Identifying Progress Toward Ethnoracial Achievement Equity across U.S. School Districts: A New Approach

Allison Atteberry, PhD*
Assistant Professor
School of Education
University of Colorado-Boulder
allison.atteberry@colorado.edu

Kendra Bischoff, PhD
Associate Professor
Department of Sociology
Cornell University
kbischoff@cornell.edu

Ann Owens, PhD
Associate Professor
Department of Sociology
University of Southern California
annowens@usc.edu

September 2020
Version accepted for publication in *Journal of Research on Educational Effectiveness*

The project was supported in by a research grant from the Russell Sage Foundation. We are grateful for their support. All errors are solely attributable to the authors.

*=Corresponding author

Abstract

We draw on novel district-level test score data to describe novel approaches for measuring ethnoracial achievement gaps and assessing trends toward achievement equity from 2009 to 2016. Using SEDA data, we estimate gap trends for each grade over time in each district. We measure trends in both within-district gaps—comparing Black or Hispanic to White students in the same district—and national gaps—comparing a district's Black or Hispanic students to White students nationally. Within-district ethnoracial gaps shrunk in one-third to two-thirds of districts, depending on subject and ethnoracial dyad. Across subjects and ethnoracial dyads, national gaps shrunk in more than half of districts, indicating that non-White students gained on White students nationally, but not in their own districts. Our findings add complexity to the achievement gap literature by (1) estimating gaps at the district level; (2) noting considerable variation in the magnitude of gap shrinkage across districts; (3) pointing to the importance of comparison group and imperfect correspondence of within-district and national gap trends in districts; and (4) identifying variation in gap trends across grades and subjects.

Introduction

Ethnoracial inequality in educational achievement and attainment is a long-standing feature of the American education system. For example, national reading assessments show that White fourth graders scored nearly a standard deviation higher than Black or Hispanic students on average in 2015 (NCES, 2015). The measurement and analysis of differences in achievement among ethnoracial groups are imperfect means to fully understand and appreciate the depth of racial inequality in the U.S. education system. However, unequal outcomes, while lacking nuance, can shine a light on unequal opportunities. Instead of framing educational disparities through a deficit lens, attributing racial/ethnic gaps to the shortcomings of Black or Hispanic students, one can frame educational inequality as the result of opportunity gaps, resulting from disparate social and educational opportunities available to students of different ethnoracial groups (Carter & Welner, 2013). On their own, statistics about racial and ethnic test score gaps provide a blunt description of trends and patterns in these racialized opportunity gaps. Coupled with demographic, institutional, and community-level data, information about differences in achievement can be used to more fully understand ethnoracial inequality in the American education system. Observing and experiencing persistent ethnoracial disparities in school can shape students' beliefs about ethnoracial groups and create barriers to diverse friendships (Allport, Clark, & Pettigrew, 1954; Moody, 2001; Tyson, Darity Jr, & Castellino, 2005). Moreover, achievement outcomes predict long-term educational attainment, employment, and earnings (Chetty, Friedman, & Rockoff, 2014; Currie & Thomas, 2001; Jencks & Phillips, 1998), so these school-age educational inequalities may lead to future ethnoracial economic disparities. Therefore, understanding ethnoracial achievement gaps is imperative for documenting and addressing current and future inequalities.

Existing research on achievement inequality documents trends in national- or state-level achievement gaps or notes unequal outcomes at one point in time from survey data. Recent data

from the Stanford Education Data Archive (SEDA) allow, for the first time, a comprehensive examination of achievement scores for nearly every public school district in the U.S. These data show that achievement inequalities exist within districts as well as nationally, with White students outperforming their Black and Hispanic peers, on average, in nearly every district in the U.S. (Reardon, Kalogrides, & Shores, 2019). In this article, we contribute to existing literature in three ways. First, we estimate trends over time in test score gaps for each grade in each school district, which we interpret as indicative of district-level change. This is in contrast to measuring changes in achievement equity within cohorts of students over time (Reardon, 2019; Reardon, Weathers, Fahle, Jang, & Kalogrides, 2019). Second, we conceptualize definitions of achievement gap trends that consider how the given district's non-White subgroup fares in relation to White peers within their own school district ("within-district gap trends") as well as to White students nationally ("national gap trends"). Documenting inequalities both within districts and at a macro scale is critical to identifying the pathways through which gaps emerge and the consequences of relative academic achievement for later outcomes, such as the college admissions process. One might reach different conclusions about achievement gap patterns depending on the reference group. Third, we establish methods for detecting meaningful degrees of ethnoracial achievement gap shrinkage over time (again, both within districts and relative to peers nationwide) to understand how much progress toward racial equity is occurring. To do so, we address several methodological and measurement challenges, including scaling of achievement outcomes, estimation approach, data quality thresholds, and precision weighting. Our results provide a comprehensive and rich description of ethnoracial test score gaps in grades 3 through 8 from 2009 to 2016 by addressing the following three descriptive research questions:

(1) What is the prevalence of ethnoracial achievement gap shrinkage across the U.S., and how much do gap trends vary in magnitude?

(2) Do 'within-district gap trends' covary with 'national gap trends', as defined above?

(3) What is the extent of variation in gap shrinkage by grade and subject within a district?

The paper proceeds as follows: (I) We contextualize the unique attributes of the SEDA dataset alongside prior data used to study ethnoracial achievement gap trends. (II) We differentiate between two main ways that gaps have been tracked over time in prior research and describe our own conceptual definition of gap shrinkage. (III) We present the statistical models we use to produce parameters that capture gap shrinkage and summarize the methodological challenges and choices therein. (IV) We describe the data conditions that produce reliable gap shrinkage estimates, thereby defining our analytic sample. (V) We present descriptive analyses related to our three research questions, and (VI) conclude with a discussion of findings, limitations, and next steps.

I. Leveraging Data to Measure Achievement Gap Trends

Scholars and policymakers have long been interested in tracking ethnoracial achievement gaps over time to understand trends in inequality. Yet exactly how researchers have done so has been constrained by the nature of the data available to them.

National Assessment of Educational Progress (NAEP)

NAEP, which is administered by the National Center for Education Statistics (NCES), is the longest-running source of nationally-representative data on ethnoracial achievement gaps. The NAEP test is administered to a stratified random sample of students in certain years, and scores are aggregated overall and by student subgroups to the state and national level.[1] From these, one can produce age-specific trends in a static set of math and reading skills available since the early 1970s and grade-specific trends in a more fluid set of academic content knowledge available since

[1] There are three different forms of NAEP: NAEP Long-Term Trends (LTT) has been administered every four years since 1971 to students age 9, 13, and 17 and holds constant the content assessed over time. Main NAEP has been administered every two years since 1990 to students in grades 4, 8 and 12. Finally, NAEP Trial Urban District Assessment (TUDA) began in 2002 with six urban school districts (now 27 districts in 2019) to explore the feasibility of using NAEP to report on the performance of public school students at the district level.

1990. NAEP is particularly useful for examining gap trends over time because of its longevity, national representativeness, subgroup sampling, and because all students nationally take the same test. However, NAEP also has several limitations: It only samples students at certain ages/grades (e.g., grades 4, 8, and 12), results cannot be used to characterize trends for individual districts or schools, and it does not track individual students over time.

NCES Longitudinal Cohort Datasets

In contrast, several NCES studies have been designed to follow specific cohorts of individual students as they move through school. For instance, the National Education Longitudinal Study of 1988 (NELS:1988) follows a nationally-representative cohort of eighth graders in 1988, while the Early Childhood Longitudinal Study - Kindergarten Class (ECLS-K) follows nationally-representative cohorts of kindergarteners from 1998-99 and 2010-11 as they move through eighth grade. These datasets allow researchers to track and compare the achievement levels of specific students in more-advantaged subgroups to their less-advantaged peers. These datasets have some unique advantages for measuring gap trends: First, because the studies follow individual students over time, researchers can address compositional changes in the sample. Second, NELS and ECLS contain vertically-scaled assessments of academic achievement, producing scale scores that are comparable across grade levels. Third, these datasets contain rich information about students' schooling experiences and home lives that can be used to predict and explain achievement gaps. However, because NCES datasets use sampling frames designed to represent individuals, the data are too sparse to capture representative gap trends within specific schools or districts (e.g., in ECLS-K only about 20 kindergartens were sampled from each district).

Data Generated by Educational Testing Companies

A handful of educational testing companies have also amassed achievement data on individual U.S. students over time (e.g., College Board's PSAT/SAT data). The Northwest

Evaluation Association (NWEA), for instance, has administered the Measures of Academic Progress (MAP) assessment to over 18 million students over the last decade. Its database of vertically-scaled achievement test scores could be used to track gap trends both within cohorts of individual students as they move through school, and repeated cross-sectional grade-level gap trends. The major drawbacks of these data sources are that they are not readily available to researchers and test-takers are not randomly sampled.

Federally-Mandated Statewide Assessment Systems

Since *No Child Left Behind* (NCLB) was passed in 2002, states have been required to administer a single statewide test at the end of each school year to virtually all public school students in certain grades and report results both publicly and to the federal government. The U.S. Department of Education, in turn, maintains EDFacts, a repository of annual state achievement test data.[2] These statewide data systems are intended to monitor changes in student achievement in *every* U.S. public school and district. Since states must report results disaggregated by student subgroups, these data systems are also used to track ethnoracial and other achievement gaps over time. However, states are not required to maintain a dataset of individual student data records that track unique students over time or report achievement means and standard deviations (most states instead report percent of students in three to five proficiency categories). In addition, states typically do not use assessments that are vertically-scaled, limiting comparisons of student growth trajectories. Finally, because each state chooses its own standardized assessment and sets its own proficiency benchmarks, these test score data have been of limited use for national studies.

Current Study: Stanford Education Data Archive (SEDA)

No dataset has all of the desired properties to study achievement gap trends—individual-level data collected across multiple cohorts of students tracked longitudinally on a single,

[2] The EdFacts raw data includes no suppressed cells, nor do they have a minimum cell size for reporting.

vertically-scaled test administered in all U.S. public schools.[3] However, SEDA Version 3.0 addresses the main shortcomings of the EDFacts database. The SEDA research team developed a method to convert coarsened EDFacts proficiency rate data to achievement means and standard deviations for all districts, reconstructing achievement distributions for a given subject-grade-subgroup (e.g., ethnoracial or gender). SEDA researchers also used NAEP data to norm test score results across states to a common scale,[4] thus supporting comparisons across states and over time.

SEDA Data Structure and Coverage

SEDA provides district × year × grade × subject × subgroup estimates of achievement means and standard deviations for most U.S. public school districts. As a result of required data suppression[5] and—to a lesser degree—data quality,[6] about 67.5% cases (of the approximately 8.1 million that go into the estimation procedure) are ultimately available in the SEDA long files.

II. Conceptualizing Gap Trends

There are two distinct ways to think about tracking achievement gap trends, each with its own advantages and challenges: (1) Students within the *same cohort* may exhibit mean ethnoracial achievement gap closure as they progress through school together; or (2) mean *grade-specific* achievement gaps could shrink over time. The former seeks to follow a cohort longitudinally, while the latter calls for repeated cross-sections of the same grade in different years. Both approaches

[3] A complete examination of achievement inequality in the U.S. would include private school students, who comprised about 10% of the elementary and secondary school population during our study period. White students are slightly overrepresented while Black and Hispanic students are underrepresented in private schools. Private schools are not required to administer state tests and therefore comparable achievement data is not available.

