Administrative School Quality Documentation

Last Updated by Audrey R Murchland 09/16/20

Table of Contents

A1. Administrative Data Collection and Documentation Information

State level data

Data for selected years from 1905 to 1973 were pulled from multiple sources including the Federal *Digest of Educational Statistics* (sequence starts in 1962), *Biennial Survey of Education in the United States* including *Statistical Summary of Education in the United States*, and also *Statistical Abstract of the United States*. The years with data available were driven partly by the publication of the Digest, which typically occurred every other year but sometimes at other intervals. In selecting years to collect data, we were also influenced by the goal to collect indicators across a wide range of years. For each year, four data elements were collected for overall grades/levels in each state (i.e. data broken down further by grade/level were not available): number of teachers (sometimes described as instructional staff or classroom teachers), number of enrolled students, number of days in the academic year (term length), and average daily attendance. From these four elements, three indicators of school quality were derived: term length, percent attendance, and student teacher ratio. Prior to the Brown vs Board of Education decision in 1954, many states maintained racially segregated schools and each data element was typically reported separately for schools serving black children and schools serving white children. After 1954 all states, and prior to 1954 non-segregated states, reported a single value for all children. When available, we used data from the segregated schools, but for white children prior to 1954 we filled in missing values with the data for all children in that state/year. After 1954, we filled in missing data for either white or black children with data for all children in that state/year. States were categorized according to known Census Bureau FIPS codes, which have remained constant over time.

County level data

State statistical reports were collected for selected years from 1905 to 1973 for all available states and counties, obtained through the Harvard Gutman Education School Library or inter-library loans. Copies were made of relevant pages, sent to the University of Alabama, and data was entered there by hand with an estimated error rate of approximately 3.8%. Error rate was calculated using the number of flagged anomalies during data re-checking out of the total number of entries. For available years, four data elements were collected for reported counties for each schooling levels (elementary, middle, and high school): number of teachers (sometimes described as instructional staff or classroom teachers), number of enrolled students, number of days in the academic year (term length), and average daily attendance. Data formats differed across states and years (reporting formats were constant across counties but varied at the state-level), so the values were picked which were closest to these four elements. These data augment data posted by Lleras Muney [*Lleras-Muney A. Compulsory attendance and child labor state laws, STATA format. 2005.*]. From the four elements, three indicators of school quality were derived: term length, percent attendance, and student teacher ratio. Prior to the Brown vs Board of Education decision in 1954, many states maintained racially segregated schools and each data element was typically reported separately for schools serving black children and schools serving white children. After 1954 all states, and prior to 1954 non-segregated states, reported a single value for all children. When available, we used data from the segregated schools, but for white children prior to 1954 we filled in missing values with the data for all children in that state/year. After 1954, we filled in missing data for either white or black children with data for all children in that state/year. Counties were categorized according to contemporary Census FIPS codes. To the extent that census track FIPS codes have changed, there will be inaccuracies in some linkages because we did not use cohort specific

boundary linkages. Please see the Census Bureau for more information on changes to census track FIPS codes over time: https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.1970.html

A2. Administrative Data Interpolation Procedures

State level data

To interpolate between years, using a data set which included one observation per state per year from 1900 to 1975 (note out of range predictions, raw data ranged from 1905-1973), we specified mixed models for predicting each of the three quality variables with random intercepts for region and states within regions and random slopes for year and year squared. Models also included an indicator for whether this was a state that was segregated in that year (i.e., a segregated state prior to 1954). This specification was chosen after comparing simple models with only random intercepts but fixed slopes and models including 4 cubic spline terms for year. Model performance was generally similar, and no single model performed the best for all 3 outcomes. We chose a model that performed well and was relatively interpretable. Predicted values were derived for each state and year based on the fixed effects predictions plus the random effects for region, each state, and the year and year squared terms (times their respective year values). When the original school quality values were regressed upon these predicted values, the r-squared values were 0.89 (term length in black schools), 0.88 (term length in white schools), 0.84 (percent attendance in black schools), 0.87 (percent attendance in white schools), 0.92 (student teacher ratio in black schools), and 0.80 (student teacher ratio in white schools).

To account for the likely period in which school quality is likely to influence students, we created state-level school quality variables with varying lead times (1 year, 6 years, 10 years, and 14 years). For example, a 6-year lead time variable will assume that school quality values in 1950 would apply to children born in 1944. To use lead time values, merge year of the state school quality measures with participants' birth year and select the school quality lead variable most appropriate for the school period of interest (e.g. 6 years of age, 10 years of age, 14 years of age).

