

NBER WORKING PAPER SERIES

INTERGENERATIONAL MOBILITY IN AFRICA

Alberto Alesina
Sebastian Hohmann
Stelios Michalopoulos
Elias Papaioannou

Working Paper 25534
<http://www.nber.org/papers/w25534>

NATIONAL BUREAU OF ECONOMIC RESEARCH

1050 Massachusetts Avenue
Cambridge, MA 02138

February 2019, Revised August 2020

We are grateful to three anonymous referees, and the editor Fabrizio Zilibotti, for extremely useful comments. Also, we thank our discussants in various conferences Oriana Bandiera, Sara Lowes, and Nathaniel Hendren as well as to Francesco Caselli, Raj Chetty, and Antonio Ciccone for detailed comments. We are thankful for useful discussions and feedback to James Fenske, Larry Katz, Nathan Nunn, David Weil, Elisa Cavatorta, Shaun Hargreaves, Jim Robinson, Sam Asher, Paul Novosad, Charlie Rafkin, and Yining Geng. We also got useful feedback from conference and seminar participants at the University of Zurich, UPF, Monash University, Warwick, Brown, Harvard, King's College, Tufts, Copenhagen, Sheffield, Williams, the NBER Summer Institute, and the PSE Conference on Culture, Institutions, and Economic Prosperity. We also thank Remi Jedwab, Adam Storeygard, Julia Cagé, Valeria Rueda and Nathan Nunn for sharing their data. Papaioannou gratefully acknowledges financial support from the European Research Council (Consolidator Grant ORDINARY), LBS Wheeler Institute for Business and Development, and RAMD. All errors and omissions are our responsibility. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2019 by Alberto Alesina, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Intergenerational Mobility in Africa

Alberto Alesina, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou

NBER Working Paper No. 25534

February 2019, Revised August 2020

JEL No. I24,J62,P16

ABSTRACT

We examine intergenerational mobility (IM) in educational attainment in Africa since independence using census data. First, we map IM across 27 countries and more than 2,800 regions, documenting wide cross-country and especially within-country heterogeneity. Inertia looms large as differences in the literacy of the old generation explain about half of the observed spatial disparities in IM. The rural-urban divide is substantial. Though conspicuous in some countries, there is no evidence of systematic gender gaps in IM. Second, we characterize the geography of IM, finding that colonial investments in railroads and Christian missions, as well as proximity to capitals and the coastline are the strongest correlates. Third, we ask whether the regional differences in mobility reflect spatial sorting or their independent role. To isolate the two, we focus on children whose families moved when they were young. Comparing siblings, looking at moves triggered by displacement shocks, and using historical migrations to predict moving-families' destinations, we establish that, while selection is considerable, regional exposure effects are at play. An extra year spent in a high-mobility region before the age of 12 (and after 5) significantly raises the likelihood for children of uneducated parents to complete primary school. Overall, the evidence suggests that geographic and historical factors laid the seeds for spatial disparities in IM that are cemented by sorting and the independent impact of regions.

Alberto Alesina
Department of Economics
Harvard University
Littauer Center 210
Cambridge, MA 02138
and IGER
N/A user is deceased

Sebastian Hohmann
Stockholm School of Economics SITE
Sveavagen 65
Box 6501
SE 113 83
Stockholm
Sweden
sebas.hohmann@gmail.com

Stelios Michalopoulos
Brown University
Department of Economics
64 Waterman Street
Providence, RI 02912
and CEPR
and also NBER
smichalo@brown.edu

Elias Papaioannou
London Business School
Regent's Park
Sussex Place
London NW1 4SA
United Kingdom
and CEPR
papaioannou.elias@gmail.com

An online appendix is available at <http://www.nber.org/data-appendix/w25534>

1 Introduction

There is rising optimism about Africa, a continent with 1.2 billion opportunities, as the Economist (2016) touted not long ago. The formerly “hopeless continent” is gradually becoming the “hopeful” one (Economist (2000, 2011)). Educational attainment is rising, health is improving, and the income of many Africans is growing. Some even speak of an African “growth miracle” (Young (2012)). However, anecdotal evidence indicates widespread inequalities, uneven progress, and poverty traps, suggesting that the “miracle” may not be for all. A comprehensive assessment is lacking.

We take the first step toward mapping, exploring, and explaining intergenerational mobility across the continent. We look at educational attainment using census data covering more than 16 million individuals across 27 African countries and 2,846 regions. Reconstructing the joint distribution of parental and offspring education since the 1960s, when most of Africa becomes independent, allows us to shed light on a variety of questions. Where is the land of educational opportunity? Are differences in intergenerational mobility across countries and regions small, moderate, or wide? How large are gender disparities? How big is the rural-urban gap? Which elements of a region’s history and geography correlate with educational mobility? Do regions matter for mobility or do districts with higher mobility attract families more eager to climb the social ladder?

1.1 Results Preview

In the first part of the paper, we compile new country and regional-level measures of educational opportunity. As recent works on intergenerational mobility in income (e.g., Chetty et al. (2017)) and education (Card, Domnisoru, and Taylor (2018)), we construct measures of absolute upward intergenerational mobility (IM) defined as the likelihood that children born to parents that have not completed primary schooling manage to do so. Similarly, we map absolute downward mobility, defined as the likelihood that the offspring of parents with completed primary education fail to do so. To account for “selection on cohabitation”, we focus on ages between 14 and 18, as in this age range children have largely finished primary school and still reside with parents or older relatives.

We document large cross-country differences in upward and downward mobility. The likelihood that children born to parents with no education complete primary schooling exceeds 70% in South Africa and Botswana; the corresponding statistic in Sudan, Ethiopia, Mozambique, Burkina Faso, Guinea, and Malawi hovers below 20%. Most importantly, there is substantial within-country variation. In Kenya, a country with a close-to-average upward IM of 50%, the likelihood that children of illiterate parents will complete primary education ranges from 5% (in the Turkana region in the Northwest) to 85% (in Westlands in Nairobi). Upward IM is higher in urban as compared to rural areas. While there is a gender gap in educational levels, intergenerational mobility is, on average, similar for boys and girls, though there is a non-negligible gender gap in the Sahel and North Africa. Spatial disparities in mobility exhibit inertia: Upward IM is higher in countries and regions with higher literacy among the old. Variation in the latter accounts for roughly half of the observed IM variability. Downward mobility is also linked to the literacy of the old generation, but the association is weaker.

In the second part of the paper, we characterize the geography of IM in Africa by looking at geographical and historical variables that have been linked to regional development. Upward IM is higher and downward IM is lower in regions close to the coast and the capital, with rugged terrains and low malaria. Among the historical legacies, colonial transportation investments and missionary activity are the strongest correlates of mobility. These correlations are present when we exploit within-province variation and when we estimate LASSO to account for multicollinearity and measurement error. While these associations do not identify causal effects, they suggest how historical contingencies, related to colonization and geography, have influenced not only initial conditions (the literacy of the old generation) but also the trajectories of regional economies.

The observed differences in regional IM may be the result of two forces. On the one hand, regions may exert a causal impact on mobility, for example, providing higher-quality infrastructure, more and better schools. On the other hand, there may be sorting, as families with higher ability and/or valuation of education move to areas with better opportunities. In the third part, we assess the relative magnitudes of these two factors employing the approach of Chetty and Hendren (2018a). The methodology exploits differences in the age at which children of migrant households move to distinguish “*selection*” from “*regional childhood exposure effects*.” Both forces are at play. Selection is present; families’ sorting into better (worse) locations correlates strongly with child attainment. The analysis also uncovers sizable “*regional exposure effects*” both for boys and girls. An additional year in the higher mobility region before the age of 12, and especially between 5 – 11, increases the likelihood that children of households without any education manage to complete primary schooling.

To advance on the identification of regional exposure effects, we conduct three exercises, separately and jointly. First, we explore whether the educational attainment of siblings whose family moved is proportional to their age difference interacted with differences in mobility between the permanent residents in origin and destination districts. The regional childhood exposure estimates from the household-fixed-effects specifications are similar to the baseline ones. Second, we look at moves taking place in periods of abnormal outflows, as these likely reflect displacement shocks exogenous to households. We continue finding considerable regional exposure effects for moving children in the critical-for-primary schooling age (5 – 11) and somewhat smaller before 5. Third, we use historical migration to project -and account for- households’ endogenous destination choice. The regional childhood exposure estimates remain significant.

Overall, the analysis suggests that the vast spatial differences in mobility reflect both sorting and regional exposure effects. The uncovered inertia, coupled with the strong association between mobility (and old’s literacy) with historical and geographic traits, suggests that these features have shaped regional dynamics post-independence.

1.2 Related Literature

Our work blends two strands of literature that have, thus far, moved in parallel. The first is the growing research studying intergenerational mobility (see Solon (1999) and

Black and Devereux (2011) for reviews).¹ Card, Domnisoru, and Taylor (2018) use the US population census of 1940 to map absolute educational mobility looking at children residing with at least one parent. They document rising mobility during the first half of the 20th century, which differs across race and states.² Chetty et al. (2014) provide a mapping of IM in income across US counties and explore its correlates. Chetty and Hendren (2018a,b) use matched parents-children administrative tax records of moving families to isolate the effect of neighborhood exposure on income IM from sorting. Our work relates to Asher, Novosad, and Rafkin (2020) and Geng (2018), who also map and study educational mobility across Indian and Chinese regions, respectively. In parallel work, the World Bank compiles measures in intergenerational mobility in education and income for many countries using survey data (Narayan et al. (2018)). Our main contribution to this research is to compile new statistics and characterize the educational mobility for many African countries and regions, distinguishing also between gender and rural-urban residence. Moreover, we estimate regions' independent influence on mobility, showing at the same time that bidirectional sorting (from higher to lower opportunity regions and vice versa) is considerable.

The second strand is the research on the origins of African development that provides compelling evidence of historical continuity as well as instances of rupture in the evolution of the economy and polity (see Michalopoulos and Papaioannou (2020) for a review). An open question is whether the correlation between deeply rooted factors and current outcomes reflects the one-time effect of the former on initial conditions or if historical shocks have altered the transmission of opportunity across generations. By building data on IM across African regions and exploring its correlates, we begin answering such questions. Moreover, by isolating the role of regions on mobility from sorting, we start unbundling the mechanisms linking geography-history to contemporary development.

Structure In Section 2, we present the census data on educational attainment and detail the construction of the intergenerational mobility measures. Section 3 describes IM across African countries and regions. Section 4 explores the geographic, historical, and at-independence correlates of educational mobility. In Section 5, we exploit differences in ages-at-move among migrant children to isolate regional childhood exposure effects from sorting. In Section 6, we summarize and discuss avenues for future research.

2 Data and Methods

2.1 Why Education?

We focus on education for several reasons. First, income data are available for a tiny share of the African population and a handful of countries. For instance, Alvaredo et al. (2017) report that for Ghana, Kenya, Tanzania, Nigeria, and Uganda, income data encompass

¹Early studies on intergenerational mobility in education include Bowles (1972), Blake (1985), and Spady (1967). Hertz et al. (2008) estimate country-level IM coefficients across 42 countries. Hilger (2017) studies trends in educational IM in the United States over the 20th century, while Chetty et al. (2017) and Davis and Mazumder (2020) study the dynamics of absolute IM in income in the US.

²A strand of the US-focused literature looks at racial differences in mobility (e.g., Chetty et al. (2020b), Davis and Mazumder (2018), Derenoncourt (2018)). These studies relate to our companion work Alesina et al. (2020b,a), where we explore ethnic and religious differences in educational mobility across Africa.

less than 1% of the adult population, while for most African countries tax records do not exist. Moreover, consumption data are noisy and cover small samples. In contrast, education is available at a fine geographic resolution. Second, measurement error in educational attainment is a lesser concern compared to that of reported income, wealth, or consumption. Third, education is useful in mapping intergenerational mobility, as people tend to complete primary schooling, which is the key educational achievement across most of Africa, by the age of 12–14. Hence, unlike lifetime earnings, the analysis can start when individuals are early in the life cycle. Fourth, parental investment in children’s education is at the heart of theoretical work in intergenerational linkages (e.g., Becker and Tomes (1979), Loury (1981)). Fifth, a voluminous research in labor economics shows that education causally affects lifetime income (e.g., Card (1999)). Individual returns to schooling are sizable in low-income (African) countries.³ Sixth, in the Appendix (section C.2), using geo-referenced Demographic and Health Surveys (DHS) and Afrobarometer Surveys, we present evidence of a strong correlation between educational attainment and various proxies of well-being in Africa, including living conditions, child mortality, attitudes toward domestic violence, political and civic engagement.

2.2 Sample

2.2.1 Countries & Regions

We use individual records, retrieved from 69⁴ national censuses from 27 countries: Benin, Botswana, Burkina Faso, Cameroon, Egypt, Ethiopia, Ghana, Guinea, Kenya, Lesotho, Liberia, Malawi, Mali, Morocco, Mozambique, Nigeria, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, South Sudan, Tanzania, Togo, Uganda, Zambia, and Zimbabwe. We obtain the data from IPUMS (Integrated Public Use Microdata Series) International, hosted at the University of Minnesota Population Centre, that reports harmonized representative samples, typically 10%.⁵ As of 2015, the sample countries were home to about 850 million people, representing around 75 percent of Africa’s population and GDP. IPUMS also reports residence, allowing us to assign individuals to “coarse” and “fine” administrative units. Our sample spans 367 provinces (admin-1) and 2,846 districts (admin-2 or 3 units) of a mean (median) size of 5206 (1578) sqm.⁶

³Most studies suggest higher returns to education in low income countries, as compared to the “consensus” estimate of 6.5% – 8.5% in high income countries (e.g., Psacharopoulos (1994), Caselli, Ponticelli, and Rossi (2014)). Young (2012) estimates Mincerian returns of about 11.3% (OLS) to 13.9% (2SLS) across 14 Sub-Saharan African countries using DHS data, higher than in 11 non-SSA low income countries [range of 8.7% (OLS) - 10.4% (2SLS)]. Montenegro and Patrinos (2014) estimate Mincerian returns of about 12.4% in Africa, compared to 9.7% for the rest of the world. Four of the top-5 countries are in Africa. Psacharopoulos and Patrinos (2004) document a mean increase in wages for those with completed primary of 37.6% across 15 Sub-Saharan African countries in the 1980s and 1990s, as compared to 26.5% for secondary and 27.8% for tertiary.

⁴We start from 74 censuses. We discard Burkina Faso (1985), Kenya (1979), and Liberia (1974), as they lack identifiers to match children to older relatives. We also remove Togo (1960 and 1970), as they do not cover all regions.

⁵In Nigeria data come from household surveys conducted in consecutive years between 2006 and 2010. As the number of observations is small, we aggregate the survey waves and count them as one census-year.

⁶For Botswana, Lesotho, and Nigeria, IPUMS reports one level of administrative units. In Ghana after 1984, Burkina Faso in 1985, Ethiopia in 1984, Malawi in 1987, and South Africa after 1996, districts change, as administrative boundaries are redrawn. We have harmonized these countries’ boundaries.

2.2.2 Education

IPUMS records education for around 93 million individuals. Dropping those younger than 14 to allow for primary school completion leaves about 66.8 million observations.⁷ Appendix Figure A.1 portrays the evolution of the pan-African distribution of educational attainment across cohorts. Education rises, mostly reflecting increasing completion of primary schooling. The share of Africans with tertiary education is minuscule even for the 1980s-born, while secondary education has increased modestly.⁸ We need to observe education for children and at least one individual of the immediately older generation. This requirement brings the sample to 25.8 million. Appendix table B.1 gives details on sample construction.

For a first look at the data, we construct 4×4 attainment transition matrices for individuals older than 25 years. Figure 1 (a) shows the Africa-wide transition matrix using all censuses, while figures 1 (b) and (c) zoom in Mozambique and Tanzania, respectively. The vertical axis indicates the likelihood that the child has the respective education, conditional on the older generation attainment, depicted on the horizontal axis. 81.5% of the “old” generation across the continent has not completed primary schooling. 19% of African children, whose parents have not completed primary schooling, manage to do; 9.5% finish high-school, and 2.5% get a college degree. The figure also illustrates the sharp differences between the two Eastern African countries. In Tanzania, 47% of children whose parents have not finished primary school manages to do so; in Mozambique, the corresponding share is 12%.

2.3 Methodology

We construct measures of *absolute* IM that reflect the likelihood that children complete a strictly higher or lower education level than members of the immediately previous generation in the household (parents and/or extended family members, such as aunts and uncles). For the education of the “old”, we take the average attainment of individuals one generation older in the household, rounded to the nearest integer (results are similar if we take the minimum or maximum).⁹ As the relevant dimension for Africa during this period regards the completion of primary schooling, we focus on this aspect.¹⁰

⁷We validated the IPUMS data across country-cohorts with the Barro and Lee (2013) statistics and at the regional level using DHS; correlations exceed 0.9 (Appendix Section C)

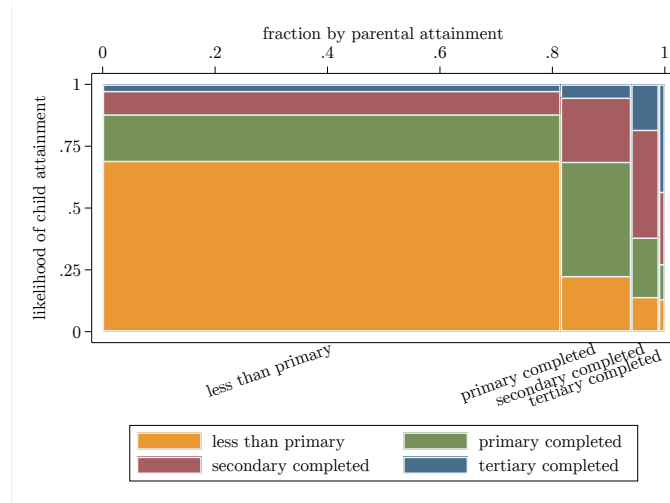
⁸There are four attainment categories: (i) no schooling and less than completed primary; (ii) completed primary (and some secondary); (iii) completed secondary (and some tertiary); and (iv) completed tertiary (and higher). We use attainment, rather than years of schooling, for many reasons. First, the attainment data have wider coverage than years of schooling. In the raw IPUMS data, there are about 25.5 million records with attainment, but without years of schooling. The latter is missing altogether for four countries and several censuses. Second, there is likely less noise on completion data as compared to schooling years, which are often inferred from the former. Third, looking at children, whose parents have not completed primary schooling, allows for a common across countries, simple to grasp baseline.

⁹Some studies use data that match children to either mothers or fathers (e.g., Asher, Novosad, and Rafkin (2020)). Others, like we do, take the average (e.g., Hilger (2017)), while some take the highest value (e.g., Geng (2018)). Taking the mean, maximizes coverage (see also Davis and Mazumder (2020)).

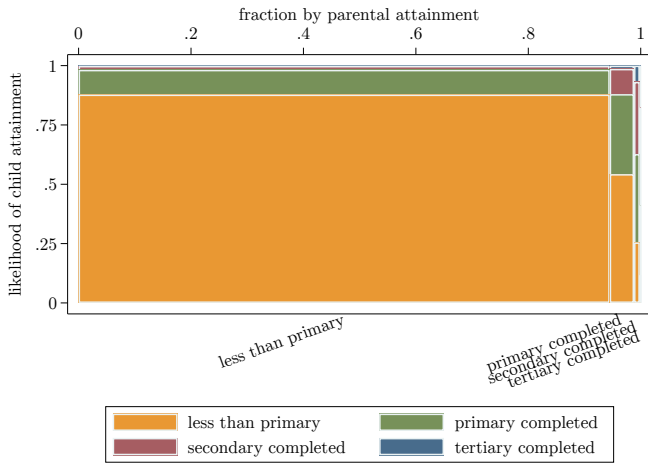
¹⁰The intergenerational mobility literature has employed various measures (see Black and Devereux (2011)). Many studies focus on (one minus) the intergenerational coefficient obtained from a regression of children on parental schooling (e.g., Hertz et al. (2008)); others work with rank-rank correlation coefficients and intergenerational rank movements (e.g., Asher, Novosad, and Rafkin (2020), Geng (2018), Chetty et al. (2014)). While rank-based measures isolate the relative movement of children in the distribution compared to the older generation from the overall increase, they may be sensitive to measurement error (see

Figure 1: Educational Attainment Transition Matrices

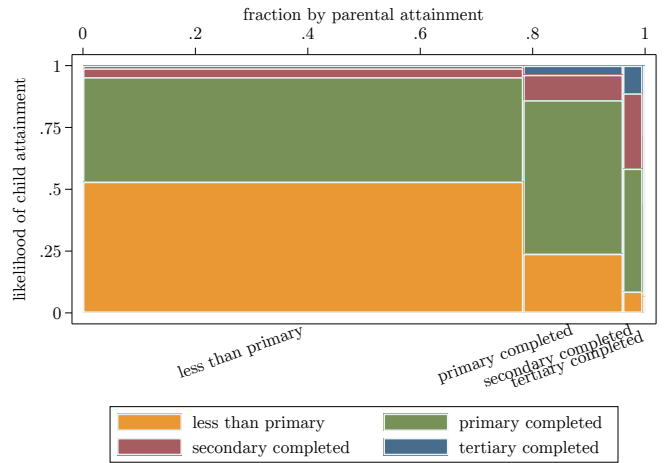
(a) Africa, 27 countries, 69 censuses



(b) Mozambique, 1997, 2007 census



(c) Tanzania, 1988, 2002, 2012 census



The figure shows the transition matrices for four educational attainment categories for Africa, Mozambique and Tanzania. The sample consists of individuals aged 25 and older, co-residing with at least one individual of an older generation.

To construct absolute IM measures, we first define the following indicator variables:

- lit_par_{ibct} equals 1 if the parent of individual i born in birth-decade b in country c and observed in census-year t is literate and zero otherwise. We label “illiterate” those who have not completed primary education and “literate” those who have.
- IM_up_{ibct} equals 1 if a child i born to illiterate parents in birth-decade b in country c and observed in census-year t is literate and zero otherwise.
- IM_down_{ibct} equals 1 if a child i born to literate parents in birth-decade b in country c and observed in census-year t is illiterate and zero otherwise.

Mogstad et al. (2020)). Other studies (e.g., Card, Domnisoru, and Taylor (2018), Davis and Mazumder (2020) and Chetty et al. (2017)) focus, as we do, on absolute transition likelihoods. Gottschalk and Spolaore (2002) provide a theoretical exploration of different mobility measures. The absolute IM measures correlate strongly with the IM coefficient across both countries and regions. The correlation of the absolute IM statistics with the intergenerational correlation is though small.

Then, we estimate the following specifications, pooling observations across all censuses and countries:

$$\text{lit_par}_{ibct} = \alpha_c^o + [\gamma_b^o + \delta_b^y + \theta_t] + \epsilon_{ict} \quad (1)$$

$$\text{IM_up/down}_{ibct} = \alpha_c^y + [\gamma_b^o + \delta_b^y + \theta_t] + \epsilon_{ict}, \quad (2)$$

For parental literacy (equation (1)), we compute means among all individuals for whom we observe their parents' (older generation relatives) attainment, netting birth-decade fixed effects for the “young” (δ_b^y) and the “old” (γ_b^o) and census-year fixed effects (θ_t). For upward IM, we estimate equation (2) for children whose parents have not completed primary education; thus the country fixed effects ($\hat{\alpha}_c^y$) reflect the conditional likelihood that children of illiterate parents become literate, netting cohort and census effects. For downward IM, we estimate (2) for children whose parents have completed at least primary; so $\hat{\alpha}_c^y$ measure the conditional likelihood that children of literate parents do not complete primary schooling netting census-year and cohort effects.

For the regional analysis, we run similar specifications at the district level, country-by-country, and extract the demeaned literacy of the old generation, upward IM, and downward IM (conditioning on cohort and census fixed-effects).

$$\text{lit_par}_{ibcrt} = \alpha_r^o + [\gamma_b^o + \delta_b^o + \theta_t] + \epsilon_{ibcrt} \quad (3)$$

$$\text{IM_up/down}_{ibcrt} = \alpha_r^y + [\gamma_b^o + \delta_b^o + \theta_t] + \epsilon_{ibcrt}. \quad (4)$$

2.4 Cohabitation Selection

Estimating the IM of individuals who reside with at least one older family member (usually a biological parent) raises cohabitation-selection concerns, as the transmission of education may differ between children living with older family member(s) and those that do not. This issue is less pressing for young children, as almost all of them cohabit with their parents. The younger the child, however, the higher the risk of misclassifying her attainment as “less-than-primary” when in fact she would complete primary education a few years after we observe her in the census. Hence, following Card, Domnisoru, and Taylor (2018), we focus on “children” aged 14 – 18 years, as by then primary education is mostly completed and cohabitation rates are still high (see also Hilger (2017)).

We use census information on the “relationship to household head” to recover the “old” generation and take the average of their educational attainment. Appendix Section D provides details, discussing also how we deal with heterogeneity in family structure (e.g., nuclear families, presence of young wives). The Appendix reports statistics for each census, as their detail differ. On average, cohabitation with *any relatives* for children aged 14 – 18 is around 94.5%. However, the “*relationship to household head*” variable is coarsely documented in some censuses.¹¹ To maximize coverage and avoid misclassifying

¹¹An extreme example is the Togo 2010 census, which classified 92.9% of individuals 14 – 18 years as cohabitating with some relative. Due to the census' sparse categorization of the relationship to family head, about half of the children are classified as residing with “*other relatives*.” Some censuses distinguish between biological, adopted, and step-children (e.g., Nigeria, South Africa, Zambia), but most do not.

coresidence with older family member(s) due to census coarseness, we assign “*other relatives (not elsewhere classified)*” to the “old” generation if they are at least 15 and less than 40 years older than the child. [This imputation affects about 10% of the sample and does not affect the results.]

For individuals aged between 14 and 18 years, the coresidence rate across all censuses with an older generation relative is 84% (see appendix table D.2). Cohabitation rates with an older family member exceed 90% in 11 censuses; it is between 85% – 90% for 15 and between 80% – 85% for 17. The lowest coresidence rate is recorded in Kenya in 1969 (63.3%), in Malawi in 1987 (68.9%), and in Botswana in 1991 and 2011 (around 70%). As a reference point, Card, Domnisoru, and Taylor (2018) report coresidence rates for African Americans and whites in the US 1940 census of about 78% and 89%, respectively.

We also work with individuals aged 14-25, as this increases the sample considerably, including also high-school and college graduates, while cohabitation is still reasonably high (around 70%). The Appendix (Section D) gives details and also reports the distribution of district-level cohabitation rates; the mean (median) is 82% (82.5%). Cohabitation rates have slightly risen, though this most likely reflects improvements in census details.

3 Intergenerational Mobility across Countries and Regions

3.1 IM across African Countries

3.1.1 Baseline Measures

Table 1 shows simple (unconditional) country-level estimates of intergenerational mobility (columns (1)-(4)) alongside the number of children (young) for the 14 – 18 and the 14 – 25 sample. (The series are strongly correlated, $\rho > .97$). On average, less than forty percent of children of illiterate parents have managed to complete primary education. Downward IM is considerable, as approximately one out of four children born to literate parents does not complete primary education.

