Stanford Education Data Archive

Technical Documentation

Version 3.0

July 2019

Erin M. Fahle, St. John's University

Benjamin R. Shear, University of Colorado Boulder

Demetra Kalogrides, Stanford University

Sean F. Reardon, Stanford University

Belen Chavez, Stanford University

Andrew D. Ho, Harvard University

Suggested citation:

Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., Chavez, B., & Ho, A. D. (2018). Stanford Education Data Archive: Technical Documentation (Version 3.0). Retrieved from http://purl.stanford.edu/db586ns4974.

Contents

I. What is SEDA?

The Stanford Education Data Archive (SEDA) is part of the Educational Opportunity Project at Stanford University (https:\\edopportunity.org), an initiative aimed at harnessing data to help scholars, policymakers, educators, and parents learn how to improve educational opportunities for all children. SEDA includes a range of detailed data on educational conditions, contexts, and outcomes in schools, school districts, counties, commuting zones, and metropolitan statistical areas across the United States. Available measures differ by aggregation; see Sections I.A. and I.B. for a complete list of files and data.

By making the data files available to the public, we hope that anyone who is interested can obtain detailed information about U.S. schools, communities, and student success. We hope that researchers will use these data to generate evidence about what policies and contexts are most effective at increasing educational opportunity, and that such evidence will inform educational policy and practices.

The construction of SEDA has been supported by grants from the Institute of Education Sciences, the Spencer Foundation, the William T. Grant Foundation, the Bill and Melinda Gates Foundation, the Overdeck Family Foundation, and by a visiting scholar fellowship from the Russell Sage Foundation. Some of the data used in constructing the SEDA files were provided by the National Center for Education Statistics (NCES). The findings and opinions expressed in the research and reported here are those of the authors alone; they do not represent the views of the U.S. Department of Education, NCES, or any of the aforementioned funding agencies.

I.A. Overview of Test Score Data Files

SEDA 3.0 contains test score data files for schools, geographic school districts (GSDs), counties, commuting zones (CZs), and metropolitan statistical areas (metros). Test score data files contain information about the average academic achievement as measured by standardized test scores administered in 3rd through 8th grade in mathematics and English/Language Arts (ELA) over the 2008-09 through 2015-16 school years. The exact measures reported differ by these levels of aggregation.

School Files. There are two school-level test score data files, corresponding to the two different metrics in which the data are released: the cohort standardized (CS) scale and the grade cohort standardized (GCS) scale. In each file there are variables corresponding to the average test score in the middle grade of the data, the average "learning rate" across grades (grade slope), the "trend" in the test scores across cohorts (cohort slope), and the difference between math and ELA (math slope). Each measure is included along with its respective standard error. Estimates are reported for all students; no estimates are provided by demographic subgroup.

Geographic District, County, Commuting Zone, and Metropolitan Statistical Area Files. Twenty-four test score files are released corresponding to the four units (GSDs, counties, CZs, and metros) by two scales (CS and GCS) by three pooling levels (long, pooled by subject, and pooled overall). "Long" files contain estimates for each grade and year separately; "pooled by subject" (or poolsub) files contain estimates that are averaged across grades and years within subjects; and "pooled overall" (or pool) files contain estimates that are averaged across grades, years, and subjects. In the long files there are variables corresponding to test score means by subgroup and their respective standard errors in each grade, year and subject. In the two types of pooled files, there are variables corresponding to the average test score mean (averaged across grades, years, and subjects), the average "learning rate" across grades and the average "trend" in the test scores across cohorts, along with their standard errors. In the pooled overall file, there is also a variable that indicates the average difference between math and ELA and its standard error. Estimates are reported for all students and by demographic subgroups.

Table 1 lists the files and file structures. Lists of variables can be found in the codebook that accompanies this documentation.

I.B. Covariate Data

SEDA 3.0 also provides estimates of socioeconomic, demographic and segregation characteristics of schools, geographic school districts, counties and metros. The measures included in the district, county, and metro covariates files come primarily from two sources. The first is the American Community Survey (ACS) detailed tables which we obtained from the National Historical Geographic Information System (NHGIS) web portal.[1] These data include demographic and socioeconomic characteristics of individuals and households residing in each unit. The second is the Common Core of Data (CCD) which is an annual survey of all public elementary and secondary schools and school districts in the United States. The data includes basic descriptive information on schools and school districts, including demographic characteristics. [2] The measures included in the school covariates file come from the CCD as well as the Civil Rights Data Collection (CRDC). The CRDC includes data about school demographics, teacher experience, school expenditures, high school course enrollments as well as other information not used here.[3]

Nine files (three per aggregation) in SEDA 3.0 contain CCD and ACS that data have been curated for use with the geographic school district-level, county-level, and metro-level achievement data. These data include raw measures as well derived measures (e.g., a composite socioeconomic status measure, segregation measures). Each of the three covariate files we construct for each unit contain the same variables, but differ based on whether they report these variables separately for each grade and year, average across grades (providing a single value per unit per year) or average across grades and years (providing a single value per unit). A single data file is provided for schools with one observation for each school in each year. The Covariate Data Construction section of the documentation describes more detail about the construction of these data files and the computation of derived variables. Table 2 lists the names and file structures of the covariate data files.

[1] The ACS data is available for download from the NHGIS website at: https://www.nhgis.org/

[2] The CCD raw data can be accessed at https://nces.ed.gov/ccd/.

[3] More information about the Civil Rights Data Collection can be found here: https://ocrdata.ed.gov/

I.C. Data Use Agreement

Prior to downloading the data, users must sign the data use agreement, shown below.

You agree not to use the data sets for commercial advantage, or in the course of for-profit activities. Commercial entities wishing to use this Service should contact Stanford University's Office of Technology Licensing (info@otlmail.stanford.edu).

You agree that you will not use these data to identify or to otherwise infringe the privacy or confidentiality rights of individuals.

THE DATA SETS ARE PROVIDED "AS IS" AND STANFORD MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED. STANFORD SHALL NOT BE LIABLE FOR ANY CLAIMS OR DAMAGES WITH RESPECT TO ANY LOSS OR OTHER CLAIM BY YOU OR ANY THIRD PARTY ON ACCOUNT OF, OR ARISING FROM THE USE OF THE DATA SETS.

You agree that this Agreement and any dispute arising under it is governed by the laws of the State of California of the United States of America, applicable to agreements negotiated, executed, and performed within California.

You agree to acknowledge the Stanford Education Data Archive as the source of these data. In publications, please cite the data as:

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., Chavez, B., Buontempo, J., & DiSalvo, R. (2019). Stanford Education Data Archive (Version 3.0). Retrieved from http://purl.stanford.edu/db586ns4974.

Subject to your compliance with the terms and conditions set forth in this Agreement, Stanford grants you a revocable, non-exclusive, non-transferable right to access and make use of the Data Sets.

II. Achievement Data Construction

II.A. Source Data

The SEDA 3.0 achievement data is constructed using data from the ED*Facts* data system housed by the U.S. Department of Education (USEd), which collects aggregated test score data from each state's standardized testing program as required by federal law. The data include assessment outcomes for eight consecutive school years from the 2008-09 school year to the 2015-16 school year in grades 3 to 8 in English Language Arts (ELA) and math.

Under federal legislation, each state is required to test every student in grades 3 through 8 (and in one high school grade) in math and ELA each year. States have the flexibility to select (or design) and administer a test of their choice that measures student achievement relative to the state's standards. States then each set their own benchmarks or thresholds for the levels of performance or "proficiency" in each grade and subject. States are required to report the number of students scoring who are "proficient," both overall and disaggregated by certain demographic subgroups, for each school. More often, states report the number of students scoring at each of a small number (usually 3-5) of ordered performance levels, where one or more levels represent "proficient" grade-level achievement.

When states report this information to the USEd, it is compiled into the ED*Facts* database. The ED*Facts* database reports the number of students disaggregated by subgroup scoring in each of the ordered performance categories, for each grade, year and subject; *no individual student-level data is reported*. The student subgroups include race/ethnicity, gender, and socioeconomic disadvantage, among others. In 2013-2016, the data is further broken out by assessment type: regular assessments, regular assessments with accommodations, and alternate assessments with grade-level standards, modified standards and alternate standards. However, in 2009-2012, we cannot distinguish students taking regular from alternate assessments; these counts were combined in the reported data. Therefore, for consistency in all years, we use all performance data reported in ED*Facts*, including results of students taking both regular and alternate assessments. The raw data include no suppressed cells, nor do they have a minimum cell size for reporting.

Each row of data corresponds to a school-subgroup-subject-grade-year cell. The raw data include no suppressed cells, nor do they have a minimum cell size for reporting. Table 3 illustrates the structure of the raw data from ED*Facts* prior to use in constructing SEDA 3.0.

II.B. Definitions

Commuting Zone (CZ): Regions defined by the geographic boundaries of a local economy. We use the 2000 boundary definitions (https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas/), which are the most recent commuting zone definitions.

Geographic School District (GSD): The aggregate of all public schools, regardless of type and administrative control, residing in a geographic catchment area defined by a traditional public school district. GSDs allow linking of achievement data to demographic and economic information from EDGE/ACS, which is reported for students living in GSD boundaries regardless of where they attend school.

Group: A subgroup-unit (as defined below). For schools, the only available subgroup is all students. For GSDs, counties, CZs, and MSAs, data for subgroups are available when estimates are sufficiently precise.

Metropolitan Statistical Area (metro): A county or group of counties with a population exceeding 50,000 and encompassing an urban area, combined with any surrounding counties with strong commuting ties to the urban area (https://www.census.gov/programs-surveys/metro-micro/about/glossary.html). The U.S. Census Bureau revises the metropolitan statistical area definitions after each decennial census. We use the 2013 U.S. Census Bureau definitions, which are the definitions based on the 2010 census (https://www.census.gov/programs-surveys/metro-micro/geographies/geographic-reference-files.2013.html). We make one modification to the definitions: The Census defines very large metropolitan areas as Consolidated Metropolitan Statistical Areas (CMSAS); each CMSA is subdivided into Metropolitan Area Divisions. We treat each Division as a separate metropolitan area for analysis purposes, as the CMSAs generally quite large.

Subgroup: The term “subgroup” refers to the group of students to which an estimate pertains. This may be: all, white, black, Hispanic, Asian, male, female, economically disadvantaged, or not economically disadvantaged students.

Unit: The term “unit” refers to the aggregation of the data. This may be a school, GSD, county, CZ, or metro.

II.C. Construction Overview

The construction process produces mean test score estimates for schools, GSDs, counties, CZs and metros on two nationally comparable scales in a series of ten steps, outlined in Figure 1. We provide a brief conceptual description of each step here. We then provide substantial description and technical details about each step in Section II.D.

Step 1: Creating the Crosswalk. This step assigns each public school district to a GSD and links each GSD uniquely to a county, CZ, and metro.

Step 2: Data Cleaning. This step removes data for states and units in particular subjects, grades, and years for which we cannot produce any estimates. We also remove any identified errors in the raw data here.

Step 3: Estimating and Linking Cutscores. This step uses Heteroskedastic Ordered Probit (HETOP) models to estimate the state-grade-subject-year cutscores from the GSD proficiency count data for all students. It links the estimated cutscores to the NAEP scale and then standardizes the linked cutscores to the Cohort Standardized (CS) scale. The resulting cutscores are comparable across states and years.

Step 4: Exclude and Prepare Data. This step excludes data for *unit-subgroup-subject-grade-year* cases with low participation in the assessment or high percentages of students taking alternate assessments.

Step 5: Estimating School and District Means. This step uses the pooled HETOP model to estimate school and GSD subgroup-subject-grade-year means and standard deviations, along with their standard errors, based on the cutscores from Step 3 and the data prepared in Step 4.

Step 6: Aggregating to County, CZ, and MSA Means. This step aggregates the GSD-subgroup estimates from Step 5 to counties, CZs, and metros. From this point onward, we have test score estimates for five units: schools, GSDs, counties, CZs, and metros. Subsequent steps are equivalent for all units unless otherwise noted.

Step 7: Scaling Across Grades. This step creates grade cohort standardized (GCS) estimates for all units. From this point onward, we have two scales of the data for all units: CS and GCS. Subsequent steps are equivalent for both scales unless otherwise noted.

Step 8: Calculating Achievement Gaps. This step estimates white-black, white-Hispanic, white-Asian, male-female, and nonpoor-poor achievement gaps for GSDs, counties, CZs, and metros in each subject-grade-year where there is sufficient data.

Step 9: Pooling Mean and Gap Estimates. This step estimates the average achievement, learning rate, and trend in test scores by subject and overall for each unit and scale. From this point onward, we have three levels of the data for all units: long (not pooled by grade, year, or subject), pooled by subject (poolsub), and pooled overall (pool).

Step 10: Suppressing Data for Release. The step suppresses estimates that are too imprecise to be useful or do not reflect the performance of at least 20 unique students in both long and pooled files for all units and scales. For estimates reported in the long files, this step also adds a small amount of random noise to meet the reporting requirements of the US Department of Education.

II.D. Detailed Construction Overview

Notation

In the remainder of the documentation, we use the following mathematical notation:

- Mean estimates are denoted by $\hat{\mu}$ and standard deviation estimates by $\hat{\sigma}$.
- The cutscore estimates are denoted as $\hat{c}_1, \dots, \hat{c}_K$. There are K total cutscores in each state-subject-grade-year.
- A subscript indicates the aggregation of the estimate. We use the following subscripts:
  - u = unit (generic)
    - n = school
    - d = GSD
    - c = county
    - z = CZ
    - m = metro
    - f = state
  - r = subgroup
    - all = all students
    - wht = white
    - blk = black
    - hsp = Hispanic
    - asn = Asian
    - mal = male
    - fem = female
    - ecd = economically disadvantaged
    - nec = not economically disadvantaged
    - wbg = white-black gap
    - whg = white-Hispanic gap
    - mfg = male-female gap
    - neg =not economically disadvantaged-economically disadvantaged gap
  - y = year
  - b = subject
  - g = grade
- A superscript indicates the scale of the estimate. The metric is generically designated as $\boldsymbol{x}$**.** There are four scales. The first two scales are only used in construction. The latter two scales are reported:
  - $state$ = state-referenced metric
  - $naep$ = NAEP test score scale metric
  - cs = cohort scale metric
  - gcs = grade within cohort scale metric

Step 1. Creating the Crosswalk & Defining Geographic School Districts

The primary purpose of the crosswalk is to assign schools to GSDs. Each traditional public school district in the U.S. is defined by a geographic catchment area; the schools that fall within this geographic boundary make up the GSD. Commonly, public school districts have administrative control over the traditional public schools that fall within their specific geographic boundaries. However, there may be some schools physically located within the geographic boundary of a school district that are not under its administrative control. For example, there may be charter schools located within the boundaries of a traditional public school district that are operated by a charter school network (which has no associated geographic boundary). Any school that is not affiliated with one of the traditional public school districts is assigned to a GSD based on its geographic location; the assigned GSD will be the traditional public school district in whose geographic boundaries the school is physically located. The GSD, therefore, contains all of the public school students living within the geographic boundaries of the school district. The motivation for this assignment is to better align the test scores for students living within school district boundaries with the demographic and socioeconomic data that we retrieve from other sources that report data by geographic school district boundaries.

Below are the GSD-assignment rules for common types of schools that are operated by a local education agency (LEA) without a straightforward geographic boundary.

Charter schools: If a charter school is operated by an administrative district that only has charter schools or is authorized by a state-wide administrative agency, it is geolocated and assigned to a GSD based on its location.[4] If a charter school is operated by a traditional public school district, we use that as its GSD regardless of the school's location.

Schools operated by high school districts: In the cases where schools in high school districts serve students in grades 7 and 8, the high schools are assigned to the elementary school district in which they are geographically located.

[4] Geographic location is determined by the latitude and longitude coordinates of a school's physical address as listed in the CCD. The GSD of charter schools sometimes varies from year to year for approximately 5.45% of the roughly 8,612 charter schools. In these cases, we use the GSD the charter is assigned to in the most recent year it is observed. 18 charter schools cannot be geolocated using the provided latitude/longitude information. All such schools are assigned to a single GSD with no geographic boundary.

Virtual schools: By their nature, most virtual schools do not draw students from within strict geographic boundaries. We therefore assign all the virtual schools within a state to a single "virtual school district". We identify schools as virtual using CCD data from 2013-14 through 2015-16 CCD data. The virtual school identifier did not exist in earlier years of data, so we flag schools as virtual in all years of our data if they are identified as virtual by the later year CCD indicators.[5] Additionally, we identify virtual schools by searching school names for terms such as "virtual", "cyber", "online", "internet", "distance", "extending", "extended", "on-line", "digital" and "kaplan academy". Since schools may change names, if we identify a school as virtual by this approach in one year, we flag the school as virtual in all years.[6] Note that virtual schools are retained in the estimation of state cutscores, but no mean estimates are produced or reported in SEDA 3.0 for virtual schools or virtual school districts (these are removed from the data in **Step 4**).

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

The second purpose of the crosswalk is to identify a stable district ID for cases where school districts restructure or are reported differently in different data sets during the time period of our data. These cases are discussed below.

Schools in districts that restructure: Some districts changed structure during the time period covered by SEDA 3.0 data. We have identified a small number of these cases. In California, two Santa Barbara 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 Tennessee, Memphis Public

[5] In 2013-2015, we identified 12 non-virtual schools in Alabama identified as "virtual" by the CCD indicator. We treat these as regular schools in all subsequent steps.

[6] 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 GSD in which they are located.

Schools and Shelby County Public Schools (LEA IDs: 4702940, 4703810) merged. In Texas, North Forest ISD merged with Houston ISD (LEA IDs: 4833060, 482364). For all cases, SEDA 3.0 contains estimated test score distributions for the combined GSDs.

Schools in New York City: The CCD assigns schools in New York City to one of thirty-two districts or one "special schools district." We aggregate all New York City Schools to the city level and give them all the same GSD code, creating one unified New York City GSD code.

