STANFORD EDUCATION DATA ARCHIVE

TECHNICAL DOCUMENTATION

Joseph C. Van Matre, Kenneth Shores, Demetra Kalogrides, and Sean F. Reardon
Version 1.1
28 July 2016

CONTENTS

INTRODUCTION AND DATA DESCRIPTION 2

LINKING SCHOOLS TO GEOGRAPHIC SCHOOL DISTRICT BOUNDARIES 3

District Level Analysis 3
Traditional Public Schools 3
Charter Schools 4
Virtual Schools 4
District Changes 4
Overlapping Districts 4
Districts Without SDDS Data 5
Geographic Crosswalks and Shape Files 5

ESTIMATION 5

Data Anomalies 5
Estimating Means and Standard Deviations 6
Estimating Means 6
Estimating Gaps in Racial Achievement 7

DATA SUPPRESSION AND PRIVACY PROTECTION 7

LINKING STATE TEST SCORES TO THE NAEP SCALE 7

POOLING ESTIMATES 8

VERSIONING AND PUBLICATION 9

REFERENCES 9

INTRODUCTION AND DATA DESCRIPTION

This document describes the source data and procedures used to prepare the data sets available on the Stanford Education Data Archive (SEDA; seda.stanford.edu)[1]. The data contained in the archive include:

1. Mean test scores and standard errors (on state standardized tests) for nearly all school districts in the United States,
2. Test score gaps between white and black students and white and Hispanic students for nearly all school districts in the United States, and
3. Socioeconomic, demographic and segregation characteristics of geographical school districts.

The test score source data come from the EDFacts data system at the U.S. Department of Education (USEd), which collects aggregated test score data from each state's standardized testing program. The data were obtained through a restricted use license from the National Center for Education Statistics (NCES). The EDFacts data collected for SEDA include over 200 million standardized assessment outcomes for students in SY 2008-2009, 2009-2010, 2010-2011, 2011-2012, and 2012-2013 (the year in the data is indicative of the spring semester; year 2009 indicates SY 2008-2009); grades 3 to 8; and test subjects English/Language Arts (ELA) and Math. Every state administered standardized assessments to all students in public schools in those years, grades, and subjects under the No Child Left Behind (NCLB) Act's amendments to the Elementary and Secondary Education Act (ESEA)[2].

EDFacts provides the school-by-grade-level count of students who are assigned to each performance level in ELA and Mathematics in each state. The assessments used to determine a student's performance level in each subject, the number of possible performance levels, and the cut scores associated with each performance level are determined by states, subject to federal regulations and oversight. For example, some states categorized students' scores into four categories: "below basic," "basic," "proficient," and "advanced." The EDFacts data report the number of students in a school-year-grade-subject that scored in each of the (four, in this case) categories.

In addition to the counts of students scoring in the respective proficiency category for a specific year, grade and subject, each EDFacts file provided counts of student subgroups falling in the respective category. The student subgroups include race/ethnicity, gender, socioeconomic disadvantage, among others. Thus, for each school, year, grade, and test subject we know the total number of students scoring in each performance category, as well as the total number of white, black, Hispanic, etc. students scoring in each of the categories. The raw data include no suppressed cells, nor do they have a minimum cell size.

---

[1] Suggested citation for data: Sean F. Reardon, Demetra Kalogrides, Andrew Ho, Ben Shear, Kenneth Shores, Erin Fahle. (2016). Stanford Education Data Archive (Version 1.1 File Title). http://purl.stanford.edu/db586ns4974.

Suggested citation for technical documentation: Joseph C. Van Matre, Kenneth Shores, Demetra Kalogrides, Sean F. Reardon. (2016). *Stanford Data Archive: Technical Documentation* (Version 1.1). http://purl.stanford.edu/db586ns4974.

[2] These data do not include outcomes for students with disabilities who participated in alternative assessments under alternative achievement standards (i.e. "the 1% rule").

No individual student-level data is included in the EDFacts data.

Some measures of school and community characteristics (e.g., district-level income inequality measures) are available in SEDA. The non-achievement data come from the NCES Common Core of Data (CCD), the School District Demographics System (SDDS), and the US Census Bureau's American Community Survey (ACS).

The following sections outline the process used to create the academic achievement and gap data found on SEDA from the EDFacts data provided by NCES.