[4] For more on how SEDA researchers equated test scores into NAEP scores across states, see the section of the SEDA 3.0 technical documentation on Cutscore Estimation and Linking (Fahle et al., 2019).

[5] The U.S. Department of Education stipulates that SEDA must drop estimates derived from fewer than 20 students and adds a small amount of random noise, which is described as roughly equivalent to randomly removing one student's score from each unit- subgroup- subject- grade- year estimate. Raw counts of students therefore cannot be recovered from published estimates. For full details, refer to pg. 38 and Table 14 of Fahle et al. (2019).

[6] The research team first excluded about 6.7% of cases during an initial round of data cleaning. Common reasons cases were discarded include: Students took incomparable tests within the state-subject-grade-year, or the state (or district or a subgroup) had participation lower than 95% in a subject-grade-year. For details, see Fahle et al. (2019).

trace the development of ethnoracial inequalities, but they evoke different kinds of explanations for the patterns observed.

Longitudinal Cohort Gap Trends

Following a cohort over time draws attention to students' learning trajectories as they move through their school-age years. Analyses that examine within-cohort gap closure over time often use longitudinal student-level data to compare the growth trajectories of White students to their non-White peers. An advantage of following individual students over time is that one can attend to possible attrition from the sample over time. However, this kind of analysis must also confront several important methodological challenges related to how one compares achievement scores of students across different grades, when they are learning fundamentally different material. Some datasets include vertically- and interval-scaled test scores (e.g., ECLS-K and NWEA)[7], which place students on a single achievement scale regardless of grade level, though designing and calibrating a vertical scale is notoriously difficult (Briggs, 2013; Briggs & Dadey, 2015; Briggs & Domingue, 2013; Briggs & Weeks, 2009). The issue of test score scaling is no small matter when examining within-cohort gap trends: For instance, von Hippel and Hamrock (2019) show that the gap in achievement between White and Black children in ECLS-K:99 appears to grow by 536% in reading from first through eighth grade (137% in math) when using the IRT-based scale scores, but those same gaps appear to only grow by 22% in reading (and no significant growth in math) when they instead use IRT-based theta versions of the scores, which are designed to be vertically scaled and to facilitate comparisons across grades.

In the absence of student-level longitudinal data, one can infer within-cohort gap closure from repeated cross sectional data of adjacent years and adjacent grades, as Reardon and

[7] When collected seasonally—thus allowing researchers to separate school-years from summers—following individual children as they move through school can shed light on whether schooling experiences exacerbate or ameliorate achievement gaps, though much has been written about the difficulties in getting these estimates right (Quinn, 2014; Quinn & McIntyre, 2017; Reardon, 2008; von Hippel & Hamrock, 2019).

colleagues have done (Reardon & Hinze-Pifer, 2017; Reardon, Kalogrides, et al., 2019). Since SEDA provides group achievement *means*—rather than individual student level data—this approach will yield valid estimates of gap closure trends under the assumption that underlying changes in cohort composition are unrelated to achievement patterns (Reardon, Papay, et al., 2019). The SEDA dataset provides achievement means translated into a grade equivalency scale to facilitate comparisons across grades. They derive White-Black (W-B) gap closure estimates for a given district by tracking changes in mean grade-level equivalency scores for White students from grade 3 in 2009, to grade 4 in 2010, grade 5 in 2011, and so on (which they refer to as a growth or learning rate) to that same growth rate for Black students. In essence, the approach compares how many grade levels of learning each subgroup acquires, on average, as they move through school (Reardon & Hinze-Pifer, 2017).[8] Reardon finds that W-B test score gaps grow by about a quarter of a grade level between grades 3 - 8. In contrast, he finds that White-Hispanic (W-H) test score gaps narrow slightly (about 0.12 grade levels) from grades 3 – 8 as Hispanic students' growth rates accelerate relative to White students.

Grade-Specific, Cross-Sectional Gap Trends

Another analytic approach is to compare achievement scores of *same* age/grade children year after year, as is reported in The Nation's Report Card using NAEP data (NCES, 2013). This approach avoids the vertical scaling issue noted above, but makes it challenging to compare the achievement outcomes of different sets of students because compositional changes in student populations could partly account for apparent gap trends.

Cross-sectional trend data on gaps indicate that both W-B and W-H gaps have followed a similar pattern over the past few decades: Gaps narrowed from the early 1970s through the late 1980s, stagnated somewhat during the 1990s, and then began to narrow again in the late 1990s

[8] Reardon (2019) interprets each subgroup's grade 3 – 8 mean test score growth as a measure of "middle childhood educational opportunities available to children [in a] district when they are roughly age 9 to 14" (pg. 41, 2019).

(Ferguson, 1998; Grissmer, Flanagan, & Williamson, 1998; Hedges & Nowell, 1998; Hedges & Nowell, 1999; Jencks & Phillips, 1998; NCES, 2013; Neal, 2005; Reardon, Cimpian, & Weathers, 2014; Reardon, Valentino, Kalogrides, Shores, & Greenberg, 2013). Broadly speaking, Black-White gaps have been slightly larger in magnitude, while White-Hispanic gaps have narrowed more slowly.

In the current study, we leverage the extensive coverage provided by SEDA to produce grade-specific ethnoracial achievement gaps over time for each school district in the U.S. More specifically, we use the SEDA data to compare annual snapshots of a district's achievement gaps from 2008-09 to 2015-16 in the same grade and subject. We aim to observe if, for instance, the eight-year trend in a district's grade 4 math achievement for Black students is steeper than that of the district's grade 4 White students over the same period. When identifying districts that are advancing equity, we intentionally describe a district's gaps as *shrinking* rather than *closing*. We do so to reinforce the fact that we do not track, within a district, whether a given cohort of students' racial achievement gap closes as they move from one grade to the next, but rather we assess the changing magnitude of a district's ethnoracial achievement gap in a particular grade over time.

Gap Trends Relative to White Students Within Same District and/or Nationally

We track two distinct kinds of gap shrinkage. The first involves a within-district comparison, wherein gaps are considered shrinking if the performance of, say, the district's grade 4 Black students approaches that of the *district's* grade 4 White students over time. The second compares the trend in performance of a given district's grade 4 Black students to that of grade 4 White students *nationally*. For simplicity, we refer to the former as "within-district shrinkage" and the latter as "national shrinkage." In Section (III) we explain how we compute both estimates from the SEDA data.

To further clarify the difference between within-district and national gap shrinkage, see Figure 1, introduced here as a visual aid and explained more fully in the Methods section. Figure 1 presents 4th grade achievement in a hypothetical district: In the left panel, test scores for both Black students (blue) and White students (red) start below the 2009 national average for all students (grey dashed line at Y=0). However, suppose the district enacts a policy that particularly stimulates the scores of the district's Black students, so the math achievement of each successive year of the district's Black 4th graders trends upwards slightly faster than the scores of each successive year of White 4th graders. Throughout the eight years of the panel, we would characterize this district as exhibiting *within-district* gap shrinkage for 4th graders. In the right panel of Figure 1, we depict the achievement trend of the district's Black students (blue) *now* compared to White students nationally (green dashed line at Y=0). This panel shows that Black 4th graders in this district have exhibited significant increases in achievement that outpace increases among White 4th graders nationally and thus could contribute to national W-B gap shrinkage. This particular district has experienced both within-district and national Black-White gap shrinkage among 4th graders.

[Insert Figure 1 about Here]

However, it is possible for grade 4 gaps in a given district to exhibit within-district shrinkage, national shrinkage, neither, or both. For example, if White students nationally are improving faster than White students in the district, it is possible that within-district gap shrinkage is occurring, but national gap shrinkage is not. Social science and educational research often rely on comparisons to draw conclusions, with the choice of which entities or groups to compare being critical to the interpretation of the findings. For example, ethnoracial achievement gap trends from the NAEP compare the test scores of a national sample of Black and White children, which is useful for understanding how average ethnoracial inequality in the United States has changed over

time. NAEP trends, however, are not useful for understanding how the context and practices of individual schools and school districts might affect ethnoracial achievement gaps. It could be the case, for example, that national ethnoracial test score gaps are large while within-district gaps are smaller under conditions of extensive racial and economic segregation between school districts. This would suggest that policies to reduce the most prevalent form of inequality should target between school-district disparities as opposed to within-district policies and practices.

As described above, in our analysis we consider two comparison groups to Black and Hispanic students in each school district—White students who share a local educational context and White students who share a national context. Documenting inequalities both within districts and at a more macro scale is useful for identifying the pathways through which gaps emerge, as the example above that compared between-district disparities vs within-district disparities shows, as well as the consequences of different definitions of relative academic achievement. The complicated reality is that educational achievement may affect individual students through many pathways, with comparisons to proximate and non-proximate students potentially influencing students in different ways. While a student may be compared to proximate students for enrollment in selective courses, for example, that same student may be compared to a national set of students in the college admissions process. A strength of this study is that we are able to identify the prevalence of different forms of achievement gap shrinkage and their covariance across districts and by grade level nationally.

In sum, this article seeks to advance the understanding of ethnoracial achievement gaps by applying a NAEP-style gap trend analysis to each public school district across the U.S. In doing so, we analyze *variability* in (and covariability between) two types of gap trends across districts in a recent period by comparing Black and Hispanic students to White students within their own districts and to White students nationally. In contrast to a longitudinal cohort-based approach

(Reardon & Hinze-Pifer, 2017; Reardon, Kalogrides, et al., 2019), our approach draws attention to cross-sectional changes over time, which could point to changes in districts' programs, policies, or local context. We conclude with an examination of the relationship among gap trends across the six grade levels and two subjects within each district to explore the prevalence and contours of variation in ethnoracial achievement gap trends within districts. As depicted in Figure 1, and as described in more detail below, our analytic focus is on districts in which Black or Hispanic students' achievement is improving, and not simply where it is declining more slowly than comparable White students.

III. Methods

To introduce our modeling approach, we focus on the W-B grade 4 math achievement gap in one example district. In practice, all models are estimated separately for every district-grade for each racial dyad gap (W-B or W-H) and subject (English/Language Arts (ELA) or math). While the basic model structure is the same for both within-district and national gap shrinkage estimates, we make different test score scaling choices and focus on different model coefficients to capture these two phenomena. The basic idea of this model is illustrated in Figure 1 for the W-B gap for one hypothetical district d in grade $g = 4$ and subject $b = math$. Figure 1 visually summarizes the differences between within-district versus national gap shrinkage (as described above) and serves as a companion to the following description of the analytic model.

