

Convergence or Reversal? The Evolution of the Gender Gap in Education in Sub-Saharan Africa

Replication package for the paper by Hardi Ahmed, Damien de Walque, and Carolina Lopez.

Overview

This package replicates all tables, figures, and maps in the paper and its online appendix. The analysis covers individuals born 1970--2009 across 35 Sub-Saharan African countries, using DHS and MICS microdata to study trends in gender gaps in primary attendance, secondary attendance, and years of education, as well as PASEC and HLO data for learning outcomes.

All statistical analysis is conducted in Stata 18. The pipeline runs in two stages: (1) data preparation and analysis in Stata, producing all regression tables, descriptive statistics, coefficient-plot figures, country-level trend figures, and country-level CSV files; and (2) map production in R 4.3, reading the CSV files and writing the five choropleth map figures. **All microdata (DHS, MICS, and PASEC) are included in this replication package.** No separate data download is required.

Data Sources

This paper uses six data sources. **All datasets are included in this replication package.** The HLO database, Africa shapefiles, population data, and World Bank education duration data are publicly available and included. The DHS, MICS, and PASEC microdata are also included; these are normally restricted-access datasets requiring registration or an application, so recipients of this package should not redistribute the raw files.

1. DHS Surveys

Source: ICF / The DHS Program

Access: Included in this package. DHS data are normally available following free registration and a data use request at dhsprogram.com. Recipients of this package should not redistribute the raw files.

File location: data/datain/DHSMICS/dhs/<country>/, preserving the original DHS subfolder structure (e.g., data/datain/DHSMICS/dhs/benin/benin_dhs_2017/BJIR71DT/BJIR71FL.DTA). IR (Individual Recode, women aged 15--49) and MR (Men's Recode) files in Stata .DTA format are included for all 27 countries.

Countries covered by DHS (27 countries): Angola, Benin, Burkina Faso, Burundi, Cameroon, Côte d'Ivoire, Gabon, Gambia, Ghana, Guinea, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Nigeria, Rwanda, Senegal, Sierra Leone, South Africa, Tanzania, Uganda, Zambia, Zimbabwe. Angola and Mauritania are cleaned by a separate script (AngolaMauritania.do) due to differences in DHS file naming conventions. The complete list of survey rounds used per country is in Table 1.

2. MICS Surveys

Source: UNICEF (Multiple Indicator Cluster Surveys)

Access: Included in this package. MICS data are normally available following free registration at mics.unicef.org. Recipients of this package should not redistribute the raw files.

File location: data/datain/DHSMICS/mics/MICS/, with one subfolder per country preserving the original UNICEF folder names:

```
data/datain/DHSMICS/mics/MICS/
├── Central African Republic MICS6 Datasets/
├── Chad MICS6 Datasets/
├── Comoros MICS6 SPSS Datasets/
├── DR Congo MICS6 SPSS Datafiles/
├── Eswatini MICS6 SPSS Datasets/
├── Guinea Bissau MICS6 Datasets/
├── Sao Tome and Principe MICS6 Datasets/
└── Togo MICS6 SPSS Datasets/
```

Note: MICS data are provided in SPSS format (.sav). Before running the main pipeline, run Code/Stata/DataPrep/DHSMICS/mics/SPSSStata.do once to convert them to .dta format.

Countries covered by MICS (8 countries): Central African Republic, Chad, Comoros, Congo, Dem. Rep., Eswatini, Guinea-Bissau, São Tomé and Príncipe, Togo. These countries are used because no later DHS survey is available for them.

3. PASEC Microdata

Source: CONFEMEN (Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN)

Access: Included in this package. PASEC microdata are normally restricted and available on request at pasec.confemen.org. Recipients of this package should not redistribute the raw files.

File location: data/datain/Learning/Pasec/, with one file per wave and grade level:

```
data/datain/Learning/Pasec/
├── PASEC2014_GRADE2.dta
├── PASEC2014_GRADE6.dta
├── PASEC2019_GRADE2.dta
└── PASEC2019_GRADE6.dta
```

Countries covered by PASEC (14 Francophone SSA countries): Benin, Burkina Faso, Burundi, Cameroon, Congo, Côte d'Ivoire, Gabon, Guinea, Madagascar, Niger, Congo Dem. Rep., Senegal, Chad, and Togo. PASEC is used as the primary source for learning outcomes for these countries; HLO is used for all remaining countries (see Section 4).

4. HLO Database (Harmonized Learning Outcomes)

Source: World Bank **Access:** Publicly available; **included in this package.** No download required.

File location: data/datain/Learning/WB/ --- contains hlo_database.xlsx (the raw database) and learning.dta (the Stata version used by the pipeline). Both are included.