The pan-African mean masks sizable variation. The likelihood that children of illiterate parents will complete at least primary education ranges from an abysmal 4% in South Sudan and 11% in Mozambique to 80% in South Africa and 70% in Botswana. The lowest upward IM is in the Sahel (Sudan, Burkina Faso and to a lesser extent Mali and Senegal) and the highest in Southern Africa (Botswana, Zambia, Zimbabwe, and South Africa) with Western and Eastern African countries in the middle. Downward mobility is negatively correlated with upward mobility. Downward IM is the highest in countries plagued by long-lasting conflicts, such as Rwanda (0.47), Liberia (0.54), Mozambique (0.51), and South Sudan (0.77). Downward IM is below 10 percent in more stable ones like Botswana, South Africa, Egypt, and Nigeria. The uncovered cross-country heterogeneity in absolute IM across Africa is considerably larger than the cross-Indian state and cross-Chinese province variability in relative IM documented by Asher, Novosad, and Rafkin (2020) and Geng (2018), respectively.¹²

¹²Geng (2018) documents a province range in IM rank-rank coefficients of 0.25 to 0.5 in the 2000 Chinese Census. The range across (340) prefectures is between -0.033 to 0.661 . Asher, Novosad, and Rafkin (2020) estimate a range of relative educational mobility of 0.17 to 0.72 across 124 Indian districts and 0.26 to 0.60 across 25 states. Yet, as our statistics reflect absolute rather than relative changes of children’s

Table 1: Country-Level Estimates of Intergenerational Mobility (IM)

mobility / N age range	census years	(1)	(2)	(3)	(4)	(5)	(6)
		upward 14-18	upward 14-25	downward 14-18	downward 14-25	N with e_0 obs 14-18	N with e_0 obs 14-25
South Africa	1996, 2001, 2007, 2011	0.791	0.814	0.068	0.049	1,047,243	1,944,362
Botswana	1981, 1991, 2001, 2011	0.704	0.716	0.069	0.058	44,516	76,211
Zimbabwe	2012	0.664	0.738	0.146	0.108	49,855	79,290
Egypt	1986, 1996, 2006	0.637	0.628	0.071	0.066	2,128,269	4,056,814
Nigeria	2006, 2007, 2008, 2009, 2010	0.63	0.65	0.084	0.074	38,885	63,868
Tanzania	1988, 2002, 2012	0.595	0.636	0.177	0.151	860,096	1,358,638
Ghana	1984, 2000, 2010	0.566	0.556	0.159	0.142	489,957	845,090
Togo	2010	0.51	0.526	0.19	0.179	46,958	83,442
Cameroon	1976, 1987, 2005	0.509	0.506	0.117	0.115	270,300	443,222
Zambia	1990, 2000, 2010	0.486	0.507	0.2	0.182	307,043	484,973
Kenya	1969, 1989, 1999, 2009	0.454	0.523	0.219	0.169	624,501	1,016,810
Lesotho	1996, 2006	0.437	0.496	0.289	0.231	38,310	71,965
Morocco	1982, 1994, 2004	0.414	0.393	0.107	0.122	397,451	785,159
Benin	1979, 1992, 2002, 2013	0.376	0.354	0.232	0.231	192,949	326,478
Uganda	1991, 2002	0.358	0.393	0.311	0.277	345,215	518,395
Rwanda	1991, 2002, 2012	0.292	0.35	0.472	0.383	237,006	388,219
Senegal	1988, 2002	0.255	0.256	0.243	0.234	158,517	283,080
Sierra Leone	2004	0.248	0.245	0.368	0.35	42,905	72,534
Liberia	2008	0.221	0.297	0.538	0.418	31,437	55,981
Mali	1987, 1998, 2009	0.205	0.197	0.262	0.27	267,300	433,470
Guinea	1983, 1996	0.193	0.179	0.402	0.403	84,865	144,991
Burkina Faso	1996, 2006	0.184	0.189	0.267	0.253	201,788	294,456
Malawi	1987, 1998, 2008	0.155	0.225	0.48	0.384	246,463	383,502
Ethiopia	1984, 1994, 2007	0.129	0.152	0.302	0.273	851,496	1,300,687
Sudan	2008	0.119	0.174	0.394	0.274	466,630	799,231
Mozambique	1997, 2007	0.111	0.158	0.512	0.419	267,367	419,569
South Sudan	2008	0.041	0.07	0.767	0.646	48,071	83,835
mean / total		0.381	0.405	0.276	0.239	9,785,393	16,814,272

Columns (1) and (2) give upward-IM estimates. They reflect the likelihood that children, aged 14-18 and 14-25, whose parents have not completed primary schooling to complete at least primary education. Columns (3) and (4) give downward-IM estimates. They reflect the likelihood that children, aged 14-18 and 14-25, whose parents have completed primary schooling or higher fail to complete primary education. Columns (5) and (6) give the number of observations (children whose parental education is reported in the censuses). Countries are sorted from the highest to the lowest level of upward IM in the 14-18 sample (column (1)). “Mean” gives the unweighted average of the 27 country-estimates.

Given heterogeneity in family structures across the continent, we estimated different IM statistics for children co-residing with biological parents, other older generation relatives, and both. Appendix E.1 reports the cross-country measures. Upward IM is somewhat higher and downward IM lower for children co-residing with biological parents. However, the various measures are strongly correlated (0.95) and the country rankings not much affected by family structure.

3.1.2 Rural-Urban Residence

We compiled IM separately for rural and urban households. Appendix Table E.2 reports the statistics across countries. The correlation between rural and urban IM is 0.85 for both the upward and downward measures. Setting aside South Sudan, an outlier, upward IM in urban places ranges from 0.21 in Mozambique to around 0.85 in Zimbabwe and South Africa (mean 0.53 and st. dev. 0.2). The variability in rural upward IM relative to the mean is wider (mean 0.33 and st. dev. 0.22), hovering around 0.06 in Mozambique, Ethiopia, South and North Sudan but exceeding 0.6 in Nigeria, Egypt, Zimbabwe, Botswana, and South Africa. Overall, the rural-urban gap in mobility is the highest in poor

position in the educational distribution, the estimates’ ranges are not directly comparable. The variability of educational mobility across the US is lower than the pan-African one illustrated here. Fletcher and Han (2018) report IM schooling coefficients ranging from 0.3 till 0.6 across US states (median 0.45) using survey data in 1982, 1992, and 2004. Hilger (2017) reports a coefficient of variation of around 0.3 for educational mobility across US states.

countries (Appendix Figure E.2 (b)). In Figure 2 we explore the evolution of rural-urban gaps. Upward IM is on average 18% higher for urban, as compared to rural households, for all cohorts and countries, but Egypt in the 1960s and 1970s. The rural-urban gap is the highest in countries with low levels of mobility and literacy. For example, there is a gap of about 40 percentage points between rural and urban places in Ethiopia and Burkina Faso; the rural-urban gap is below 10 percentage points in South Africa and Botswana.

Figure 2: Upward IM Urban-Rural Gap



The figure plots the difference in upward IM between individuals aged 14-18 residing in urban and rural locations by country and birth decade. The criteria for the rural-urban classification vary. In some countries, statistical agencies rely solely on population cutoffs, while others use localities' economic activity. In a few instances, the statistical codebook does not provide precise information. Rural-urban status is not reported for Morocco.

3.1.3 Gender

We also estimate IM separately for boys and girls. Appendix Table E.2 gives the country means. The correlation of the IM measures for boys and girls exceed .90 and, as such, the cross-country ranking is similar. Figure 3 shows the evolution of male-female differences in upward-IM. There is a gender gap for the 1960s cohorts (especially when we exclude Botswana) that disappears for the 1980s and the 1990s cohorts. To be sure, there are countries where boys fare much better than girls: the gender gap is salient in North Africa (Morocco and Egypt) and the Sahel (Senegal, Togo, Mali, and Ethiopia). However, girls born to illiterate parents in many Southern and Eastern African countries, like Lesotho, Botswana, Tanzania, and South Africa, enjoy a small edge in completing primary schooling over boys. Gender differences in mobility are not related to GDP per capita (Appendix Figure E.2 (a)).

Figure 3: Upward IM Male-Female Gap



The figure plots the difference (gap) in upward IM between male and female young individuals aged 14-18 by country and birth decade.

3.2 Mapping the African Land of Opportunity

3.2.1 Cross-Sectional Patterns

Figure 4 illustrates social mobility across the continent, mapping Africa’s land of opportunity. Panel (a) shows the distribution of absolute upward IM across (mostly admin-2) districts and Panel (b) plots absolute downward IM.

Figure 4: District-level Upward and Downward IM

(a) upward; brighter colors → higher ↗ IM (b) downward; brighter colors → higher ↘ IM

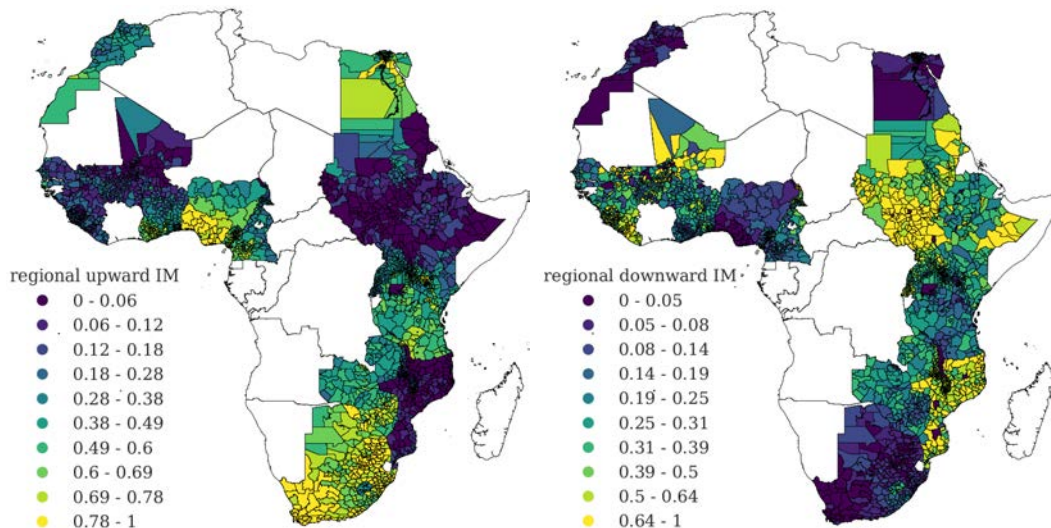


Table 2: Summary Statistics: District-Level Estimates of IM

country	districts	upward					downward				
		mean	median	stdev	min	max	mean	median	stdev	min	max
South Africa	216	0.788	0.802	0.07	0.565	0.897	0.081	0.073	0.038	0.018	0.217
Zimbabwe	88	0.734	0.746	0.136	0.428	1.0	0.161	0.161	0.086	0.02	0.462
Botswana	23	0.71	0.717	0.083	0.5	0.826	0.076	0.077	0.027	0.0	0.133
Nigeria	37	0.7	0.772	0.21	0.301	0.957	0.094	0.083	0.051	0.02	0.189
Egypt	236	0.673	0.683	0.108	0.392	0.914	0.076	0.068	0.039	0.013	0.242
Tanzania	113	0.615	0.619	0.096	0.391	0.836	0.182	0.181	0.068	0.056	0.369
Ghana	110	0.577	0.637	0.157	0.176	0.803	0.214	0.198	0.077	0.101	0.557
Cameroon	230	0.539	0.58	0.208	0.083	0.895	0.228	0.182	0.144	0.035	0.812
Kenya	173	0.504	0.523	0.189	0.054	0.872	0.261	0.269	0.108	0.041	0.586
Togo	37	0.493	0.506	0.13	0.235	0.687	0.252	0.242	0.093	0.092	0.543
Zambia	72	0.48	0.472	0.123	0.282	0.771	0.275	0.28	0.096	0.084	0.483
Morocco	59	0.429	0.422	0.14	0.158	0.702	0.144	0.13	0.066	0.062	0.375
Lesotho	10	0.421	0.423	0.057	0.318	0.497	0.328	0.337	0.06	0.235	0.419
Uganda	161	0.373	0.374	0.124	0.019	0.659	0.382	0.38	0.118	0.152	0.933
Benin	77	0.36	0.369	0.126	0.105	0.597	0.274	0.264	0.079	0.123	0.594
Rwanda	30	0.302	0.283	0.061	0.228	0.468	0.501	0.53	0.095	0.255	0.623
Senegal	34	0.253	0.183	0.151	0.078	0.592	0.316	0.282	0.132	0.149	0.793
Sierra Leone	107	0.219	0.17	0.143	0.032	0.667	0.563	0.581	0.189	0.142	1.0
Ethiopia	94	0.207	0.123	0.223	0.008	0.81	0.427	0.412	0.195	0.0	1.0
Malawi	227	0.195	0.16	0.111	0.049	0.562	0.533	0.551	0.122	0.179	0.8
Liberia	47	0.187	0.194	0.079	0.032	0.348	0.613	0.594	0.115	0.397	1.0
Guinea	34	0.156	0.151	0.072	0.06	0.432	0.441	0.44	0.098	0.25	0.68
Sudan	129	0.155	0.104	0.142	0.001	0.556	0.549	0.545	0.177	0.27	1.0
Burkina Faso	45	0.144	0.138	0.077	0.03	0.501	0.328	0.328	0.1	0.0	0.609
Mali	241	0.142	0.126	0.093	0.014	0.538	0.455	0.406	0.223	0.0	1.0
Mozambique	144	0.094	0.066	0.084	0.017	0.67	0.641	0.625	0.158	0.141	1.0
South Sudan	72	0.043	0.021	0.055	0.0	0.31	0.849	0.864	0.138	0.5	1.0
total	2,846	0.403	0.375	0.267	0.0	1.0	0.337	0.294	0.235	0.0	1.0

This table shows summary statistics for district level estimates of IM. “Total” shows the unweighted summary statistics across all districts.

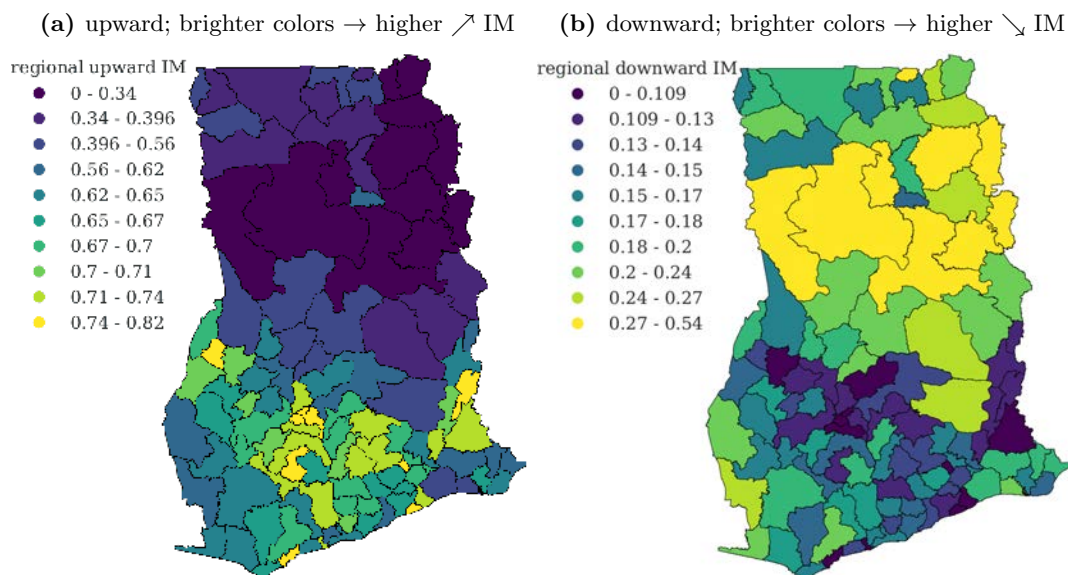
Table 2 reports summary statistics by country. The district-level (unweighted) average and median for upward (downward) IM across the 2,846 regions are 0.40 (0.34) and 0.375 (0.294), respectively, close to the cross-country values.¹³ As an example of the large within country variation, Figures 5 (a) and (b) portray upward and downward IM across 110 regions in Ghana. While average upward IM is 0.58, regional IM ranges from 0.18 to 0.82 with rates below 0.4 in the Northern regions and above 0.7 in the South. The mean downward mobility is 0.20, but it varies from 0.08 to 0.50. This north-south gradient mirrors both the country’s religious geography as well as colonial-era missionary activity and transportation investments, topics we return to below.

IM varies greatly across regions in many countries.¹⁴ In Burkina Faso, for example, the average upward-IM of 0.132 masks a regional range from 0.03 to 0.50. In Uganda, the upward-IM range is wider [0.015 – 0.69]. Spatial differences in IM are wider in countries with lower levels of mobility, a pattern that adds to the literature showing that underdevelopment moves in tandem with regional inequalities (see Kanbur and Venables (2005) for review).

¹³As in some countries, like Nigeria, districts are large, the map misses within-region spatial variation in IM that is likely non-negligible.

¹⁴For some districts mobility is either zero or one. These extremes reflect the small number of observations. The mean (median) district estimate is based on 1,936 (891) children (*st.dev* = 3,287). The patterns are similar if we restrict to regions with many observations.

Figure 5: Ghana: District-level Upward and Downward IM



3.3 Trends

In Table 3 we examine how average IM evolves for Africans born in the 1960's, 1970's, 1980's and 1990's.¹⁵ The within-country and within-district estimates show a mild increase in upward IM in the 1970s and 1980s. Upward IM is about 12 percentage points higher for the 1990s-born as compared to those born in the 1960s. Downward IM is falling over time, though at a weaker and more heterogeneous pace.¹⁶

Table 3: Evolution of IM across cohorts

	(1)	(2)	(3)	(4)
	IM up	IM down	IM up	IM down
1970s cohort	0.0549 (0.034)	-0.00812 (0.030)	0.0171 (0.028)	-0.00536 (0.049)
1980s cohort	0.0572 (0.040)	0.00713 (0.029)	0.0567 (0.047)	-0.0271 (0.049)
1990s cohort	0.117** (0.042)	-0.0295 (0.028)	0.124*** (0.041)	-0.0752* (0.043)
R2	0.908	0.855	0.919	0.710
within R2	0.221	0.064	0.228	0.038
N	71	71	7551	7147
level	country	country	district	district

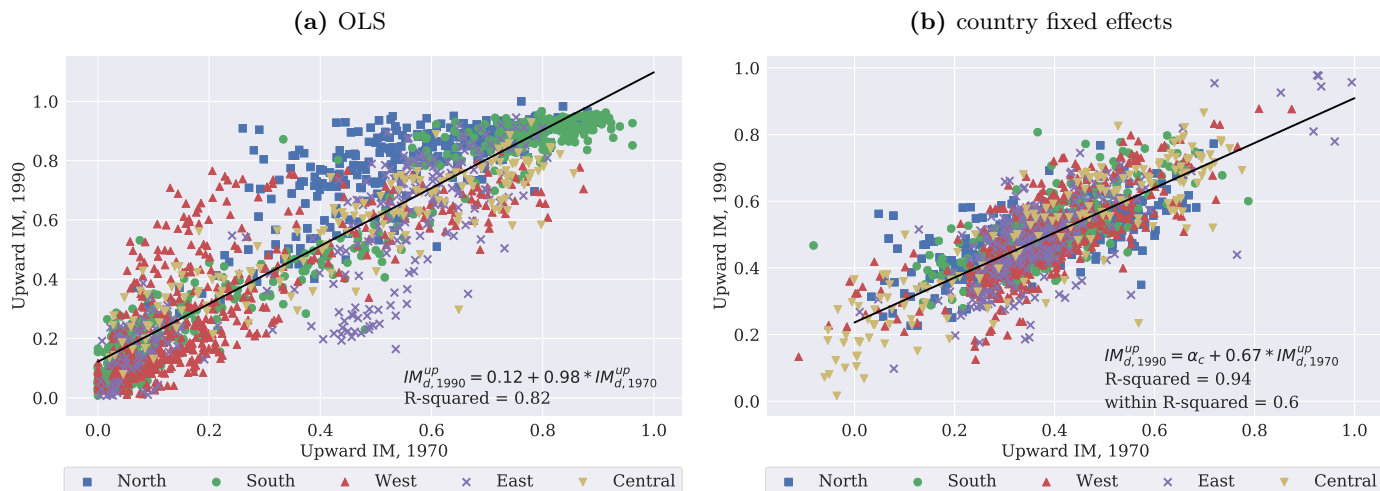
The table reports OLS estimates associating cohort-level upward IM (in columns (1) and (3)) and downward IM (in (2) and (4)) across countries (in (1)-(2)) and across regions (in (3)-(4)) with cohort indicators; the 1960s cohort serves as the omitted category. Specification (2) includes country constants (not reported) and specification (4) includes region constants (not reported). Standard errors clustered at the country-level are reported in parentheses. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

¹⁵Appendix E.3 portrays the distribution of regional IM across cohorts. The standard deviation of upward IM is roughly constant though the distribution becomes less skewed over time. The standard deviation of downward IM falls slightly.

¹⁶There is some relation of these patterns with the ones that Hilger (2017) presents for the US. He finds that the share of children with strictly higher educational attainment than their parents increased for the 1930s, 1940s, and 1950s born cohorts, but started falling after. The increase was acute for African Americans, though the decline applied to both whites and blacks.

Figure 6 illustrates the correlation of regional upward-IM for the 1990s and the 1970s cohorts. There is an almost one to one link with a strong fit. The slope decreases to .67 in the country fixed-effects specification.

Figure 6: District-level Upward IM over Time



The figures visualize the link between district-level upward IM for the 1990s to the 1970s cohorts. Panel (a) shows the simple linear regression fit; panel (b) shows the regression with country fixed effects fit. Dots are color-coded by African region following the classification of Nunn and Puga (2012).

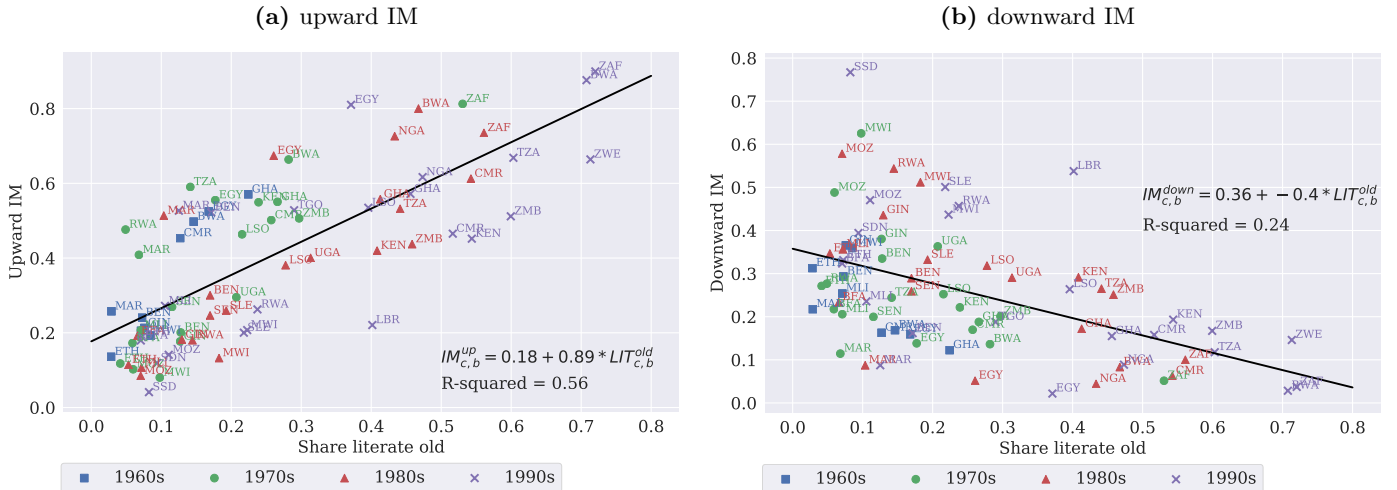
3.4 Literacy of the Old and IM

Motivated by evidence from the recent research agenda on intergenerational mobility (e.g., Chetty et al. (2020a)) showing that upward mobility is higher in regions with better outcomes (wealth, education, income) and research on African growth stressing poverty traps and slow convergence (e.g., Gunning and Collier (1999)), we examine the association between IM and literacy rates of the “old generation”. While these correlations do not have a causal interpretation, they allow us to explore inertia.

3.4.1 Cross-Country Patterns

Figure 7, panel (a), plots the relationship between country-level IM across cohorts and the literacy rate of the old generation of the respective cohort. A strong positive association emerges. In Ethiopia, Burkina Faso, Mozambique, North, and South Sudan, where for all cohorts the share of literate “old” is less than 20%, the likelihood that children from illiterate parents will complete primary school is below or close to 20%. The analogous statistic for Botswana and South Africa, where the old-cohorts’ literacy rate exceeds 50%, hovers around 70%. A one-percentage-point increase in the literacy of the old is associated with a .89 percentage points increase in upward IM; and variation in the former explains 56% of the cross-country-cohort variation in upward IM. Figure 7 Panel (b) uncovers a similar though attenuated relationship between the literacy of the “old” generation and downward IM. A one percentage point increase in the “old” generation’s literacy maps into a 0.4 decline in downward IM; the old generation’s literacy explains about a fourth of the variation in downward IM. Compared to upward IM, downward IM appears more sensitive to cohort-specific civil conflict (e.g., Sudan, Liberia, Sierra Leone).

Figure 7: Literacy of the Old Generation and Intergenerational Mobility across Countries



The figures plot upward-IM and downward-IM across country-birth-cohorts against the share of the “old” generation that has completed primary education. The figures also report the unweighted OLS regression fit.

3.4.2 Regional Patterns

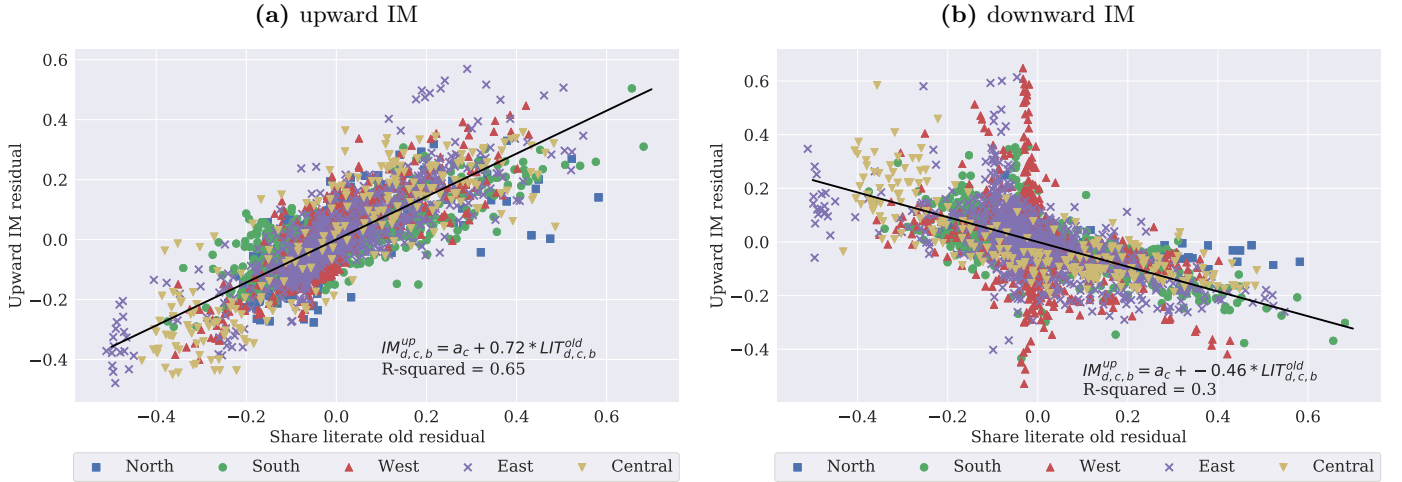
Figures 8 (a) and (b) plot the district-level association between upward and downward IM and mean literacy of the “old” generation, netting country-cohort and census effects. We observe a strong association between the literacy of the “old” and upward IM across African regions. Likewise, there is a negative -but less steep- correlation between downward IM and the literacy of the old. A 10 percentage points increase in the literacy of the “old” is associated with a roughly 7 percentage points increase in the likelihood that children of illiterate parents will complete primary and a 4.5 percentage points lower chance that kids of literate parents will fall below parental literacy. The estimates retain statistical significance and decline modestly when we replace the country constants with admin-1 fixed effects to account for relatively local features. This pattern is similar to Asher, Novosad, and Rafkin (2020) that a state’s/region’s mean education is the strongest correlate of upward educational mobility in India. Similarly, Güell et al. (2018) document a significantly positive correlation between IM in well-being and education across Italian regions.

Hence, disadvantaged (from non-educated) families children are more likely to complete primary school in regions with relatively higher literacy. Path dependence can reflect various mechanisms. First, poverty trap dynamics that are especially salient in subsistence agriculture rural Africa. Second, sunk costs in large-scale investments and infrastructure. Third, persistent spatial disparities in schools may be a contributing factor. Fourth, inertia may result from internal migration and spatial sorting. Fifth, the estimates may partly reflect human capital externalities (as Wantchekon (2019) shows in Benin).

3.4.3 Heterogeneity

We explored heterogeneity in the old’s literacy-IM association in terms of the child’s gender and the rural-urban household residence. The analysis, reported for brevity in Appendix E.4, reveals two noteworthy patterns. First, the association between IM and the share of

Figure 8: Literacy of the Old Generation and IM at the District Level



The figures plot district-level upward IM (left panel) and downward IM (right panel) against the share of the “old” generation with completed primary education ($\hat{\alpha}_{cr}^o$) net of census and cohort effects. The figures also show the unweighted linear regression line fit, net of country fixed effects; $\hat{\alpha}_{cr}^y = \alpha_c + \beta \times \hat{\alpha}_{cr}^o + \epsilon_{cr}$. Dots are color-coded by African region.

literate old applies to both genders though it is somewhat stronger for girls. Second, while inertia is present for both rural and urban households, the educational fate of the young generation appears more sensitive to the old’s heritage in rural places.

