**Stanford Education Data Archive (SEDA)**

**Summary of Data Cleaning, Estimation, and Scaling Procedures**

Sean F. Reardon, Demetra Kalogrides, Kenneth Shores

Version 1.0: 27 April 2016

1. **Data Documentation**

This document describes the data—and procedures used to produce the data—that are available through the Stanford Education Data Archive (SEDA; [seda.stanford.edu](file:///C%3A%5CUsers%5Csreardon%5CAppData%5CLocal%5CMicrosoft%5CWindows%5CTemporary%20Internet%20Files%5CContent.Outlook%5CYQ8LMH13%5Cseda.stanford.edu)). The data contained in this archive describe 1) average test scores (on state standardized tests) for nearly all school districts in the United States; 2) academic achievement gaps between whites and blacks and whites and Hispanics for the majority of ethnic minority students in the United States; and 3) socioeconomic, demographic and segregation characteristics of these districts. The raw test score data come from the EdFacts data system at the U.S. Department of Education, which collects aggregated test score data from each state’s standardized testing program. No individual student-level data is included in the raw data. The EdFacts data includes aggregated data based on over 200 million standardized tests in English/Language Arts (ELA) and Math taken by students in grades 3-8 from the 2008-2009 to 2012-2013 school years. The non-achievement data comes from two publicly available sources: the Common Core of Data and the American Community Survey. From these data we have constructed measures of school and community characteristics (e.g., district-level income inequality measures); these measures are available in the SEDA Archive.

* 1. **Data Cleaning**

The EdFacts data include data for years 2009, 2010, 2011, 2012, and 2013 (the year indicates the Spring of the academic year, so “2009” refers to the 2008-09 school year); grades 3 to 8; and test subjects ELA and Math. Every state administered standardized tests to all students in public schools in those years/grades/subjects under the No Child Left Behind (NCLB) Act.

The raw EdFacts data contain counts of students scoring in each of the state’s proficiency categories. For instance, a state may categorize students’ scores into 4 categories: “below basic,” “basic,” “proficient,” and “advanced.” The EdFacts data record the number of students in a school-year-grade-subject that scored in each of the respective 4 categories. Each file that we received contains counts of students scoring in the respective proficiency category for a specific year, grade and subject. Within each file, we also observe counts of student subgroups falling in the respective category. That is, for any given school, year, grade, and test subject we know the total number of students scoring in 1 of the 4 hypothetical categories, as well as the total number of white, black, Hispanic, etc. students scoring in the categories. The raw data include no suppressed cells nor do they have a minimum cell size. Not all schools report the same subgroups. The data presented in the archive has information about all students (total students), whites, blacks and Hispanics.

We first define a set of crosswalks that link schools to districts, districts to counties, and counties to commuting zones and metros. The following describes those data cleaning decisions and the crosswalks that were constructed.

* + 1. **Charter Crosswalk**

Many charter schools have a unique school district identifier (an LEAID, in the data) that is different than the LEAID of the traditional school district in which they are geographically located. Rather than treat charter schools as separate school districts, we assign charter schools to the traditional school district in which they are physically located. We do so in order that we can relate average district test scores to local community and socio-demographic characteristics (the sociodemographic data come from the American Community Survey data, which are tabulated by school district geography; there are no ACS tabulations for schools that do not have a geographic catchment area). Moreover, we are interested in whether the presence of charter schools, and patterns of charter school enrollment, leads to higher or lower overall academic achievement among students in a given community. One could analyze the data differently (and in the future we will likely provide data that reports average test scores separately for charter schools and traditional districts), but that is not what we have done in constructing the current SEDA files.

To link charter schools to local traditional school districts, we constructed a crosswalk using data from the Common Core of Data (CCD). For every charter school that has an LEAID that does not correspond to a traditional school district, we assign a local LEAID that is the LEAID of the traditional district in which it is geographically located. The geographically located district-level identifier is constructed based on the latitude and longitude coordinates for the charter school in question, available from the CCD.