Countries covered by HLO (after dropping PASEC countries): Angola, Botswana, Comoros, Eswatini, Ethiopia, Gambia, Ghana, Kenya, Lesotho, Liberia, Malawi, Mali, Mauritania, Mauritius, Mozambique, Namibia, Nigeria, Rwanda, Seychelles, Sierra Leone, South Africa, South Sudan, Sudan, Tanzania, Uganda, Zambia, Zimbabwe (27 countries). In learning_gap.do, the 14 PASEC countries are dropped from HLO so that each

country appears in only one source. The combined PASEC + HLO sample covers 39 countries in reading and 28 in mathematics.

5. Population Data

Source: World Bank

Access: Publicly available; included in this package.

File location: data/datain/Other/ (included). No download required.

6. World Bank Education Duration Data

Source: World Bank (used to define primary and secondary cycle lengths by country in MICS)

Access: Publicly available; included in this package.

File location: data/datain/DHSMICS/mics/WBEDUC/ (included). No download required.

7. Africa Shapefiles

Source: Included in this package.

File location: maps/Africa_Boundaries-shp/. Required by all R map scripts.

Software Requirements

Stata

Version: Stata 18 (MP or SE). The code uses reghdfe, which requires Stata 14 or later; Stata 18 is recommended.

Required user-written packages (install via `ssc install <package>`):

Package Purpose Install command

reghdfe High-dimensional ssc install reghdfe fixed effects OLS

ftools Required ssc install ftools dependency for
reghdfe

estout Export regression ssc install estout tables to LaTeX

coefplot Coefficient plots ssc install coefplot

listtex Export frequency ssc install listtex tables to LaTeX

require Check package ssc install require installation

To install all at once, run:

```
ssc install reghdfe
ssc install ftools
ssc install estout
ssc install coefplot
```

```
ssc install listtex
ssc install require
```

R

Version: R 4.3 or later.

Required packages (install via `install.packages()`):

Package Version tested Purpose

sf 1.0-14 Read and manipulate shapefiles

ggplot2 3.4.4 Map and figure production

dplyr 1.1.4 Data manipulation

cowplot 1.1.3 Multi-panel figure layout

scales 1.3.0 Axis and color scaling

To install all at once, run in R:

```
install.packages(c("sf", "ggplot2", "dplyr", "cowplot", "scales"))
```

Directory Structure

Top-level layout:

```
GenderGapReplication/
├── README.md
├── code/           ← all Stata and R scripts
├── data/           ← raw inputs (datain/) and processed outputs (dataout/)
├── tables/         ← LaTeX (.tex) and CSV outputs
├── figures/        ← PNG figures
└── maps/           ← Africa shapefiles
```

Code:

```
code/
├── Stata/
│   ├── Master.do      ← main entry point
│   ├── DataPrep/
│   │   ├── PrepMaster.do ← runs all data preparation
│   │   └── DHSMICS/
```

```

├── CombinedDHSMICS.do  ← merges DHS + MICS
├── dhs_data/
│   ├── DHSMaster.do
│   ├── DHS.do          ← cleans all 27 DHS surveys
│   └── AngolaMauritania.do
├── mics/
│   ├── MICSMaster.do
│   ├── MICS.do         ← cleans all 8 MICS countries
│   └── SPSSStata.do    ← converts .sav to .dta (run once)
├── Pasec/
│   └── PasecClean.do
├── Analysis/
│   ├── Analysis.do      ← calls all sub-masters
│   ├── 1970/           ← full sample (born 1970–2009)
│   │   ├── Master1970.do
│   │   ├── Surveys.do   → Table 1
│   │   ├── SummaryStatistics.do → Table 2
│   │   ├── BaseEducSSA.do → Table 3 + Figure 1
│   │   ├── AbsoluteComparisonsAggregate.do → Figure 8
│   │   ├── AbsoluteComparisonsCountry.do → Figures A-7–A-10
│   │   ├── CountriesCohort.do → CSV for R maps (Figures 2–3)
│   │   ├── CountriesComparison.do → Table A-4
│   │   ├── ResidenceEducSSA.do → Table 5 + Figure 5
│   │   ├── WealthEducSSA.do → Table 6 + Figure 6
│   │   ├── MuslimChristian.do → Table 7 + Figure 7
│   │   └── NonMigrants.do → Table A-10 + Figure A-11
│   ├── 2000/           ← younger sample (born 2000–2009)
│   │   └── [mirrors 1970/] → Appendix tables and figures
│   └── Learning/
│       ├── MasterLearning.do
│       ├── learning_gap.do → Table A-6 + CSV for Figure 4
│       └── pasec_regression.do → Table 4
├── RFile/
│   ├── Master.R        ← main R entry point
│   ├── 1970Gap.R       → Figure 2
│   ├── 1970Cohort.R    → Figure 3
│   ├── 2000Gap.R       → Figure A-2
│   ├── 2000Cohort.R    → Figure A-3
│   └── LearningGap.R   → Figure 4