Within-District: Estimating Gap Trends Relative to White Peers in Same District

To produce parameters that will allow us to identify within-district gap shrinkage, we run the following OLS regression model to compare subgroup achievement trends in each district-grade:

$$Gr4Math_{ry}^{Y09-scale} = \beta_0 + \beta_1(whtgrp_{ry}) + \beta_2(year_{ry}^{*}) + \beta_3(whtgrp_{ry} \times year_{ry}^{*}) + \varepsilon_{ry}$$

(Eq. 1)

In the model above, the outcome $Gr4Math_{ry}^{Y09\text{-}scale}$ represents the up to 16 SEDA mean 4th grade achievement estimates for up to $y = 8$ years of the panel and the $r = 2$ subgroups, White and Black students, in district *d*. Instead of using the CS-scaled achievement means provided by SEDA,[9] in Equation (1) we re-standardized estimates relative to the nationwide NAEP mean achievement in 2009 (in each grade and subject). We refer to this as the Y09-scale. Here, we intentionally construct the scale so that 0 represents a fixed point in time (2009), so that an estimated mean achievement of above 0 in any year would indicate that the given subgroup is performing better than students were nationally in 2009. By standardizing relative to the initial year of the panel, we ensure that positive trends in outcomes represent absolute improvements for a given subgroup of students. This avoids the concern that a positive slope in achievement scores for the non-White subgroup could reflect an increase only relative to the White subgroup's trend while actually decreasing in an absolute sense (i.e., Black students' achievement declining less than White students'). For a complete discussion of the scaling of the SEDA estimates, see Online Appendix B.

For our example, we model mean achievement estimates as a function of $whtgrp_{ry}$ (a dummy variable set equal to 1 if the given achievement mean belongs to White 4th graders and set equal to 0 if it belongs to Black 4th graders, $year_{ry}^{*}$ (linear time, with the asterisk indicating it was centered in the first year of the panel, 2009) and $whtgrp_{ry} \times year_{ry}^{*}$ (an interaction between the two). This model allows both subgroups to have their own achievement starting point and

[9] In the SEDA dataset, mean achievement estimates are reported in what is called the cohort-standardized (CS) scale. Fahle et al. (2019) provide the following guidance to interpreting the original CS-scale: "...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...such that all means...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." (pp. 21-22).

achievement trend. As shown in Figure 1 (left panel), district *d*'s β_0 will capture the initial mean achievement of Black students in 2009 in that district, and β_1 will capture the estimated W-B gap in 2009. The district's coefficient β_2 will capture the linear trend in grade 4 math achievement score for Black students. Finally, and most importantly, the district's β_3 coefficient will capture the difference in White and Black achievement trends between 2009 and 2016. This last coefficient allows us to capture within-district gap shrinkage: For districts in which $\boldsymbol{\beta_3}$ **is negative** (as illustrated in the left panel of Figure 1), within-district gap shrinkage has occurred over time. If β_3 is positive, gaps have widened in the district: gap expansion.[10] Finally, to characterize a district's gaps as shrinking, we also require that the Black subgroup's absolute achievement level must be increasing from 2009 through 2016 (i.e., β_2 must also be positive), which eliminates cases in which within-district gaps are shrinking simply because Black achievement is falling less precipitously than White achievement. This is a substantive choice, as our analysis is focused on identifying places of promise—that is, places where gaps are potentially shrinking because some progress is occurring for historically-underserved ethnoracial groups.[11]

By estimating the OLS achievement trend models separately for each district-grade-subject combination, we produce gap trend estimates for every district in each grade and subject, for both ethnoracial dyads. Note that there can be, at most, 16 observations in each regression model, but the majority of district-grades do not actually possess all eight annual estimates for both racial subgroups (see Online Appendix Table A1 for more on this),[12] which raises concerns about the

[10] In order to interpret the coefficients as described, we actually also require that the White subgroup must be initially outperforming the Black subgroup in 2009 (i.e., β_1 must also be positive). However, as will be shown in Table 2, in practice this is the case in over 99% of district-grade-subject observations.

[11] While it is true that inequality can decline as both groups do worse, this phenomenon must be interpreted differently; this form of gap trend shrinking may be interesting in its own right but is not the focus of the current study. In Table 2, we present the number and percent of places that are excluded as based on each requirement. Depending on the subject and racial dyad, this choice (β_2 must be positive) excludes between 13% and 35% of district-grades where shrinkage occurs.

[12] There are many possible reasons a given mean achievement estimate might be missing, but the primary cause of is the U.S. Department of Education's requirement that estimates cannot be shared from cells that contain fewer than 20 students. See our introduction of the SEDA data for an overview of missing estimates, and see Fahle et al. (2019) for a complete explanation.

validity of estimated coefficients from these models. To address this, we considered using a multilevel random effects model to produce empirical-Bayes shrunk (EBS) estimates for each district-grade (separate models for each subject). However, we found these estimates were unsuitably over-shrunk toward their grand mean. See Online Appendix B for documentation of how EBS-based gap change estimates compared to our preferred OLS-based estimates. We address potential concerns about sensitivity to sparse data by imposing a minimum data requirement for a district-grade to be included in the analytic sample (more on this below).

Finally, in our preferred specification, we use the standard errors on the mean achievement estimates provided in the SEDA data to precision-weight the data points; however, the precision-weights have little effect on the gap trend estimates. As we document in Online Appendix B, the weighted and unweighted estimates are correlated with one another at or above 0.99.

National-Level: Estimating Gap Trends Relative to White Students Nationally

To capture whether a district's Black students are narrowing gaps with White students' national mean (national gap shrinkage), we estimate the same model shown in Equation (1), but re-standardize the mean achievement outcomes around the grade-year-subject specific national NAEP mean for White students (rather than all students in 2009). We refer to this as the White-standardized (WS) scale:

$$Gr4Math_{ry}^{WS-scale} = \beta_0 + \beta_1(whtgrp_{ry}) + \beta_2(year_{ry}^{*}) + \beta_3(whtgrp_{ry} \times year_{ry}^{*}) + \varepsilon_{ry} \quad \text{(Eq. 2)}$$

The right-hand panel of Figure 1 visualizes the effect of this re-standardization on the coefficients in Equation (2). With the focus shifted to national gap shrinkage, the trend among the given district's White students is no longer relevant. This means that our focus shifts away from the coefficient β_3 (from which we identified within-district shrinkage) to the coefficient β_2 (national

gap shrinkage). Re-standardizing onto the WS-scale makes this possible: Now, if $\boldsymbol{\beta_2}$ **is positive**, it indicates that the district's Black students' mean achievement is trending upward relative to White students nationally[13]—national gap shrinkage (and if β_2 is negative, gap expansion is occurring relative to the White national mean). As summarized in Figure 1, from here forward, we use Y09-scaled estimates for within-district shrinkage and WS-scaled estimates for national shrinkage. Again, see Online Appendix B for a full description of the various scalings of the SEDA achievement estimates.

IV. Distribution of SEDA Estimates and Analytic Sample

Distribution of SEDA's Achievement Estimates Across the U.S.

For simplicity we describe the case of W-B gaps in English/Language Arts (ELA) in this section. There are just over 13,500 school districts in the U.S., which could each have mean achievement estimates for both White and Black students in a maximum of six grade levels in up to eight school years. Due to the reporting issues and data suppression described above, achievement means are not available in all district-grade-years. The SEDA dataset includes 476,462 unique district-grade-year observations and 92% of those cells possess an ELA estimate for either White students, Black students, or both (Table 1, column 1).

To appreciate the role of both ethnoracial segregation and, relatedly, the suppression of small data cells, we calculate the percentage of district-grade-year cells that possess estimates for both White and Black students (Table 1, row 1C). Across the U.S., only 20% of those cells have mean ELA achievement estimates for *both* White and Black students (i.e., there are at least 20 White students and 20 Black students in a district-grade-year). It may seem limiting that only 20% of districts in a given grade and year have estimates for both groups. However, nearly 87% of all

[13] The nationwide achievement trend for White students did not decrease during this period. This ensures that a positive β_2 indicates that Black mean achievement is increasing in an absolute sense. See Appendix Table B2.

U.S. Black public school students are covered in those 20% of district-grade-year cells (Table 1, row 1D). Patterns are similar when identifying cases with both White and Hispanic students: Only 25% of SEDA's district-grade-year observations include estimates for both subgroups, but 85% of all Hispanic public school students are represented in those district-grade-years. The availability of data affects our ability to estimate gap trends for a given district in each of its six grade levels. Next, we lay out decision rules for selecting cases for our analytic sample, implications for the representativeness of our analyses, and the robustness of results to our decisions.

[Insert Table 1 about Here]

Analytic Sample in Current Study

For each subject, our gap trend estimates are produced at the district-grade level so that a given district could have up to six unique W-B ELA gap trend estimates (one per grade). Because the unit of analysis is the district-grade in our subject-specific OLS models, we define decision rules for analytic sample inclusion at that level. Row (2A) of Table 1 shows that, in total, there are 16,043 district-grade observations with at least one year with ELA estimates for both White and Black students. This is the universe of cases where there is *any* information about both subgroups.

To produce valid estimates of a district's achievement gap trends, we need reliable estimates of each subgroup's achievement trend in the same district in a given grade. Refer to Online Appendix Figure A1 for a visual representation of one district's subgroup achievement estimates in ELA for each of its six grade levels. Ideally, there would be 16 estimates for each district-grade (i.e., each scatterplot in Figure A1), one per year each for White and Black students. However, most district-grades do *not* have all 16 estimates, and, the sparser the data, the more extreme and easily influenced by outliers the trends become (see again Online Appendix Figure A1). Trends also become noisier when data points are clustered among consecutive school years in the dataset.

To address these issues, we impose a "minimum data requirement" (MDR) for a district-grade to be included in analyses (a given district could meet the definition in some grades but not others). We require that the district-grade has mean achievement estimates for both subgroups either in the first and last year of the data panel or in at least 6 of the 8 panel years. We also check the sensitivity of our results to how we define the MDR.[14] We ultimately think this MDR definition balances the need for sufficient data with efforts to minimize the number of district-grades that are excluded. Row (2B) of Table 1 shows that 63% of district-grades meet the MDR, and row (2C) shows that about 73% of all Black student observations are represented in these 10,046 district-grades. This is the analytic sample for all subsequent analyses. The sample size for W-H gap shrinkage in ELA is 10,872 district-grades, in which 69% of all Hispanic students are enrolled.

Since inclusion decisions are made at the district-grade level, a given district might meet the MDR in some grades but not others, as shown in Table 1. For ELA, 2,048 districts have at least one grade level that meets the MDR (row 3A). Of those, only 64% meet the MDR in all six grade levels, while 86% meet the MDR in at least three of six grade levels. The percentages tend to be slightly lower in math. In general, *patterns* in the size of the analytic samples are similar for W-H gap analyses, but about ten percentage points lower than for W-B gap analyses.

[14] In Online Appendix A, we explore the impact of using six different MDR definitions that vary in terms of restrictiveness/inclusiveness, as well as not using any MDR. The percentage of district-grades that are identified as exhibiting district gap shrinkage is relatively consistent across MDR definitions. However, the standard deviations (or 1^{st}-99^{th} percentile ranges) are quite different due to how the MDRs address outlier trends. For instance, Appendix Table A3 shows that for W-B ELA district gap trends, the 1^{st}-99^{th} range of estimates is from -0.60 to +0.68 with MDR definition #1. However, using the least restrictive MDR definition #6 (column 6), we observe a 1^{st}-99^{th} percentile range of estimates from -1.53 to +1.68. However, when we visually examined district-grades with gap trend estimates larger in magnitude than +/- 0.70 SDs, we almost always find these are cases with less annual achievement data available with influential data points possibly skewing the trend estimates. We generally do not think these district-grades have sufficient evidence of how their gaps are trending during this period.

V. Results

RQ1: What is the Prevalence of Ethnoracial Achievement Gap Shrinkage across the U.S., and How Much Do Gap Trends Vary in Magnitude?

