Article

Measuring and forecasting progress towards the education-related SDG targets

https://doi.org/10.1038/s41586-020-2198-8

Received: 25 June 2019

Accepted: 18 March 2020

Published online: 15 April 2020

Open access

Check for updates

Joseph Friedman[1,2], Hunter York[1], Nicholas Graetz[1,3], Lauren Woyczynski[1], Joanna Whisnant[1], Simon I. Hay[1,4] & Emmanuela Gakidou[1,4]✉

Education is a key dimension of well-being and a crucial indicator of development[1–4]. The Sustainable Development Goals (SDGs) prioritize progress in education, with a new focus on inequality[5–7]. Here we model the within-country distribution of years of schooling, and use this model to explore educational inequality since 1970 and to forecast progress towards the education-related 2030 SDG targets. We show that although the world is largely on track to achieve near-universal primary education by 2030, substantial challenges remain in the completion rates for secondary and tertiary education. Globally, the gender gap in schooling had nearly closed by 2018 but gender disparities remained acute in parts of sub-Saharan Africa, and North Africa and the Middle East. It is predicted that, by 2030, females will have achieved significantly higher educational attainment than males in 18 countries. Inequality in education reached a peak globally in 2017 and is projected to decrease steadily up to 2030. The distributions and inequality metrics presented here represent a framework that can be used to track the progress of each country towards the SDG targets and the level of inequality over time. Reducing educational inequality is one way to promote a fairer distribution of human capital and the development of more equitable human societies.

The value of education is well-recognized, both as a primary human right and as a key driver of progress in economic development, health, fertility, politics, social empowerment, and human capital[3–13]. The international community recognized educational attainment as a key development priority in the Millennium Development Goals (MDGs), which became a key focus for a large variety of global actors. The education-related MDG targets focused largely on expanding primary education up to 2015[14], and great progress in this regard was seen as a result. In the SDGs—the follow-up to the MDGs with a target year of 2030—education was again highly prioritized, with a wider scope that emphasized reducing inequalities.

Increases in global schooling rates

SDG target 4.1 calls for universal primary schooling. Progress towards this goal has been, and is projected to continue to be, substantial (Fig. 1). Globally, the proportion of 25–29-year olds with at least 6 years of schooling rose from 50.1% (95% uncertainty interval: 49.3–51.0%) in 1970 to 83.2% (82.1–84.0%) in 2018 and is projected to reach 89.4% (87.4–91.0%) by 2030. Even as far back as 1970, countries in high-income regions and in eastern Europe and central Asia had on average already achieved near universal primary attainment. In the remaining regions, rates of primary attainment have risen substantially. Although this progress is to be celebrated, important gaps remain in a subset of nations that are not projected to achieve near universal levels of primary attainment by 2030, largely due to gaps in schooling among women (see Extended Data Fig. 1).

SDG target 4.1 also calls for universal secondary schooling. However, secondary attainment estimates reveal a much more heterogeneous picture. In 1970, countries generally fell into one of two categories; nearly 50% of the global population aged 25–29 residing in highly educated regions had already attained 12 years of schooling, whereas the rest of the world saw rates at or below 10%. Although global attainment of at least 12 years of schooling has risen steadily since 1970, no major world region has achieved near universal levels. All regions have seen progress, yet the inter-regional disparities remain massive in 2018 and are projected to decrease only slightly in the coming years.

SDG target 4.3 addresses tertiary education, calling for 'equal access' for all individuals. Tertiary education exhibited a substantial scale-up between 1970 and 2018 that is projected to continue in the coming decade, although global completion rates remain low. Similar to the trend in secondary education, the high-income and eastern European and central Asian regions exhibit substantially higher rates throughout the time period shown, and are projected to achieve about half of their population completing tertiary education by 2030. The remaining regions have also seen progress, with much of the growth seen after 2000. The increase is particularly notable in North Africa and the Middle East as well as in Southeast Asia, East Asia, and Oceania.

In summary, regional disparities in tertiary education completion are increasing over time and are projected to continue to do so, whereas secondary gaps are expected to decrease only slightly. The success of narrowing the global gap for primary education has not been extended to higher levels of education, which raises concerns

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA. [2]Center for Social Medicine and Humanities, University of California Los Angeles, Los Angeles, CA, USA. [3]Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Health Metrics Sciences, School of Medicine, University of Washington, Seattle, WA, USA. ✉e-mail: gakidou@uw.edu

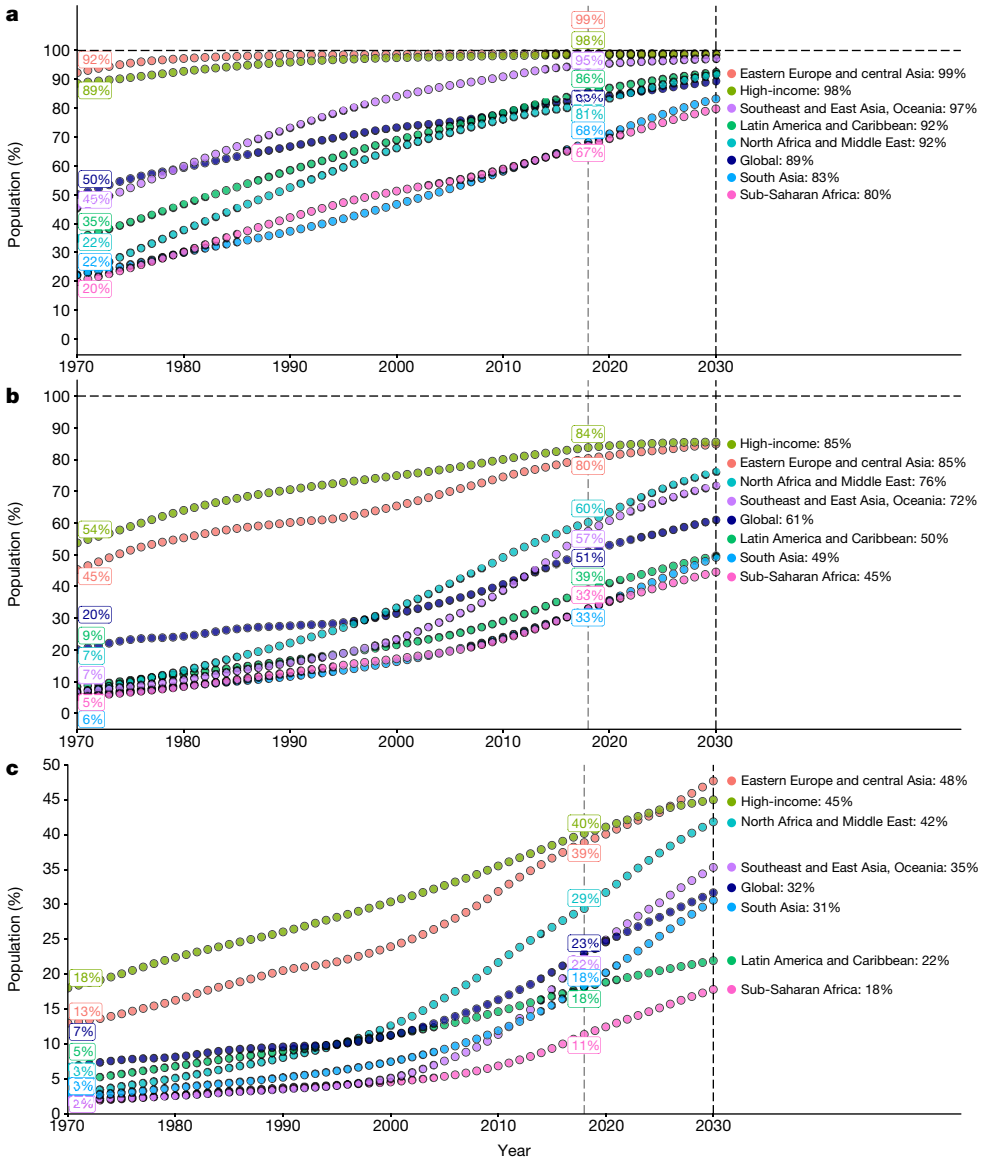


Fig. 1 | Regional attainment of primary, secondary, and tertiary schooling from 1970 to 2030. a–**c**, Attainment rates of 6+ (**a**), 12+ (**b**), and 15+ (**c**) years of schooling are shown. All trends reflect 25–29-year-old individuals separated by major world region. The vertical dashed lines indicate 2018, when the forecasts begin, and 2030, the target year for the SDGs.

about gaps in opportunities amplifying across regions in the coming decade.