```

Data:

```

data/
├── datain/
│   ├── DHSMICS/
│   │   ├── dhs/      ← DHS surveys by country (included)
│   │   └── mics/     ← MICS6 .sav files + WBEDUC data (included)
│   ├── Learning/
│   │   ├── Pasec/    ← PASEC microdata (included)
│   │   └── WB/       ← HLO database (included)
│   └── Other/        ← population data (included)
└── dataout/         ← produced by the pipeline

```

```

├── DHSMICS/
│   ├── dhs/      ← DHS.dta
│   ├── mics/     ← MICS.dta
│   └── combined/ ← DHSMICS.dta (main analysis dataset)
└── Learning/
    └── Pasec/     ← pasec_clean.dta

```

Outputs:

tables/

```

├── tex/ — 1970/ 2000/ Learning/ ← LaTeX tables
└── csv/ — 1970/ 2000/ Learning/ ← CSVs for R maps

```

figures/ — 1970/ 2000/ Learning/ ← PNG figures

Maps input:

maps/

```

└── Africa_Boundaries-shp/ ← shapefile (included)

```

Files and Contents

A. Data Preparation (Code/Stata/DataPrep/)

PrepMaster.do --- Calls CombinedDHSMICS.do and PasecClean.do in sequence. Produces all processed datasets needed for analysis.

DHSMICS/CombinedDHSMICS.do --- Calls DHSMaster.do and the MICS cleaning scripts, then appends DHS.dta and MICS.dta into a single file. Saves the merged dataset as data/dataout/DHSMICS/combined/DHSMICS.dta. This is the main analysis dataset used by all analysis scripts.

DHSMICS/dhs_data/DHS.do --- Reads individual (IR) and men's (MR) recode files for each DHS country. Harmonizes variable names and coding across survey rounds (education years, primary attendance, secondary attendance, birth year, age, rural, wealth quintile, religion, household size, migration status). Stacks all countries and saves data/dataout/DHSMICS/dhs/DHS.dta. Angola and Mauritania are handled in AngolaMauritania.do due to differences in file naming.

DHSMICS/mics/SPSSStata.do --- Scans data/datain/DHSMICS/mics/MICS/ recursively and converts every UNICEF .sav file to Stata .dta format. **Must be run before** MICS.do. Only needs to be run once per machine.

DHSMICS/mics/MICS.do --- Reads the converted MICS6 .dta files for all 8 MICS countries (CAR, Chad, Comoros, Congo Dem. Rep., Eswatini, Guinea-Bissau, São Tomé and Príncipe, Togo). Harmonizes variable names and coding to match DHS conventions (education years, primary/secondary attendance, birth year, age, rural, wealth quintile, religion, household size). Merges in population data and education duration data from WBEDUC. Saves data/dataout/DHSMICS/mics/MICS.dta.

Pasec/PasecClean.do --- Reads the raw PASEC wave files, constructs plausible value averages for reading and mathematics, harmonizes country codes, and assigns survey weights. Saves data/dataout/Learning/Pasec/pasec_clean.dta.

B. Analysis --- Full Sample (Code/Stata/Analysis/1970/)

All scripts load DHSMICS.dta unless noted. Each saves output to tables/tex/1970/ (LaTeX), tables/csv/1970/ (CSV), or figures/1970/ (PNG) as noted.

Surveys.do --- Tabulates the number of surveys, survey rounds, and sample sizes by country and data source (DHS vs. MICS). Exports 1970country_surveys_summary.tex (Table 1).

SummaryStatistics.do --- Computes weighted summary statistics for the full sample, separately for males and females. Key outcomes: attended_primary, attended_secondary (restricted to ages ≥ 16), educ_years (restricted to ages ≥ 25). Covariates: rural, poor, christian, muslim, above_median_size. Exports 1970summary_statistics.tex (Table 2).

BaseEducSSA.do --- Estimates OLS regressions of educational outcomes on an indicator for female, with country, survey year, birth year, and data source fixed effects, as well as a household fixed effects specification. Exports 1970base_results.tex (Table 3). Also produces 1970cohort_base.png (Figure 1): coefficients on female by five-year birth cohort.

AbsoluteComparisonsAggregate.do --- Computes gender-specific means by five-year birth cohort for primary attendance, secondary attendance, and years of education, applying censoring corrections to avoid composition bias in the youngest cohorts. Produces 1970gender_levels_by_cohort_censorfix.png (Figure 8, male and female education levels by cohort with censoring correction applied). Note: the censoring threshold requires that the youngest member of each cohort is at least 16 (for primary/secondary) or 25 (for years) at survey date.