County level data

To interpolate between years, using a data set which included one observation per county per year from 1900 to 1975 (note out of range predictions, raw data ranged from 1905-1973). We first imputed information across school levels (high school, middle school, and elementary) and segregation groups (schools serving white and black children before 1954) based on the nearest possible school type. Then, we specified mixed models for predicting each of the three quality variables with a random intercept for county and random slopes for year and year squared. Models also included race-specific state quality measures, indicators for school level (high school, middle school, or elementary), and an indicator for whether this was a state that was segregated in that year (i.e., a segregated state prior to 1954). This specification was chosen after comparing simple models with only random intercepts but fixed slopes and models including quadratic or cubic spline terms for year. Model performance was generally similar; no single model performed the best for all 3 outcomes. We chose a model that performed well and was relatively interpretable. Predicted values were derived for each county and year based on the fixed effects predictions plus the random effects for county and the year and year squared terms (times their

respective year values). When the original school quality values were regressed upon these predicted values, the r-squared values were 0.81 (term length in black schools), 0.70 (term length in white schools), 0.52 (percent attendance in black schools), 0.47 (percent attendance in white schools), 0.63 (student teacher ratio in black schools), and 0.64 (student teacher ratio in white schools).

A3. Calculation of Policy Predicted Years of Education (PPYEd)

We restricted the Census based analyses to match the birth years found in the REGARDS cohort (1908 or later). In addition, because we wanted to estimate predicted years of schooling among those who were likely to have completed their education, we included only people 25 years or older in the respective censuses. For people in the 1980 5% sample, we restricted the analysis to those who were born between 1908 and 1952. For people in the 1990 5% sample, we restricted the analysis to those born between 1908 and 1962.

Another consideration was the lead time to apply to the school quality and CSL measures since school quality and CSLs most likely do not influence the participant at the year of birth. We chose a lead time of 6 years for the school quality measures, e.g., we assumed that school quality values in 1950 would apply to children born in 1944. For the compulsory schooling laws, a lead time of 6 years was chosen for mandatory age at school enrollment, and a lead time of 14 years was chosen for the youngest age individuals could legally drop out of school and youngest age individuals could receive a work permit. For each respondent, years of compulsory schooling were calculated by taking the difference between enrollment age when respondents were 6 years old and minimum drop-out age (CSL) or minimum work permit age (CSLw) when the respondents were 14 years old. For PPYEd, interpolated CSL measures that calculated using imputation models (described under Administrative Data Interpolation Procedures) were used. Additional models were explored that instead used time updated CSL measures based on most recently reported state CSLs following first available state CSLs. Additional details on model comparisons are available below.

PPYEd was calculated using:

$$\text{Predicted(education)} = \beta_0 + \beta_1\text{CSL} + \beta_2\text{CSLw} + \beta_3\text{CSL*Black} + \beta_4\text{CSLw*Black} + \beta_5\text{TLb} + \beta_6\text{TLw} + \beta_7\text{Attb} + \beta_8\text{Attw} + \beta_9\text{Strb} + \beta_{10}\text{Strw} + \beta_{11}\text{TLb*Black} + \beta_{12}\text{TLw*Black} + \beta_{13}\text{Attb*Black} + \beta_{14}\text{Attw*Black} + \beta_{15}\text{Strb*Black} + \beta_{16}\text{Strw*Black} + \beta_{17}\text{C}$$

C = cubic splines for birth year, birth year*Black interactions, Black, sex, and state indicators
CSL= minimum drop_out age – enrollment age
CSLw = minimum work permit age – enrollment age
TL=term length; Att=attendance; Str=student-teacher ratio

The predicted values from this model (PPYEd) were then linked to individual level data in the REGARDS cohort based on state of residence at age 6, year of birth, race, and sex.

Summary of PPYEd Model Comparisons

Using the combined 1980 and 1990 census 5% sample, we ran a series of models to compare model fit across different operationalizations of school quality measures and compulsory schooling laws (CSLs). The base model included cubic spline for birth year, birth year by race interactions, race, sex, and state indicators. All subsequent models included the covariates from this base model. Model 1 added linear operationalization of the school quality or CSL measures. Model 2 included linear and quadratic terms for school quality or CSLs. Model 3 included cubic splines for school quality or CSLs measures. Model 4 included linear operationalization of school quality and interaction between school quality/CSL and race. Model 5 included linear and quadratic operationalization of school quality/CSL and interactions between school quality/CSL and race. Model 6 included cubic splines for the school quality/CSL measures and interaction between school quality/CSLs and race.

