Mapping the Link between Life Expectancy and Educational Opportunity | *Methods*

About the data

Data for this project came from four sources: the National Historical Geographic Information System (NHGIS, 2011-2015), the U.S. Small-area Life Expectancy Estimates Project (USALEEP, 2010-2015), the Civil Rights Data Collection (CRDC; 2011-2012, 2013-2014, 2015-2016), and the Common Core of Data (CCD, 2015-2016).

Life expectancy estimates came from the USALEEP (National Center for Health Statistics, 2018). USALEEP was created through a partnership between the National Center for Health Statistics, of the Centers for Disease Control and Prevention; the Robert Wood Johnson Foundation; and the National Association for Public Health Statistics and Information Systems. This first-of-its-kind dataset contains life expectancy estimates at the census-tract level for 2010 through 2015—a much lower level of geographic aggregation than was previously available. Data regarding census tracts and their demographics from 2011 to 2015 were obtained from the NHGIS. The NHGIS is an open database of over 200 years of U.S. Census information (National Historical GIS, n.d.).

School-level data on educational opportunity came from the CRDC. The United States Department of Education has administered the CRDC since 1968 to collect student and program information at the school level, most of which is disaggregated by race/ethnicity, sex, limited English proficiency, and disability status. Data are collected on a variety of educational, teacher, and student topics; since 2011, these data have been collected from a census of all public schools in the country (U.S. Department of Education, 2018). To improve the relevance of these estimates to the predicted life expectancy (which was estimated for the period 2010-2015), CRDC data from the 2011-2012, 2013-2014 and 2015-2016 school years were combined when possible. In addition to these main data sources, the CCD for the 2015-2016 school year were used to identify each school's geographic location.

Measuring educational opportunity

A recent report by the National Academies of Sciences, Engineering, and Medicine (2019) describes inequality of educational opportunity in four main areas: exposure to racial, ethnic, and economic segregation; access to high-quality early learning programs; access to high-quality curricula and instruction; and access to supportive school and classroom environments. Adapting this framework, we focused on two of these areas—access to high-quality curricula and instruction and access to supportive school and classroom environments. We did not focus on access to high-quality early learning programs since our project examines educational opportunity at the high school level. Similarly, we did not focus on exposure to racial, ethnic, and economic segregation because we wanted to study facets of educational opportunity that are more easily malleable through educational policy interventions. The National Academies of Sciences, Engineering, and Medicine identifies access to effective teaching, access to and enrollment in rigorous course work, disparities in curricular breadth, and disparities in access to academic supports as the four key components of high-quality curricula and instruction. In the area of supportive learning environments, they identify disparities in school climate, discipline practices, and nonacademic supports as key indicators.

In addition, our final model incorporates the recommendations of the education experts we consulted: This work would not have been possible without the expert guidance of Child Trends' experts, including Deb Temkin, David Murphey, and Natalia Pane; as well as Dan Losen of the Center for Civil Rights Remedies, Chad Aldeman of Bellwether Education Partners, and Elizabeth Ross of the National Council on Teacher Quality. We also reviewed the literature on educational effectiveness and considered the availability of data. Drawing on all of this research, we chose four main domains to measure access to educational opportunity: 1) access to rigorous academics, 2) access to supportive conditions for learning (including discipline), 3) access to appropriate nonacademic supports, and 4) access to effective teaching.

Scores for each domain were calculated from a combination of relevant variables, as described in Table 1 below. When data for an indicator were available for multiple years, we calculated the average value across years.