Progress towards gender equity

Gender equity has been a central focus of the SDG targets. SDG target 5 calls for gender equity broadly, and target 4.5 calls for the elimination of all gender disparities in education. We find that great strides have been made in reversing educational disparities for women globally, and in all regions of the world.

To benchmark the progress of each country towards gender parity in education, we calculate the absolute gap in the mean years of schooling, and assess the contribution of primary, secondary, and tertiary schooling to these gaps (Fig. 2). In 1970, men aged 25–29 years had completed on average 1.7 (1.6–1.8) additional years of education compared with women of the same age. By 2018, this gap had nearly closed, falling to only 0.3 (−0.2–0.8) years, and is projected to reverse by 2030. Previous modelling studies of global gender differences in educational attainment that have focused on all adults 25 and older show progress, but note that women are not yet close to catching up to men[15]. By focusing only on young women and men, we show that among the most recently educated members of societies, women had in fact nearly closed the gender gap in 2018. Young men had statistically significantly higher levels of attainment compared with women, at the 95% confidence level, in 142 countries in 1970, 27 countries in 2018, and only 4 countries by 2030. For 2030, the countries in which women's education is predicted to still lag behind that of men are predominantly in sub-Saharan Africa, Southeast Asia, East Asia, and Oceania. In addition, by 2030, women are expected to achieve statistically significantly higher mean years of schooling than men in 18 countries—a tremendous reversal of the global landscape that was observed in 1970.

In absolute terms, the largest component of this reduction has been observed in primary education. In 1970, men aged 25–29 completed 0.9 (0.9–1.0) additional years of primary schooling compared with women, which fell to only 0.3 (0.2–0.4) years in 2018. This reflects progress in nearly every region; all had primary education gaps favouring men in 1970. By 2018, these gaps had shrunk by considerable margins in every region, and many disappeared entirely. Nevertheless, a small number

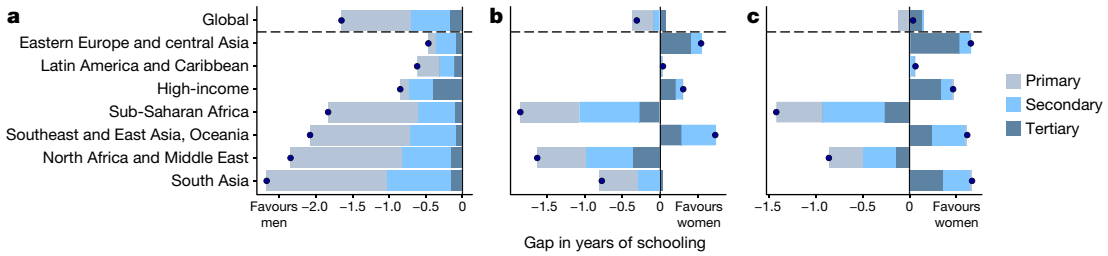


Fig. 2 | Regional gender gaps in primary, secondary, tertiary, and total schooling. a–c, The gender gap is shown for 1970 (**a**), 2018 (**b**), and 2030 (**c**). The total gap in years of schooling is represented by a dot, for individuals aged 25–29, separated by each regional group. The grey, light blue, and dark blue bars represent the contributions of primary, secondary, and tertiary schooling, respectively, to the total gender gap.

of countries are forecast to have persistent gaps in attainment of at least 6 years of schooling, largely in North Africa and the Middle East, as well as sub-Saharan Africa (Supplementary Fig. 14).

Secondary and tertiary education both show a more heterogeneous pattern, in which women are overtaking men in most regions of the world, whereas large-magnitude disparities seen in sub-Saharan Africa and North Africa and the Middle East are projected to persist. Our estimates indicate that in 2012, women aged 25–29 overtook men in the global average of tertiary attainment, and they are forecast to do so for secondary attainment in 2026. Unlike primary attainment, which has largely converged globally in a place of gender parity, women have overtaken men by substantial margins in many nations in Latin America, Asia, and Europe. This phenomenon has been reported for many nations in the Organisation for Economic Co-operation and Development (OECD)[16,17] and elsewhere[18–20], in which boys increasingly fall behind girls in schooling as nations develop. Our results indicate the commonality of this trend for many regions of the world, and show how these advances have contributed to closing the overall gender gap. Notably, our results indicate that these gaps are projected to grow with time.

Assessment of inequalities in education

Although gender equity is of crucial importance, it only captures one dimension of inequality in education. Beyond gender, SDG target 4.5 calls for broad social equity in educational attainment, across lines of ethnicity, race, socio-economic status, ability, and other identities[21]. The particular social groupings that are relevant vary across countries, but insight between countries can be gleaned by assessing the total inequality.

To facilitate benchmarking between nations and a global assessment of trends in educational inequality, we use a metric of the total within-country inequality in education, the average interpersonal difference (AID), which represents the average difference between any two individuals in a population. Results and discussion using alternative metrics of inequality, including relative measures such as the Gini coefficient, are presented in the Supplementary Information.

Globally, inequality rose steadily before peaking in 2017 with a 4.6-year (4.5–4.7) average within-country difference between any two given individuals (Fig. 3a). Subsequently, inequality has been decreasing and is projected to continue to do so up to 2030. Looking at the arc of inequality in education over time across regions and countries, a consistent Kuznets curve can be observed in almost every setting. A Kuznets curve describes a development trend in which progress is associated with first increased and then decreased inequality, creating an inverse-U-shaped curve[22].

We observe substantial variation in the maximum level of inequality reached during each period, which in some cases reflect threefold differences in the degree of equality for a given average level of schooling. In this way, these curves provide a valuable tool for comparing the level of inequality of each country compared with their neighbours, relative to their overall level of progress.

Latin America and the Caribbean had the highest levels of inequality in 1970, with an AID of 4.5 years (4.4–4.6) (Fig. 3a). Over time, however, Latin America and the Caribbean has had an only intermediate-height Kuznets curve, despite substantial progress in the mean years of educational attainment (Fig. 3b). Latin America and the Caribbbean stands out as having less inequality in education at each point in the development arc compared with regions such as South Asia or North Africa and the Middle East, as shown by a lower overall Kuznets curve. This result highlights the need to assess inequality for each region with respect to its level of development by looking across decades to understand variation in the arc of educational expansion.