Finally, the crosswalk links the GSD estimates to counties, CZs, and metros. No additional geolocation is done in support of this aspect of the crosswalk. GSDs are assigned to counties, metros, and CZs based on the county codes provided in CCD. A small number of counties restructure during the time frame of our data, meaning that we observe some districts belonging to two different counties over the course of our data. To avoid this issue, we create a stable ID for this county that is equivalent to the county definition in the most recent year of data. Districts are always assigned to this stable county ID, regardless of the year of the data. We use the 2013 metropolitan statistical area definitions.

The crosswalk and the shape files used to locate schools within each geographic unit are available in the SEDA database. The county, metro, and CZ 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. Note that in the data repository the shape files are labeled as "v21". No updates were made to these files in this release; their version number was not edited.

Step 2. Data Cleaning

In this step, we first merge the ED*Facts* data (described under **II.A. Source Data**, above) by NCES school ID and year with the crosswalk developed in **Step 1**. This merge provides us with counts of students scoring in each proficiency category by school-subgroup-subject-grade-year that is linked to GSDs, counties, CZs, and metros. As noted above, in 2008-09 through 2011-12, we cannot distinguish students taking regular from alternate assessments; these counts were combined in the reported data. Therefore, for consistency in all years, we combine the performance data for regular and alternate assessments as reported in ED*Facts*. Notably, in a small number of cases that the state's alternate assessments have one additional performance category relative to the regular assessment.[7] Because our estimation uses combined counts of students scoring in each performance category across all assessments, this leads to the bottom or top proficiency category of the data having a very small number of observations. To avoid issues during estimation, we collapse the sparse bottom or top category with the adjacent category in these state-subject-grade-year cases. The affected state, subject, grade, and year cases include: Arkansas, math and ELA, grades 3-8, years 2012, 2013, 2014 and 2016; Colorado, math and ELA, grades 3-8, years 2012, 2013, and 2014; Iowa, math and ELA, grades 3 through 8, years 2015 and 2016; New York, math, grades 3-6, years 2013 and 2014; Oregon, math and ELA, grades 3-8 in 2013 and 2014; and South Carolina, math and ELA, grades 3-8, years 2012, 2013, and 2014.

Next, we remove all data[8] for *state-subject-grade-year* cases that do not meet the requirements of our estimation. A general description of these cases follows, and a list of specific cases can be found in Table 4:

Students took incomparable tests within the state-subject-grade-year: There are two common ways this appears within the data. First, there are cases where districts were permitted to administer locally selected assessments. This occurred in Nebraska during SY 2008-2009 (ELA and Math) and SY 2009-2010 (Math). Second, students take end-of-

[7] The ED*Facts* documentation notes these discrepancies in years after 2011-12.

[8] For all subgroups and all schools in the state. In other words, no estimates will be available for these state-subject-grade-year cases.

course rather than end-of-grade assessments. This is the case in some or all years for 7th and 8th grade math for California, Virginia and Texas (among other states, reported in Table 5). The problem is that assessments were scored on different scales and using different cut scores. Therefore, proficiency counts cannot be compared across districts or schools within these state-subject-grade-year cases.

The state had participation lower than 95% in the tested subject-grade-year: Using the ED*Facts* data, we are able to estimate a participation rate for all state-subject-grade-year cases in the 2012-13 through 2014-15 school years. This participation data file is not available prior to the 2012-13 school year, and therefore we cannot calculate participation rates prior to 2012-13. Participation is the ratio of the number of test scores reported to the number enrolled students in a given state-subject-grade-year:

$$\widehat{part}_{fygb} = \frac{numscores_{fygb}}{numenrl_{fygb}} \tag{2.1}$$

for each state f, year y, grade g, and subject b.

This state-level suppression is important because both the quality of the estimates and the linkage process depends on having the population of student test scores for that state-subject-grade-year. State participation may be low due to a number of factors, including student opt out or pilot testing. Note that we do not suppress any entire state-subject-grade-year cases prior to the 2012-13 school year as enrollment data are not available in ED*Facts*. However, opt out was low in 2012-13 (no state was excluded based on this threshold), which suggests states met 95% threshold in prior years when data are not available.

Insufficient data was reported to ED*Facts*: Some states reported no data in certain years: Wyoming did not report any assessment outcomes in 2009-10. Others reported data from which we cannot recover reliable estimates. In the 2008-09, 2009-10, and 2010-11 school years, Colorado reported data in only two proficiency categories, and a large majority of the data (88% across subjects, grades, and years) fall into a single category. These data do not provide sufficient information to estimate means and/or standard deviations in most regions. In the 2014-15 and 2015-16 school years, New Mexico reported data in on two proficiency categories. We remove these cases because the two years are consecutive and fall at the end of the time series of our data.

In addition to the exclusion of *state-subject-grade-year* cases, we also remove idiosyncratic data errors. These were identified by looking at the distribution of students across proficiency categories. When the distribution changed too abruptly for the given cohort in the given year compared with their performance in the prior and subsequent years, as well as compared with other cohorts in the GSD, these data were determined to be entry errors and were removed. These cases are listed in Table 5.

Step 3. Cutscore Estimation and Linking

In this step, we use HETOP models and the all-student GSD proficiency count data to estimate state-subject-grade-year cutscores on a common scale linked to NAEP. To address practical challenges that can arise in linking and the HETOP estimation framework, within a specific state-subject-grade-year we:

Rearrange GSDs. We reconfigure GSDs that meet certain criteria within a state-subject-grade-year in order to improve the HETOP estimation process. First, we combine vectors of counts that have fewer than 20 students into “overflow” groups because estimates based on small sample sizes can be inaccurate. Second, in some vectors with more than 20 students the pattern of counts does not provide enough information to estimate a mean or a standard deviation; we also place these count vectors into the “overflow” group. If the resulting overflow groups have parameters that cannot be estimated via maximum likelihood, they are removed from the data. This reconfiguration allows us to retain the maximum possible number of test scores in the estimation sample for the cutscores. This is important as the linking methods we use later in this step rely on having information about the full population in each state-grade-year-subject.

Constrain GSDs. For groups not in the “overflow” group, we always estimate a unique mean. But we can sometimes obtain more precise and identifiable estimates by placing additional constraints on group standard deviation parameters in the HETOP model. We constrain standard deviation parameter estimates for groups that meet the following conditions during estimation:

- There are fewer than 50 student assessment outcomes in a GSD.
- There are not sufficient data to estimate both a mean and standard deviation (all student assessment outcomes fall in only two adjacent performance level categories; all student assessment outcomes fall in the top and bottom performance categories; or all student assessment outcomes fall in a single performance level category).

After these data processing steps, we estimate a separate HETOP model for each state-subject-grade-year and save the cutscore estimates. For state-grade-year-subjects with only two proficiency categories, we cannot estimate unique GSD standard deviations and instead we use the model with a single, fixed standard deviation parameter (the HOMOP model). We denote the estimated cutscores as $\widehat{c}_{1fygb}^{state}, \ldots, \widehat{c_{K-1}}_{fygb}^{state}$, for a state f, year y, grade g, and subject b, where the proficiency data are reported in K categories. These cutscores are expressed in units of their respective state-year-grade-subject student-level standardized distribution. The HETOP model estimation procedure also provides standard errors of these cutscore estimates, denoted $se\left(\widehat{c}_{k fygb}^{state}\right)\ for\ k = 1, .., K-1$, respectively (Reardon, Shear, Castellano, & Ho, 2017). Note that we do not use the group-specific means or standard deviations that are simultaneously estimated along with the cutscores; mean estimation is described in **Steps 5** and **6**. See Reardon et al. (2017) and the description in **Step 5** below for additional details about the HETOP model.

To place these cutscores on a common scale across states, grades, and years we use data from the National Assessment of Educational Progress (NAEP). NAEP data provide estimates of 4th and 8th grade test score means and standard deviations for each state on a common scale, denoted $\hat{\mu}_{fygb}^{naep}$ and $\hat{\sigma}_{fygb}^{naep}$, respectively, as well as their standard errors.[9] Because NAEP is administered only in 4th and 8th grades in odd-numbered years, we interpolate and extrapolate linearly to obtain estimates of these parameters for grades (3, 5, 6, and 7) and years (2010, 2012, 2014, and 2016) in which NAEP was not administered. First, within each NAEP-tested year (2009, 2011, 2013, 2015, and 2017) we linearly interpolate between grades 4 and 8 to grades 5, 6, and 7 and extrapolate to grade 3. Next, for all grades 3-8, we linearly interpolate between the odd NAEP-tested years to estimate parameters in 2010, 2012, 2014 and 2016, using the interpolation/extrapolation formulas here:

$$\begin{aligned} \hat{\mu}_{fygb}^{naep} &= \hat{\mu}_{fy4b}^{naep} + \frac{g-4}{4}\left(\hat{\mu}_{fy8b}^{naep} - \hat{\mu}_{fy4b}^{naep}\right), \quad \text{for g} \in \{3, 5, 6, 7\} \\ \hat{\mu}_{fygb}^{naep} &= \frac{1}{2}\left(\hat{\mu}_{f[y-1]gb}^{naep} + \hat{\mu}_{f[y+1]gb}^{naep}\right), \quad \text{for y} \in \{2010, 2012, 2014, 2016\} \end{aligned} \tag{3.1}$$

[9] Note that the NAEP scales are not comparable across math and reading, but they are comparable across states, grades and years within each subject.

We do the same to interpolate/extrapolate the state NAEP standard deviations. The reported NAEP means and standard deviations, along with interpolated values, by year and grade, are reported in Table 6.

We then use these state-specific NAEP estimates to place each state's cutscores on the NAEP scale. The methods we use—as well as a set of empirical analyses demonstrating the validity of this approach—are described in more detail by Reardon, Kalogrides, and Ho (Forthcoming). We provide a brief summary here. Because GSD test score moments and the cutscores are expressed on a state scale with mean 0 and unit variance, the estimated mapping of $\widehat{c}_{k_{fygb}}^{state} for\ k = 1, \ldots, K-1$ to the NAEP scale is given by Equation (3.2) below, where $\hat{\rho}_{fygb}^{\text{state}}$ is the estimated reliability of the state test. This mapping yields an estimate of the k^{th} cutscore on the NAEP scale; denoted $\widehat{c}_{k_{fygb}}^{naep}$.

$$\widehat{c}_{k_{fygb}}^{naep} = \hat{\mu}_{fygb}^{naep} + \frac{\widehat{c}_{k_{fygb}}^{state}}{\sqrt{\hat{\rho}_{fygb}^{state}}} \cdot \hat{\sigma}_{fygb}^{naep} \tag{3.2}$$

The intuition behind Equation (3.2) is straightforward: cutscores in states with relatively high NAEP averages should be placed higher on the NAEP scale. The reliability term, $\hat{\rho}_{fygb}^{\text{state}}$, in Equation (3.2) is necessary to account for measurement error in state accountability test scores. Note that cutscores on the state scale are expressed in terms of standard deviation units of the state score distribution. The state scale cutscores are biased toward zero due to measurement error. They must be disattenuated during mapping to the NAEP scale, given that the NAEP scale accounts for measurement error due to item sampling. We disattenuate the means by dividing them by the square root of the state test score reliability estimate, $\hat{\rho}_{fygb}^{\text{state}}$. The reliability data used to disattenuate the estimates come from Reardon and Ho (2015) and were supplemented with publicly available information from state technical reports. For cases where no information was available, test reliabilities were imputed using data from other grades and years in the same state.

Finally, we standardize the NAEP-linked cutscores relative to a reference cohort of students. This standardization is accomplished by subtracting the national grade-subject-specific mean and dividing by the national grade-subject-specific standard deviation for a reference cohort. We use the average of the three national cohorts that were in 4th grade in 2009, 2011, and 2013. We rescale at this step such that all means recovered in Step 5 will be interpretable as an effect size relative to the average of the three national cohorts that were in 4th grade in 2009, 2011, and 2013.

For each grade, year and subject we calculate:

$$\begin{aligned} \hat{\mu}^{naep}_{avg,gb} &= \sum_{Y\in\{2005,2007,2009\}} \frac{1}{3}\mu^{naep}_{(y=Y+g)gb} \\ \hat{\sigma}^{naep}_{avg,gb} &= \sum_{Y\in\{2005,2007,2009\}} \frac{1}{3}\sigma^{naep}_{(y=Y+g)gb} \end{aligned} \tag{3.3}$$

In Equation (3.3), Y refers to the year in which the cohort was in the spring of kindergarten. For the 2009 4th grade cohort, this is equal to 2005 (or 2009 minus 4).

Then we standardize each cutscore:

$$\widehat{c_k}^{cs}_{fygb} = \frac{\widehat{c_k}^{naep}_{fygb} - \hat{\mu}^{naep}_{avg,gb}}{\hat{\sigma}^{naep}_{avg,gb}} \tag{3.4}$$

The resulting cutscores are on the CS scale, standardized to this nationally averaged reference cohort within subject, grade, and year.

Step 4. Selecting Data for Mean Estimation

In Step 5, we estimate a model separately for each unit-subgroup that draws only on the subject-grade-year data for that unit-subgroup. In some subjects, grades, and years, we are less confident in the quality of the unit-subgroup data and do not want leverage it in estimation as it may bias the parameter estimates.[10] These cases are described below:

The participation rate is less than 95%. In these cases, the population of tested students on which the mean and standard deviation estimates are based may not be representative of the population of students in that school). Therefore, we remove all unit-subgroup-subject-grade-year cases where participation was lower than 95%. Participation is defined as:

$$\widehat{part}_{urygb} = \frac{numscores_{urygb}}{numenrl_{urygb}}. \tag{4.1}$$

This measure can be constructed in the 2012-13 through 2015-16 school years; we do not remove data based on this rule in earlier years. If the participation rate for "all students" is less than 95%, we do not report any estimates for demographic subgroups regardless of whether the subgroup-specific participation rate was greater than 95% because we are concerned about data quality.

Insufficient data reported by student demographic subgroups. There are a small number of cases where the total number of test scores reported by race or gender is less than 95% of the total reported test scores for all students. For example, there may be 50 test scores reported for all students, but only 20 test scores for male students and 20 test scores for female students. In this case, we would not report the male or female test score means because insufficient test scores were reported by gender. We calculate the reported percentage as:

[10] This logic of this data selection differs from the cleaning done in Step 2 to support cutscore estimation. For the cutscore estimation, we wanted to keep as much data as possible in the estimation process because the linking procedure at the end of the Step 3 requires population-based data. Moreover, the cutscore are not particularly sensitive to low-quality data for individual GSDs. In contrast, the school/GSD estimates will be strongly affected by low quality data (due to the factors described above). First, those parameters may not accurate reflect the academic performance in the unit. Second, in the model that we use (described more below), we "borrow" information across grades and years in some cases. If we include these low-quality data cases, we may be borrowing from "bad" information.

$$\widehat{rep}_{urygb} = \frac{\sum_r numscores_{urygb}}{numscores_{u,all,ygb}}. \tag{4.2}$$

This measure can be constructed in all years.

More than 40% of students take alternate assessments. We are concerned that we are getting a biased estimate in unit-subgroup-subject-grade-year cases where over 40% of the students take alternate assessments. These assessments typically differ from the regular assessment and have different proficiency thresholds. This flag can be constructed in the 2012-13 through 2015-16 school years; we do not remove data based on this rule in earlier years.

Students scored only in the top or only in the bottom proficiency category. We cannot obtain maximum likelihood estimates of unique means for these cases and therefore remove them prior to estimation. This flag can be constructed in every year.

We next flag and remove schools-subgroups and GSD-subgroups that do not meet the minimum estimation requirements, described below. First, we create a "type flag" for each unit-subgroup-subject-grade-year case. It is considered "deficient" if the case meets one of the following conditions: a) has all observations in a single category; b) has all observations in only 2 adjacent categories; c) has only 2 proficiency categories (one cut score); or, d) has all observations in only the top and bottom categories (e.g., X-0-0-X or X-0-X). Otherwise, cases are flagged as "sufficient". Constraints on the parameter estimates for "deficient" cases are needed during estimation because they do not provide sufficient data to freely estimate both a mean and a standard deviation. Second, we construct a "size flag." We flag unit-subgroup-subject-grade-year cases as "small" if they have fewer than 100 test scores; otherwise, cases are flagged as "large". Each unit-subgroup-subject-grade-year, then, has two associated flags. These flags will be used again during estimation to place constraints on the standard deviation estimates for individual unit-subgroup-subject-grade-year cases. If a unit-subgroup has only one "deficient" or "small" unit-subgroup-subject-grade-year case, then that case is dropped from the data. We also drop entire unit-subgroups that have no "sufficient" unit-subgroup-subject-grade-year cases.

Our estimation methods, described in the next step, cannot produce a standard deviation estimate when all subject-grade-year cases for a given unit when these conditions are met.

Finally, we select not to perform the mean estimation for a subset of whole schools and GSDs (across all subgroups, subjects, grades and years). These include: (1) virtual schools and GSDs (described in Step 2); (2) charter schools that could not be geolocated; and (3) schools and GSDs with more than 20% of all students taking alternate assessments. Note that while we technically perform this data selection only for schools and GSDs in this Step, we apply a subset of these rules to counties, CZs, and metros during the aggregation process. Table 7 shows the cases that are excluded based on these rules for all geographies.

Step 5. Estimating Means for Schools and Districts

The goal of this step is to estimate the mean and standard deviation of test scores for each subgroup in each unit (school or district) across subjects, grades, and years. We have two pieces of information that we use for this process: the observed proficiency counts for each subgroup-unit-state-grade-year-subject from **Step 4** and the estimated cutscores separating the proficiency categories in the associated state-grade-year-subject from **Step 3**. We use these data and a pooled HETOP model (Shear and Reardon, 2019; Reardon et al., 2017) to estimate μ_{urygb}^{cs} and σ_{urygb}^{cs}, the mean and standard deviation of achievement on the CS scale for unit u (school or GSD), subgroup r, year y, grade g, and subject b. As described below, the pooled HETOP model is run separately for each unit-subgroup-subject, but combines data across grades and years when estimating these parameters. Combining data across grades and years allows us to get better estimates of σ_{urygb}^{cs} for years and grades in which sample sizes are small or where the proficiency count data are limited.