LINKING SCHOOLS TO GEOGRAPHIC SCHOOL DISTRICT BOUNDARIES

Each public school in the U.S. can be thought of as belonging to both an "administrative school district" (the local education agency that has administrative control over the school) and a "geographic school district" (a geographic catchment area defined by an administrative district). We define each school's administrative district based on the school's local education agency (LEA) as reported by the EDFacts and CCD data.

Most traditional LEAs have a geographic boundary that defines their district. For all schools that are geographically located within the boundaries of a given traditional LEA, we define this LEA as the school's geographic district. Most traditional public schools have the same geographic and administrative district. Many charter schools or schools administered by the state (e.g. state magnet schools, schools for the blind, etc.) do not belong to an administrative district that has a corresponding geographic boundary. Many of these schools are located in the geographic boundaries of a different local education agency. Such schools have different administrative and geographic districts.

The current SEDA data release contains estimates based on geographic school districts. That is, estimates are based on test scores of students attending schools that fall within a given geographic boundary; district test score distribution estimates in SEDA will therefore, in some cases, be based on test scores from schools from multiple administrative districts.

Schools are placed into a geographic boundary to allow for comparisons between district-level achievement and demographic and economic information. Demographic and economic data come from the American Community Survey (ACS), available from the School District Demographics System (SDDS, National Center for Educational Statistics, US Department of Education). These data do not distinguish between administrative and geographic districts; ACS contains economic and demographic information about all children living in a geographic school district. Thus, in order to compare local economic information about a district and that district's achievement, it is necessary to merge all schools in that geographic area into a common local education agency that can be merged to the ACS data.

Certain decision rules are followed to determine whether a school is located in a particular geographic boundary. The following sections describe the decision rules that were used to classify schools into distinct units of analysis or to exclude anomalous data.

DISTRICT LEVEL ANALYSIS

TRADITIONAL PUBLIC SCHOOLS

Traditional public schools were placed into their administrative district as listed in the CCD (which corresponds to the geographic boundaries of a traditional school district).

Charter Schools

If a charter school's administrative district is listed in the CCD as belonging to a traditional public school district (a district that includes non-charter public schools), it is placed into its administrative district as listed in the CCD; it is treated as if it were a traditional public school. If a charter school is listed in the CCD as belonging to a district that only has charter schools or is authorized by a state-wide administrative agency, it is coded as belonging to the geographic school district in which it is located.[3]

Virtual Schools

By their nature, most virtual schools do not draw students from within strict geographic boundaries in a state (or even from within a single state). The CCD does not identify which schools are virtual schools in the years that are included in SEDA. The CCD does specifically identify virtual schools in the 2013-14 school year. We identify all schools that are listed as virtual schools in the 2013-14 school year as virtual schools in the 2008/09-2012/13 school years. We identify additional virtual schools by searching school names for the terms "virtual", "cyber", "online", "internet", "distance", "extending", and "extended". Some naming or classification of schools was ambiguous. When the type of school was unclear, research staff consulted school and district websites for additional details. Schools whose primary mode of instruction was online but that required regular attendance at a computer lab or school building were coded as belonging to the geographic school district in which they are located. For purposes of estimating district test score means, virtual schools are retained in the estimation, but are assigned their own "geographic district" ID ([fips state id]99999), so that their students' scores are included in the estimation procedures, but are not included in any geographic district's score distribution.

District Changes

Some districts changed shape over the 2008/09-2012-13 school years. In California, two Santa Barbra districts (LEA IDs: 0635360, 0635370) joined to become the Santa Barbara Unified School District. In South Carolina, two districts joined to become the Sumter School District (LEA IDs: 4503720, 4503690). In both cases, SEDA contains estimates of test score means for the two original school districts in all years in order to link them to covariate data from the SDDS. A single estimate for the new combined district can be obtained by computing the weighted average of the means within each grade, year, and subject.

The CCD assigns schools in New York City to one of thirty-two geographic districts or one "special schools district". All New York City Schools are aggregated to the city level and given the same district code, creating one unified New York City district code.

Overlapping Districts

Districts overlapped in two ways. In some states, many districts have separate elementary and high school districts (i.e. there are several geographically disjoint elementary school districts that feed into a high school district that covers the union of the elementary school districts). In some cases, the high school districts include students in grades 7 and 8. In such cases, elementary schools are assigned to

[3] Geographic location is determined by the latitude and longitude coordinates of a school's physical address as listed in the CCD. The location of charter schools sometimes varies from year to year. This can result in the charter school being placed in different geographic districts in different years.

their administrative district as listed in the CCD and high schools with grades 7 and 8 are assigned to the elementary district in which they are geographically located.