Between 1970 and 2018, sub-Saharan Africa and South Asia saw great advances in education, and also large increases in inequality. South Asia had the highest level of educational inequality globally in 2018, with an AID value of 6.0 (5.7–6.3). Its Kuznets curve is largely similar to that of North Africa and the Middle East. If sub-Saharan Africa continues to develop at its current trajectory, we expect its trend to look similar to that of Latin America and the Caribbean, which is approximately 30 years further along the development arc. Taking a more granular look, substantial variation can be seen between nations in sub-Saharan Africa (Fig. 3c). In 2018, several countries in western sub-Saharan Africa displayed the highest inequality values in the world, well above the ninetieth percentile mark for their level of mean attainment. Nevertheless, several nations in southern Africa are below the tenth percentile of inequality values for their mean attainment.

The region of Southeast and East Asia and Oceania is noteworthy for having the flattest Kuznets curve, and therefore the least unequal trajectory of development among low- and middle-income countries. Eastern Europe and central Asia underwent rapid gains in education from 1970 to 1995, achieving mean values similar to high-income countries by 2018, with lower overall inequality.

Centring equality in global progress

Educational inequalities exist in many different forms and need to be addressed in order for societies to maximize well-being and the potential for education to facilitate economic development. Gender gaps are projected to persist for girls in much of the developing world and widen for boys in a subset of developed countries[16,23]. Disparities can also be found along dimensions of wealth, ethnicity, race, ability, and other social groupings[20,24]. Previous work has shown substantial inequalities in education between urban and rural areas[5,25,26], and along lines of wealth[20]. These inequalities are easy to miss when drawing on national average measures of attainment.

The distributions and inequality metrics presented here provide a framework that can be used to track the progress of each country towards the SDG targets and levels of inequality over time. Once detected, inequalities can be reduced with the implementation of specific policies. For example, eliminating school fees, improving local access to schools, increasing the number of years of compulsory schooling, and providing food, stipends, and other resources for children at school are known to increase participation among the most economically disadvantaged children, and the creation of special governmental bodies can reduce gaps for children of minority ethnic groups[25,27,28]. It is therefore essential to examine progress in average levels of attainment with an understanding

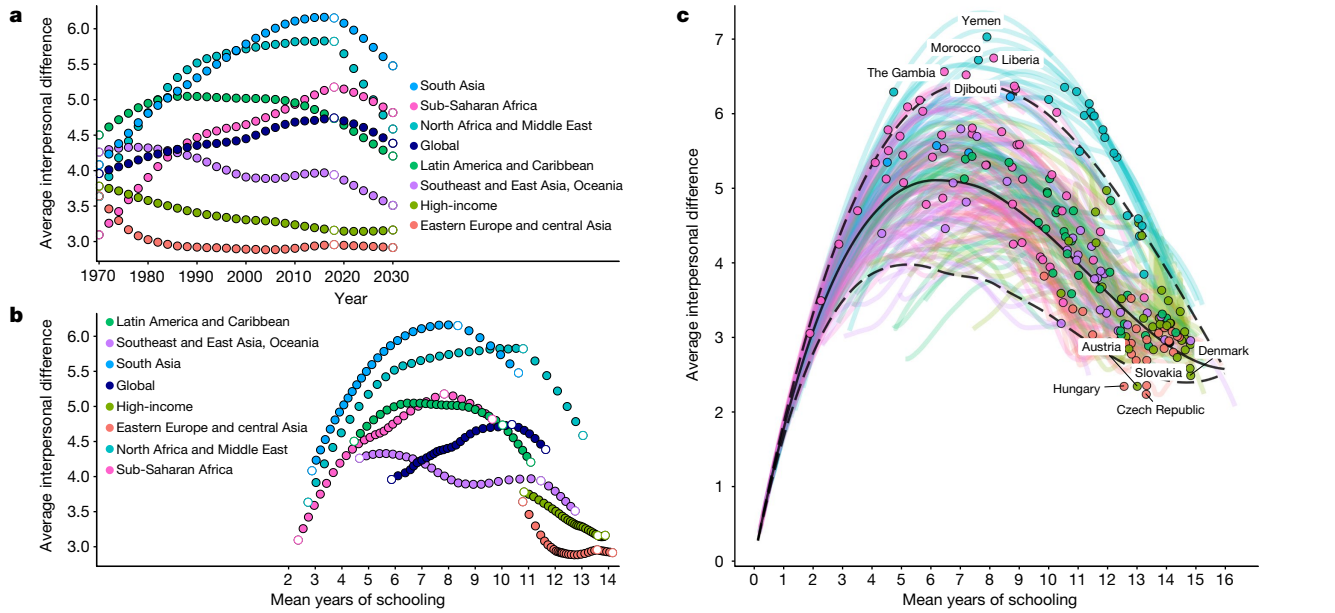


Fig. 3 | Trajectories in educational inequality. a, Trends in educational inequality are shown over time, with labels indicating the rank of inequality levels in 2030. **b**, Trends in educational inequality are shown with respect to mean years of schooling, with labels indicating the rank of inequality levels in 1970. Results are shown globally and regionally for every second year from 1970 to 2030, for individuals aged 25–29. The white dots mark 1970, the beginning of the estimates, 2018, the beginning of forecasts, and 2030, the SDG target year. **c**, National trends in the AID and mean years of schooling are shown from 1970 to 2030, with the value for 2018 shown as a bold point. The five highest and lowest values in 2018 are labelled. The solid line shows the median level of inequality for a given degree of mean years of schooling, across all years of data from 1970 to 2030, and the dashed lines show the smoothed ninety-fifth and fifth percentiles. Quantiles were calculated over modelled estimates from $n = 195$ countries.

of the full within-country distribution and inequality. Gains in education are linked to improvements in numerous other sectors of society[3,4,13]. Ensuring equality in education will translate into positive effects in the equality of human productivity, health, and well-being.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2198-8.