We use a pooled HETOP model in order to overcome three practical challenges. The challenges are: 1) in some states, years, and grades, $K = 2$ and there is not sufficient information to estimate both a mean and a standard deviation for each unit-subgroup-grade-year-subject; 2) if $K \geq 3$ but there are sampling zeros because test scores were not observed in all K categories for a particular grade and year, there may not be sufficient information to estimate both a mean and a standard deviation; and 3) when the sample size n_{kurygb} is small, prior simulations (e.g., Reardon et al., 2017; Shear & Reardon, 2019) have shown that estimates of standard deviations can be biased and contain excessive sampling error.

We estimate a pooled HETOP model (Shear & Reardon, 2019) for each unit, separately for each subject and subgroup, by “pooling” data across all available grades and years, and maximizing the joint log likelihood function given by:

$$L = \ln\left[P\left(\mathbf{N}_{urb} \middle| \mathbf{M}_{urb}^{cs}, \mathbf{H}_{urb}^{cs}, \mathbf{C}_{fb}^{cs}\right)\right] = \sum_{y=1}^{Y}\sum_{g=1}^{G}\sum_{k=1}^{K} n_{kurygb} \ln\left(\pi_{kurygb}\right)$$

$$= \sum_{y=1}^{Y}\sum_{g=1}^{G}\sum_{k=1}^{K_{gy}} n_{kurygb} \ln\left(\Phi\left(\frac{\mu_{urygb}^{cs} - c_{k-1_{fygb}}^{cs}}{\exp\left(h_{urb}(g,y)\right)}\right) - \Phi\left(\frac{\mu_{urygb}^{cs} - c_{k_{fygb}}^{cs}}{\exp\left(h_{urb}(g,y)\right)}\right)\right),$$

where $\mathbf{N}_{urb}$ is a matrix of proficiency counts across all available grades (G) and years (Y) for unit u, subgroup r and subject b, $\mathbf{M}_{urb}^{cs}$ is a vector of estimated means across grades and years, $\mathbf{H}_{urb}^{cs}$ is a vector of estimated parameters for the function $h(\quad)$ described below, and $\mathbf{C}_{fb}^{cs}$ is a matrix of cutscores across grades and years. The cutscores are treated as fixed here, using the values estimated in Step 3. We have replaced σ_{urygb}^{cs} in the above equation with $\exp\left(h_{urb}(g,y)\right)$, where $h_{urb}(g,y)$ is a unit-specific function used to model the natural log of the standard deviations as a function of grade and year:

$$h_{urb}(g,y) = \ln\left(\sigma_{urygb}^{cs}\right) = \gamma_{urygb}^{cs}.$$

We do this for two reasons. First, estimating $\gamma_{urygb}^{cs} = \ln\left(\sigma_{urygb}^{cs}\right)$ rather than σ_{urygb}^{cs} directly ensures that the ML estimate will be positive. Second, defining γ_{urygb}^{cs} as a function of grade and year, rather than allowing this value to be unique in each grade and year, defines the pooled HETOP model. If we place no constraints on the model and allow $h_{urb}(g,y) = \gamma_{urbgy}$ to take on a unique value in each grade and year, maximization of this likelihood will result in identical estimates to those obtained by maximizing the likelihood separately for each grade and year.

To leverage the data available across multiple grades and years and overcome the limitations noted above, we define $h_{urb}(g,y)$ in the following way. First, we allow γ_{urygb} to be freely estimated in each grade-year cell that is both “sufficient” and “large”, by the flags defined above. For all other grade-year cells, we constrain $h_{urb}(g,y)$ such that the estimate of γ_{urygb} is equal to the mean of the $\hat{\gamma}_{urygb}$ estimates across the freely estimated cells. That is, we estimate a common “pooled” standard deviation across the grades and years in which there are “deficient” data and/or “small” cell sizes.

More formally, for all years and grades in which $n_{urygb} < 100$ and/or in which there are insufficient data to estimate both a mean and a standard deviation, we constrain $h_{urb}(g,y) =$

γ_{urb}^{cs} to be equal, while allowing $h_{urb}(g, y) = \gamma_{urygb}^{cs}$ to be freely estimated in grades and years with at least 100 test scores and sufficient data to estimate both a mean and standard deviation. We further constrain the model such that the "pooled" log standard deviation is equal to the (unweighted) mean of the unconstrained log standard deviations by defining the constraint:

$$\gamma_{urb}^{cs} = \frac{\sum_{g=1}^{G}\sum_{y=1}^{Y}\left(I_{urygb}^{100} \cdot I_{urygb}^{S} \cdot \gamma_{urygb}^{cs}\right)}{\sum_{g=1}^{G}\sum_{y=1}^{Y}\left(I_{urygb}^{100} \cdot I_{urygb}^{S}\right)},$$

where I_{urygb}^{100} is the size indicator flag (equal to 1 if size is "large") and $I_{urygb}^{S}$is the sufficient data indicator flag (equal to 1 if there are sufficient data). If I_{urygb}^{100} and I_{urygb}^{S} are equal to 1 for all cells in a unit, then we estimate a unique mean and standard deviation for each cell. For all other units, there will be a mix of freely estimated and constrained standard deviation parameters. Recall in Step 4 that we removed unit-subgroups where $I_{urygb}^{S} = 0$ for all cells because we are unable to estimate a standard deviation parameter.

Summary

The models described here are used to produce ML estimates of μ_{urygb}^{cs} and σ_{urygb}^{cs} (where $\hat{\sigma}_{urygb}^{cs}$ may be constrained to be equal in some grades and years), as well as estimated standard errors $se\left(\hat{\mu}_{urygb}^{cs}\right)$ and $se\left(\hat{\sigma}_{urygb}^{cs}\right)$ and the estimated sampling covariances $cov\left(\hat{\mu}_{urygb}^{cs}, \hat{\sigma}_{urygb}^{cs}\right)$, where unit can be either a GSD d, or a school n. This process is applied separately for each district-subgroup-subject or school-subgroup-subject within each state. The estimates are on the CS scale described elsewhere, and can be transformed to other scales, such as the GCS scale.

Step 6. Aggregating GSD-subgroup estimates to Counties, CZs and Metros

We adopt a different approach to estimate the mean and standard deviation of achievement in counties, CZs and MSAs in a given year y, grade g, and subject b. We use the estimates for the GSDs from **Step 5** that correspond to a given county, CZ or metro within a subject-grade-year to estimate an overall mean and variance for that unit. As noted above, we use stable county identifiers in cases where we observe that a district is placed in multiple counties during the years in our sample. The district is assigned to the county it is observed in during the 2015-16 school year (the last year of our data).

We describe the process here for counties, but it also applies to CZs and MSAs. Suppose there are a set of C counties, each of which contains one or more unique GSDs. These higher-level units are defined geographically and are non-overlapping. Hence, each GSD falls within exactly one county. The county mean is estimated as the weighted average of GSD means across all D_c GSDs in county c, computed as

$$\hat{\mu}^{cs}_{crygb} = \sum_{d=1}^{D_c} p_{dc}\hat{\mu}^{cs}_{drygb}, \tag{6.1}$$

where p_{dc} is the proportion of county c represented by GSD d. The estimated county standard deviation is estimated as the square root of the estimated total variance between and within GSDs within a county,

$$\hat{\sigma}^{cs}_{crygb} = \sqrt{\hat{\sigma}^2_{B_c} + \hat{\sigma}^2_{W_c}} \tag{6.2}$$

where $\hat{\sigma}^2_{B_c}$ is the estimated variance between GSDs in county c and $\hat{\sigma}^2_{W_c}$ is the estimated variance within GSDs in county c. The formulas used to estimate $\hat{\sigma}^2_{B_c}$ and $\hat{\sigma}^2_{W_c}$ are based on equations in Reardon et al. (2017). These formulas and formulas for estimating the standard errors of the county means and standard deviations, $\hat{\mu}^{cs}_{crygb}$ and $\hat{\sigma}^{cs}_{crygb}$, are included in Appendix A1.

Step 7. Scaling the Estimates

As described in **Step 3**, we standardize the cutscores prior to estimation such that all mean estimates are produced on the CS scale. In the step, we establish a second scale: The **Grade Cohort Standardized (GCS) scale.** We recommend CS-scaled estimates for research purposes and the GCS scale for low-stakes reporting to non-research audiences.

Recall that the CS scale is standardized within subject and grade, relative to the average of the three cohorts in our data who were in 4th grade in 2009, 2011 and 2013. We use the average of three cohorts as our reference group because they provide a stable baseline for comparison. This metric is interpretable as an effect size, relative to the grade-specific standard deviation of student-level scores in this common, average cohort. For example, a GSD with a mean of 0.5 on the CS scale represents a GSD where the average student scored approximately one half of a standard deviation higher than the national reference cohort scored in that same grade. GSD means reported on the CS scale have an overall average near 0 as expected. Note that this scale retains information about absolute changes over time by relying on the stability of the NAEP scale over time. This scale does not enable absolute comparisons across grades, however.

The GCS scale standardizes the unit means relative to the average difference in NAEP scores between students one grade level apart. The average grade-level difference in national NAEP scores is estimated as the within-cohort grade-level change (separately by subject b), for the average of three cohorts of students in 4th grade in 2009, 2011, and 2013 (see detail on how $\hat{\mu}^{\text{naep}}_{avg,\text{g}b}$ and $\hat{\sigma}^{naep}_{avg,gb}$ are calculated in **Step 3**). It is denoted $\hat{\gamma}_{avg,b}$:

$$\hat{\gamma}_{\text{avg},b} = \frac{\hat{\mu}^{\text{naep}}_{avg,8b} - \hat{\mu}^{\text{naep}}_{avg,4b}}{4} \tag{7.1}$$

We then identify the linear transformation that sets the grade 4 and 8 averages for this cohort at the "grade level" values of 4 and 8 respectively. Then transform unit means, standard deviations, and their variances accordingly:

$$\hat{\mu}^{gcs}_{urygb} = 4 + \frac{\hat{\mu}^{\text{naep}}_{avg,\text{g}b} - \hat{\mu}^{\text{naep}}_{avg,4b}}{\gamma_{avg,b}} + \frac{\hat{\sigma}^{naep}_{avg,gb}}{\gamma_{avg,b}} \mu^{cs}_{urygb} \tag{7.2}$$

$$\hat{\sigma}^{gcs}_{urygb} = \frac{\hat{\sigma}^{naep}_{avg,gb}}{\gamma_{avg,b}}\hat{\sigma}^{cs}_{urygb}$$

$$var\left(\hat{\mu}^{gcs}_{dygb}\right) = \left(\frac{4\sigma_{gb}}{\mu_{8b} - \mu_{4b}}\right)^2 var\left(\hat{\mu}^{cs}_{dygb}\right) = \left(\frac{\sigma_{gb}}{\gamma\gamma_b}\right)^2 var\left(\hat{\mu}^{cs}_{dygb}\right)$$

$$var\left(\hat{\sigma}^{gcs}_{dygb}\right) = \left(\frac{4\sigma_{gb}}{\mu_{8b} - \mu_{4b}}\right)^2 var\left(\hat{\sigma}^{cs}_{dygb}\right) = \left(\frac{\sigma_{gb}}{\gamma_b}\right)^2 var\left(\hat{\sigma}^{cs}_{dygb}\right)$$

Then, $\hat{\mu}^{gcs}_{urygb}$ can be interpreted as the estimated average national "grade-level performance" of students in unit u, subgroup r, year y, grade g, and subject b. For example, if $\hat{\mu}^{gcs}_{ury4b} = 5$, 4th-grade students in unit u, subgroup r, and year y are one grade level ($\hat{\gamma}_{2009b}$) above the 4th grade 2009-2013 national average ($\hat{\mu}^{naep}_{avg,4b}$) in performance on the tested subject b.

GSD means reported on the GCS scale have an overall average near 5.5 (midway between grades 3 and 8) as expected. This metric enables absolute comparisons across grades and over time, but it does so by relying not only on the fact that the NAEP scale is stable over time but also that it is vertically linked across grades 4 and 8 and linear between grades. This metric is a simple linear transformation of the NAEP scale, intended to render the NAEP scale more interpretable. As such, this metric is useful for descriptive research to broad audiences not familiar with interpreting standard deviation units. However, we do not advise it for analyses where the vertical linking across grades and the linear interpolation assumptions are not required or defensible.

Step 8. Calculating Achievement Gaps

We provide achievement gap estimates in SEDA 3.0 for all units except schools. Gaps are estimated as the difference in average achievement between subgroups, using the mean estimates from **Steps 5**, **6** and **7**. We provide white-black (wbg), white-Hispanic (whg), white-Asian (wag), male-female (mfg), and nonECD-ECD (neg) achievement.

In each scale, the unit-subject-grade-year gap is given by the difference in the means, e.g., the white-black gap is given by:

$$\widehat{wbg}^{x}_{uygb} = \hat{\mu}^{x}_{u(r=wht)ygb} - \hat{\mu}^{x}_{u(r=blk)ygb} \tag{9.1}$$

where x denotes a particular scale (CS, GCS) described in Steps **3** and **7** above. The standard error of the gap is given by:

$$se\left(\widehat{wbg}^{x}_{uygb}\right) = \sqrt{se\left(\hat{\mu}^{x}_{u(r=wht)ygb}\right)^{2} + se\left(\hat{\mu}^{x}_{u(r=blk)ygb}\right)^{2}} \tag{9.2}$$

The gaps can be interpreted similarly to the means in the units defined by the CS and GCS scales. If one or both of the subgroup means needed for the calculation is excluded in a given unit-subject-grade-year, the gap estimate will also be excluded.

Step 9. Pooled Mean and Gap Estimates

Pooled Mean Estimates

For each unit-subgroup, we have up to 96 subject-grade-year mean estimates (8 years, 6 grades, 2 subjects). We pool the estimates within a unit using precision-weighted random-coefficient models. These models provide more precise estimates of average performance in a unit (across grades and cohorts), as well as estimates of the grade slope (the "learning rate" at which scores change across grades, within a cohort) and cohort slope (the "trend" or rate at which scores change across student cohorts, within a grade). For GSDs, counties, CZs and metros, we provide both subject-specific and overall pooled estimates. For schools we provide only overall pooled estimates.

Subject-Specific Pooled Estimates. This model allows each unit-subgroup to have a subject-specific intercept (average test score), a subject-specific linear grade slope (the "learning rate"), and a subject-specific cohort trend (the "trend"). We fit the following model for GSDs, counties, CZs, and metros:

$$\begin{aligned}
\hat{\mu}^{x}_{urygb} = {} & [\beta_{0md} + \beta_{1md}(cohort_{urygb} - 2006.5) \\
& + \beta_{2md}(grade_{urygb} - 5.5)]M_b \\
& + [\beta_{0ed} + \beta_{1ed}(cohort_{urygb} - 2006.5) \\
& + \beta_{2ed}(grade_{urygb} - 5.5)]E_b + \epsilon_{urygb} + e_{urygb} \\
\beta_{0mu} &= \gamma_{0m0} + v_{0mu} \\
\beta_{1mu} &= \gamma_{1m0} + v_{1mu} \\
\beta_{2mu} &= \gamma_{2m0} + v_{2mu} \\
\beta_{0eu} &= \gamma_{0e0} + v_{0eu} \\
\beta_{1eu} &= \gamma_{1e0} + v_{1eu} \\
\beta_{2eu} &= \gamma_{2e0} + v_{2eu} \\
e_{uygb} \sim N(0, \omega^2_{uygb});\ & \epsilon_{uygb} \sim N(0, \sigma^2); \begin{bmatrix} v_{0mu} \\ \vdots \\ v_{2eu} \end{bmatrix} \sim MVN(0, \boldsymbol{\tau}^2).
\end{aligned} \tag{9.1}$$

In this model, M_b is an indicator variable equal to 1 if the subject is math and E_b is an indicator variable equal to 1 if the subject is ELA. β_{0bu} represents the mean test score in subject b, in unit u, in grade 5.5 for cohort 2006.5. $cohort$ is defined as $year - grade$, so this pseudocohort and pseudo-grade represents the center of our data's grade and cohort ranges, since the middle year is 2012 and the middle grade is 5.5. The β_{1bu} parameter indicates the average within-grade (cohort-to-cohort) change per year in average test scores in unit u in subject b; and, the β_{2bu} indicates the average within-cohort change per grade in average test scores in unit u in subject b.

If the model is fit using one of the scales that standardizes scores within grades (the cs scale), the coefficients will be interpretable in NAEP student-level standard deviation units (relative to the specific standard deviation used to standardize the scale). Between-unit differences in β_{0bu}, β_{1bu}, and β_{2bu} will be interpretable relative to this same scale. If the model is fit using the grade-level scale (gcs), the coefficients will be interpretable as test score differences relative to the average between-grade difference among students.

Overall Pooled Estimates. SEDA 3.0 also provides estimates pooled across grades, years, and subjects. For GSDs, counties, CZs, and metros, this model is as follows:

$$\begin{aligned}
\hat{y}^x_{uygb} = \beta_{0u} + \beta_{1u}\big(cohort_{uygb} - 2006\big) + \beta_{2u}\big(grade_{uygb} - 5.5\big) \\
+ \beta_{3u}(M_b - .5) + \epsilon_{uygb} + e_{uygb} \\
\beta_{0u} = \gamma_{00} + v_{0u} \\
\beta_{1u} = \gamma_{10} + v_{1u} \\
\beta_{2u} = \gamma_{20} + v_{2u} \\
\beta_{3u} = \gamma_{30} + v_{3u} \\
e_{uygb} \sim N\big(0, \omega^2_{uygb}\big);\ \epsilon_{uygb} \sim N(0, \sigma^2); \begin{bmatrix} v_{0u} \\ v_{1u} \\ v_{2u} \\ v_{3u} \end{bmatrix} \sim MVN(0, \boldsymbol{\tau}^2).
\end{aligned} \tag{9.2}$$

This model allows each unit to have a unit-specific intercept (average test score, pooled over subjects), linear grade slope (the "learning rate" at which scores change across grades, within a cohort, pooled over subjects), cohort trend (the "trend," or rate at which scores change across student cohorts, within a grade, pooled over subjects), and the math-ELA difference.