Within-district gap shrinkage. The top panel of Table 2 presents the prevalence of W-B and W-H within-district gap shrinkage for the district-grades that met the MDR. Recall that we impose two additional requirements for our measure of within-district gap shrinkage. First, the non-White group must initially perform lower than the White subgroup in 2009, an analytic decision that eliminates almost no district-grades in our sample. Second, non-White outcomes must improve over time and not just decrease less precipitously than outcomes for the White subgroup. With these two restrictions, 59% and 48% of district-grades can be assessed for within-district W-B gap shrinkage in ELA and math, respectively (row 3). Seventy-five percent and 55% of district-grades are eligible for valid within-district W-H gap reductions in ELA and math, respectively. Table 2, row 4, shows that among these district-grade observations, W-B gap shrinkage occurred in only 37% in ELA (N=3,704) and 30% in math (N=2,715) (that is, $\beta 3$ is negative). W-H gap shrinkage occurred in 57% of eligible district-grades for ELA (N=6,184) and only 36% in math (N=3,451).

[Insert Table 2 about Here]

Rows 5-9 of Table 2 present the prevalence of five different magnitudes of gap shrinkage—at least 0.10, 0.20, 0.30, 0.40, and 0.50 standard deviations (SDs). The higher the threshold of within-district gap shrinkage, the less often district-grades meet the criteria. For example, depending on the racial dyad and subject, between 8% and 16% of district-grades exhibited gap shrinkage of at least 0.30 SDs between 2009 and 2016. Shrinkage of at least 0.50 SDs is extremely rare, occurring in fewer than 5% of district-grades across racial dyads and subjects.

Figure 2 shows the distribution of possible W-B and W-H within-district gap changes from 2009 through 2016 in ELA and math (1^{st}-99^{th} percentile range), highlighting the wide variation in gap trends across district-grades. Negative estimates correspond to gap shrinkage and positive estimates to gap expansion. The darker the shaded bars, the greater the shrinkage. During this 8-year period, most of the gap trend estimates fall between +/-0.70 SDs, depending on subject and gap dyad. Although the mean within-district gap trend estimates for both W-B and W-H gaps are near zero in both subjects over this period, the mean is slightly positive for W-B gaps (left column) and slightly negative for W-H gaps (right column), which shows that, on average, W-B within-district gaps expanded and W-H within-district gaps shrunk.

[Insert Figure 2 about Here]

National gap shrinkage. The lower panel of Table 2 shows that, broadly speaking, more district-grades exhibit national gap shrinkage (comparing non-White subgroups in a district to White students nationally) than within-district shrinkage. In ELA, 37% of district-grades exhibit W-B within-district shrinkage whereas 52% exhibit national gap shrinkage. In math, the difference is even more striking: While only 30% of district-grades exhibit W-B within-district shrinkage, 58% exhibit national shrinkage. National shrinkage is also more common for W-H gaps (69% in ELA, 63% in math). Rows 3-7 present the prevalence of magnitudes of national gap shrinkage. These results show that a greater share of district-grades exhibit national shrinkage of large magnitudes than is true for within-district shrinkage. For example, for W-B gaps, 18% and 25% of district-grades in ELA and math, respectively, and for W-H gaps, approximately 30% of district-grades in both subjects exhibit national shrinkage of at least 0.30 SDs, compared to between 9 and 16% for within-district shrinkage. We discuss the correspondence of within-district and national shrinkage in the next section.

Figure 3 presents the distribution of national gap trend estimates.[15] On average, national gaps shrank in all subjects and racial dyads, though most notably in W-H national gaps, wherein the average district's Hispanic students decreased their gap with White students nationally by 0.14 SDs (ELA) and 0.12 SDs (math). The range of national gap trend estimates is also generally broader than for within-district gap trend estimates. For instance, W-B gaps in math generally range from decreasing by as much as nearly 1 SD to increasing by as much as nearly 0.80 SDs.

[Insert Figure 3 about Here]

RQ2: Do Within-District Gap Trends Covary with National Gap Trends?

Within-district and national gap trends may not move in the same direction. Declining ethnoracial inequality within districts may be the result of new policies or practices, but all students in the district may still lag national progress. Districts where both White and non-White students are outperforming the national trend may be preparing students well for national competition in college admissions or the labor market, but persistent inequity within the district might create future barriers for non-White students in, for example, admission to AP classes or other district opportunities.

Figures 4 (W-B gaps) and 5 (W-H gaps) present scatterplots that juxtapose within-district gap trend estimates with national gap trend estimates. The first quadrant (shaded gray) in each graph contains district-grade observations where neither within-district nor national gaps shrank (achievement gaps grew), while the third quadrant (shaded green) in each graph contains district-grade observations where both within-district and national gaps shrank (achievement gaps narrowed). District-grade observations in the first and third quadrants thus show conclusive

[15] Recall that the SEDA achievement estimates are rescaled for national gap shrinkage analyses such that a positive β_2 coefficient on the variable $year_{ry}^{*}$ captures a shrinking gap between the district's non-White students and White students nationally over time. However, for the sake of clarity, we reverse scale these coefficients so that—as with district shrinkage estimates—more positive estimates represent widening gaps, and more negative estimates represent narrowing gaps (as noted in Figure 3).

evidence of movement toward worsening or improving achievement equity, respectively, for Black and Hispanic students.

Observations in the second and fourth quadrants, however, highlight more complex narratives. The second quadrant (shaded pink) contains district-grades where only within-district shrinkage occurred: non-White students' achievement improved relative to White students in their own districts, but not relative to White students nationally. This implies that all ethnoracial groups in the district experienced smaller achievement gains than the mean gains for White students nationally, with non-White students experiencing larger gains than White students in their district. The fourth quadrant (shaded blue) contains district-grades where only national shrinkage occurred: non-White students' achievement improved relative to Whites nationally, but not relative to White students within their own districts. This implies that all ethnoracial groups in the district experienced larger gains than the mean gains for White students nationally, but White students' gains outpaced Black students' gains in the same district.

For the W-B gap-change estimates shown in Figure 4, within-district and national gap trend estimates are correlated at 0.615 in ELA and at 0.479 in math. In ELA, we find that 34% of these district-grades exhibit both national and within-district shrinkage (green dots), while 36% exhibit both national and within-district expansion (gray dots). In 11% of district-grades, W-B within-district shrinkage occurs while W-B national shrinkage does not (pink dots). In the remaining 18% of districts, W-B national shrinkage occurs, while within-district shrinkage does not. Results are similar for math W-B gaps (lower panel of Figure 4), though a slightly smaller percentage (30%) exhibit both within-district and national gap expansion.

[Insert Figure 4 about Here]

Figure 5 shows W-H gap-change estimates and for ELA, within-district and national gap shrinkage corresponds much more strongly than is true for W-B estimates. In 54% of district-

grades, the W-H gap shrank both within-district and compared to White achievement nationally (versus just 34% for W-B gaps). In math, 40% of district-grades experienced W-H gap shrinkage both within districts and nationally compared to 34% for W-B gaps. Compared to W-B gaps, similar proportions of district-grades experienced W-H gap shrinkage either only within districts or only nationally in both ELA and math.

[Insert Figure 5 about Here]

RQ3: What is the Extent of Variation in Gap Shrinkage by Grade and Subject within a District?

It would perhaps be most intuitive to think of gap shrinkage as a district-level phenomenon that occurs in response to changes in a district's programs, practices, and residential community that promote achievement in a uniform way regardless of grade level or subject. It would also simplify any attempt to identify districts that are improving ethnoracial equity if gap shrinkage tended to be similar across grade levels or subjects. Moreover, consistency in gap shrinkage across a district's grade levels or subjects would signal that local conditions or district policies are, in fact, contributing to those promising trends.

In practice, however, we find that gap trends across a district's six grades or its subjects do not often all move in the same direction. Column percentages in Table 3 (W-B within-district gaps) and Table 4 (W-H within-district gaps) show how uncommon it is to observe districts that have consistent gap shrinkage (of different magnitudes) across grades. For instance, consider the 1219 districts where, in at least one grade, the W-B within-district ELA gap shrinks by at least 0.10 SDs (column 2 of Table 3). Only 1.2% (N=15) of those districts exhibit gap shrinkage of at least 0.10 SDs in all six grades. In approximately 44% of those districts (N=541), the gap shrinks by 0.10+ SDs (i.e., at least 0.10 SDs) in only a single grade. Only 13.6% of those districts exhibit 0.10+ SD gap shrinkage in four or more of the district's six SEDA grades. As the specified threshold of gap

shrinkage increases across the columns, fewer districts experience consistent improvement across grades. For both ELA and math, there are no districts where W-B gaps shrink by 0.30+ SDs in all six of its grades (column 4 of Table 3).

[Insert Table 3 about Here]

Consistency in within-district gap shrinkage across all six grades is slightly more common for W-H gaps (Table 4). In ELA, for instance, of the 2033 districts where W-H gaps shrink in at least one grade, 9.1% of them exhibit shrinkage in all six grades, compared to 3.8% for W-B gaps (column 1 of Table 3 versus Table 4). However, even for W-H within-district gaps, it is uncommon that all six of a district's grades have gap shrinkage of a given magnitude (only 3.5% for gap shrinkage of 0.10+ SDs, 1.2% for 0.20+ SDs, and 0.4% for 0.30+ SDs). For math (lower panel of Table 4), consistent gap shrinkage across grades is even more elusive, as it was for W-B gaps.

[Insert Table 4 about Here]

We conduct the same analyses presented in Tables 3 and 4 for national gap shrinkage (results available upon request), and the overall pattern of findings is quite similar: It is uncommon for districts to exhibit national gap shrinkage in all six grades at or above a given magnitude. The heterogeneity in within-district gap shrinkage estimates across grades suggests that it may not be appropriate to characterize an *entire* district as uniformly experiencing gap shrinkage. Very few districts would be classified as having gap shrinkage if we require shrinkage for all grades.

Of course, this does not mean that the direction and magnitude of gap trends across grades in the same district are entirely unrelated to one another. Table 5 shows that the correlations across a district's grade-specific shrinkage estimates are generally positive and are between 0.40 and 0.53 for adjacent grades.[16] Although the correlations are positive across all grade levels, racial dyads, and subjects, they decline, sometimes substantially, for grade levels that are further apart. In the

[16] These models include district fixed effects. We also find these correlations are quite similar if we use MDR definition 1 instead of our preferred MDR definition 2, or if we limit the analysis to districts in which all 6 grades meet the MDR (see row 3F of Table 1). Results available upon request.

most extreme case, the correlation between shrinkage estimates in grades 3 and 8 for the same district is only as high as 0.14 (W-B gaps in ELA) and can be as low as 0.04 (W-B gaps in math and W-H gaps in ELA).

[Insert Table 5 about Here]

Figure 6 (W-B gap) and Figure 7 (W-H gap) convey the degree of variability in gap trend estimates across grades within the same district. To accomplish this, we randomly sampled 75 districts, rank-ordered them by their grade 3 within-district gap trend estimates (x-axis), and plotted each of their grade-specific gap shrinkage estimates (scaled in SDs) as points on one vertical line per district. Recall that negative estimates indicate gap shrinkage whereas positive estimates indicate gap expansion. These graphs visualize how uncommon it is for any given district to exhibit uniformly positive or negative estimates across all six grades, indicated by the fact that most lines straddle the horizontal zero line. The range between the highest and lowest estimate is often around 0.40 to 0.60 SDs. These graphs also highlight patterns in grades that tend to have higher or lower estimates across districts. For example, in the upper panel of Figure 6 (W-B gap estimates in ELA), it appears that grades 7 and 8 are over-represented among the gaps that are widening the most (that is, most positive), while grade 3 and 4 often experience the greatest gap shrinkage. Indeed, Table 5 shows that mean W-B gap trends are largest (more positive) in middle school grades and smaller (more negative) in earlier grades. The grade-specific pattern is less clear for W-H within-district gaps. We focus here on within-district gap shrinkage, but results for national gap shrinkage are quite similar.