Table 1. Domains, indicators, and variables included in the educational opportunity measure

Domain (alpha value)	Indicator	CRDC Variables Used	Years of Data	Indicator Type
Access to rigorous academics (alpha = .67)	Were Advanced Placement (AP) courses offered?	AP enrollment indicator and AP enrollment count	2011-12, 2013-14 and 2015-16	Yes/No
	Was dual enrollment was offered?	Dual enrollment Indicator	2013-14 and 2015-16	Yes/No
	Number of advanced mathematics courses offered per 100 students	Number of advanced mathematics courses offered, total enrollment	2011-12, 2013-14 and 2015-16	Continuous
Supportive conditions for learning (alpha = .54)	Proportion of students who were chronically absent (reversed)	Number of students chronically absent, total enrollment	2013-14 and 2015-16	Continuous
	Number of days lost to out-of-school suspension per 100 students (reversed)	Days missed due to out-of-school suspension, total enrollment	2015-16	Continuous
	Proportion of students experiencing out-of-school suspension	Number of students experiencing out-of-school suspension, total enrollment	2011-12, 2013-14 and 2015-16	Continuous
Non-academic supports (alpha = .58)	Number of school counselors per student	Number of full-time equivalent school counselors, total enrollment	2011-12, 2013-14 and 2015-16	Continuous
	Number of school nurses per student	Number of full-time equivalent nurses, total enrollment	2015-16	Continuous
	Number of psychologists per student	Number of full-time equivalent psychologists, total enrollment	2015-16	Continuous
	Number of social workers	Number of full- time equivalent social	2015-16	Continuous

Domain (alpha value)	Indicator	CRDC Variables Used	Years of Data	Indicator Type
		workers, total enrollment		
Effective teaching (single item scale, no alpha)	Proportion of teachers who are inexperienced	Number of teachers in their first year, number of teachers in their second year, total number of teachers	2015-16	Continuous

Some additional indicators, shown in Table 2 below, were considered and ultimately not included.

Table 2. Domains and indicators excluded from the educational opportunity measure

Domain	Indicator	Reason for Exclusion
Access to rigorous academics	Were International Baccalaureate (IB) courses offered?	IB classes were much less common, and almost all schools offering IB classes also offered AP classes.
Supportive conditions for learning	Ratio of school law enforcement officers to school support staff	In the 2015-16 survey administration (the only year in which schools were required to report on law enforcement officers), there was an issue with question administration, and 14,394 high schools (out of the ~ 24,000 included in the study) were not asked the item about law enforcement officers (Office for Civil Rights, 2018).
	Proportion of students who were expelled	Expulsion is mandatory for many serious offenses, so this measure may reflect some expulsions that are not at the discretion of the school, and thus be less reflective of school climate.
	Proportion of students experiencing in-school suspension	The documented link between in-school suspension (ISS) and negative outcomes is not as strong as the link between out-of-school suspension and negative outcomes. A wide range of activities take place during ISS, which may include therapeutic activities that may not be harmful to students. Expert advice suggests that that ISS is not inherently harmful but may lead to out-of-school suspension, which might explain the weak relationship between ISS and negative outcomes (D. Losen, personal communication, October 9, 2019).
Effective teaching	Teacher turnover rate	This data was not available in the CRDC data.

Next, each indicator was scaled on a 0-to-10 scale (with 0 being the lowest value and 10 the highest). For dichotomous (e.g., Yes/No) indicators, all schools with a 'No' response were coded as 0, and all schools with a 'Yes' response were coded as 10. All continuous variables were translated to a 0-to-10 score with linear rescaling. If necessary, negative indicators were reversed so that all indicators had the same positive directionality. Reversing negative indicators and rescaling variables allowed us to average indicators while accounting for different measurement scales. In addition, because the CRDC is prone to outliers (Office for Civil Rights, 2013), prior to rescaling, the continuous values were truncated at 5 percent and 95 percent. This means that (before linear rescaling) all values lower than the 5th percentile of that indicator were rounded up to the 5th percentile, and all values higher than the 95th percentile were rounded down to the 95th percentile.

To aggregate indicator scores within domains and calculate domain-level scores, we calculated factor scores for each domain. This is equivalent to performing factor analysis to calculate the relative importance of each indicator within the domain, and then taking the average of all indicators across a domain weighted by their importance. For ease of presentation and understanding, the final domain score for each domain was categorized into five quintiles, with a score of 'one' representing the lowest-scoring one-fifth of schools and a score of 'five' representing the highest scoring one-fifth. The overall educational opportunity measure is the average of the four domain scores.