AbsoluteComparisonsCountry.do --- Repeats the cohort-level means decomposition by country. Saves per-country PNG figures to figures/1970/CountryTrends/. These country-level trend plots are assembled into Figures A-7--A-10 in the Online Appendix.

CountriesCohort.do --- Computes the female--male gap in each educational outcome by country and birth cohort. Saves tables/csv/1970/attainment_gap.csv and tables/csv/1970/1970Combined.csv. These CSV files are read by the R map scripts (1970Gap.R, 1970Cohort.R) to produce Figures 2 and 3.

CountriesComparison.do --- Runs country-by-country regressions of educational outcomes on female. Exports the coefficient table as tables/tex/1970/attainment_gap.tex (Table A-4) and writes the gap CSV used for map color coding.

ResidenceEducSSA.do --- Estimates OLS regressions separately for rural and urban sub-samples, and a pooled regression with a female $\times$ rural interaction term. Exports 1970residence.tex (Table 5) and 1970residence.png (Figure 5).

WealthEducSSA.do --- Estimates OLS regressions separately for poor and non-poor sub-samples, and a pooled regression with a female $\times$ poor interaction. Exports 1970wealth.tex (Table 6) and 1970wealth.png (Figure 6).

MuslimChristian.do --- Estimates OLS regressions separately for Muslim, Christian, and other-religion sub-samples, and pooled regressions with religion interaction terms. Exports 1970religion.tex (Table 7) and 1970MuslimsChristians.png (Figure 7).

NonMigrants.do --- Repeats the main OLS specification restricting to individuals who have never migrated (never_migrated == 1). Exports non_migrants_base_results.tex (Table A-10) and non_migrants_cohort_base.png (Figure A-11).

C. Analysis --- Younger Sample (Code/Stata/Analysis/2000/)

Scripts mirror those in 1970/ but restrict to individuals born 2000--2009 (ages 15--21 at survey). Each produces an analogous set of output files with 2000 prefix. Key outputs:

Script Output

SummaryStatistics.do 2000summary_statistics.tex (Table A-2)

BaseEducSSA.do 2000base_results.tex + 2000cohort_base.png

AbsoluteComparisonsAggregate.do 2000gender_levels_by_cohort.png

CountriesCohort.do csv/2000/attainment_gap.csv (input for R maps)

CountriesComparison.do 2000attainment_gap.tex

ResidenceEducSSA.do 2000residence.tex + 2000residence.png

WealthEducSSA.do 2000wealth.tex + 2000wealth.png

MuslimChristian.do 2000religion.tex + 2000MuslimsChristians.png

Surveys.do 2000country_surveys_summary.tex

D. Analysis --- Learning Outcomes (Code/Stata/Analysis/Learning/)

learning_gap.do --- Loads pasesc_clean.dta and the HLO database. Computes gender gaps in reading and mathematics test scores by country: PASEC gaps from plausible value means, HLO gaps from the World Bank database. Exports tables/tex/Learning/learning_gap_sources.tex (Table A-6) and tables/csv/Learning/learning_gap.csv (read by LearningGap.R to produce Figure 4).

pasesc_regression.do --- Estimates gender gaps in PASEC test scores using OLS with school, grade, wave, and country fixed effects (Rubin's Rules applied across five plausible values). Exports tables/tex/Learning/pasesc_results.tex (Table 4).

E. Map Production (Code/RFile/)

All R scripts read CSV files produced by Stata and the Africa shapefile from maps/Africa_Boundaries-shp/. They write PNG figures to figures/.

Script Input CSV Output figure Paper exhibit

1970Gap.R csv/1970/attainment_gap.csv figures/1970/1970AttainmentGapSSA.png Figure 2

1970Cohort.R csv/1970/attainment_gap.csv figures/1970/1970ReversalCohortSSA.png Figure 3

2000Gap.R csv/2000/attainment_gap.csv figures/2000/2000AttainmentGapSSA.png Figure A-2

2000Cohort.R csv/2000/attainment_gap.csv figures/2000/2000ReversalYearSSA.png Figure A-3

LearningGap.R csv/Learning/learning_gap.csv
 figures/Learning/LearningGapSSA.png Figure 4

How to Run

Important: Steps 1--2 require the raw DHS, MICS, and PASEC data to have been downloaded and placed in the correct folders (see *Data Sources* above). Steps 3 onward can run once Step 2 is complete.