[Insert Figure 6 & Figure 7 about Here]

Turning from a focus on correspondence across grades to correspondence across subjects, Table 6 shows that about 70% of district-grades exhibit math and ELA gap trends that move in the same direction (either both widening or both shrinking), and math-ELA correlations tend to be

around 0.60. Finally, we bring grades and subjects together to look for consistency across the (up to 12) gap trend estimates for each district (Figure 8). In only 10.4% of districts do W-B gaps shrink in at least 90% of their estimates (18.6% for W-H gaps).

[Insert Table 6 and Figure 8 about Here]

Taken together, the analyses of achievement gap trends by grade and subject show that there is variability across educational dimensions within districts in gap shrinkage. The implementation of a fairly restrictive minimum data requirement ensures that we have only reported results for districts with strong data to support our inferences, and guards against inconsistencies due to noisy estimates. Though gap trend estimates across grades in the same district are positively correlated, the degree of across-grade variability is substantively large. Few districts experience gap shrinkage in all grades or in all subjects. Our findings suggest that further research is needed to understand why gap trends in different grades in the same district might be moving in different directions.

VI. Conclusions, Limitations, and Next Steps

In this article, we develop novel conceptual and methodological approaches for measuring achievement gap shrinkage. We then apply these measures to explore the degree to which school districts in the U.S. are making progress in shrinking achievement gaps between White students and Black and Hispanic students from 2009-2016. We measure within-district shrinkage—whether Black or Hispanic students' achievement is (1) increasing; and (2) approaching that of White students in their district —and national shrinkage—whether a district's Black or Hispanic students' achievement is approaching that of White students nationally. Newly released data from SEDA enable a comprehensive analysis of achievement gap trends over time, across grades, and across subjects. First, we find that White-Black within-district gaps shrank in 37% and 30% of district-

grades in ELA and math, respectively. White-Hispanic ELA within-district gaps shrank in 57% of district-grades, while math gaps shrank in 36% of district-grades. National gap shrinkage was more prevalent than within-district gap shrinkage for all subjects and racial dyads, which suggests that the achievement trends for White students increased more in district-grades with sufficient numbers of White and Black or Hispanic students than for White students nationally.

Second, most within-district gap shrinkage is modest in size. Fewer than about 5% of achievement gaps across racial dyads and subjects shrank by a half standard deviation or more. Between 21-25% of within-district gaps shrank by at least 0.10 SDs, with the exception of ELA White-Hispanic gaps (42%). Those percentages are somewhat higher for national gaps (39-56% across racial dyads and subjects). Across the district-grades in our sample, the average White-Black within-district gap widened slightly (by about 0.024 SDs), while the average White-Hispanic within-district gap narrowed slightly (by up to 0.08 SDs in ELA). National gap shrinkage shows more progress, with W-H gaps declining, on average, in both math and ELA over this period. In fact, the average White-Hispanic gap shrank by 0.12 to 0.14 SDs. We also note the significant variability around these averages: For instance, though the average White-Black ELA within-district gap grew by 0.025 SDs, the 1-99th percentile range was -0.63 to +0.69.

Third, we find that the correlation between within-district and national gap trends is between 0.46 and 0.62. In about 12% of district-grades, the achievement gap shrank within districts but the gains for Black or Hispanic students were not steep enough to reduce gaps relative to White students nationally. In 16% to 23% of district-grades, non-White subgroups gained ground on the national mean among White students while within-district racial gaps did not narrow. This approach highlights the value of measuring gaps carefully in two different ways, as both explanations and policy recommendations for gap shrinkage might vary depending on the reference group of interest.

Fourth, we find that most districts do not exhibit gap shrinkage in all six grades or both subjects. Only up to 9% of districts with any gap shrinkage (in any grade) also exhibited gap shrinkage in all six grade levels. Despite this lack of consistency across grades, within-district gap trend estimates within the same district were not random. Rather, district gaps were correlated at around 0.44 to 0.53 in adjacent grades, down to weak correlations (0.04 to 0.14) for within-district gap trends 5 grade levels apart (grade 3 vs. grade 8). On the other hand, district-grade gap trends exhibited stronger correlations across math and ELA estimates of about 0.60. The notable variability (across subjects and grades) among gap trends that belong to the same district complicates the question of what particular places might be doing to address these gaps.

While we present a comprehensive picture of achievement gap shrinkage in the U.S., our current analyses are limited in several ways. First, we cannot estimate gaps for every single district in the U.S. due to missing data and suppression of small sample sizes by SEDA. As we show in Table 1, however, we capture achievement trends for more than 65% to 73% of each non-White subgroup. Second, between approximately 50% (White subgroup) and 75% (Hispanic subgroup) of district-grades do not have achievement means in all eight years of the SEDA data, and while we address this carefully in our gap trend measures, it prevents more fine-grained analyses of timing, e.g., consistent versus rapid shrinkage over time, monotonic versus non-monotonic trends. Third, district level analyses may mask heterogeneity between schools. The gap shrinkage we see could be occurring widely across all schools in a district or concentrated in just a few, either because within-district segregation means that students of color only attend a few schools or because a few schools are very effective at gap shrinkage. SEDA recently released school-level data, though currently achievement estimates are pooled across all grades, years, and subjects. If school-level estimates are later disaggregated by year, our procedures for estimating gap trends could be applied at the school level (though data suppression may present a larger problem at the

school level). Gap shrinkage could be conceptualized as occurring within schools, within districts, and/or relative to national trends. Finally, while not a limitation of our work, we take a different approach to identifying trends in achievement gaps than other researchers using SEDA. One fruitful next step might be to understand how these different conceptualizations and measurements of gap closure are related.

This article describes the distribution of gap shrinkage in districts across the U.S. We model one way to identify school districts that have made substantial progress in shrinking ethnoracial achievement gaps. A next step in this intellectual agenda is to understand why gap shrinkage occurs. Within-district-level gap shrinkage could be attributable to two broad categories of explanations. First, gap shrinkage could occur due to demographic change. Whereas we are comparing the same ethnoracial categories over time, the characteristics of population in these categories could change in ways that are correlated with achievement. For example, we found that the White-Hispanic gap in ELA shrank in more than half of district-grades. These rapid gains in Hispanic students' language skills could be due to changing demography within the Hispanic student population in terms of immigrant generation, parent assimilation, national origin, or settlement patterns in old versus new destinations. Second, gap closure could occur due to changing conditions in the district. These conditions could either be policies or practices implemented by the district—e.g., free early childhood education—or social conditions in the district—e.g., declines in violent crime—that differentially benefit non-White students' learning. To address long-standing achievement inequality in the U.S., we must have a complete understanding of where gap shrinkage is occurring and why. Lessons from these promising districts may be scalable to produce large-scale change.

Online Appendix A. Analytic Sample

The Minimum Data Requirement (MDR) Decision

A key analytic decision for modeling gap shrinkage is to establish a minimum data quality requirement for a district-grade to be included in the analysis. We make data inclusion decisions at the district-grade level, and analyses are run separately by subject (ELA, math) and racial dyad (W-B, W-H). For sake of explication, we describe analytic sample decision-making for ELA W-B gap shrinkage, however we follow the same procedure for all four gap shrinkage estimates.

The goal in choosing an analytic sample is to balance maximum district-grade inclusion while not producing and interpreting unreliable estimates of gap shrinkage. At the district-grade level, the primary source of unreliability in these estimates comes from missing any of the up to 8 years of achievement scores (2009 - 2016) for a given subgroup, which are used to establish the subgroup's linear achievement trend during the panel. To estimate within-district gap shrinkage, we then compare the two subgroups' trends to one another. If one (or both) of those two groups only possesses a handful of achievement estimates across the 8-year panel, the linear trend may be noisy and susceptible to influence from any outlier year. In other words, the lower the number of yearly observations, the less precise the estimated trend will be. In addition, if one of the two subgroups does not possess achievement scores in the very first and last year of the panel, we may be hesitant to make conclusions about how the W-B gap in 2009 compares to the W-B gap in 2016. We refer to 2009 and 2016 as the "anchor years" in the sense that they help pin down any estimates of gap trends during the panel time frame. In other words, the greater the spread in the year predictor, the more precise the estimated trend will be.

In Table A1 we show number and the percentage of district-grades that have exactly 1 year, 2 years, etc., up to 8 years with both achievement means. Among district-grades that have at least one annual achievement mean estimate in the given subject for the given subject (the denominator

of the column percentages), it is most common for the district-grade to possess an estimate in all 8 years. For instance, 49.5% of district-grades with at least one ELA achievement estimate for White students in fact have all 8. However, the other half of district-grades have less than the full set of 8 years. Perhaps one should not be concerned about estimating a trend when, for instance, 7 of 8 years is available. However, 21.3% of district-grades have achievement estimates in 5 or fewer years. Trends based on fewer years may be more susceptible to the influence of unusual years.

[Insert Table A1 about Here]

It follows that the availability of annual achievement means is less consistent for Black and Hispanic subgroups, since these groups of students are simply smaller and therefore may not have a reported mean achievement estimate in SEDA due to the suppression of estimates based on fewer than 20 students. For instance, only 27.3% of all district-grades with at least one ELA mean achievement estimate for Hispanic students have estimates in all 8 years. Table A1 highlights the decision the researcher must make with this information to define minimum data requirements (MDR) for inclusion in the analysis. Moreover, it is important to justify what primary definition is used and how this choice affects overall findings.

Seven Different Options for MDR Definition

Again, using the example of estimating W-B gap in ELA, a given district-grade must possess at least one year with both a White and Black mean achievement estimate, at an absolute minimum, to be considered for inclusion. We can think of this as the total possible population of district-grades for which we can estimate gaps (e.g., 16,063 district-grades for ELA W-B gap).

We conducted all analyses on seven different MDR definitions, some of which are more restrictive (fewer districts receive estimates, but those estimates are more reliable) and some of which are more inclusive (more districts receive estimates, but not all estimates will be reliable).

The six MDR definitions are summarized in Table A2, along with the seventh option of specifying no MDR.

[Insert Table A2 about Here]

The MDR definitions are ordered from most restrictive (MDR definition #1) to least restrictive (definition #7), with the percent of district-grades that meet the MDR. In our most restrictive MDR, only between 42 and 57% of district-grades will receive gap shrinkage estimates, depending on the subject and racial dyad. The definitions are based on combinations of several factors including, the number of the 8 possible years a district-grade possess, whether the subgroups possess achievement estimates in the "anchor years" specifically (2009 and 2016), or a minimum number of years between the first and last achievement estimate. Examining seven different MDR definitions allowed us to make an informed decision about which MDR definition to select as our primary specification and also examine the sensitivity of the results to the choice of MDR definition.

Choosing a Preferred MDR Definition

In Table A3, we present basic results using each of the seven MDR definitions. This includes the percent of district-grades that meet the MDR definition, the percentage of the non-White group nationally that are in these district-grades, and basic descriptive statistics about the distribution of within-district gap trend estimates produced using the given MDR sample.