1. UNESCO. *Migration, Displacement and Education: Building Bridges, not Walls*. https://en.unesco.org/gem-report/report/2019/migration (2018).
2. Marmot, M., Friel, S., Bell, R., Houweling, T. A. & Taylor, S. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* **372**, 1661–1669 (2008).
3. Lim, S. S. et al. Measuring human capital: a systematic analysis of 195 countries and territories, 1990–2016. *Lancet* **392**, 1217–1234 (2018).
4. Gakidou, E., Cowling, K., Lozano, R. & Murray, C. J. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *Lancet* **376**, 959–974 (2010).
5. Graetz, N. et al. Mapping local variation in educational attainment across Africa. *Nature* **555**, 48–53 (2018).
6. UNESCO Institute for Statistics. *Quick Guide to Education Indicators for SDG 4*. http://uis.unesco.org/sites/default/files/documents/quick-guide-education-indicators-sdg4-2018-en.pdf (2018).
7. Nilsson, M., Griggs, D. & Visbeck, M. Policy: map the interactions between sustainable development goals. *Nature* **534**, 320–322 (2016).
8. Lutz, W., Cuaresma, J. C. & Sanderson, W. Economics. The demography of educational attainment and economic growth. *Science* **319**, 1047–1048 (2008).
9. Basu, A. M. & Stephenson, R. Low levels of maternal education and the proximate determinants of childhood mortality: a little learning is not a dangerous thing. *Soc. Sci. Med.* **60**, 2011–2023 (2005).
10. Bicego, G. T. & Boerma, J. T. Maternal education and child survival: a comparative study of survey data from 17 countries. *Soc. Sci. Med.* **36**, 1207–1227 (1993).
11. Hatt, L. E. & Waters, H. R. Determinants of child morbidity in Latin America: a pooled analysis of interactions between parental education and economic status. *Soc. Sci. Med.* **62**, 375–386 (2006).
12. Lutz, W. & Kc, S. Global human capital: integrating education and population. *Science* **333**, 587–592 (2011).
13. The European Commission. *Demographic and Human Capital Scenarios for the 21st Century* (eds Lutz, W. et al.) (2018).
14. United Nations. *The Millenium Development Goals Report 2015*. http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf (2015).
15. Goujon, A. et al. A harmonized dataset on global educational attainment between 1970 and 2060 – an analytical window into recent trends and future prospects in human capital development. *J. Demogr. Economics* **82**, 315–363 (2016).
16. OECD. *PISA 2015 Results (Volume I)* (2016).
17. OECD. *The Pursuit of Gender Equality* (2017).
18. Lopus, S. & Frye, M. Visualizing Africa's educational gender gap. *Socius* **4**, 237802311879595 (2018).
19. Grant, M. J. & Behrman, J. R. Gender gaps in educational attainment in less developed countries. *Popul. Dev. Rev.* **36**, 71–89 (2010).
20. Jones, G. W. & Ramchand, D. S. Closing the gender and socio-economic gaps in educational attainment: a need to refocus: closing the gender and socio-economic gaps in educational attainment. *J. Int. Dev.* **28**, 953–973 (2016).
21. United Nations. *#Envision2030 Goal 4: Quality Education*. https://www.un.org/development/desa/disabilities/envision2030-goal4.html (accessed November 2018)
22. Kuznets, S. Economic growth and income inequality. *Am. Econ. Rev.* **45**, 1–28 (1955).
23. UNESCO. *Achieving Gender Equality in Education: Don't Forget the Boys*. https://unesdoc.unesco.org/ark:/48223/pf0000262714 (2018).
24. Lewis, M. & Lockheed, M. *Inexcusable Absence: Why 60 Million Girls Still Aren't in School and What to do About It* (Center for Global Development, 2006).
25. UNICEF. *Magic Box - School Mapping*. http://school-mapping.azurewebsites.net/ (accessed November 2018).
26. Graetz, N., Woyczynski, L., Wilson, K. F., Hay, S. I. & Gakidou, E. Mapping persistent local disparity in educational attainment. *Nature* **555**, 48–53 (2019).
27. Urbina, D. R. Intergenerational educational mobility during expansion reform: evidence from Mexico. *Popul. Res. Policy Rev.* **37**, 367–417 (2018).
28. Cohen, A. K. & Syme, S. L. Education: a missed opportunity for public health intervention. *Am. J. Public Health* **103**, 997–1001 (2013).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Overview

Our study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER)[29]. We use a multi-stage model to estimate the average years of schooling, and the single-year distribution of educational attainment, for 1970 to 2018, and create projections to 2030. These models draw on a database of 3,180 nationally representative censuses and surveys. Estimates are created for the 195 counties and territories examined in the Global Burden of Disease 2017 study[30]. In the first stage, we model mean years of schooling and the proportion of the population without any formal schooling from 1970 to 2018. This is performed using a cohort extrapolation model and a subsequent age period model with Gaussian process regression to synthesis all data and create final estimates with uncertainty. The second stage entails an ensemble *K*-nearest neighbours algorithm to estimate the distribution of education from 1970 to 2018, drawing on previously estimated quantities. Finally, trends in these distributions are projected to 2030 using a rate of change approach, and mean years of schooling values for 2019–2030 are calculated from the resulting distributions. All analyses are run using 1,000 draws to propagate model and data uncertainty through to subsequent steps. All estimation steps are validated, and all hyper-parameters are optimized, using out of sample predictive validity.

Data sources

We compiled a database of 3,180 nationally representative surveys and censuses describing the distribution of years of schooling by age and sex. Data sources providing single years of schooling are used directly, while those providing larger bins of educational attainment, for example 'some primary attainment' are probabilistically split into single-year proportions using a previously published crosswalk model[31]. Data are top-coded to 18 years, as it is a common choice among providers of single-year education data[32], and it is reasonable to assume that the importance of education for health or social capital diminishes greatly after the completion of 18 years, which represents 2 to 3 years of post-university education in most educational systems.

Data adjustment model

Data are adjusted for systematic biases between data providers in a regional and location-specific fashion. Gold-standard data are identified using expert knowledge of the high-volume data providers that have robust processes in place to ensure data quality. In almost all cases, census data obtained from the IPUMS data repository are considered as the gold standard, or Demographic Health Survey data where IPUMS are not available. Supplementary Table 3 lists the location-specific gold-standard data providers. Regional effects are applied to all data to adjust them to the gold standard available in that region. Subsequently, in countries that had gold-standard data available, country-specific effects are used to adjust for within-county biases between data sources. This has the benefit of being able to correct for biases in all countries, even when gold-standard data are not available in that country, using regional effects. Country-specific effects ensure consistent time trends with minimal discontinuities.

We use a mixed-effects regression model with random effects for data provider and nested random effects for data provider within country. This model is run separately for each region, and is formulated as follows:

$$\text{logit}(P_{Q,A,S,Y,L}) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{location} + \beta_4 \times \text{year} + u_{\text{dataprovider}} + u_{\text{location:dataprovider}}$$

in which $P_{Q,A,S,Y,L}$ is the quantity of interest, either proportion of the population with no education or $\frac{\text{mean years of educational attainment}}{18}$ for a given age, sex, year, and location. $u_{\text{data provider}}$ is a region-specific random effect that captures the average bias between data providers across all countries within that region, and $u_{\text{location:data provider}}$ is a nested location-specific random effect that captures the additional bias between a location-specific gold standard (where applicable) and the other sources present in that location.

To calculate source adjustments for each data provider, the $u_{\text{data provider}}$ value for each data provider is compared with the regional gold standard, and the difference is applied to all surveys. Subsequently, in locations that have gold-standard data present, $u_{\text{location:data provider}}$ effects are applied in the same fashion.

Cohort extrapolation

We use an age-cohort modelling process to project cohorts through time, leveraging the stability of cohort-specific educational attainment after age 25. To model the changes by age within cohorts, we use data from all available cohorts with multiple observations at or after age 25. For each quantity being modelled, we calculated y_{Q,L,S,C,A_x}, which is the logit difference of the $P_{Q,A,S,Y,L}{}'$ (the adjusted input data) at time x and at time y, for all possible combinations of repeat cohort observations. We restrict repeat cohort observations to those that are less than or equal to 10 years apart and to those where both observations occur after 1990 to avoid the attribution of differences in measurements to mortality as opposed to advances in survey and census design. In addition, we normalize all repeat cohort observation pairings so that the average change at 65 years of age is 0 to account for systematic bias between survey iterations (such as improvements in sampling). This is similar to other previously described approaches[33], in which only excess mortality beyond the age of 65 is considered. This calculation is shown below:

$$y_{Q,L,S,C,A_x} = \text{logit}(P_{L,S,C,A,\text{Src}_x}{}') - \text{logit}(P_{L,S,C,A,\text{Src}_y}{}') - \text{bias}_{L,P,S,C,\text{Src}}$$

in which Q is the quantity being modelled, L is location, S is sex, C is cohort, A is age, Src is data provider, and $\text{bias}_{L,P,S,C,\text{Src}}$ is the average change for cohorts as they age from 60 to 70 between the two surveys. This is the age period for which we expect the educational attainment of a cohort to be least prone to changes due to migration and mortality, and any changes observed during this period are therefore used as a measure of inherent bias between multiple waves of a survey or census.

These logit differences were examined with respect to several predictor variables. We then modelled the logit difference using a number of linear mixed-effects models, which were evaluated using out-of-sample predictive validity (see Supplementary Information). The best performing model specification is displayed here:

$$y_{Q,L,S,C,A_x} = I + u_{\text{location :super region}}$$

in which I is a natural spline with a knot at age 70 intended to capture the potential nonlinearity in the rate of change of differential mortality by education over age. $u_{\text{location:super region}}$ are random intercepts on location, nested within super-regional random intercepts.