Step 1 --- Install software dependencies

Install Stata packages:

```
ssc install reghdfe
ssc install ftools
ssc install estout
ssc install coefplot
ssc install listtex
ssc install require
```

Install R packages:

```
install.packages(c("sf", "ggplot2", "dplyr", "cowplot", "scales"))
```

Step 2 --- Set the root path (Stata)

The root path is **auto-detected by Stata username** --- no manual change is needed if your username is already registered in Master.do. If you are a new user, open Code/Stata/Master.do and add an else if block with your username (printed in the error message on first run) and the path to this folder:

```
else if "`c(username)'" == "your_username" {
    global root "/path/to/GenderGapReplication"
}
```

The same block must be added to any sub-file you want to run directly (or always run via Master.do, which sets the root for all downstream scripts).

For R, the root path is also **auto-detected by username** in Code/RFile/Master.R. If you are a new user, add an else-if block with your username and path, mirroring the Stata block above:

```
} else if (user == "your_username") {
  root <- "/path/to/GenderGapReplication"
}
```

Step 3 --- Data

All raw microdata (DHS, MICS, PASEC) are included in this package under data/datain/. No data placement is required. The HLO database, population data, shapefiles, and World Bank education duration data are also included.

Step 4 --- Run data preparation and analysis (Stata)

Open Stata, navigate to the package root, and run:

```
do "Code/Stata/Master.do"
```

This executes the full pipeline in order:

1. **Data preparation** (PrepMaster.do)

- Cleans and harmonizes DHS surveys across all 27 DHS countries → DHS.dta
- Cleans MICS surveys from UNICEF for 8 non-DHS countries → MICS.dta
- Merges DHS and MICS into a single cross-country dataset → DHSMICS.dta
- Cleans PASEC data → pasec_clean.dta

2. **Full sample analysis** --- 1970/Master1970.do calls all scripts in Analysis/1970/, producing all tables and country-trend figures for individuals born 1970--2009. CSV files for R maps are also written at this stage.

3. **Younger sample analysis** --- 2000/Master2000.do calls all scripts in Analysis/2000/, producing appendix tables and figures for individuals born 2000--2009.

4. **Learning outcomes** --- Learning/MasterLearning.do calls learning_gap.do and pasec_regression.do, producing learning gap tables and the CSV file for the R map.

Runtime: The pipeline prints per-step and total elapsed time automatically at the end of each stage. On a MacBook Pro Max running Stata 19.5, the full pipeline (data preparation + analysis) runs in approximately 20--40 minutes. Runtimes will vary on other hardware.

Step 5 --- Produce maps (R)

After Step 4, all CSV inputs for the R scripts are in place. Run in R:

```
source("Code/RFile/Master.R")
```

Or run each script individually if only a subset of maps needs to be regenerated. Each script is self-contained and reads only the CSV files produced by Stata.

Runtime: Each R script takes approximately 1--3 minutes.

Variable Construction

The main analysis dataset (DHSMICS.dta) is a harmonized individual-level file combining DHS and MICS microdata. Each row represents one individual. Key variables:

Variable Type Description

female Binary 1 = female, 0 = male

attended_primary Binary Ever attended primary school (1 = yes, 0 = no)

attended_secondary Binary Ever attended secondary school (1 = yes, 0 = no). Regressions restrict to individuals aged ≥ 16 at survey to avoid censoring.

educ_years Continuous Total years of completed education. Regressions restrict to individuals aged ≥ 25 to allow completion. Values above 25 recoded to missing.

birth_year Integer Respondent's year of birth

hhage Integer Respondent's age at time of survey

cohort Integer Five-year birth cohort label (1970, 1975, ..., 2005). Created within analysis scripts.

rural Binary 1 = rural residence, 0 = urban

poor Binary 1 = bottom two quintiles of the household wealth index, 0 = otherwise

christian Binary 1 = respondent identifies as Christian

muslim Binary 1 = respondent identifies as Muslim

above_median_size Binary 1 = household size above sample median

never_migrated Binary 1 = respondent has never migrated (from DHS migration module; missing for MICS)

country_string String Country name

country_code String ISO-3 country code

country_phase String Country-phase identifier (DHS country code + survey round)

data String Data source: "DHS" or "MICS"

survey_year Integer Calendar year of the survey

cluster_id String Survey cluster identifier (used for clustering standard errors)

household_id String Unique household identifier (used for household fixed effects)

sample_strata String Survey stratum (used for stratified standard errors)

population_weight1 Float Survey sampling weight (normalized to sum to one within country-round)

religion Categorical Religion: 1 = Islam, 2 = Christianity, 3 = Other

wealth_index Categorical DHS/MICS wealth quintile (1 = poorest, 5 = wealthiest). poor is derived from this.