[Insert Table A3 about Here]

We chose MDR definition #2 (column 2 of Table A3) to construct our primary analytic sample in the main narrative. We did so for several reasons: First, this is a relatively restrictive MDR definition (e.g., for ELA W-B gap shrinkage, 63% of possible district-grades meet this MDR), and since we ultimately hope to identify places that exhibit truly-promising within-district gap shrinkage, we prefer to err on the side of caution. Second, despite being a fairly restrictive

definition, most (73%) of the nation's Black public students attend schools in these district-grades (and 72% of Hispanic students). Third, we conducted extensive visual inspections of estimated group-specific trends in achievement superimposed on a scatterplot of raw SEDA achievement scores for hundreds of individual district-grades and found that, in the MDR definition #2 sample, the trends generally matched the achievement data well. As we used increasingly inclusive MDR definitions, these visual inspections often revealed linear trends that were unusually extreme due to the influence of individual data points.

To illustrate these issues, Appendix Figures A1 and A2 each show one example district's ELA achievement estimates and trends for White and Black students, separately by grade (OLS-based trend estimates shown in solid lines). In Figure A1, the example district does not meet any MDR definition. We can see that the within-district gap shrinkage estimates are often implausibly large (e.g., 2 SDs in grade 6) because they are based on too few data points and/or there is insufficient data toward the end of the panel. See, for example, that the trend in Black achievement in grade 3 is based on only data from 2009 and 2012. Take as another example grade 4; this district's OLS-based gap shrinkage estimate is -1.04 (see lower right text in grade 4 illustration). This shows greater gap shrinkage than any district-grade show in our main results (see Figure 2). The extreme gap shrinkage in grade 4 may be driven by the 2009 achievement estimate for Black students (-0.49), which is exerting undue influence on the Black achievement trend since there is no data for Black students in grade 4 after 2013. Several of the grades in Appendix Figures A1 illustrate why we hesitate to conclude that gaps have truly closed without imposing an MDR.

[Insert Figure A1 about Here]

In Appendix Figure A2, this example district meets our preferred MDR definition #2 in all six grades. Here, we see that within-district gap trend estimates (reported in lower right corner of each graph space) are generally smaller in magnitude (the largest is 0.24 SDs in grade 6), and they

follow the general pattern of the data much more closely. In a district-grade with sufficient data, are more comfortable making conclusions about the direction of gap shrinkage during this time period.

[Insert Figure A2 about Here]

Returning to Appendix Table A3, we can now compare the descriptive findings with respect to within-district gap trend estimates across the seven MDR definitions. In fact, the mean of within-district gap trend estimates is relatively consistent. The percentage of district-grades that are identified as exhibiting within-district gap shrinkage (i.e., equivalent to row 4 of Table 2 in the main narrative) is also relatively consistent. For instance, for W-B gaps in ELA, that percentage is always between 36 and 38%. As expected, however, the standard deviations (or 1st-99th percentile ranges) are quite different due to how the MDRs address outlier trends. For instance, with MDR definition #1, the 1st-99th range of estimates is from -0.60 to +0.68. In other words, almost all district-grades exhibit absolute changes in gap sizes between at most 0.60 to 0.68 SDs. Practically-speaking, this actually seems like a range of relatively large shifts in the size of a gap over an 8 year period. However, as we included more district-grades with fewer annual achievement scores (moving from the most restrictive MDR #1 to least restrictive MDR #6), this range expanded significantly. Using the least restrictive MDR definition #6 (column 6), we observe a 1st-99th percentile range of estimates from -1.53 to +1.68. The benefit, of course, of MDR definition #6 is the number of district-grades that receive an estimate (93%).

The difference in the 1st-99th percentile range across MDR definitions is most dramatic for W-H gap shrinkage estimates in math (fourth panel of Table A3). While the most restrictive MDR definition #1 yields a set of gap trend estimates between -0.62 and +0.59 (1st-99th percentile range), the least restrictive MDR definition #6 yields a 1st-99th percentile range of -2.34 to +2.40 SDs—more than three times the range of possible gap size changes over time than MDR definition #1.

However, when we visually examined district-grades with gap trend estimates larger in magnitude than +/- 0.70 standard deviations, we almost always find these are cases with less annual achievement data available with influential data points possibly skewing the trend estimates. We generally do not think these district-grades have sufficient evidence of how their gaps are trending during this period.

We ultimately decided that imposing a thoughtful MDR definition was the best approach to addressing the fact that some districts have fewer annual datapoints and therefore noisy gap trend estimates. We considered additional or complementary approaches, as well. For instance, we considered using an empirical Bayes approach to estimating coefficients in our models, and we considered precision-weighting the individual achievement estimates. The first proved untenable (see section 2 of Online Appendix B), and the latter proved to have little effect on our estimates (see section 3 of Online Appendix B). In our primary specification, then, we only use MDR Definition #2 to constrain our OLS-based estimates of gap trends.

Online Appendix B. Technical Appendix

In Appendix B, we expand upon three technical decisions made for the primary analyses for the interested reader. This includes a more in-depth description of: (1) how we re-scaled the SEDA achievement estimates to yield regression coefficients that capture *within-district* versus *national* gap trends; (2) our decision to use slope coefficients from traditional OLS models rather than empirical-Bayes shrank slope coefficients from a multilevel random effects model; and (3) the finding that the choice of precision-weighting achievement estimates had little impact on results.

(1) Scaling SEDA Achievement Estimates for Analysis.

SEDA Approach to Scaling. The scaling of the district-grade-year mean achievement estimates is central to our interpretation of slope coefficients, which in turn yield our estimates of both within-district and national gap trends, each of which required a different approach to standardizing mean achievement estimates. Because all standardization occurs within subject-grade, one can simplify the consideration of scaling mean achievement estimates by thinking of only one grade and subject (e.g., grade 4 math). This serves as a more detailed discussion of the comparison made in Figure 1 of the main narrative.

The SEDA mean achievement estimates that are provided publicly are expressed in what Fahle et al. (2019) refer to as the cohort-standardized (CS) metric. Once all achievement outcomes had been placed on the NAEP scale, those NAEP scores were standardized relative to a reference cohort of students. More specifically, the CS metric was produced by "subtracting the national grade-subject-specific mean and dividing by the national grade-subject-specific standard deviation for a reference cohort" (pg. 21-22). Fahle et al. (2019) opt to use the average of the three national cohorts that were in 4th grade in 2009, 2011, and 2013. As a result, estimated means reported in

the CS-scale can be interpreted "as an effect size relative to the average of the three national cohorts that were in 4th grade in 2009, 2011, and 2013" (pg. 22).

For our purposes, however, the CS-scale is not ideal, because we do not want a mean achievement estimate of 0 to represent a "moving target" over time for a blended cohort. We therefore first un-standardize the CS-scale achievement estimates to return them to their original NAEP scale. The relevant means and standard deviations used to create the CS scale are provided in Table 6 of Fahle et al. (2019). We then re-standardize the NAEP mean achievement estimates in two different ways, one for within-district and one for national gap trends. For sake of explication, we focus on W-B gap trends, however the same logic applies to W-H gap trends.

Y09-Scale for Within-District Gap Trends. When estimating within-district gap trends, we are focusing on comparing a given district's White and Black achievement trends over time, relative to one another. For within-district gap shrinkage, we therefore standardize mean NAEP achievement estimates (at the grade (g)– year (y)– subject (b)– subgroup (r) level) relative to the national mean NAEP score (across all students) in 2009 for a given subject and grade.[17] We refer to this as the year-2009-standardized (Y09)-scale. On the Y09-scale, a mean achievement estimate of 0 in any given year indicates that the given racial group in that district is performing at the same level as all students nationally were in 2009. A positive trend over time (the coefficient on the linear year predictor) in Y09-scaled estimates for a given district's Black students indicates that mean NAEP performance is increasing among the district's Black students *in an absolute sense*. If the year-slope for Black students is positive *and* the difference in year-slopes between White and Black students is negative, it indicates that the W-B gap within the district has shrunk, because the district's Black student mean achievement improved over time in an absolute sense (that is,

[17] We retrieve national average scale scores and standard deviations for grade 4 and 8 mathematics and reading for all students among public schools in 2009, 2011, 2013, 2015, and 2017 from the NAEP Data Explorer (NAEP Data Explorer, 2009, 2011, 2013, 2015, 2017). We follow the procedures outlined by Fahle et al. (2019) on pp. 20-21 to interpolate and extrapolate linearly to obtain mean achievement estimates for grades 3, 5, 6, and 7, in which NAEP is not administered.

not because Black students are *losing ground less quickly* in a relative sense to some other group over time). In Appendix Table B1, we document the NAEP achievement scores we used from Fahle et al. (2019), as well as the interpolated NAEP scores for non-tested grades in 2009.

[Insert Table B1 about Here]

WS-Scale for National Gap Trends. When estimating national gap shrinkage estimates, we are focusing on how the district's own trend in mean achievement scores for, say, Black students compares to the national trend in achievement scores for White students (in the same grade and subject). When estimating national W-B gap trends, we therefore standardize mean NAEP achievement estimates (at the g-y-b-r level) relative to the nationwide mean achievement *of White students* in the same year (and grade and subject).[18] We refer to this as the White-subgroup standardized (WS)-scale. On the WS-scale, a mean Black achievement estimate of 0 in any given year would indicate that Black students in the district are performing as well as White students nationally. A positive trend over time (the coefficient on the linear year predictor) in WS-scaled estimates for a given district's Black students indicates that mean NAEP performance among the district's Black students is getting closer to White student achievement nationally in a relative sense. In other words, a positive year-slope for Black students on the WS-scale captures national W-B gap shrinkage. In Appendix Table B2, we document the NAEP mean achievement scores for White students used to standardize, which were retrieved from the online NAEP Data Explorer (2009, 2011, 2013, 2015, 2017), as well as the NAEP scores we imputed for White students in non-tested grades and years, following the imputation practices of Fahle et al. (2019) used in SEDA, described on pp. 20-21.

[18] We retrieve national average scale scores and standard deviations for grade 4 and 8 mathematics and reading for White, students in public schools in 2009, 2011, 2013, 2015, and 2017 from the NAEP Data Explorer (2009, 2011, 2013, 2015, 2017). We follow the procedures outlined by Fahle et al. (2019) on pp. 20-21 to interpolate and extrapolate linearly to obtain mean achievement estimates for White students in grades 3, 5, 6, and 7 and even school years, in which NAEP is not administered. The Data Explorer notes that Black includes African American, Hispanic includes Latino, and Pacific Islander includes Native Hawaiian. Race categories exclude Hispanic origin. Prior to 2011, students in the "two or more races" category were categorized as "unclassified."

[Insert Table B2 about Here]

In Appendix Figure B1, we present boxplots that capture the median and spread in these three different scalings of the NAEP scores over time and separately by subgroup (for grade 4 ELA only). Recall that the SEDA population may differ to some extent from the NAEP population. The three scalings of the NAEP scores are highly correlated with one another (above 0.99). However, these re-scalings ensure that interpretations of coefficients are aligned with our efforts to measure within-district and national gap shrinkage.