A few school districts overlap state borders. In this case, schools on either side of the state border take different accountability tests. We treat these districts as two districts, each one coded as part of the state in which it resides.

Districts Without SDDS Data

Some schools belong to county-run districts that do not have data in SDDS. All such schools in a given state are given a special geographic district identifier ([fips state id]99998).

There are some traditional public school districts that that do not have data in SDDS. All such schools in a given state are given a special identifier, distinct from the county-run district identifier ([fips state id]99997).

Geographic Crosswalks and Shape Files

The crosswalk used to place schools into their geographic unit of analysis are available on the SEDA website. While every effort is made to ensure schools are placed in the proper geographic unit based on the decision rules described in the previous sections, if you believe that a crosswalk contains an error, please contact sedasupport@stanford.edu.

The shape files used to locate schools within each geographic unit are also available online. The county, MSA, and commuting zone shape files are original from the US Census Bureau. A district level shape file was created using the U.S. Census Bureau's 2010 TIGER/Line Files. These files were from the National Historical Geographic Information System (NHGIS). The Census Bureau provides three shape files: elementary district boundaries, high school district boundaries, and unified district boundaries. Research staff merged the elementary and unified shape files to conform to the decision rules outlined above.

ESTIMATION

Data Anomalies

There are some idiosyncratic instances that make specific school-year-subject level observations in the EDFacts data inappropriate for inclusion in the estimation process or required specific modifications for inclusion. The following describes these instances:

- Districts were permitted to administer locally-selected assessments in Nebraska during SY 2008-2009 (ELA and Math) and SY 2009-2010 (Math). Because these assessments were scored on different scales and using different cut scores, proficiency counts cannot be compared across Nebraska districts or schools in these cases. EDFacts assessment data from Nebraska for these specific subjects and years are not included in SEDA.
- Students in grades 7 and 8 in California take Math assessments corresponding to the course they are enrolled in, not their grade level. Because all students in a given grade do not take a common assessment, proficiency counts cannot be compared across districts or schools in these cases. EDFacts assessment data from California for these grades are not included in SEDA.
- For one district, grade and year in Arkansas and Louisiana, respectively, the reported scores were implausible given the available data for other grades and years. In particular, the distribution of students across proficiency categories for the given cohort changed too abruptly in

the given year compared with their performance in the prior and subsequent years, as well as compared with other cohorts in the district, to be believable change. These data determined to be entry errors and were removed.
- In SY 2012-2013, there were fewer than 10 students in the lowest proficiency category for grade 3 Math assessments in South Dakota. The lowest proficiency category was combined with the second lowest proficiency category prior to estimation for this case.
- Wyoming did not report any assessment outcomes in SY 2009-2010. Estimates are not available for these years.

Estimating Means and Standard Deviations

Estimating Means

The mean and standard error (SE) of the mean are estimated for each geographic unit of analysis. The mean and SE of the mean are estimated from the proficiency counts reported in the EDFacts data by fitting an ordered probit model as outlined in Reardon, Shear, Castellano and Ho (2016). The following outlines the type of model used under various conditions:

- A partially heteroskedastic ordered probit (PHOP) model is used in cases when there are more than 3 proficiency categories. When there are 50 or more assessment outcomes for a particular observation (district-subgroup-year-grade-subject), the SD is freely estimated. When there are fewer than 50 assessment outcomes for a particular observation, the SDs of scores are constrained such that they are constant for all districts (within a state-grade-year-subject cell) with fewer than 50 tested students. See Reardon, Shear, Castellano and Ho (2016) for details on the PHOP model.
- A homoscedastic ordered probit (HOMOP) model is used when there are only two proficiency categories, as the PHOP model cannot be estimated in those cases. See Reardon, Shear, Castellano and Ho (2016) for details on the HOMOP model.

Means and standard deviations were not individually estimated for district-grade-year-subject cells with fewer than 20 students. Instead, we combine all observations in a state with fewer than 20 students into a “residual district” and estimate the mean and SD of test scores in the residual district. Including residual districts allows for all student scores to be used in the estimation. The estimates of the mean and SD in the residual district are not included in the publicly released files as they do not correspond to any single identifiable place. As in the case of general estimation, if the residual district has fewer than 20 observations, the residual district is not included in the estimation.