Tables 8 and 9 report the variance and covariance terms from the estimated $\boldsymbol{\tau}^2$ matrices from the pooling models for GSDs, counties, CZs, and metros. Tables 10 and 11 report the estimated reliabilities from these models.

For schools, we estimate the same general model as shown in equation (9.2). However, we use different grade and cohort centering. Specifically, we center relative to the middle grade of the school. We define the middle grade as the middle grade for which we have test score estimates from **Step 5**, regardless of whether or not the school serves additional grades or tested in other grades for which we could not produce estimates. For each school, the middle grade is: $mg_n = \frac{\max(grade)_n + \min(grade)_n}{2}$. Cohort is centered at: $mc_n = (2012.5 - mg_n)$. Note that 2012.5 is the middle year of our data: $\frac{2016+2009}{2} = 2012.5$. We use this same middle year, regardless of whether or not the school was observed over that whole time period. For reference, the schools in our sample tend to serve common grade spans: grades 3-5 (26,572 schools); grades 3-6 (13,330 schools); grades 3-8 (10,549 schools); grades 6-8 (12,729 schools); and, grades 7-8 (5,426 schools). In total, schools serving these grade spans make up 85% of all schools in our sample.

Tables 12 and 13 report the variance and covariance terms from the estimated $\boldsymbol{\tau^2}$ matrices, as well as the reliabilities, from the school pooling models.

Pooled Gap Estimates

We use the same models to pool gaps in GSDs, counties, CZs, and metros; however, the interpretation of the parameters differs. From these models, we recover the average test score gap across grades and years, the rate of the gap changes over grades within cohorts, and the trend in the gap across cohorts within grades.

Notably the pooled gaps are not identical to the difference in the pooled mean estimates. For users interested in analyzing pooled achievement gaps, it is important to use the pooled gap estimates rather than taking the difference between pooled estimates of group-specific means. For example, the pooled white-black gap estimate in unit u is obtained by 1) computing the gap (the difference in mean white and black scores) in each unit-grade-year-subject; 2) fitting model 10.1 or 10.2 above using these gap estimates on the left-hand side; and 3) constructing $\hat{\beta}_{0u}^{ols}$ and $\hat{\beta}_{0u}^{eb}$ from the estimates. This is the preferred method of computing the average gap in unit u. The alternative approach (taking the difference of pooled white and black mean scores) will not yield the same estimates. That is, this preferred approach will not yield identical estimates of pooled gaps as: 1) fitting model 10.1 or 10.2 above using the white mean estimates on the left-hand side; 2) constructing $\hat{\beta}_{0u(r=wht)}^{ols}$ and $\hat{\beta}_{0u(r=w)}^{eb}$ for white students from the estimates; 3) doing the same with black student mean scores to construct $\hat{\beta}_{0u(r=blk)}^{ols}$ and $\hat{\beta}_{0u(r=blk)}^{eb}$ for black students; and then 4) estimating gaps by subtracting $\hat{\beta}_{0u(r=wht)}^{ols} - \hat{\beta}_{0u(r=blk)}^{ols}$ and $\hat{\beta}_{0u(r=wht)}^{eb} - \hat{\beta}_{0u(r=blk)}^{eb}$. In particular, the EB shrunken mean of the gaps is not in general equal to the difference in the EB shrunken means. The former is preferred.

OLS and EB Estimates from Pooled Models

SEDA 3.0 contains two sets of estimates derived from the pooling models described in Equations (9.1) and (9.2). First are what we refer to as the OLS estimates of $\beta_{0u}, \dots, \beta_{3u}$. Second are the Empirical Bayes (EB) shrunken estimates of $\beta_{0u}, \dots, \beta_{3u}$. The OLS estimates are the estimates of $\beta_{0u}, \dots, \beta_{3u}$ that we would get if we took the fitted values from Model (9.1) or (9.2) and added in the residuals $v_{0u}, \dots, v_{3u}$. That is $\hat{\beta}_{0u}^{ols} = \hat{\gamma}_{00} + \hat{v}_{0u}$, for example. These are unbiased estimates of $\beta_{0u}, \dots, \beta_{3u}$, but they may be noisy in small units. We obtain standard errors of these as described in Appendix A2.

The EB estimates are based on the fitted model as well, but they include the EB shrunken residual. That is, $\hat{\beta}_{0u}^{eb} = \hat{\gamma}_{00} + \hat{v}_{0u}^{eb}$, for example, where $\hat{v}_{0u}^{eb}$ is the EB residual from the fitted model. The EB estimates are biased toward $\hat{\gamma}_{00}$, but have statistical properties that make them suited for inclusion as predictor variables or when one is interested in identifying outlier GSDs. We report the square root of the posterior variance of the EB estimates as the standard error of the EB estimate.

For a small number of cases, we were unable to recover an estimate of the OLS SE for a given parameter. For these, we report only the EB estimates of the parameter and standard error.

In general, the EB estimates should be used for descriptive purposes and as predictor variables on the right-hand side of a regression model; they are the estimates shown on the website (https://edopportunity.org). They should not be used as outcome variables in a regression model because they are shrunken estimates. Doing so may lead to biased parameter estimates in fitted regression models. The OLS estimates are appropriate for use as outcome variables in a regression model. When using the OLS estimates as outcome variables, we recommend fitting precision-weighted models that account for the known error variance of the OLS estimates.

Replicating the Pooled Estimates

Notably, we pooled non-noised long-form estimates prior to data suppression in **Step 10** (see below). Users will not be able to identically replicate our pooled estimates given two differences between the public long files and the ones used to create the pooled estimates: added noise and fewer estimates. However, the results should be largely similar.

Step 10. Suppressing Data for Release

Long Form Files

For the GSD, county, CZ, and metro long-form files, our agreement with the US Department of Education requires (1) that all reported cells reflect at least 20 students; and (2) that a small amount of random noise is added to each estimate in proportion to the sampling variance of the respective estimate. We (1) drop any estimate that does not reflect at least 20 students and (2) adjust the SEs of the means to account for the additional error.

The added noise is roughly equivalent to randomly removing one student's score from each unit-subgroup-subject-grade-year estimate. These measures are taken to ensure that the raw counts of students in each proficiency category cannot be recovered from published estimates. The random error added to each to unit-subgroup estimate is drawn from a normal distribution $\mathcal{N}(0, (1/n) * \widehat{\omega^2})$ 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. SEs of the mean are adjusted to account for the additional error. The added noise is roughly equivalent to the amount of error that would be introduced by randomly removing one student's score from each unit-subgroup-grade-year estimate.

In addition, we remove any imprecise individual estimates where the CS scale standard error greater than 2 standard deviations. Any individual estimate with such a large standard error is too imprecise to use in analysis. Table 14 summarizes the cases removed in the GSD, county, CZ, and metro long files.

Pooled Files

In the interest of discouraging the over-interpretation of imprecisely estimated parameters, SEDA 3.0 does not report EB or OLS estimates of β_u when OLS reliabilities are below 0.7. We compute the reliability of OLS estimate $\hat{\beta}_{ku}^{ols}$ as $\frac{\hat{\tau}_k^2}{\hat{\tau}_k^2+\hat{V}_{ku}}$, where $\hat{\tau}_k^2$ is the k^{th} diagonal element of the estimated $\boldsymbol{\tau^2}$ matrix (the estimated true variance of β_{kd}) and $\hat{V}_{ku}$ is the square of the estimated standard error of $\hat{\beta}_{ku}^{ols}$. That is, we do not report $\hat{\beta}_{ku}^{ols}$ if $\hat{V}_{ku} > \frac{3}{7}\hat{\tau}_k^2$. For subgroups, we use the same procedure. However, we use the standard error threshold determined for all students to censor estimates rather than calculate a subgroup-specific threshold.

II.E. Additional Notes

Gender Mean and Gap Estimates. Recent research reported by Reardon, Kalogrides, et al. (2019) suggests that the magnitude of gender achievement gaps can be impacted by the proportion of test items that are multiple-choice versus constructed-response. As a result, differences in gender gaps across states (or across time when a state changes the format of its test) may confound true differences in achievement with differences in the format of the state test used to measure achievement. See Reardon, Fahle, et al. (2019) for a description of an analytic strategy that can be used to adjust for these potential effects.

III. Covariate Data Construction

SEDA 3.0 contains CCD and ACS data that have been curated for use with the school, GSD, county, and metro achievement data. SEDA 3.0 differs from the prior version of SEDA in that it uses the new crosswalk files to aggregate the covariates to GSDs and counties, as well as releases school and metro covariate data.

III.A. ACS Data and SES Composite Construction

For GSDs, counties and metros, we use data from the ACS to construct measures of median family income, proportion of adults with a bachelor's degree or higher, proportion of adults that are unemployed, the household poverty rate, the proportion of households receiving SNAP benefits, and the proportion of households with children that are headed by a single mother. We also combine these measures to construct a single socioeconomic status composite.

ACS data for districts and counties are available as 5-year pooled samples, from which we use samples from 2006-2010 through 2012-2016. The samples we use here reflect data for the total population of residents in each unit. In select years, district-level tabulations are also available for families who live in each school district in the U.S and who have children enrolled in public school. However, the most recent sample of this data that has all of the information we need is the 5-year 2007-2011 sample. We prefer to use the total population tabulation data from more recent years. We have compared measures constructed using the total population samples and the relevant children enrolled in public schools samples in years where both samples are available and the measures are highly correlated ($r > 0.99$) and not sensitive to which sample we use.

The construction of our derived measures from the ACS data occurs in a variety of steps, which we describe below. Our derivation of these measures is complicated by the fact that we use the ACS-reported margins of error to compute empirical Bayes shrunken versions of our key ACS measures. The shrunken measures help account for attenuation bias that results from the fact that smaller units' measures include more measurement error due to smaller sample sizes. Appendix B2 describes the problems of measurement error and attenuation bias in detail. Below we describe the steps we take to create our derived measures from the raw ACS data:

Step 1: We download and clean the raw ACS data for each year and unit, saving the measures of interest along with their margins of error. We use data from the 2006-2010, 2007-2011, 2008-2012, 2009-2013, 2010-2014, 2011-2015, and 2012-2016 samples. We were unable to locate all the necessary margins of error for the 2005-2009 sample so do not use those data here. In Appendix B1 we provide a list of the raw ACS data tables we downloaded and use to compute each derived measure.

Step 2: Some of our derived measures require combining various fields from ACS in order to compute our desired metric. For example, in order to compute the proportion of adults with a bachelor's degree or higher we sum the number of men with a bachelor's degree, a master's degree or a professional degree with the number of women with a bachelor's degree, a master's degree or a professional degree and divide that sum by the total number of adults in the unit. Each of these component measures is reported with its own margin of error in the raw ACS data. We use the margins of error from each component measure to generate a single standard error for the combined bachelor's degree attainment rate variable (and do the same for all 6 socioeconomic measures that make up the SES composite). Appendix B3 describes our methodology for computing the sampling variance of sums of ACS variables in detail.

Step 3: After constructing the 6 SES measures and their standard errors we impute some missing data using Stata's –**mi impute chained–** routine, which fills in missing values iteratively by using chained equations. We reshape the data from long (one observation for each unit and race group [all, white, black and Hispanic] in each year) to wide (one observation for each unit and a separate variable for each of the 6 SES by race measures in each year). We use both the 6 SES measures and their standard errors in the imputation model as well as the total population count in each unit. The imputation model, therefore, includes median income, proportion of adults with a bachelor's degree or higher, child poverty rate, SNAP receipt rate, single mother headed household rate, and unemployment rate for each race group (all, white, black, Hispanic) in each of 7-year spans for both the estimates and their standard errors. We estimate the imputation model 5 times.

Step 4: Next we use the imputed data to compute the SES composite. This is done 5 times for each imputed data set and then we take the average. This measure is computed as the first principal component score of the following measures (each standardized): median income, percent of adults ages 25 and older with a bachelor's degree or higher, child poverty rate, SNAP receipt rate, single mother headed household rate, and employment rate for adults ages 16-64. We use the logarithm of median income in these computations. We calculate the component loadings by conducting the analysis in 2008-2012 at the GSD level and weighting by GSD enrollment. We then use the loadings from this principal component analysis to calculate SES composite values for different subgroups, years and units. Note that only observations without any imputed ACS data are used in the computation of the factor weights.

Table 15 shows the component loadings for the socioeconomic status composite as well as the mean and standard deviation of each measure it includes. The "standardized loadings" indicate the coefficients used to compute the overall GSD SES composite score from the 6 standardized indicator variables in 2008-2012, resulting in an SES composite that has an enrollment-weighted mean of 0 and standard deviation of 1 across all GSDs in 2008-2012 without any imputed data. The "unstandardized loadings" are re-scaled versions of the coefficients that are used to construct an SES composite score from the raw (unstandardized) indicator variables, but which is on the same scale as the standardized SES composite scores.

To provide context for interpreting values of the SES composite, Table 16 reports average values of the indicator variables at different values of the SES composite.

Step 5: The next step is to construct a standard error of the SES composite. We discuss our methodology in detail in Appendix B4.

Step 6: The final step is to do the empirical Bayes shrinking for the SES composites as well as for each of the 6 SES measures that go into making the composite. In addition to the time-varying versions of the SES composite, we also create an SES composite that is the average of SES in the 2007-2011 and 2012-2016 ACS (i.e., using years with non-overlapping samples). The shrinkage is done using a random effects meta-analysis regression model weighted by the standard error of each measure.

III.B. Common Core of Data Imputation

School-level data from the CCD are available from Fall 1987 until Fall 2015. There is some missing data on racial composition and free/reduced price lunch receipt for some schools in some years. We therefore impute missing data on race/ethnicity and free/reduced priced lunch counts at the school level prior to aggregating data to the GSD, county, or metro level. The imputation model includes school-level data from the 1991-92 through 2015-16 school years and measures of total enrollment, enrollments by race (black, Hispanic, white, Asian, and Native American), enrollments by free and reduced-priced lunch receipt (note that reduced-priced lunch is only available in 1998 and later), an indicator for whether the school is located in an urban area, and state fixed effects. To improve the imputation of free and reduced-priced lunch in more recent years we also use the proportion of students at each school that are classified as economically disadvantaged in the ED*Facts* data for 2008-09 through 2015-16 in the imputation model. Different states use different definitions of economically disadvantaged but these measures are highly correlated with free lunch rates from the CCD (r=.90). The imputations are estimated using predictive mean matching in Stata's –**mi impute chained–** routine, which fills in missing values iteratively by using chained equations. The idea behind this method is to impute variables iteratively using a sequence of univariate imputation models, one for each imputation variable, with all variables except the one being included in the prediction equation on the right-hand side. This method is flexible for imputing data of different types. For more information, see: https://www.stata.com/manuals13/mi.pdf.

Prior to the imputation, we make three changes to the reported raw CCD data. First, for states with especially high levels of missing free and reduced-price lunch data in recent years, we searched state department of education websites for alternative sources of data. We were only able to locate the appropriate data for Oregon and Ohio. For these states we replace CCD counts of free and reduced-price lunch receipt with the counts reported in state department of education data for 2008-09 through 2015-16. In Ohio, 8% of schools were missing CCD free lunch data in 4 or more of the 7 ED*Facts* years. In Oregon, 5% of schools were missing CCD free lunch data in 4 or more of the 7 ED*Facts* years. Other states with high rates of missing free lunch data in the CCD during the ED*Facts* years are Alaska, Arizona, Montana, Texas, and Idaho.

Unfortunately, we were unable to locate alternative data sources for these states, and rely on the imputation model to fill in missing data.

Second, starting in the 2011-12 school year some states began using community eligibility for the delivery of school meals whereby all students attending schools in low-income areas would have access to free meals regardless of their individual household income. Free lunch counts in schools in the community eligibility program are not reported in the same way nation-wide in the CCD. In community eligible schools, some schools report that all of their students are eligible for free lunch while others report counts that are presumably based on the individual student-level eligibility. Because reported free lunch eligible rates of 100 percent in community eligible schools may not accurately reflect the number of children from poor families in the school, we impute free lunch eligible rates in these schools. We replace free and reduced priced lunch counts as equal to missing if the school is a community eligible program school in a given year and their reported CCD free lunch rate is 100 percent. We then impute their free lunch eligible rate as described above.

Third, and finally, prior to imputation we replaced free and reduced-price lunch counts as missing if the count was equal to 0. Anomalies in the CCD data led some cases to be reported as zeros when they should have been missing so we preferred to delete these 0 values and impute them using other years of data from that school.

The structure of the data prior to imputation is wide – that is, there is one variable for each year for any given measure (i.e., total enrollment 1991, total enrollment 1992, total enrollment 1993, …, total enrollment 2015) for all the measures described above. The exception are time invariant measures – urbanicity and state. We impute 6 datasets and use the average of the 6 imputed values for each school in each year.

IV. Versioning and Publication

New or revised data will be posted periodically to the SEDA website. SEDA updates that contain substantially new information are labeled as a new version (e.g. V1.0, V2.0, etc.). 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, please include the version number in the citation. All versions of the data will remain archived and available on the SEDA website to facilitate data verification and research replication.

SEDA 3.0 makes the following additions to data contained in SEDA 2.1, we now release:

- Pooled estimates of the average test scores in schools with at least 20 students across grades and years.
- Subject-grade-year (long) estimates of the average test scores for all students and by student subgroups for metropolitan statistical areas and commuting zones.
- Subject-grade-year (long) estimates of the average test scores by economic disadvantage, including estimated achievement gaps between non-disadvantaged and disadvantaged students.