3.5 Summary

The mapping of the spatial distribution of educational opportunity across Africa reveals new regularities. First, there are wide differences in IM across countries. Second, within-country regional disparities in IM are large, especially in low education/income countries. Third, upward mobility is higher and downward IM lower for urban households. Forth, gender disparities are, on average, small, but in the Sahel and North Africa, it is harder for girls of uneducated parents to complete primary schooling. Fifth, upward IM is strongly linked to the average parental education in the region. Likewise, downward IM is negatively correlated to the literacy of the old generation, though this association is less strong. Sixth, inertia is more substantial for rural, as compared to urban households. These patterns suggest slow convergence,¹⁷ as improvements in educational attainment among illiterate households are larger in regions with relatively higher human capital levels. Persistence may stem either from regions’ independent impact on educational mobility or from spatial sorting. We return to this question in section 5.

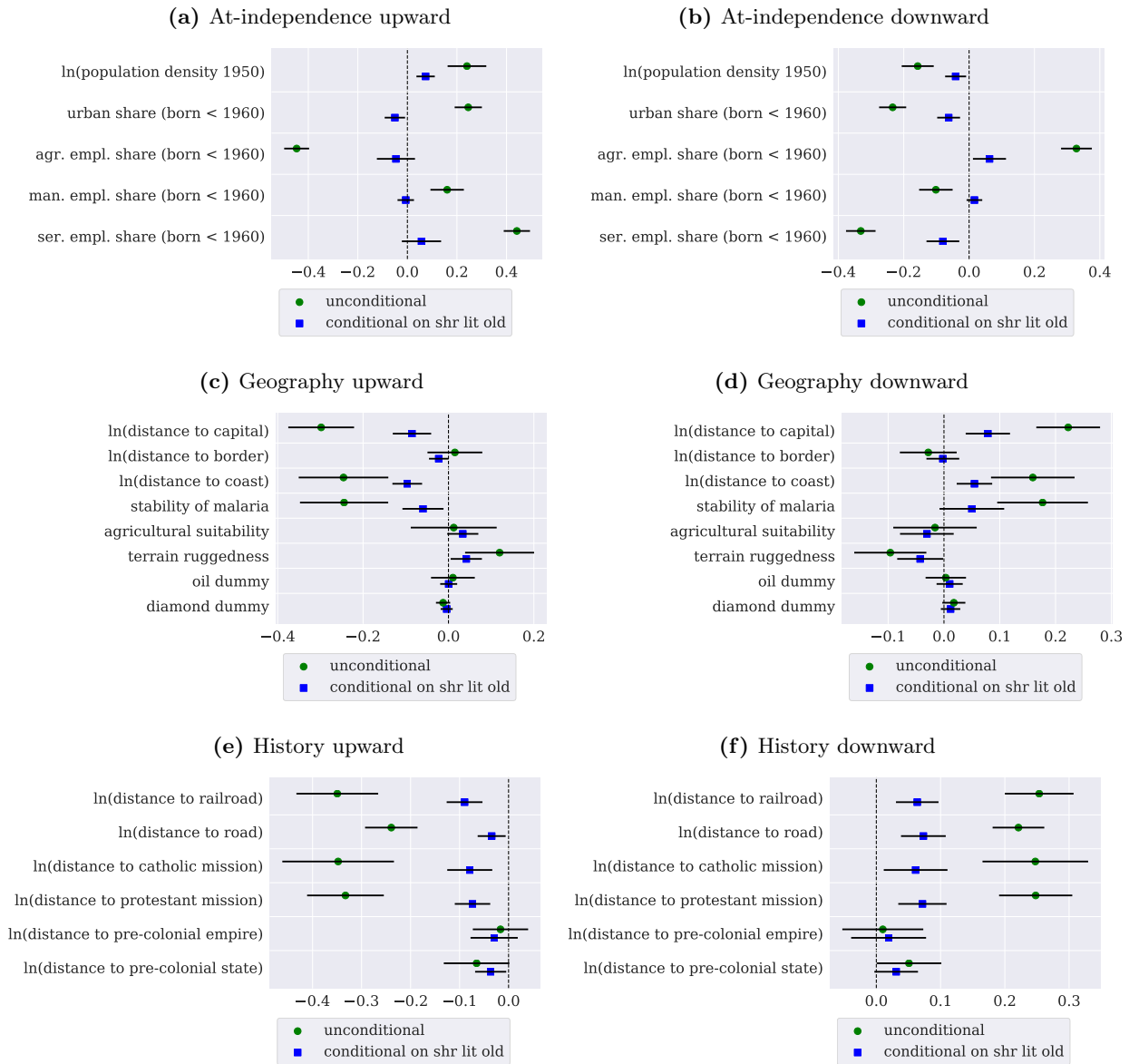
4 Correlates of Intergenerational Mobility

In this Section, we explore the correlates of regional IM, aiming to characterize its geography. We run univariate specifications linking IM to geographical, historical, and at-independence variables, discussed in the research on the origins of African development and studies on mobility outside Africa. [Appendix F provides variable definitions and

¹⁷In Alesina et al. (2020b) we show that terms typically estimated in education-growth-convergence regressions have a natural connection to absolute upward and downward IM. Our approach therefore connects to studies on educational convergence.

sources.] As the literacy of the old generation correlates strongly with IM, we also report specifications conditioning on it. The correlational analysis, albeit simple, is useful to illustrate whether the geographic and historical factors are associated with contemporary IM only through their correlation with initial conditions (education of the old) that still matter due to inertia, or whether they correlate with the rate at which educational endowments are transmitted intergenerationally above and beyond their association with the initial conditions. Figures 9 plot the (unweighted) within-country standardized correlation (“beta”) coefficients between upward and downward IM with the various features. Standard errors are clustered at the country level. The Appendix reports permutations: (i) adding province constants to condition on more localized, time-invariant features. (ii) dropping North African countries, as their historical development differs from Sub-Saharan Africa; (iii) excluding regions with cohabitation below 80%.

Figure 9: Within-Country Correlates of Regional IM



4.1 Development At Independence

We commence examining the association between IM and proxies of economic development in the 1950s-1960s when most African countries turn independent. Figures 9 (a)-(b) plot the correlations. We first explore how IM relates to (the log of) population density in 1950, that we take as a proxy for local development. Population density correlates positively and significantly with upward IM and negatively with downward IM. This result may not be surprising, as population density and the literacy of the “old” generation are strongly correlated. Coefficients decline once we account for the latter, though they retain significance. Population density correlates more strongly with upward -as compared to downward- IM (“beta” coefficients of 0.074 and -0.04). A similar pattern obtains when we look at urbanization.

Motivated by the literature on structural transformation in Africa (e.g., McMillan, Rodrik, and Verduzco-Gallo (2014)), we explore the correlation between IM and the employment shares across broad economic sectors.¹⁸ Agricultural employment is negatively correlated with upward mobility and positively correlated with downward mobility; these patterns hold when we condition on the literacy of the “old”. The specifications using the labor share in services or manufacturing on the RHS yield a “mirror” image.¹⁹

4.2 History

Figures 9 (c)-(d) plot the correlations between IM and historical variables.

Colonial Roads and Railroads Colonial railroads and roads have played an important role in African countries’ post-independence development (e.g., Jedwab and Moradi (2016)). Log distance to colonial railroads is significantly related to both upward and downward IM, even conditional on the old’s literacy. Districts that are one standard deviation closer to colonial railroads have, on average, 0.08 standard deviation higher levels of upward and lower levels of downward mobility. The estimates are virtually unchanged when we explore within-province variation.

Colonial Missions Earlier studies uncover positive effects of Christian missionary activity on education (e.g., Wantchekon, Klačnja, and Novta (2015)). We examine the correlation between IM and proximity to colonial missions using data from Nunn (2010) and Cagé and Rueda (2016). There are 1,321 (361 Catholic, 933 Protestant, 27 British and Foreign Bible Society) and 723 (Protestant only) missions in these datasets, respectively. Proximity to Christian missions correlates significantly with “old’s” literacy rates (results not shown). The Figures illustrate a significantly positive (negative) association between proximity to missions with upward (downward) IM. When we condition on the literacy of the “old”, the distance coefficient declines in absolute value but retains significance (beta 0.07). While data on missions are coarse (Jedwab, zu Selhausen, and Moradi (2018)), the

¹⁸We use data for individuals born before 1960. To abstract from migration, we focus on individuals residing in their birth district (the results are similar if we use all individuals). As we lack migration data for Lesotho, Nigeria, and Zimbabwe, the sample spans 24 countries.

¹⁹These results square with the concurrent analysis of Asher, Novosad, and Rafkin (2020), who document higher relative upward educational mobility rates in urban -manufacturing-service-oriented Indian districts as compared to those specializing in agriculture.

analysis suggests that investments by Christian missions have lasting consequences, both by shaping initial literacy which in turn increases educational mobility and by directly influencing mobility.

Precolonial Political Centralization We then explored the correlation between IM and pre-colonial political centralization that correlates with regional contemporary development (Michalopoulos and Papaioannou (2013)). We associate IM with log distance to the centroid of the nearest pre-colonial kingdom/empire using data from Brecke (1999) and to pre-colonial states using Murdock (1967) (though data are missing for parts of the continent). Distance to pre-colonial states is not a robust correlate of IM.

4.3 Geography

Figures 9 (e)-(f) plot the within-country correlations between IM and geographic, location, and ecological features.

Distance to the Capital Much evidence documents the limited ability of African states to broadcast power outside the capitals (e.g., Michalopoulos and Papaioannou (2014)). During colonization, the limited public goods were confined to the capital and a few urban hubs. The literacy of the “old” is much higher in the capital than the hinterlands; similarly upward IM also declines further from the capital city. The standardized coefficient drops, once we condition on the literacy of the “old”, from -0.29 to -0.094 , though it remains precisely estimated. The patterns are similar with downward mobility.

Distance to the Border African borders appear unruly and conflict prone, as they often partition ethnic groups (e.g., Alesina, Easterly, and Matuszeski (2011)). Nevertheless, there is no systematic association between IM and distance to the border.

Distance to the Coast Economic activity in Africa is concentrated along the coastline. Thus, literacy falls once one moves inland (results not shown). Proximity to the coast relates to the presence of Europeans and associated investments during colonization, but also to the intensity of slave raids. Upward (downward) educational mobility is significantly higher (lower) in coastal areas. The coefficient retains significance when we condition on the literacy of the old.

Malaria We associate IM with an index reflecting a district’s malaria ecology that has been linked to Africa’s underdevelopment (e.g., Gallup and Sachs (2001)). Malaria correlates strongly with IM; the association operates above and beyond initial differences in literacy (that correlate with malaria).

Land Quality for Agriculture Upward IM is somewhat higher and downward IM is lower in regions with high-quality land, but the correlations do not pass standard statistical significance thresholds.

Ruggedness We then examined the association between IM and ruggedness that correlates positively with cross-country economic performance in Africa, as rugged terrain shielded regions from slave raids (Nunn (2008)).²⁰ Moreover, as malaria is pervasive in the lowlands, populations in mountainous terrains are less affected. There is a positive and significant association between terrain ruggedness and the literacy of the “old” generation. Upward IM is significantly higher and downward IM is lower in rugged regions. The correlations remain significant when we control for the old generation’s literacy, which is higher in regions with rugged topography. These results add to Nunn and Puga (2012) that across African countries ruggedness correlates positively with output.

Natural Resources The “natural resource curse” literature links conflict and underdevelopment to oil, diamonds, and precious minerals (e.g., Berman et al. (2017)). The association between IM and the presence of oil fields or diamond mines is weak and never passes significance thresholds. This most likely reflects opposing mechanisms, as natural resource wealth also spurs human capital accumulation and structural transformation in Africa (Hohmann (2018b)).

4.4 LASSO Estimates

We also employed LASSO (Least Absolute Shrinkage and Selection Operator), a simple machine learning method that is useful in detecting robust predictors in the presence of multi-collinearity and measurement error. The LASSO analysis -reported in Appendix F.2- reveals some interesting patterns that complement the univariate correlations. First, distance to colonial railroads and distance to the capital are the most important features predicting IM; this result suggests that colonial transportation investments, though overall small and mostly connecting ports with mineral rich interior areas, had lasting consequences. Second, proximity to natural resource and precolonial states have minimal power predicting IM. Third, terrain ruggedness, distance to the coast, and malaria ecology lie in-between, carrying some modest power predicting regional IM. Fourth, proximity to Protestant missions is a robust predictor of IM, while proximity to Catholic missions drops out of the empirical model once regularization increases.

4.5 Summary

Colonial railroads, proximity to the capital, and to (Protestant) missions correlate strongly with mobility. Geographic aspects, terrain ruggedness and malaria ecology are also relevant in characterizing educational mobility. In contrast, natural resources, proximity to borders and precolonial statehood do not seem to play a role. As these variables also correlate with the old generation’s literacy, which is the most influential covariate of mobility, when we condition on it, the coefficients drop roughly by two-thirds. These patterns suggest that geography and history mostly matter by shaping at-independence development

²⁰We also run specifications using regional proxies of slave trade intensity using data from Nunn (2008)). The data are, however, not well-suited for our analysis. First, the data are at the ethnicity rather than the region level. Assigning them to contemporary regions overlapping historical homelands using ethnographic maps introduces error. Second, the ethnicity data do not cover the Trans-Saharan and the Red Sea slave trades that are relevant for Ethiopia, North, and South Sudan, Mali, Kenya, Nigeria and Senegal.

(education of the “old”), which appears quite persistent across most African countries.²¹

5 Regional Childhood Exposure Effects

Does the environment “cause” mobility? To answer this question, we follow the approach of Chetty and Hendren (2018a) and exploit differences in the timing of children’s moves across districts to isolate regional childhood exposure effects from sorting. This approach compares the educational attainment of children whose families moved to a better/worse region -in terms of average mobility- at different ages to identify the rate at which their attainment converges to that of permanent residents. If regions affect individual mobility, this effect should be stronger, the longer the exposure to the new environment.

We first describe the semi-parametric specification, discuss the identifying assumptions, and report the results. Second, we present parametric estimates, explore heterogeneity, and summarize the sensitivity checks. Third, we isolate moves due to displacement shocks in the origin and use past migration destinations to “instrument” for the location of moving families to advance on causation.

5.1 Baseline Semi-Parametric Estimates

5.1.1 Specification

For children who moved from place of birth o to destination region d at age m , their attainment can be expressed as follows:

$$\begin{aligned} \text{IM_up}_{ihbmcod} = & [\psi_h +] \alpha_{ob} + \alpha_m + \sum_{m=1}^{18} \beta_m \times \mathbb{I}(m_i = m) \times \Delta_{odb} \\ & + \sum_{b=b_0}^B \kappa_b \times \mathbb{I}(b_i = b) \times \Delta_{odb} + \epsilon_{ihbmcod}, \end{aligned} \quad (5)$$

The dependent variable equals one if child, i , born in cohort b in country c to illiterate household h , completes primary education (or higher) and zero otherwise (upward IM). The variable of interest, Δ_{odb} , denotes the difference between upward educational mobility of permanent residents in the destination minus origin for children born in cohort b :

$$\Delta_{odb} = \widehat{\text{IM_up}}_{bd}^{\text{nm}} - \widehat{\text{IM_up}}_{bo}^{\text{nm}}.$$

Average region-cohort upward IM, computed among non-movers (individuals residing in their place of birth at the time of census), is a sufficient statistic summarizing the economic and social environment that shapes educational decisions. We estimate a different slope, β_m , for each age of move (years 1 to 18) controlling for any direct effect via age of move constants, α_m ; these capture disruption effects and any other age-specific unobserved feature that affects the education trajectory. Origin-region \times birth-decade fixed effects, α_{ob} , account for unobserved factors of the child’s birthplace at the time of birth.

²¹Two caveats apply here. First, these correlations do not imply causal effects. Second, the correlations may reflect differential measurement error across the various regressors and the education of the old.

We add interactions of destination-origin differences in cohort-specific IM with cohort effects, to partly account for potential differential measurement error across cohorts and other trends (this has no effect). The intuition of the above specification is that if children move from regions with worse to places with better educational opportunities ($\Delta_{odb} > 0$), and exposure matters, the earlier the move, the greater the effect of the region. Since the specification includes (3, 231) origin-cohort fixed effects, variation comes from children born in the same place in the same decade, who move to regions with different mobility.²² Modeling exposure effects in proportion to years spent in destination follows Chetty and Hendren (2018a), who derive a similar parsimonious relationship from a generic setting of exposure effects.

The age-specific slopes, β_m , are identified even in the presence of sorting; i.e., parents without primary schooling, but with a higher propensity to educate their children, are more likely to move to regions with better opportunities. The identifying assumption is that the *timing* of the move is uncorrelated with latent children’s ability. In other words, parents more likely to invest in their children’s education can move from worse to better environments, on average; but the more “ambitious” parents should not move earlier. Since this is a restrictive assumption, we relax it estimating a household fixed-effects variant of equation (5), with ψ_h . In these models, the age-specific slopes, β_m , reflect the extent to which educational attainment differences between siblings relate to the regional gap at the age of move, interacted with differences in the mobility of permanent residents between origin and destination, $(m1 - m2)\Delta_{odb}$. The identifying assumption is that households who move to places with higher (lower) upward mobility do not do so to favor some of their children. We return to this issue below.

5.1.2 Sample and Descriptive Statistics

In this section, we work with a sample of 16 countries (11, 169, 357 matched-to-parents children, aged 14 – 25, whose household moved before 18) where IPUMS records the current and birth region, and years in the current residence.²³ Overall, the average (median) migrant outflow share [number of migrants leaving a region divided by total residents] during census years (where we observe total population) is 0.081 (0.038), while the corresponding mean (median) inflow share is 0.058 (0.038). These statistics are broadly in line with the survey evidence in FAO (2017) and United Nations Conference on Trade and Development (2018). Hohmann (2018a) uses IPUMS data to estimate migration gravity equations within African countries. He documents distance elasticities of about 1, quite close to the estimates of migration flows across U.S. states 2005-2016.²⁴

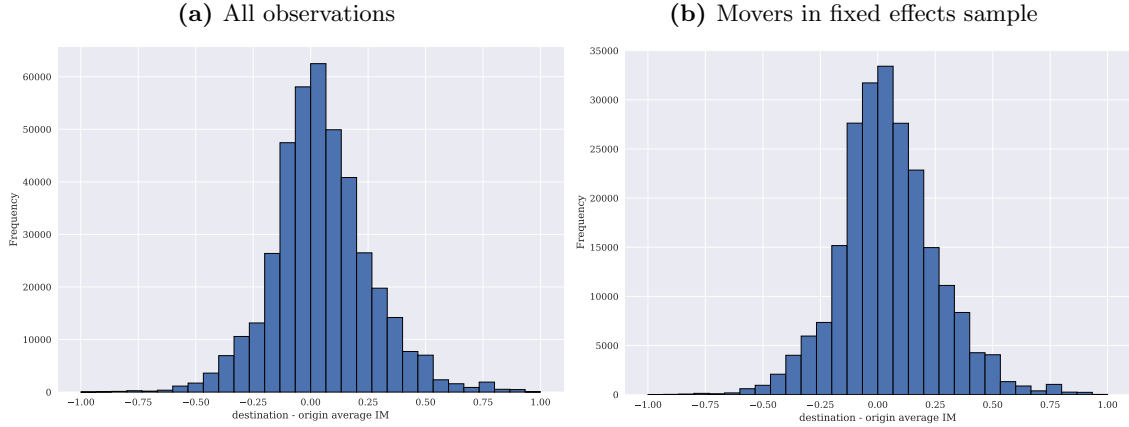
²²The only difference vis a vis Chetty and Hendren (2018a) is that we are not interacting the origin-cohort effects α_{ob} with age-at-move m . Doing so would require adding more than 100,000 fixed-effects, 1,084 (regions) \times 5 (cohorts) \times 18 (age at move).

²³The countries (number of observations) [number of regions] are Benin (28,076) [76], Cameroon (38,415) [230], Egypt (81,525) [27], Ethiopia (16,340) [87], Ghana (20,259) [10], Guinea (8,718) [34], Kenya (21,177) [177], Morocco (27461) [58], Mali (22,256) [47], Malawi (10,986) [30], Rwanda (11,687) [103], Sudan (24,276) [25], Togo (7,616) [36], Uganda (28,764) [56], South Africa (13,483) [9], and Zambia (45,136) [72]. For some countries, birth is at admin-1 level, whereas residence is at admin-2 level. In other countries, region of residence and birth are at the same level. We harmonized residence and birth region at the finest level and end up with 1,084 “*birth/current residence regions*”.

²⁴The literacy rate of the old generation for households moving to regions with higher mobility exceeds that of the non-moving households by 21 percentage points. Moving households in destinations with lower than the origin IM also have an old-generation literacy edge over non-moving households of 0.07. As our

Figure 10 plots the histogram of Δ_{odb}^{nm} . Panel (a) looks across the entire sample of moving children (406, 175); panel (b) looks at children of moving families that we consider in the within-household specifications (226, 739). The mean and median are positive, .05 and .034, respectively; on average, families move to regions with higher levels of upward mobility, though migration flows both ways. Roughly 58% move to regions with higher upward IM. These statistics complement Young (2013) who documents substantial bidirectional urban-rural migration flows across African regions with survey data.

Figure 10: Differences in Intergenerational Mobility between Destination and Origin



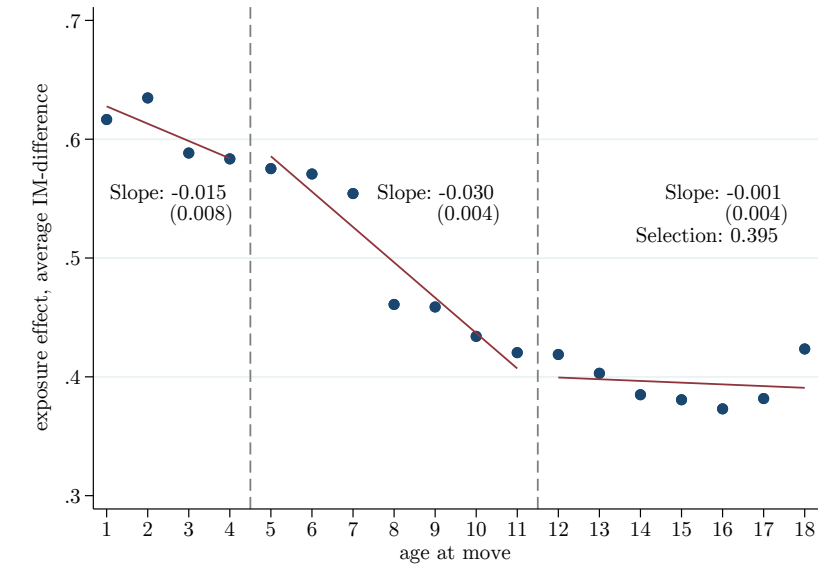
The figures plot the distribution of Δ_{odb}^{nm} – the destination minus origin differences in cohort-region average non-migrant IM. Panel (a) plots the distribution for all migrant children, aged 14-25. $\mu = 0.049$, $p_{50} = 0.034$, $\sigma = 0.214$. Panel (b) plots the distribution for migrant children, aged 14-25, residing in a household with at least two children of different ages at the time of the move. $\mu = 0.048$, $p_{50} = 0.034$, $\sigma = 0.214$.

5.1.3 Results

Figure 11 plots the age-specific exposure effects, $\hat{\beta}_m$, against the child’s age at the time of the move. The figure uncovers two regularities: “*regional exposure effects*” that are particularly strong for children aged 5 – 11 and “*selection effects*.” First, the slopes are significantly positive for children moving at *all* ages. This applies even for children who move at the age of 13 – 18 ($\hat{\beta}_m \approx 0.40$). Since the destination is rather unlikely to have a causal effect on primary school completion for children moving after the age of 14, the estimates reflect *selection*. Households moving to regions with higher (lower) IM have unobservable characteristics translating into a higher (lower) propensity that children complete primary school. The degree of selection does not vary with children’s age after the age of 13–14. Children who move to regions where permanent residents have one percentage point higher upward IM have a 0.4 higher likelihood to complete primary education purely due to spatial sorting.

Second, the estimates reveal *regional exposure effects*, since moving to a better (worse) district early in life, roughly before the age of 12, translates into a higher (lower) likelihood of upward educational mobility. The estimates are around 0.65 for children whose family moved before they turn 5 years old; the likelihood to complete primary schooling is 30 percentage points higher if parents move to regions with 0.5 higher levels of IM (mean analysis focuses on children of households where the old generation has not completed primary education, we effectively condition on such differences).

Figure 11: Semi-parametric Childhood Exposure Effects on Primary Education, Observational Estimates



$IM = .6$, standard deviation = $.49$). As the pure selection effect is around 0.4, regional exposure effects total around 0.25 for children moving shortly after their birth. The relationship between age at move and exposure effects is negative, but not very steep for children moving before 5 – 6; moving to regions with higher mobility yields almost equally large benefits (likelihood to complete primary schooling) for children who are between 1 and 4 years old. The age at move estimates for children moving between ages 5 – 12 decline approximately linearly, revealing that the differential impact of moves in high mobility regions is especially large for younger kids. Chetty and Hendren (2018a) define the *regional exposure effect* as $\gamma_m = \hat{\beta}_{m+1} - \hat{\beta}_m$. Regressing the slopes on the age at move for ages 5 to 11, we obtain an estimate of about $-.03$. That is, for every additional year spent in this age bracket, a child of illiterate parents sees her chances for completing primary increase by roughly 3 percentage points. If instead we run a regression pooling across ages 1-11, we obtain a slope of -0.022 (0.002).

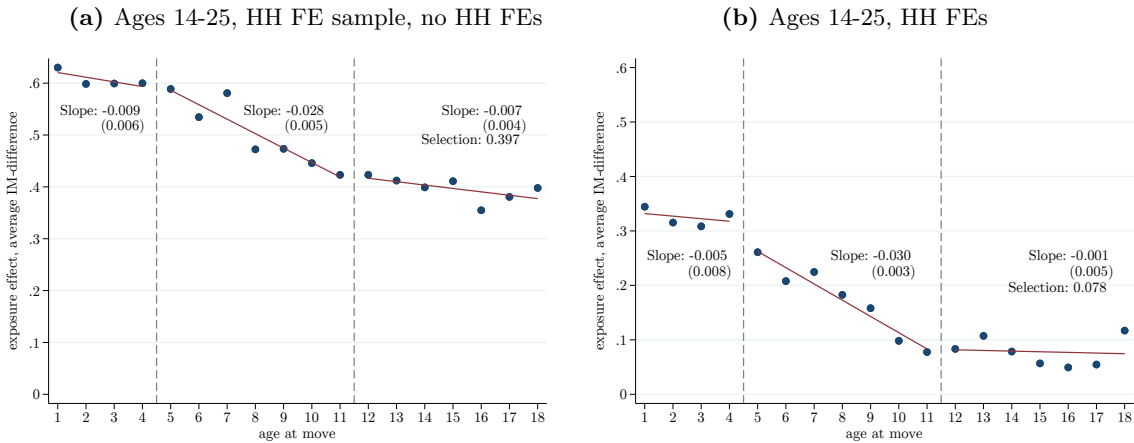
5.1.4 Household Fixed-Effects Estimates

We then add to regression equation (5) family-specific constants, ψ_h , to exploit variation among children belonging to the *same* household, who moved at different ages. Doing so, we relax the assumption that latent family characteristics are orthogonal to the move.

Figure 12, panel (b) plots the age-specific exposure effects, $\hat{\beta}_m$, obtained when comparing siblings that moved at different ages; panel (a) omits them to allow comparability of the cross-sectional and the within-household estimates in the same sample (226,739 children from 90,022 households with more than one child in-between 14 – 25). First, the selection/sorting effect, captured by the slopes after age 12, drops significantly, once we account for unobserved family features, from 0.40 (panel (a)) to 0.078 (panel (b)). The 90% confidence intervals (not shown) include 0 for all age-of-move slopes after 12. Family constants account almost fully for selection/sorting.