Estimates are not reported in SEDA under the following conditions:

- District-grade-year-subject observations where all students were in a single proficiency category (the mean and SD cannot be estimated in this case).
- The mean is estimated with extreme imprecision (the SE of the state referenced mean is greater than 2).
- A cell (district-grade-year-subject) includes test scores of fewer than 20 students.
- Grade 7 and 8 mathematics in California.
- 2008-09 ELA and math estimates and 2009-10 math assessments in Nebraska.
- 2008-09 ELA and math estimates in Wyoming.

- District-level EDFacts data contained obvious errors (e.g., the two districts in Arkansas and Louisiana mentioned above).

The resulting estimates are scaled in units of state-grade-year-subject student-level test score standard deviations.

Estimating Gaps in Racial Achievement

Achievement gaps are estimated using the V-statistic described by Ho and Reardon (2012) and Reardon and Ho (2015). Our agreement with the U.S. Department of Education restricts publication of average scores or gaps to cases where the data contain at least 20 assessment outcomes (in each group reported).

Estimates are reported with a corresponding flag that indicates the percentage of students in the district for whom race data was provided. Some districts did not report race data for all students; this flag provides a measure of the quality of the gap estimates. In almost all districts, race data is available for more than 95% of the test scores.

DATA SUPPRESSION AND PRIVACY PROTECTION

The raw EDFacts data contain no suppressed cells. However, our agreement with USEd restricts publication of average scores or gaps to cells where the data contain at least 20 assessment outcomes (in each group reported). In each reported cell, a small amount of random noise is added to each estimate in proportion to the sampling variance of the respective estimate. This is done to ensure that the raw counts of students in each proficiency category cannot be recovered from published estimates.

The random error added to each estimate is drawn from a normal distribution $\mathcal{N}(0, \widehat{\omega^2}/n)$ where $\widehat{\omega^2}$ is the squared estimated standard error of the estimate and n is the number of student assessment outcomes to which the estimate applies. Imprecise estimates, typically for smaller districts, have greater noise added, and more precise estimates (typically for larger districts) have less noise added. SEs of the mean are adjusted to account for the additional error. The added noise is roughly equivalent to randomly removing one student's score from each cell (each district-grade-year).

LINKING STATE TEST SCORES TO THE NAEP SCALE

The estimated means and standard deviations produced by the ordered probit model are scaled relative to their state-specific distributions. The National Assessment of Educational Progress (NAEP) tests provide a state's mean and standard deviation on the same scale, each state's estimated test scores can be placed on the NAEP scale. Reardon, Kalogrides, and Ho (2016) describe the method used to link the state-specific estimates to the NAEP scale. SEDA reports both state-specific cores and the NAEP scale score.

In order to make these linked estimates usefully interpretable, they are standardized in three ways. The first takes the linked NAEP scores and standardizes them to the national NAEP distribution, within each grade-year-subject. This metric is interpretable as an effect size, within a grade-year-subject. The second standardizes the linked means by dividing by the national grade-subject-specific standard deviation for the middle cohort of our data. This metric is interpretable as an effect size, relative to the grade-specific standard deviation of scores in one cohort. This has the advantage of being able to describe aggregated changed over time in test scores. The third standardization divides the linked scores by the average

difference in NAEP scores between students one grade level apart. A one-unit difference in this grade-equivalent unit scale is interpretable as equivalent to the average difference in skills between students one grade level apart in school. The standardization and interpretation of the scores is described in more detail in Reardon, Kalogrides and Ho (2016).

POOLING ESTIMATES

SEDA provides grade-year-subject specific estimates, as well as estimates pooled across grades and years (within districts, for each subject separately). Pooling provides more precise estimates of district (or other geographic area) test score patterns than do individual district-grade-year-subject estimates. SEDA provides pooled estimates based on random coefficient (multi-level) models. These models are based on up to 60 subject-grade-year estimates for a given district, adjusting for grade and cohort. The models weight the estimates by the precision of each of the 60 estimates. They allow each district to have a district-subject-specific intercept (average score), a district subject-specific linear grade slope (rate at which scores change across grades, within a cohort), and a district subject-specific cohort trend (the rate at which scores change across student cohorts, within a grade). The model is as follows:

$$\hat{Y}_{usgy} = \left[(\alpha_{0m0} + v_{um0}) + (\alpha_{0m1} + v_{um1})(g_g - 5.5) + (\alpha_{0m2} + v_{um2})(y_y - 2010.5)\right]M_s$$
$$+\left[(\alpha_{0e0} + v_{ue0}) + (\alpha_{0e1} + v_{ue1})(g_g - 5.5) + (\alpha_{0e2} + v_{ue2})(y_y - 2010.5)\right]E_s + e_{usgy} + \epsilon_{usgy}$$

$$\epsilon_{usgy} \sim N\left[0, \hat{\omega}^2_{usgy}\right]$$

$$e_{usgy} \sim N[0, \sigma^2]$$

$$\begin{bmatrix} v_{um0} \\ \vdots \\ v_{ue2} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{m0} & \cdots & \tau_{m0e2} \\ \vdots & \ddots & \vdots \\ \tau_{m0e2} & \cdots & \tau_{e2} \end{pmatrix}\right] = N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\tau}\right]$$

Here, $\hat{Y}_{usgy}$ is the estimated achievement for district *u* in subject *s* in grade *g* in year *y*. M_s and E_s indicate whether the estimate is for math or ELA, respectively. Sampling error is ϵ_{usgy}, with variance $\hat{\omega}^2_{usgy} = var\left(\hat{Y}_{usgy}\right)$. Within-unit variation not captured by the subject-specific grade and year trends are indicated by e_{usgy}, with a constant variance σ^2 which is estimated. Unit-specific average math and ELA levels, grade trends, and cohort trends may vary among units, and are assumed to follow a multivariate normal distribution $\boldsymbol{\tau}$ which must be estimated.

SEDA also provides estimates pooled across grades, years, and subjects. This model is as follows:

$$\hat{Y}_{usgy} = \left[(\alpha_{00} + v_{u0}) + (\alpha_{01} + v_{u1})(g_g - 5.5) + (\alpha_{02} + v_{u2})(y_y - 2010.5) + (\alpha_{03} + v_{u3})(M_s - 0.5)\right]M_s$$
$$+ e_{usgy} + \epsilon_{usgy}$$

$$\epsilon_{usgy} \sim N\left[0, \hat{\omega}^2_{usgy}\right]$$

$$e_{usgy} \sim N[0, \sigma^2]$$

$$\begin{bmatrix} v_{u0} \\ \vdots \\ v_{u3} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0 & \cdots & \tau_{03} \\ \vdots & \ddots & \vdots \\ \tau_{03} & \cdots & \tau_3 \end{pmatrix}\right] = N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\tau}\right]$$

This model allows each district to have a district-specific intercept (average score, pooled over subjects), a district-specific linear grade slope (rate at which scores change across grades, within a cohort, pooled

over subjects), and a district-specific cohort trend (the rate at which scores change across student cohorts, within a grade, pooled over subjects), and a district-specific math-ELA difference.

The Empirical Bayes (EB) estimates of the intercept, grade slope, and cohort trends from these models included in the SEDA data.

VERSIONING AND PUBLICATION

New or revised data is posted periodically to SEDA. If you indicate that you would like to be notified about new postings when filling out the data use agreement, you will receive an email notifying you of any updates.

SEDA updates that contain substantially new information are labeled as a new version (e.g. V1.0, V2.0). Updates that make corrections or minor revisions to previously posted data are labeled as a subsidiary of the current version (e.g. V1.1, V1.2, etc.) When citing any SEDA data set for presentation, publication or use in the field, include the version number in the citation. All versions of the data will remain available on SEDA to facilitate data verification and research replication.

REFERENCES

Ho, A.D., & Reardon S.F. Estimating Achievement Gaps From Test Scores Reported in Ordinal "Proficiency" Categories. Journal of Educational and Behavioral Statistics August 2012 37: 489-517, first published on October 26, 2011doi:10.3102/1076998611411918

Reardon, S.F., & Ho, A.D. Practical Issues in Estimating Achievement Gaps From Coarsened Data. Journal of Educational and Behavioral Statistics April 2015 40: 158-189, doi:10.3102/1076998615570944

Reardon, S.F., Kalogrides, D., & Ho, A. (2016). Linking U.S. School District Test Score Distributions to a Common Scale, 2009-2013 (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09

Reardon, S.F., Shear, B.R., Castellano, K.E., & Ho, A.D. (2016). Using Heteroskedastic Ordered Probit Models to Recover Moments of Continuous Test Score Distributions from Coarsened Data (CEPA Working Paper No.16-02). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-02