SEDA 3.0 makes the following modifications to the procedures used in SEDA 2.1:

- We changed the estimation procedure for all units to use the pooled HETOP model rather than the original HETOP model. When constraining estimates, this model draws on information from the same unit, rather than different units. We believe that this improves our mean estimates in units where some cells do not have sufficient data to estimate a unique standard deviation.
- We do not add any additional noise to the “pool” files per a revised agreement with the NCES. We also now release pooled estimates for units with at least 20 unique students (across grades/years), rather than requiring at least 20 students within each grade/year.

- Prior to estimation, we now remove cases where more than 40% of students take alternate assessments. We also do not report estimates for unit-subgroups with more than 20% of students taking alternate assessments.
- All test score and covariate data files have been updated to reflect updates to the crosswalk file (described in **Step 1**), including:
    - Minor corrections.
    - A new policy for districts that reorganize during the time frame of our data.
    - We use stable county identifiers, in cases where we observe that a district is placed in multiple counties during the years in our sample. The district is assigned to the county it is observed in during the 2015-16 school year.

References

Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2019). Gender Achievement Gaps in U.S. School Districts. *American Educational Research Journal*, 000283121984382. https://doi.org/10.3102/0002831219843824

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*(2), 158–189. https://doi.org/10.3102/1076998615570944

Reardon, S. F., Kalogrides, D., & Ho, A. D. (Forthcoming). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics.*

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. Educational Researcher, 1–11. https://doi.org/10.3102/0013189X18762105

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, 42(1), 3–45. https://doi.org/10.3102/1076998616666279

Shear, B. R. & Reardon, S. F. (2019) Using Pooled Heteroskedastic Ordered Probit Models to Improve Small-Sample Estimates of Latent Test Score Distributions. *CEPA Working Paper No. 19-05.* Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp19-05

Tables

Table 1. Test Score Files

File Name	Form	Metric	Unit					Disaggregated by			Subgroups						
			School	Geographic District	County	Metro	commzone	Year	Grade	Subject	Means				Gaps		
											All	Race	Gender	ECD	Race	Gender	ECD
SEDA_school_pool_cs_v30	Pooled	CS	X								X						
SEDA_school_pool_gcs_v30	Pooled	GCS	X								X						
SEDA_geodist_long_cs_v30	Long	CS		X				X	X	X	X	X	X	X	X	X	X
SEDA_geodist_long_gcs_v30	Long	GCS		X				X	X	X	X	X	X	X	X	X	X
SEDA_geodist_poolsub_cs_v30	Pooled	CS		X						X	X	X	X	X	X	X	X
SEDA_geodist_poolsub_gcs_v30	Pooled	GCS		X						X	X	X	X	X	X	X	X
SEDA_geodist_pool_gcs_v30	Pooled	CS		X							X	X	X	X	X	X	X
SEDA_geodist_pool_cs_v30	Pooled	GCS		X							X	X	X	X	X	X	X
SEDA_county_long_cs_v30	Long	CS			X			X	X	X	X	X	X	X	X	X	X
SEDA_county_long_gcs_v30	Long	GCS			X			X	X	X	X	X	X	X	X	X	X
SEDA_county_poolsub_cs_v30	Pooled	CS			X					X	X	X	X	X	X	X	X
SEDA_county_poolsub_gcs_v30	Pooled	GCS			X					X	X	X	X	X	X	X	X
SEDA_county_pool_cs_v30	Pooled	CS			X						X	X	X	X	X	X	X
SEDA_county_pool_gcs_v30	Pooled	GCS			X						X	X	X	X	X	X	X
SEDA_metro_long_cs_v30	Long	CS				X		X	X	X	X	X	X	X	X	X	X
SEDA_metro_long_gcs_v30	Long	GCS				X		X	X	X	X	X	X	X	X	X	X
SEDA_metro_poolsub_cs_v30	Pooled	CS				X				X	X	X	X	X	X	X	X
SEDA_metro_poolsub_gcs_v30	Pooled	GCS				X				X	X	X	X	X	X	X	X
SEDA_metro_pool_cs_v30	Pooled	CS				X					X	X	X	X	X	X	X
SEDA_metro_pool_gcs_v30	Pooled	GCS				X					X	X	X	X	X	X	X
SEDA_commzone_long_cs_v30	Long	CS					X	X	X	X	X	X	X	X	X	X	X
SEDA_commzone_long_gcs_v30	Long	GCS					X	X	X	X	X	X	X	X	X	X	X
SEDA_commzone_poolsub_cs_v30	Pooled	CS					X			X	X	X	X	X	X	X	X
SEDA_commzone_poolsub_gcs_v30	Pooled	GCS					X			X	X	X	X	X	X	X	X
SEDA_commzone_pool_cs_v30	Pooled	CS					X				X	X	X	X	X	X	X
SEDA_commzone_pool_gcs_v30	Pooled	GCS					X				X	X	X	X	X	X	X

Notes:

Metric: CS = Cohort Scale; GCS = Grade Scale
Unit *Metro = Metropolitan Statistical Area; CZ = Commuting Zone*
Academic Years: 2008/09 – 2014/16
Grades: 3 – 8
Subjects: Math, ELA
Race: white, black, Hispanic, and Asian
Race Gaps: white-black, white-Hispanic, white-Asian
Gender: male, female
Gender Gaps: male-female
ECD: economically disadvantaged, not disadvantaged (as defined by states)
ECD Gaps: not disadvantaged-economically disadvantaged

Table 2. Covariate Data Files

File Name	Form	Disaggregated by		
		Unit	Year	Grade
SEDA_cov_school_pooled_v30	Pooled	X		
SEDA_cov_geodist_long_v30	Long	X	X	X
SEDA_cov_geodist_poolyr_v30	Pooled	X	X	
SEDA_cov_geodist_pool_v30	Pooled	X		
SEDA_cov_county_long_v30	Long	X	X	X
SEDA_cov_county_poolyr_v30	Pooled	X	X	
SEDA_cov_county_pool_v30	Pooled	X		
SEDA_cov_metro_long_v30	Long	X	X	X
SEDA_cov_metro_poolyr_v30	Pooled	X	X	
SEDA_cov_metro_pool_v30	Pooled	X		

Table 3. Example ED*Facts* Data Structure

School	Group	Subject	Grade	Year	Number of students scoring at…			
					Level 1	Level 2	Level 3	Level 4
1	All Students	Math	3	2009	26	87	185	32
1	All Students	ELA	3	2009	13	102	195	20
2	All Students	Math	3	2009	35	238	192	7
2	All Students	ELA	3	2009	7	278	187	0

Table 4. State-Subject-Year-Grade Data Not Included in SEDA 3.0

State Abbreviation	Reason for Missing	Cases missing (gyb)
AK	No EdFacts data submitted by state	2016: ELA 3-8, Math 3-8
AR	Math tests vary by course	2009: Math 8; 2010: Math 8; 2015: Math 8
CA	Incomplete data due to pilot testing	2014: Math 7-8
CA	Math tests vary by course	2009: Math 7-8; 2010: Math 7-8; 2011: Math 7-8; 2012: Math 7-8; 2013: Math 7-8; 2014: Math 7-8
CA	Participation below 95%	2014: ELA 3-8, Math 3-6
CO	Participation below 95%	2015: ELA 5-8, Math 4-8; 2016: ELA 5-8, Math 4-8
CO	State had 1 cutscore	2009: ELA 3-8, Math 3-8; 2010: ELA 3-8, Math 3-8; 2011: ELA 3-8, Math 3-8
CT	Participation below 95%	2014: ELA 3-8, Math 3-8
DC	Participation below 95%	2015: ELA 8, Math 8
FL	Participation below 95%	2014: Math 3-8
ID	Participation below 95%	2014: ELA 3-8, Math 3-8
IL	Participation below 95%	2015: ELA 8, Math 8
KS	No EdFacts data submitted by state	2014: ELA 3-8, Math 3-8
MD	Participation below 95%	2014: ELA 3-7, Math 3-7
ME	Participation below 95%	2015: ELA 7-8, Math 6-8
MO	Math tests vary by course	2013: Math 8; 2014: Math 8; 2015: Math 8; 2016: Math 8
MT	Participation below 95%	2014: ELA 3-8, Math 3-8; 2015: ELA 3-8, Math 3-8
ND	Math tests vary by course	2015: Math 6
ND	Participation below 95%	2015: ELA 5-8, Math 7-8
NE	Each district allowed to have their own test	2009: Math 3-8; 2010: Math 3-8
NH	Participation below 95%	2015: ELA 8, Math 8; 2016: ELA 8
NJ	Participation below 95%	2015: ELA 3-8, Math 3-8; 2016: ELA 3-8, Math 3-8
NM	State had 1 cutscore	2015: ELA 3-8, Math 3-8; 2016: ELA 3-8, Math 3-8
NV	No EdFacts data submitted by state	2015: ELA 3-8, Math 3-8
NV	Participation below 95%	2014: ELA 3-8, Math 3-8
NY	Participation below 95%	2014: ELA 3-8, Math 3-8; 2015: ELA 3-8, Math 3-8; 2016: ELA 3-8, Math 3-8
OH	Math tests vary by course	2015: Math 8
OK	Math tests vary by course	2012: Math 8; 2013: Math 8
OR	Participation below 95%	2014: ELA 3-8, Math 3-8
RI	Participation below 95%	2015: ELA 5-8, Math 6-8
SD	Participation below 95%	2014: ELA 3-8, Math 3-8
TN	Math tests vary by course	2014: Math 8
TN	Testing problems, computers	2016: ELA 3-8, Math 3-8
TX	Math tests vary by course	2012: Math 7-8; 2013: Math 7-8; 2014: Math 7-8; 2015: Math 7-8; 2016: Math 7-8
UT	Math tests vary by course	2009: Math 8; 2010: Math 8; 2011: Math 8; 2012: Math 8; 2013: Math 8
UT	Participation below 95%	2016: Math 8
VA	Math tests vary by course	2009: Math 5-8; 2010: Math 5-8; 2011: Math 5-8; 2012: Math 5-8; 2013: Math 5-8; 2014: Math 5-8; 2015: Math 5-8; 2016: Math 5-8
VT	Participation below 95%	2014: ELA 3-8, Math 3-8
WA	Participation below 95%	2014: ELA 3-8, Math 3-8; 2015: ELA 3-8, Math 3-8; 2016: ELA 3-8, Math 3-8
WV	Participation below 95%	2014: Math 3-7; 2016: Math 3-7
WY	Greater than 10% more tests than enrollment	2012: ELA 3-8, Math 3-8
WY	No EdFacts data submitted by state	2010: ELA 3-8, Math 3-8
WY	Participation below 95%	2013: Math 3-8; 2014: ELA 3-8, Math 3-8

Note: Year is spring of year, so 2016 is the 2015-16 school year.

Table 5. Individual GSDs Removed Prior to Estimation due to Data Errors

District ID	District Name	State	Grade	Year	Subject
0200003	Lower Yukon School District	AK	3	2015	ela
0509750	Mena School District	AR	6	2009	math
0509750	Mena School District	AR	6	2009	ela
2201470	St. Helena Parish	LA	4	2010	ela
3910019	Marietta City	OH	7	2014	math

Table 6. NAEP Means and Standard Deviations by Year and Grade.

		Reading/English Language Arts									
	Grade	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Means	8	259.1	260.1	260.9	261.7	263.3	264.8	263.9	263.0	263.5	264.0
	7	248.5	249.3	250.0	250.7	252.1	253.4	252.8	252.3	252.6	252.9
	6	237.9	238.6	239.2	239.8	240.9	242.0	241.7	241.5	241.6	241.8
	5	227.3	227.8	228.3	228.8	229.7	230.5	230.6	230.8	230.7	230.6
	4	216.7	217.0	217.4	217.8	218.5	219.1	219.6	220.0	219.8	219.5
	3	206.0	206.2	206.5	206.8	207.3	207.7	208.5	209.3	208.8	208.4
SDs	8	36.8	36.3	36.0	35.8	35.5	35.3	35.5	35.8	36.4	36.9
	7	37.1	36.6	36.5	36.3	36.2	36.1	36.2	36.3	36.9	37.4
	6	37.5	37.0	36.9	36.9	36.9	36.9	36.9	36.9	37.4	38.0
	5	37.9	37.4	37.4	37.4	37.5	37.6	37.5	37.4	38.0	38.5
	4	38.2	37.7	37.8	37.9	38.2	38.4	38.2	38.0	38.5	39.0
	3	38.6	38.1	38.2	38.4	38.8	39.2	38.9	38.6	39.0	39.5
		Math									
	Grade	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Means	8	279.1	280.1	280.8	281.4	282.1	282.7	281.6	280.4	280.6	280.9
	7	268.8	269.6	270.2	270.8	271.5	272.1	271.1	270.1	270.2	270.2
	6	258.5	259.1	259.7	260.3	260.9	261.6	260.7	259.8	259.7	259.6
	5	248.2	248.6	249.2	249.7	250.4	251.0	250.2	249.4	249.2	248.9
	4	238.0	238.1	238.7	239.2	239.8	240.4	239.8	239.1	238.7	238.3
	3	227.7	227.6	228.1	228.7	229.2	229.8	229.3	228.8	228.2	227.7
SDs	8	37.7	37.6	37.3	37.1	37.1	37.1	37.3	37.5	38.5	39.6
	7	35.7	35.6	35.4	35.2	35.3	35.4	35.6	35.8	36.8	37.8
	6	33.8	33.7	33.5	33.4	33.5	33.7	33.8	34.0	35.0	35.9
	5	31.8	31.7	31.6	31.6	31.8	32.0	32.1	32.3	33.2	34.1
	4	29.8	29.8	29.8	29.7	30.0	30.3	30.4	30.5	31.4	32.3
	3	27.9	27.8	27.9	27.9	28.2	28.6	28.7	28.8	29.6	30.5

Note: Table shows the interpolated national NAEP estimates. We use the expanded population estimates, which may differ slightly from those reported publicly on the website.

Table 7. Subject-Grade-Year Cases Removed Pre-Estimation

Cases Dropped Pre Estimation	Districts (dygbr)	Metros	Counties	Commuting Zones
Cases dropped virtual districts	21,881 (0.25%)	Does not happen at metro level	Does not happen at county level	Does not happen at commuting zone level
Cases dropped because noGEO	2,124 (0.02%)			
Manual cases dropped	377,776 (4.33%)			
Cases dropped because state participation <95% or > 105%	380,097 (4.35%)			
Cases dropped because participation of "all" students <95% or > 105%	469,563 (5.38%)	37,829 (4.97%)	72,284 (3.20%)	40,969 (7.22%)
Cases dropped because participation of case itself <95% or > 105%	502,322 (5.75%)	81,662 (10.73)	155,002 (6.86%)	86,964 (15.31%)
Cases dropped because subgroup category total is not within 5% of all students (gender, race representation)	373,809 (4.28%)	Not at metro level	Not at county level	Not at cz level
Cases dropped because alternative assessments >40%	31,486 (0.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Cases dropped because pathological	351,788 (4.03%)	Does not happen at metro level	Does not happen at county level	Does not happen at cz level
Cases dropped because only PS cells	19,149 (0.22%)			
Cases dropped because no NP cells	254,361 (2.91%)			
Total cases dropped for any reason	582,060 (6.67%)	84,869 (11.15%)	162,130 (7.17%)	89,714 (15.80%)
Total cases not dropped	8,149,877 (93.33%)	676,185 (88.85%)	2,098,449 (92.83%)	478014 (84.20%)
Total number of cases	8,731,937 (100.00%)	761,054 (100.00%)	2,260,579 (100.00%)	567,728 (100%)