Connecting educational opportunity and life expectancy

Because life expectancy is measured at the census-tract level and educational opportunity is measured at the school level, we needed to transform the data. To account for life expectancy of adolescents in all neighborhoods across the country, we considered census tracts to be the primary unit of analysis.

To match census-tract data to school data, we first used data from the CCD to identify the geographic location of each school by latitude and longitude. We then implemented a two-step matching process:

1) Any census tract that had a school (or schools) in it was matched to the school(s).
2) Any census tract without a school in it was matched to the closest school or schools that lies in the same school district.[1]

Finally, for tracts that matched with more than one school (which occurred in 4,500 census tracts—either because there was more than one school within a census tract, or because two schools share the same location and are thus equally close to a census tract border), we calculated the average educational opportunity for all matched schools, weighted by the average enrollment at the school. The analysis file therefore contained each census tract in the USALEEP life expectancy data matched with an educational opportunity score.

Regression analysis

The central research question of this study was: "What is the association between educational opportunity at public high schools and youth life expectancy?" To answer this question, we used linear regression analysis, treating the census tract as the unit of analysis. The simplest model can be expressed as:

$$Life\ expectancy = Intercept + Educational\ opportunity * \beta + \epsilon$$

where β is the coefficient for educational opportunity, and ϵ is the error term. This model was estimated using ordinary least square regression. To account for the fact that the dependent variable (life expectancy) was measured with uncertainty, we estimated the standard errors of the model robustly using a sandwich estimator.

In addition, demographic differences between census tracts may be strongly related to both life expectancy and the quality of educational opportunities offered. Thus, our final model accounts for both poverty and race/ethnicity. The final model on which our visualizations and narrative are based is:

$$\begin{aligned} Life\ expectancy \\ &= Intercept + Educational\ opportunity * \beta + \text{Percent living in poverty} \\ &+ \text{Percent Hispanic} + \text{Percent Black} + \text{Percent American Indian} + \text{Percent Pacific Islander} \\ &+ \text{Percent Asian} + \text{Percent two or more races} + \text{Percent other race} + \epsilon \end{aligned}$$

[1] Information on the school district(s) associated with each census tract comes from the National Center for Education Statistics' geographic crosswalks (https://nces.ed.gov/programs/edge/Geographic/RelationshipFiles). We used the 2015 file that corresponds to the 2013-14 school year and matches to 2015 census data. Here, closest is defined as the school with the minimal straight line (i.e., Euclidean) distance to the census tract boundary. More than one school may be identified as matching due to schools' sharing a geographic location.

In this model, non-Hispanic and White were treated as baseline levels, which is captured within the intercept term, and thus are not included in the model formula.

To validate our analytic decisions, we conducted a series of sensitivity analyses. We made three key decisions in choosing the data and model: 1) to include magnet and charter schools, 2) to exclude virtual and online schools, and 3) to exclude interactions between demographics and educational opportunity from the model. To assess the impact of these decisions on our conclusions, we ran the analysis with and without each decision and compared the results. We found that the model variations created using these three decisions made little difference in the percent of variance explained by the model (r-squared), or in the coefficients of all variables. Thus, we chose to include schools that were connected by local geography because of our research question. Therefore, the final data set includes all physical schools, including charter and magnet schools, but excludes virtual and online schools. We also chose to exclude interaction effects because they did not contribute substantially to the model and added complexity.