This charter-LEAID crosswalk is then merged onto the school-level achievement data file we received from EdFacts. We use the local LEAID variable when we aggregate school data to the district level for our analyses (so charter schools are aggregated into their local school district). In the data files, this local LEAID is named “leaidC”; the “C” indicates that charter schools are combined with local school districts.

* + 1. **New York City Crosswalk**

EdFacts achievement data are disaggregated into New York City’s 33 supervisory school districts. We combine these disaggregated supervisory districts into a single district for NYC Public Schools. To do this, we construct a crosswalk with the 33 unique supervisory district IDs and a corresponding aggregated identifier. This crosswalk is merged onto the school-level file, and the new aggregated identifier replaced the 33 original identifiers. Any charter schools that were geographically located in one of the 33 supervisory districts now also belong in the aggregated NYC Public Schools identifier.

* + 1. **Santa Barbara, CA and Sumter, SC Crosswalks**

Santa Barbara and Sumter became a unified school districts in 2012, meaning they consolidated from an elementary and secondary district to a unified district. Because we wish to merge these district-level data to the district-level demographic information from the American Community Survey, and because the American Community Survey treats these districts as unconsolidated—elementary and secondary instead of unified—we constructed a crosswalk that placed schools in years 2012 and 2013 in their the original (pre-2012) elementary and secondary districts.

* + 1. **County Crosswalk & Permanent County ID**

We constructed a crosswalk that links district-level identifiers to their counties. We merge this onto the charter-included district-level identifier (“leaidC” from above) so that our data now has an additional geographic coordinate. In total, we have the school-id, the charter included district id, the original district id, and now a county id. Three issues arise with this linking.

The first is that some districts stop existing or the counties to which they were attached in one year change in the following year. So that each district is associated with only one county, we take the county id associated with the district in the last year we observe the district (the vast majority of districts are observed in the last year that we have data, 2013). This county id is then used as the permanent county id associated with the district, which we refer to as “conum\_perm.”

The second issue was that some district county ids were incorrect. This occurred almost exclusively with charter schools that were members of large, multi-state charter organizations. For these schools, most of which were found in Arizona and Indiana, the county id placed schools in the wrong states, which meant that scores attributed to those schools would be benchmarked to the wrong test. To identify the correct county id and state to which the school belonged, NCES records were inadequate. We ended up doing an online search to find the school—often times going to the school’s website directly. From this information, we could find out the school’s geographic address and with this address we could find its county of origin. This county of origin was manually inserted for the respective districts. In total, this affected 18 districts.

Finally, Hawaii and DC are treated as consolidated school districts, but their associated schools have multiple county ids. We gave schools in Hawaii and DC the primary county ids associated with them.

* + 1. **Miscellaneous Corrections**

A few additional states and schools had data that were not amenable for cross-country or cross-time comparisons. For one district, grade and year in Arkansas and Louisiana, respectively, scores changed too abruptly for that year to be plausible. These data were removed. In California, in grades 7 and 8 in Math, not all students are administered the same math test (they take the math test corresponding to the level of math course they are enrolled in). Because students take different tests, math proficiency category counts are not amenable for comparison. In Nebraska, in 2009 (for both Math and ELA) and in 2010 (in Math), districts were permitted to administer locally-designed or determined tests. For these years in Nebraska, data are not amenable for comparison or estimation and were therefore dropped. Finally, in South Dakota, in 2013 grade 3 and math, too few students scored in the lowest proficiency category for our models to estimate an average score. Students scoring in the bottom category were added to the category above. This affected a very small number of students—fewer than 10 in the state.

* + 1. **Metropolitan and Commuting Zone Crosswalk**

We now have a dataset with one county identifier per district, as well as multiple achievement scores. These data are then merged onto a metropolitan and commuting zone crosswalk dataset that contains a unique commuting zone identifier and three metropolitan identifiers. There are three metropolitan area identifiers (coded as 03, 09, and 13), because the Census published new metropolitan area definitions in the years 2003, 2009, and 2013.