Table 8. GSD and County Variances and Covariances

Identifiers			Pooled			Math			ELA		
Unit	Metric	Subgroup	tau(int)	tau(grd)	cov(int,grd)	tau(int)	tau(grd)	cov(int,grd)	tau(int)	tau(grd)	cov(int,grd)
GSD	CS	all	0.12292	0.00216	0.00165	0.13408	0.00322	0.00268	0.11941	0.00179	0.00091
GSD	CS	asn	0.16496	0.00222	0.00183	0.18427	0.00295	0.00344	0.15681	0.00188	0.00072
GSD	CS	blk	0.07583	0.00245	0.00185	0.07936	0.00318	0.00229	0.07879	0.00200	0.00145
GSD	CS	f	0.11420	0.00198	0.00143	0.11889	0.00299	0.00225	0.11808	0.00159	0.00092
GSD	CS	hsp	0.07418	0.00248	-0.00019	0.07467	0.00326	0.00086	0.08214	0.00218	-0.00116
GSD	CS	m	0.12437	0.00222	0.00163	0.14057	0.00313	0.00293	0.11543	0.00188	0.00066
GSD	CS	nam	0.07908	0.00324	-0.00055	0.08150	0.00439	0.00013	0.08306	0.00236	-0.00096
GSD	CS	wht	0.09295	0.00205	0.00153	0.10619	0.00309	0.00213	0.08723	0.00164	0.00117
GSD	CS	mfg	0.00550	0.00012	0.00006	0.00543	0.00009	0.00003	0.00726	0.00016	0.00020
GSD	CS	wag	0.07389	0.00121	0.00076	0.08298	0.00142	0.00077	0.06994	0.00106	0.00089
GSD	CS	wbg	0.05351	0.00072	0.00126	0.05514	0.00084	0.00182	0.05413	0.00064	0.00073
GSD	CS	whg	0.04610	0.00074	0.00049	0.04583	0.00075	0.00109	0.04941	0.00079	-0.00023
GSD	GCS	all	1.22844	0.02301	0.04534	1.31389	0.03786	0.10785	1.25258	0.01885	-0.00810
GSD	GCS	asn	1.64260	0.02342	0.05363	1.80836	0.03775	0.14361	1.64213	0.02000	-0.01552
GSD	GCS	blk	0.75371	0.02521	0.03532	0.77895	0.03468	0.07150	0.82639	0.02091	0.00319
GSD	GCS	f	1.14097	0.02103	0.04058	1.16545	0.03468	0.09476	1.23848	0.01678	-0.00775
GSD	GCS	hsp	0.73619	0.02485	0.01411	0.72976	0.03380	0.05530	0.86276	0.02347	-0.02466
GSD	GCS	m	1.23901	0.02359	0.04570	1.37686	0.03754	0.11386	1.21079	0.01992	-0.01022
GSD	GCS	nam	0.78209	0.03185	0.01231	0.79322	0.04322	0.05405	0.87316	0.02552	-0.02283
GSD	GCS	wht	0.92914	0.02164	0.03843	1.04118	0.03504	0.08626	0.91484	0.01713	-0.00068
GSD	GCS	mfg	0.05532	0.00125	0.00177	0.05285	0.00109	0.00349	0.07584	0.00167	0.00119
GSD	GCS	wag	0.73492	0.01300	0.02467	0.81246	0.01703	0.05649	0.73170	0.01106	-0.00088
GSD	GCS	wbg	0.53025	0.00765	0.02398	0.53919	0.01170	0.05005	0.56784	0.00671	-0.00027
GSD	GCS	whg	0.45891	0.00759	0.01484	0.45003	0.00974	0.03804	0.51819	0.00840	-0.00992
County	CS	all	0.05916	0.00118	0.00033	0.06845	0.00179	0.00088	0.05583	0.00101	0.00003
County	CS	asn	0.11471	0.00194	0.00144	0.12985	0.00239	0.00296	0.11050	0.00183	0.00053
County	CS	blk	0.04639	0.00159	0.00040	0.05143	0.00217	0.00084	0.04715	0.00132	0.00013
County	CS	f	0.05432	0.00115	0.00046	0.06071	0.00174	0.00091	0.05470	0.00096	0.00028
County	CS	hsp	0.03820	0.00164	-0.00117	0.04172	0.00218	-0.00032	0.04235	0.00150	-0.00187
County	CS	m	0.06539	0.00122	0.00027	0.07738	0.00175	0.00097	0.05935	0.00108	-0.00015
County	CS	nam	0.07598	0.00227	-0.00099	0.07795	0.00287	-0.00049	0.08114	0.00193	-0.00143
County	CS	wht	0.04517	0.00114	0.00031	0.05672	0.00176	0.00091	0.03935	0.00095	-0.00003
County	CS	mfg	0.00393	0.00008	0.00002	0.00380	0.00006	-0.00001	0.00574	0.00011	0.00017
County	CS	wag	0.08266	0.00128	0.00249	0.08902	0.00140	0.00278	0.08148	0.00126	0.00243
County	CS	wbg	0.05218	0.00061	0.00151	0.05440	0.00079	0.00232	0.05209	0.00053	0.00081
County	CS	whg	0.04887	0.00061	0.00109	0.04927	0.00066	0.00193	0.05187	0.00065	0.00014
County	GCS	all	0.59082	0.01260	0.01775	0.67144	0.02023	0.05105	0.58588	0.01082	-0.00804
County	GCS	asn	1.13821	0.02085	0.04049	1.27245	0.02934	0.10564	1.15619	0.01959	-0.01043
County	GCS	blk	0.46152	0.01607	0.01481	0.50183	0.02247	0.04014	0.49516	0.01404	-0.00591
County	GCS	f	0.54228	0.01226	0.01766	0.59658	0.01943	0.04669	0.57391	0.01020	-0.00525
County	GCS	hsp	0.37783	0.01615	-0.00316	0.40498	0.02102	0.02330	0.44569	0.01656	-0.02626
County	GCS	m	0.65110	0.01301	0.01880	0.75813	0.02021	0.05704	0.62292	0.01159	-0.01046
County	GCS	nam	0.75061	0.02171	0.00575	0.75397	0.02759	0.04323	0.85426	0.02120	-0.02693
County	GCS	wht	0.44921	0.01209	0.01463	0.55691	0.01953	0.04435	0.41308	0.01017	-0.00624
County	GCS	mfg	0.03984	0.00080	0.00109	0.03710	0.00068	0.00230	0.05990	0.00117	0.00089
County	GCS	wag	0.82376	0.01455	0.04459	0.87330	0.01881	0.07916	0.85150	0.01303	0.01381
County	GCS	wbg	0.51805	0.00667	0.02505	0.53216	0.01157	0.05395	0.54621	0.00550	0.00087
County	GCS	whg	0.48710	0.00656	0.02085	0.48581	0.00992	0.04811	0.54385	0.00692	-0.00621

Note: GSD = Geographic district; CZ = Commuting zone; CS = cohort scale; GCS = grade-cohort scale; wht = white; blk = black; hsp = Hispanic; asn = Asian; m = male; f = female; wag = white-Asian gap; wbg = white-black gap; whg = white-Hispanic gap; mfg = male-female gap; tau = variance; rel = reliability

Table 9. CZ and Metro Variances and Covariances

Identifiers			Pooled			Math			ELA		
Unit	Metric	Subgroup	tau(int)	tau(grd)	cov(int,grd)	tau(int)	tau(grd)	cov(int,grd)	tau(int)	tau(grd)	cov(int,grd)
CZ	CS	all	0.04371	0.00068	-0.00015	0.04844	0.00099	0.00015	0.04402	0.00062	-0.00039
CZ	CS	asn	0.10533	0.00193	0.00149	0.11778	0.00246	0.00340	0.10141	0.00190	0.00032
CZ	CS	blk	0.03406	0.00107	0.00053	0.03935	0.00135	0.00104	0.03301	0.00099	0.00005
CZ	CS	f	0.04058	0.00067	0.00003	0.04315	0.00095	0.00028	0.04398	0.00063	-0.00009
CZ	CS	hsp	0.02380	0.00097	-0.00083	0.02743	0.00125	-0.00019	0.02628	0.00098	-0.00129
CZ	CS	m	0.04824	0.00068	-0.00025	0.05515	0.00096	0.00009	0.04620	0.00062	-0.00045
CZ	CS	nam	0.06789	0.00177	-0.00076	0.06554	0.00221	-0.00023	0.07730	0.00163	-0.00140
CZ	CS	wht	0.02489	0.00066	0.00011	0.03341	0.00098	0.00040	0.02080	0.00059	-0.00009
CZ	CS	mfg	0.00293	0.00005	0.00000	0.00265	0.00005	-0.00002	0.00459	0.00008	0.00013
CZ	CS	wag	0.09367	0.00158	0.00342	0.10339	0.00176	0.00468	0.08846	0.00158	0.00268
CZ	CS	wbg	0.04179	0.00047	0.00090	0.04493	0.00067	0.00178	0.04017	0.00039	0.00009
CZ	CS	whg	0.03422	0.00045	0.00088	0.03518	0.00051	0.00153	0.03590	0.00049	0.00014
CZ	GCS	all	0.43687	0.00726	0.00871	0.47429	0.01114	0.03125	0.46244	0.00679	-0.01057
CZ	GCS	asn	1.04500	0.01964	0.03813	1.14931	0.02851	0.10056	1.06205	0.02058	-0.01118
CZ	GCS	blk	0.33916	0.01106	0.01300	0.38438	0.01463	0.03389	0.34667	0.01053	-0.00435
CZ	GCS	f	0.40586	0.00723	0.00972	0.42329	0.01070	0.02942	0.46177	0.00682	-0.00752
CZ	GCS	hsp	0.23504	0.00948	-0.00331	0.26483	0.01202	0.01497	0.27675	0.01087	-0.01758
CZ	GCS	m	0.48110	0.00725	0.00915	0.53938	0.01105	0.03472	0.48535	0.00685	-0.01156
CZ	GCS	nam	0.67131	0.01658	0.00406	0.63241	0.02044	0.03616	0.81457	0.01821	-0.02563
CZ	GCS	wht	0.24846	0.00690	0.00785	0.32792	0.01078	0.02475	0.21856	0.00636	-0.00403
CZ	GCS	mfg	0.02987	0.00053	0.00059	0.02582	0.00056	0.00152	0.04787	0.00083	0.00066
CZ	GCS	wag	0.93173	0.01742	0.05626	1.01407	0.02345	0.10398	0.92435	0.01641	0.01501
CZ	GCS	wbg	0.41903	0.00495	0.01786	0.43799	0.00941	0.04319	0.42174	0.00426	-0.00474
CZ	GCS	whg	0.34397	0.00486	0.01606	0.34637	0.00766	0.03592	0.37641	0.00521	-0.00381
Metro	CS	all	0.04346	0.00083	0.00023	0.05038	0.00118	0.00048	0.04199	0.00077	0.00003
Metro	CS	asn	0.10567	0.00157	0.00089	0.11923	0.00205	0.00255	0.10203	0.00150	-0.00018
Metro	CS	blk	0.03705	0.00135	0.00089	0.04160	0.00181	0.00133	0.03762	0.00115	0.00043
Metro	CS	f	0.04124	0.00083	0.00037	0.04578	0.00117	0.00055	0.04295	0.00077	0.00033
Metro	CS	hsp	0.03095	0.00124	-0.00101	0.03481	0.00157	-0.00014	0.03417	0.00125	-0.00174
Metro	CS	m	0.04719	0.00087	0.00010	0.05635	0.00121	0.00048	0.04345	0.00080	-0.00019
Metro	CS	nam	0.05726	0.00209	-0.00118	0.05750	0.00260	-0.00134	0.06257	0.00195	-0.00123
Metro	CS	wht	0.03475	0.00082	0.00051	0.04416	0.00119	0.00089	0.03042	0.00074	0.00019
Metro	CS	mfg	0.00245	0.00006	0.00000	0.00231	0.00005	-0.00001	0.00415	0.00010	0.00015
Metro	CS	wag	0.08297	0.00108	0.00241	0.09149	0.00120	0.00327	0.07934	0.00108	0.00178
Metro	CS	wbg	0.04533	0.00054	0.00128	0.04808	0.00075	0.00211	0.04427	0.00046	0.00051
Metro	CS	whg	0.04104	0.00049	0.00090	0.04106	0.00055	0.00184	0.04421	0.00053	-0.00016
Metro	GCS	all	0.43534	0.00895	0.01272	0.49331	0.01342	0.03559	0.44058	0.00822	-0.00599
Metro	GCS	asn	1.05078	0.01676	0.03358	1.16632	0.02495	0.09433	1.06759	0.01632	-0.01671
Metro	GCS	blk	0.36846	0.01395	0.01710	0.40513	0.01928	0.03813	0.39464	0.01220	-0.00123
Metro	GCS	f	0.41324	0.00893	0.01340	0.44903	0.01320	0.03362	0.45048	0.00820	-0.00299
Metro	GCS	hsp	0.30525	0.01222	-0.00382	0.33690	0.01534	0.02009	0.35998	0.01391	-0.02359
Metro	GCS	m	0.47100	0.00926	0.01242	0.55111	0.01379	0.03913	0.45612	0.00866	-0.00852
Metro	GCS	nam	0.56501	0.01968	-0.00098	0.55526	0.02285	0.02252	0.65935	0.02142	-0.02217
Metro	GCS	wht	0.34735	0.00879	0.01401	0.43252	0.01369	0.03567	0.31946	0.00782	-0.00252
Metro	GCS	mfg	0.02501	0.00060	0.00052	0.02257	0.00052	0.00136	0.04324	0.00102	0.00097
Metro	GCS	wag	0.82684	0.01249	0.04457	0.89904	0.01731	0.08460	0.82852	0.01130	0.00705
Metro	GCS	wbg	0.45410	0.00587	0.02211	0.46955	0.01064	0.04822	0.46416	0.00479	-0.00101
Metro	GCS	whg	0.41043	0.00542	0.01684	0.40460	0.00858	0.04234	0.46365	0.00582	-0.00820

Note: GSD = Geographic district; CZ = Commuting zone; CS = cohort scale; GCS = grade-cohort scale; wht = white; blk = black; hsp = Hispanic; asn = Asian; m = male; f = female; wag = white-Asian gap; wbg = white-black gap; whg = white-Hispanic gap; mfg = male-female gap; tau = variance; rel = reliability

Table 10. GSE and County Reliabilities

Identifiers			Pooled		Math		ELA	
Unit	Metric	Subgroup	rel(int)	rel(grd)	rel(int)	rel(grd)	rel(int)	rel(grd)
GSD	CS	all	0.989	0.863	0.981	0.827	0.98	0.761
GSD	CS	asn	0.958	0.573	0.925	0.485	0.928	0.418
GSD	CS	blk	0.936	0.664	0.894	0.579	0.899	0.517
GSD	CS	f	0.985	0.819	0.974	0.779	0.975	0.69
GSD	CS	hsp	0.95	0.686	0.912	0.608	0.922	0.543
GSD	CS	m	0.986	0.829	0.976	0.777	0.975	0.716
GSD	CS	nam	0.899	0.564	0.832	0.475	0.845	0.369
GSD	CS	wht	0.982	0.829	0.972	0.788	0.969	0.707
GSD	CS	mfg	0.819	0.298	0.724	0.158	0.768	0.241
GSD	CS	wag	0.917	0.447	0.864	0.339	0.862	0.308
GSD	CS	wbg	0.911	0.434	0.856	0.334	0.86	0.308
GSD	CS	whg	0.917	0.437	0.859	0.313	0.869	0.33
GSD	GCS	all	0.988	0.867	0.981	0.849	0.981	0.765
GSD	GCS	asn	0.957	0.577	0.926	0.541	0.928	0.422
GSD	GCS	blk	0.936	0.668	0.894	0.602	0.899	0.518
GSD	GCS	f	0.985	0.824	0.974	0.804	0.976	0.693
GSD	GCS	hsp	0.95	0.686	0.912	0.623	0.922	0.55
GSD	GCS	m	0.986	0.834	0.976	0.806	0.975	0.72
GSD	GCS	nam	0.898	0.56	0.832	0.481	0.846	0.377
GSD	GCS	wht	0.982	0.832	0.972	0.809	0.969	0.709
GSD	GCS	mfg	0.82	0.303	0.723	0.184	0.768	0.238
GSD	GCS	wag	0.916	0.459	0.864	0.38	0.861	0.307
GSD	GCS	wbg	0.91	0.44	0.856	0.396	0.861	0.307
GSD	GCS	whg	0.916	0.439	0.859	0.366	0.869	0.333
County	CS	all	0.996	0.91	0.987	0.866	0.992	0.837
County	CS	asn	0.926	0.544	0.872	0.451	0.883	0.427
County	CS	blk	0.929	0.695	0.879	0.616	0.894	0.577
County	CS	f	0.994	0.882	0.981	0.834	0.988	0.789
County	CS	hsp	0.934	0.696	0.88	0.619	0.902	0.576
County	CS	m	0.994	0.883	0.983	0.825	0.989	0.799
County	CS	nam	0.888	0.547	0.821	0.458	0.842	0.406
County	CS	wht	0.991	0.882	0.977	0.836	0.982	0.794
County	CS	mfg	0.907	0.412	0.82	0.235	0.885	0.363
County	CS	wag	0.904	0.47	0.838	0.364	0.853	0.365
County	CS	wbg	0.93	0.543	0.876	0.454	0.895	0.42
County	CS	whg	0.94	0.523	0.883	0.409	0.907	0.419
County	GCS	all	0.996	0.914	0.987	0.881	0.992	0.843
County	GCS	asn	0.924	0.548	0.874	0.494	0.883	0.431
County	GCS	blk	0.929	0.696	0.881	0.628	0.895	0.582
County	GCS	f	0.994	0.886	0.982	0.851	0.989	0.794
County	GCS	hsp	0.934	0.693	0.88	0.622	0.903	0.586
County	GCS	m	0.994	0.887	0.984	0.846	0.989	0.806
County	GCS	nam	0.887	0.537	0.821	0.46	0.843	0.416
County	GCS	wht	0.991	0.884	0.978	0.851	0.983	0.801
County	GCS	mfg	0.908	0.42	0.82	0.264	0.885	0.363
County	GCS	wag	0.904	0.491	0.84	0.418	0.853	0.363
County	GCS	wbg	0.929	0.546	0.878	0.52	0.895	0.42
County	GCS	whg	0.939	0.527	0.884	0.486	0.908	0.423

Note: GSD = Geographic district; CZ = Commuting zone; CS = cohort scale; GCS = grade-cohort scale; wht = white; blk = black; hsp = Hispanic; asn = Asian; m = male; f = female; wag = white-Asian gap; wbg = white-black gap; whg = white-Hispanic gap; mfg = male-female gap; tau = variance; rel = reliability