Age-period model

Age-period models were fit on all values of $P_{Q,A,S,Y,L}{}''$, which reflect the adjusted input data after cohort extrapolation, to interpolate data for observed cohorts, and to extrapolate to all parts of the desired time series, producing $P_{Q,S,Y,L}{}'''$, single-year estimates of attainment from 1970 to 2018. Several linear mixed-effects models were used and evaluated using out-of-sample predictive validity (see Supplementary Information). Separately for each sex, and region grouping used in the GBD study, the quantity of interest of the country–age–year-specific population, $P_{Q,A,S,Y,L}{}'''$ was estimated:

$$\text{logit}(P_{Q,A,S,Y,L}'') = \beta_{s,r} + \delta_{s,r}\text{year} + I_{s,r} + \alpha_{c,a,s}$$

in which $\beta_{s,r}$ is a sex- and region-specific intercept; $\delta_{s,r}$ captures the linear secular trend for each sex and region; $I_{s,r}$ is a natural spline on age to capture the nonlinear age pattern by sex and region, with knots at 45 and 65 years of age; and $\alpha_{c,a,s}$ is a country-sex-specific random intercept.

Gaussian process regression

Gaussian process regression (GPR) was used to ensure final model results are consistent with input data and incorporate model and data uncertainty to produce uncertainty intervals. GPR has been used extensively as a data synthesis tool[34]. GPR uses a covariance function to smooth the residuals from the age-period model, taking into account the uncertainty in each data point. GPR also synthesizes both data and model uncertainty, in order to produce estimate uncertainty intervals. GPR assumes that the trend in the underlying data follows a Gaussian process, which is defined using a mean function $m(\cdot)$ and a covariance function $\text{Cov}(\cdot)$. Therefore, separately for each Q quantity being estimated, the location–sex–age–year-specific outcome measures are defined:

$$\text{logit}(y_{Q,L,S,C,A}) = g_{Q,L,S,A,Y} + \epsilon_{Q,L,S,A,Y}$$

Where the error term is normally distributed:

$$\epsilon_{Q,L,S,A,Y} = \text{normal}(0, \sigma_p^2)$$

The error variance, σ_p^2 is composed of the squared standard error of the observed data point, as well as the prediction errors from the age-cohort imputation process. The mean function of the model is defined as the age-period model predictions, as detailed above. The covariance function of the model is derived using a Matérn covariance function, consistent with prior applications of GPR:

$$M(y,y') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{d(y,y')\sqrt{2\nu}}{l} \right)^{\nu} K_{\nu} \left(\frac{d(y,y')\sqrt{2\nu}}{l} \right)$$

where $d(\cdot)$ is a distance function, σ^2 is the marginal variance, ν is a smoothness hyper parameter defining the differentiability of the function, l is a link-scale parameter approximately equivalent to the number of years at which two points are no longer correlated, K_{ν} is the Bessel function, and $\Gamma(\cdot)$ is the gamma function. Similar to previous applications of GPR, we approximate σ_p^2 as the super-region and sex-specific residual from the mean function, with ν set to 2 and l to 40, to reflect the inherent smoothness of educational attainment trends over time.

Ensemble *K*-nearest neighbours distribution model

To create a full time-series of distributions of single-years of educational attainment to 2018, we used a K-nearest neighbours algorithm to reconstruct an ensemble distribution for each location–age–sex–year (LASY) combination. To pick K candidate distributions for each LASY combination, we used two modelled entities produced by the above methods, mean educational attainment and proportion of the LASY population with 0 years of schooling, to find the most similar distributions in our database of 3,180 surveys and censuses. The metric used to find the most similar distributions was the Mahalanobis distance:

$$D_M^i(H^i) = \sqrt{(H^i - I^{\text{LASY}})^T S^{-1} (H^i - I^{\text{LASY}})}$$

in which H^i is a multivariate vector $\left(\text{logit}\left(\frac{\text{mean}^i}{18} \right), \text{logit}(\text{prop}_0^i) \right)$ corresponding to a survey–age–sex–year entry in our educational database, I^{LASY} is a multivariate vector $\left(\text{logit}\left(\frac{\text{mean}^{\text{LASY}}}{18} \right), \text{logit}(\text{prop}_0^{\text{LASY}}) \right)$ representing the modelled entities described above, and S^{-1} is the covariance matrix between vectors $\text{logit}\left(\frac{\text{mean}^i}{18} \right)$ and $\text{logit}(\text{prop}_0^i)$.

For each I^{LASY}, K distributions with the smallest Mahalanobis distances are chosen as candidate distributions for the final ensemble distribution. To collapse K distributions to a final ensemble distribution, we use a weighted average of the candidate distributions based on a location, age, and cohort distance defined as:

$$\text{Distance}^i = (P_{\text{age}} \times \text{Distance}_{\text{age}}^i)^{\psi} + (P_{\text{cohort}} \times \text{Distance}_{\text{cohort}}^i)^{\psi} + (P_{\text{space}} \times \text{Distance}_{\text{location}}^i)^{\psi}$$

All values of P and Distance are rescaled to lie between 0.001 and 1. $\text{Distance}_{\text{location}}^i$ is 0.001 for same country, 0.33 for same region, 0.66 for same super-region, and 1 otherwise.

$$\text{Weights}^i = \frac{1}{\text{Distance}^i}$$

ψ is a hyperparameter controlling how sharply weights decrease as Distance^i increases. To collapse K distributions to a final ensemble distribution for each LASY combination we calculated:

$$\text{Proportion}_{\text{eduyrs}}^{\text{LASY}} = \frac{\sum_{i=1}^{K} \text{Weights}^i \times \text{proportion}_{\text{eduyrs}}^i}{\sum_{i=1}^{K} \text{Weights}^i}$$

in which $\text{Proportion}_{\text{eduyrs}}$ is the proportion in each educational bin, 0–18.

Final ensemble distributions were then smoothed by bin using a Loess smoother with a span of η over time to ensure plausible time series for each draw. All hyperparameters were optimized using out-of-sample predictive validity (detailed in the Supplementary Information), and chosen values include: $K = 80$; $P_{\text{age}} = 0.25$; $P_{\text{cohort}} = 0.85$; $P_{\text{space}} = 0.7$; $\psi = 2.5$; $\eta = 0.5$.

Rate of change distribution forecasting model

To forecast the distribution of education and mean years of schooling, we use a rate of change (ROC) model at the single-year bin level. This has the benefit of producing projections of mean attainment that respect the nonlinear dynamics of distributional growth. The model is fit in a timeseries-specific fashion, separately by sex and country. For each single-year bin, we derive a ROC using a weighted average of the ROC for the last 15 years:

$$\text{ROC}_{\text{eduyr}}^{\text{LAS}} = \sum_{i=2004}^{2018} \frac{\text{logit}(\text{proportion}_{\text{eduyr}})^{\text{LAS}^i} - \text{logit}(\text{proportion}_{\text{eduyr}})^{\text{LAS}^{i-1}}}{15}$$

Where $\text{ROC}_{\text{eduyr}}^{\text{LAS}}$ is the average rate of change over the last 15 years within each location–age–sex (LAS) combination for each single-year bin of education (0–18).

The ROC model was leveraged only where the cohort extrapolation model could not inform our estimates. This begins in 2019 for 25–29-year-olds, 2024 for 30–34-year-olds, and 2029 for 35–40-year-olds. For the results presented in the main text, for 25–29-year-olds, this method was used for 2019 onwards.