Metropolitan areas and commuting zones, in some cases, cross state lines. Because the estimation of student means is relative to state-specific scores, and because measures of achievement gaps are estimated using information between groups (whites and blacks and whites and Hispanics) taking the same test, we construct a second identifier for metros and commuting zones that specifies the metro/commuting zone identifier as well as the state in which it falls. Thus, two districts in the same metro may have different state identifiers, and this variable captures that. For metros and commuting zones, this metro/commuting zone-by-state id variable is used to collapse the data for metropolitan area and commuting zone analyses.

* + 1. **Collapsing**

We construct multiple datasets for years, grades and subjects for different levels of geographic aggregation. The lowest level of aggregation we have is the school. These are the primary files. We then collapse using five other levels of aggregation: counties [1], metropolitan area-by-state id using metropolitan coding in years 2003, 2009, and 2013 [2-4], and commuting zones-by-state ids [5]. The collapse command sums the counts of students scoring in the various proficiency categories. These data can then be used to estimate means and achievement gaps for different levels of geographic aggregation in the United States.

* 1. **Estimating Means**

We estimate district-level means for approximately 13,000 school districts in the United States. We do so by fitting heteroskedastic ordered probit (HETOP) models to the ordered proficiency counts of each district in a give state-grade-year-subject, using the methods described in [Reardon, Shear, Castellano and Ho (2016)](http://cepa.stanford.edu/content/using-heteroskedastic-ordered-probit-models-recover-moments-continuous-test-score-distributions-coarsened-data). Specifically, we use their PHOP model, constraining district’s standard deviations of scores to be constant for cells with 50 or fewer tested students. When there are only two proficiency categories, we use the HOMOP model instead, since the PHOP model required 3 or more categories. The resulting estimates are scaled in units of state-grade-year-subject student-level test score standard deviations. [Reardon, Shear, Castellano and Ho (2016)](http://cepa.stanford.edu/content/using-heteroskedastic-ordered-probit-models-recover-moments-continuous-test-score-distributions-coarsened-data) describe these methods in more detail.

* + 1. **Observations where Estimates are Impossible**

We do not estimate means in cells (e.g., district-year-grade-subject cells) in which the number of students is fewer than 20. We also drop estimated mean scores in cases where the estimated standard errors are greater than 2 (meaning the confidence interval of a district’s mean is 4 student-level standard deviations wide; essentially meaning we have no information). This happens very rarely, and is generally a result of insufficient information for the PHOP model to identify the district’s test score distribution.

* + 1. **Adding Noise to Estimates**

Our agreement with the Department of Education requires us to add a small amount of random noise to each published estimate in order to make recovery of specific cell counts—i.e., the specific counts of students scoring in a specific proficiency category for a given district-year-grade-subject—impossible. The noise we add is based on the number of students in the district, as well as the sampling variance of the estimated achievement gap and its standard error. Random noise is added to each estimate in proportion to the sampling variance of the respective estimate; thus, districts with less precision will have greater noise added, and districts with greater precision (and more students) will have less noise added. Specifically, we add random error to each estimate, where the error is drawn from a normal distribution with mean 0 and variance $\hat{ω}^{2}/n$ (where $\hat{ω}^{2}$ is the squared estimated standard error of the estimate and $n$ in the number of students in the cell to which the parameter applies).

* + 1. **Placing State-Specific Means on National Scale**

The method described by [Reardon, Shear, Castellano and Ho (2016)](http://cepa.stanford.edu/content/using-heteroskedastic-ordered-probit-models-recover-moments-continuous-test-score-distributions-coarsened-data) estimates a district’s average achievement and standard deviation relative to a state-(and grade-year-subject-)specific distribution. To make these state-specific estimates comparable to those in other states, grades, and years, [Reardon, Kalogrides and Ho (2016)](https://cepa.stanford.edu/content/linking-us-school-district-test-score-distributions-common-scale-2009-2013) describe and validate a linking method that links the distribution of state test scores to the distribution of state-specific scores on the National Assessment of Educational Progress (NAEP). Because we know a given state’s mean and standard deviation from the NAEP, and because the same NAEP test is taken across all states, each state’s estimated test scores can be placed on a NAEP metric.