Models with alternating compositions with and without CSLs and school quality measures were compared using R^2. Comparisons of models using imputed CSLs are found in Table 1 and comparisons of models using time-updated CSLs and an indicator for missing CLS are found in Table 2.

For clarification, the fifth column in the tables below contains models with both CSLs and school quality measures with 6-year leads. The last column in each table uses a split sample design where a random variable between 0-1 is generated. Models 0-6 (for the model containing both CSLs and school quality measures with 6-year leads) are fitted for the first half of the sample (random variable <0.5). Predicted values are obtained for each model. Then, we fit a regression of years of education on predicted value of education for those with random variable >0.5 (the other half of the sample). In comparing the models, it appears that Model 4 appears to be the best fit for both imputed CSLs and time updated CLSs with a reasonable interpretability.

Table 1. Model comparisons for State School Quality Measures and Imputed CSLs

	1980 + 1990 School quality, 10 year lead	1980 + 1990 School quality, 14 year lead	1980 + 1990 School quality, 6 year lead	1980 + 1990 CSLs*	1980 + 1990 school quality + CSLs*	Split sample, school quality + CSLs*
	R^2	R^2	R^2	R^2	R^2	R^2
Model 0			0.1516	0.1516	0.1516	0.1482
Model 1			0.1536	0.1525	0.1537	0.1502
Model 2			0.1539	0.1529	0.1541	0.1501
Model 3			0.1543	0.1532	0.1545	0.1496
Model 4			0.1547	0.1525	0.1548	0.1515
Model 5			0.1551	0.1529	0.1553	0.1514
Model 6	0.1555	0.1555	0.1556	0.1534	0.1559	0.1510

Model 0: without school quality measures: just cubic splines for birthyear, birthyear*race interactions, race, sex, and state indicators
Model 1: Model 0 + linear operationalization of school quality measures
Model 2: linear and quadratic terms for school quality measures + Model 0
Model 3: cubic splines for school quality measures + Model 0
Model 4: linear operationalization of school quality, interaction between school quality and race + Model 0
Model 5: linear and quadratic operationalization of school quality, interaction between school quality and race + Model 0
Model 6: cubic splines for school quality measures, interaction between school quality and race + Model 0
Holding N's same for Models with just CSLs or school quality
CSLs: comyrs = drop_age with 14 year lead, enrollment age with 6 year lead
CSLs: childcom = work age with 14 year lead - enrollment age with 6 year lead

Table 2. Model comparisons for State School Quality Measures and Time-Updated CSLs

	1980 + 1990 School quality, 10 year lead	1980 + 1990 School quality, 14 year lead	1980 + 1990 School quality, 6 year lead	1980 + 1990 CSLs*	1980 + 1990 school quality + CSLs*	Split sample, school quality + CSLs*
	R^2	R^2	R^2	R^2	R^2	R^2
Model 0			0.1516	0.1516	0.1516	0.1482
Model 1			0.1536	0.1521	0.1536	0.1502
Model 2			0.1539	0.1525	0.1541	0.1502
Model 3			0.1543	0.1526	0.1544	0.1499
Model 4			0.1547	0.1522	0.1548	0.1515
Model 5			0.1551	0.1526	0.1553	0.1516
Model 6	0.1555	0.1555	0.1556	0.1527	0.1558	0.1513

Model 0: without school quality measures: just cubic splines for birthyear, birthyear*race interactions, race, sex, and state indicators
Model 1: Model 0 + linear operationalization of school quality measures
Model 2: linear and quadratic terms for school quality measures + Model 0
Model 3: cubic splines for school quality measures + Model 0
Model 4: linear operationalization of school quality, interaction between school quality and race + Model 0
Model 5: linear and quadratic operationalization of school quality, interaction between school quality and race + Model 0
Model 6: cubic splines for school quality measures, interaction between school quality and race + Model 0
Holding N's same for Models with just CSLs or school quality
CSLs: comyrs = drop_age with 14 year lead, enrollment age with 6 year lead
CSLs: childcom = work age with 14 year lead - enrollment age with 6 year lead
All models include an indicator for missing CLSs