Second, the household-fixed-effects specifications also yield significant regional expo-

Figure 12: Semi-parametric Childhood Exposure Effects on Primary Education, Observational and Within-family Estimates



sure effects. The slopes for children moving during ages 1 – 4 are around 0.35; two siblings moving to a region with higher IM when they are 1 and 4, respectively, have, on average, the same increase in the likelihood of completing primary schooling. If the difference between the destination and the origin (Δ_{odb}^{nm}) is close to one standard deviation (0.5), the increase in upward-IM is around 18 percentage points for both siblings. The age-of-move slopes, $\hat{\beta}_m^{fe}$, fall for children moving when they are between ages 5 and 12. The estimate of the exposure effects for ages 5 – 11 is $\gamma_m^{fe} = \hat{\beta}_{m+1}^{fe} - \hat{\beta}_m^{fe} = -.03$.

The comparison of the cross-sectional to the within-household specifications reveals that sorting is considerable; around two-thirds of the total magnitude. The marginal impact of moving to areas with higher (lower) mobility is the same when we look across all moving children and when we compare children of the same family. The fact that the household constants reduce the magnitude of the age at move coefficients, but do not affect their slope suggests that where families choose to move does not vary with children’s age.²⁵

5.2 Parametric Estimates

5.2.1 Specification

Regression equation (5) is demanding, as it includes thousands of origin-cohort fixed effects; this issue becomes more challenging when we add household constants. Following Chetty and Hendren (2018a) we estimate a parametric variant of specification (5).

²⁵We run pairwise tests of coefficient equality (see Greene (2011), section 5.4) for all ages-at-move (see Appendix Figure G.1). The difference between the coefficients of ages 1 – 4 and 12 – 18 is significantly different from zero across most permutations.

$$\begin{aligned}
\text{IM_up}_{ihbmcod} = & [\psi_h +] \sum_{b=b_0}^B \mathbb{I}(b_i = b) \times \left(\alpha_b^1 + \alpha_b^2 \times \widehat{\text{IM_up}}_{ob}^{\text{nm}} \right) + \\
& \sum_{m=1}^{18} \zeta_m \times \mathbb{I}(m_i = m) + \sum_{b=b_0}^B \kappa_b \times \mathbb{I}(b_i = b) \times \Delta_{odb} + \\
& \mathbb{I}(m_i < 5) \times (\beta_0 + (18 - m_i) \times \beta_1) \times \Delta_{odb} + \\
& \mathbb{I}(5 \leq m_i \leq 11) \times (\gamma_0 + (18 - m_i) \times \gamma_1) \times \Delta_{odb} + \\
& \mathbb{I}(m_i \geq 12) \times (\delta_0 + (18 - m_i) \times \delta_1) \times \Delta_{odb} + \epsilon_{ihbmcod}. \tag{6}
\end{aligned}$$

Instead of origin-cohort fixed effects, α_{ob} , equation (6) includes birth-cohort constants interacted with a linear-in-origin-IM term. The equation also imposes a piecewise linear structure, allowing the regional exposure effects to differ for pre-school years (ages 1 – 4), the ages relevant for primary school (5 – 11), and post-primary education years (12 – 18).

5.2.2 Results

Table 4 reports the results. Column (1) shows that the *marginal* exposure effect for children whose families moved when the children were more than 12 years old is zero and statistically insignificant. The marginal exposure effect for children moving before 5 is 0.019 and weakly significant. Exposure to areas with higher mobility is especially strong for children whose (illiterate) parents move when they are in the ages critical for primary school, roughly between 5 and 11. Reassuringly, the estimate (0.031) is similar to the semi-parametric estimates (obtained in two steps). Column (2) shows that the coefficients for the three age-of-move brackets are similar in the smaller sample of individuals included in the household-fixed-effects specifications, reported in column (3). The marginal exposure for children whose families moved when they were older than 12 is zero. The slope for moves before 5 years is 0.006, statistically indistinguishable from zero. The slope is 0.0305, tightly estimated for children moving between 5 and 11.

Table 4: Parametric Estimates of Regional Childhood Exposure Effects

	(1)	(2)	(3)
	IM	IM	IM
β : 1-4	0.0189*	0.0128	0.00643
	(0.011)	(0.017)	(0.016)
γ : 5-11	0.0309***	0.0292***	0.0305***
	(0.005)	(0.005)	(0.006)
δ : 12-18	-0.000462	0.00159	0.00198
	(0.006)	(0.006)	(0.004)
R-squared	0.142	0.119	0.679
N	406175	226739	226739
age at mig FE	yes	yes	yes
birth decade FE	yes	yes	yes
hh FE	no	no, hhfe sample	yes
age range	14-25	14-25	14-25

The dependent variable in all specifications is an indicator variable that takes the value of one for children of parents without completed primary education who have completed at least primary education and zero otherwise (upward IM). The independent variables comprise a linear origin-average-IM (calculated for the birth-cohort relevant to the individual among non-movers) term, age-at-move indicator variables, birth-decade \times destination indicators interacted with destination-minus-origin differences in upward IM, all of which are not reported, and three linear terms for destination-minus-origin differences in the relevant-birth-cohort-non-mover average IM for moves taking place when the child moves, ages 1-4, 5-11, and 12-18. Double clustered at the origin and at the destination district standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

5.2.3 Heterogeneity

We examined heterogeneity across children moving to regions with higher (lower) IM than their place of birth and heterogeneity across gender, augmenting equation (6) with interactions between the linear-in-age-at-move regional exposure effects for the three age-of-move brackets with the respective indicator variables. For brevity we report these results in the Online Appendix Table G.1.

There is not much heterogeneity on regional exposure effects between moves to higher and lower IM regions. The estimates are small and statistically insignificant for moves after 12 or during ages 1 – 4 for both sets of children. Regional exposure effects are around 0.03 for moves to either worse or higher IM regions. The educational losses for children moving to worse regions before the age of 12 are roughly equal to the gains of children moving to regions with higher IM.

Regional exposure effects for both boys and girls moving before the age of 5 and after the age of 12 are unstable. The regional exposure effect for primary school age for boys is around 0.023, somewhat smaller than the baseline of 0.03. The interaction of Δ_{odb} with the female indicator for ages 5 – 11 is 0.01, suggesting that girls benefit (lose) somewhat more when moving to higher (lower) mobility regions.

5.2.4 Sensitivity Analysis

The uncovered regional exposure effects and sorting are robust to various permutations (Appendix Table G.2). These include: (i) dropping multi-generational households; (ii) looking only at children matched to biological parents; (iii) dropping North Africa.

Measurement Error We considered the possibility of measurement error in Δ_{odb}^{nm} that is not unlikely. First, we miss multiple moves, as censuses just report birth and current region. We also lack information on temporary migration. Second, as districts are large and there is likely within-district variation, IM captures imperfectly the relevant environments at birth and current residence. Third, as Chetty et al. (2020a) argue, the “noise to signal ratio is likely amplified since it is identified purely from residual variation in Δ_{odb}^{nm} , controlling for origin quality” (with the origin-cohort constants). To account for classical measurement error in differences in regional IM we employed a 2SLS estimator based on a sample split (see Appendix Figure G.2). The estimates increase by about 10%. In line with error-in-variables, we obtain somewhat larger estimates when we drop regions with few observations.

Household Income Shocks We also looked solely across rural households, where the old generation works in agriculture to assuage concerns that the uncovered regularities reflect income shocks triggering the move and, at the same disproportionately, affecting younger children. While we cannot control for household income at the timing of the move, as we do not observe it, such income effects are likely to be at best moderate for rural African households often engaged in subsistence farming. Appendix Table G.3 shows that an extra year in regions with higher than the origin IM increases the likelihood that children of rural/agriculture households will complete primary schooling in the 5 – 11 age bracket; the marginal effect of moving after 12 is tiny, as is for moves before 5.

5.3 Endogeneity

While the inclusion of household constants accounts for time-invariant family features that affect investments in education, time-varying factors may jointly drive household moves and children’s educational investments in proportion to exposure to the region with higher mobility. We address this -and related- concerns exploiting “push shocks” and using historical migration to predict the destination of moving households.

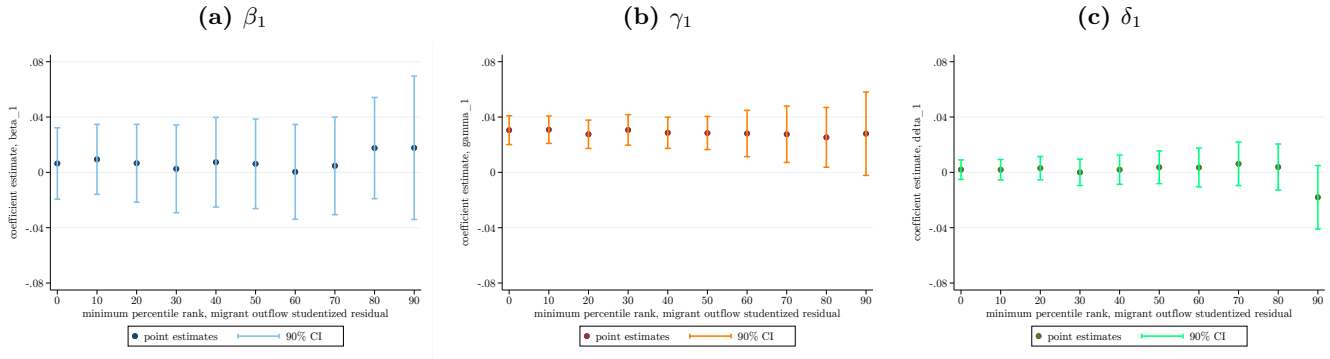
5.3.1 Displacement (Push) Shocks

As a starting point, we look at moves that are more likely to reflect (push/displacement) shocks exogenous to household decisions. To pinpoint anomalous periods of outflows from the origin, we first construct an origin-district-year migration panel for each country that covers roughly the period from 1965 until the last census year. Second, for each district, we regress outflows on a constant and a linear time trend and obtain residuals. Third, we sort the (standardized) residuals from highest to lowest. High (positive) residuals indicate years of abnormally large out-migration from a given district, while low (negative) residuals denote below trend outflows. The latter are more likely to reflect a household’s choice to move, while the former capture irregular district out-migration shocks that are more likely to be exogenous from the household’s viewpoint.

Figures 13 (a)-(c) plot the parametric regional exposure effects for the three age groups (1 – 4; 5 – 11; 12 – 18). Conservatively, we report the within-household specifications (results are similar when we use all data and omit the household constants). In the within-household estimates, looking at moves in years of unusually large outflows mitigates

concerns that the timing-of-move is chosen to favor some of the siblings. As we move from left to right, we successively drop observations focusing more narrowly on children whose families move in abnormal years. The left-most observation for each panel reports the benchmark estimates. The 50th percentile looks at moves that took place in years when flows have been above the historical district-specific median. The estimates of the 90th percentile look at moves that occurred during the two to five years with the highest outflows. This is because for most countries, we have outflow data for roughly 40 years.

Figure 13: Outflow (Displacement) Anomalies, Household Fixed-Effects Estimates



The figure shows parametric regression estimates of childhood regional childhood exposure effects concentrating, successively, on district-years that experienced increasingly larger migration outflows. Each point reports the marginal effect of an additional year of exposure in the relevant age-at-move range (with $\Delta_{odb}=1$). Panel (a) shows the marginal effect for ages-at-move between 1 – 4, panel (b) for 5 – 11, and panel (c) for 12 – 18. The left-most point for each graph shows the baseline estimates, where no observations are dropped. The next observation uses observations from district-years with studentized outflow residuals above the 10th percentile, etc. All regressions include household fixed effects. 90% confidence bands are constructed from double clustered standard errors at the origin and destination district.

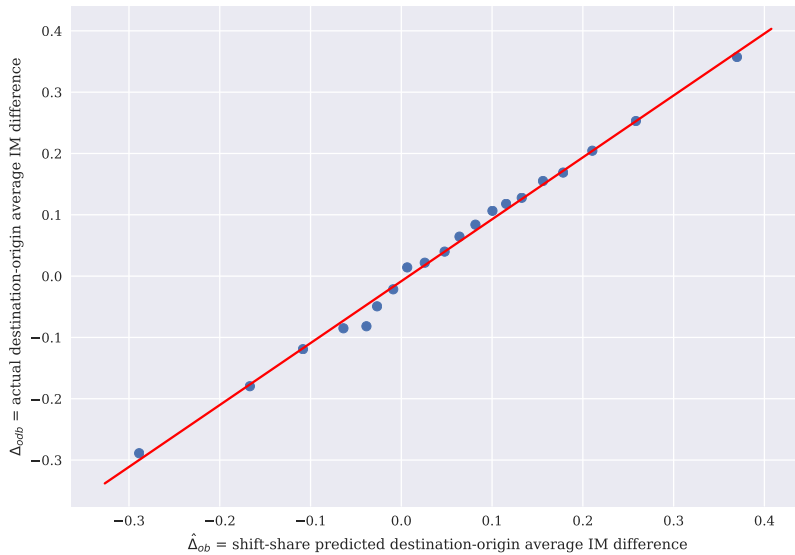
The marginal effects for moves before the age of 5 (panel (a)) and after 12 (panel (c)) are small and statistically insignificant. The marginal exposure effect is significantly positive for moves when kids are between 5 and 11. γ retains economic and statistical significance when we look at moves that most likely reflect origin-specific shocks, even when we drop 90% of the sample. These results suggest that the baseline estimates reflect regions’ independent impact on children’s educational attainment rather than unobserved time-varying household factors. Moreover, the estimates hint that the effects of moving are similar for families who decide to move for idiosyncratic reasons and displaced households (Chetty and Hendren (2018a) present similar patterns in the US).

5.3.2 Expected Destination of Moving Households

Moving households even when they relocate due to exogenous reasons, decide where to settle. If a household’s endogenous choice of destination also relates to differential investments into some children, then the estimates may be biased. While the household fixed effects partially account for such concerns, the choice of destination may still be correlated with unobserved child features. We use past migration destinations from each origin to predict where moving households will settle with a “shift-share” design. The idea behind this approach is that migrants tend to settle in regions where earlier migrants from their

community have moved to (e.g., Derenoncourt (2018)).²⁶

Figure 14: Predicted and Actual Migration. Binned Scatterplot: Δ_{odb} on $\hat{\Delta}_{ob}$



Binned scatterplot of actual destination-minus-origin differences in non-migrant IM (Δ_{odb}) on the vertical axis and the historical-migration-destination-share weighted-average at the origin $\hat{\Delta}_{ob}$ on the horizontal axis. A regression yields: $\Delta_{odb} = -0.01$ (*s.e.* = 0.01) + 1.01 (*s.e.* = 0.046) $\hat{\Delta}_{ob}$, R^2 : 0.53.

Figure 14 shows a binned scatterplot of actual and historical-predicted migration. The elasticity is one and precisely estimated. We then estimated the parametric specification, replacing actual Δ_{odb} with the historical-predicted difference $\hat{\Delta}_{ob}$. Table 5, columns (1)-(3), report the “reduced-form” estimates, while columns (4)-(6) report 2SLS that combine the “reduced-form” estimates with the “first stage”. Columns (1)-(4) and (2)-(5) report cross-sectional estimates in the full sample of moving children and in the sample where we compare siblings, respectively. Columns (3) and (6) report household fixed-effects estimates. As the first-stage slope is approximately one, the reduced form and 2SLS estimates are similar. The marginal impact of an additional year in the region with higher (lower) mobility in the critical for primary schooling age is significantly positive; the estimate is somewhat larger than the OLS, most likely because instrumentation reduces classical measurement error (see Section 5.2.4). The corresponding to column (4) OLS estimate (in the sample of 391,371 obs)) with actual, rather than projected, $\hat{\Delta}_{ob}$ is 0.035, about 15% lower than the IV.²⁷ The exposure effect is small and statistically indistinguishable from zero for moves after the age of 12 and before 5.

²⁶For any migration year y , we compute the destination- d share from origin o as $\sigma_{ody} = \frac{\sum_{x=T_0}^{y-w} \text{migrants}_{odx}}{\sum_{d=1}^D \sum_{x=T_0}^{y-w} \text{migrants}_{odx}}$ where D is the total number of districts in the country, T_0 is the first year for which we observe a migrant and w is a time window; we set $w = 10$ to avoid migration flows reflecting the delayed response to past shocks). For individuals who migrate in year y from o to d , we compute “predicted” $\hat{\Delta}_{od}$ as the historic share-weighted analog, $\hat{\Delta}_{ob} = \sum_{d=1}^D \Delta_{odb} \times \sigma_{ody}$. Δ_{odb} depends on the average IM of non-migrants in the migrating children’s birth decade in origin and destination. σ_{ody} depends on the number of people who moved from o to d up to w years prior to year y .

²⁷In these specifications, Guinea and Egypt (1986) drop, because the residence tenure variable has a maximum of 10 years and we cannot therefore construct the historical migrant share.

Table 5: Parametric Estimates of Regional Childhood Exposure Effects: Shift-share Instrument for Origin-Destination Differences, Reduced Form and IV Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
	IM	IM	IM	IM	IM	IM
beta: 1-4	0.0189 (0.013)	-0.00393 (0.019)	-0.00478 (0.020)	0.0185 (0.013)	-0.00362 (0.018)	-0.00579 (0.020)
gamma: 5-11	0.0413*** (0.006)	0.0470*** (0.007)	0.0534*** (0.009)	0.0407*** (0.007)	0.0466*** (0.008)	0.0547*** (0.011)
delta: 12-18	0.0145* (0.008)	0.0113 (0.008)	0.00515 (0.006)	0.0132* (0.008)	0.0112 (0.007)	0.00447 (0.006)
R-squared	0.124	0.103	0.679	0.132	0.110	0.007
N	391372	219210	219210	391372	219210	219210
age at mig FE	yes	yes	yes	yes	yes	yes
birth decade FE	yes	yes	yes	yes	yes	yes
hh FE	no	no, hhfe sample	yes	no	no, hhfe sample	yes
age range	14-25	14-25	14-25	14-25	14-25	14-25
estimator	OLS	OLS	OLS	2SLS	2SLS	2SLS

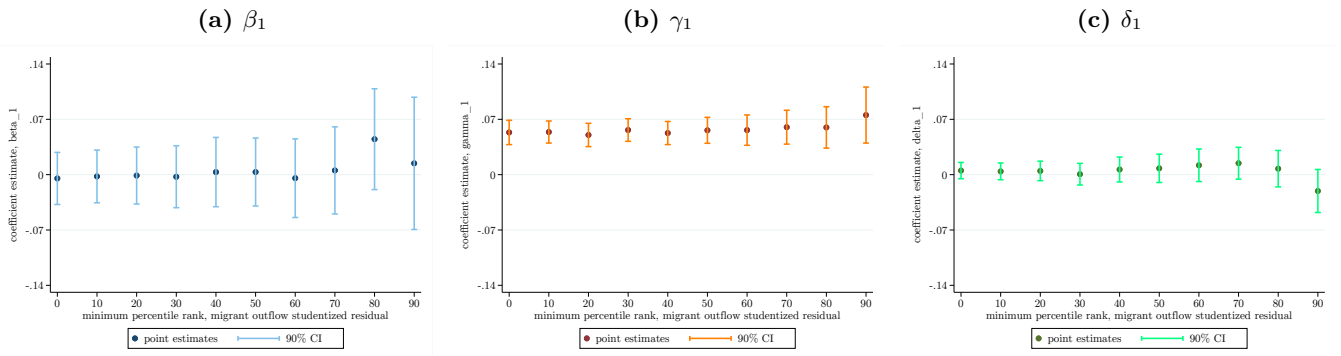
The dependent variable in all specifications is an indicator that takes the value of one if the child of parents who have not finished primary education has completed at least primary schooling and zero otherwise (upward IM). The independent variables comprise a linear origin-average-IM (calculated for the birth-cohort relevant to the individual among non-movers) term, age-at-move indicator variables, birth-decade \times destination indicators interacted with destination-minus-origin differences in upward IM, all of which are not reported, and three linear terms for destination-minus-origin differences in the relevant-birth-cohort-non-mover average IM for moves taking place when the child moves, ages 1-4, 5-11, and 12-18. Columns (1)-(3) report “reduced-form” estimates, using differences in upward mobility between origin and destination district projected by past migration. Columns (4)-(6) report 2SLS (two-stage-least-squares) estimates, where actual differences in upward IM between origin and destination district for moving children is “instrumented” with differences in upward IM projected based on historical migration. Standard errors double clustered at the origin and at the destination district level are reported in parentheses below the coefficients. $*p < 0.1$, $**p < 0.5$, $***p < 0.01$.

5.3.3 Blending “Push” Shocks with Expected Destination

In a demanding test that blends the two approaches, we replace (or instrument) actual differences in IM between origin and destination (Δ_{odb}), with those predicted from historical migration ($E[\Delta_{odb}|t-10]$) and sequentially keep observations of moves taking place in anomalous origin-district years. Figures 15 (a)-(c) plot the “reduced-form” estimates for the marginal exposure effects from the parametric specification (6) for the three age brackets. The regional exposure effect for moving children after the age of 12 is zero and tightly estimated. The estimate for moves before the age of five is also centered around zero, although the standard error bands are wide. The regional exposure estimate for kids moving in-between 5 and 12 is positive, around 0.045. The coefficient retains significance even when we drop 90% of the observations, effectively looking at children whose families moved in the two-five most abnormal years of out-migration from their place of birth. These estimates – that jointly account for the endogeneity of the move from district o , by looking at years of abnormal outflows, and households’ choices of destination d , by using historical (lagged by 10 years) migration – advance the causal interpretation of regional childhood exposure effects.

Sensitivity Analysis Appendix, Section G.3, reports additional results and sensitivity checks focusing on the more demanding approach that focuses on moves in years of displacement shocks and projects households’ choices of destination with historical migration. First, the 2SLS estimates are similar to the reduced-form ones, as the first-stage is approximately 1. Second, we obtain similar results when we omit the household constants

Figure 15: Regional Exposure Effects from Displacement Shocks. Household Fixed-Effects Reduced-Form Estimates



This figure shows parametric regression estimates of regional exposure effects looking, successively, on district-years that experienced increasing larger migration outflows and using predicted by historical migration differences between origin and destination, $\hat{\Delta}_{ob}$. Each point gives the marginal effect of an additional year of exposure with $\hat{\Delta}_{ob}=1$. Panel (a) shows the coefficients for ages-at-move 1-4, panel (b) for 5-11, and panel (c) those for 12-18. The left-most point for each graph shows the baseline estimates, where no observations are dropped. The next point uses observations from district-years with studentized outflow residuals ranked above the 10th percentile, etc. All regressions include household fixed effects. 90% confidence bands are constructed from double clustered standard errors at the origin and destination district.

and focus on the larger sample that covers children from all moving households. Third, we define “large outflows” at the country- rather than at the region-level. The regional exposure effects for moving children after the age of 12 is zero. The estimates for moves before the age of five are positive, but statistically insignificant. The regional exposure estimate for kids moving in-between 5 and 12 hovers around 0.045.

5.4 Summary

This Section reveals two results: First, sorting is considerable. Second, regions matter. Children who move earlier in life to regions where residents have higher intergenerational mobility are more likely to complete primary schooling. This pattern also applies when we compare siblings. Regional childhood exposure effects are present, even when we look at moves triggered by displacement shocks at the origin and when we account for the potentially endogenous destination using past migration. Compared to the US evidence on region’s impact on relative (rather than absolute) intergenerational income (rather than education) mobility of Chetty and Hendren (2018a), sorting in Africa appears higher. However, regional exposure effects are similar. We also explored differences across relatively rich and poor African countries, finding similar in both groups selection and regional exposure effects (Appendix Figure G.3).

6 Conclusion

We conduct a systematic exploration of intergenerational mobility in education across African countries and districts since independence.

In the first part, we compile new estimates of absolute intergenerational mobility in educational attainment across African countries and regions, distinguishing by gender and rural-urban residence. Opportunities for upward mobility vary substantially across the continent and regions in the same country. The literacy of the “old” generation is a strong

predictor of both upward and downward mobility, pointing to inertia and slow convergence. Persistence is more substantial for rural than urban places. Second, we explored the geographic and historical correlates of regional mobility. Upward mobility is higher and downward IM is lower in regions with colonial investments in railroads and those close to Christian, mainly Protestant missions. Distance to the coast and the capital and an ecology favorable to malaria correlate negatively with upward IM and positively with downward IM. Upward mobility is higher in regions that were more developed at independence, with higher urbanization and employment in services-manufacturing. In the third part, we distinguish between spatial sorting and regions' independent influence on educational mobility. We find that both sorting and regional childhood exposure effects are at play. Boys and girls whose families move to regions with higher (lower) upward mobility have a significantly higher (smaller) likelihood to complete primary schooling when the move takes place before the age of 12 (and after 5). This pattern also applies when we compare siblings, look at moves triggered by regional displacement shocks, and use historical migration patterns to predict moving households' destination regions. Thus, regions matter crucially for education in Africa, both because households with a latent propensity to invest in their children's future move to high mobility (high literacy) places and because the environment exerts an independent impact on educational mobility. Regional disparities are wide and unlikely to disappear unless policies specifically target them.

Our analysis here -as well as in our companion papers Alesina et al. (2020b,a) where we study ethnic and religious differences in educational mobility- opens several avenues for future research. A first avenue is to examine the causal effects of historical factors on educational mobility. Such work could combine the newly compiled IM statistics with quasi-experimental variation to explore the economic mechanisms underlying path dependence, including colonial-era investments. A second avenue is to examine the role of nationwide educational policies, like laws on compulsory primary education, and school construction programs (like Figueiredo Walter (2020)) for social mobility, topics largely unexplored in the context of Africa. A third avenue is to construct measures of each region's impact on IM following the approach of Chetty and Hendren (2018b) and explore regional heterogeneity. Fourth, future work should investigate how the diverse set of family structures across Africa mediate the transmission of education from one generation to the next. It is also important to examine differences in the transmission of human capital from mothers, fathers, and other relatives, distinguishing between boys and girls. Fifth, as data on income start to become available, future work could study interconnections between education and income mobility. Sixth, using finer resolution data on income, consumption, and education one could examine their inter-relations and isolate the relative change in children's position in the distribution from the general increase that most African countries have experienced (using for example the bounds approach of Asher, Novosad, and Rafkin (2020)). Finally, one could link the regional statistics to political variables (e.g., electoral competition and participation) and leaders' characteristics, to study jointly regional, ethnic, and religious favoritism and discrimination.

References

- Alesina, Alberto, William Easterly, and Janina Matuszeski. 2011. “Artificial States.” *Journal of the European Economic Association* 9 (2):246–277.
- Alesina, Alberto, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou. 2020a. “Ethnicity and Social Mobility in Africa.” *mimeo* .
- . 2020b. “Religion and Social Mobility in Africa.” *mimeo* .
- Alvaredo, Facundo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2017. “Global Inequality Dynamics: New Findings from WID.World.” *American Economic Review* 107 (5):404–409.
- Asher, Sam, Paul Novosad, and Charlie Rafkin. 2020. “Intergenerational Mobility in India: Estimates from New Methods and Administrative Data.”
- Barro, Robert J. and Jong Wha Lee. 2013. “A New Data Set of Educational Attainment in the World, 1950–2010.” *Journal of Development Economics* 104 (C):184–198.
- Becker, Gary S. and Nigel Tomes. 1979. “An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility.” *Journal of Political Economy* 87 (6):1153–1189.
- Berman, Nicolas, Mathieu Couttenier, Dominic Rohner, and Mathias Thoenig. 2017. “This Mine Is Mine! How Minerals Fuel Conflicts in Africa.” *American Economic Review* 107 (6):1564–1610.
- Black, Sandra E. and Paul J. Devereux. 2011. “Recent Developments in Intergenerational Mobility.” In *Handbook of Labor Economics*, vol. 4B. Elsevier, 1487–1541.
- Blake, Judith. 1985. “Number of Siblings and Educational Mobility.” *American Sociological Review* 50 (1):84–94.
- Bowles, Samuel. 1972. “Schooling and Inequality from Generation to Generation.” *Journal of Political Economy* 80 (3, Part 2):S219–S251.
- Brecke, Peter. 1999. “Violent Conflicts 1400 AD to the Present in Different Regions of the World.” In *1999 Meeting of the Peace Science Society, Unpublished Manuscript*.
- Cagé, Julia and Valeria Rueda. 2016. “The Long-Term Effects of the Printing Press in Sub-Saharan Africa.” *American Economic Journal: Applied Economics* 8 (3):69–99.
- Card, David. 1999. “The Causal Effect of Education on Earnings.” In *Handbook of Labor Economics*, vol. 3, Part A. Elsevier, 1801–1863.
- Card, David, Ciprian Domnisoru, and Lowell Taylor. 2018. “The Intergenerational Transmission of Human Capital: Evidence from the Golden Age of Upward Mobility.” :102.
- Caselli, Francesco, Jacopo Ponticelli, and Federico Rossi. 2014. “A New Dataset on Mincerian Returns.” *unpublished* .

- Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2020a. “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.” *NBER Working Papers* 25147.
- Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. 2017. “The Fading American Dream: Trends in Absolute Income Mobility since 1940.” *Science* 356 (6336):398–406.
- Chetty, Raj and Nathaniel Hendren. 2018a. “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects.” *The Quarterly Journal of Economics* 133 (3):1107–1162.
- . 2018b. “The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates.” *The Quarterly Journal of Economics* 133 (3):1163–1228.
- Chetty, Raj, Nathaniel Hendren, Maggie R. Jones, and Sonya Porter. 2020b. “Race and Economic Opportunity in the United States: An Intergenerational Perspective.” *The Quarterly Journal of Economics* 135 (2):711–783.
- Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. “Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States.” *The Quarterly Journal of Economics* 129 (4):1553–1623.
- Davis, Jonathan and Bhashkar Mazumder. 2018. “Racial and Ethnic Differences in the Geography of Intergenerational Mobility.” *SSRN* 3138979.
- . 2020. *The Decline in Intergenerational Mobility After 1980*.
- Derenoncourt, Ellora. 2018. “Can You Move to Opportunity? Evidence from the Great Migration.” *Harvard University, mimeo* .
- Economist, The. 2000. “Hopless Africa.” *The Economist* .
- . 2011. “Africa Rising. The Hopeful Continent.” *The Economist* .
- . 2016. “1.2 Billion Opportunities.” *The Economist* .
- FAO. 2017. “Evidence on Internal and International Migration Patterns in Selected African Countries.” .
- Figueiredo Walter, Torsten. 2020. “Misallocation of State Capacity? Evidence from Two Million Primary Schools.” *mimeo* .
- Fletcher, Jason and Joel Han. 2018. “Intergenerational Mobility in Education: Variation in Geography and Time.” Working Paper 25324, National Bureau of Economic Research.
- Gallup, John Luke and Jeffrey D. Sachs. 2001. *The Economic Burden of Malaria*. American Society of Tropical Medicine and Hygiene.
- Geng, Yining. 2018. “Intergenerational Mobility in China across Space and Time.” *Universitat Pompeu Fabra, mimeo* .

- Gottschalk, Peter and Enrico Spolaore. 2002. "On the Evaluation of Economic Mobility." *The Review of Economic Studies* 69 (1):191–208.
- Greene, William H. 2011. *Econometric Analysis*. Boston: Pearson, 7 edition ed.
- Güell, Maia, Michele Pellizzari, Giovanni Pica, and José V. Rodríguez Mora. 2018. "Correlating Social Mobility and Economic Outcomes." *The Economic Journal* 128 (612):F353–F403.
- Gunning, Jan Willem and Paul Collier. 1999. "Explaining African Economic Performance." *Journal of Economic Literature* 37 (1):64–111.
- Hertz, Tom, Tamara Jayasundera, Patrizio Piraino, Sibel Selcuk, Nicole Smith, and Alina Verashchagina. 2008. "The Inheritance of Educational Inequality: International Comparisons and Fifty-Year Trends." *The B.E. Journal of Economic Analysis & Policy* 7 (2).
- Hilger, Nathaniel. 2017. "The Great Escape: Intergenerational Mobility in the United States, 1930-2010." *unpublished manuscript* .
- Hohmann, Sebastian. 2018a. "Internal Migration in Africa: Evidence from Census Data." *mimeo* .
- . 2018b. "Natural Resources, Trade, and Structural Transformation in Africa." *unpublished manuscript* .
- Jedwab, Remi and Alexander Moradi. 2016. "The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa." *The Review of Economics and Statistics* 98 (2):268–284.
- Jedwab, Remi, Felix Meier zu Selhausen, and Alexander Moradi. 2018. "The Economics of Missionary Expansion: Evidence from Africa and Implications for Development." Tech. Rep. 2018-07, Centre for the Study of African Economies, University of Oxford.
- Kanbur, Ravi and Anthony J. Venables. 2005. "Spatial Inequality and Development: Overview of UNU-WIDER Project." *Cornell University, Department of Applied Economics and Management Working Papers* 127127.
- Loury, Glenn C. 1981. "Intergenerational Transfers and the Distribution of Earnings." *Econometrica* 49 (4):843–867.
- McMillan, Margaret, Dani Rodrik, and Inigo Verduzco-Gallo. 2014. "Globalization, Structural Change, and Productivity Growth, with an Update on Africa." *World Development* 63 (C):11–32.
- Michalopoulos, Stelios and Elias Papaioannou. 2013. "Pre-Colonial Ethnic Institutions and Contemporary African Development." *Econometrica* 81 (1):113–152.
- . 2014. "National Institutions and Subnational Development in Africa." *The Quarterly Journal of Economics* 129 (1):151–213.

- . 2020. “Historical Legacies and African Development.” *Journal of Economic Literature* 58 (1):53–128.
- Mogstad, Magne, Joseph P. Romano, Azeem Shaikh, and Daniel Wilhelm. 2020. “Inference for Ranks with Applications to Mobility across Neighborhoods and Academic Achievement across Countries.” *NBER Working Papers* 26883.
- Montenegro, Claudio E. and Harry Anthony Patrinos. 2014. “Comparable Estimates of Returns to Schooling around the World.” Tech. Rep. WPS7020, The World Bank.
- Murdock, George Peter. 1967. “Ethnographic Atlas: A Summary.” *Ethnology* 6 (2):109–236.
- Narayan, Ambar, Roy Van der Weide, Alexandru Cojocaru, Silvia Redaelli, Christoph Lakner, Daniel Gerszon Mahler, Rakesh Ramasubbaiah, and Stefan Thewissen. 2018. *Fair Progress?: Economic Mobility Across Generations Around the World*. World Bank.
- Nunn, Nathan. 2008. “The Long-Term Effects of Africa’s Slave Trades.” *The Quarterly Journal of Economics* 123 (1):139–176.
- . 2010. “Religious Conversion in Colonial Africa.” *American Economic Review* 100 (2):147–152.
- Nunn, Nathan and Diego Puga. 2012. “Ruggedness: The Blessing of Bad Geography in Africa.” *The Review of Economics and Statistics* 94 (1):20–36.
- Psacharopoulos, George. 1994. “Returns to Investment in Education: A Global Update.” *World Development* 22 (9):1325–1343.
- Psacharopoulos, George and Harry Anthony Patrinos. 2004. “Returns to Investment in Education: A Further Update.” *Education Economics* 12 (2):111–134.
- Solon, Gary. 1999. “Intergenerational Mobility in the Labor Market.” In *Handbook of Labor Economics*, vol. 3, Part A. Elsevier, 1761–1800.
- Spady, William G. 1967. “Educational Mobility and Access: Growth and Paradoxes.” *American Journal of Sociology* 73 (3):273–286.
- United Nations Conference on Trade and Development. 2018. *Economic Development in Africa Report 2018. Migration for Structural Transformation*. United Nations.
- Wantchekon, Leonard. 2019. “The Strength of Weak Ties: Educational Externalities and Intergenerational Mobility in Benin.” *Princeton University, mimeo* .
- Wantchekon, Leonard, Marko Klašnja, and Natalija Novta. 2015. “Education and Human Capital Externalities: Evidence from Colonial Benin.” *The Quarterly Journal of Economics* 130 (2):703–757.
- Young, Alwyn. 2012. “The African Growth Miracle.” *Journal of Political Economy* 120 (4):696–739.
- . 2013. “Inequality, the Urban-Rural Gap, and Migration.” *The Quarterly Journal of Economics* 128 (4):1727–1785.

Intergenerational Mobility in Africa

Supplementary Online Appendix

Alberto Alesina*
Harvard University, IGIER, CEPR and NBER

Sebastian Hohmann†

Stelios Michalopoulos‡
Brown University, CEPR and NBER

Elias Papaioannou§
London Business School and CEPR

This draft: July 2020

Abstract

The Online Appendix provides additional evidence and sensitivity checks. Section A shows the distribution of schooling by country-birth-decade. Section B details the sample construction process. In Section C we report validation checks of the IPUMS data on education with alternative cross-country and regional statistics. We also explore the relationship between educational attainment and various proxies of well-being across African regions using all available Demographic and Health as well as Afrobarometer Surveys. Section D provides details on family structure and cohabitation patterns across the millions of African households in our dataset. In Section E we report and discuss intergenerational mobility (IM) statistics, distinguishing between gender and rural-urban household residence. In Section F we present variable definitions and LASSO estimates for the correlational analysis where we associate educational mobility to at-independence, historical, and geographical features. In section G we describe sensitivity checks and additional results pertaining to the regional childhood exposure estimates.

Keywords: Africa, Development, Education, Inequality, Intergenerational Mobility.

JEL Numbers. N00, N9, O10, O43, O55

*Harvard University, IGIER, CEPR and NBER aalesina@harvard.edu <https://scholar.harvard.edu/alesina/home>

†sebas.hohmann@gmail.com

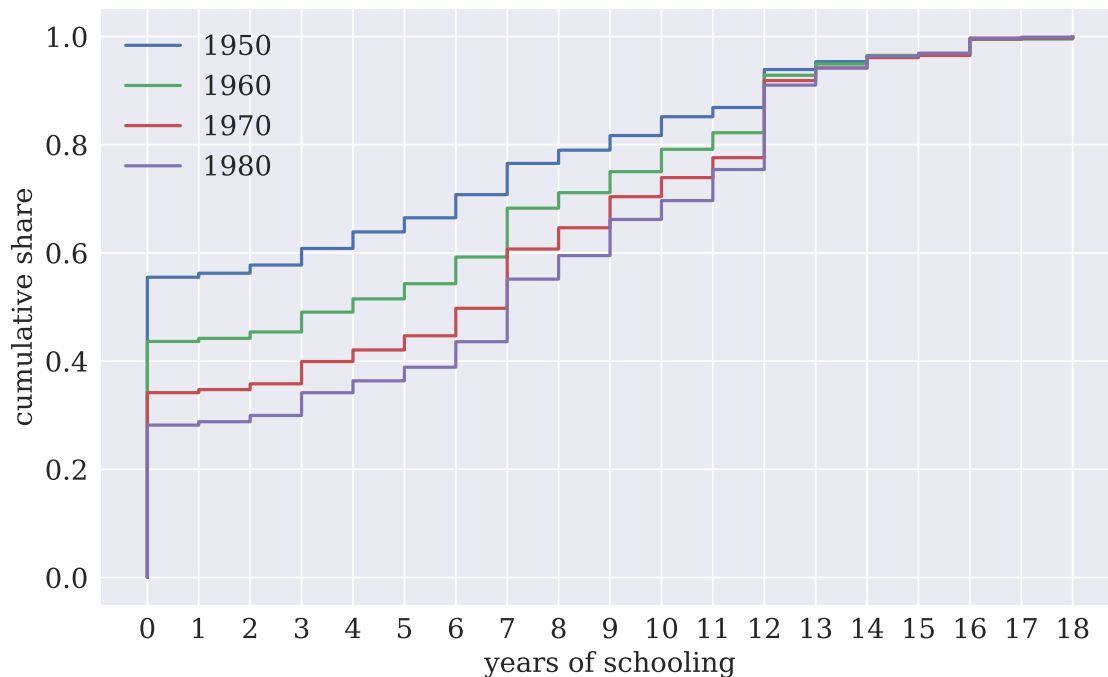
‡Brown University, CEPR and NBER smichalo@brown.edu <https://sites.google.com/site/stelioecon/>

§London Business School and CEPR eliasp@london.edu <https://sites.google.com/site/papaioannouelias/>

A Education Data and Statistics

This section presents the distribution of years of schooling by birth-cohort in Africa without imposing any sample restrictions, related to observing younger and older generation in the same household. We include all individuals aged 25 and older at the time of the census, using all available censuses for a given country.

Figure A.1: Cumulative distribution of years of schooling by birth decade



This figure shows the cumulative distribution function for years of schooling for four birth decades since 1950 for ages 25+. Note that the sample is unbalanced in the sense that not all countries in the sample have a census such that each country is represented in every birth decade.

B Sample Construction

The table below gives the number of observations for each country-census, detailing how we compile the sample that we use in the main analysis from the raw IPUMS data (see please the Table notes).

Table B.1: Sample construction

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
country	year	fraction	N_{all}	N_{age}	N_{owned}	$N_{age \geq 14}$ <i>owned</i>	$N_{14 \leq age \leq 25}$ <i>owned</i>	$N_{14 \leq age \leq 18}$ <i>owned</i>	$N_{age \geq 14}$ <i>olded</i>	$N_{14 \leq age \leq 25}$ <i>olded</i>	$N_{14 \leq age \leq 18}$ <i>olded</i>	$N_{age \geq 14}$ <i>olded, no mqhh</i>	$N_{14 \leq age \leq 25}$ <i>olded, no mqhh</i>	$N_{14 \leq age \leq 18}$ <i>olded, no mqhh</i>
Benin	1979	10	331,049	329,784	244,898	171,690	62,112	24,322	72,597	37,048	18,381	39,210	30,663	17,044
Benin	1992	10	498,419	498,107	435,652	256,763	101,543	44,735	109,234	65,432	35,000	57,957	56,106	33,190
Benin	2002	10	685,467	685,467	612,658	373,452	155,832	69,048	160,458	104,331	57,364	87,164	90,874	54,655
Benin	2013	10	1,009,693	1,009,693	911,604	559,525	240,049	108,694	244,182	170,580	93,329	141,282	148,917	88,090
Botswana	1981	10	97,238	96,187	72,951	50,399	20,258	9,533	19,054	12,165	6,847	12,048	10,821	6,413
Botswana	1991	10	132,623	132,623	113,172	78,814	32,680	15,830	26,842	18,264	10,979	16,608	15,573	10,013
Botswana	2001	10	168,676	168,134	159,257	109,649	44,806	20,616	42,629	29,159	16,102	26,541	25,017	14,834
Botswana	2011	10	201,752	201,235	190,212	138,375	48,926	20,677	47,177	28,874	14,302	29,888	24,861	13,288
Burkina Faso	1985	10	884,797	883,447	484,384	410,398	159,162	75,374	0	0	0	0	0	0
Burkina Faso	1996	10	1,081,046	1,075,824	803,264	552,402	226,436	114,148	250,977	157,907	95,710	119,954	127,332	88,869
Burkina Faso	2006	10	1,417,824	1,410,123	1,244,291	770,161	321,384	151,393	327,195	211,275	123,364	153,253	167,136	112,927
Cameroon	1976	10	736,514	736,320	605,749	413,814	157,287	72,886	153,664	93,046	52,858	94,677	90,900	50,900
Cameroon	1987	10	897,211	896,649	763,652	481,727	191,552	90,805	178,841	114,345	67,018	117,877	104,219	65,397
Cameroon	2005	10	1,772,359	1,772,359	1,542,200	1,018,632	438,407	199,054	433,774	299,540	164,047	261,360	257,094	154,003
Egypt	1986	14.1	6,799,093	6,794,386	5,418,332	4,262,426	1,609,719	722,024	1,931,495	1,345,174	693,323	1,397,036	1,276,970	679,214
Egypt	1996	10	5,902,243	5,901,839	4,453,382	3,810,835	1,471,285	718,874	1,616,808	1,230,963	695,795	1,302,668	1,211,167	692,249
Egypt	2006	10	7,282,434	7,282,434	5,739,722	5,096,618	1,977,932	785,619	2,046,232	1,590,965	759,450	1,706,597	1,568,687	756,807
Ethiopia	1984	10	3,404,306	3,398,027	2,733,575	1,800,650	620,022	303,780	556,877	360,470	234,377	397,880	344,904	230,979
Ethiopia	1994	10	5,044,598	5,044,597	4,201,616	2,833,214	1,224,762	614,179	1,034,238	788,117	498,607	743,852	765,183	494,430
Ethiopia	2007	10	7,434,086	7,434,086	1,097,614	744,744	331,544	161,226	259,645	200,774	128,818	187,293	190,833	126,197
Ghana	1984	10	1,309,352	1,309,351	1,050,813	747,642	302,953	142,526	340,180	219,918	121,103	174,954	182,110	110,717
Ghana	2000	10	1,894,133	1,894,133	1,730,902	1,152,128	434,882	200,000	489,201	309,485	167,556	309,535	282,693	161,277
Ghana	2010	10	2,466,289	2,466,289	2,262,894	1,575,528	603,020	270,162	636,999	424,323	229,128	404,083	380,287	217,963
Guinea	1983	10	457,837	457,778	364,805	275,065	99,816	44,129	120,722	68,409	33,806	61,363	57,736	32,395
Guinea	1996	10	729,071	727,246	551,619	397,137	148,064	69,165	202,166	113,623	58,386	90,316	87,334	52,520
Guinea	1969	6	659,310	659,310	659,310	394,835	167,003	67,260	107,690	73,634	42,565	74,096	68,169	41,487
Kenya	1979	6.7	1,033,769	1,031,996	853,843	593,682	267,515	132,599	0	0	0	0	0	0
Kenya	1989	5	1,074,098	1,072,777	828,512	578,099	259,837	125,884	220,398	167,963	103,453	168,343	163,583	102,768
Kenya	1999	5	1,407,547	1,407,547	1,191,268	832,083	378,922	176,867	293,668	229,678	137,385	237,572	224,347	136,227
Kenya	2009	10	3,841,935	3,841,935	3,402,695	2,246,737	955,548	432,424	779,027	599,114	354,687	583,560	560,842	344,093
Lesotho	1996	10	187,795	187,795	165,960	121,446	50,160	24,283	57,938	40,652	21,687	41,310	37,454	20,514
Lesotho	2006	10	180,208	180,208	171,947	123,644	50,609	22,361	52,223	37,257	18,405	39,656	34,511	17,796
Liberia	1974	10	150,256	150,256	127,442	91,811	34,393	16,014	0	0	0	0	0	0
Liberia	2008	10	348,057	348,057	294,517	210,111	87,459	38,854	86,523	60,197	32,411	58,878	55,981	31,437
Malawi	1987	10	798,669	798,193	657,998	447,247	176,370	81,029	122,477	89,555	55,837	101,627	88,444	55,755
Malawi	1998	10	991,393	991,393	826,197	582,694	251,873	114,846	170,361	131,703	83,465	140,075	128,509	82,947
Malawi	2008	10	1,341,977	1,341,046	1,161,773	736,175	307,167	135,833	216,020	170,362	108,154	179,356	166,624	107,806
Mali	1987	10	785,384	773,407	582,678	422,837	162,820	76,364	191,000	117,544	63,569	106,596	104,041	61,368
Mali	1998	10	991,330	986,822	734,156	519,001	207,852	102,961	239,952	151,320	85,538	126,625	132,131	81,840
Mali	2009	10	1,451,856	1,424,140	1,262,277	776,333	326,105	158,458	366,650	242,131	135,111	174,224	198,373	124,732
Morocco	1982	5	1,012,873	1,012,873	948,008	571,980	242,307	115,031	294,399	204,348	106,327	201,143	201,143	106,113
Morocco	1994	5	1,294,026	1,293,171	1,293,171	842,330	322,163	149,529	447,785	283,778	141,311	308,463	269,180	138,034
Morocco	2004	5	1,482,720	1,481,076	1,481,076	1,052,531	363,627	161,892	563,135	328,002	155,722	406,132	314,978	153,357
Mozambique	1997	10	1,551,517	1,550,505	1,248,483	879,255	370,427	167,753	276,001	205,248	126,675	200,488	188,002	121,938

Table B.1: Sample construction, continued

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
country	year	fraction	N_{all}	N_{age}	N_{owned}	$N_{age \geq 14}$ <i>owned</i>	$N^{14 \leq age \leq 25}$ <i>owned</i>	$N^{14 \leq age \leq 18}$ <i>owned</i>	$N^{age \geq 14}$ <i>olded</i>	$N^{14 \leq age \leq 25}$ <i>olded</i>	$N^{14 \leq age \leq 18}$ <i>olded</i>	$N_{age \geq 14}$ <i>olded, no mghh</i>	$N^{14 \leq age \leq 25}$ <i>olded, no mghh</i>	$N^{14 \leq age \leq 18}$ <i>olded, no mghh</i>
Mozambique	2007	10	2,047,048	2,047,048	1,616,853	1,103,596	439,299	193,512	340,331	250,911	150,964	252,302	231,567	145,429
Nigeria	2006	.06	83,700	83,700	82,740	49,282	18,063	8,803	20,096	13,901	7,991	11,172	12,308	7,667
Nigeria	2007	.06	85,183	85,182	84,122	49,102	18,013	8,811	20,104	13,875	7,990	11,727	12,389	7,709
Nigeria	2008	.07	107,425	107,425	105,944	62,151	23,183	11,453	26,700	18,548	10,570	15,708	16,985	10,279
Nigeria	2009	.05	77,896	77,880	77,650	45,988	16,676	8,050	17,545	12,180	6,815	10,302	10,547	6,482
Nigeria	2010	.05	72,191	71,991	58,973	41,830	15,485	7,534	17,736	12,400	6,857	10,988	11,639	6,748
Rwanda	1991	10	742,918	742,918	535,602	372,386	146,839	71,287	121,757	98,452	59,490	101,356	93,882	58,192
Rwanda	2002	10	843,392	843,392	629,146	473,714	221,106	109,367	160,426	139,716	87,351	137,927	132,864	85,472
Rwanda	2012	10	1,038,369	1,038,369	938,201	624,155	250,162	112,248	206,411	169,300	95,112	176,081	161,473	93,342
Senegal	1988	10	700,199	699,981	527,462	378,289	153,541	68,971	200,537	118,387	58,295	100,231	104,889	56,535
Senegal	2002	10	994,562	994,562	911,891	594,599	260,317	124,706	374,034	223,678	112,749	150,679	178,191	101,982
Sierra Leone	2004	10	494,298	492,922	395,788	291,916	120,773	55,346	148,389	92,052	48,263	71,802	72,943	43,093
South Africa	1996	10	3,621,164	3,578,019	3,055,995	2,328,067	840,077	376,601	883,678	594,898	308,241	614,616	538,065	295,656
South Africa	2001	10	3,725,655	3,725,655	3,353,684	2,598,672	915,973	421,066	1,014,988	673,692	351,871	686,654	607,967	336,252
South Africa	2007	2	1,047,657	1,047,657	842,103	665,305	233,345	105,048	268,806	177,267	88,805	181,994	158,798	84,647
South Africa	2011	8.6	4,418,594	4,418,594	3,845,633	3,101,908	1,020,126	422,182	1,104,821	709,491	347,343	790,453	641,430	331,670
South Sudan	2008	7	542,765	542,765	542,333	295,567	120,639	57,922	132,069	89,636	49,765	91,682	83,835	48,071
Sudan	2008	16.6	5,066,530	5,066,530	3,902,071	2,790,992	1,197,729	560,663	1,207,911	843,100	476,833	831,591	799,231	466,630
Tanzania	1988	10	2,310,424	2,304,474	1,911,308	1,322,841	556,836	278,218	483,701	336,938	210,110	349,716	325,632	208,887
Tanzania	2002	10	3,732,735	3,732,735	3,123,724	2,190,557	903,114	416,283	684,743	479,128	293,410	482,029	446,969	284,588
Tanzania	2012	10	4,498,022	4,498,022	3,918,823	2,603,099	1,036,707	491,497	897,469	641,322	382,527	622,879	586,677	366,998
Togo	1960	10	13,759	13,759	13,758	7,842	3,329	1,374	2,930	1,619	775	1,293	1,265	704
Togo	1970	1	23,680	23,617	23,609	12,262	4,146	1,627	4,186	2,216	1,113	2,440	1,901	1,042
Togo	2010	10	584,859	584,859	517,900	339,447	132,399	58,429	131,016	85,981	46,328	98,071	80,277	45,213
Uganda	1991	10	1,548,460	1,547,604	1,242,885	855,537	378,505	179,263	282,250	205,997	128,937	215,330	199,233	127,667
Uganda	2002	10	2,497,449	2,497,449	2,042,838	1,355,857	601,101	289,123	438,634	337,594	223,066	328,581	319,184	217,556
Zambia	1990	10	787,461	787,461	664,239	460,486	216,756	108,294	209,029	157,907	94,631	159,787	155,451	94,493
Zambia	2000	10	996,117	996,117	825,110	570,022	259,096	119,089	235,708	177,260	102,847	151,980	151,774	95,117
Zambia	2010	10	1,321,973	1,321,973	1,028,628	704,471	307,786	147,933	263,135	205,511	126,500	178,608	177,748	117,433
Zimbabwe	2012	5	654,688	653,276	587,010	397,356	157,602	74,305	120,052	86,481	52,191	89,333	79,337	49,874
total			117,902,003	117,764,561	93,086,552	66,791,915	26,575,313	12,238,046	25,805,626	18,096,111	10,085,172	18,018,282	16,819,190	9,788,011

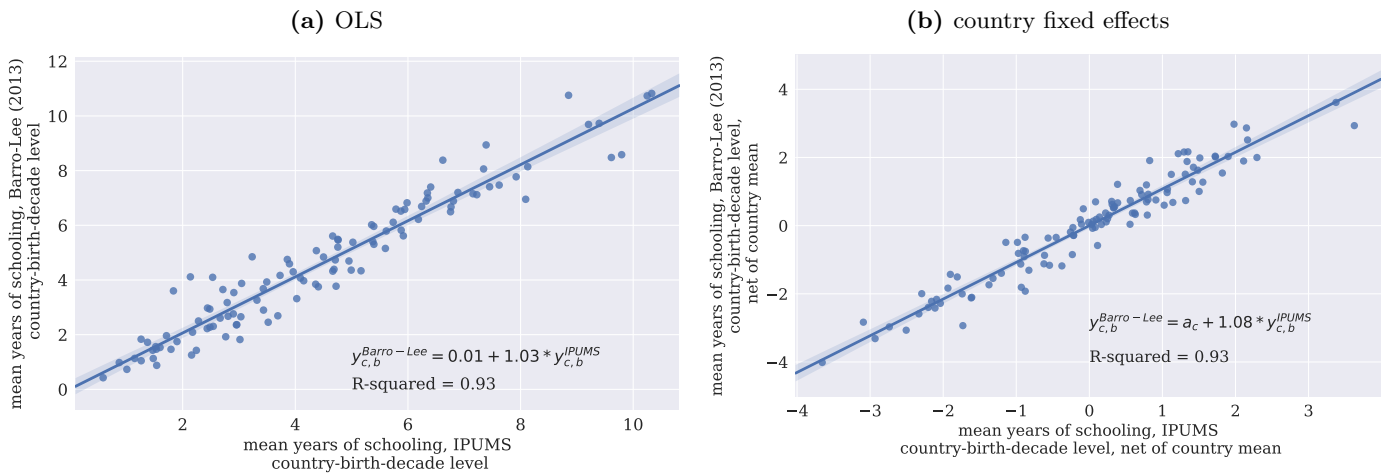
The table shows how we proceed from the raw IPUMS data to the sample employed in our analysis. Columns (1) and (2) give the country and census year, respectively. Column (3) reports coverage, the fraction of the census obtained by IPUMS. Column (4) gives the total number of observations in the IPUMS data. Column (5) gives the number of observations with information on individual's age. Column (6) gives the number of observations for which IPUMS reports both the age and education. Column (7) gives the number of observations/individuals, older than 14 years old with available information on age and education. Column (8) gives the number of observations/individuals in the 14-25 age range with available information on age and education, while column (9) reports the corresponding number for individuals in the 14-18 age range. Columns (10)-(12) are similar to (7)-(9), but also require that IPUMS reports the age and education of at least one member of the older generation. Columns (13)-(15) are similar to (10)-(12), but omit individuals residing in multigenerational households.

C Data Validation and Relevance of Education

Section C.1 validates the IPUMS data across country-cohorts with the widely-used Barro and Lee (2013) statistics and at the regional level using 109 geo-referenced Demographic and Health Surveys from 22 countries. Section C.2 reports the correlational analysis of a variety of “good outcomes” and educational attainment. The correlational analysis uses all georeferenced DHS surveys (3,457,367 observations from 134 surveys in 33 countries) and all geo-referenced rounds of the Afrobarometer surveys (246,723 observations from 37 countries).