Portraying uncertainty

Traditionally, the uncertainty in the results of regression analyses can be portrayed through confidence or prediction intervals, which allow a visual representation of how certain we can be about the regression line, or the predictions generated from that regression line, respectively. The prediction interval for predicted values $\hat{y}$ can be expressed as:

$$\hat{y} \pm T_{\alpha,N-p-1} * \sqrt{Var[\hat{y}]}$$

However, this calculation does not fully account for the known variability in the measurement of the true y values (i.e., the known variability in life expectancy). To account for this variability in our confidence intervals, we first calculated the average variance in y, and then added that to the variance of $\hat{y}$.

$$\mu_{Var(y)} = \frac{1}{N}\sum_{i=1}^{N}(SE_{y_i})^2$$

Thus:

$$\hat{y} \pm T_{\alpha,N-p-1} * \sqrt{Var[\hat{y}] + \mu_{Var(y)}}$$

We acknowledge that this method is unconventional, and that little work has been done on accounting for known uncertainty in y variables. This estimate for the width of the confidence interval is conservative because there is some overlap between the variance of y and $\hat{y}$ (due to y being used to estimate $\hat{y}$). This means the confidence interval presented here may be wider than the true confidence interval.

Data visualization and beta testing

The results from the regression analysis were used to create an interactive data visualization website in order to communicate our results to a wide audience. The methodology for each section of the website is described below.

Average neighborhood. In the explanation of the educational opportunity measure, we provide examples of each indicator in the "average neighborhood." Each indicator in this average neighborhood is calculated as the mean value for the neighborhoods in the middle quintile of overall educational opportunity.

Scatterplots. The model and prediction interval described above are also key components of the data visualization. The scatter plots that start the visualization display life expectancy (on the y-axis) by

educational opportunity (on the x-axis). Each blue dot represents the density of census tracts at that point on the graph (because representing each of the 65,000 census tracts as an individual dot was not practical). The regression line shows the slope of educational opportunity from the full model, which controls for poverty, race, and ethnicity. This means that the slope can be interpreted as the following: A one-point increase in educational opportunity corresponds to a 0.21 year increase in life expectancy for the census tract after controlling for race, poverty, and ethnicity. The shaded region represents the modified confidence interval (as described in the section above), which accounts for both uncertainty in model parameters and in the measurement of life expectancy, and is calculated assuming that new census tracts have average demographics and varying educational opportunities.

Interactive maps. In the interactive map, school districts are mapped using unified school districts from 2015.[2]

Open source. To promote open access and research transparency, we used open source software (R for data management and analysis and Javascript for the visualization) and made all of our code available for public use. Complete source code and data for this project are available at https://github.com/child-trends/educational opportunity. This code is licensed under a GNU GPL license and we encourage others to use and build on this work.

Beta testing. In December 2018, beta testing was conducted over GoToMeeting with seven state health and education policy officials from six states. The beta testing participants took part in a combination of exploration and structured prompts while sharing the website on their screen for 30 minutes. The feedback from the beta testers was compiled and implemented into the website. We wish to express our deep appreciation to the state health and education policy officials who served as beta testers for the data visualization website. Their feedback was invaluable in making this resource as useful and user-friendly as possible.

[2] Information on the school district boundaries is available at https://gis-server.data.census.gov/arcgis/rest/services/Hosted/VT_2015_970_00_PY_D1/VectorTileServer/.

References

National Academies of Sciences, Engineering, and Medicine. (2019). Monitoring Educational Equity. The National Academies Press. https://doi.org/10.17226/25389

National Center for Health Statistics. (2018). U.S. Small-area Life Expectancy Estimates Project – USALEEP. Retrieved December 9, 2019 from https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html

National Historical GIS. (n.d.). IPUMS NHGIS. Retrieved October 31, 2019 from www.nhgis.org

Office for Civil Rights. (2013). Civil Rights Data Collection: State and National Estimation Data Notes. Civil Rights Data Collection. Retrieved December 9, 2019 from https://ocrdata.ed.gov/downloads/EstimationDataNotes.docx

Office for Civil Rights. (2018). 2015-2016 Data Notes. Retrieved December 9, 2019 from https://ocrdata.ed.gov/Downloads/Data-Notes-2015-16-CRDC.pdf

United States Department of Education. Office for Civil Rights. (2018). Civil Rights Data Collection (CRDC) for the 2015-16 School Year. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E103004V1