Constructed variables created within analysis scripts (not saved to DHSMICS.dta): -

attended_secondary_r = attended_secondary restricted to hhage >= 16 - educ_years_r = educ_years restricted to hhage >= 25 - fem_rural = female × rural interaction (in ResidenceEducSSA.do) - fem_poor = female × poor interaction (in WealthEducSSA.do) - cohort_code = numeric cohort code 1--8 (created in aggregate comparison scripts)

Censoring correction (aggregate figures only): The aggregate level figures apply a censoring restriction to avoid composition bias in younger cohorts: for secondary attendance, the cohort is included only if the youngest person in the cohort (i.e., those born in the last year of the cohort window) was at least 16 at the time of the survey. This means the 2000--04 and 2005--09 cohorts are censored out for secondary attendance in many survey rounds. See AbsoluteComparisonsAggregate.do for implementation.

Relevant Sample

Countries: 35 Sub-Saharan African countries (full sample, born 1970--2009). DHS microdata is used for 27 countries; UNICEF MICS microdata is used for the 8 remaining countries without DHS coverage (Central African Republic, Chad, Comoros, Congo Dem. Rep., Eswatini, Guinea-Bissau, São Tomé and Príncipe, Togo). Angola and Mauritania are within the DHS pipeline but handled by a separate cleaning script. The younger-cohort sample (born after 1999) covers the same 35 countries (27 DHS + 8 MICS).

Individuals: Respondents aged 15--49 at the time of the survey (DHS/MICS standard sampling frame), born between 1970 and 2009. The **full sample** analysis uses all individuals born 1970--2009 (ages 15--49). The **younger sample** analysis restricts to individuals born 2000--2009 (ages 15--21 at a 2021 survey, or similar).

Survey years: The most recent available survey round per country is used. Survey years by country are documented in Table 1.

Learning outcomes sample: PASEC covers 14 Francophone SSA countries. The HLO database covers a broader set of countries and years, combined with PASEC to produce the cross-country learning gap estimates.

Map Color Scheme

All gap maps (Figures 2, 4, A-2) and cohort maps (Figures 3, A-3) use the same three-category classification per country-outcome:

Color Meaning

Blue Female advantage: females exceed males in at least one cohort, and the maximum absolute female advantage exceeds 0.4 percentage points (or years). The gap map shows the magnitude at the earliest qualifying cohort; the cohort map shows the birth year of that cohort.

Red Male advantage: no cohort shows a female advantage exceeding 0.4 pp, but at least one cohort shows a male advantage. The gap map shows the magnitude at the most recent cohort.

Yellow No meaningful gap: the gap never exceeds 0.4 pp in either direction across any cohort.

Gray No data available for that country-outcome combination.

List of Exhibits

The table below maps each paper exhibit to the script that produces it and the output file name.

Main Text

Exhibit Description Produced by Output file

Table 1 Survey coverage and 1970/Surveys.do tex/1970/1970country_surveys_summary.tex sample size (full sample, 35 countries)

Table 2 Summary statistics 1970/SummaryStatistics.do tex/1970/1970summary_statistics.tex (full sample)

Table 3 Gender and education 1970/BaseEducSSA.do tex/1970/1970base_results.tex in SSA --- main OLS (full sample)

Table 4 Gender gap in test Learning/pasec_regression.do tex/Learning/pasec_results.tex scores: PASEC evidence

Table 5 Gender and education 1970/ResidenceEducSSA.do tex/1970/1970residence.tex in SSA by geographic area (rural/urban)

Table 6 Gender and education 1970/WealthEducSSA.do tex/1970/1970wealth.tex in SSA by household wealth

Table 7 Gender and education 1970/MuslimChristian.do tex/1970/1970religion.tex in SSA by religion

Figure 1 Gender gap in 1970/BaseEducSSA.do figures/1970/1970cohort_base.png educational attainment by birth cohort in SSA

Figure 2 Geographic R: 1970Gap.R figures/1970/1970AttainmentGapSSA.png distribution of gender gaps in education (full sample maps)

Figure 3 Timing of gender gap R: 1970Cohort.R figures/1970/1970ReversalCohortSSA.png reversals in education (full sample)

Figure 4 Geographic R: LearningGap.R figures/Learning/LearningGapSSA.png distribution of gender gaps in test scores

Figure 5 Educational attainment 1970/ResidenceEducSSA.do figures/1970/1970residence.png by cohort and residence (rural vs. urban)

Figure 6 Educational attainment 1970/WealthEducSSA.do figures/1970/1970wealth.png by cohort and household wealth

Figure 7 Educational attainment 1970/MuslimChristian.do figures/1970/1970MuslimsChristians.png by cohort and religion
(Christian vs. Muslim)

Figure 8 Male and female 1970/AbsoluteComparisonsAggregate.do figures/1970/1970gender_levels_by_cohort_censorfix.png educational attainment by birth cohort
(censoring-corrected levels)