Table 11. CZ and Metro Reliabilities

Identifiers			Pooled		Math		ELA	
Unit	Metric	Subgroup	rel(int)	rel(grd)	rel(int)	rel(grd)	rel(int)	rel(grd)
CZ	CS	all	0.998	0.929	0.995	0.902	0.996	0.877
CZ	CS	asn	0.94	0.621	0.9	0.546	0.904	0.515
CZ	CS	blk	0.927	0.71	0.893	0.645	0.89	0.622
CZ	CS	f	0.996	0.906	0.992	0.873	0.993	0.845
CZ	CS	hsp	0.952	0.758	0.919	0.693	0.923	0.664
CZ	CS	m	0.997	0.902	0.993	0.864	0.993	0.836
CZ	CS	nam	0.902	0.594	0.848	0.52	0.865	0.478
CZ	CS	wht	0.993	0.906	0.988	0.877	0.986	0.843
CZ	CS	mfg	0.941	0.517	0.884	0.376	0.93	0.483
CZ	CS	wag	0.931	0.588	0.887	0.491	0.89	0.483
CZ	CS	wbg	0.934	0.604	0.897	0.548	0.9	0.492
CZ	CS	whg	0.961	0.649	0.927	0.552	0.935	0.549
CZ	GCS	all	0.998	0.931	0.995	0.914	0.996	0.885
CZ	GCS	asn	0.939	0.613	0.902	0.578	0.904	0.522
CZ	GCS	blk	0.926	0.712	0.894	0.662	0.89	0.627
CZ	GCS	f	0.996	0.91	0.992	0.888	0.993	0.852
CZ	GCS	hsp	0.951	0.754	0.918	0.694	0.924	0.675
CZ	GCS	m	0.997	0.905	0.993	0.881	0.994	0.846
CZ	GCS	nam	0.901	0.58	0.848	0.515	0.866	0.49
CZ	GCS	wht	0.993	0.905	0.989	0.889	0.986	0.85
CZ	GCS	mfg	0.941	0.526	0.884	0.401	0.93	0.484
CZ	GCS	wag	0.93	0.6	0.889	0.544	0.89	0.483
CZ	GCS	wbg	0.932	0.592	0.898	0.602	0.9	0.499
CZ	GCS	whg	0.96	0.645	0.928	0.623	0.935	0.555
Metro	CS	all	0.998	0.94	0.995	0.916	0.996	0.893
Metro	CS	asn	0.951	0.578	0.914	0.499	0.911	0.452
Metro	CS	blk	0.948	0.744	0.918	0.683	0.917	0.631
Metro	CS	f	0.997	0.927	0.994	0.9	0.995	0.873
Metro	CS	hsp	0.97	0.8	0.949	0.741	0.948	0.707
Metro	CS	m	0.997	0.926	0.994	0.895	0.994	0.87
Metro	CS	nam	0.882	0.576	0.824	0.493	0.831	0.45
Metro	CS	wht	0.996	0.923	0.992	0.896	0.991	0.865
Metro	CS	mfg	0.952	0.543	0.901	0.345	0.945	0.512
Metro	CS	wag	0.939	0.511	0.893	0.406	0.889	0.396
Metro	CS	wbg	0.954	0.613	0.923	0.546	0.923	0.482
Metro	CS	whg	0.975	0.664	0.953	0.563	0.955	0.559
Metro	GCS	all	0.998	0.942	0.995	0.927	0.996	0.9
Metro	GCS	asn	0.95	0.58	0.914	0.541	0.911	0.459
Metro	GCS	blk	0.948	0.747	0.918	0.698	0.917	0.635
Metro	GCS	f	0.997	0.93	0.994	0.912	0.995	0.878
Metro	GCS	hsp	0.969	0.796	0.949	0.744	0.949	0.718
Metro	GCS	m	0.997	0.929	0.994	0.909	0.995	0.877
Metro	GCS	nam	0.881	0.564	0.824	0.48	0.832	0.46
Metro	GCS	wht	0.995	0.923	0.992	0.91	0.991	0.87
Metro	GCS	mfg	0.953	0.542	0.9	0.367	0.945	0.512
Metro	GCS	wag	0.938	0.535	0.894	0.475	0.889	0.396
Metro	GCS	wbg	0.953	0.609	0.924	0.609	0.923	0.484
Metro	GCS	whg	0.974	0.666	0.953	0.649	0.956	0.567

Note: GSD = Geographic district; CZ = Commuting zone; CS = cohort scale; GCS = grade-cohort scale; wht = white; blk = black; hsp = Hispanic; asn = Asian; m = male; f = female; wag = white-Asian gap; wbg = white-black gap; whg = white-Hispanic gap; mfg = male-female gap; tau = variance; rel = reliability

Table 12. School Pooling Model Variances and Covariances

Identifiers			Pooled			Math			ELA		
Unit	Metric	Subgroup	tau(int)	tau(grd)	cov(int,grd)	tau(int)	tau(grd)	cov(int,grd)	tau(int)	tau(grd)	cov(int,grd)
School	CS	all	0.20433	0.00437	0.00268	0.21708	0.00653	0.00512	0.20113	0.00320	0.00068
School	GCS	all	2.01802	0.04495	0.07390	2.03390	0.07005	0.17465	2.13233	0.03432	-0.02105

Note: CS = cohort scale; GCS = grade-cohort scale

Table 13. School Pooling Model Reliabilities

Identifiers			Pooled		Math		ELA	
Unit	Metric	Subgroup	rel(int)	rel(grd)	rel(int)	rel(grd)	rel(int)	rel(grd)
School	CS	all	0.975	0.697	0.96	0.642	0.963	0.515
School	GCS	all	0.975	0.704	0.959	0.666	0.964	0.526

Note: CS = cohort scale; GCS = grade-cohort scale

Table 14. Suppressed Estimates by Unit Post-Estimation, Long Form Data for GSDs, Counties CZs, and Metros

Cases Dropped Post Estimation	Districts (dygbr)	Counties	Metro	Commuting Zone
Drop if cs_sdse > 2	14,007 (0.17%)	51,969 (2.48%)	10912 (1.62%)	7,206 (1.51%)
Drop if cs_mnse > 2	14,803 (0.18%)	120 (0.01%)	29 (0.00%)	16 (0.00%)
Suppress if alt asmt > 20%	15,584 (0.19%)	503 (0.02%)	102 (0.02%)	0 (0.00%)
Drop if small (totgrd <20)	2,779,477 (34.10%)	337,327 (16.10%)	81,400 (12.05%)	49,012 (10.27%)
Resulting total cases in SEDA Long Files	5,354,173 (67.50%)	1,712,869 (81.73%)	583,238 (86.33%)	420,940 (88.22%)
Cases going into estimation	8,149,877 (100.00%)	2,095,764 (100.00%)	675,553 (100.00%)	477,160 (100.00%)

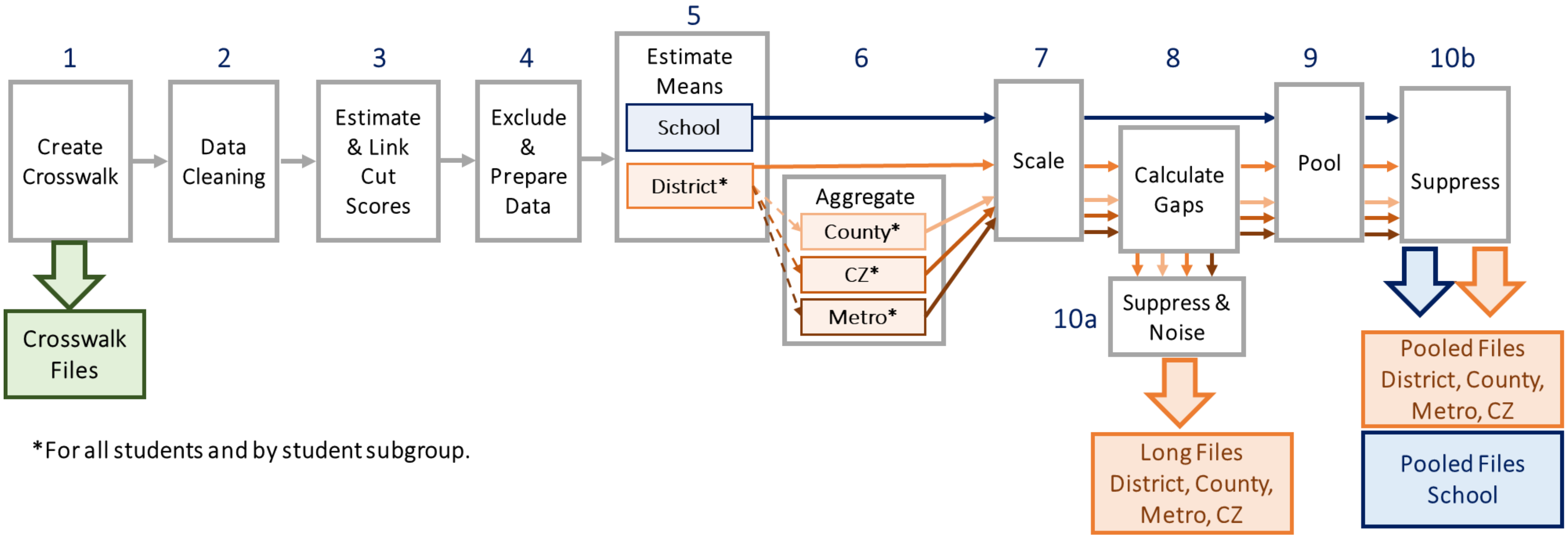
Table 15. Component Loadings and Summary Statistics for Socioeconomic Status Composite Construction.

	Standardized Loadings	Unstandardized Loadings	Mean	SD
log(Median Family Income)	0.904	0.641	10.899	0.329
% with BA or Higher	0.721	1.227	0.28	0.137
Poverty Rate	-0.921	-1.892	0.195	0.113
SNAP Eligibility Rate	-0.925	-2.997	0.121	0.072
Unemployment Rate	-0.778	-5.13	0.095	0.035
Single Mother Headed Household Rate	-0.805	-2.333	0.195	0.08

Table 16. Summary Statistics at Different Values of the Socioeconomic Status Composite.

	SES Composite						
	below -2.5	-2.5 to -1.5	-1.5 to -.5	-.5 to .5	.5 to 1.5	1.5 to 2.5	above 2.5
log(Median Family Income)	10.22	10.43	10.64	10.84	11.18	11.58	12.10
% with BA or Higher	0.12	0.15	0.20	0.24	0.36	0.58	0.80
Poverty Rate	0.47	0.38	0.28	0.19	0.09	0.04	0.02
SNAP Eligibility Rate	0.36	0.26	0.18	0.11	0.06	0.03	0.01
Unemployment Rate	0.19	0.14	0.11	0.09	0.07	0.05	0.05
Single Mother Headed Household Rate	0.43	0.33	0.25	0.19	0.14	0.10	0.07

Figures



*For all students and by student subgroup.

Figure 1. SEDA 3.0 Construction Process.

Appendices

Appendix A: Additional Detail on Statistical Methods

1. Estimating County-Level Means and Standard Deviations

This section briefly describes how means, standard deviations, and standard errors are estimated for counties and metros. As described above, we first estimate GSD-level means and standard deviations. We then estimate the county, CZ, and metro means as weighted averages of the GSD means and the county, CZ, and metro standard deviations as estimates of total variance within a county, CZ, or metro based on the GSD means and standard deviations.

The county, CZ, and metro aggregates are estimated within subjects, grades, and years. Let $\hat{\mu}_d$ and $\hat{\sigma}_d$ be the estimated means and standard deviations for the D GSD units $d = 1, \ldots,$ that will be aggregated for a given county, CZ, or metro. We also have estimates of the standard errors for each mean and standard deviation, $se(\hat{\mu}_d)$ and $se(\hat{\sigma}_d)$. We do not include grade, subject, year, or state subscripts here for clarity.

We estimate aggregate county, CZ, or metro means independently for each aggregate unit. To estimate the aggregate parameters we make the simplifying assumption that $cov(\hat{\mu}_i, \hat{\mu}_j) = cov(\hat{\sigma}_i, \hat{\sigma}_j) = cov(\hat{\mu}_i, \hat{\sigma}_i) = 0$ for $i \neq j$. The derivations for these expressions are based on the formulas in the appendix of Reardon et al. (2017) used to estimate to overall mean and variance of a set of groups in the HETOP model. Let

$$p_d = \frac{n_d}{\sum_{d=1}^{D} n_d} = \frac{n_d}{N_c}$$

be the proportion of all students in the aggregate unit c that are in GSD d. We estimate the aggregate mean for aggregate unit c as the weighted average of the GSD estimated means,

$$\hat{\mu}_c = \sum_{d=1}^{D} p_d \hat{\mu}_d,$$

with an estimated standard error of

$$se(\hat{\mu}_c) = \sqrt{\sum_{d=1}^{D} [p_d^2 \cdot se(\hat{\mu}_d)^2]}.$$

We estimate the standard deviation for aggregate unit a as the square root of the sum of the estimated between and within-GSD variance,

$$\hat{\sigma}_c = \sqrt{\sum_{d=1}^{D} [p_d(\hat{\mu}_d - \hat{\mu}_c)^2 + q_d\hat{\sigma}_d^2]},$$

with the associated estimated standard error

$$se(\hat{\sigma}_c) = \sqrt{z_c * \left(\frac{1}{\hat{\sigma}_c}\right)}.$$

In these expressions we define

$$q_d = \left(\frac{p_d + (n_d - 1)}{n_d}\right)\left(\frac{p_d}{1 + 2\left(\frac{1}{2\tilde{n}_c}\right)}\right),$$

$$\tilde{n}_c = \left[\left(\frac{1}{D}\right)\sum_{d=1}^{D}\left(\frac{1}{n_d - 1}\right)\right]^{-1},$$

and

$$z_c = \sum_{d=1}^{D} [(p_d^2(\hat{\mu}_d - \hat{\mu}_c)^2 se(\hat{\mu}_d)^2) + (q_d^2 \cdot \hat{\sigma}_d^2 \cdot se(\hat{\sigma}_d)^2)].$$

2. Constructing OLS Standard Errors from Pooled Models

In the SEDA 3.0 data, we release the OLS and EB estimates of the intercept and grade slope, as well as their standard errors, from the pooled models described in Section 9. The recovery of the OLS SEs is not straightforward from HLM. In order to recover these, we perform the estimation in two steps and calculate the OLS SEs post-estimation.

The remainder of this section describes the method and computational implementation. The equations are written to correspond to the pooling model shown in equation 9.2; however, this procedure is the same for the other variant of our pooling models.

Step 1. We estimate σ^2 using the three-level model described in equation 9.2 and define:

$$\hat{\phi}^2_{drygb} = \hat{\sigma}^2 + \omega^2_{drygb} \tag{A-2.1}$$

Where ω^2_{drygb} is the variance of the $\hat{y}^x_{drygb}$ estimate (either μ or σ). We assume that $\hat{\sigma}^2$ is a very precise estimate because of the large amount of data in the model.

Step 2. We then reweight the data and estimate a two-level HLM model:

Level-1:

$$\hat{\phi}^{-1}_{drygb}\hat{y}^x_{drygb} = [\beta_{0d} \quad \beta_{1d} \quad \beta_{2d} \quad \beta_{3d}] \begin{bmatrix} \hat{\phi}^{-1}_{drygb} \\ \hat{\phi}^{-1}_{drygb}(cohort_{drygb} - 2006.5) \\ \hat{\phi}^{-1}_{drygb}(grade_{drygb} - 5.5) \\ \hat{\phi}^{-1}_{drygb}(math_{drygb} - .5) \end{bmatrix}$$

$$+ \hat{\phi}^{-1}_{drygb} e_{drygb} \tag{A-2.2}$$

Level-2:

$$\begin{aligned} \beta_{0d} &= \gamma_{00} + \nu_{0d} \\ \beta_{0d} &= \gamma_{10} + \nu_{1d} \\ \beta_{0d} &= \gamma_{20} + \nu_{2d} \\ \beta_{0d} &= \gamma_{30} + \nu_{3d} \end{aligned}$$

After estimation, the HLM residual file contains the OLS and EB estimates, as well as the posterior variance matrices, $\boldsymbol{V}^{EB}_d$, for each GSD. From the model, we also recover an estimate of $\boldsymbol{\tau}^2$. Using $\boldsymbol{V}^{EB}_d$ and $\hat{\boldsymbol{\tau}}^2$, we can calculate the standard errors of the OLS estimates for each GSD as the inverse of:

$$(\boldsymbol{V}_d^{OLS})^{-1} = (\boldsymbol{V}_d^{EB})^{-1} - \hat{\boldsymbol{\tau}}^{-2}. \tag{A-2.3}$$

Appendix B: Covariates

1. List of Raw ACS Tables Used for SES Composite

Table Description	Table ID	Universe	Description	Usage	Derived Construct
Median household income	B19013	Households	median family income in the past 12 months	we adjust the reported median income for inflation (2012 constant dollars)	Median Income
Median household income	B19013B	Families with a householder who is Black or African American alone	median family income in the past 12 months	we adjust the reported median income for inflation (2012 constant dollars)	White Median Income
Median household income	B19013H	Families with a householder who is white alone (not Hispanic or Latino)	median family income in the past 12 months	we adjust the reported median income for inflation (2012 constant dollars)	Hispanic Median Income
Median household income	B19013I	Families with a householder who is Hispanic or Latino	median family income in the past 12 months	we adjust the reported median income for inflation (2012 constant dollars)	Black Median Income
Sex by Educational Attainment for the Population 25 and Older	B15002	Population 25 years and over	counts of number of individuals that fall into each of 16 educational attainment categories, by sex	we use the counts of men and women with a bachelor's degree or higher along with the total count to generate the BA+ rate	Bachelor's Degree Rate
Sex by Educational Attainment for the Population 25 and Older	C15002B	Black or African American alone population 25 years and over	counts of number of individuals that fall into each of 4 educational attainment categories, by sex	we use the counts of men and women with a bachelor's degree or higher along with the total count to generate the BA+ rate	Black Bachelor's Degree Rate
Sex by Educational Attainment for the Population 25 and Older	C15002H	White alone, not Hispanic or Latino population 25 years and over	counts of number of individuals that fall into each of 4 educational attainment categories, by sex	we use the counts of men and women with a bachelor's degree or higher along with the total count to generate the BA+ rate	White Bachelor's Degree Rate
Sex by Educational Attainment for the Population 25 and Older	C15002I	Hispanic or Latino population 25 years and over	counts of number of individuals that fall into each of 4 educational attainment categories, by sex	we use the counts of men and women with a bachelor's degree or higher along with the total count to generate the BA+ rate	Hispanic Bachelor's Degree Rate