C.1 Validation IPUMS Education Data

Figure C.1: IPUMS and Barro-Lee (2013) Years of Schooling



This figure plots years of schooling at the country-birth-decade level derived from Barro and Lee (2013) on the equivalent measure computed from the IPUMS data for all cohorts since 1940 for individuals aged 20+, as this is age reported by Barro and Lee (2013). The figure also reports the OLS regression fit.

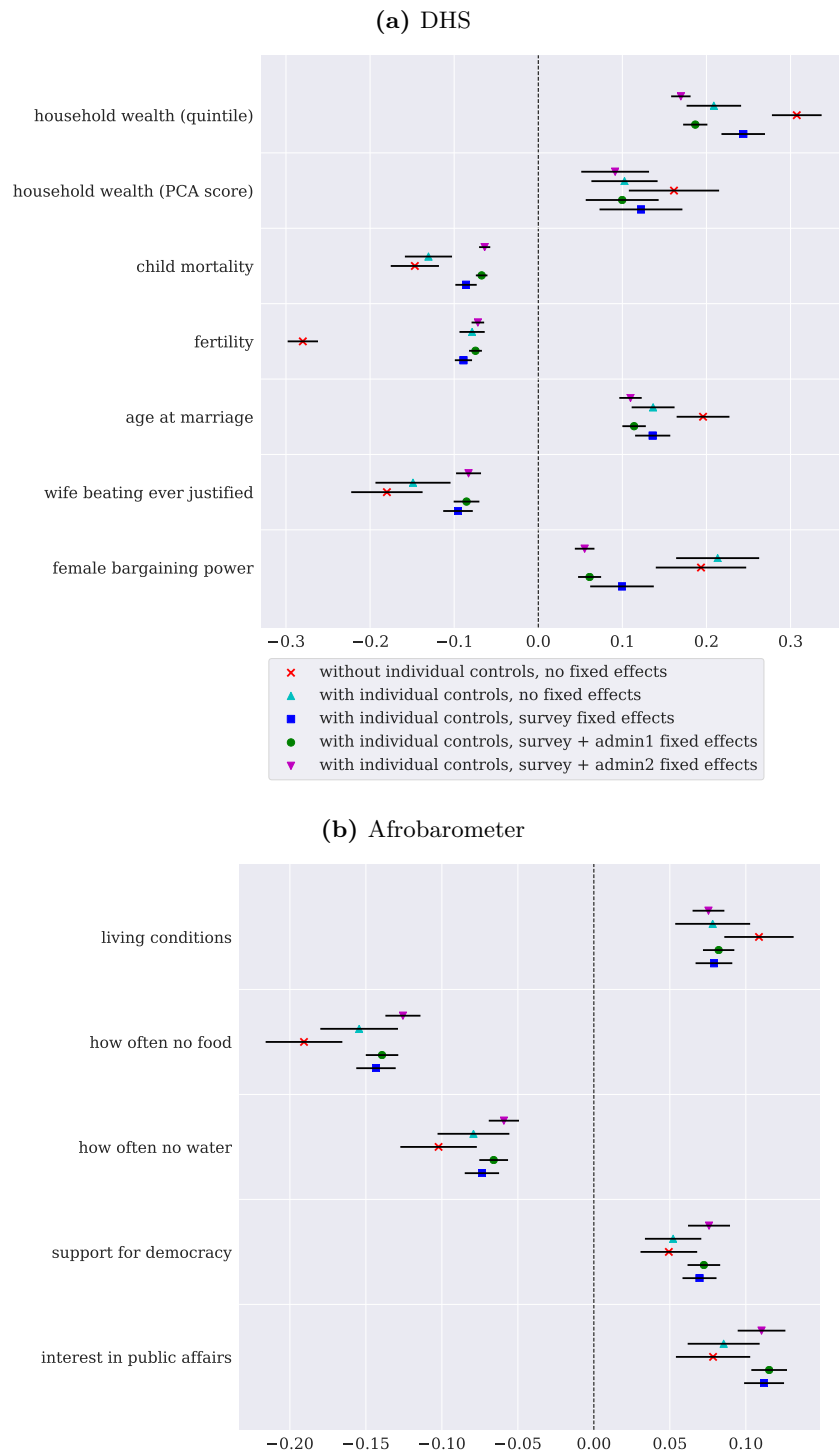
Table C.1: IPUMS and DHS years of schooling, 1980s birth cohort

	(1)	(2)	(3)
	ipums	ipums	ipums
dhs	0.986***	0.997***	0.948***
	(0.0446)	(0.0256)	(0.0270)
level	country	province	district
R-squared	0.861	0.933	0.895
N	22	306	2335

This table shows results for regressions of mean years of schooling computed for individuals aged 18+ for three different levels of aggregation – country, province (admin-1) and district (admin-2). The LHS variable in all regressions is the variable computed from IPUMS census data. The RHS variable is the corresponding measure computed from DHS survey data. Column (1) shows the country-level, column (2) the province level, and column (3) the district-level result. Standard errors [clustered at the country-level in columns (2) and (3)] in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

C.2 Education and “Good Outcomes”

Figure C.2: Correlation of at-least primary educational attainment and good outcomes



The figures report OLS regression results (standardized coefficients) associating various outcomes reported in DHS (panel (a), 3,457,367 observations, 134 surveys, 33 countries) and Afrobarometer surveys (panel (b), 246,723 observations, rounds 1-7, 37 countries) on an indicator variable that takes the value of one for individuals who have completed at least least primary educational attainment. The figure also plots 95% confidence bands, based on double-clustered standard errors at the survey (=country×year) and at the admin-1 levels. In the DHS specifications (panel (a)) the set of individual controls consists of age, age-squared, the natural logarithm of the number of household members, a gender dummy, an indicator for male household head, and an urban location dummy. In the Afrobarometer Survey specifications (panel (b)) the set of individual controls consists of age, age-squared, a gender dummy, and an urban location indicator.

D Household Structure and Cohabitation Selection

In this Section we describe the estimation of cohabitation rates and, alongside, we provide an overview of the household structure across countries and censuses. We estimate intergenerational mobility (IM) for boys and girls (“children”) between ages 14 and 18. We choose this age interval because children in this age range (i) have most likely completed primary schooling –the relevant educational cutoff for Africa in the period we study (Figure A.1)– and (ii) still largely cohabit with their parents and other older generation relatives. We also compile IM statistics for (young) individuals aged 14-25, as this doubles the sample size, while cohabitation rates remain still reasonably high.

D.1 Household Structure

To estimate IM we need to observe the educational attainment of the young generation (14 – 18 or 14 – 25-year-old) and that of the immediately older generation in a given household.

The starting point for the intergenerational matching is census information for each individual’s “*relationship to household head*” (we denote the head’s generation as $G(0)$). Across the different censuses this variable can take as many as three dozen distinct values including: children, parents, grandparents, nieces and nephews, grandchildren, siblings, in-laws, etc. Some censuses provide fine classifications, distinguishing between biological children, adopted, foster, and step-children, as, for example, the post-1995 censuses in South Africa, Cameroon, Burkina Faso, Botswana, and Zambia. However, most censuses have a single “children” category. Likewise, some censuses distinguish clearly between nieces/nephews, uncles/aunts, cousins, sibling in-laws, in-laws whereas others report a composite “*other relatives (not elsewhere classified)*” category that subsumes various family relationships.¹ Illustrative examples include the 1976 and the 1987 Censuses in Cameroon that, beside the household head, only identify clearly the spouse, children, parents, and then all other possible family relationships appear in a single “*other relatives*” category. In an even more extreme case, the 1991 census in Botswana and the 1988 census in Senegal report a single category of “*other relatives and non relatives*”.

Using the “*relationship to household head*” we assign household members to generations.

- $G(-2)$, that we rarely observe, consists of the household head’s grandparents and grand-uncles/aunts.
- $G(-1)$ consists of the household head’s parents, parents-in-law, and uncles/aunts (in-law).
- $G(0)$ consists of the household head, typically male, as well as his spouse, siblings, cousins, siblings and cousins in-law (in the censuses with fine disaggregation).
- $G(1)$ consists of children (biological, adopted, etc.) as well as nieces and nephews.
- $G(2)$ consists of the head’s grandchildren (and in a few occasions grandnephews/nieces).

¹The level of disaggregation also differs among non-relatives who are part of the household. These individuals cannot be matched and so count as not co-resident.

Individuals classified as “*other relatives*” cannot be unequivocally assigned to generations; the prevalence of this category depends (partially) on the census-specific detail. One extreme case is that of the 2010 Census in Togo, where half of 14–18 year old are classified as “*other relatives*”. To avoid misclassification due to differences in census coarseness, we use the age of those classified as “*other relatives*” to assign them to generations. Specifically, we count individuals as being one generation apart if they are in-between 15 and 40 years older.

D.2 Household Arrangements for 14-18 Year Old

Table D.1 reports the relationship to household head for all 14 – 18 years old individuals in each census. In total there are 13,005,949 individuals. The first column reports the percentage of those between 14 and 18 years of age that we can assign to a generation using only censuses’ classification on “*relationship to household head*”, i.e., without using their age. The remaining columns give the breakdown across generations ($G(0)$, $G(1)$, $G(2)$), further distinguishing by the main categories of “*relationship to household head*”. The last row gives the pan-African averages (across 13 million observations). The following patterns emerge.

First, about 70% are children of the household head, and so belong to generation $G(1)$. An additional 1% are nieces or nephews (of the household head).

Second, 4.8% are grandchildren of the household head, $G(2)$; approximately half of them cohabit with one of their parents, who is the son or daughter of the household head (not reported). Some of the grandchildren also reside with an uncle, aunt, or another relative of the older generation, $G1$.

Third, a few, around 2%, of 14 – 18 year old, almost exclusively men, appear as household heads. An additional 4.8% are spouses, reflecting the low age of female marriage in some countries, such as Mali, Burkina Faso, Mozambique, Cameroon, and Guinea. For about 10% of these individuals (who are in $G(0)$), at least one of their parents lives in the household, ($G(-1)$).

Fourth, 10% are classified as “other relatives”. As such we cannot directly assign them to generations; they could, for example, be nieces/nephews or son/daughter in laws, but also young cousins, siblings, etc.

Fifth, around 2.5% are unrelated to the household head. 1% are “friends”, “employees”, “visitors”, or live in “group quarters”. 1.5% are assigned to a catch-all “non-relatives” category.

D.3 Cohabitation Rates

We then estimate cohabitation rates for each country-census. Table D.2 gives the estimates. Column (1) reports the number of observations. In total, we have about 13 million 14 – 18 year old across the 68 censuses spanning 26 countries plus the 5 household surveys from Nigeria. The sample jumps to 28,154,990 for the sample of 14 – 25 years old. Column (2) shows that 94.5% of 14 – 18 year olds cohabit with some relative(s); the corresponding share is 92.6% in the larger sample.

Column (3) shows the share of 14–18 year old individuals, who cohabit with at least

one relative of the immediately older generation (assigned either with the “relationship to head” classification or using the age of “other relatives”). 83.8% of 14 – 18 year olds cohabit with an older generation relative. Coresidence rates exceed 90% in Egypt, Morocco, and Nigeria. For most censuses cohabitation rates hover between 80% and 90%. As the quality of census information is worse in the early decades, the lowest cohabiting rate is in Kenya in 1969 (63.3%); in the 1989, 1999, and 2009 Kenyan censuses cohabitation rates hover around 80%. For all other censuses cohabitation rates exceed 69%. Cohabitation rates in the 14 – 25 sample are, unsurprisingly lower, on average 69%. Setting aside the Kenyan census of 1969, cohabitation rates for 14 – 25 year old individuals range from about 55% to 85%.

Appendix Figures D.1 (a)-(b) zoom on the family arrangements of young individuals who reside with at least one relative from an older generation, i.e., the sample on which IM rates are constructed. Roughly half (51% and 47%) cohabit with both their mother and father and an additional 8% – 9% coreside with both parents and other relatives of the immediately older generation (e.g., uncles and aunts). 25% cohabit with one parent, mostly the mother, though often other older generation relatives are present. About 16% – 20% live with other relatives only.

Appendix Figure D.2 (a) reports the distribution of cohabitation rates across the 27 countries for the 14 – 18 age sample. Cross-country cohabitation rates average 82.5%; the standard deviation is quite small. The countries with the highest cohabitation rates are Egypt (96.5%) and Morocco (94.5%) while the countries with the lowest cohabitation rates are Zimbabwe (70.5%) and Botswana (72.9%).

Appendix Figure D.2 (b) plots the distribution of cohabitation rates across 2,846 districts (mostly admin-2 units), pooling across all censuses. The mean (median) is 82.6% (82.5%). The standard deviation is less than a tenth of the mean, 0.09. While there are some outliers, mostly in regions with limited coverage, the p_{10} – p_{90} range is 0.71 – 0.95.

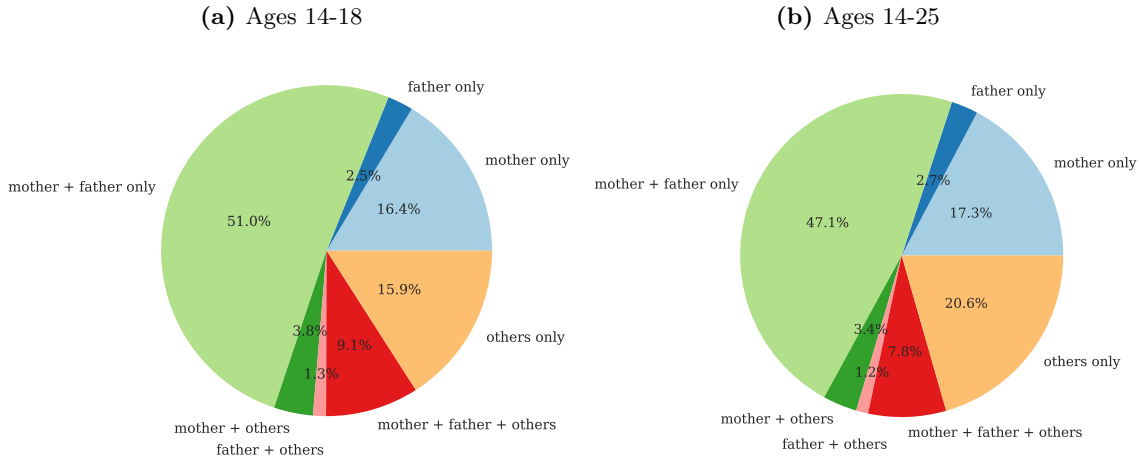
Appendix Figure D.2 plots the evolution of cohabitation rates across districts for the 1970, 1980, and 1990 cohorts. Panel (a) looks at an unbalanced sample of districts. Panel (b) looks at a balanced sample of 1,760 districts from 17 countries for which we have data for all three cohorts. The average coresidence rate increases slightly from 0.804 for the 1970 cohort, to 0.82 for the 1980 cohort and to 0.855 for the 1990-cohort. The increase in the median is smaller (from 0.82 to 0.855), as in the earlier cohorts there are some outliers. The standard deviation of cohabitation rates also falls and over time the distribution becomes less skewed.

Table D.2: Cohabitation rates by census, ages 14-18 and 14-25

iso	year	Ages 14-18			Ages 14-25		
		(1) N	(2) living with relative	(3) living w/ old or fam mbr 15-40 yrs older	(4) N	(5) living with relative	(6) living w/ old or fam mbr 15-40 yrs older
BEN	1979	24,439	90.0	75.9	62,680	88.0	59.9
BEN	1992	46,233	87.9	77.9	105,192	87.5	64.4
BEN	2002	70,033	92.6	83.0	158,251	90.7	67.0
BEN	2013	108,694	94.9	85.9	240,049	93.5	71.1
BFA	1996	122,125	95.7	84.6	250,759	95.8	71.8
BFA	2006	154,448	94.2	81.6	329,175	94.2	66.1
BWA	1981	9,798	87.9	75.0	21,736	82.2	63.9
BWA	1991	15,831	82.3	69.4	32,681	73.7	55.9
BWA	2001	20,650	93.9	78.3	44,939	87.2	65.2
BWA	2011	20,693	79.4	69.3	49,130	79.0	59.1
CMR	1976	75,676	87.5	73.3	163,629	85.6	60.0
CMR	1987	94,750	90.9	74.8	200,469	89.1	60.7
CMR	2005	203,450	95.6	82.8	449,515	93.6	68.7
EGY	1986	723,853	99.3	96.1	1,611,636	98.7	83.6
EGY	1996	718,875	99.5	96.8	1,471,286	98.9	83.7
EGY	2006	785,676	98.7	96.7	1,978,864	97.8	80.4
ETH	1984	309,641	93.8	76.9	633,338	92.6	58.3
ETH	1994	627,895	93.4	80.7	1,254,576	91.9	64.2
ETH	2007	969,639	94.6	81.5	1,949,874	92.7	63.3
GHA	1984	142,526	93.8	85.0	302,954	91.3	72.6
GHA	2000	200,000	94.2	83.8	434,882	91.8	71.2
GHA	2010	270,162	94.2	84.8	603,020	89.1	70.4
GIN	1983	44,316	84.6	76.9	100,250	84.4	68.8
GIN	1996	71,439	92.8	84.4	153,619	91.8	76.7
KEN	1969	67,260	78.3	63.3	167,003	70.2	44.1
KEN	1989	127,055	94.0	81.9	263,585	89.8	64.2
KEN	1999	176,867	92.5	77.7	378,922	90.0	60.6
KEN	2009	438,897	93.6	82.6	968,723	89.9	63.2
LBR	2008	38,854	94.1	83.4	87,459	92.2	68.8
LSO	1996	24,395	93.9	87.2	50,274	93.7	81.1
LSO	2006	22,361	94.8	82.3	50,625	92.9	73.7
MAR	1982	118,922	97.3	93.6	251,487	96.4	85.5
MAR	1994	149,529	96.4	94.5	322,163	95.3	88.1
MAR	2004	161,892	97.8	96.2	363,627	96.7	90.2
MLI	1987	79,031	95.9	84.0	168,635	95.6	73.0
MLI	1998	113,202	94.6	84.1	225,998	94.8	74.2
MLI	2009	159,705	94.9	85.4	328,411	95.3	74.4
MOZ	1997	172,377	96.7	75.8	380,792	95.5	56.0
MOZ	2007	199,970	97.6	78.4	453,004	96.6	57.8
MWI	1987	81,070	93.0	68.9	176,442	91.7	50.8
MWI	1998	114,846	94.3	72.7	251,873	93.6	52.3
MWI	2008	144,751	93.2	76.3	324,412	93.8	54.2
NGA	2006	8,868	99.2	91.2	18,236	97.1	77.3
NGA	2007	8,886	98.7	91.0	18,194	96.4	77.4
NGA	2008	11,572	99.2	92.9	23,462	97.5	80.6
NGA	2009	8,058	90.7	84.9	16,702	89.0	73.3
NGA	2010	7,914	99.1	93.1	16,452	97.5	81.0
RWA	1991	78,334	91.6	82.7	163,796	90.2	67.2
RWA	2002	117,490	87.2	76.2	235,909	85.7	60.6
RWA	2012	113,386	91.8	84.9	252,856	89.5	67.8
SDN	2008	591,921	96.3	86.4	1,258,800	95.4	71.7
SEN	1988	71,362	88.2	84.9	158,443	85.0	77.6
SEN	2002	124,706	93.8	90.4	260,317	92.8	85.9
SLE	2004	55,764	96.6	87.7	121,925	95.4	76.8
SSD	2008	57,942	95.5	85.9	120,722	94.3	74.3
TGO	2010	59,583	92.9	79.5	136,454	90.9	65.3
TZA	1988	278,539	91.5	75.5	557,552	88.8	60.5
TZA	2002	416,283	84.5	70.5	903,115	82.6	53.1
TZA	2012	491,497	89.7	77.9	1,036,707	87.6	61.9
UGA	1991	179,474	92.0	72.0	381,144	89.7	54.5
UGA	2002	289,123	94.4	77.2	601,101	92.0	56.2
ZAF	1996	384,919	93.3	83.1	865,987	88.8	71.7
ZAF	2001	421,066	96.1	83.6	915,973	92.4	73.6
ZAF	2007	114,829	90.5	78.8	256,019	87.7	70.5
ZAF	2011	431,062	94.2	80.7	1,052,039	88.1	67.6
ZMB	1990	108,294	98.5	87.4	216,756	97.1	72.9
ZMB	2000	119,089	98.4	86.4	259,096	97.3	68.5
ZMB	2010	159,684	98.8	87.6	333,273	97.5	69.4
ZWE	2012	74,478	92.9	70.5	158,021	90.3	55.2
total		13,005,949	94.5	83.8	28,154,990	92.6	69.0

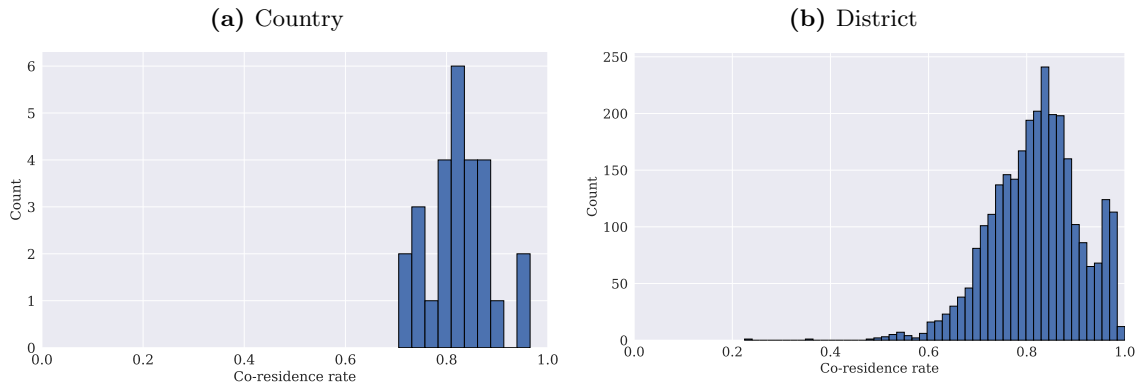
Cohabitation rates by census for all individuals aged 14-18 and 14-25. Columns (1) and (4) report the number of individuals for each census. Columns (2) and (5) report the fraction of individuals living with a relative of any generation. Columns (3) and (6) report the fraction of individuals living with at least one relative of the immediately older generation; these are identified with the “relationship to household head” classification or with the age for “other relatives (not elsewhere classified).

Figure D.1: Family Structures. All Censuses



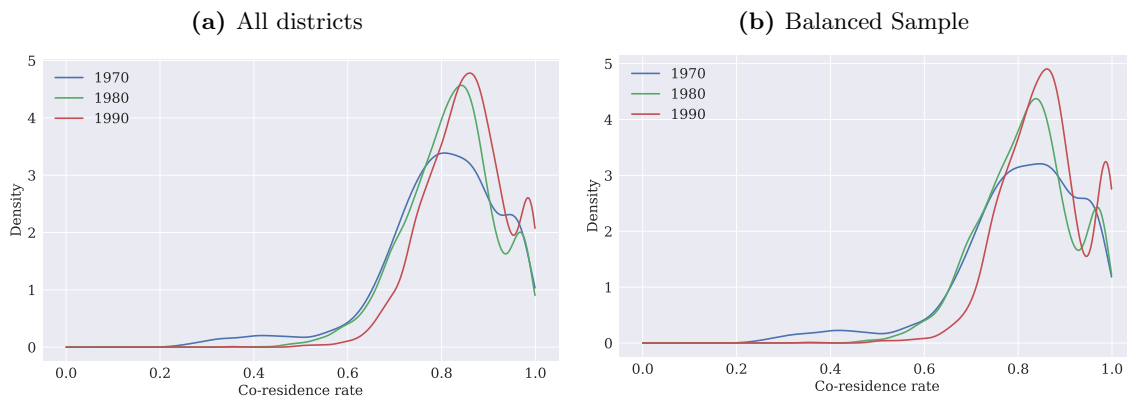
The chart plots the distribution of household structure for individuals aged 14-18 (panel (a)) and 14-25 (panel (b)), who reside with at least one relative from the immediately older generation.

Figure D.2: Country & District Co-habitation Rates, individuals aged 14-18



Distribution of co-habitation rates for all individuals aged 14-18 at the country (panel (a)) and at the district (panel (b)) level, pooling observations across all cohorts and census-years. Country: Min=0.705, q10=0.742, median=0.826, q90=0.894, max=0.965, std=0.06. District: Min=0.225, q10=0.706, median=0.826, q90=0.947, max=1, std=0.09

Figure D.3: District Co-habitation across Cohorts, individuals aged 14-18



The figure plots the distribution of co-habitation rates for individuals aged 14-18 at the district level across birth-cohorts. Panel (a) uses all birth-cohort-district observations. Panel (b) uses observations from districts with data for all cohorts.

sample	cohort	N	min	q10	q20	median	mean	q80	q90	max	std	skew
all	1970	1,996	0.247	0.658	0.719	0.813	0.8	0.905	0.957	1.0	0.13	-1.246
all	1980	2,352	0.405	0.691	0.74	0.823	0.817	0.894	0.953	0.995	0.096	-0.45
all	1990	2,598	0.355	0.74	0.777	0.852	0.85	0.924	0.977	1.0	0.087	-0.383
balanced	1970	1,760	0.247	0.656	0.723	0.821	0.804	0.919	0.96	1.0	0.134	-1.308
balanced	1980	1,760	0.466	0.69	0.737	0.823	0.819	0.904	0.966	0.995	0.098	-0.305
balanced	1990	1,760	0.355	0.747	0.782	0.855	0.855	0.928	0.988	1.0	0.087	-0.338

E IM Across Countries and Regions

This Section reports additional measures of intergenerational mobility (IM), distinguishing across gender, rural-urban residence, and family structure. It also gives further descriptive evidence.

E.1 IM by Household Structure

We compute upward IM and downward IM for the 1990s cohort, which has the broadest country-coverage, for individuals who cohabit with “older” generations of different types. In particular, we compute IM for children cohabiting (i) only with biological parents, (ii) only with older relatives other than biological parents (e.g., aunts, uncles), and (iii) with biological parent(s) and other older generation relative(s). For comparability, figure E.1 below also reports the baseline IM measure, estimated for young individuals residing in all types of households. Table E.1 gives the cross-country correlations of IM statistics across the different types of family structure.

Figure E.1: IM by Household Structure, individuals ages 14-18

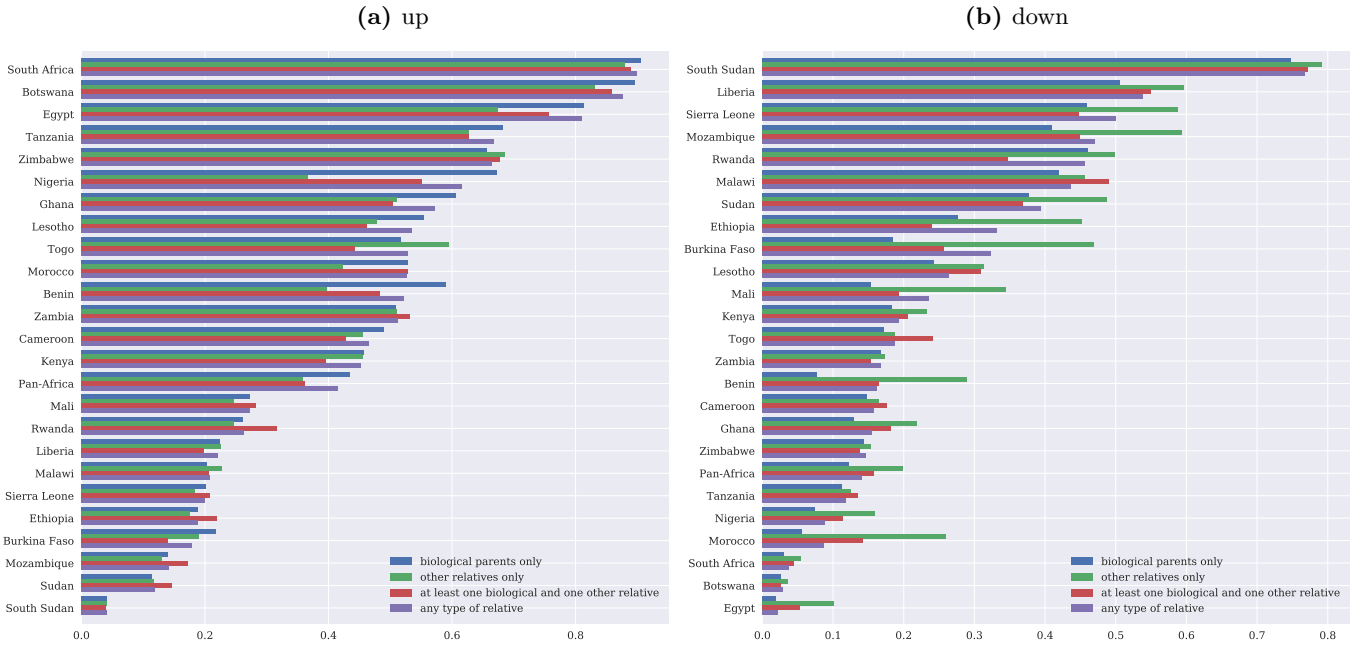


Table E.1: Correlation matrix for country-level IM, ages 14-18, different family structures

Panel A: upward IM					Panel B: downward IM				
	other only	biological only	other and biological	all (baseline)		other only	biological only	other and biological	all (baseline)
other only	1.0	0.95	0.961	0.966	other only	1.0	0.92	0.927	0.968
biological only	0.95	1.0	0.981	0.997	biological only	0.92	1.0	0.971	0.984
other and biological	0.961	0.981	1.0	0.989	other and biological	0.927	0.971	1.0	0.973
all (baseline)	0.966	0.997	0.989	1.0	all (baseline)	0.968	0.984	0.973	1.0

This table shows a correlation matrix for country-level IM, computed for individuals aged 14-18, 1990s birth cohort, who live in different family arrangements. Panel A shows the results for upward IM, panel B for downward IM.

E.2 IM across Gender and Rural and Urban Households

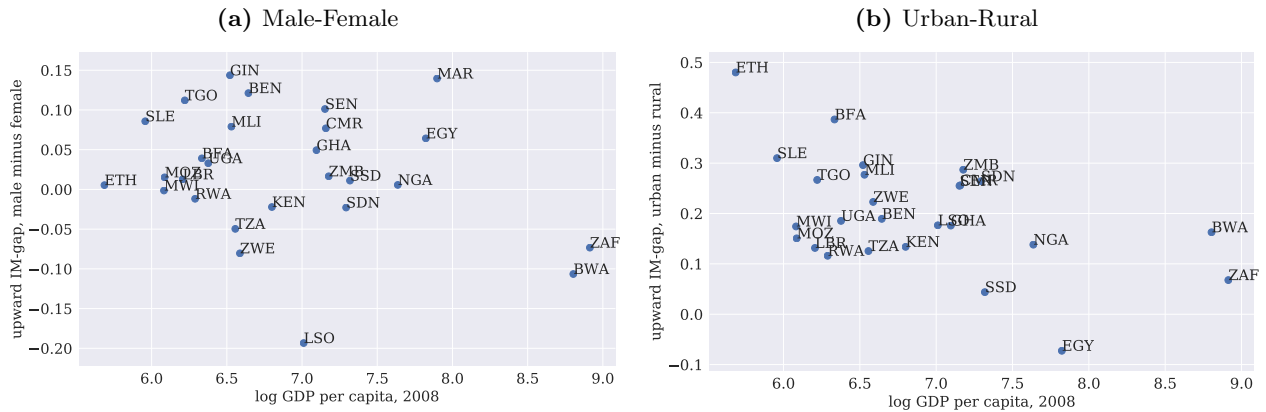
Appendix Table E.2 reports upward IM separately for girls and boys as well as rural and urban households and the associated male-female and urban-rural IM gaps. Appendix Figure E.2 plots the cross-country association between the gender and urban-rural gaps and economic development.

Table E.2: Country-level estimates of IM, male/female, urban/rural

gender/location	census years	(1)	(2)	(3)	(4)	(5)	(6)
		male	female	male - female	urban	rural	urban - rural
Lesotho	1996, 2006	0.339	0.532	-0.193	0.595	0.418	0.177
Botswana	1981, 1991, 2001, 2011	0.649	0.756	-0.107	0.769	0.606	0.163
Zimbabwe	2012	0.627	0.707	-0.08	0.867	0.644	0.223
South Africa	1996, 2001, 2007, 2011	0.754	0.827	-0.073	0.838	0.77	0.068
Tanzania	1988, 2002, 2012	0.571	0.621	-0.05	0.695	0.57	0.125
Sudan	2008	0.109	0.132	-0.023	0.348	0.084	0.264
Kenya	1969, 1989, 1999, 2009	0.444	0.466	-0.022	0.586	0.452	0.134
Rwanda	1991, 2002, 2012	0.286	0.297	-0.012	0.327	0.211	0.116
Malawi	1987, 1998, 2008	0.154	0.156	-0.001	0.315	0.141	0.174
Ethiopia	1984, 1994, 2007	0.131	0.126	0.005	0.545	0.065	0.48
Nigeria	2006, 2007, 2008, 2009, 2010	0.632	0.626	0.006	0.745	0.606	0.138
South Sudan	2008	0.046	0.035	0.011	0.076	0.032	0.044
Liberia	2008	0.227	0.214	0.013	0.305	0.174	0.132
Mozambique	1997, 2007	0.118	0.103	0.015	0.213	0.063	0.151
Zambia	1990, 2000, 2010	0.494	0.478	0.017	0.689	0.402	0.287
Uganda	1991, 2002	0.373	0.34	0.033	0.534	0.349	0.185
Burkina Faso	1996, 2006	0.202	0.163	0.039	0.502	0.115	0.387
Ghana	1984, 2000, 2010	0.589	0.54	0.049	0.681	0.505	0.176
Egypt	1986, 1996, 2006	0.667	0.603	0.064	0.593	0.665	-0.073
Cameroon	1976, 1987, 2005	0.546	0.469	0.077	0.707	0.451	0.256
Mali	1987, 1998, 2009	0.242	0.163	0.079	0.423	0.146	0.277
Sierra Leone	2004	0.291	0.205	0.086	0.458	0.148	0.31
Senegal	1988, 2002	0.307	0.206	0.101	0.415	0.16	0.255
Togo	1960, 1970, 2010	0.561	0.449	0.112	0.719	0.452	0.267
Benin	1979, 1992, 2002, 2013	0.432	0.311	0.121	0.509	0.32	0.189
Morocco	1982, 1994, 2004	0.484	0.344	0.14			
Guinea	1983, 1996	0.262	0.118	0.144	0.403	0.107	0.296
mean		0.39	0.37	0.02	0.533	0.333	0.2

This table reports upward IM measures (estimated without cohort effects) separately for male and female individuals and those living in urban and rural locations aged 14-18. Columns (1) and (2) give the estimates for upward IM for boys and girls, respectively. They reflect the likelihood that children, whose parents have not completed primary schooling will manage to complete at least primary education. Columns (4) and (5) give the corresponding estimates for individuals residing in urban and rural locations at the time of the census. Column (5) shows the mobility gap, subtracting the female from the male estimate and column (6) shows the urban minus rural gap. Countries are sorted from the lowest to highest male-female IM gap (column (5)). “mean” gives the simple average of the 27 country-estimates.

Figure E.2: Male/female and Urban/Rural Gaps in upward IM and GDP per capita

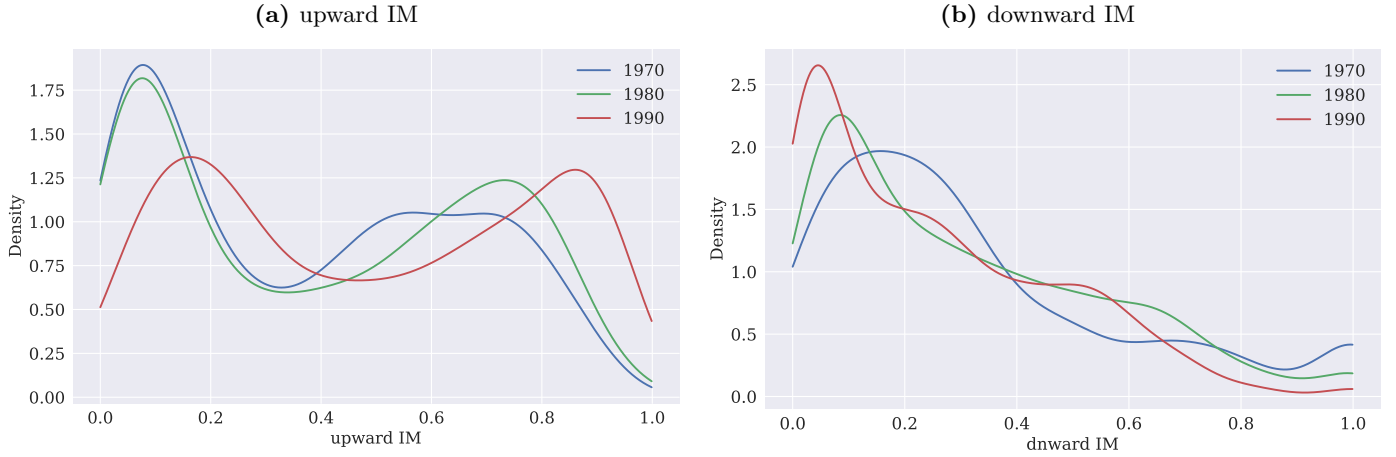


The figures plot the gap in upward IM between boys and girls (panel (a)) and individuals growing up in urban and rural households (panel (b)) against the logarithm of real GDP per capita in 2008 using World Bank data (2008 is the earliest year for which data for all countries, including South Sudan, is available).

E.3 IM across Cohorts

Appendix Figure E.3 plots the distribution of regional upward IM (panel (a)) and downward IM (panel (b)) for the 1970s, 1980s, and 1990-born cohorts. There are 17 countries with information on the three cohorts; Benin, Burkina Faso, Botswana, Cameroon, Egypt, Ethiopia, Ghana, Kenya, Lesotho, Morocco, Mali, Mozambique, Malawi, Rwanda, Tanzania, South Africa, Zambia. Appendix Table E.3 reports summary statistics.

Figure E.3: Distribution district \times cohort level IM, balanced sample



Distribution of district upward (panel (a)) and downward (panel (b)) IM for individuals aged 14-18 for different cohorts. Sample restricted to the districts for which IM is available for all cohorts.

Table E.3: Summary-statistics, district-cohort-level IM, individuals aged 14-18

direction	cohort	count	min	q10	q20	median	mean	q80	q90	max	std	skew
up	1970	1,759	0.0	0.04	0.076	0.363	0.378	0.683	0.769	0.962	0.282	0.22
up	1980	1,759	0.0	0.043	0.075	0.403	0.403	0.725	0.801	0.971	0.298	0.117
up	1990	1,759	0.006	0.098	0.164	0.488	0.49	0.833	0.898	1.0	0.304	0.026
down	1970	1,620	0.0	0.049	0.095	0.25	0.326	0.5	0.75	1.0	0.273	1.084
down	1980	1,620	0.0	0.047	0.077	0.247	0.312	0.545	0.667	1.0	0.252	0.838
down	1990	1,620	0.0	0.02	0.04	0.198	0.251	0.461	0.571	1.0	0.219	0.848

Balanced sample of 17 countries for which we have data for 1970, 1980, and 1990 birth cohort. The number of districts is different for the two directions since not every region in every cohort has recorded in the census an individual with the required configuration (right age range, literate or illiterate parents).

E.4 Heterogeneity. IM and Literacy of the Old Generation

Appendix Table E.4 reports un-weighted within-country regression specifications associating upward and downward IM with the old generation's literacy distinguishing between rural and urban households. The table also reports a Chow test of coefficient equality. The old's literacy - upward-IM correlation is considerably stronger for rural, 0.75 (in (2)), as compared to urban 0.525 (in (1)) households. Columns (5)-(6) replace the country fixed-effects with province constants to account for more localized features. The estimate in the urban sample is 0.40, while in the rural sample it is 0.62, a significant difference. The specifications with downward IM in columns (3)-(4), (7)-(8) yield similar patterns. The slope is twice as large in rural as compared to urban households (0.27 versus 0.53).

Table E.4: Old literacy and IM at the district-level, urban/rural, ages 14-18

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	IM up	IM up	IM down	IM down	IM up	IM up	IM down	IM down
share literate old	0.525*** (0.044)	0.746*** (0.052)	-0.270*** (0.026)	-0.532*** (0.049)	0.402*** (0.048)	0.623*** (0.055)	-0.209*** (0.039)	-0.436*** (0.038)
R-squared	0.750	0.863	0.700	0.764	0.825	0.918	0.774	0.821
N	1997	2633	1982	2618	1997	2633	1982	2618
sub-sample	urban	rural	urban	rural	urban	rural	urban	rural
country FEs	yes	yes	yes	yes	no	no	no	no
province FEs	no	no	no	no	yes	yes	yes	yes
p: coeff-equal	0.0000		0.0000		0.0000		0.0000	

The dependent variable is the district-level share of literate kids of illiterate parents (estimated net of census year and old and young birth decade fixed effects). The independent variable is the district-level share of literate parents (also estimated net of fixed effects). Standard errors clustered at the admin-1 (province)-level in parentheses. $*p < 0.1$, $**p < 0.5$, $***p < 0.01$. p -values for coefficient equality in the urban/rural sub-samples are from a Chow-test ($\sim \chi^2$ under H_0).

Appendix Table E.5 associates – again in an unweighted manner – upward and downward IM to the literacy of the “old” generation at the regional level, distinguishing across boys and girls. The country (province) fixed-effects coefficient in the upward IM specification for boys is 0.66 (0.53), while for girls it is 0.78 (0.62). A Chow test of coefficient equality suggests that the difference is statistically different than zero. Likewise, the downward IM - “old” generation literacy association is somewhat steeper for girls, as compared to boys. The difference is 0.12 in the country fixed-effects estimation, but it drops to 0.047 in the province fixed-effects specifications.

Table E.5: Old literacy and IM at the district-level, male/female, ages 14-18

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	IM up	IM up	IM down	IM down	IM up	IM up	IM down	IM down
share literate old	0.657*** (0.036)	0.782*** (0.042)	-0.400*** (0.028)	-0.521*** (0.043)	0.529*** (0.039)	0.623*** (0.043)	-0.397*** (0.033)	-0.444*** (0.039)
R-squared	0.867	0.892	0.718	0.746	0.921	0.940	0.770	0.800
N	2846	2845	2829	2829	2846	2845	2829	2829
sub-sample	male	female	male	female	male	female	male	female
country FEs	yes	yes	yes	yes	no	no	no	no
province FEs	no	no	no	no	yes	yes	yes	yes
p: coeff-equal	0.0000		0.0028		0.0000		0.0474	

The dependent variable is the district-level share of literate kids of illiterate parents (estimated net of census year and old and young birth decade fixed effects). The independent variable is the district-level share of literate parents (also estimated net of fixed effects). Standard errors clustered at the admin-1 (province)-level in parentheses. $*p < 0.1$, $**p < 0.5$, $***p < 0.01$. p -values for coefficient equality in the male/female sub-samples are from a Chow-test ($\sim \chi^2$ under H_0).

F Correlates of IM

This Section complements the correlational analysis (Section 5 of the paper). First, we provide variable definitions and sources for all correlates of regional mobility. Second, we discuss and present LASSO estimates.

F.1 Variable definitions

ln(distance to the capital) The natural logarithm of the geodesic distance from the district centroid to the national capital. Computed using GIS software.

ln(distance to the border) The natural logarithm of the geodesic distance from the district centroid to closest point on the national border. Computed using GIS software.

ln(distance to the coast) The natural logarithm of the geodesic distance from the district centroid to closest point on the coastline. Computed using GIS software.

ln(1+malaria stability) The natural logarithm of 1 + mean stability of malaria transmission in the district. The latter variable is computed, using GIS software, as the within-district zonal statistic of a raster provided by Kiszewski et al. (2004), which we resample to a resolution of 30 arc-seconds prior to computing the statistic.

ln(1+agricultural suitability) The natural logarithm of 1 + mean agricultural suitability in the district. The latter variable is computed, using GIS software, as the within-district zonal statistic of a raster provided by Ramankutty et al. (2002), which we resample to a resolution of 30 arc-seconds prior to computing the statistic.

ln(terrain ruggedness) The natural logarithm of terrain ruggedness. The latter is computed using cell-level data on elevation at 30 arc-second resolution from Survey (1996). Given the grid cell data, picture a 3×3 block of 9 cells and let $e_{r,c}$ be the elevation of the cell in row r , column c of the grid. Following Nunn and Puga (2012), we compute ruggedness as $\sqrt{\sum_{i=r-1}^{r+1} \sum_{j=c-1}^{c+1} (e_{i,j} - e_{r,c})^2}$, that is, the square root of the sum of all the squared differences in elevation between the middle cell and the surrounding 8 cells.

oil field dummy A dummy = 1 if the district is intersected by an oil field, and zero otherwise. Data on oil fields come from Lujala, Rød, and Thieme (2007)

diamond mine dummy A dummy = 1 if the district is intersected by a diamond mine, and zero otherwise. Data on oil fields come from Lujala, Gleditsch, and Gilmore (2005)

ln(population density 1950) The natural logarithm of mean population density in the district in 1950. The latter variable is computed, using GIS software, as the within-district zonal statistic of a raster provided by Klein Goldewijk, Beusen, and Janssen (2010), which we resample to a resolution of 30 arc-seconds prior to computing the statistic.

ln(distance to railroad) The natural logarithm of the geodesic distance from the district centroid to closest point on a colonial railroad. Computed using GIS software. Data on colonial railroads come from Jedwab and Moradi (2016).

ln(distance to road) The natural logarithm of the geodesic distance from the district centroid to closest point on a colonial road. Computed using GIS software. Data on colonial roads come from Jedwab and Storeygard (2018).

ln(distance to Catholic mission) The natural logarithm of the geodesic distance from the district centroid to the closest Catholic Mission. Computed using GIS software. Data on missions come from Nunn (2010).

ln(distance to Protestant mission) The natural logarithm of the geodesic distance from the district centroid to the closest Protestant Mission. Computed using GIS software. Data on missions come from Nunn (2010) and Cagé and Rueda (2016).

ln(distance to precolonon. empire) The natural logarithm of the geodesic distance from the district centroid to the closest pre-colonial empire. Computed using GIS software. Data on pre-colonial empires come from Besley and Reynal-Querol (2014).

ln(distance to precolonon. state) The natural logarithm of the geodesic distance from the district centroid to the closest pre-colonial state. Data on pre-colonial states are obtained combining maps of pre-colonial ethnic homelands with the levels of jurisdictional hierarchy beyond the local community level in Murdock (1967). Societies with 3 or 4 levels are classified as states.

urban share (born < 1960) The share of the (non-migrant) district population born prior to 1960 classified as urban. IPUMS census data.

agri. empl share (born < 1960) The share of the (non-migrant) district population born prior to 1960 and working in agriculture. IPUMS census data.

manuf. empl share (born < 1960) The share of the (non-migrant) district population born prior to 1960 and working in manufacturing. IPUMS census data.

serv. empl share (born < 1960) The share of the (non-migrant) district population born prior to 1960 and working in services. IPUMS census data.

F.2 LASSO

Many of the variables that we consider in Section 5 to characterize the spatial distribution of educational mobility contain measurement error and are correlated with each other. We thus apply LASSO, a machine learning method that via regularization selects the variables (shrinkage) most predictive of upward and downward IM (Tibshirani (1996)). Before presenting the results, we should stress that LASSO does not aim to identify causal effects nor provide unbiased estimates of linear regression coefficients. The objective is providing parsimonious empirical models that predict well regional IM.

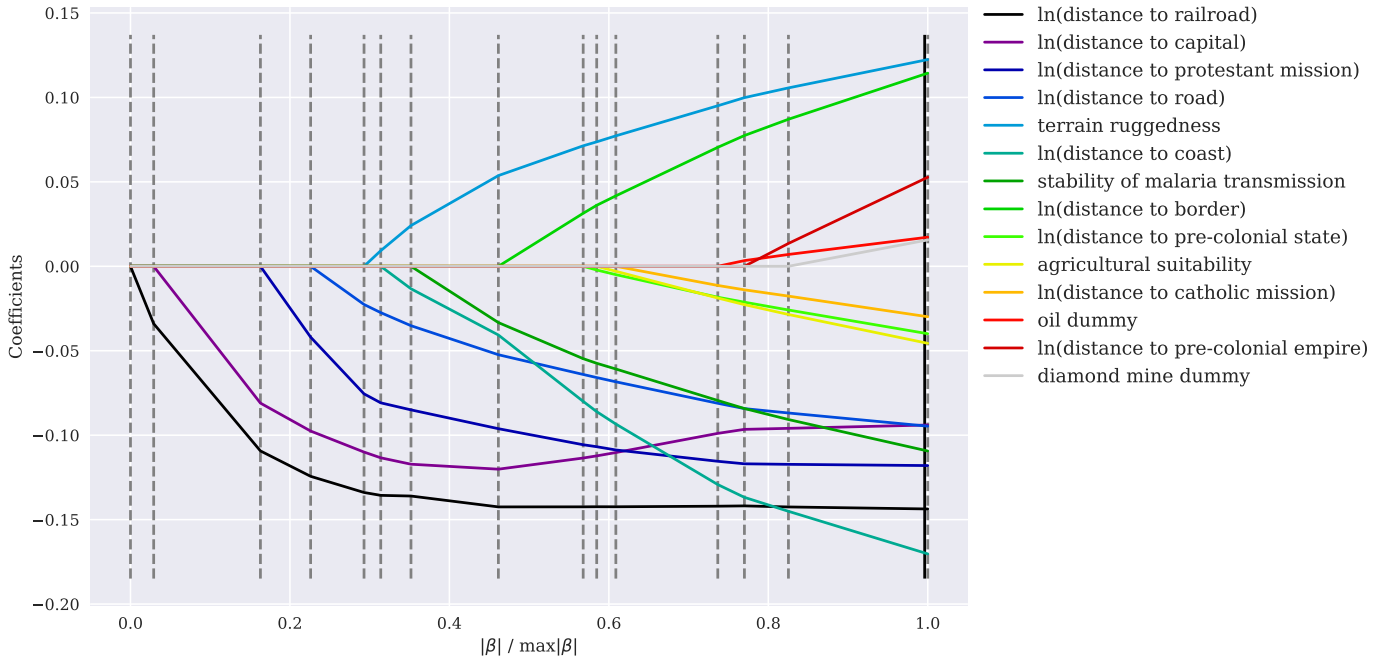
We compute the LASSO path using Least Angle Regression (LAR, Efron et al. (2004)), allowing λ to range from 0 (OLS) to infinity (all coefficients are set to zero).²

Figures F.1 (a)-(b) plot the entire LASSO path for upward and downward IM, looking at the role of the geographic and historical features. The vertical axis plots the coefficients of the standardized variables. The vertical axis gives the ratio of the sum of the absolute values of the coefficients for a given λ to the OLS sum (that takes the maximum value). The graphs should be read from right, no regularization (OLS), to left (regularization results in all parameters to be zero). Moving from right to left, more variables drop from the model; the last variables that drop out are the “strongest” predictors of mobility.

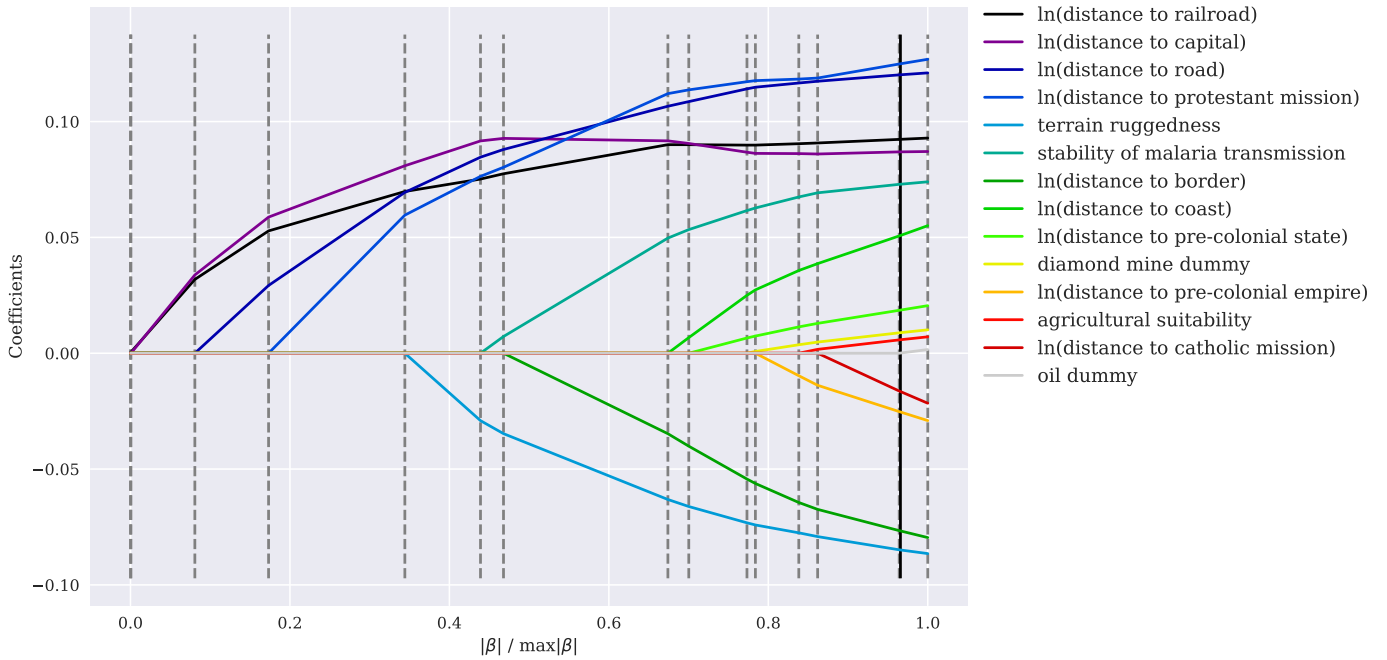
²We also computed the “optimal” degree of regularization through K -fold cross-validation. We split the data into 10 (5) equal size bins of districts; for a given value of λ we fit (train) the model in 9 (4) bins, then take out-of-sample fits in the remaining bin, and compute the forecast accuracy with the mean square error criterion. Repeating this process for a grid of 100 values of λ , we take the value that yields the lowest error. The cross-validation routine yields $\beta / \max |\beta|$ around 0.8.

Figure F.1: Lasso estimates

(a) Upward



(b) Downward



While the number of observations is quite small for “machine learning” type algorithms, LASSO is useful pinpointing the relatively more important predictors of mobility. The LASSO analysis with upward IM yields some noteworthy features. First, in line with the univariate correlational analysis, natural resource indicators and proximity to pre-colonial states are the variables with the least explanatory power. Second, proximity to colonial railroads is the most important predictor of upward mobility, hinting that these investments, though overall small, had sizable lasting consequences. Third, proximity to the capital is also an important predictor of upward IM. Fourth, proximity to Protes-

tant missions is a robust predictor of upward IM, way more than proximity to Catholic missions that drops out of the empirical model once regularization increase. While these results do not have a causal interpretation, they add to the literature assessing the legacy of Protestant and Christian missions. Note that the univariate correlations are quite similar, showing the usefulness of complementing OLS with LASSO. Fifth, ecological features, malaria stability, terrain ruggedness, and distance to coast are in-between suggesting some non-negligible explanatory power.

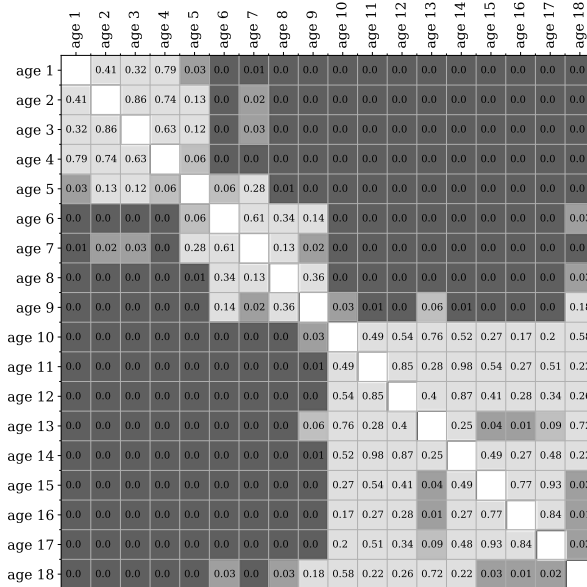
The patterns when we run LASSO on downward IM are similar. Distance to colonial railroads and the capital carry the highest prediction power, while natural resources, soil suitability for agriculture, and distance to precolonial states are the least robust predictors of downward IM. The graph also illustrates the difference Protestant and Catholic missions; the former being among the most important predictors of downward IM, the latter the (second) least important.