Online Appendix

Exhibit Description Produced by Output file

Table A-1 Survey coverage and 2000/Surveys.do tex/2000/2000country_surveys_summary.tex sample size (younger sample, born after 1999, 35 countries)

Table A-2 Summary statistics 2000/SummaryStatistics.do tex/2000/2000summary_statistics.tex (younger sample)

Table A-3 Gender and 2000/BaseEducSSA.do tex/2000/2000base_results.tex education in SSA
--- main OLS
(younger sample)

Table A-4 Gender gap in 1970/CountriesComparison.do tex/1970/attainment_gap.tex educational attainment by country (full sample)

Table A-5 Gender gap in 2000/CountriesComparison.do tex/2000/2000attainment_gap.tex educational attainment by country (younger sample)

Table A-6 Data sources and Learning/learning_gap.do tex/Learning/learning_gap_sources.tex gender gap in learning outcomes
(PASEC + HLO)

Table A-7 Gender and 2000/ResidenceEducSSA.do tex/2000/2000residence.tex education in SSA by geographic area
(younger sample)

Table A-8 Gender and 2000/WealthEducSSA.do tex/2000/2000wealth.tex education in SSA by household wealth
(younger sample)

Table A-9 Gender and 2000/MuslimChristian.do tex/2000/2000religion.tex education in SSA by religion (younger sample)

Table A-10 Selective migration 1970/NonMigrants.do tex/1970/non_migrants_base_results.tex robustness (non-migrants only)

Figure A-1 Gender gap in 2000/BaseEducSSA.do figures/2000/2000cohort_base.png educational attainment by birth year in SSA (younger sample, annual)

Figure A-2 Geographic R: 2000Gap.R figures/2000/2000AttainmentGapSSA.png distribution of gender gaps (younger sample maps)

Figure A-3 Timing of gender R: 2000Cohort.R figures/2000/2000ReversalYearSSA.png gap reversals (younger sample)

Figure A-4 Educational 2000/ResidenceEducSSA.do figures/2000/2000residence.png attainment by birth year and residence (younger sample)

Figure A-5 Educational 2000/WealthEducSSA.do figures/2000/2000wealth.png attainment by birth year and wealth (younger sample)

Figure A-6 Educational 2000/MuslimChristian.do figures/2000/2000MuslimsChristians.png attainment by birth year and religion (younger sample)

Figures Country-level 1970/AbsoluteComparisonsCountry.do figures/1970/CountryTrends/ A-7--A-10 trends in gender gaps and attainment levels

Figure A-11 Selective 1970/NonMigrants.do figures/1970/non_migrants_cohort_base.png migration: educational attainment by cohort

Exhibit numbers follow the current draft of the paper. If numbers change in revision, update only the label column of this table.

Data Availability Statement

Statement of Rights

The authors confirm that they have legitimate access to and permission to use all datasets included in this replication package. The DHS microdata were obtained through the DHS Program's standard registration and data use request process. The MICS microdata were obtained through UNICEF's standard registration process. The PASEC microdata were obtained under a data sharing agreement with CONFEMEN. The HLO database, World Bank population data, World Bank education duration data, and Africa shapefiles are publicly available datasets reproduced here in accordance with their respective access conditions.

All data used in this paper are included in this replication package. The DHS and MICS microdata are normally available following free registration; PASEC microdata are normally available on application. The remaining datasets are publicly available. Recipients of this package should not redistribute the raw DHS, MICS, or PASEC files to third parties.

1. DHS Surveys

- **Filename(s):** IR (Individual Recode) and MR (Men's Recode) .DTA files for 27 countries, located under data/datain/DHSMICS/dhs/<country>/<survey_round>/. File names follow the original DHS naming convention (e.g., BJIR71FL.DTA for Benin IR Round 7). The complete list of countries and rounds is in Table 1 of the paper.
- **Source:** ICF International / The DHS Program
- **URL:** dhsprogram.com. Registration is free. After registration, submit a data use request selecting each country and the most recent available survey round. Request the IR (Individual Recode, women aged 15--49) and MR (Men's Recode) files in Stata (.DTA) format. Survey rounds used per country are listed in Table 1.
- **Access date:** January 2024 -- February 2026 (surveys downloaded at different times as rounds became available)
- **Citation:** ICF. Various years. *Demographic and Health Surveys*. Rockville, Maryland, USA. Available at dhsprogram.com.
- **License:** Free registration and data use agreement required. Data are licensed for non-commercial research use only. Do not redistribute the raw files to third parties.