[Insert Figure B1 about Here]

(2) OLS versus Empirical-Bayes Shrunk (EBS) Estimates

We next compare results of analyses from two possible modeling approaches—OLS or EBS—using as an example, grade 6 ELA W-B achievement gap trends. As a reminder, all models are executed separately for each combination of grade, subject, and racial dyad.

In the first approach we considered, we used a three-level model that nested the up to 16 subgroup-year specific achievement estimates (L1) within districts (L2), which were in turn nested in states (L3). See Appendix Figure B2 for an overview of this model. This random-effects model yields district-specific (L2) empirical-Bayes (EB) estimates of gap shrinkage, under the assumption that those estimates come from a normal distribution.

[Insert Figure B2 about Here]

We initially pursued this approach because we were concerned that OLS regression results from a single district-grade would be too noisy (more on this below). The EBS approach, in principal, produces district-specific gap trend estimates that are a weighted combination of both the findings for the given district, and—when those findings are less reliable—the overall grand mean of gap trends. In essence, this approach partially relies on evidence from the population to make better estimates for any given district. The extent to which a given district's gap trend

estimates are weighted toward the district information versus the population information is determined by the reliability of the district-specific estimates. That is, the less reliable the district-specific information, the more the EBS estimates will look like the grand mean. This approach may be particularly useful when the individual district estimates are not the focus of the inquiry. The EBS approach is often efficient but can be overly conservative (Raudenbush & Bryk, 2002).

In the second approach, we used a (non-nested) OLS regression model that takes on the same right-hand structure as L1 of the HLM model. However, we now execute a separate OLS model for each district-grade (in a given subject and for a given racial dyad). Each OLS model could therefore have at most 16 observations (8 mean achievement scores for the White subgroup and 8 mean achievement scores for the Black (or Hispanic) subgroup). Because each model has so few observations, the coefficients are somewhat noisy. A potential advantage of the EBS approach over the OLS approach, would have been that the estimates are less noisy since they "borrow" information from the grand mean.

In practice, however we found that the district-grade specific EB estimates were *over*-shrunk toward the grand mean. When making a visual inspection of the data and the regression lines fit by the HLM or OLS model, we found that the EB estimates often did not reflect the patterns of achievement data observed in a specific district-grade, but rather the overall mean across all districts. For instance, in Appendix Figure B3, we present one example district's W-B gap trend estimates in ELA, separately by grade level; the OLS-based estimates are shown with solid lines and EBS estimates shown with dashed lines. Here one can see that, in most of the grades, the OLS-based estimate of this district's gap trend is negative (reported in lower right corner of each graph space), while the EBS estimates shown directly above the OLS estimates are positive and closer to zero. Moreover, the EBS-estimated trends to not closely follow the data in the given

district-grade—that is, the look quite similar to one another regardless of grade level (see especially grade 7 Black achievement trend).

[Insert Figure B3 about Here]

As a result of the over-shrinkage, the overall distribution of EBS-based gap trends across all district-grades was implausibly narrow, and most estimates fell just on either side of 0. This primarily reflects the fact that almost *any* longitudinal regression model run with only 16 observations will produce coefficients with large standard errors (in the HLM setting, low reliability) and thus the EBS approach will heavily weight the grand mean estimates.

While the EBS approach proved untenable, we also observed that OLS estimates often did not reflect the underlying achievement data for a given district-grade if that given district-grade was missing a higher proportion of the up to 8 yearly achievement scores. In these cases, we concluded there was insufficient information to assess whether a given gap had been widening or closing over time. We therefore tempered the noisiness of the OLS estimates by imposing a minimum data requirement (MDR) for inclusion in the analytic sample. For a complete discussion of analytic decisions regarding the MDR definition, see Online Appendix A.

In Appendix Figure B4, we compare the median and variability of (a) the OLS-estimated gap trends using the MDR, (b) the EBS-estimated gap trends, (c) and a non-parametric gap trend estimate we calculated by subtracting the raw difference in White and Black achievement means in 2009 from the raw difference in White and Black achievement means in 2016. These boxplots quickly illustrate just how shrunk the EB estimates of within-district gap shrinkage are.

[Insert Figure B4 about Here]

(3) Precision-Weighting

The SEDA dataset provides a standard error for each mean achievement estimate at the grade-year-subject-subgroup-district level. In our primary models presented in the main tables, we

opt to precision-weight the achievement estimates when running our regression models. We also re-produced our results without precision-weighting. In Appendix Figure B4, we compare gap trend estimates produced both with and without the precision weighting, as well as the correlation between the two. For ELA W-H gap trends (upper right panel), the SD of the estimates produced with the precision weights is 0.25, the SD was 0.26 without the weights, and the correlation between the two is 0.994. These results suggest that our results are not very sensitive to the decision regarding precision-weighting.

[Insert Figure B5 about Here]

References

Allport, G. W., Clark, K., & Pettigrew, T. (1954). The nature of prejudice.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50*(2), 204-226.

Briggs, D. C., & Dadey, N. (2015). Making sense of common test items that do not get easier over time: Implications for vertical scale designs. *Educational Assessment, 20*(1), 1-22.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*(6), 551-576.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

Carter, P. L., & Welner, K. G. (2013). *Closing the opportunity gap: What America must do to give every child an even chance*: Oxford University Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633-2679.

Currie, J., & Thomas, D. (2001). Early test scores, school quality and SES: Longrun effects on wage and employment outcomes. *Research in Labor Economics, 20*, 103-132.

Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., Belen, C., & Ho, A. D. (2019). *Technical Documentation Version 3.0*. Retrieved from Stanford, CA: https://edopportunity.org/papers/SEDA_documentation_v30_DRAFT09212019.pdf

Ferguson, R. (1998). Test Score Trends along Racial Lines, 1971 to 1996: Popular Culture and Community Academic Standards. 348-390. America Becoming: Racial Trends and Their Consequences, vol. 1. National Research Council: Washington, DC, National Academy Press.

Grissmer, D., Flanagan, A., & Williamson, S. (1998). Why did the Black–White score gap narrow in the 1970s and 1980s?

Hedges, L. V., & Nowell, A. (1998). Black-White test score convergence since 1965 *The Black White Test Score Gap*: The Brookings Institution.

Hedges, L. V., & Nowell, A. (1999). Changes in the black-white gap in achievement test scores. *Sociology of Education*, 111-135.

Jencks, C., & Phillips, M. (1998). *The black-white test score gap*: Brookings Institution Press.

Moody, J. (2001). Race, school integration, and friendship segregation in America. *American Journal of Sociology, 107*(3), 679-716.

NAEP Data Explorer. (2009, 2011, 2013, 2015, 2017). *National Assessment of Educational Progress (NAEP), 2009, 2011, 2013, 2015, and 2017 Mathematics and Reading Assessments*. Retrieved from: https://www.nationsreportcard.gov/ndecore/xplore/NDE

NCES. (2013). *The Nation's Report Card: Trends in Academic Progress 2012. NCES 2013-456*. National Center for Education Statistics: ERIC Clearinghouse.

NCES. (2015). *The Nation's Report Card: 2015 Mathematics and Reading Assessments*. Retrieved from https://www.nationsreportcard.gov/reading_math_2015/#reading?grade=4

Neal, D. (2005). Why Has Black-White Skill Convergence Stopped? NBER Working Paper No. 11090. *National Bureau of Economic Research*.

Quinn, D. M. (2014). Black-White Summer Learning Gaps: Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis*. doi:10.3102/0162373714534522

Quinn, D. M., & McIntyre, J. (2017). Do learning rates differ by race/ethnicity over kindergarten? Reconciling results across gain score, first-difference, and random effects models. *Economics of Education Review, 59*, 81-86.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1): Sage Publications, Inc.

Reardon, S. F. (2008). Thirteen ways of looking at the black-white test score gap. *Stanford Institute for Research on Education Policy & Practice, Working Paper, 8*.

Reardon, S. F. (2019). Educational Opportunity in early and middle childhood: Using full population administrative data to study variation by place and age. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 5*(2), 40-68.

Reardon, S. F., Cimpian, J., & Weathers, E. S. (2014). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps *Handbook of Research in Education Finance and Policy, Second Edition* (pp. 491-509): Taylor and Francis.

Reardon, S. F., & Hinze-Pifer, R. (2017). Test Score Growth among Chicago Public School Students, 2009-2014. *Stanford Center for Education Policy Analysis*.

Reardon, S. F., Kalogrides, D., & Shores, K. (2019). The geography of racial/ethnic test score gaps. *American Journal of Sociology, 124*(4), 1164-1221.

Reardon, S. F., Papay, J. P., Kilbride, T., Strunk, K. O., Cowen, J., An, L., & Donohue, K. (2019). *Can Repeated Aggregate Cross-Sectional Data Be Used to Measure Average Student Learning Rates? A Validation Study of Learning Rate Measures in the Stanford Education Data Archive*. Stanford. Stanford Center for Education Policy Analysis. Retrieved from http://cepa.stanford.edu/wp19-08

Reardon, S. F., Valentino, R. A., Kalogrides, D., Shores, K. A., & Greenberg, E. H. (2013). Patterns and trends in racial academic achievement gaps among states, 1999-2011. *Unpublished Working Paper. Center for Education Policy Analysis, Stanford University*.

Reardon, S. F., Weathers, E. S., Fahle, E., Jang, H., & Kalogrides, D. (2019). *Is separate still unequal? New evidence on school segregation and racial academic achievement gaps*. Retrieved from

Tyson, K., Darity Jr, W., & Castellino, D. R. (2005). It's not "a black thing": Understanding the burden of acting white and other dilemmas of high achievement. *American Sociological Review, 70*(4), 582-605.

von Hippel, P. T., & Hamrock, C. (2019). Do test score gaps grow before, during, or between the school years? Measurement artifacts and what we can know in spite of them. *Sociological Science, 6*, 43-80.

Tables

Table 1. Analytic Sample: SEDA Observations with Sufficient Data for Gap Trend Analysis at Various Units of Analysis

	White-Black Gap Analysis		White-Hispanic Gap Analysis	
	ELA	Math	ELA	Math
1. DISTRICT-GRADE-YEAR LEVEL (N=476462)				
A. Total # of D-G-Y's in Dataset	476,462 *(100%)*	451,000 *(100%)*	476,462 *(100%)*	451,000 *(100%)*
B. D-G-Y's w/ At Least 1 of 2 Group Means	437,266 *(92%)*	415,091 *(92%)*	442,349 *(93%)*	418,594 *(93%)*
C. D-G-Y's that Have Both Group Means	97,403 *(20%)*	90,016 *(20%)*	117,206 *(25%)*	108,580 *(24%)*
D. % of Non-Wht Group Observed in D-G-Y's in	86.6%	86.7%	84.9%	85.6%
2. DISTRICT-GRADE LEVEL (N=69373)				
A. D-G's w/ At Least 1 Year w/ Both Means	16,043 *(100%)*	15,508 *(100%)*	21,088 *(100%)*	20,654 *(100%)*
B. Among (2A), D-G's that Meet MDR	10,046 *(63%)*	8,987 *(58%)*	10,872 *(52%)*	9,589 *(46%)*
C. % of Non-Wht Group Observed in D-G's in (2B)	72.7%	72.1%	68.9%	65.4%
3. DISTRICT LEVEL (N=12052)				
A. D's w/ At Least 1 Grade that Meets MDR	2,048 *(100%)*	2,018 *(100%)*	2,370 *(100%)*	2,344 *(100%)*
B. Among (3A), D's w/ >= 2 Grades that Meet MDR	1,863 *(91%)*	1,823 *(90%)*	2,075 *(88%)*	2,054 *(88%)*
C. Among (3A), D's w/ >= 3 Grades that Meet MDR	1,760 *(86%)*	1,629 *(81%)*	1,910 *(81%)*	1,825 *(78%)*
D. Among (3A), D's w/ >= 4 Grades that Meet MDR	1,593 *(78%)*	1,487 *(74%)*	1,735 *(73%)*	1,637 *(70%)*
E. Among (3A), D's w/ >= 5 Grades that Meet MDR	1,475 *(72%)*	1,133 *(56%)*	1,522 *(64%)*	995 *(42%)*
F. Among (3A), D's w/ All 6 Grades Meet MDR	1,307 *(64%)*	897 *(44%)*	1,260 *(53%)*	734 *(31%)*

Footnote: D = District; G = Grade, Y = Year. MDR = Minimum Data Requirements (see pg. 16).