SDG progress and inequality metrics

Drawing on these estimates of the distribution of years of schooling, we calculate several metrics detailing global progress towards the SDG 4 targets. We calculate the proportion of the population of individuals age 25–29 who have completed primary, secondary, and tertiary education, defined as completing at least 6, 12, and 15 years of schooling, respectively. We describe gender equality using the ratio of female to

male attainment of primary and secondary education, as well as the gap in mean years of schooling between men and women. Aggregate measures at the national level for both sexes, and at the regional level were calculated, using projected population estimates drawn from the World Population Prospects dataset[35]. We also present a novel index of educational inequality among young people in each country, the average AID. This index is defined as the average value of the absolute differences between all possible pairs of individuals in the population. The AID is also mathematically equivalent to the Gini coefficient, multiplied by two times the mean of the distribution[36].

Predictive validity

The main aims of this analysis are predictive in nature, and we therefore assessed each stage of our model, and each model selection decision, with respect to predictive capacity. We focused mainly on 'out-of-sample' predictive ability, which reflects how well the model predicts data that was not directly available. This most mimics the true task that we want our model to accomplish, that is, to make accurate predictions for the geographies and time periods that do not have input data available. To assess out-of-sample predictive validity, we followed the general strategy of dividing our database into 'training' and 'testing' data. The model was fit on the training data, and the results were compared with the testing data. The 'error' of the model represents the average amount that our model was incorrect compared with the 'true' data that was held out. Each step of the modelling process was assessed for how well it predicted (out-of-sample) the mean years of schooling for a given population, as well as other aspects of the distribution, such as the proportion with 0 years of schooling. We also assessed the degree to which predictive validity varied by time period, across regions, and by which type of data source was held out. There were small differences in predictive validity across these dimensions, for example, models tended to perform slightly better in the 2000–2018 period where the most data are available; however, they were generally modest. Furthermore, we found that the best performing models tended to perform optimally across almost all geographies/time periods, so it was not necessary to use multiple models for a single step. All predictive validity results, and a discussion of their implications, can be found in the Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

This study used data that are available from public online repositories, most of which require a straightforward registration process and usage agreement with the data provider. A detailed table of data sources and availability can be found in the Supplementary Information. Although the authors are restricted from providing the data directly in most cases, specific datasets may be made available by request and with permission from the data provider. The authors may be contacted for assistance in acquiring data for the replication of this study. All maps presented in this study have been produced by the authors and no permissions are required for publication. Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset[37].

Code availability

All code used for these analyses is available here: https://github.com/Joseph-Friedman/education_inequality.

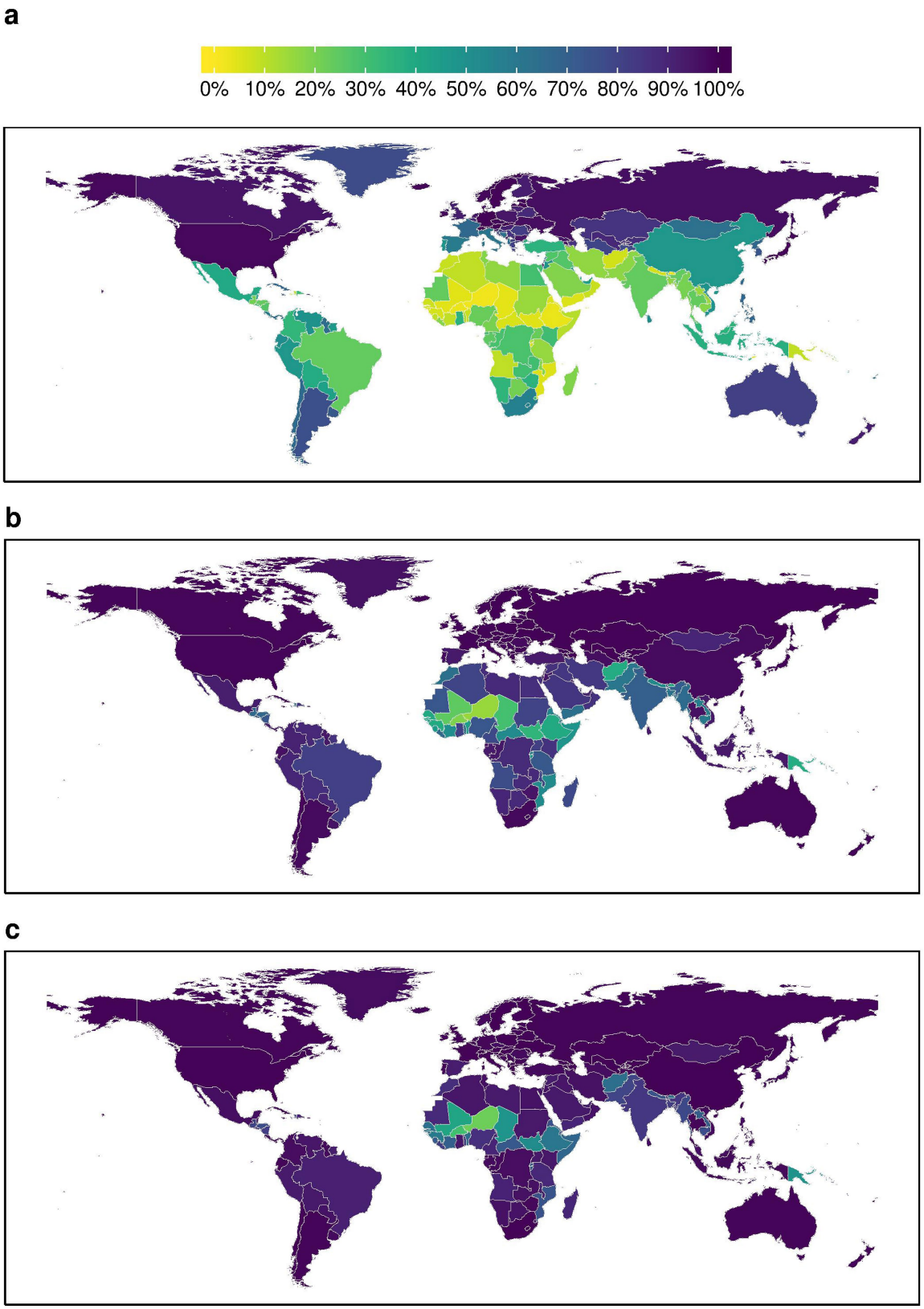
29. Stevens, G. A. et al. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. *PLoS Med*. **13**, e1002056 (2016).
30. Foreman, K. J. et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. *Lancet* **392**, 2052–2090 (2018).
31. Friedman, J., Graetz, N. & Gakidou, E. Improving the estimation of educational attainment: new methods for assessing average years of schooling from binned data. *PLoS One* **13**, e0208019 (2018).
32. IPUMS International. *YRSCHOOL*. https://international.ipums.org/international-action/variables/YRSCHOOL#comparability_section (accessed November 2018).
33. Barro, R. J. & Lee, J. W. A new data set of educational attainment in the world, 1950–2010. *J. Dev. Econ*. **104**, 184–198 (2013).
34. GBD 2016 Mortality Collaborators. Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1084–1150 (2017).
35. United Nations. *World Population Prospects*. https://population.un.org/wpp/ (accessed November 2018).
36. Gakidou, E. *Health Inequality: Definition, Measurement, and Determinants* (Harvard Univ., 2001).
37. GeoNetwork. *Global Administrative Unit Layers (GAUL)*. http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691 (2007).