2. MICS Surveys

- **Filename(s):** MICS6 SPSS (.sav) datasets for 8 countries, located under data/datain/DHSMICS/mics/MICS/, one subfolder per country (e.g., "Central African Republic MICS6 Datasets/"). Original UNICEF folder and file names are preserved.
- **Source:** UNICEF
- **URL:** mics.unicef.org. Registration is free. After registration, navigate to each country page and download the MICS6 dataset package. Countries used: Central African Republic, Chad, Comoros, Congo Dem. Rep., Eswatini, Guinea-Bissau, São Tomé and Príncipe, Togo.
- **Access date:** 21 June 2026
- **Citation:** UNICEF. Various years. *Multiple Indicator Cluster Surveys (MICS6)*. New York, USA. Available at mics.unicef.org.
- **License:** Free registration required. Data are licensed for research use only. Do not redistribute the raw files to third parties.

3. PASEC Microdata

- **Filename(s):** data/datain/Learning/Pasec/PASEC2014_GRADE2.dta, PASEC2014_GRADE6.dta, PASEC2019_GRADE2.dta, PASEC2019_GRADE6.dta
- **Source:** CONFEMEN (Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN)
- **URL:** pasec.confemen.org. Data are restricted and available on application. Contact CONFEMEN directly to request the 2014 and 2019 wave microdata (Grades 2 and 6) for the 14 Francophone SSA countries: Benin, Burkina Faso, Burundi, Cameroon, Congo, Côte d'Ivoire, Gabon, Guinea, Madagascar, Niger, Congo Dem. Rep., Senegal, Chad, Togo.
- **Access date:** 9 March 2026
- **Citation:** CONFEMEN. Various years. *PASEC: Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN*. Dakar, Senegal. Available at pasec.confemen.org.
- **License:** Restricted. Access is subject to a data sharing agreement with CONFEMEN. Do not redistribute the raw files to third parties.

4. HLO Database (Harmonized Learning Outcomes)

- **Filename(s):** data/datain/Learning/WB/hlo_database.xlsx, data/datain/Learning/WB/learning.dta (Stata version used by the pipeline)
- **Source:** World Bank
- **URL:** datacatalog.worldbank.org (search "Harmonized Learning Outcomes")
- **Access date:** 8 March 2026
- **Citation:** Angrist, N., S. Djankov, P. Goldberg, and H. Patrinos. 2021. "Measuring human capital using global learning data." *Nature* 592: 403--408.
- **License:** Publicly available. No redistribution restrictions.

5. Population Data

- **Filename(s):** data/datain/Other/population.dta, data/datain/Other/Country_ID.xlsx, data/datain/Other/75ef169b-ed2e-4a3d-96fa-85d60cd6e9b5_Data.csv, data/datain/Other/75ef169b-ed2e-4a3d-96fa-85d60cd6e9b5_Series - Metadata.csv
- **Source:** World Bank
- **URL:** data.worldbank.org/indicator/SP.POP.TOTL
- **Access date:** 8 February 2026
- **Citation:** World Bank. Various years. *World Development Indicators*. Washington, DC: World Bank. Available at data.worldbank.org.
- **License:** Publicly available under Creative Commons Attribution 4.0 International (CC BY 4.0). No redistribution restrictions.

6. World Bank Education Duration Data

- **Filename(s):** data/datain/DHSMICS/mics/WBEDUC/WBEDUCDATA.dta
- **Source:** World Bank
- **URL:** World Development Indicators, multiple indicators: [SE.PRM.DURS](#) (primary duration), [SE.SEC.DURS](#) (secondary duration), [SE.SEC.DURS.LO](#) (lower secondary duration)
- **Access date:** 27 January 2026
- **Citation:** World Bank. Various years. *World Development Indicators*. Washington, DC: World Bank. Available at data.worldbank.org.
- **License:** Publicly available under Creative Commons Attribution 4.0 International (CC BY 4.0). No redistribution restrictions.

7. Africa Shapefiles

- **Filename(s):** maps/Africa_Boundaries-shp/Africa_Boundaries.shp (and associated .cpg, .dbf, .prj, .shx files)
- **Source:** ArcGIS Hub (geoduck)
- **URL:** hub.arcgis.com/datasets/geoduck::africa-boundaries/about
- **Access date:** 28 January 2021
- **Citation:** Africa Boundaries. ArcGIS Hub. Available at hub.arcgis.com/datasets/geoduck::africa-boundaries/about.
- **License:** See ArcGIS Hub page for license terms.

References

H. Ahmed, de Walque, D., and C. Lopez "Convergence or Reversal? The Evolution of the Gender Gap in Education in Sub-Saharan Africa." Working Paper.

ICF. Various years. *Demographic and Health Surveys*. Rockville, Maryland, USA. Available at dhsprogram.com.

UNICEF. Various years. *Multiple Indicator Cluster Surveys (MICS6)*. New York, USA. Available at mics.unicef.org.

CONFEMEN. Various years. *PASEC: Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN*. Dakar, Senegal. Available at pasec.confemen.org.