* + 1. **Standardizing Variables for Interpretability**

In order to make these linked estimates usefully interpretable, they are standardized in three ways. The first takes the linked NAEP scores and standardizes 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 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)](https://cepa.stanford.edu/content/linking-us-school-district-test-score-distributions-common-scale-2009-2013).

* + 1. **Pooling Estimates**

In order to make a more parsimonious dataset, as well as to “shrink” unreliable estimates of district means towards the average, we fit a set of random coefficient (multi-level) models. These models take the up to 60 estimates and pool them, adjusting for grade and cohort, and weighting by the precision of each of the 60 estimates. The models allow each district to have a district-specific intercept (average score), a district-specific linear grade slope (rate at which scores change across grades, within a cohort), and a district-specific cohort trend (the rate at which scores change across student cohorts, within a grade). We include the Empirical Bayes (EB) estimates of the intercept, grade slope, and cohort trend in the SEDA archice. These EB estimates correspond to the shrunken average achievement scores in a district for the median grade and cohort in our data (students in grade 5.5 in 2011). We fit these models separately for ELA and Math, and once pooling both subjects to obtain pooled estimates.

* 1. **Estimating Achievement Gaps**

We estimate achieve gaps using the V-statistic described by [Ho and Reardon](http://cepa.stanford.edu/content/practical-issues-estimating-achievement-gaps-coarsened-data) (2012) ; [Reardon and Ho 2015)](http://cepa.stanford.edu/content/estimating-achievement-gaps-test-scores-reported-ordinal-proficiency-categories). Of the approximately 13,000 school districts in the United States, we can estimate a white-black and white-Hispanic achievement gap for approximately 2,600 and 2,900 districts, respectively. The remaining districts have insufficient minority students per grade for an achievement gap to be reliably estimated and reported. Our agreement with the Department of Education restricts publication of average scores or gaps to cases where at least 20 students’ test scores are available (in each group reported).

* + 1. **Adding Noise to Estimates**

As above, we add a small amount of noise to the gap estimates in order to make recovery of specific cell counts—i.e., the specific counts of students scoring in a specific proficiency category for a given district-year-grade-subject—impossible. The procedure is the same as described above.

* + 1. **Pooling Estimates**

For a given district, we have up to 60 estimated white-black and white-Hispanic achievement gaps (up 5 years, 6 grades, and 2 subjects, so long as there are at least 20 white and 20 black/Hispanic students tested in the district-year-grade-subject). These data—with noise added—are available in the SEDA archive. However, we also aggregate these observations in order to have a more reliable, precisely estimated and parsimonious description of achievement inequality for a given district. To this end, we construct meta-analytic averages of a district’s achievement gap, pooling across years and grades (and, in half the cases, across subjects).

The meta-analytic average is estimated by using Stata’s –metareg- command for each district, where the outcome variable is a district’s achievement gap, and the dependent variables are the (centered) grade, year, and subject. We provide three estimates of the average achievement gap: the average gaps in math and in ELA (averaged over years and grades), and a pooled estimate (the average over all grades, years, and both subjects). The meta-analytic averages are regression adjusted to account for differences in among districts in the pattern of which grades, years, and subjects have available achievement gaps.

**References**

Ho, Andrew D., and Sean F. Reardon. 2012. "Estimating Achievement Gaps From Test Scores Reported in Ordinal 'Proficiency' Categories." *Journal of Educational and Behavioral Statistics* 37(4):489-517.

Reardon, Sean F., and Andrew D. Ho. 2015. "Practical Issues in Estimating Achievement Gaps From Coarsened Data." *Journal of Educational and Behavioral Statistics* 40(2):158-89.

Reardon, Sean F., Demetra Kalogrides, and Andrew D. Ho. 2016. "Linking U.S. School District Test Score Distributions to a Common Scale, 2009-2013." in *Unpublished manuscript*.

Reardon, Sean F., Benjamin R. Shear, Katherine E. Castellano, and Andrew D. Ho. 2016. "Using Heteroskedastic Ordered Probit Models to Recover Moments of Coarsened Test Score Distributions." in *Unpublished manuscript*.