Acknowledgements This work was primarily supported by grant OPP1152504 from the Bill & Melinda Gates Foundation. J.F. received support from the UCLA Medical Scientist Training program (NIH NIGMS training grant GM008042). We thank S. B. Munro for assisting with the preparation of the manuscript.

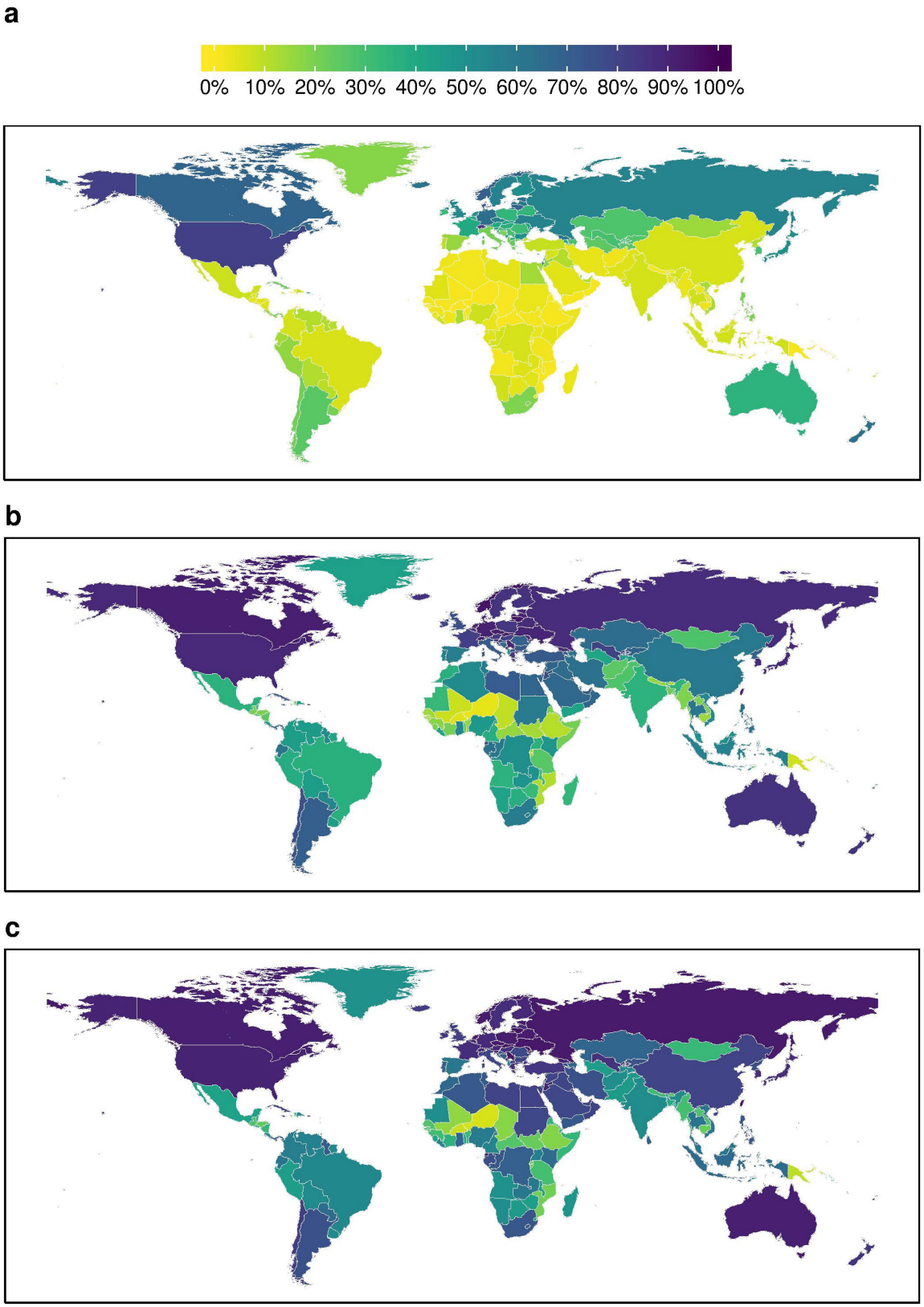
Author contributions J.F., H.Y., N.G. and E.G. conceived and planned the study. J.F., H.Y. and J.W. obtained, extracted and processed educational attainment data. J.F. and H.Y. wrote the computer code and designed and carried out the statistical analyses, with substantial intellectual and methodological inputs from E.G., N.G., L.W. and S.I.H. J.F., H.Y. and E.G. wrote the first draft of the manuscript and all authors contributed to subsequent revisions.

Competing interests The authors declare no competing interests.

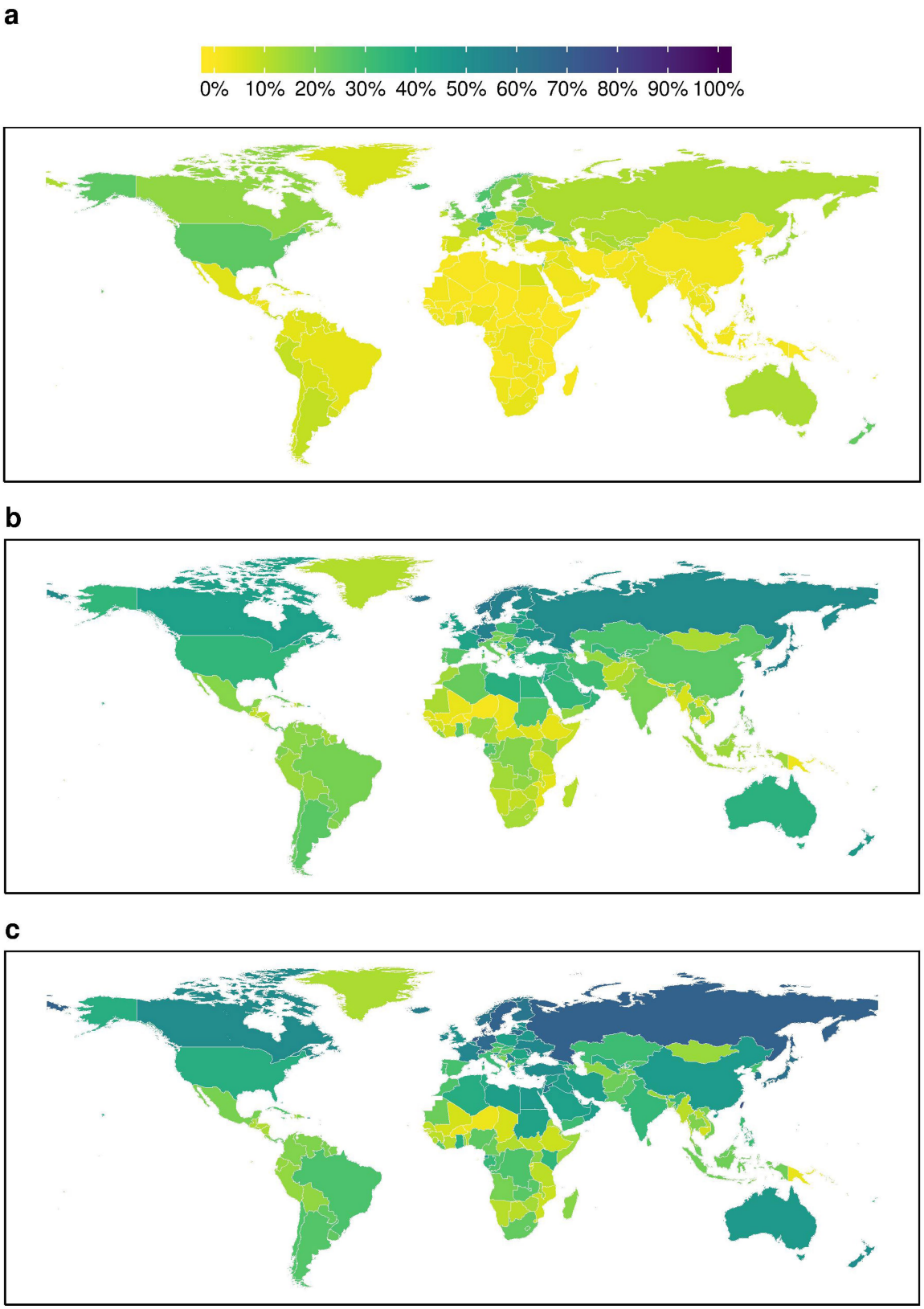
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2198-8.
Correspondence and requests for materials should be addressed to E.G.
Peer review information *Nature* thanks Noam Angrist and Monica Grant for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