G Exposure Effects

This Section’s results complement the analysis in Section 5 of the paper. The sub-sections of this appendix follow the sub-sections of the main paper. In particular, section G.1 reports tests of coefficient equality for the age-specific regional effects and some sensitivity checks of the semi-parametric estimates in section 5.1. Section G.2 complements the parametric estimates of section 5.2, presenting the heterogeneity analysis with respect to gender and the direction of the move and further sensitivity checks. Section G.3 reports additional results that exploit displacement shocks and use historical migration to predict moves supporting the evidence in section 5.3.

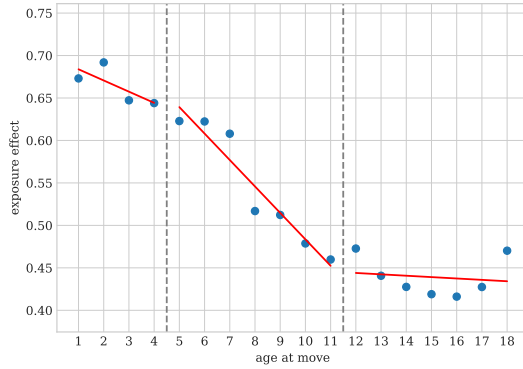
G.1 Baseline Semi-Parametric Estimates

Figure G.1: Visualization of coefficient equality pairwise F -tests



This figure visualizes p -values of pairwise coefficient equality F -tests for all coefficients in the household fixed effects specification in Figure 12 (tests for the observational estimates in the full and household-fixed effects samples yield similar results (not reported)). Cells are shaded according to significance level: ■ $p < 0.01$, ■ $0.01 < p < 0.05$, ■ $0.05 < p < 0.1$, ■ $p > 0.1$.

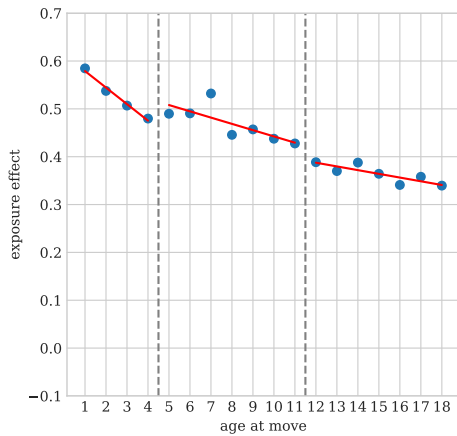
Figure G.2: Instrumenting Δ_{odb} with sample split



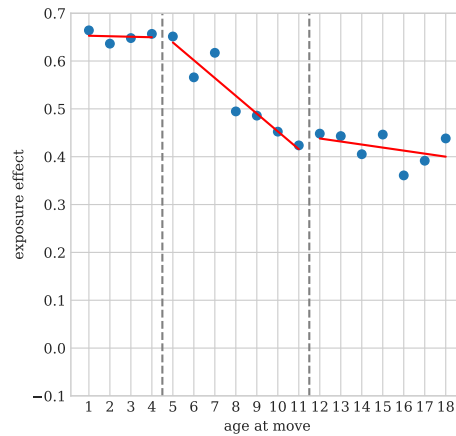
This figure is based on the same specification as Figure 11 except that instead of using Δ_{odb} , we split the sample of non-movers in every cohort-district in two at random and construct two measures of Δ_{odb} , one per sample split. We then re-estimate the specification using all terms involving Δ_{odb} from one of the halves as in instrument for the other half.

Figure G.3: Graphs of semi-parametric exposure effects, rich/poor countries

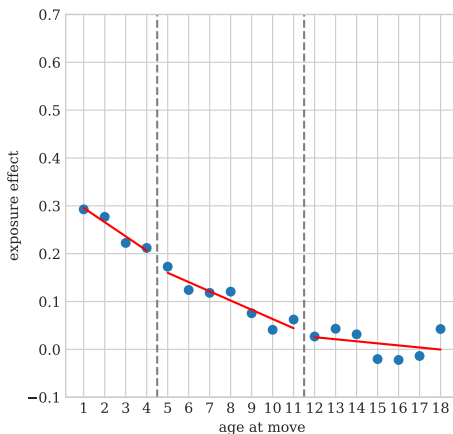
(a) Rich: HH FE sample, no HH FEs



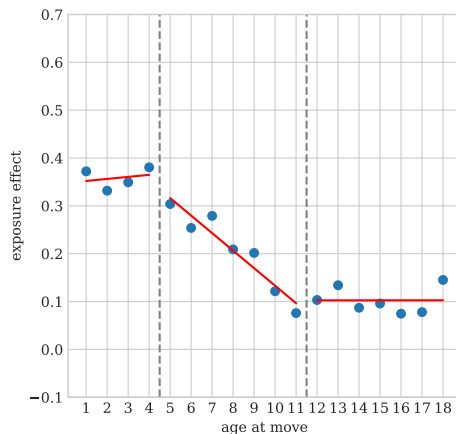
(b) Poor: HH FE sample, no HH FEs



(c) Rich: HH FEs



(d) Poor: HH FEs



This figure plots the estimates from the baseline semi-parametric exposure effects specification estimated, separately for “rich” and “poor” countries. The left column shows the estimates for rich countries, the bottom two those for poor countries. We use World Bank data for real GDP per capita in 1986 (the earliest year for which data for all countries in sample used for the exposure effects analysis are available) and use the median to make the rich/poor split. Accordingly, rich are, from least to most, Sudan, Ghana, Kenya, Zambia, Egypt, Morocco, Cameroon, South Africa. Poor are, from least to most, Benin, Togo, Guinea, Mali, Rwanda, Malawi, Uganda, Ethiopia.

G.2 Parametric Estimates

Table G.1: Parametric exposure effects estimates, heterogeneity

	(1)	(2)	(3)		(4)	(5)	(6)
	IM	IM	IM		IM	IM	IM
β : 1-4, Δ_{odb}^m	0.00973 (0.018)	0.00672 (0.024)	-0.00504 (0.022)	β : 1-4, Δ_{odb}^+	0.0141 (0.019)	0.0288 (0.030)	0.0180 (0.027)
γ : 5-11, Δ_{odb}^m	0.0233*** (0.006)	0.0255*** (0.006)	0.0257*** (0.006)	γ : 5-11, Δ_{odb}^+	0.0339*** (0.008)	0.0311*** (0.010)	0.0271*** (0.009)
δ : 12-18, Δ_{odb}^m	-0.00966 (0.007)	-0.00324 (0.008)	-0.00272 (0.006)	δ : 12-18, Δ_{odb}^+	0.00788 (0.009)	0.00365 (0.011)	-0.0000797 (0.006)
β : 1-4, $\Delta\Delta_{odb}^f$	0.0175 (0.020)	0.0126 (0.026)	0.0262 (0.025)	β : 1-4, $\Delta\Delta_{odb}^-$	0.0116 (0.032)	-0.0429 (0.047)	-0.0309 (0.042)
γ : 5-11, $\Delta\Delta_{odb}^f$	0.0168*** (0.006)	0.00999 (0.008)	0.0133 (0.008)	γ : 5-11, $\Delta\Delta_{odb}^-$	-0.00758 (0.013)	-0.00523 (0.017)	0.00820 (0.015)
δ : 12-18, $\Delta\Delta_{odb}^f$	0.0180** (0.009)	0.0125* (0.007)	0.0115 (0.007)	δ : 12-18, $\Delta\Delta_{odb}^-$	-0.0220 (0.014)	-0.00237 (0.017)	0.00513 (0.012)
R-squared	0.150	0.123	0.683	R-squared	0.143	0.120	0.679
N	406175	226739	226739	N	406175	226739	226739
age at mig FE	yes	yes	yes	age at mig FE	yes	yes	yes
birth decade FE	yes	yes	yes	birth decade FE	yes	yes	yes
hh FE	no	no, hhfe sample	yes	hh FE	no	no, hhfe sample	yes
age range	14-25	14-25	14-25	age range	14-25	14-25	14-25

The dependent variable in all regression is a dummy = 1 if the child has completed at least primary, and zero otherwise (i.e. a dummy for IM). The independent variables comprise a linear origin-average-IM (calculated for the birth-cohort relevant to the individual among non-movers) term, age-at-move dummies, birth-decade \times destination dummies interacted with destination-minus-origin opportunity differences (to capture differences in measurement error across locations and cohorts), all of which not reported, as well as three linear terms for destination-minus-origin differences in relevant-birth-cohort-non-mover average IM for move-ages 1-5, 6-12, and 13-18. In columns (1) - (3), Coefficient estimates Δ_{odb}^m show the estimates for the reference group (male children). $\Delta\Delta_{odb}^f$ show estimates of differential effects for female children. In columns (4) - (6), coefficient estimates Δ_{odb}^+ show the estimates for the reference group (movers to better places). $\Delta\Delta_{odb}^-$ show estimates of differential effects for movers to worse places. Standard errors clustered at origin- and destination-levels in parentheses. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

Table G.2: Parametric exposure effects estimates, robustness

	(1) IM	(2) IM	(3) IM	(4) IM	(5) IM	(6) IM	(7) IM	(8) IM	(9) IM
beta: 1-4	0.0174 (0.013)	0.0169 (0.022)	-0.000658 (0.019)	0.0186* (0.011)	0.0109 (0.017)	0.00980 (0.015)	0.0144 (0.012)	0.000246 (0.014)	-0.000107 (0.018)
gamma: 5-11	0.0236*** (0.005)	0.0225*** (0.005)	0.0218*** (0.005)	0.0302*** (0.005)	0.0274*** (0.006)	0.0307*** (0.006)	0.0313*** (0.006)	0.0323*** (0.008)	0.0331*** (0.008)
delta: 12-18	0.00147 (0.006)	0.00493 (0.007)	0.00473 (0.004)	-0.000455 (0.006)	0.00139 (0.006)	0.00196 (0.004)	-0.000639 (0.006)	0.00228 (0.006)	-0.00319 (0.007)
R-squared	0.167	0.148	0.675	0.142	0.115	0.683	0.099	0.084	0.692
N	297189	158438	158438	392079	213863	213863	276864	164071	164071
age at mig FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
birth decade FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
hh FE	no	no, hhfe sample	yes	no	no, hhfe sample	yes	no	no, hhfe sample	yes
age range	14-25	14-25	14-25	14-25	14-25	14-25	14-25	14-25	14-25
inc. Egypt + Morocco	no	no	no	yes	yes	yes	yes	yes	yes
incl mig hhs	yes	yes	yes	no	no	no	no	no	no
old for IM	all	all	all	all	all	all	biolog. only	biolog. only	biolog. only

The dependent variable in all regression is a dummy = 1 if the child has completed at least primary, and zero otherwise (i.e. a dummy for IM). The independent variables comprise a linear origin-average-IM (calculated for the birth-cohort relevant to the individual among non-movers) term, age-at-move dummies, birth-decade \times destination dummies interacted with destination-minus-origin opportunity differences (to capture differences in measurement error across locations and cohorts), all of which not reported, as well as three linear terms for destination-minus-origin differences in relevant-birth-cohort-non-mover average IM for move-ages 1-4, 5-11, and 12-18. Columns (1)-(3) exclude Egypt and Morocco. Columns (4)-(6) exclude individuals in multigenerational households. Columns (7)-(9) exclude individuals in multigenerational households and use only the education of biological parents to assess individual IM status. Standard errors clustered at origin- and destination-levels in parentheses. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

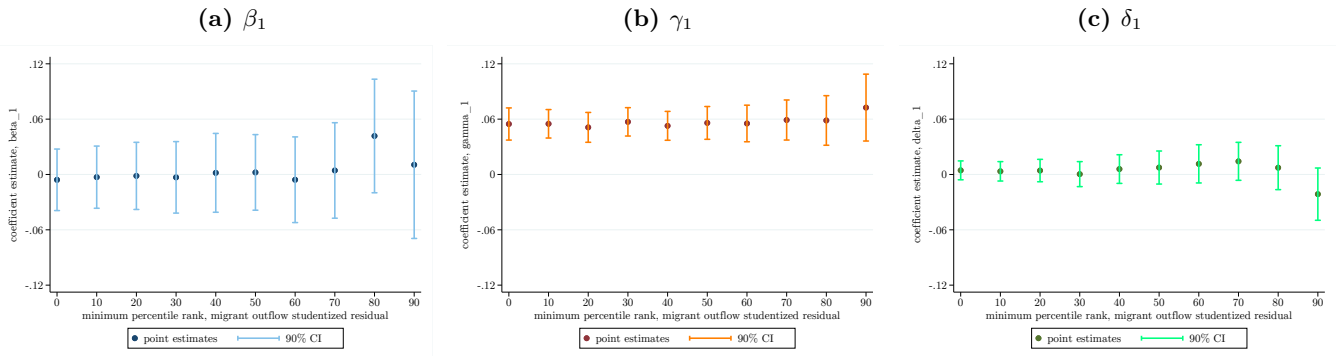
Table G.3: Parametric exposure effects estimates for average IM: parents working in agriculture and migrants in rural destination

	(1) IM	(2) IM	(3) IM	(4) IM	(5) IM	(6) IM	(7) IM	(8) IM
beta: 1-4	-0.00574 (0.030)	-0.0260 (0.028)	-0.00296 (0.022)	-0.0101 (0.021)	-0.00364 (0.039)	-0.0344 (0.040)	-0.0159 (0.038)	-0.0463 (0.039)
gamma: 5-11	0.0437*** (0.011)	0.0345*** (0.010)	0.0432*** (0.008)	0.0373*** (0.008)	0.0431*** (0.014)	0.0407*** (0.015)	0.0433*** (0.013)	0.0424*** (0.014)
delta: 12-18	0.0112 (0.009)	0.0159* (0.008)	-0.000312 (0.008)	0.00330 (0.007)	0.00721 (0.010)	0.00372 (0.011)	0.00731 (0.010)	0.0103 (0.012)
R-squared	0.677	0.675	0.692	0.685	0.671	0.668	0.664	0.661
N	48871	57952	105224	97388	36234	31899	43234	38917
age at mig FE	yes	yes	yes	yes	yes	yes	yes	yes
birth decade FE	yes	yes	yes	yes	yes	yes	yes	yes
hh FE	yes	yes	yes	yes	yes	yes	yes	yes
age range	14-25	14-25	14-25	14-25	14-25	14-25	14-25	14-25
ind/occ	industry	occupation	all obs	all obs	industry	industry	occupation	occupation
rural	all obs	all obs	household	region	household	region	household	region

The dependent variable in all regression is a dummy = 1 if the child has completed at least primary, and zero otherwise (i.e. a dummy for IM). The independent variables comprise a linear origin-average-IM (calculated for the birth-cohort relevant to the individual among non-movers) term, age-at-move dummies, birth-decade \times destination dummies interacted with destination-minus-origin opportunity differences (to capture differences in measurement error across locations and cohorts), all of which not reported, as well as three linear terms for destination-minus-origin differences in relevant-birth-cohort-non-mover average IM for move-ages 1-4, 5-11, and 12-18. The first two columns focus on children of agricultural workers defined by industry (column (1)) or occupation (column (2)). The next two columns focus instead on children of parents who migrated to rural areas defined by rural residence of the household (column (3)) or by a greater than sample median rural share (column (4)). Columns (5)-(8) focus more narrowly on the intersections of these two types of sample restrictions: Children of parents working in agriculture (as defined by industry) and having migrated to rural areas as defined by household (column (5)) or greater than median rural resident share (column (6)); and children of parents working in agriculture (as defined by occupation) and having migrated to rural areas as defined by household (column (7)) or greater than median rural resident share (column (8)). Standard errors clustered at origin- and destination-levels in parentheses. * $p < 0.1$, ** $p < 0.5$, *** $p < 0.01$.

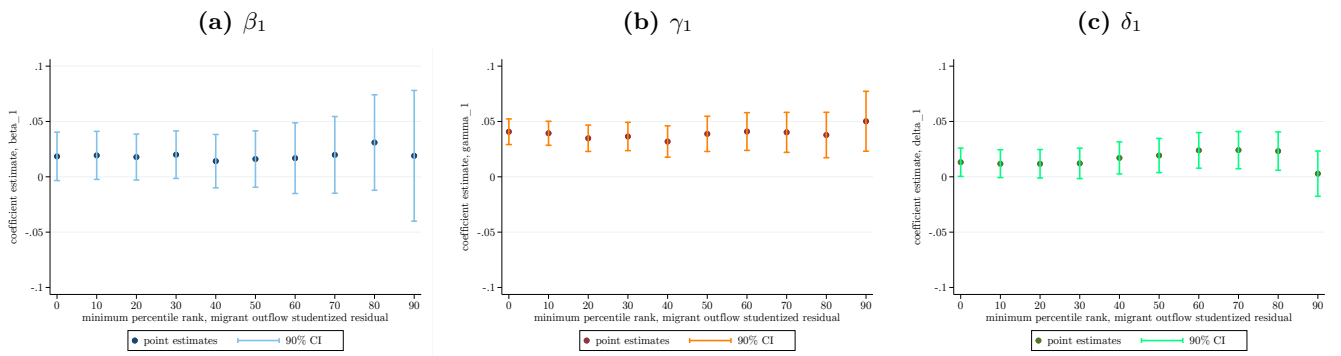
G.3 Endogeneity

Figure G.4: Combined strategies, region-level outflow anomalies, household fixed effects, 2SLS



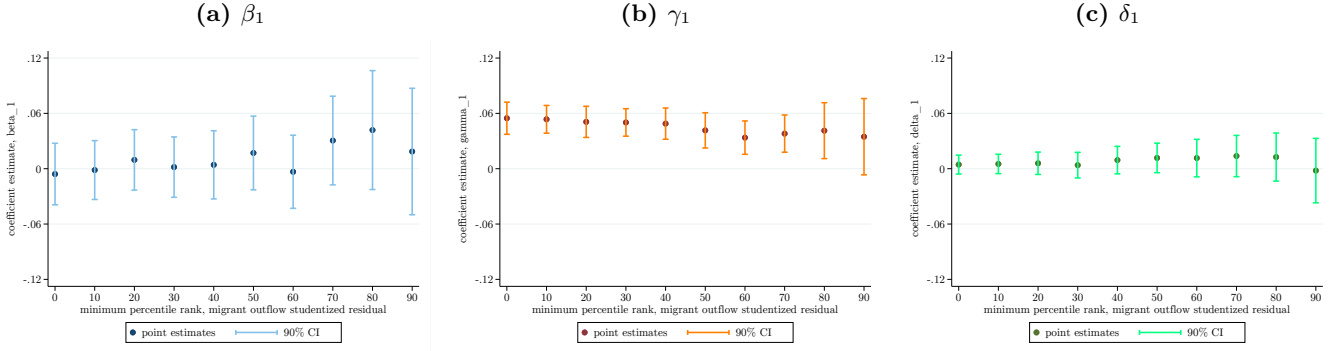
This figure shows point estimates from parametric regression estimates of exposure effects concentrating, successively, on district-years that experienced anomalously larger and large migration outflows and instrumenting the treatment Δ_{odb} with the predicted treatment $\hat{\Delta}_{ob}$ from the shift-share instrument (reduced form regressions). Each point represents the marginal effect of one additional year of treatment (in the relevant age-at-move range) with $\hat{\Delta}_{ob}=1$. Panel (a) shows the coefficient estimates for ages-at-move 1-4, panel (b) those for 5-11, and (c) those for 12-18. To identify outflow-years as anomalous, we construct a district-year migration outflow panel. For each district, we then regress these migration outflows on a constant and a linear time trend and obtain the studentized residuals from the regression. We then rank these residuals, within a district, from lowest to highest. District-years with high positive residuals experienced migration outflows above trend, and those with negative residuals experienced outflows below trend. We then run the baseline parametric regressions but successively focus on only those district-year observations that have a minimum percentile rank of at least X. The left-most observation for each graph shows the baseline estimates where no observations are dropped. The next observation requires the observations to come from district-years with studentized outflow residuals ranked at least at the 10th percentile within a district etc. All regressions include household fixed effects. 90% confidence bands constructed from standard errors clustered at origin- and destination-levels are shown.

Figure G.5: Combined strategies, region-level outflow anomalies, observational estimates, 2SLS



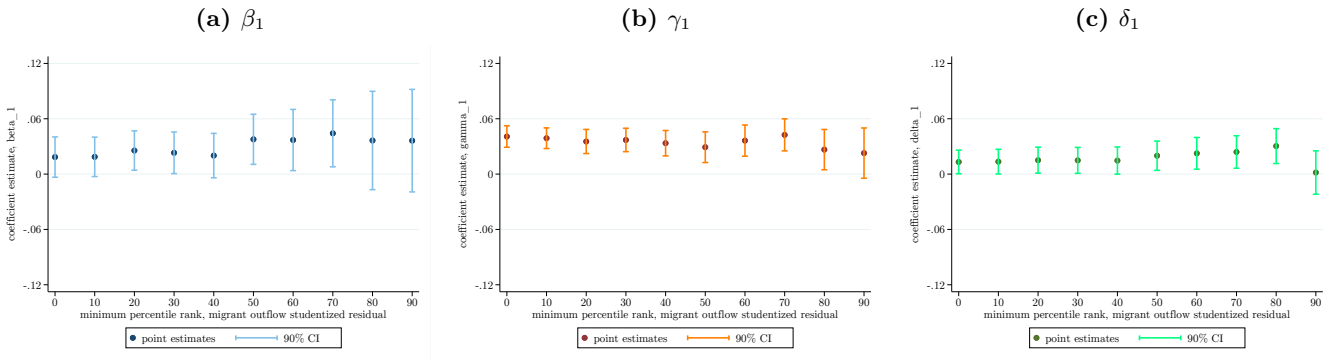
This figure shows point estimates from parametric regression estimates of exposure effects concentrating, successively, on district-years that experienced anomalously larger and large migration outflows and instrumenting the treatment Δ_{odb} with the predicted treatment $\hat{\Delta}_{ob}$ from the shift-share instrument (reduced form regressions). Each point represents the marginal effect of one additional year of treatment (in the relevant age-at-move range) with $\hat{\Delta}_{ob}=1$. Panel (a) shows the coefficient estimates for ages-at-move 1-4, panel (b) those for 5-11, and (c) those for 12-18. To identify outflow-years as anomalous, we construct a district-year migration outflow panel. For each district, we then regress these migration outflows on a constant and a linear time trend and obtain the studentized residuals from the regression. We then rank these residuals, within a district, from lowest to highest. District-years with high positive residuals experienced migration outflows above trend, and those with negative residuals experienced outflows below trend. We then run the baseline parametric regressions but successively focus on only those district-year observations that have a minimum percentile rank of at least X. The left-most observation for each graph shows the baseline estimates where no observations are dropped. The next observation requires the observations to come from district-years with studentized outflow residuals ranked at least at the 10th percentile within a district etc. 90% confidence bands constructed from standard errors clustered at origin- and destination-levels are shown.

Figure G.6: Combined strategies, country-level outflow anomalies, household fixed effects, 2SLS



This figure shows point estimates from parametric regression estimates of exposure effects concentrating, successively, on country-years that experienced anomalously larger and large migration outflows and instrumenting the treatment Δ_{odb} with the predicted treatment $\hat{\Delta}_{ob}$ from the shift-share instrument (reduced form regressions). Each point represents the marginal effect of one additional year of treatment (in the relevant age-at-move range) with $\hat{\Delta}_{ob}=1$. Panel (a) shows the coefficient estimates for ages-at-move 1-4, panel (b) those for 5-11, and (c) those for 12-18. To identify outflow-years as anomalous, we construct a country-year migration outflow panel (we sum, country-by-country, for each year, the migration outflows from all districts). We then regress these migration outflows on a constant and a linear time trend and obtain the studentized residuals from the regression. We then rank these residuals, within a country, from lowest to highest. Country-years with high positive residuals experienced migrant movements above trend, and those with negative residuals experienced movements below trend. We then run the the baseline parametric regressions but successively focus on only those country-year observations that have a minimum percentile rank of at least X. The left-most observation for each graph shows the baseline estimates where no observations are dropped. The next observation requires the observations to come from country-years with studentized outflow residuals ranked at least at the 10th percentile within a country etc. All regressions include household fixed effects. 90% confidence bands constructed from standard errors clustered at origin- and destination-levels are shown.

Figure G.7: Combined strategies, country-level outflow anomalies, observational estimates, 2SLS



This figure shows point estimates from parametric regression estimates of exposure effects concentrating, successively, on country-years that experienced anomalously larger and large migration outflows and instrumenting the treatment Δ_{odb} with the predicted treatment $\hat{\Delta}_{ob}$ from the shift-share instrument (reduced form regressions). Each point represents the marginal effect of one additional year of treatment (in the relevant age-at-move range) with $\hat{\Delta}_{ob}=1$. Panel (a) shows the coefficient estimates for ages-at-move 1-4, panel (b) those for 5-11, and (c) those for 12-18. To identify outflow-years as anomalous, we construct a country-year migration outflow panel (we sum, country-by-country, for each year, the migration outflows from all districts). We then regress these migration outflows on a constant and a linear time trend and obtain the studentized residuals from the regression. We then rank these residuals, within a country, from lowest to highest. Country-years with high positive residuals experienced migrant movements above trend, and those with negative residuals experienced movements below trend. We then run the the baseline parametric regressions but successively focus on only those country-year observations that have a minimum percentile rank of at least X. The left-most observation for each graph shows the baseline estimates where no observations are dropped. The next observation requires the observations to come from country-years with studentized outflow residuals ranked at least at the 10th percentile within a country etc. 90% confidence bands constructed from standard errors clustered at origin- and destination-levels are shown.

References

- Barro, Robert J. and Jong Wha Lee. 2013. "A New Data Set of Educational Attainment in the World, 1950–2010." *Journal of Development Economics* 104 (C):184–198.
- Besley, Timothy and Marta Reynal-Querol. 2014. "The Legacy of Historical Conflict: Evidence from Africa." *American Political Science Review* 108 (2):319–336.
- Cagé, Julia and Valeria Rueda. 2016. "The Long-Term Effects of the Printing Press in Sub-Saharan Africa." *American Economic Journal: Applied Economics* 8 (3):69–99.
- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." *The Annals of Statistics* 32 (2):407–499.
- Jedwab, Remi and Alexander Moradi. 2016. "The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa." *The Review of Economics and Statistics* 98 (2):268–284.
- Jedwab, Remi and Adam Storeygard. 2018. "The Average and Heterogeneous Effects of Transportation Investments: Evidence from Sub-Saharan Africa." *mimeo* .
- Kiszewski, Anthony, Andrew Mellinger, Andrew Spielman, Pia Malaney, Sonia Ehrlich, and Jeffrey Sachs. 2004. "A Global Index Representing the Stability of Malaria Transmission." *American Journal of Tropical Medicine and Hygiene* 70 (5):486–98.
- Klein Goldewijk, Kees, Arthur Beusen, and Peter Janssen. 2010. "Long-Term Dynamic Modeling of Global Population and Built-up Area in a Spatially Explicit Way: HYDE 3.1." *The Holocene* 20 (4):565–573.
- Lujala, Päivi, Nils Petter Gleditsch, and Elisabeth Gilmore. 2005. "A Diamond Curse? Civil War and a Lootable Resource." *Journal of Conflict Resolution* 49 (4):538–562.
- Lujala, Päivi, Jan Ketil Rød, and Nadia Thieme. 2007. "Fighting over Oil: Introducing A New Dataset." *Conflict Management and Peace Science* 24 (3):239–256.
- Murdock, George Peter. 1967. "Ethnographic Atlas: A Summary." *Ethnology* 6 (2):109–236.
- Nunn, Nathan. 2010. "Religious Conversion in Colonial Africa." *American Economic Review* 100 (2):147–152.
- Nunn, Nathan and Diego Puga. 2012. "Ruggedness: The Blessing of Bad Geography in Africa." *The Review of Economics and Statistics* 94 (1):20–36.
- Ramankutty, Navin, Jonathan Foley, John Norman, and Kevin McSweeney. 2002. "The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change." *Global Ecology and Biogeography* 11 (5):377–392.
- Survey, U.S. Geological. 1996. "GTOPO30." *Sioux Falls, SD: U.S. Geological Survey Center for Earth Resources Observation and Science* .
- Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1):267–288.