Poverty Status in the Last 12 Months by Age	B17020	Population for whom poverty status is determined	counts of number of individuals living in households above and below the poverty line in various age bins	we use the counts of those living in poverty that are school aged (6-17 years old)	Poverty Rate, 6-17 Year Olds
Poverty Status in the Last 12 Months by Age	B17020B	Black or African American alone population for whom poverty status is determined	counts of number of individuals living in households above and below the poverty line in various age bins	we use the counts of those living in poverty that are school aged (6-17 years old)	Black Poverty Rate, 6-17 Year Olds
Poverty Status in the Last 12 Months by Age	B17020H	White alone, not Hispanic or Latino population for whom poverty status is determined	counts of number of individuals living in households above and below the poverty line in various age bins	we use the counts of those living in poverty that are school aged (6-17 years old)	White Poverty Rate, 6-17 Year Olds
Poverty Status in the Last 12 Months by Age	B17020I	Hispanic or Latino population for whom poverty status is determined	counts of number of individuals living in households above and below the poverty line in various age bins	we use the counts of those living in poverty that are school aged (6-17 years old)	Hispanic Poverty Rate, 6-17 Year Olds
Sex by Age by Employment Status for the Population 16 and Over	B23001	Population 25 to 64 years	counts of individuals by age, labor market status and employment status	we use the count of those employed divided by the count of those in the labor market for civilians ages 16-64 to compute an unemployment rate	Unemployment Rate
Sex by Age by Employment Status for the Population 16 and Over	C23002B	Black or African American alone, not Hispanic or Latino population 16 years and over	counts of individuals by age, labor market status and employment status	we use the count of those employed divided by the count of those in the labor market for civilians ages 16-64 to compute an unemployment rate	Black Unemployment Rate
Sex by Age by Employment Status for the Population 16 and Over	C23002H	White alone, not Hispanic or Latino population 16 years and over	counts of individuals by age, labor market status and employment status	we use the count of those employed divided by the count of those in the labor market for civilians ages 16-64 to compute an unemployment rate	White Unemployment Rate
Sex by Age by Employment Status for the Population 16 and Over	C23002I	Hispanic or Latino population 16 years and over	counts of individuals by age, labor market status and employment status	we use the count of those employed divided by the count of those in the labor market for civilians ages 16-64 to compute an unemployment rate	Hispanic Unemployment Rate

Receipt of Food Stamps/SNAP in the past 12 months by poverty status in the past 12 months for households	B22003	Households	counts of households receiving food stamps/SNAP benefits by poverty status	we use the counts of households receiving SNAP divided by the total number of households to compute the SNAP rate	SNAP Rate
Receipt of Food Stamps/SNAP in the past 12 months by poverty status in the past 12 months for households	B22005B	Households with a householder who is Black or African American alone	counts of households receiving food stamps/SNAP benefits by poverty status	we use the counts of households receiving SNAP divided by the total number of households to compute the SNAP rate	Black SNAP Rate
Receipt of Food Stamps/SNAP in the past 12 months by poverty status in the past 12 months for households	B22005H	Households with a householder who is White alone, not Hispanic or Latino	counts of households receiving food stamps/SNAP benefits by poverty status	we use the counts of households receiving SNAP divided by the total number of households to compute the SNAP rate	White SNAP Rate
Receipt of Food Stamps/SNAP in the past 12 months by poverty status in the past 12 months for households	B22005I	Households with a householder who is Hispanic or Latino	counts of households receiving food stamps/SNAP benefits by poverty status	we use the counts of households receiving SNAP divided by the total number of households to compute the SNAP rate	Hispanic SNAP Rate
Household Type	B11001	Households	counts of different types of households	we use the count of family households with a female householder, no husband present divided by the total number of family households	Female Headed Household Rate
Household Type	B11001B	Households with a householder who is Black or African American alone, not Hispanic or Latino	counts of different types of households	we use the count of family households with a female householder, no husband present divided by the total number of family households	Black Female Headed Household Rate
Household Type	B11001H	Households with a householder who is White alone, not Hispanic or Latino	counts of different types of households	we use the count of family households with a female householder, no husband present divided by the total number of family households	White Female Headed Household Rate
Household Type	B11001I	Households with a householder who is Hispanic or Latino	counts of different types of households	we use the count of family households with a female householder, no husband present divided by the total number of family households	Hispanic Female Headed Household Rate

2. Measurement Error, Attenuation Bias and Solutions

Formally, attenuation bias can be specified as follows. As an example, consider the true relationship between race-specific achievement and socioeconomic status we would like to estimate:

$$Y_g = \beta_{0g} + \beta_{1g}(SES_g) + \varepsilon_g \tag{B-2.1}$$

Where *Y* is white or non-white minority achievement in a unit (district, county, or metropolitan area) (g indexes group), and *SES* is the average socioeconomic status of the group. Race specific SES is measured with error and measurement error will be larger in units with relatively smaller sample sizes of non-white minorities. Thus, the data we observe are $W_g = SES_g + \varepsilon_g$. In this case, the bias in $\beta_{1g}$is known as attenuation bias. This bias can by quantified by multiplying by the variable's reliability $\lambda = \frac{var(SES_g)}{var(SES_g)+\sigma_1^2}$, i.e. the true variance of the variable SES_g relative to the true variance plus the variance of the measurement error.

To address attenuation bias, we use regression calibration, which makes use of the fact that the measurement error in SES_g (and consequently $SESGap$) are known from Census data.[11] Regression calibration is a method that replaces the error-prone variable W with its best linear prediction (blp). The best linear predictor of $SESGap$ can be defined as:

$$SESp_g^{blp} = E(SES_g) + \frac{cov(SES_g, W_g)}{var(W_g)}\left(W_g - E(W_g)\right)$$

$$= \mu + \frac{cov(SES_g, SES_g + \varepsilon_g)}{\sigma_{SES_g}^2 + \sigma_g^2}(W_g - \mu)$$

[11] Specifically, the ACS reports margins of error which can be easily converted standard errors for each Census variable. Appendix B3: Computing the sampling variance of sums of ACS variables provides a full description of how standard errors for cross-tabulated Census data are constructed.

$$= \mu + \lambda\left(W_g - \mu\right) \tag{B-2.2}$$

Note that $SES_g^{blp}$is "shrunken" towards the mean value of SES_g as a function of λ which, recall, is equal to the reliability of the variable SES_g and can be estimated as a random effect (or empirical Bayes estimate) from a generalized linear model.

Now, we show that regressing Y_g on SES_g^{blp}results in consistent estimates of β_{1g}.

$$\begin{aligned}\frac{cov\left(Y_g, \mu + \lambda\left(W_g - \mu\right)\right)}{var\left(\mu + \lambda\left(W_g - \mu\right)\right)} &= \frac{cov\left(Y_g, \lambda W_g\right)}{\lambda^2\left(\sigma_{SES_g}^2 + \sigma_g^2\right)} \\ &= \frac{cov\left(Y_g, SES_g\right)}{\lambda\left(\sigma_{SES_g}^2 + \sigma_g^2\right)} \\ &= \frac{cov\left(Y_g, SES_g\right)}{\sigma_{SES_g}^2} = \beta_{1g}\end{aligned}$$

(B-2.3)

3. Computing the sampling variance of sums of ACS variables

In each unit we are given counts in K cells: $\widehat{n1}_d, \widehat{n2}_d, \dots, \widehat{nK}_d$; we also know total counts t_d; we also have margins of error of the counts

$$MoE(\widehat{n1}_d), MoE(\widehat{n2}_d), \dots, MoE(\widehat{nK}_d).$$

We then compute the sampling variances of the

$$var(\widehat{nk}_d) = \left[\frac{MOE(\widehat{nk}_d)}{1.645}\right]^2$$

from these we compute

$$\widehat{pk}_d = \frac{\widehat{nk}_d}{t_d}$$

and

$$var(\widehat{pk}_d) = \frac{var(\widehat{nk}_d)}{t_d^2}.$$

We do not know the sampling rate in unit d; let's call it r_d. If the estimates come from a simple random sample, we would have

$$var(\widehat{pk}_d)^* = \frac{pk_d(1 - pk_d)}{r_d t_d}$$

The estimated design effect in district d for variable k is then

$$\widehat{Dk}_d = \frac{var(\widehat{pk}_d)}{var(\widehat{pk}_d)^*}$$

We can compute the average design effect in unit d as

$$D_d = \frac{1}{K}\sum_{k=1}^{K} \widehat{Dk}_d$$

Now we compute

$$\hat{P}_d = \frac{1}{t_d}\sum_{k=1}^{K}\widehat{nk}_d = \sum_{k=1}^{K}\widehat{pk}_d$$

We want to know $var\left(\hat{P}_d\right)$. If we had a simple random sample, we would have

$$var\left(\hat{P}_d\right)^* = \frac{P_d(1-P_d)}{r_d t_d}$$

Given the design effect in unit d, however, we would expect this to be inflated by a factor D_d.

So, we have:

$$\begin{aligned}
var\left(\hat{P}_d\right) &= D_d var\left(\hat{P}_d\right)^* \\
&= D_d\frac{P_d(1-P_d)}{r_d t_d} \\
&= \left[\frac{1}{K}\sum_{k=1}^{K}\widehat{Dk}_d\right]\frac{P_d(1-P_d)}{r_d t_d} \\
&= \left[\frac{1}{K}\sum_{k=1}^{K}\frac{var\left(\widehat{pk}_d\right)}{var\left(\widehat{pk}_d\right)^*}\right]\frac{P_d(1-P_d)}{r_d t_d} \\
&= \left[\frac{1}{K}\sum_{k=1}^{K}\frac{r_d t_d var\left(\widehat{pk}_d\right)}{pk_d(1-pk_d)}\right]\frac{P_d(1-P_d)}{r_d t_d} \\
&= \left[\frac{1}{K}\sum_{k=1}^{K}\frac{var\left(\widehat{pk}_d\right)}{pk_d(1-pk_d)}\right]P_d(1-P_d) \\
&= \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{nk_d}\right]P_d(1-P_d) \\
&= \frac{1}{\tilde{n}_d}P_d(1-P_d)
\end{aligned}$$

where nk_d=$\frac{pk_d(1-pk_d)}{var\left(\widehat{pk}_d\right)}$ is the effective sample size in cell k in unit d (the sample size nk_d such that $\frac{pk_d(1-pk_d)}{nk_d} = var\left(\widehat{pk}_d\right)$), and $\tilde{n}_d = \left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{nk_d}\right)^{-1}$ is the harmonic mean of the effective sample sizes across cells within unit d. Note that $\frac{\tilde{n}_d}{t_d} = \tilde{r}_d$ is the harmonic mean of the effective sampling rate across cells within d.

An alternate approach is to assume a common design effect across units

$$\begin{aligned} var(\hat{P}_d) &= D_d var(\hat{P}_d)^* \\ &= D_d \frac{P_d(1-P_d)}{r_d t_d} \\ &= D \frac{P_d(1-P_d)}{r_d t_d} \end{aligned}$$

where $D = \frac{1}{T}\sum_{j=1}^{J} t_j D_j$ is the average design effect across units (weighted by unit size to increase precision). We can write

$$\begin{aligned} D &= \frac{1}{T}\sum_{j=1}^{J} t_j D_j \\ &= \frac{1}{T}\sum_{j=1}^{J} t_j \left[\frac{1}{K}\sum_{k=1}^{K} \frac{r_j t_j}{nk_j}\right] \\ &= \sum_{j=1}^{J} \frac{t_j}{T}\frac{r_j}{\tilde{r}_j} \end{aligned}$$

So then,

$$\begin{aligned} var(\hat{P}_d) &= D_d var(\hat{P}_d)^* \\ &= D_d \frac{P_d(1-P_d)}{r_d t_d} \\ &= D \frac{P_d(1-P_d)}{r_d t_d} \end{aligned}$$

$$= \left[\sum_{j=1}^{J} \frac{t_j}{T} \frac{r_j}{\tilde{r}_j}\right] \frac{P_d(1-P_d)}{r_d t_d}$$

$$= \left[\sum_{j=1}^{J} \frac{t_j}{T} \frac{r_j t_d}{\tilde{r}_j t_d}\right] \frac{P_d(1-P_d)}{r_d t_d}$$

Assume r_j is constant across units and assume the effective sampling rate in unit j is independent of the unit size t_j; then this simplifies to

$$var(\hat{P}_d) = \frac{P_d(1-P_d)}{t_d \tilde{r}},$$

where

$$\tilde{r} = \left[\sum_{j=1}^{J} \frac{t_j}{T} \frac{1}{\tilde{r}_j}\right]^{-1}$$

is the (weighted) harmonic mean of the effective sampling rates. We can compute $\tilde{r}$ without knowing the actual sampling rates:

$$\tilde{r} = \left[\sum_{j=1}^{J} \frac{t_j}{T} \frac{1}{\frac{1}{t_j}\left(\frac{1}{K}\sum_{k=1}^{K} \frac{var(\widehat{pk}_j)}{pk_d(1-pk_j)}\right)^{-1}}\right]^{-1}$$

$$= \left[\sum_{j=1}^{J} \frac{t_j^2}{T} \left(\frac{1}{K}\sum_{k=1}^{K} \frac{var(\widehat{pk}_j)}{pk_d(1-pk_j)}\right)\right]^{-1}$$

To recap, we have two approaches to compute the sampling variance of $\hat{P}_d$:

1. For each unit, compute the harmonic mean of the effective sample size

$$\tilde{n}_d = \left(\frac{1}{K}\sum_{k=1}^{K} \frac{var(\widehat{pk}_d)}{pk_d(1-pk_d)}\right)^{-1}$$

then

$$Var(\hat{P}_d) = \frac{P_d(1 - P_d)}{\tilde{n}_d}.$$

Or:

2. Compute the weighted harmonic mean of the effective sampling rate across units (using any of these formulas, all identical):

$$\tilde{r} = \left[\sum_{j=1}^{J} \frac{t_j}{T} \frac{1}{\tilde{r}_j}\right]^{-1}$$

$$= \left[\sum_{d=1}^{D} \frac{t_d^2}{T} \left(\frac{1}{K} \sum_{k=1}^{K} \frac{var(\widehat{pk}_d)}{pk_d(1 - pk_d)}\right)\right]^{-1}$$

$$= \left[\frac{1}{(1.645^2)TK} \sum_{d=1}^{J} \sum_{k=1}^{K} \frac{MoE(\widehat{nk}_d)^2}{pk_d(1 - pk_d)}\right]^{-1}$$

then

$$Var(\hat{P}_d) = \frac{P_d(1 - P_d)}{\tilde{r}t_d}.$$

The first approach allows a different design effect in each unit, but the design effect is probably noisily estimated, so will have more noise in the estimated sampling variances. The second assumes a common design effect across units. Our decision criteria for generating sampling variances is as follows:

1. When $K = 1$ and $P_d > 0$, use the sampling variance provided by ACS, i.e., $var(\hat{p}_d) = \frac{var(\hat{n}_d)}{t_d^2}$

2. When $K = 1$ and $P_d = 0$, use the sampling variance method 2, i.e., $Var(\hat{P}_d) = \frac{P_d(1-P_d)}{\tilde{r}t_d}$, where $P_d = \frac{1}{t_d}$.

3. When $K > 1$ and $P_d > 0$, use the sampling variance method 2, i.e., $Var(\hat{P}_d) = \frac{P_d(1-P_d)}{\tilde{r}t_d}$

4. When $K > 1$ and $P_d = 0$, use the sampling variance method 2, i.e., $Var(\hat{P}_d) = \frac{P_d(1-P_d)}{\tilde{r}t_d}$, where $P_d = \frac{1}{t_d}$.

4. Estimating sampling variance of composite SES measures

Let $\bar{\widehat{\mathbf{X}}}_d$ be the vector of 6 variables we use to construct the SES composite in unit d. Let $\mathbf{W}_d$ be the diagonal matrix containing the standard errors of $\widehat{\mathbf{X}}_d$.[12]

Our estimated SES composite (S) in unit d is

$$\hat{S}_d = \bar{\widehat{\mathbf{X}}}_d \mathbf{B},$$

where $\mathbf{B}$ is a 6×1 vector of unstandardized coefficients. The sampling variance of $\hat{S}_d$ is

$$var(\hat{S}_d) = \mathbf{B}' \mathbf{V}_d \mathbf{B},$$

where $\mathbf{V}_d$ is the covariance matrix of $\widehat{\mathbf{X}}_d$. We know the diagonal elements of $\mathbf{V}_d$ ($\mathbf{W}_d$); but not the off-diagonals. We need to know $\mathbf{V}_d$ to get the standard error of $\hat{S}_d$. How can we compute $\mathbf{V}_d$?

Define $\mathbf{R}_d$**,** the correlation matrix describing the correlations of the estimates $\widehat{\mathbf{X}}_d$. If we knew $\mathbf{R}_d$, then we can get

$$\mathbf{V}_d = \mathbf{W}_d \mathbf{R}_d \mathbf{W}_d.$$

The key is getting an estimate of $\mathbf{R}_d$. We can use PUMS data to estimate $\mathbf{R}$ empirically (via bootstrapped samples). We do this as follows:

a. Set $N = 5{,}000$, and $J = 1{,}000$ (or some other values)

b. Pick PUMA k.

c. From all families in PUMA k, draw a random sample of N families.

[12] Note that we get the standard errors of these variables from ACS. The exception is ln(median income), as we get a standard error for median income. Let $\widehat{M}_d$ be the estimated median income in unit d. The Delta method gives us

$$se[\ln(\widehat{M}_d)] \approx \frac{1}{\widehat{M}_d} se(\widehat{M}_d).$$

d. Compute $\widehat{\mathbf{X}}_k$ from the micro-data (so if $\mathbf{X}$ includes ln(median income), then estimate ln(median income) in PUMA k from the sample, and likewise for the 6 variables we include in $\mathbf{X}$).

e. Repeat (c) and (d) J times for PUMA k.

f. Estimate $\widehat{\mathbf{R}}_k^B$ from the J samples

g. Repeat (b)-(f) for all PUMAs $k = 1, \dots, K$.

h. Repeat (b)-(g) for each race/ethnic group r to get $\widehat{\mathbf{R}}_{kr}^B$. We might need to set $N = 1{,}000$ for race-ethnic groups, because race samples are smaller in each PUMA.

Next we examine how $\widehat{\mathbf{R}}_k$ and $\widehat{\mathbf{R}}_{kr}$ vary across PUMAs and race/ethnic groups. If $\widehat{\mathbf{R}}_k$ and $\widehat{\mathbf{R}}_{kr}$ are relatively constant across PUMAs and subgroups, we can just use a single common value of $\widehat{\mathbf{R}}$ for all units and subgroups. We find that they are generally similar, so we use a common $\widehat{\mathbf{R}}$ in all PUMAs.