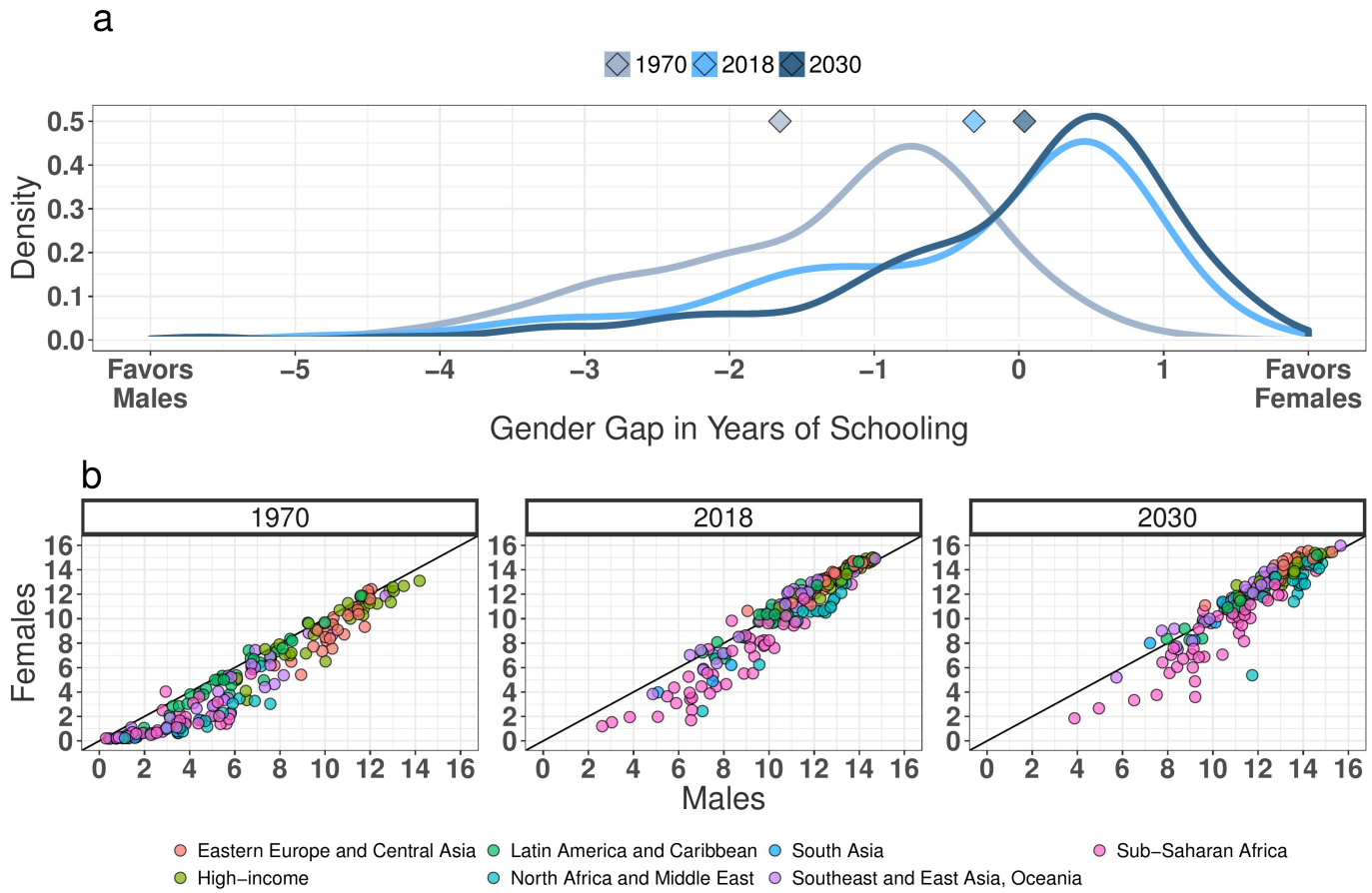
Extended Data Fig. 1 | Completion of 6 or more years of schooling. a–c, The percentage of the population aged 25–29 completing at least 6 years of schooling is shown by country, for 1970 (**a**), 2018 (**b**), and 2030 (**c**). Maps were produced using R v.3.5.0.



Extended Data Fig. 2 | Completion of 12 or more years of schooling. a–**c**, The percentage of the population aged 25–29 completing at least 12 years of schooling is shown by country, for 1970 (**a**), 2018 (**b**), and 2030 (**c**). Maps were produced using R v.3.5.0.



Extended Data Fig. 3 | Completion of 15 or more years of schooling. a–**c**, The percentage of the population aged 25–29 completing at least 15 years of schooling is shown by country, for 1970 (**a**), 2018 (**b**), and 2030 (**c**). Maps were produced using R v.3.5.0.



Extended Data Fig. 4 | Years of schooling among men and women. a, The distribution of the gap in mean years of schooling between men and women, aged 25–29, is shown for 1970, 2018, and 2030, with the population-weighted mean for each time point represented with a diamond. Means were calculated over modelled estimates from $n = 195$ countries. **b**, Years of schooling is represented for men on the *x* axis and women on the *y* axis for 1970, 2018, and 2030, in which each point indicates the value for one country, colour-coded by regional grouping. A point above the line indicates additional schooling for women relative to their male counterparts.

nature research

Corresponding author(s): Emmanuela Gakidou

Last updated by author(s): Feb 4, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No primary data collection was carried out for this analysis.
Data analysis	All analyses were conducted using R version 3.1.3 and Python 2.7.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

This study used data that are available from public online repositories, but which in most cases require a straightforward registration process and usage agreement with the data provider. A detailed table of data sources and availability can be found in the supplement. Although the authors are restricted from providing the data directly in most cases, specific data sets may be made available by request and with permission from the data provider. The authors may be contacted for assistance in acquiring data for the replication of this study.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A descriptive, population-level, ecological study of inequality
Research sample	All available nationally representative census and survey data containing information about educational attainment.
Sampling strategy	We used all available nationally representative census and survey data containing information about educational attainment. .
Data collection	N/A - secondary analysis
Timing	N/A - Secondary analysis, data originally collected between 1950 and 2018 were included.
Data exclusions	As described in the methods section, with greater detail in the supplement, this study provides modeled estimates of the single-year distribution of years of schooling over time and by country. Data were included from 195 nations and territories that are part of the Global Burden of Disease 2017 study. Data for other areas that do not pertain to this list, or which were found to not be nationally representative, were not included. Data that did not include 5-year age groups, or which were not disaggregated by age or sex were also not included. Data from outside the 1950-2018 time period were also not included.
Non-participation	N/A - secondary analysis of nationally-representative statistics
Randomization	N/A - observational analysis, no experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging