s'|s} [\gamma(h')\widetilde{\lambda}_{\rho(e)}(h'|x_t,h)\pi(s'|s)V_{t+1}(e,i,s',h',\Phi')\ ] \tag{30}$$

subject to,

$$c_t + p_t(\Phi)x_t = z(h)y(e)s \tag{31}$$

an income shock process that follows (9).

Risky sex-producer household problem. Sex producers choose c and l, to solve:

$$V_t(e,-i,s,h,\Phi) = \max_{c_t\geq 0, 1\geq l_t\geq 0} \chi(h)u(c_t) + \beta \sum_{h'|h,s'|s} [\gamma(h')\widetilde{\lambda}_{\rho(e)}(h'|x_t,h)\pi(s'|s)V_{t+1}(e,-i,s',h',\Phi')\ ] \tag{32}$$

subject to,

$$c_t = z(h)[p_t(\Phi)l_t^{\alpha} + y(e)s(1-l_t)], \tag{33}$$

with income shock process (9).

At any point in time in the HIV Stage 1-2, the economy is summarized by the joint distribution Φ of individual states (e,i,s,h), which incorporates individual HIV status. In this Stage 1-2 of the epidemic, agents are aware of their HIV status, and that of the rest of the economy. The aggregate state variable of the economy evolves following $\Phi_{t+1} = H_t(\Phi_t)$. Notice that even though agents are aware of the average probability of infection of the rest of the economy they cannot infer the true value of ρ, moreover, even if agents are aware of the exact infection process (namely 14), having knowledge of the average probability of infection is not enough to retrieve the value of ρ.

[31] We assume normality of the prior belief to simplify the calculations, however this can be adapted to mimic more complex formulations.

An additional implication of the learning mechanism is that agents will also update their own mortality expectations and use them to make allocation decisions. This means agents keep subjective expectations in mortality since these expectations differ from their true individual probability of survival[32].

As it was the case of Stage 0, in Stage 1-2, the objective functions and prices are indexed by time which captures the nonstationarity nature the Stage. In the long-run, when both education groups have finalized their learning process of the odds of HIV infection a stationary RCE can be defined for Stage 1-2.

In this stage all agents in the economy learn about the risk of infection, a natural extension to this set up would be to make the proportion of people that learn increase gradually. This extension represents the introduction a new state variable since it will be necessary to keep track of the proportion of the population that is still miopic and the ones who are learning. This can be done in order to smooth out the transition from HIV Stage 0 to HIV Stage 1-2 and see a hump at the peak, as in the data.

[32]This is a consequence of the fact that agents learn about their risk of infection which is directly linked with their probability of survival. The lower people think their odds of infection are, (this happens at the beginning of the learning stage) the higher are their believed chances of survival.

Definition of the Stage 1-2 (Nonstationary) RCE

Given a Stage 0 joint distribution $\Phi_{t=0}(g = 0)$, and a simulated sequence of infection probabilities $\{\widetilde{\lambda}_\rho(e)\}_{t=0}^{\infty}$ by education group, a competitive equilibrium is a sequence of individual household functions $\{V_t, c_t, x_t, l_t : Z \times M \to M\}_{t=0}^{\infty}$, sequence of factor prices $\{p_t\}_{t=0}^{\infty}$, and a sequence of measures $\{\Phi_t\}_{t=0}^{\infty}$ such that, $\forall t$:

1. Given $\{p_t\}_{t=0}^{\infty}$ the policy functions $c_t(e,i,s,h)$, $x_t(e,i,s,h)$, and $l_t(e,i,s,h)$ solve the sex-consumer household problem (30) and sex-producer households problem (32).

2. All markets clear.

$$\sum_{e,i,s,h} x_t(e,i,s,h) = \sum_{e,-i,s,h} x_t(e,-i,s,h),$$

   The sex markets clear and the consumption market clears by Walras law.

3. The aggregate law of motion is,

$$\Phi_{t+1} = H_t(\Phi_t)$$

   where Φ is the joint distribution of (e,i,s,h) is induced by the equilibrium policy functions.

4. The true distribution of the HIV population endogenously evolves according to (15).

5. The beliefs of on the odds of infection by education group evolve according to Bayes rule (29) and signal (28).

Remark. The Stage 1-2 stationary RCE is the limiting case of the nonstationary RCE in which beliefs of both education groups have converged to the actual odds of infection and the cross-sectional distribution Φ does not change over time. In that case, we can drop all time subscripts.

[Stage 3-4] The Era of ARVs

We assume that ARV drugs entirely revert the negative effects of HIV on mortality and productivity, but not on felicity. This pursuits the argument that receiving treatment is not necessarily the same as being completely healthy, hence this is an implicit cost of ARV treatment that will have a positive effect on the equilibrium price of risky sex. For simplicity only those who are infected can receive treatment[33]. Let us now write and explain the nonstationary recursive problem.
Treatment is provided stochastically to the infected population, with the educated individuals having a higher probability to receive treatment $\eta_{t,e=1} > \eta_{t,e=0}$ at all t[34] . This means that ARV drugs now represent a new state variable $d \in \mathcal{D} = \{d^+, d^-\}$ as we now need to keep track of the portion of the infected population receiving treatment (d^+) and that which doesn't (d^-).
The aggregate proportion of the population that receives drugs at each t is deterministic and represented by the monotonically increasing sequence $\{\eta\}_{t=0}^{\infty}$, where $0 \leq \eta_t \leq \tilde{\eta}\ \forall t$ with $\lim_{t\to\infty} \eta_t = \tilde{\eta}$ and $\tilde{\eta} \in (0, 1]$. In addition, ARV's affect the average probability of infection by decreasing the viral load and, hence, the infectiousness of those infected. In our formulation this would be translated into a proportional increase of ρ with respect to the coverage rate.
In this stage agents keep learning about the value of $\lambda_\rho(x)$ according to (29).
Then the distribution of HIV positive population evolves according to:

$$\begin{bmatrix} \phi^-_{d^-t+1} \\ \phi^+_{d^+t+1} \\ \phi^+_{d^-t+1} \end{bmatrix} = \begin{bmatrix} \gamma_- & 0 & 0 \\ 0 & \gamma_- & 0 \\ 0 & 0 & \gamma_+ \end{bmatrix} \begin{bmatrix} 1-\lambda_{t,\rho} & 0 & 0 \\ \eta_t\lambda_{t,\rho} & \eta_t & \eta_t \\ (1-\eta_t)\lambda_{t,\rho} & 1-\eta_t & 1-\eta_t \end{bmatrix} \begin{bmatrix} 1+f & f & f \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \phi^-_{d^-t} \\ \phi^+_{d^+t} \\ \phi^+_{d^-t} \end{bmatrix} \tag{34}$$

Let us now write the nonstationary recursive problem:

Risky-sex consumer household problem. Risky-sex consumers choose c, and x to solve:

$$\begin{aligned} V_t(e,i,s,h,d,\Phi) = &\max_{c_t\geq 0, x_t \geq 0} \chi_d(h)u(c_t,x_t) \\ &+ \beta \sum_{\substack{d'|d,h'|h \\ s'|s}} [\eta_t(d'|e,d)\gamma_d(h')\lambda_{\rho(e)}(h'|x_t,h)\pi(s'|s)V_{t+1}(e,i,s',h',d,\Phi')\ ] \end{aligned} \tag{35}$$

[33]The Joint United Nations Programme on HIV/AIDS (UNAIDS) supplies ARVs coverage data only for that portion of the population that was infected and received treatment.

[34]The current data does not provide treatment composition by educational groups therefore we approximate these probabilities by $\eta_{t,e=1} = (1+\iota)\eta_t$ and $\eta_{t,e=0} = (1-\iota)\eta_t$, where η_t is the aggregate coverage rate observed in the data at t, with ι controlling the odds of treatment with respect to the aggregate by education group.

subject to,

$$c_t + p_t(\Phi)x_t = z_d(h)y(e)s \tag{36}$$

an income shock process that follows (9).

Risky sex-producer household problem. Sex producers choose c and l to solve:

$$\begin{aligned} V_t(e,-i,s,h,d,\Phi) = &\max_{c_t\geq 0, 1\geq l_t\geq 0} \chi_d(h)u(c_t) \\ &+ \beta \sum_{\substack{d'|d,h'|h \\ s'|s}} [\eta_t(d'|e,d)\gamma_d(h')\lambda_{\rho(e)}(h'|x_t,h)\pi(s'|s)V_{t+1}(e,-i,s',h',d,\Phi') \,] \end{aligned} \tag{37}$$

subject to,

$$c_t = z_d(h)[p_t(\Phi)l_t^{\alpha} + y(e)s(1-l_t)], \tag{38}$$

with income shock process (9).

Notice that if ARVs fully revert the effects of HIV on mortality rates, productivity but not preferences then:

$$\begin{aligned} \gamma_{d^+}(+) &= \gamma_- = \gamma \\ z_{d^+}(+) &= z_- = z \\ \chi_{d^+}(+) &= \chi_- = \chi_+ \end{aligned}$$

However for those who are infected but not treated, survival rates, productivity and preferences go back to that of the maturity stage:

$$\begin{aligned} \gamma_{d^-}(+) &= \gamma_+ \\ z_{d^-}(+) &= z_+ \\ \chi_{d^-}(+) &= \chi_+ \end{aligned}$$

At any point in time in the HIV Stage 3-4, the economy is summarized by the joint distribution Φ of individual states (e,i,s,h,d), which incorporates individual HIV status and treatment. In this Stage 3-4 of the epidemic, agents are know if they received drugs as well as their HIV status, and that of the rest of the economy. The aggregate state variable of the economy evolves following $\Phi_{t+1} = H_t(\Phi_t)$. Notice that, as it was the case of Stage 1-2, in Stage 3-4, the objective functions and prices are indexed by time which captures the nonstationarity nature of this stage.

In the long-run, we find a stationary RCE for Stage 3-4 that differs from Stage -1 in that some proportion of the population will be infected with HIV and a deterministic proportion of the population will receive treatment.

Definition of the Stage 3-4 (Nonstationary) RCE

Given a Stage 1-2 distribution $\Phi_0(g = 1 - 2)$ and a deterministic sequence for the treated proportion of the population $\{\eta_t\}_{t=0}^{\infty}$, a competitive equilibrium is a sequence of individual household functions $\{V_t, c_t, x_t, l_t : Z \times M \to M\}_{t=0}^{\infty}$, sequence of factor prices $\{p_t\}_{t=0}^{\infty}$, and a sequence of measures $\{\Phi_t\}_{t=0}^{\infty}$ such that, $\forall t$:

1. Given $\{p_t\}_{t=0}^{\infty}$ the policy functions $c_t(e, i, s, h, d)$, $x_t(e, i, s, h, d)$ and $l_t(e, i, s, h, d)$ solve the sex-consumer household problem (35) and sex-producer households problem (37).

2. All markets clear.

$$\sum_{e,i,s,h,d} x_t(e, i, s, h, d) = \sum_{e,-i,s,h,d} x_t(e, -i, s, h, d),$$

   The sex markets clear and the consumption market clears by Walras law.

3. The aggregate law of motion is,

$$\Phi_{t+1} = H_t(\Phi_t)$$

   where Φ is the joint distribution of (e, i, s, h, d) is induced by the equilibrium policy functions.

4. The true distribution of the HIV population endogenously evolves according to (34).

Remark. The Stage 3-4 stationary RCE is the limiting case of the nonstationary RCE in which the cross-sectional distribution Φ does not change over time. In that case, we can drop all time subscripts.

3.1 Pseudo-Algorithm Solution

We take the following steps to find a solution to our quantitative model for a given parametrization θ:

1. Find the stationary RCE for Stage -1 (Pre- HIV Era).

2. The HIV epidemic starts in Stage 0 (HIV Myopia); remember that at every period of this stage arrives an unexpected permanent shock to mortality $\widetilde{\gamma}$, productivity $\widetilde{z}$ and felicity $\widetilde{\chi}$. Then, for every period, we need to compute both the new stationary RCE associated with the new $\widetilde{\gamma}$, $\widetilde{z}$, $\widetilde{\chi}$ and the corresponding nonstationary RCE that captures the equilibrium transition from the current period (when the permanent shock occurs) to the period when the economy reaches the stationary RCE (we compute this transition backwards). Note that we are only interested in the first value function of each of those transitions, because every next period a new set unexpected permanent shocks ($\widetilde{\gamma}$, $\widetilde{z}$, $\widetilde{\chi}$) occur, which requires us to recompute again the associated stationary and nonstationary RCE.

3. Stage 1-2 (HIV Maturity) arrives as an unexpected shock that hits Stage 0 at T_0. Starting with a set of common prior beliefs $\widetilde{\rho_0}$, simulate a series of sequentially updated beliefs by education group until both convergence to the actual ρ. Solve the stationary RCE using ρ, and then solve the transition backwards using the simulated series for beliefs $\{\rho_t(e)\}_{t=T_0}^{T}$. Where T is an arbitrary large number representing the time needed to converge to the stationary RCE.

4. Stage 3-4 (ARV Era) arrives as an unexpected shock that hits the transition in Stage 1-2 after T_1 periods. We compute the stationary RCE with full ARV coverage. Then solve the transition backwards for a given monotonically increasing sequence of coverage levels $\{\eta_t(e)\}_{t=T_1}^{T}$ by education.

4 Calibration

In this Section, we discuss our calibration strategy. Some parameters which have direct observable data analogs were assigned its respective values, or values that are commonly used in the literature ($\beta, \alpha, \xi, \gamma_-, \gamma_+, f, y_{e=0}/y_{e=1}, s, z_+, \pi, \vartheta_{e=1,i}, \vartheta_{e=1,-i}, \vartheta_{e=0,i}, \tilde{\rho_o}, \phi^+_{t=0}, T_1$). Most of these parameters are calibrated to match the pre-HIV Era. The rest of the parameters ($T_0, \rho, \vartheta_{e=0,-i}, \sigma^2_\epsilon(e), \chi_+, \iota$) were picked to match several targeted moments in the data. We choose the same number of moments as of free parameters (i.e. exactly identified moment estimation). This procedure involves solving the model many times as to minimize the distance between the model generated moments and the moments observed in the data [35].
We explicitly target the cross country educational gradient coefficients documented in Santaeulalia-Llopis and Iorio [2016], and the evolution of the Malawi HIV prevalence from year 1990 to 2010. [36] The parameter governing the risk infection probability ρ, the time to reach maturity T_0 and the proportion of sex producers in the economy $\vartheta_{e=0,-i}$ are calibrated to match features of the HIV prevalence across education groups at the peak of the epidemic. The speed of learning across education groups $\sigma^2_{\epsilon,e=1}, \sigma^2_{\epsilon,e=0}$, as well as the preference shock χ_+, are calibrated to match the behavior of a mature HIV epidemic. Finally, treatment coverage by education ι and a new value of ρ, are calibrated to reflect the effects of the introduction of ARVs at the last stage of the epidemic. We now discuss our calibration strategy in detail by stage of the epidemic.

Stage -1: Pre-HIV Era

We calibrate β to 0.98 for all stages, this reflects a subjective discount rate of 2% of the economy. The relative coefficient of risk aversion ξ is set to 3 as to reflect a high risk averse country like Malawi. Survival rates $\gamma = 97.7\%$ for the pre-epidemic stage are calibrated in such a a way that the individuals have an average life expectancy of 64 years. Remember that agents enter the model when they are 18 years old. The fertility rates $f = 4\%$ reflect the average fertility rate in Malawi in 2016 according to the World Bank Sustainable Development Indicators.

Labor income is normalized to one for educated individuals, i.e., $y_{e=1} = 1$. The National Statistical Office of Malawi (2013) shows that the educational premium for someone who has completed secondary education is at least of 55% compared to someone who has not completed

[35] Our objective function is based on the SMM method (Simulated Method of Moments). Specifically, let the targeted moments be $\mathcal{M}(\theta) = [\mathbf{m} - \widehat{\mathbf{m}}(\theta)]$ where $\mathbf{m}$ is a vector of observed moments and $\widehat{\mathbf{m}}(\theta)$ is the vector of model generated moments given parametrization θ. Then, we construct the objective function $\min_\theta \ \mathcal{M}(\theta)^T W \mathcal{M}(\theta)$, where the weighting matrix W is the diagonal matrix.

[36] Although the model is not explicitly calibrated to a particular country, we use data from Malawi to discipline most of the parameters of the model.

primary school, so we set $y_{e=0} = .6452$. Households are subject to income shocks, s, that take two possible values: s equals to one in good times and s equals to $.4$ in bad times which mimics a 60% loss of household income during a period of unemployment Magalhaes and Santaeulalia-Llopis [2018]. The income shock follows a Markov process with a transition matrix π from high (i.e employment) to low (i.e unemployment) that is calibrated so that at all times 5.4% of the population is under low income[37]. Since there is no HIV in at this point, labor and sex productivity z and preference parameter χ are equal to one. We choose α such that the proportion of sex income in the aggregate economy is 7% Sulaimon et al. [2018][38]; this in turn, implies a proportion of sex income in the total income of sex producers, ω of 12%.

We also need to choose the proportion of educated and non educated individuals. We set the proportion of educated to be 13%[39] as observed in the DHS surveys for Malawi; we classify as educated those with completed primary education, this anchors the estimation of the model-generated HIV-education gradient. To see this, notice that the individuals who didn't complete primary education average 3.35 schooling years, and those with at least primary education average 8.32 schooling years. We use use this difference in schooling years across education for the estimation of the HIV-education gradient with the model-generated data.

We assume that all sex producers are non educated ($\vartheta_{e=1,-i} = 0$ i.e $\vartheta_{e=1,i} = 0.13$ from (39)). This is reasonable either because prostitutes are less educated or because teenagers that engage in sexual activity (STDs, pregnancy, etc.) are less likely to finish school Duflo et al. [2015b], Dupas [2011b]. This implies that all educated individuals are sex consumers. Consequently, we need to choose the proportion of sex producers in the economy ($\vartheta_{e=0,-i}$), which in turn delivers the proportion of sex consumers among the uneducated people ($\vartheta_{e=0,i}$). We choose this proportion to match the HIV prevalence rate at the peak of the epidemic in the next stage as we discuss next. Finally, we set the initial prevalence $\phi^{+}_{t=0}$ to 0.5%, this is number is directly linked with the calibration of the duration of the next stage (Stage 0 HIV Myopia) T_0 because the larger is $\phi^{+}_{t=0}$, the less periods necessary to reach the peak.

Stage 0: Myopic Onset of the HIV Epidemic

At this stage the ability of infected individuals to produce both sex and the consumption good at a given scale z is reduced by 65%. We choose the value of z_+ such that the proportion of

[37] This is calibrated to the average unemployment rate in Malawi in 2019, Source: ILOSTAT database

[38] Sulaimon et al. [2018] mention that in Indonesia, Malaysia, the Philippines and Thailand, the sex sector (prostitution) accounts for between 2% and 14% of the Gross Domestic Product.

[39] In the model this is the sum of the proportion of educated consumers plus educated producers $\vartheta_{e=1,i} + \vartheta_{e=1,-i} = 0.13$

sex income among sex producers at the peak is the same ω as in the pre-HIV stage. Additionally the survival probability of someone who is infected with HIV reduces to $\gamma_+ = 90\%$, which is translated to a life expectancy of 11 years after the moment of infection.

There are three additional parameters to calibrate. The first parameter carries from the previous stage, which is the proportion of sex producers in the economy $\vartheta_{e=0,-i}$. The second is a parameter, ρ, that governs the true rate of infection as a function of risky sex, $\lambda(x;\rho)$. The third is the time until the epidemic reaches maturity T_0. We choose these three parameters such that we match the HIV prevalence by education group at the peak of the epidemic as well as the number of years needed to reach the peak[40].

Once we have the proportion of individuals that are producers in the economy $\vartheta_{e=0,-i}$, we use the joint distribution of education groups and sex types (producers vs. consumers) at the pre-HIV stage ($\vartheta_{e=1,i}, \vartheta_{e=1,-i}, \vartheta_{e=0,i}, \vartheta_{e=0,-i}$) to feed the economy at each and every period (and stage) with a fertility that maintains these proportions at birth.

We finally need to choose the magnitude of the preference shock χ_+. This parameter is chosen in the next stage to match the HIV education gradient just before the introduction of ARVs.

Stage 1-2: HIV Maturity, learning the HIV Epidemic

Recall that in this stage agents are aware of the existence of HIV, but their knowledge of the degree of infection risk through sex $\lambda(x;\rho)$ is imperfect. They learn about $\lambda(x;\rho)$ (i.e., ρ) through Bayesian updates. The speed at which agents learn about the true risk of infection depends on two factors. First, the noise of the updating signal $\sigma^2_\varepsilon(e)$ which differs across education groups. Second, how far their initial prior of the infection probability $\tilde{\lambda}_o$ (i.e., $\tilde{\rho}_o$), is from the true value λ (i.e., ρ). This initial prior belief is common across education groups. [41]
The initial common prior belief $\tilde{\rho}_o$ is set to an arbitrary high number following from the fact that agents were myopic in the previous stage and their initial belief of the risk of infection through sex $\widetilde{\lambda_o} = \lambda(x, \tilde{\rho}_o)$ is approximately zero.
We choose the $\sigma^2_\varepsilon(e)$ by education group and the magnitude of the preference shock χ_+ that carries from the previous stage; such that we match the HIV-education gradient (i.e., HIV prevalence by education group) and the average time that it takes to transit from the peak of the HIV

[40]The first HIV positive patient in Malawi was detected in 1985. In 1986 the government of Malawi started implementing preventive measures against the spread of the virus Mwale [2002]. After calibration, the agents in our model become aware of the existence of the virus in year 1986. In section 6 we explore alternative scenarios in which the population starts learning about the virus in earlier stages of the epidemic.

[41]We tried a version having a common noise for the signal updates σ^2_ε and different initial prior beliefs $\tilde{\rho}_o(e)$, however this set up did not guarantee different convergence times across education groups.

epidemic to the end of Stage 1-2 (HIV Maturity). Finally we select the duration of this stage (T_1) such that $year(T_0)+T_1 = 2005$ which is the year in when ARVs were introduced in Malawi.

Stage 3-4: The Era of ARVs

In this stage, the HIV/AIDS effects disappear from the budget constraint of those who are treated, therefore they now have the same survival rate as if healthy. To inform the model about the evolution across time of the proportion of the infected population receiving ARV treatment, we use aggregate treatment data from Malawi starting in 2005 until 2018.[42] We calibrate the parameter governing the share of educated and non educated individuals receiving ARV's ι as to match the HIV gradient of this stage.

In addition, the introduction of ARV treatment reduces the overall degree of infectiousness in the economy, this translates into a reduction of the true infection rate conditional on sex consumption $\lambda(x;\rho)$(i.e an increase in ρ). We calibrate the new value of ρ as to match Malawi's average prevalence rate in 2018. Table 7 summarizes the calibration of the parameters of the model.

Table 7: List of parameters

Description	Symbol	Perceived			True	Stage
		Miopic	Onset	ART	Value	Dependent
Discount factor	β		0.98		0.98	No
Labor share of sex income	α		0.01		0.01	No
Risk aversion	ξ		3.00		3.00	No
Survival rate healthy (%)	$\gamma(-)$	$\widetilde{\gamma}$	97.7	97.7	97.7	Yes
Survival rate infected (%)	$\gamma(+)$	$\widetilde{\gamma}$	90.0	90.0	90.0	Yes
Survival rate infected but treated (%)	$\gamma_{d+}(+)$	-	-	97.0	97.0	Yes
Preference parameter healthy	$\chi(-)$	$\widetilde{\chi}$	1.0	1.0	1.0	Yes
Preference parameter infected	$\chi(+)$	$\widetilde{\chi}$	230	230	230	Yes
Preference parameter infected but treated	$\chi_{d+}(+)$	-	-	230	230	Yes
Fertility rate (%)	f		4.0		4.0	No
Understanding of epidemic $e=1$	$\rho_{e=1}$	-	$\widetilde{\rho}_{e=1}$	$\widetilde{\rho}_{e=1}$	ρ[43]	Yes
Understanding of epidemic $e=0$	$\rho_{e=0}$	-	$\widetilde{\rho}_{e=0}$	$\widetilde{\rho}_{e=0}$	ρ	Yes
Education premium (%)	$y_{e=0}/y_{e=1}$		45.0		45.0	No
Productivity if infected (%)	$z(+)$	$\widetilde{z}$	65.0	65.0	65.0	Yes

[42] Recent DHS data does not provide micro level information to distinguish if ARV treatment is higher among educated individuals, therefore we are unable to compute any gradient related to ARV treatment.

[43] $\rho = 19$ for the Miopic and Onset stages, however once ARVs are introduced the overall infectiousness reduces; this is translated into an increase of to $\rho = 44$ for the ARV stage.

Table 7 – Continued

Description	Symbol	Perceived			True	Stage
		Miopic	Onset	ART	Value	Dependent
Productivity if treated (%)	$z_{d+}(+)$	-	-	100.0	100.0	Yes
Income shock (%)	s		60.0		60	No
Transit probability from s_g to s_g	p_{gg}		0.95		0.95	No
Transit probability from s_g to s_b	p_{gb}		0.05		0.05	No
Transit probability from s_b to s_g	p_{bg}		0.90		0.90	No
Transit probability from s_b to s_b	p_{bb}		0.10		0.10	No
Initial proportion of type $e=1,i$	$\vartheta_{e=1,i}$		13.0		13.0	No
Initial proportion of type $e=1,-i$	$\vartheta_{e=1,-i}$		0.0		0.0	No
Initial proportion of type $e=0,i$	$\vartheta_{e=0,i}$		57.6		57.6	No
Initial proportion of type $e=0,-i$	$\vartheta_{e=0,-i}$		29.4		29.4	No
Variance of the signal's noise ϵ for $e=1$	$\sigma^2_{\epsilon,e=1}$	-	7.9	7.9	-	No
Variance of the signal's noise ϵ for $e=0$	$\sigma^2_{\epsilon,e=0}$	-	84	84	-	No
Initial mean of prior $\mathcal{P}_{e=1}(\lambda(x;\tilde{\rho}_o(e=1)))$	$\tilde{\rho}_o$	-	15700	-	-	No
Initial mean of prior $\mathcal{P}_{e=0}(\lambda(x;\tilde{\rho}_o(e=0)))$	$\tilde{\rho}_o$	-	15700	-	-	No
Odds of treatment parameter	ι	-	-	0.074	-	No
Initial Prevalence (%)	$\phi^+_{t=0}$	0.5	-	-	0.5	No
Duration of HIV Myopia (g=0) (Years)	T_0	∞	-	-	17	No
Duration of HIV Maturity (g=1-2) (Years)	T_1	-	∞	-	16	No

4.1 Model Fit

Table 8, presents the data moments and shows how well the benchmark model matches them[44]. The benchmark calibration matches the HIV-Education Gradient very well[45]. The model also gets pretty close to the actual values of the prevalence all along the evolution of HIV epidemic; Figure 7 compares the prevalence generated by the model with data for Malawi. Note that the Model's prevalence's is always within the 95% confidence interval. Keep in mind however, that the model was calibrated using cross country data for the gradient, and not exclusively prevalence data for

[44] In Appendix 4.2 we conduct sensitivity analysis of the estimated parameters.

[45] To illustrate, Figure 6 in the Appendix **??** shows the isomorphic representation of the estimated HIV-Educ gradient.

Malawi, therefore we consider this a very good approximation for the prevalence.

We have explicitly modelled the direct link between risky sexual behaviour and the probability of infection across education groups through $\lambda_\rho(x)$, as to capture the parallel evolution between the HIV-Education Gradient and disparities in risky sexual behaviour by education that is found the data, therefore the model is able to generate a Risky sex Education gradient that is consistent with the U-shaped pattern of the HIV education gradient and the U-shape of Risky sex Education gradient in the data. Figure 14 Panel (c) illustrates this pattern.

Additionally, the model performs remarkably well mimicking other moments relevant for the HIV epidemic. In the model the average HIV incidence at the peak is 1.66%, close to 1.6% that is reported in the data[46] for the population between 14-49 years in Malawi. The model also captures the slow down of population growth before the peak of the HIV epidemic and the increased population growth afterwards.

Table 8: Targeted Moments

Observation	**Data**	**Model**
HIV Education gradient at the peak (1999)	0.0099	0.0112
HIV Education gradient at the end of Maturity (2005)	-0.0005	-0.0015
HIV Education gradient ARV Stage (2018)	0.0046	0.0047
Prevalence at the peak (1999)	14.6%	14.1%
Prevalence at the end of Maturity (2005)	12.2%	13.1%
Prevalence ARV Stage (2018) (Stage 3-4 ARV Era)	9.2%	9.4%
Time to reach from bottom to peak (in years)	29	28
Time to reach from peak to end of Maturity (in years)	6	6

[46] Source: UNAIDS Estimates 2019.

Figure 6: HIV Education Gradient, comparison between model and data

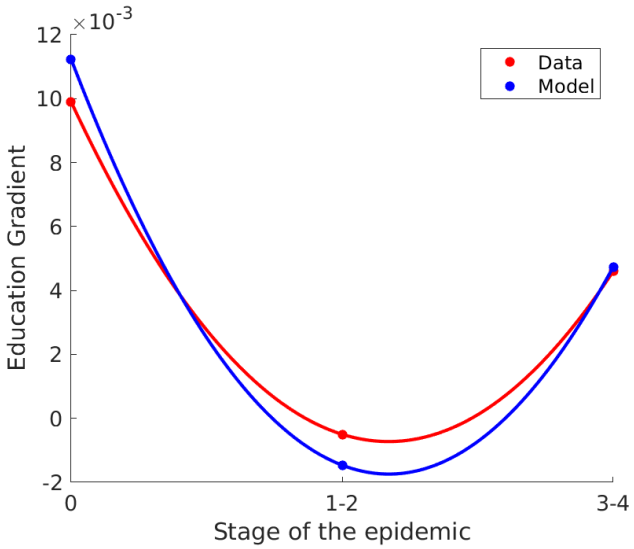
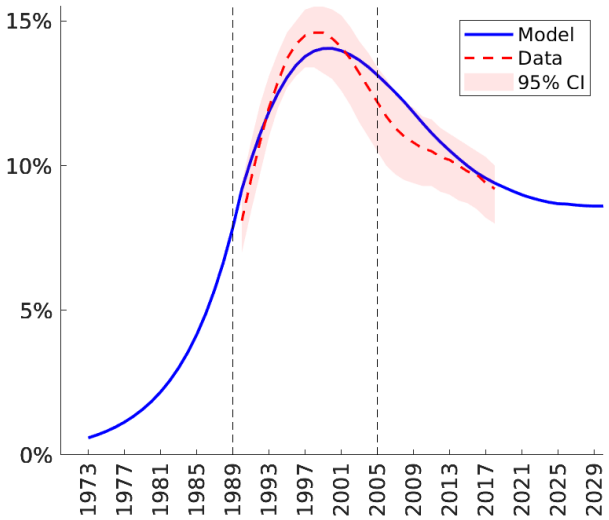


Figure 7: Evolution of the HIV Prevalence rates



Notes: The figure compares prevalence generated by the model and data for Malawi. Data source: UNAIDS estimates 2019.

4.2 Comparative Statics

This section performs comparative statics concerning the parameters estimated in Section 4. Figure 8 shows contour-maps for some of the moments of interest. These moments were computed using model generated data for given combination of parameter values. The red contour line represents the value observed in the data that is to be matched. The intersection between the red lines was used as the initial guess for the SMM estimation, this considerable speeds up the estimation. Looking at the contour plots and the red lines, we can easily see the parameters that identify each the moment. We now explore identification in some detail.

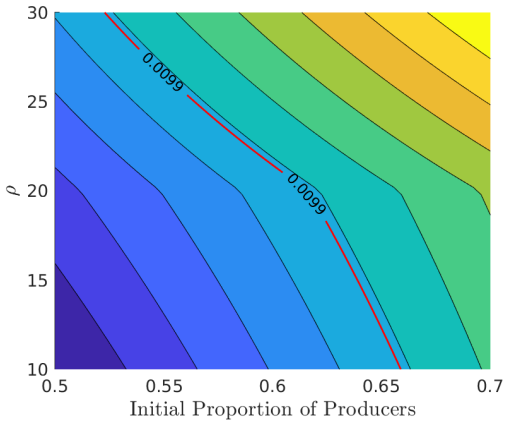
Panel (a) and (b) show that the HIV education gradient at the peak is more sensitive to variations of the initial proportion of less educated Risky-sex consumers $\vartheta_{-i,e=0}$, whereas the parameter governing the infection rates ρ has a larger impact on the prevalence's at the peak.

In Panels (c) and (d) we see that both the standard deviation of the signal's noise for less-educated $\sigma^2_{\epsilon,e=0}$ and the preference parameter χ, play a role in determining the HIV education gradient. However, Panel (d) shows that the time to reach the prevalence at the bottom is mostly sensitive to the preference parameter χ.

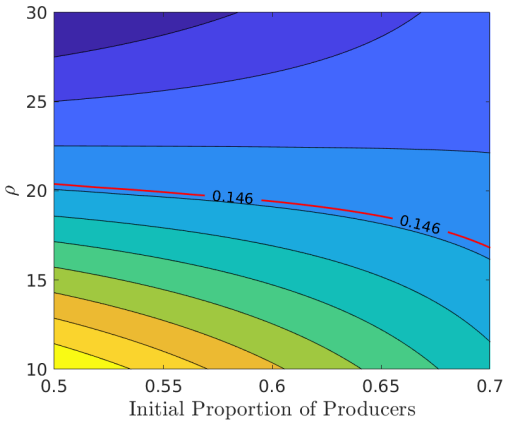
Finally, Panels (e) and (f) show that the HIV education gradient responds strongly to changes in the coverage differences by education $\iota(e)$ and that the level of the prevalence moves along with the new value of the infection rate parameter ρ.

Figure 8: Comparative Statics

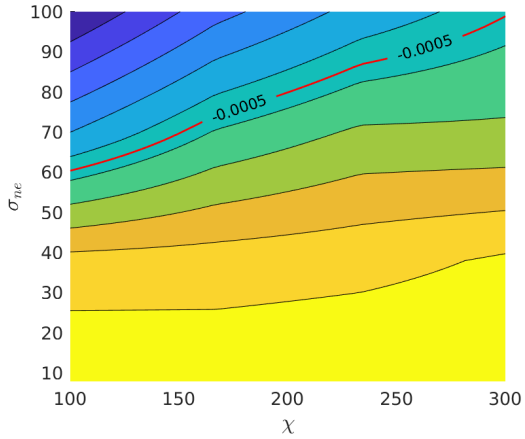
(a) Education Gradient at the peak



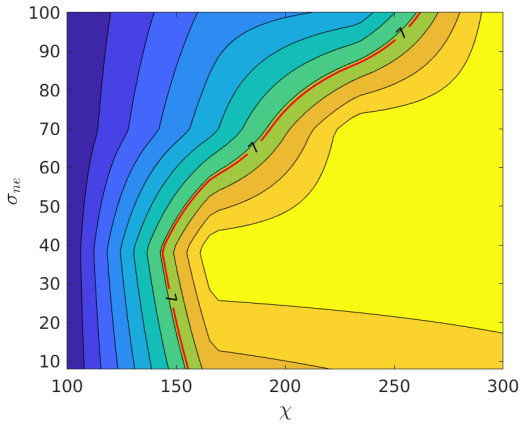
(b) Prevalence at the peak



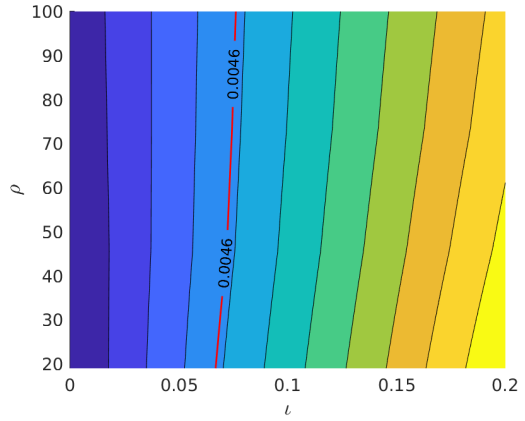
(c) Education Gradient at the end of Maturity



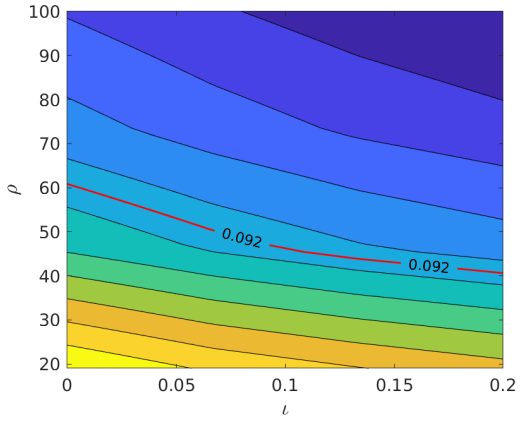
(d) Years to reach from peak to end of Maturity



(e) Education Gradient ARV stage

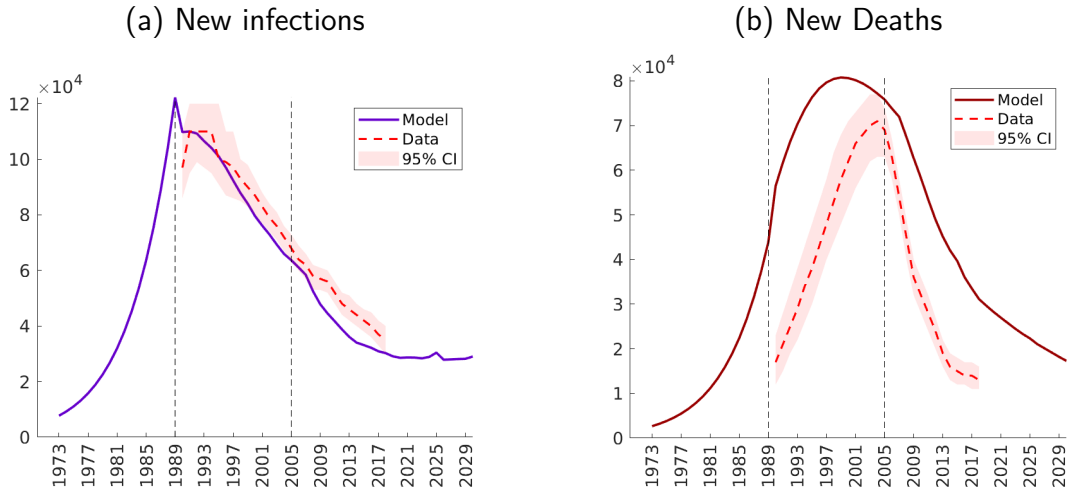


(f) Prevalence ARV stage



Notes: Panels (a) and (b) show sensitivity of the HIV Education Gradient at the peak and the prevalence at the peak with respect to the infection parameter ρ and the initial proportion of sex producers $\vartheta_{-i,e=0}$. Panels (c) and (d) display the sensitivity of the HIV Education Gradient at the end of Maturity with respect to the preference parameter χ and the variance of the signal for less educated individuals $\sigma^2_{\epsilon,e=0}$. Finally panels (e) and (f) show the sensitivity of the the gradient and prevalence after the introduction of ARV's, this time with respect to the coverage difference parameter ι and the new value of ρ.

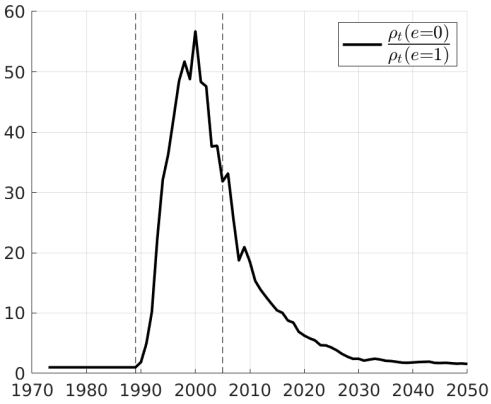
Figure 9: Non Targeted Moments



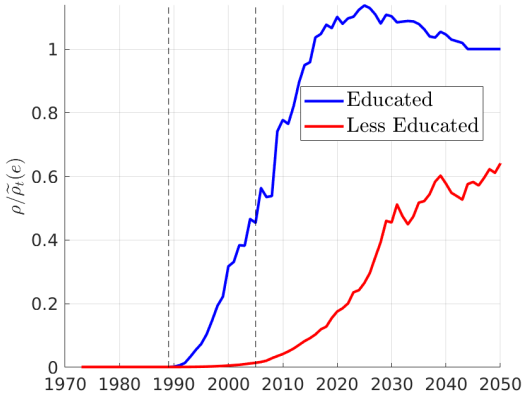
Notes: Panels (a) and (b) show the evolution of the new infections and new deaths respectively. Panel (a) shows that the model does are very good job matching infections although the series was not directly targeted.

Figure 10: Comparing beliefs (ρ) between education groups.

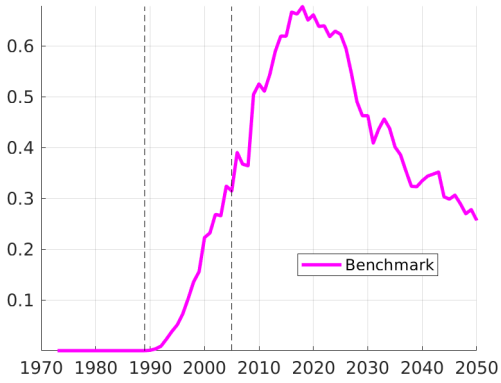
(a) Beliefs ratio $\rho_t(e=0)/\rho_t(e=1)$



(b) Relative belief convergence by education group



(c) Knowledge of ρ Education Gradient



5 What drives the epidemic?

Education affects the risky sex behaviour of individuals through three different channels along the evolution of the HIV epidemic: First, higher current income increases the amount of risky sex consumed because risky sex is a normal good. This channel is present throughout the complete duration of the HIV epidemic. Second, during Stage 1-2 more-educated individuals acquire information about the negative effects of HIV faster than their less-educated counterparts, which in turn reduces their risky sex consumption. Third, at stage 3-4, educated individuals have better access to (costly) ARV's that mitigate the effects of the disease, therefore increasing incentives for risky sex consumption. In this section we explore counterfactual experiments neutralizing the second and third channels, while keeping the first present. Note that neutralizing the first channel would be analogous to having a starting HIV Education-gradient equal to zero.
Depending on their magnitude, changes of the education composition of the population can also affect the dynamics of the HIV epidemic [47]. To account for this, we experiment by exogenously varying the education composition of the population across stages to resemble that of Malawi.

5.1 Removing Learning Asymmetry

Experiments 1-2: Removing learning differences between education groups
We explore two different ways of removing learning differences between the educated and the less-educated:

1. Making the less-educated individuals learn at the same (fast) speed as their more-educated counterparts. This means setting $\sigma^2_{\epsilon,e=0} = \sigma^2_{\epsilon,e=1} = 7.9$
2. Making the more-educated individuals learn at the same (slow) speed as their less-educated counterparts. This means setting $\sigma^2_{\epsilon,e=1} = \sigma^2_{\epsilon,e=0} = 84$

Columns 3-4 on Table 9 show the results for these two experiments. If the less-educated learn as fast as the educated (Column 3), we find that 52% of the HIV Education gradient and 53% of the Partners Education gradient is explained by the information differences, as before this effect is carried out to the next stage, where 36% of the HIV Education gradient is explained by information differences coming from the previous stage.
Similarly, if the more-educated learn at the same slow rate as the less educated (Column 4), we see that the gradient keeps increasing, to reach a peak at 0.08. In this experiment gents learn to

[47] Note that throughout the different stages of the benchmark calibration we indirectly maintained the joint distribution of education groups and sex types constant across time.

slow therefore delaying the maturity of the epidemic.
Figure 14 illustrate these counterfactual experiments against the benchmark.

5.2 Removing Differences in Access to ARVs

Experiment 3: Removing coverage access differences between education groups
This experiment involves setting $\iota = 0$. This means that educated individuals will get the same chance of treatment as everyone else in the population. Column 5 on Table 9 show the results for this experiment. We can see that the gradient continues to be negative, consistently the HIV prevalence at after ARV introduction is higher.

5.3 Accounting for the evolution of the education composition

Experiment 4: Accounting for the evolution of the education composition in Malawi
In this experiment we exogenously modify the education composition of the population to mimic that of Malawi. We achieve this by modifying the proportion of new born educated such that along the equilibrium path the resulting (endogenous) proportion of educated vs less educated resembles the evolution path observed in the data. In 1977 only 7.3% of the population above 25 years old had completed primary education. By 2000 this number increased to 19.8% [WorldBank, 2020].

Table 9: Counterfactual Experiments

Observation	Benchmark	Experiments			
		1 Fast Learning	2 Slow Learning	3 $\iota = 0$	4 Increasing Education
HIV Education gradient at end of HIV Myopia (1989)	0.0595	0.0595	0.0595	0.0595	0.0533
HIV Education gradient at the peak (1999)	0.0112	0.0527	0.0796	0.0112	-0.0177
HIV Education gradient at the end of Maturity (2005)	-0.0015	0.0310	0.0720	-0.0015	-0.0294
HIV Education gradient ARV stage (2018)	0.0047	0.0246	0.0531	-0.0027	-0.0183
Prevalence at the peak (1999)	14.1%	10.9%	16.8%	14.1%	14.1%
Prevalence at the end of Maturity (2005)	13.1%	7.1%	16.4%	13.1%	13.6%
Prevalence ARV stage (2018)	9.4%	5.4%	12.2%	9.8%	10.4%
Time to reach from bottom to peak (in years)	28	21	30	28	26
Time to reach from peak to end of Maturity (in years)	6	13	4	6	6

6 Policy Experiments

In this section we explore a set counterfactual experiments concerning the evolution of the HIV epidemic and policy interventions. We ask ourselves how would heave the HIV epidemic evolved in the following cases:

1. Early acknowledgement of the presence of the virus and its workings.
2. Improving the education level of the population.
3. Early and universal adoption of ARVs.
4. HIV prevention through Pre-exposure prophylaxis (PrEP).

Early acknowledgement of the presence of the virus and its workings

In Figure 17 panels (a) to (c) we show how the epidemic would have evolved had the learning stage started 5 years earlier and 10 years earlier than in the benchmark calibration. Panel (a) shows that the peak reaches a lower level the earlier the learning starts. Along the same line, the HIV education gradient and the sexual partners gradient start dropping at earlier stages.

Improving the education level of the population

In 2000 the UN set out a series of time-bound targets, with a deadline of 2015, that were aimed to reduce extreme poverty around the globe. These targets have later become known as the Millennium Development Goals (MDGs) and were succeeded by what we currently know as the UN Sustainable Development Goals (SDGs) to be achieved by 2030. One of the goals in the agenda (Goal 4) aims to ensure that all girls and boys complete free, equitable and quality primary and secondary education.

In our experiment we explore the possibility of achieving this goal by 2030. We do this by gradually increasing the proportion of educated new born individuals over time $\{\vartheta_{e=1,i,t}\}_{t=2000}^{\infty}$ such that $\lim_{t\to\infty} \vartheta_{e=1,i,t} = 85\%$ (the current level attained by South Africa) [48]. We start in 2000 but we also explore an alternative scenario where the starting point is 2018 instead, in this case the goal is achieved by 2048.

[48] The education composition for South Africa was chosen as a more realistic alternative to fully achieving primary education attainment as per the SDG's, that is $\lim_{t\to\infty} \vartheta_{e=1,i,t} > 85\%$. However, we explored by setting the limit to 90% and the main conclusions remained unaltered.

Figure 14 panels (d) to (f) show the results of these experiments. We see that there is no significant change in the prevalence levels however, once a larger share of the population educated, the HIV education gradient and Risky sex education gradient turn negative for the rest of the path.

Early and universal adoption of ARVs

In Figure 17 panels (d) to (f) we show how the epidemic would have evolved if ARV's were introduced at an earlier stage. If ARV's would have been introduced 6 years earlier (that is at the peak of the epidemic in 1999) universal coverage would have been attained by 2019 however the prevalence rate does not show a significant reduction. Moreover, if ARV's were introduced 16 years earlier (that is 1989 which is also the year when agents start learning about HIV) universal coverage would have been attained by 2009 and the prevalence rate would have been significantly reduced.

HIV prevention through Pre-exposure prophylaxis (PrEP)

Pre-exposure prophylaxis (PrEP) is taken on a daily basis by HIV-negative people as protection from HIV infection. Evidence shows that PrEP reduces the chances of HIV infection to near-zero (99% effectiveness) when taken consistently and correctly [Avert, 2020]. PrEP is not widely available in Malawi, although a clinical trial among HIV-positive pregnant adolescents and young women (ages 16-24) and an implementation study for at-risk adults and adolescents are underway [Avert, 2020].

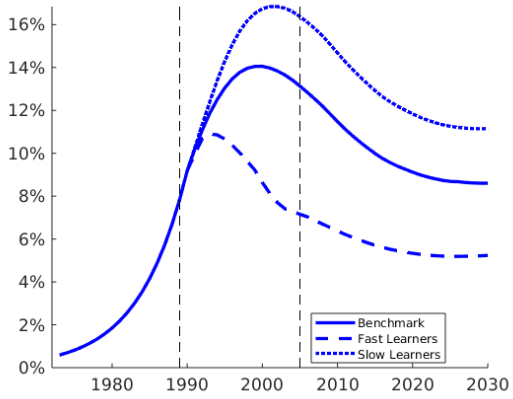
We construct an experiment where we analyze the evolution of the epidemic as if PrEP was implemented nationwide starting in 2018[49], with coverage gradually increasing and reaching full coverage by 2040; we further assume PrEP is administered at no costs and that there is no differences on PrEP take-up by education. In the model, PrEP implies a higher value of ρ for those taking the drug, we calibrate this value such that the current average probability of infection is reduced by 99%. Moreover, we ask ourselves: What if PrEP would only reduce the infection probability by 30%, 50% and 70%? would it still be worth scaling up its use?

Panels (a) to (c) of Figure **??** show that in all four cases the prevalence levels reduce considerably as well as the HIV incidence and the numbers of HIV deaths. From the model perspective we can conclude that PrEP implementation is highly recommended at any degree of effectiveness.

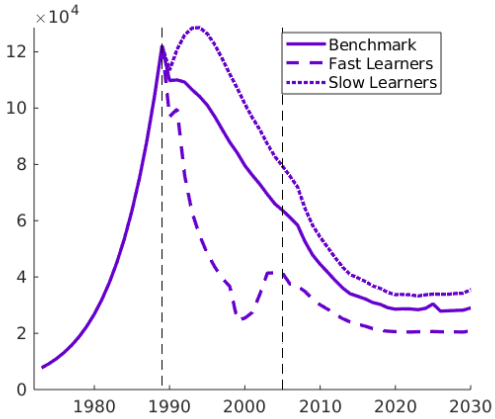
[49]This involves the introduction of a new state variable: those who are healthy can now be taking PrEP. We introduce transitions to PrEP in the same spirit as previously done with ARV's.

Figure 11: Experiment 1-2, Removing Learning differences

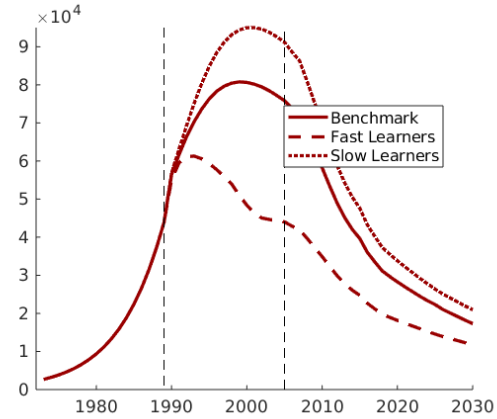
(a) Prevalence



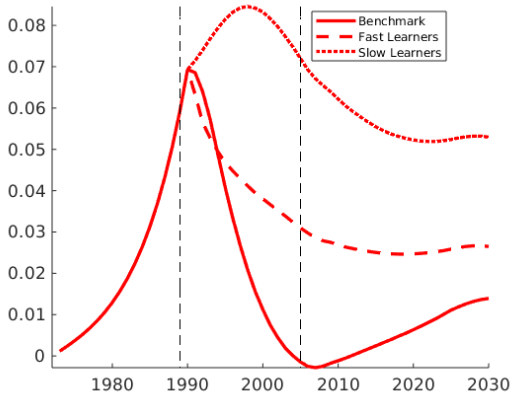
(b) New Infections



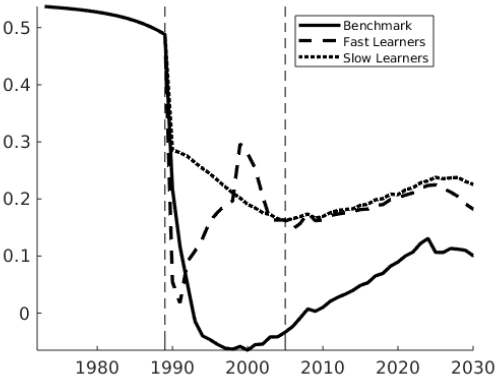
(c) New Deaths



(d) HIV Education Gradient

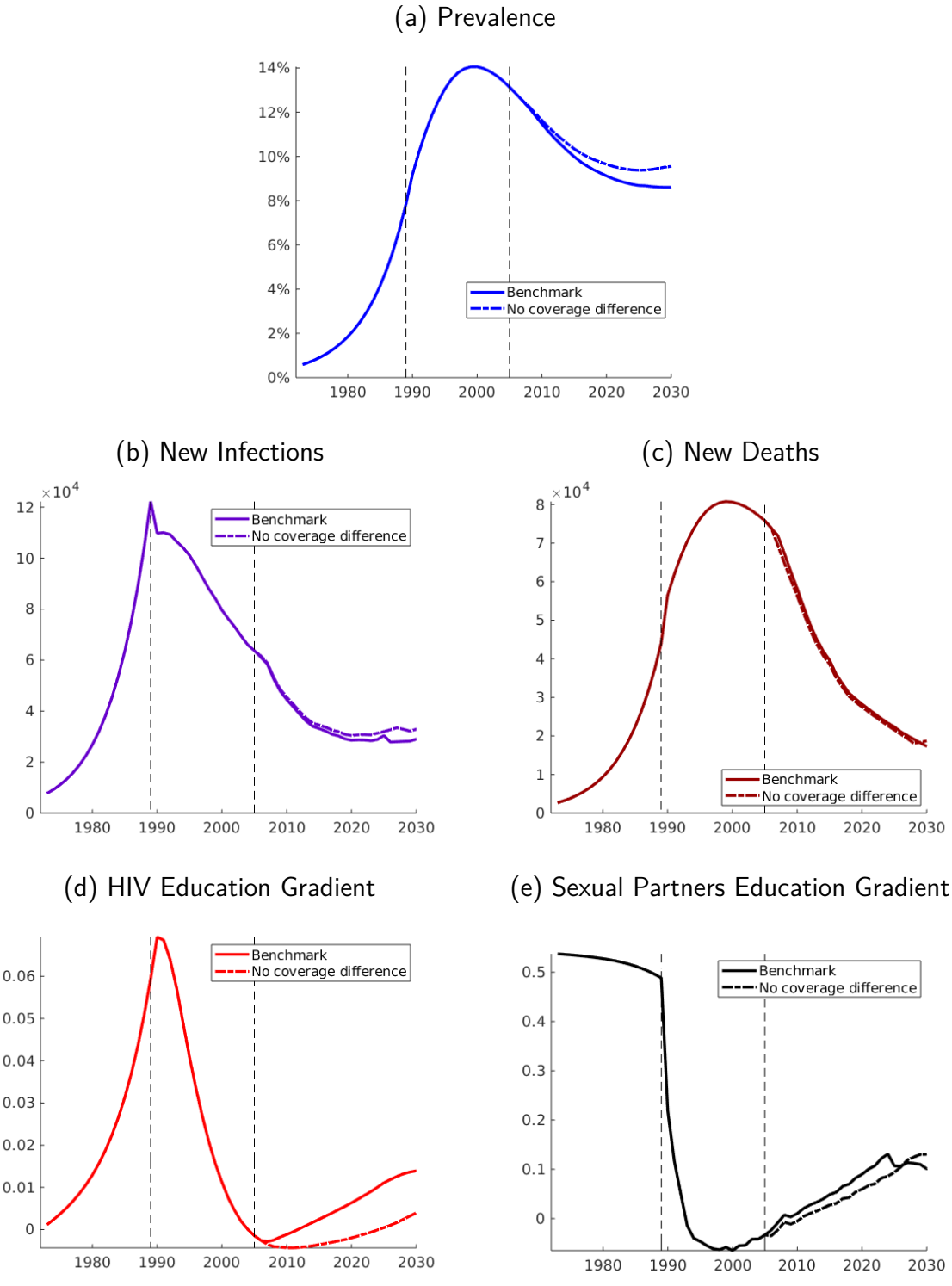


(e) Sexual Partners Education Gradient



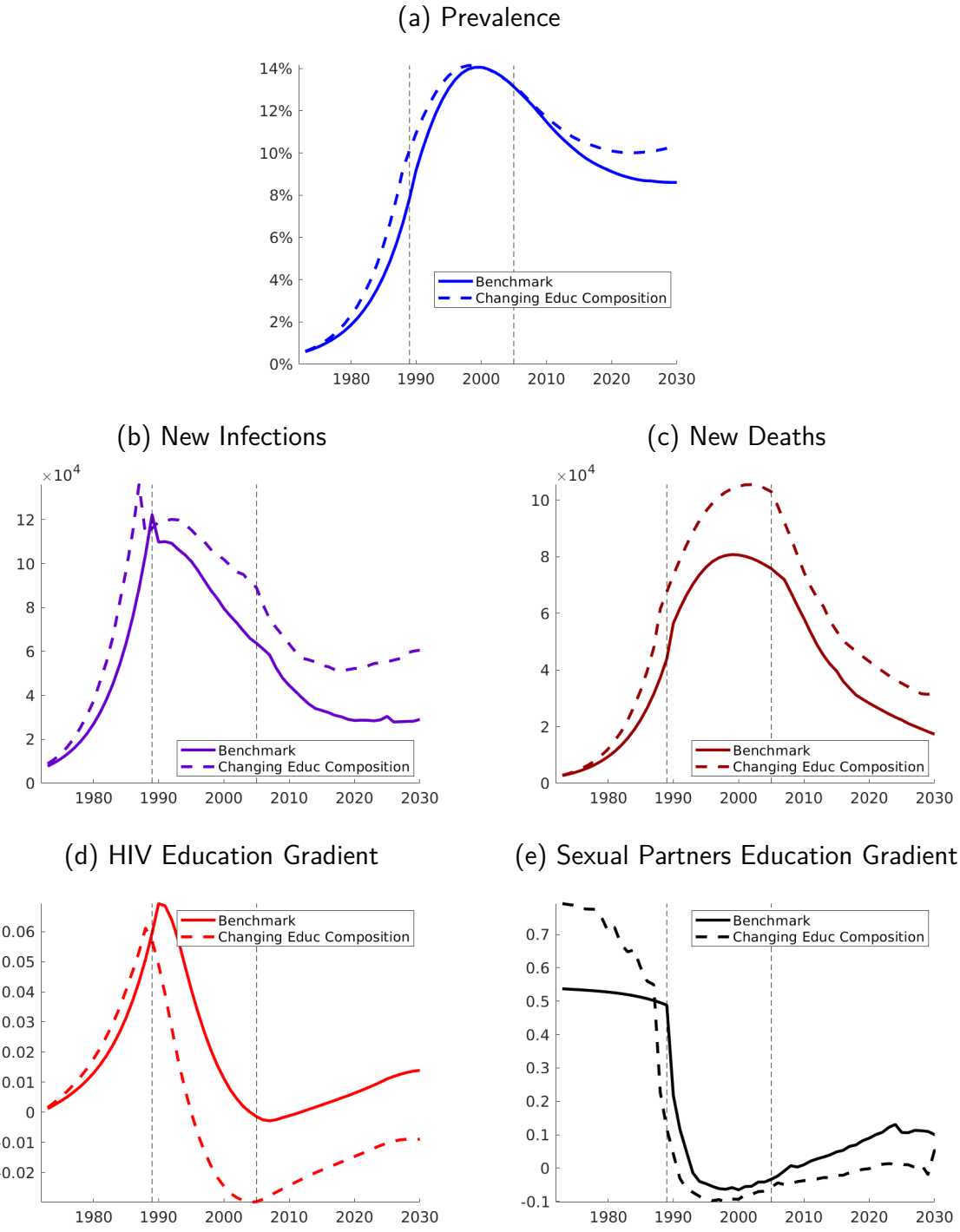
Notes: The figure shows the results of different counterfactual experiments; each row shows the effects on the time series of the prevalence rate, HIV education gradient panel and Sexual partners education gradient. Panels (a) to (c) show experiments 1-3 and panels (d) to (f) illustrate a gradual increase of the proportion educated agents. The figures also show the effects for different starting dates.

Figure 12: Experiment 3, Removing ARV accessibility differences



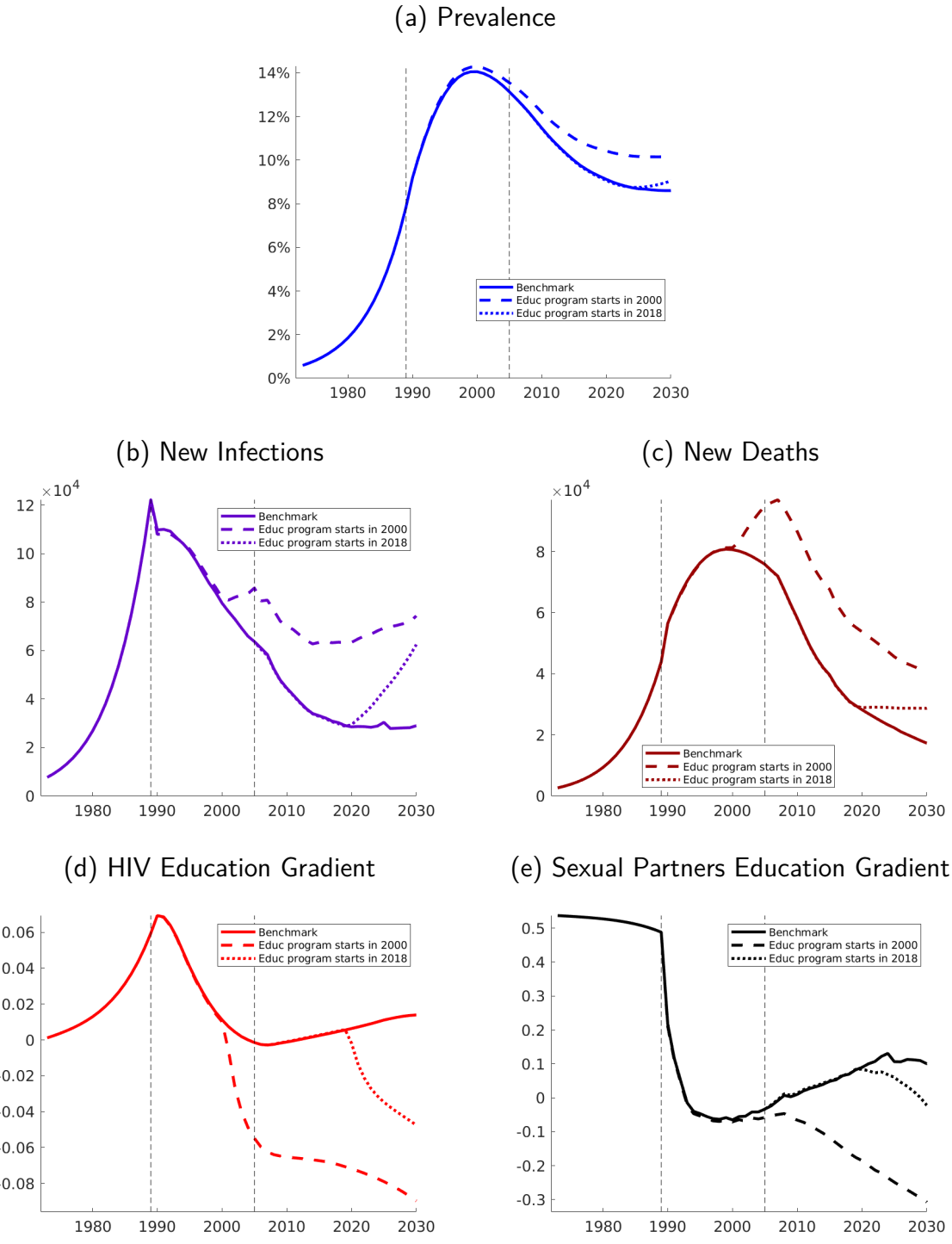
Notes: The figure shows the results of different counterfactual experiments; each row shows the effects on the time series of the prevalence rate, HIV education gradient panel and Sexual partners education gradient. Panels (a) to (c) show experiments 1-3 and panels (d) to (f) illustrate a gradual increase of the proportion educated agents. The figures also show the effects for different starting dates.

Figure 13: Experiment 4, Accounting for the evolution of the education composition



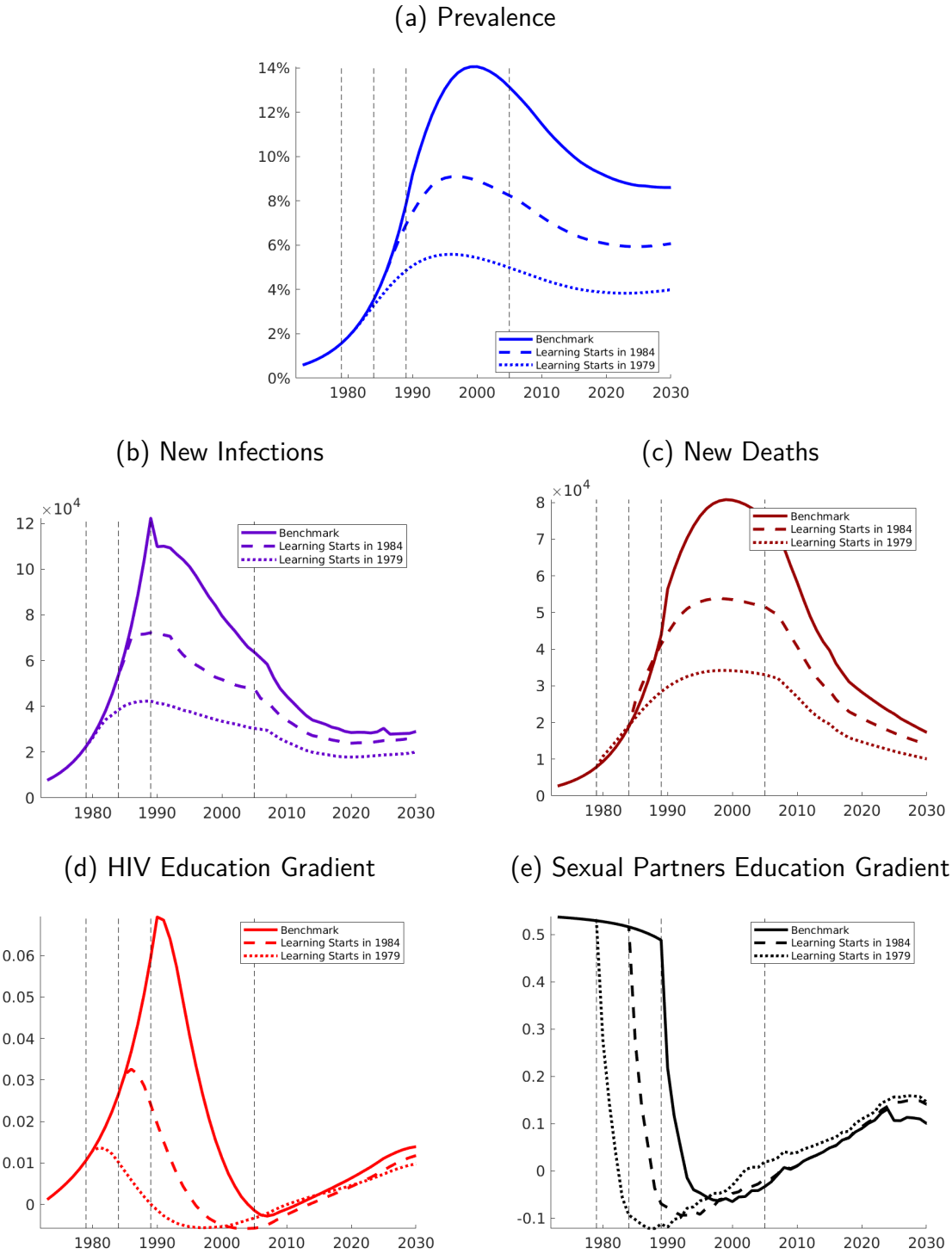
Notes: The figure shows the results of different counterfactual experiments; each row shows the effects on the time series of the prevalence rate, HIV education gradient panel and Sexual partners education gradient. Panels (a) to (c) show experiments 1-3 and panels (d) to (f) illustrate a gradual increase of the proportion educated agents. The figures also show the effects for different starting dates.

Figure 14: Policy Experiment 1, Increasing the education level of the population



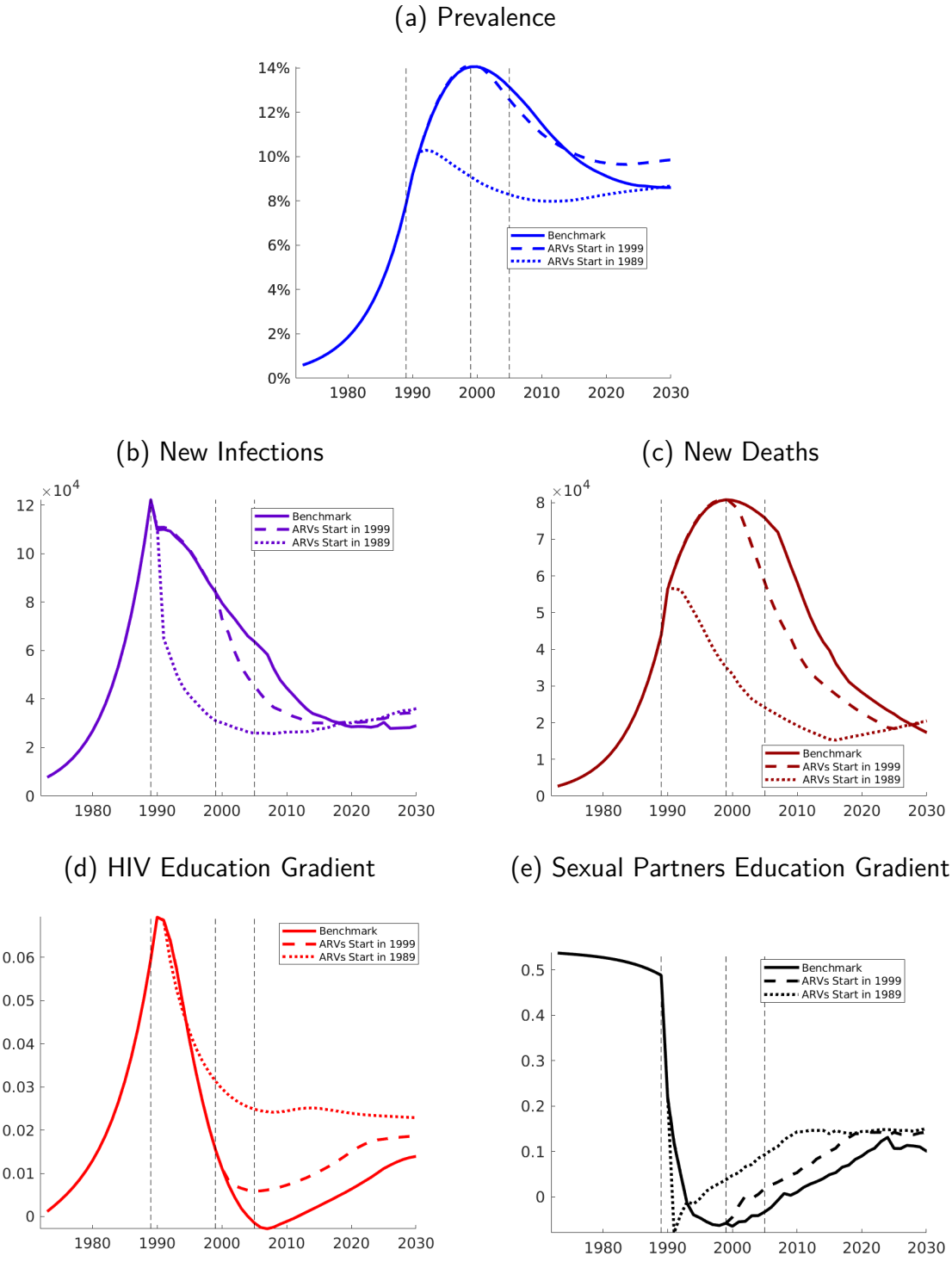
Notes: The figure shows the results of different counterfactual experiments; each row shows the effects on the time series of the prevalence rate, HIV education gradient panel and Sexual partners education gradient. Panels (a) to (c) show experiments 1-3 and panels (d) to (f) illustrate a gradual increase of the proportion educated agents. The figures also show the effects for different starting dates.

Figure 15: Policy Experiment 2, Early Learning



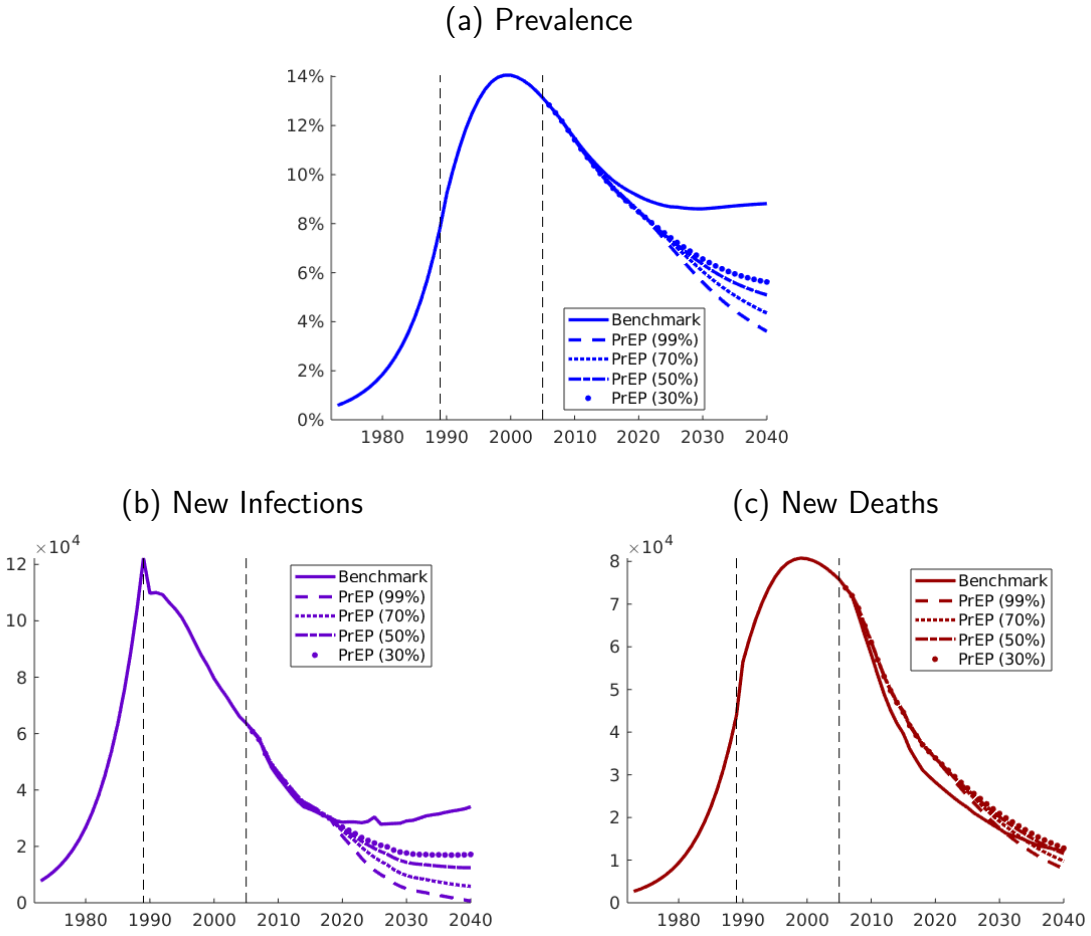
Notes: The figure shows the results of different counterfactual experiments, panels (a) to (c) show the effects of different (earlier) starting dates of the learning stage. Panels (d) to (f) simulate different dates for the introduction of ARVs.

Figure 16: Policy Experiment 3, Early Adoption of ARVs



Notes: The figure shows the results of different counterfactual experiments, panels (a) to (c) show the effects of different (earlier) starting dates of the learning stage. Panels (d) to (f) simulate different dates for the introduction of ARVs.

Figure 17: Policy Experiment 4, Preventing HIV infection with PrEP

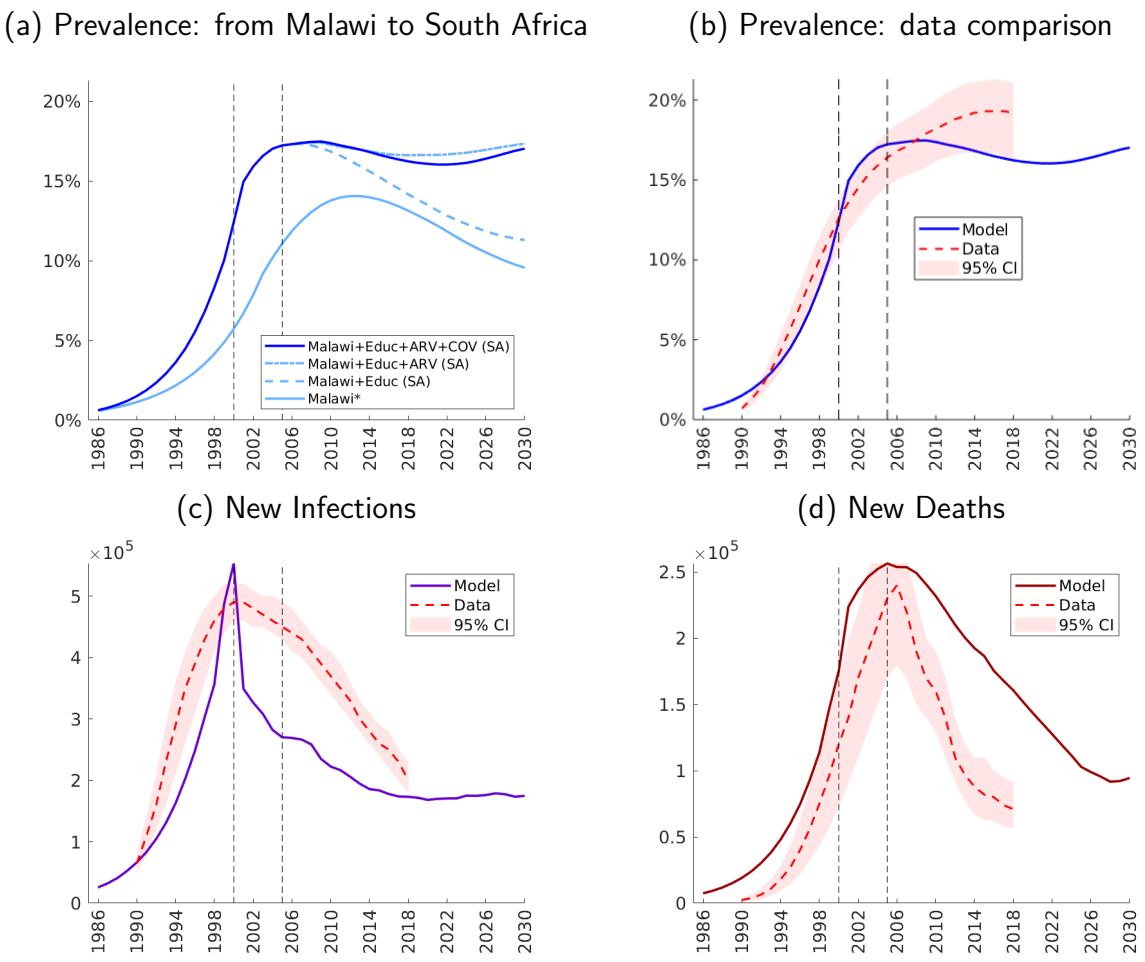


Notes: The figure shows the results of different counterfactual experiments, panels (a) to (c) show the effects of different (earlier) starting dates of the learning stage. Panels (d) to (f) simulate different dates for the introduction of ARVs.

7 Can this theory explain epidemic heterogeneity across countries?

We argue that with the adequate information, the model is certainly capable of characterizing the evolution of the HIV epidemic for individual countries. To illustrate this, we show in Figure 18 the model results characterizing the evolution of the HIV epidemic in South Africa (SA). To obtain these results the benchmark parametrization was modified to account for the South African education composition [50], the year of introduction of ARV's (2004) and the respective evolution of ARV coverage. According to AIDSinfo [2020], the ARV coverage in South Africa reached 59% in 2018, this compares to a 70% coverage in Malawi the same year.

Figure 18: Characterizing the South African (SA) HIV Epidemic



*The starting year of the epidemic in Malawi was normalized to that of South Africa(SA), that is 1986.

[50] In 1985 only 33% of the South African population had finished primary school, by 2018 it increased to 85% of the population [WorldBank, 2020].

8 Conclusion

The mixed evidence in the literature investigating the relationship between education and the probability of being HIV-positive in SSA suggests that finding which type of individuals are at greater risk of HIV infection is not an easy task. We proposed a fresh look to this question that consists of explicitly introducing the stages of the HIV epidemic into the analysis. Using nationally representative data from 39 DHS surveys to exploit variation across stages of the HIV epidemic, we showed that the relationship between completed educational attainment and individual HIV status (i.e., the HIV-Education gradient) is dynamic, and significantly evolves with the epidemic. At early stages of the epidemic more-educated individuals are more likely to be infected; however, this relationship strongly decreases as the epidemic evolves, and eventually reaches a stage where education and the probability of being HIV-positive are no longer significantly correlated. Interestingly, in the most advanced stages of the epidemic, the education gradient of HIV returns to being high and positive. We showed theoretically and empirically that the educational disparities in risky sexual behavior (in terms of extramarital partners) closely resemble the U-shaped pattern of the education gradient in HIV. In light of our findings, we proposed a quantitative frameworks of policy evaluation that incorporate the stylized dynamic relationship between education, HIV and risky sexual behavior that we document along the course of the HIV epidemic. We have found that asymmetric learning about the process of infections across education groups goes a long way in explaining the patterns in the data.

References

J. Behrman and HP. Kohler. Sexual Transmission of HIV. In Bjorn Lomborg, editor, *Rethink HIV*, chapter 1. Cambridge University Press, 2012.

Jeremy Greenwood, Philipp Kircher, Cezar Santos, and Michèle Tertilt. An equilibrium model of the african hiv/aids epidemic. Working Paper 18953, National Bureau of Economic Research, April 2013. URL http://www.nber.org/papers/w18953.

Áureo DePaula, Gil Shapira, and Petra E. Todd. How Beliefs About Hiv Status Affect Risky Behaviors: Evidence From Malawi. *Journal of Applied Econometrics*, 29(6):944–964, 09 2014. URL https://ideas.repec.org/a/wly/japmet/v29y2014i6p944-964.html.

Martina Björkman Nyqvist, Lucia Corno, Damien B. C. M. De Walque, and Jakob Svensson. Using Lotteries to Incentivize Safer Sexual Behavior: Evidence from a Randomized Controlled Trial on HIV Prevention. Policy Research Working Paper Series 7215, The World Bank, March 2015. URL https://ideas.repec.org/p/wbk/wbrwps/7215.html.

Jennifer Kates, Adam Wexler, Eric Lief, Carlos Avila, and Benjamin Gobet. Financing the response to aids in low- and middle- income countries: International assistance from donor governments in 2010. Joint United Nations Programme on HIV/AIDS and The Henry J. Kaiser Foundation, 2011.

UNAIDS. Report on the global aids epidemic. Joint United Nations Programme on HIV/AIDS, New York, 2015.

Kathleen Beegle and D. de Walque. Demographic and socioeconomic patterns of hiv/aids prevalence in africa. World Bank Policy Research Working Paper, 5076, 2009.

Robert E. Lucas. The Industrial Revolution: past and future. *Annual Report*, (May):5–20, 2004. URL http://ideas.repec.org/a/fip/fedmar/y2004imayp5-20nv18no.1.html.

Berthold Herrendorf, Richard Rogerson, and Ákos Valentinyi. Chapter 6 - growth and structural transformation. In Philippe Aghion and Steven N. Durlauf, editors, *Handbook of Economic Growth*, volume 2 of *Handbook of Economic Growth*, pages 855 – 941. Elsevier, 2014. doi: http://dx.doi.org/10.1016/B978-0-444-53540-5.00006-9. URL http://www.sciencedirect.com/science/article/pii/B9780444535405000069.

O. Galor and D.N. Weil. Population, technology, and growth: From malthusian stagnation to the demographic transition and beyond. *American Economic Review*, 90:806–828, 2000.

Ronald Lee. The Demographic Transition: Three Centuries of Fundamental Change. *Journal of Economic Perspectives*, 17(4):167–190, Fall 2003. URL http://ideas.repec.org/a/aea/jecper/v17y2003i4p167-190.html.

Jeremy Greenwood, Ananth Seshadri, and Guillaume Vandenbroucke. The baby boom and baby bust. *American Economic Review*, 95(1):183–207, 2005. doi: 10.1257/0002828053828680. URL http://www.aeaweb.org/articles.php?doi=10.1257/0002828053828680.

S. Gregson, H. Waddell, and S. Chandiwana. School education and hiv control in sub-saharan africa: From discord to armony. *Journal of International Development*, 13:467–485, 2001.

Damien de Walque. 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, 2007.

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

Damien de Walque. Does education affect hiv status? evidence from five african countries. *The World Bank Economic Review*, 23(2):209–233, 2009.

Vinod Mishra, Simona Bignami-Van-Assche, Robert Greener, Martin Vaessen, Rathavuth Hong, Peter D. Ghys, J. Ties Boerma, Ari Van-Assche, Shane-Khan, and Shea Rutstein. Hiv infection does not disproportionately affect the poorer in sub-saharan africa. *AIDS*, 21(7):S17–S29, 2007.

David Cutler and Adriana Lleras-Muney. Education and Health. Working Paper, 2011.

David Cutler, Angus Deaton, and Adriana Lleras-Muney. The determinants of mortality. *Journal of Economic Perspectives*, 20(3):97–120, 2006.

Leandro De Magalhães and Raul Santaeulàlia-Llopis. The Consumption, Income, and Wealth of the Poorest: Cross-Sectional Facts of Rural and Urban Sub-Saharan Africa for Macroeconomists. World Bank, Policy Research Working Paper 7337, 2015.

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

John Bongaarts, Thomas Buettner, Gerhard Heilig, and Francois Pelletier. Has the hiv epidemic peaked? *Population and Development Review*, 34(2):199–224, 2008.

Gary D. Hansen and Edward C. Prescott. Malthus to Solow. *American Economic Review*, 92 (4):1205–1217, 2002. doi: 10.1257/00028280260344731. URL http://www.aeaweb.org/articles.php?doi=10.1257/00028280260344731.

Douglas Gollin, Stephen L. Parente, and Richard Rogerson. The role of agriculture in development. *American Economic Review*, 92(2):160–164, 2002.

Douglas Gollin, Stephen L. Parente, and Richard Rogerson. The food problem and the evolution of international income levels. *Journal of Monetary Economics*, 54(4):1230–850, 2007.

J. Bolt and J. L. van Zanden. The First Update of the Maddison Project; Re-Estimating Growth Before 1820. *Maddison Project*, (4), 2013.

Francisco J. Buera, Joseph P. Kaboski, and Richard Rogerson. Skill biased structural change. Working Paper 21165, National Bureau of Economic Research, May 2015. URL http://www.nber.org/papers/w21165.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427, August 2008. URL https://ideas.repec.org/a/tpr/restat/v90y2008i3p414-427.html.

Tom W. Smith. A methodological analysis of the sexual behavior questions on the general social surveys. *Journal of Official Statistics*, 8:309–325, 1992.

Mark Gersovitz, Hanan G. Jacoby, F. Seri Dedy, and A. Goze Tap. The balance of self-reported heterosexual activity in kap surveys and the aids epidemic in africa. *Journal of the American Statistical Association*, 93:875–883, 1998.

Stephane Helleringer, Hans-Peter Kohler, Agnes Chimbiri, Praise Chatonda, and James Mkandawire. The likoma network study: Context, data collection, and initial results. *Demography Research*, 21:427–468, 2009.

Taryn Dinkelman, James A. Levinsohn, and Rolang Majelantle. When knowledge is not enough: Hiv/aids information and risky behavior in botswana. NBER Working Paper, 12418, 2006.

Pascaline Dupas. Do teenagers respond to hiv risk information? evidence from a field experiment in kenya. *American Economic Journal: Applied Economics*, 3(1):1–36, 2011a.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Education, HIV, and Early Fertility: Experimental Evidence from Kenya. *American Economic Review*, 105(9):2757–97, September 2015a. URL https://ideas.repec.org/a/aea/aecrev/v105y2015i9p2757-97.html.

Rose Apondi, Rebecca Bunnell, John Paula Ekwaru, David Moore, Stevens Bechange, Kenneth Khana, Rachel King, James Campbell, Jordan Tappero, and Jonathan Mermin. Sexual behavior and hiv transmission risk of ugandan adults taking antiretroviral therapy: 3 year follow-up. *AIDS*, 10:1317–1327, 2011.

WHO. Access to Antiretroviral Drugs, 2014.

T. Déirdre Hollingsworth, Roy M. Anderson, and Christophe Fraser. Hiv-1 transmission, by stage of infection. *The Journal of Infectious Diseases*, 198(5):687–693, 2008. ISSN 00221899. URL http://www.jstor.org/stable/40254093.

Bénédicte Leynaert, Angela M. Downs, Isabelle de Vincenzi, and for the European Study Group on Heterosexual Transmission of HIV. Heterosexual Transmission of Human Immunodeficiency Virus: Variability of Infectivity throughout the Course of Infection. *American Journal of Epidemiology*, 148(1):88–96, 07 1998. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a009564. URL https://doi.org/10.1093/oxfordjournals.aje.a009564.

Raul Santaeulalia-Llopis and Daniela Iorio. Education, hiv status and risky sexual behavior: How much does the stage of the hiv epidemic matter ? Economics Working Papers ECO2016/09,

European University Institute, 2016. URL https://EconPapers.repec.org/RePEc:eui:euiwps:eco2016/09.

Malawi National Statistical Office. Malawi labor and force survey key findings report, 2013.

Leandro De Magalhaes and Raul Santaeulalia-Llopis. The consumption, income, and wealth of the poorest: An empirical analysis of economic inequality in rural and urban sub-saharan africa for macroeconomists. Technical report, 2018.

Mubaraq Dele Sulaimon, Adamu Auwal Muhammad, and Oluwafunmilayo Shofoyeke. Possible health and growth implications of prostitution in Nigeria: A theoretical perspective. MPRA Paper 88402, University Library of Munich, Germany, August 2018. URL https://ideas.repec.org/p/pra/mprapa/88402.html.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Education, hiv, and early fertility: Experimental evidence from kenya. *American Economic Review*, 105(9):2757–97, September 2015b. doi: 10.1257/aer.20121607. URL http://www.aeaweb.org/articles?id=10.1257/aer.20121607.

Pascaline Dupas. Do teenagers respond to hiv risk information? evidence from a field experiment in kenya. *American Economic Journal: Applied Economics*, 3(1):1–34, January 2011b. doi: 10.1257/app.3.1.1. URL http://www.aeaweb.org/articles?id=10.1257/app.3.1.1.

B Mwale. Hiv /aids in malawi, 2002. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346002/.

WorldBank. Education attainment, at least completed primary. The World Bank World Development Indicators Data Bank, 2020. https://data.worldbank.org/.

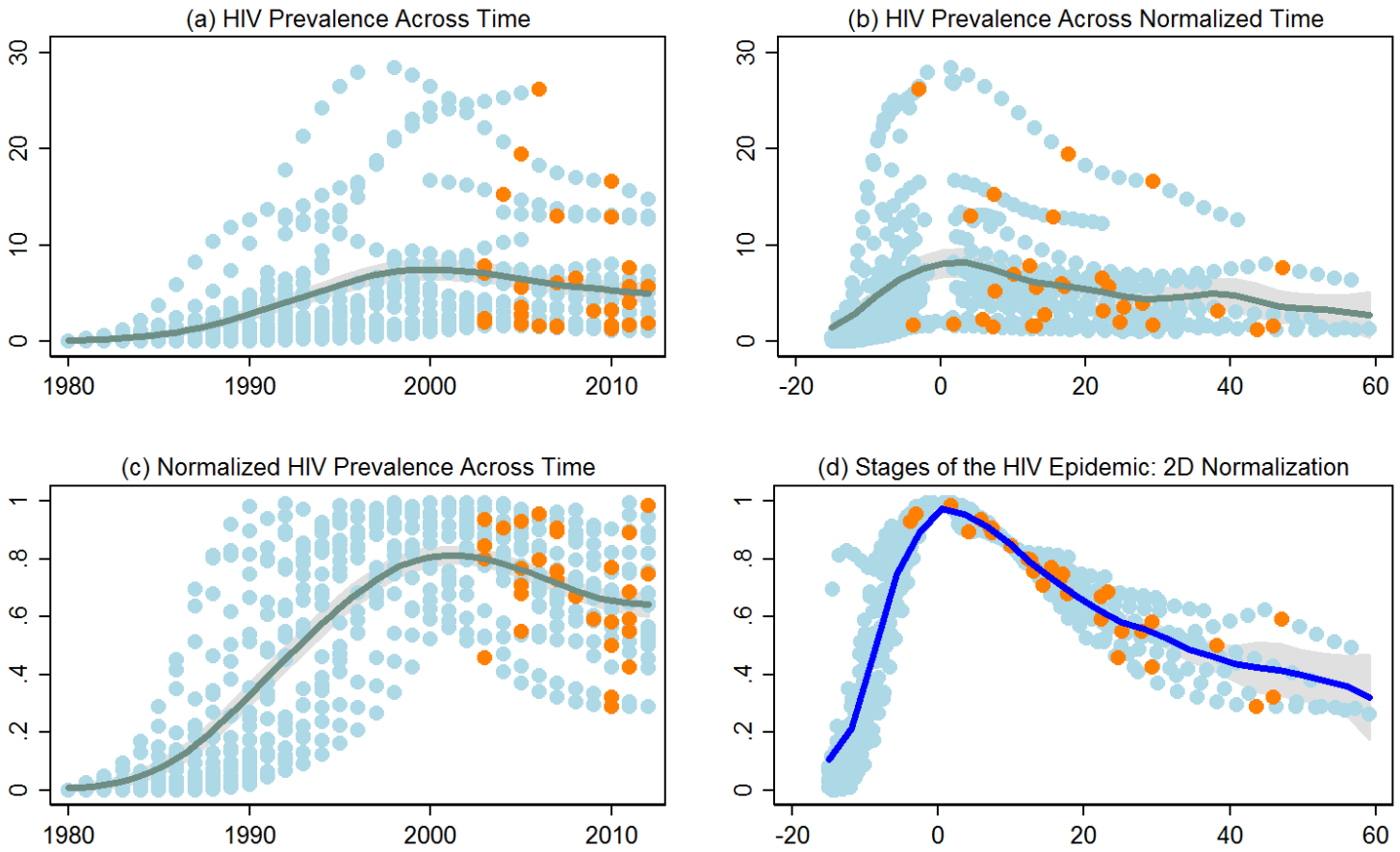
Avert. Prep for hiv prevention. global information and education on hiv and aids, 2020. URL https://www.avert.org/professionals/hiv-programming/prevention/pre-exposure-prophylaxis#footnote19_oj4uglg.

AIDSinfo. Unaids, aidsinfo dashboard, 2020.

Online Appendix

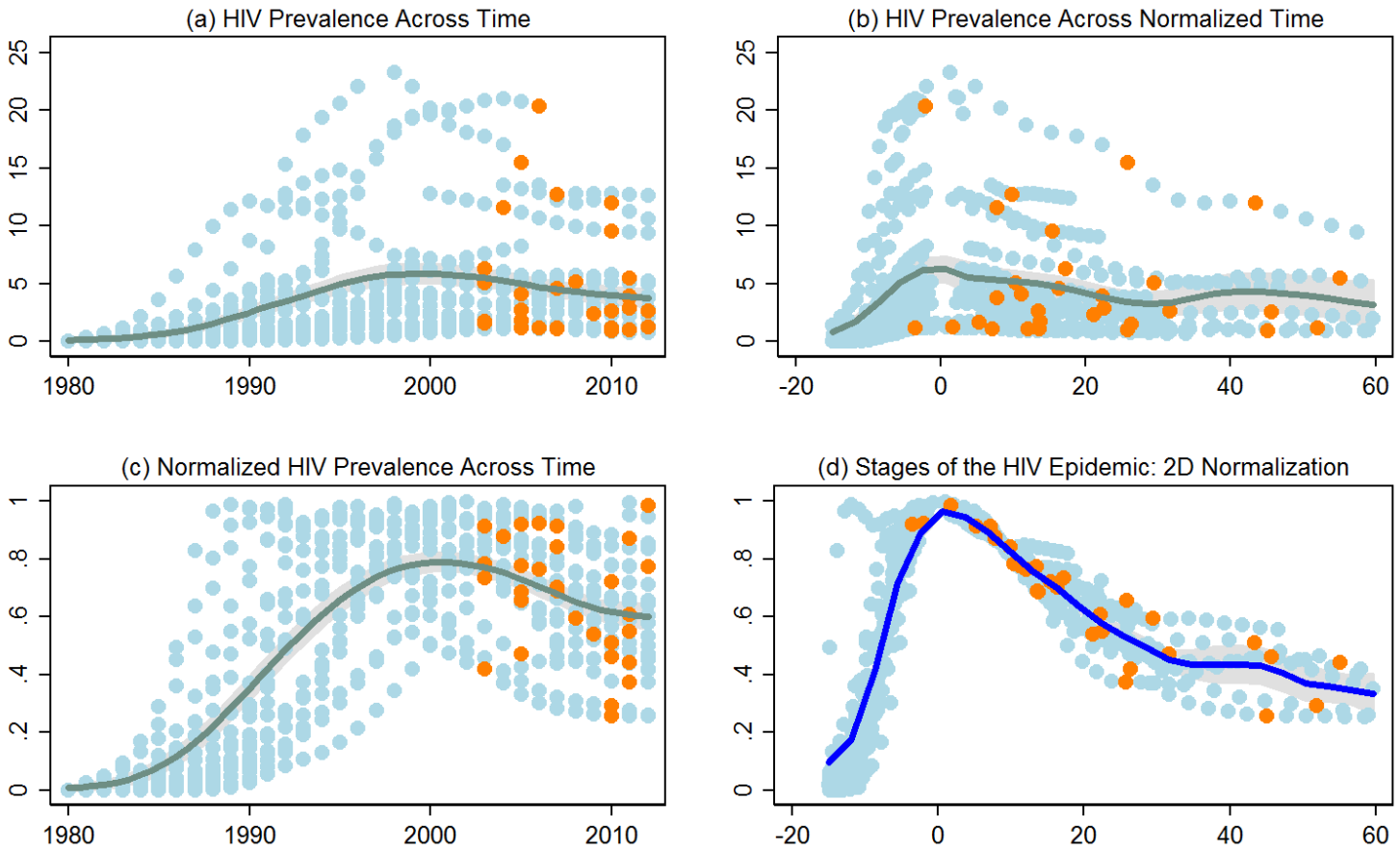
*Not for publication

Figure A-1: Stages of the HIV Epidemic for Women



Source: Location of SSA Countries (DHS Sample) on the 2D-Normalized Space at the Time of DHS Data Collection. Outcome of our 2D-normalization algorithm (Subsection 2.2.2) implemented using WPP, 2015, data. *Notes*: Each point in the scatterplot represents a DHS dataset.

Figure A-2: Stages of the HIV Epidemic for Men



Source: Location of SSA Countries (DHS Sample) on the 2D-Normalized Space at the Time of DHS Data Collection. Outcome of our 2D-normalization algorithm (Subsection 2.2.2) implemented using WPP, 2015, data. *Notes*: Each point in the scatterplot represents a DHS dataset.

Table A-1: The HIV-Education Gradient, Sexually Active Sample

HIV Status	(1)	(2)	(3)	(4)	(5)
Education	0.0048***	0.0123***	0.0107***	0.0042***	0.0039***
	(0.0009)	(0.0008)	(0.0013)	(0.0003)	(0.0004)
Education * Stage1		-0.0064***	-0.0047***	-0.0007	-0.0004
		(0.0010)	(0.0013)	(0.0004)	(0.0005)
Education * Stage2		-0.0129***	-0.0115***	-0.0029***	-0.0027***
		(0.0009)	(0.0014)	(0.0003)	(0.0004)
Education * Stage3		-0.0101***	-0.0081***	-0.0018	-0.0016
		(0.0020)	(0.0015)	(0.0015)	(0.0014)
Education * Stage4		-0.0067***	-0.0054***	-0.0019***	-0.0014***
		(0.0013)	(0.0016)	(0.0004)	(0.0004)
Male	-0.0267***	-0.0273***	-0.0272***	-0.0268***	-0.0269***
	(0.0039)	(0.0038)	(0.0025)	(0.0022)	(0.0021)
Age	0.0021***	0.0021***	0.0021***	0.0022***	0.0022***
	(0.0004)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
Urban Area	0.0260***	0.0243***	0.0276***	0.0337***	0.0344***
	(0.0059)	(0.0059)	(0.0053)	(0.0038)	(0.0038)
Stage 1	-0.0018	0.0137***	0.0097	-0.0082***	0.0099***
	(0.0062)	(0.0047)	(0.0060)	(0.0012)	(0.0023)
Stage 2	0.0132	0.0573***	0.0676***	0.0010	0.0289***
	(0.0105)	(0.0122)	(0.0136)	(0.0029)	(0.0045)
Stage 3	-0.0114	0.0198	0.0330**	-0.0129**	0.0138**
	(0.0164)	(0.0149)	(0.0128)	(0.0061)	(0.0060)
Stage 4	-0.0027	0.0143***	0.0436***	-0.0183***	0.0056
	(0.0052)	(0.0040)	(0.0097)	(0.0025)	(0.0036)
Agricultural Share	-0.0034***	-0.0034***	-0.0035***	0.0027***	-0.0006
	(0.0003)	(0.0003)	(0.0004)	(0.0005)	(0.0004)
Output per Capita	-0.0000***	-0.0000***	-0.0000***	0.0001**	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Constant	0.1128***	0.0918***	0.0554***	-0.1603***	-0.1992***
	(0.0101)	(0.0075)	(0.0098)	(0.0493)	(0.0282)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes
Sample Size	329,205	329,205	329,205	329,205	329,205

Notes: All specifications use the "Sexually Active" subsample. he underlying econometric models are as specified in the columns of Table 3. Column (1) reports the tests results for the stationary specification. Columns (2) to (5) report the tests results for the non-stationary specification. We include the same set of controls and fixed effects as in our benchmark specifications in Table 3. Standard errors are clustered at the country level using the wild cluster bootstrap from Cameron et al. [2008], and reported in parenthesis. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A-2: The Knowledge-Education Gradient: Women and Men Separately

(A) One Sexual Partner without Other Partners

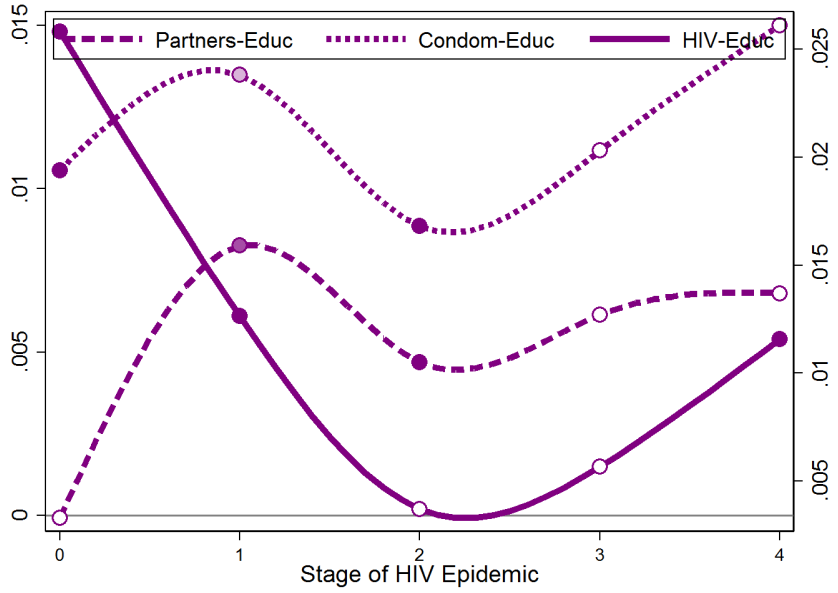
	Women					Men				
HIV Knowledge	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Education	0.0135	0.0033	0.0005	0.0074***	0.0076***	0.0076**	0.0010	0.0013	0.0078***	0.0087***
	(0.111)	(0.110)	(0.866)	(0.000)	(0.000)	(0.001)	(0.326)	(0.580)	(0.000)	(0.000)
Education * Stage1		0.0126	0.0120**	0.0047	0.0046		0.0078***	0.0059**	0.0005	-0.0001
		(0.156)	(0.043)	(0.309)	(0.292)		(0.000)	(0.039)	(0.840)	(0.952)
Education * Stage2		0.0072**	0.0093**	0.0038**	0.0033		0.0091***	0.0086***	0.0028*	0.0017
		(0.046)	(0.049)	(0.045)	(0.210)		(0.000)	(0.001)	(0.088)	(0.285)
Education * Stage3		0.0094	0.0120	0.0018	0.0020		0.0057	0.0053	-0.0034	-0.0042
		(0.612)	(0.412)	(0.912)	(0.911)		(0.386)	(0.337)	(0.444)	(0.304)
Education * Stage4		0.0104	0.0134	0.0071	0.0062		0.0069**	0.0075*	0.0019	0.0007
		(0.732)	(0.648)	(0.812)	(0.830)		(0.044)	(0.066)	(0.424)	(0.759)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes	No-No	No-No	Yes-No	No-Yes	Yes-Yes
Sample Size	213,907	213,907	213,907	213,907	213,907	167,894	167,894	167,894	167,894	167,894

(B) Always Use Condom During Sex

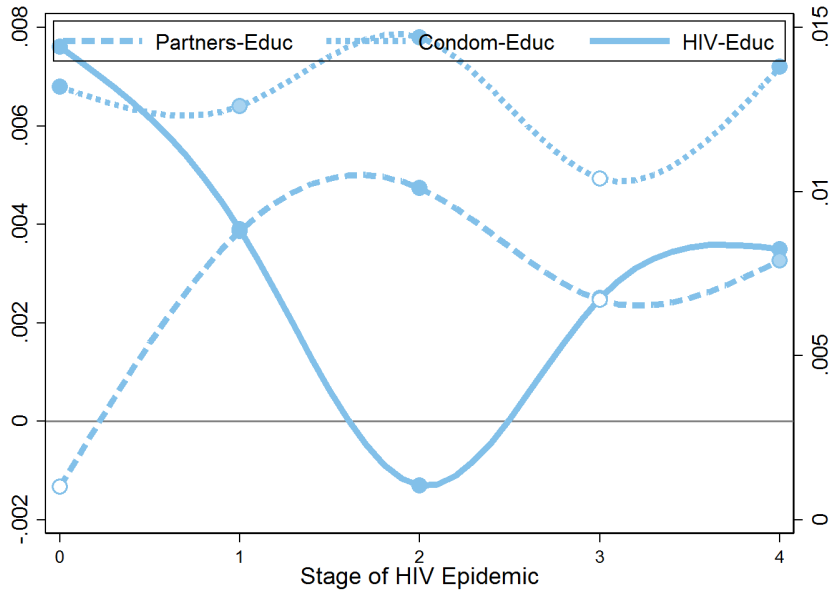
	Women					Men				
HIV Knowledge	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Education	0.0221	0.0194***	0.0169***	0.0196***	0.0196***	0.0126***	0.0132***	0.0141***	0.0187***	0.0181***
	(0.133)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
Education * Stage1		0.0044	0.0073	0.0034	0.0039		-0.0006	-0.0007	-0.0036	-0.0029
		(0.801)	(0.593)	(0.809)	(0.754)		(0.935)	(0.929)	(0.638)	(0.668)
Education * Stage2		-0.0026	0.0003	-0.0017	-0.0025		0.0015	0.0017	-0.0036	-0.0035
		(0.616)	(0.953)	(0.648)	(0.572)		(0.639)	(0.637)	(0.236)	(0.246)
Education * Stage3		0.0009	0.0039	0.0008	0.0014		-0.0028	-0.0024	-0.0073	-0.0064
		(0.956)	(0.804)	(0.952)	(0.920)		(0.770)	(0.789)	(0.415)	(0.496)
Education * Stage4		0.0067	0.0089	0.0047	0.0033		0.0006	-0.0004	-0.0046*	-0.0047**
		(0.946)	(0.929)	(0.963)	(0.974)		(0.828)	(0.910)	(0.065)	(0.029)
Year-Country Dum.	No-No	No-No	Yes-No	No-Yes	Yes-Yes	No-No	No-No	Yes-No	No-Yes	Yes-Yes
Sample Size	213,763	213,763	213,763	213,763	213,763	167,800	167,800	167,800	167,800	167,800

Notes: In panel (A) we report the coefficients of a linear model where the endogenous variable is binary for *"Can you (the respondent) reduce the chances of getting HIV by having one sex partner who has no other partners?"*. In panel (B) we report the coefficients of a linear model where the endogenous variable is binary for *"Can you (the respondent) reduce the chances of getting HIV by always wearing a condom?"*. In both panels we include the same set of controls and fixed effects as in our benchmark specifications in Table **??**. Standard errors are clustered at the country level using the wild cluster bootstrap from Cameron et al. [2008], and reported in parenthesis.* significant at 10%; ** significant at 5%; *** significant at 1%.

Figure A-3: The Knowledge-Education Gradient: Evolution Across Stages of the HIV Epidemic



(a) Women



(b) Men

Notes: The HIV-Education gradient is plotted on the left vertical axis. The HIV knowledge in terms of the Partners-Education gradient and the Condoms-Eduction gradient are plotted on the right vertical axis. For each stage j we plot $\left(\gamma_0 + \sum_{j>0} \gamma_j \mathbf{1}_j\right)$. The specification we plot is with year controls. That is, for the HIV-Education gradient we use column 3 (8) in Table **??** for respectively women (top panel) and men (bottom panel). For the HIV knowledge in terms of Partners-Education and Condoms-Education gradient we use column 3 (8) in Table A-2 for respectively women (top panel) and men (bottom panel). Significance at 10%, 5%, and 1% is represented by, respectively, markers with open circles, markers with medium transparency fill, and markers with solid fill. We use a cubic spline for interpolation across stages.

A Solution Algorithm

Solution Algorithm

Computing the recursive stationary equilibrium for Stage -1 (Pre-HIV Era)

***Algorithm No.1*: Computation of the recursive stationary equilibrium of an Aggregate Epidemic Stage:**

Step 1: Make initial guesses of price p (and prevalence ϕ^+ if not in Pre-HIV Stage).

Step 2: Compute the agents decision rules.

Step 3: Compute the stationary distribution of the population across states (follow Algorithm No.2).

Step 4: Compute aggregate sex demand and aggregate sex supply. Check the aggregate consistency conditions.

Step 5: If conditions are not met, update p and ϕ^+ and return to Step 2.

In the absence of continuous state variables, the decision rules are reduced to single values conditional on the different states
This algorithm is generic enough that can be used to compute the recursive stationary equilibrium of any aggregate stage of the epidemic.

***Algorithm No.2*: Computation of the invariant distribution of the population:**

Step 1: Make an initial guess for the (discrete) mass function ϕ_0 over the respective stage state space[51].

Step 2: For all individual states in $\mathbf{\Phi}$ compute the following expression:

if Stage: -1 (Pre-epidemic)

$$\phi_{t+1}(\mathbf{\Phi}') = \sum_{\mathbf{\Phi}} \sum_{s'|s} \gamma \pi(s'|s) \phi_t(\mathbf{\Phi}) + f \phi_t(\mathbf{\Phi}') \tag{39}$$

[51] We choose the uniform distribution as the initial values of the distribution. The algorithm should converge regardless of the choice of the initial distribution.

if Stage: 0 (HIV Myopia) or in Stage:1-2 (HIV Maturity)

$$\phi_{t+1}(\mathbf{\Phi}') = \sum_{\mathbf{\Phi}} \sum_{s'|s,h'|h} \gamma(h)\pi(s'|s)\phi_t^+ \lambda_{t,\rho}(h'|h)\phi_t(\mathbf{\Phi}) + f\phi_t(\mathbf{\Phi}') \tag{40}$$

if Stage:3-4 (ARV Era)

$$\phi_{t+1}(\mathbf{\Phi}') = \sum_{\mathbf{\Phi}} \sum_{\substack{d'|d,h'|h \\ s'|s}} \gamma(h,d)\pi(s'|s)\phi_t^+ \lambda_{t,\rho}(h'|h)\nu(d'|d)\phi_t(\mathbf{\Phi}) + f\phi_t(\mathbf{\Phi}') \tag{41}$$

Step 3: If $|\phi_{t+1} - \phi_t|$ is close to zero stop, otherwise set $\phi_t = \phi_{t+1}$ and return to Step 2.

Computing the recursive stationary equilibrium for Stage 0 (HIV Myopia)

***Algorithm No.3*: Computation of the recursive stationary equilibrium of the Myopic stage:**

Same as Algorithm No.1

Computing the Non-stationary equilibrium for Stage 0 (HIV Myopia)

***Algorithm No.4*: Computation of the solution of the Myopic stage:**
We are after a sequence of $\{\Phi_\tau\}_{\tau=T_{-1}}^{T_0}(g=0)$ where at each period $\tau \in \{T_{-1}+1, ..., T_0\}$, agents get a permanent unexpected shock to $\widetilde{\gamma}$, $\widetilde{z}$ and $\widetilde{\chi}$ following (17) and (18). T_0 being the period in which Stage 1-2 quick's in. To get each of the elements of the sequence we need to solve for an entire transition.

Step 0: Set $\tau = T_{-1} + 1$,

Step 1: Following (17) and (18) compute new values for $\widetilde{\gamma}_\tau$, $\widetilde{z}_\tau$ and $\widetilde{\chi}_\tau$ (Remember agents believe these values will be permanent).

Step 2: Compute the recursive stationary equilibrium of the Myopic stage associated with the new $\widetilde{\gamma}_\tau$, $\widetilde{z}_\tau$ and $\widetilde{\chi}_\tau$, (follow Algorithm No.2).

Step 3: Choose a very large number of transition periods $(T - \tau)$.

Step 4: Guess a time path for the prices $\{p_t\}_{t=\tau}^T$ and prevalences $\{\phi_t^+\}_{t=\tau}^T$.

Step 5: Compute the equilibrium policy (and value) functions iterating backwards in time, $t = T-1, ..., \tau$.

Step 6: Simulate the evolution of the population distribution from $t = \tau$ to $t = T$ with the help of the optimal policy functions and the initial distribution $\Phi_\tau(g = 0)$.

Step 7: Compare the simulated distribution at T with the stationary distribution function from Step 2. If they are not the same try increasing the horizon T and go back to Step 4.

Step 8: Compute the time path of excess demand for sex, and the path of prevalence's $\{\hat{\phi}_t^+\}_{t=\tau}^T$. If markets don't clear along the path, or $\{\hat{\phi}_t^+\}_{t=\tau}^T \neq \{\phi_t^+\}_{t=\tau}^T$, then update $\{p_t\}_{t=\tau}^T$ and $\{\phi_t^+\}_{t=\tau}^T$ return to Step 5.

Step 9: Record the first elements of the transition decision rules $(c_{\tau+1}, x_{\tau+1})$, prices $(p_{\tau+1})$, prevalence $(\phi_{\tau+1}^+)$ and joint distribution $(\Phi_{\tau+1})$.

Step 10: Stop if $\tau = T_0$

Step 11: Replace $\tau = \tau + 1$ and go back to Step 1.

Computing the recursive stationary equilibrium for Stage 1-2 (HIV Maturity)

***Algorithm No.5*: Computation of the recursive stationary equilibrium of the Pre-Epidemic stage:**

Same as Algorithm No.1

Computing the Non-stationary equilibrium for Stage 1-2 (HIV Maturity)

***Algorithm No.6*: Computation of the solution of the Maturity Stage:**
We are after a sequence of $\{\Phi_t\}_{t=T_0}^T (g = 1-2)$ that goes from the last period in Stage 0 (HIV Myopia $t = T_0$) to the recursive stationary equilibrium of Stage 1-2 of the epidemic (HIV Maturity).

Step 1: Choose a large number of transition periods $(T - T_0)$

Step 2: Simulate a sequence of $\{\rho_{e,t}\}_{t=T_0}^T$ by education group, following (29).

Step 3: Compute the recursive stationary equilibrium of Stage 1-2 of the epidemic. This stationary equilibrium is associated with $\lim_{t\to T} \rho_{e,t} = \rho$. That is, in the stationary equilibrium both education groups have completed learning of the true probability HIV infection risk as a function of sex.

Step 4: Guess a time path for the prices $\{p_t\}_{t=T_0}^{T}$ and prevalences $\{\phi_t^+\}_{t=T_0}^{T}$.

Step 5: Compute the equilibrium policy (and value) functions iterating backwards in time, $t = T-1, ..., T_0$.

Step 6: Simulate the evolution of the population distribution from $t = T_0$ to $t = T$ with the help of the optimal policy functions and the initial distribution $\Phi_{T_0}(g = 1)$.

Step 7: Compare the simulated distribution at T with the stationary distribution function from Step 3. If they are not the same try increasing the horizon T and go back to Step 2.

Step 8: Compute the time path of excess demand for sex, and the path of prevalence's $\{\hat{\phi}_t^+\}_{T_0}^{T}$. If markets don't clear along the path, or $\{\hat{\phi}_t^+\}_{T_0}^{T} \neq \{\phi_t^+\}_{T_0}^{T}$, then update $\{p_t\}_{T_0}^{T}$ and $\{\phi_t^+\}_{T_0}^{T}$ return to Step 5.

Computing the recursive stationary equilibrium for Stage 3-4 (ARV Era)

***Algorithm No.7*: Computation of the recursive stationary equilibrium of the ARV stage:**

Same as Algorithm No.1

Computing the Non-stationary equilibrium for Stage 3-4 (ARV Era)

***Algorithm No.8*: Computation of the solution of the ARV Stage:**
After T_1 periods in Stage 1-2 (HIV Maturity), ARV's are introduced unexpectedly. We are after a sequence of $\{\Phi_t\}_{t=T_1}^{T}(g = 3-4)$ that go from T_1 to the recursive stationary equilibrium of Stage 3-4 of the epidemic (ARV Era).

Step 1: Choose a large number of transition periods $(T - T_1)$

Step 2: Compute the recursive stationary equilibrium of Stage 3-4 of the epidemic. This stationary equilibrium is associated with $\lim_{t\to\infty} \eta_t = \widetilde{\eta} = 1$. That is, there is full coverage for all HIV infected individuals.

Step 3: Guess a time path for the prices $\{p_t\}_{t=T_1}^{T}$ and prevalences $\{\phi^+\}_{t=T_1}^{T}$.

Step 4: Given the sequences for ARV coverage by education group $\{\eta_t(e)\}_{t=T_1}^{T}$, compute the equilibrium policy (and value) functions iterating backwards in time, $t = T-1, ..., T_1$.

Step 5: Simulate the evolution of the population distribution from $t = T_1$ to $t = T$ with the help of the optimal policy functions and the initial distribution $\Phi_{t=T_1}(g = 2)$.

Step 6: Compare the simulated distribution at T with the stationary distribution function from Step 2. If they are not the same try increasing the horizon T and go back to Step 3.

Step 7: Compute the time path of excess demand for sex. and the path of prevalence's $\{\hat{\phi}_t^+\}_{t=T_1}^T$. If markets don't clear along the path, or $\{\hat{\phi}_t^+\}_{t=T_1}^T \neq \{\phi_t^+\}_{t=T_1}^T$, then update $\{p_t\}_{t=T_1}^T$ and $\{\phi_t^+\}_{t=T_1}^T$ return to Step 4.