Table 2. Cumulative Criteria for Prevalence of Within-District and National W-B and W-H Gap Shrinkage, by Subject

	White-Black Gap Analysis				White-Hispanic Gap Analysis			
	ELA		Math		ELA		Math	
Within-District Gap Trend Estimates (Y09-scaled)								
D-G's that Meet MDR	10,046	*(100%)*	8,987	*(100%)*	10,872	*(100%)*	9,589	*(100%)*
+ Non-Wht Group Initially Performs Lower than Wht Group	10,003	*(100%)*	8,958	*(100%)*	10,765	*(99%)*	9,372	*(98%)*
+ Non-Wht Group Mean Increases Across Panel*	5,932	*(59%)*	4,335	*(48%)*	8,207	*(75%)*	5,243	*(55%)*
+ Gap Shrinks (Estimate is Negative)	3,704	*(37%)*	2,715	*(30%)*	6,184	*(57%)*	3,451	*(36%)*
+ Gap Shrinks by At Least 0.10 SDs	2,500	*(25%)*	1,912	*(21%)*	4,533	*(42%)*	2,351	*(25%)*
+ Gap Shrinks by At Least 0.20 SDs	1,515	*(15%)*	1,203	*(13%)*	2,905	*(27%)*	1,445	*(15%)*
+ Gap Shrinks by At Least 0.30 SDs	856	*(9%)*	690	*(8%)*	1,697	*(16%)*	820	*(9%)*
+ Gap Shrinks by At Least 0.40 SDs	454	*(5%)*	394	*(4%)*	932	*(9%)*	423	*(4%)*
+ Gap Shrinks by At Least 0.50 SDs	219	*(2%)*	207	*(2%)*	492	*(5%)*	230	*(2%)*
National Gap Trend Estimates (WS-scaled)								
D-G's that Meet MDR	10,046	*(100%)*	8,987	*(100%)*	10,872	*(100%)*	9,589	*(100%)*
+ Gap Shrinks (Estimate is Negative)	5,257	*(52%)*	5,175	*(58%)*	7,535	*(69%)*	6,052	*(63%)*
+ Gap Shrinks by At Least 0.10 SDs	3,915	*(39%)*	4,156	*(46%)*	6,081	*(56%)*	4,949	*(52%)*
+ Gap Shrinks by At Least 0.20 SDs	2,683	*(27%)*	3,142	*(35%)*	4,516	*(42%)*	3,830	*(40%)*
+ Gap Shrinks by At Least 0.30 SDs	1,762	*(18%)*	2,244	*(25%)*	3,125	*(29%)*	2,833	*(30%)*
+ Gap Shrinks by At Least 0.40 SDs	1,097	*(11%)*	1,582	*(18%)*	1,997	*(19%)*	1,981	*(21%)*
+ Gap Shrinks by At Least 0.50 SDs	644	*(6%)*	1,097	*(12%)*	1,202	*(11%)*	1,355	*(14%)*

*Footnote: D = District; G = Grade, Y = Year. MDR = Minimum Data Requirements (see pg. 16).Y09-scaled = achievement outcomes standardized around national mean achievement in 2009. WS-scaled= achievement outcomes standardized around the national mean achievement of White students in the given year. Estimates are produced via OLS with precision weighted outcomes. * We make a substantive choice to focus on cases where outcomes are improving for the non-White subgroup. This requirement excludes between 19% (ELA) and 35% (math) of district-grades with White-Black gap shrinkage, and between 12% (ELA) and 30% (math) of district-grades with White-Hispanic shrinkage.*

Table 3. White-Black Gaps: Number of Districts with Consistent Within-District Gap Shrinkage (of Given Magnitudes) in as Few as One Grade and Up to Six Grades, by Subject

	Districts with at Least 1 Grade in which the Within-District Gap…			
	…Shrinks (Estimate <0)	…Shrinks by 0.10+ SDs	…Shrinks by 0.20+ SDs	…Shrinks by 0.30+ SDs
# of Districts with…		**ELA**		
…Exactly 1 Grade w/ Shrinkage of This Size	461	541	503	385
	(30.8)	(44.4)	(56.7)	(66.5)
…Exactly 2 Grades w/ Shrinkage of This Size	414	320	221	132
	(27.7)	(26.3)	(24.9)	(22.8)
…Exactly 3 Grades w/ Shrinkage of This Size	292	192	100	41
	(19.5)	(15.8)	(11.3)	(7.1)
…Exactly 4 Grades w/ Shrinkage of This Size	158	102	46	21
	(10.6)	(8.4)	(5.2)	(3.6)
…Exactly 5 Grades w/ Shrinkage of This Size	113	49	16	0
	(7.6)	(4.0)	(1.8)	(0.0)
…Exactly 6 Grades w/ Shrinkage of This Size	57	15	1	0
	(3.8)	(1.2)	(0.1)	(0.0)
Column Total # of Districts	1495	1219	887	579
(Column Percentage)	(100.0)	(100.0)	(100.0)	(100.0)
# of Districts with…		**Math**		
…Exactly 1 Grade w/ Shrinkage of This Size	549	545	493	375
	(41.9)	(51.1)	(63.3)	(74.1)
…Exactly 2 Grades w/ Shrinkage of This Size	366	304	193	90
	(27.9)	(28.5)	(24.8)	(17.8)
…Exactly 3 Grades w/ Shrinkage of This Size	215	134	56	30
	(16.4)	(12.6)	(7.2)	(5.9)
…Exactly 4 Grades w/ Shrinkage of This Size	121	61	30	10
	(9.2)	(5.7)	(3.9)	(2.0)
…Exactly 5 Grades w/ Shrinkage of This Size	49	19	6	1
	(3.7)	(1.8)	(0.8)	(0.2)
…Exactly 6 Grades w/ Shrinkage of This Size	10	3	1	0
	(0.8)	(0.3)	(0.1)	(0.0)
Column Total # of Districts	1310	1066	779	506
(Column Percentage)	(100.0)	(100.0)	(100.0)	(100.0)

Footnote: Table presents within-district (as opposed to national) gap trend estimates; achievement outcomes are therefore Y09-scaled (standardized around national mean achievement in 2009). Analyses restricted to district-grades that meet the MDR (minimum data requirement). Estimates produced via OLS with precision weighted outcomes. Each column is limited to districts with at least one grade in which gaps shrink by the amount indicated at the top of the column. For example, column 2 for ELA is limited to the 2,500 districts where, in at least one grade, gaps shrink by 0.10+ SDs (i.e., 0.10 or more SDs).

Table 4. White-Hispanic Gaps: Number of Districts with Consistent Within-District Gap Shrinkage (of Given Magnitudes) in as Few as One Grade and Up to Six Grades, by Subject

	Districts with at Least 1 Grade in which the Within-District Gap…			
	…Shrinks (Estimate <0)	…Shrinks by 0.10+ SDs	…Shrinks by 0.20+ SDs	…Shrinks by 0.30+ SDs
# of Districts with…		**ELA**		
…Exactly 1 Grade w/ Shrinkage of This Size	456	590	629	572
	(22.4)	(32.0)	(43.5)	(56.1)
…Exactly 2 Grades w/ Shrinkage of This Size	414	478	412	286
	(20.4)	(26.0)	(28.5)	(28.1)
…Exactly 3 Grades w/ Shrinkage of This Size	366	356	240	111
	(18.0)	(19.3)	(16.6)	(10.9)
…Exactly 4 Grades w/ Shrinkage of This Size	367	230	111	34
	(18.1)	(12.5)	(7.7)	(3.3)
…Exactly 5 Grades w/ Shrinkage of This Size	246	123	36	12
	(12.1)	(6.7)	(2.5)	(1.2)
…Exactly 6 Grades w/ Shrinkage of This Size	184	64	18	4
	(9.1)	(3.5)	(1.2)	(0.4)
Column Total # of Districts	2033	1841	1446	1019
(Column Percentage)	(100.0)	(100.0)	(100.0)	(100.0)
# of Districts with…		**Math**		
…Exactly 1 Grade w/ Shrinkage of This Size	642	632	552	437
	(40.2)	(49.3)	(60.6)	(73.3)
…Exactly 2 Grades w/ Shrinkage of This Size	425	371	235	112
	(26.6)	(29.0)	(25.8)	(18.8)
…Exactly 3 Grades w/ Shrinkage of This Size	286	171	87	35
	(17.9)	(13.3)	(9.5)	(5.9)
…Exactly 4 Grades w/ Shrinkage of This Size	152	77	26	8
	(9.5)	(6.0)	(2.9)	(1.3)
…Exactly 5 Grades w/ Shrinkage of This Size	65	24	8	2
	(4.1)	(1.9)	(0.9)	(0.3)
…Exactly 6 Grades w/ Shrinkage of This Size	28	6	3	2
	(1.8)	(0.5)	(0.3)	(0.3)
Column Total # of Districts	1598	1281	911	596
(Column Percentage)	(100.0)	(100.0)	(100.0)	(100.0)

Footnote: Table presents within-district (as opposed to national) gap trend estimates; achievement outcomes are therefore Y09-scaled (standardized around national mean achievement in 2009). Analyses restricted to district-grades that meet the MDR (minimum data requirement). Estimates produced via OLS with precision weighted outcomes. Each column is limited to districts with at least one grade in which gaps shrink by the amount indicated at the top of the column. For example, column 2 for ELA is limited to the 4,533 districts where, in at least one grade, gaps shrink by 0.10+ SDs (i.e., 0.10 or more SDs).

Table 5. Within-District Gap Trend Means and Correlations of Estimates across Grades within the same District, by Racial Dyad and Subject

	Mean Gap Trend	Gap Trend Correlation between Row Grade and…					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
				White-Black Gap, ELA			
Grade 3	0.002	1.00					
Grade 4	-0.003	0.48	1.00				
Grade 5	0.041	0.37	0.53	1.00			
Grade 6	0.034	0.22	0.31	0.44	1.00		
Grade 7	0.046	0.17	0.21	0.30	0.44	1.00	
Grade 8	0.032	0.14	0.17	0.17	0.29	0.47	1.00
				White-Black Gap, Math			
Grade 3	-0.032	1.00					
Grade 4	0.027	0.49	1.00				
Grade 5	0.031	0.30	0.46	1.00			
Grade 6	0.043	0.22	0.31	0.45	1.00		
Grade 7	0.057	0.15	0.22	0.32	0.47	1.00	
Grade 8	0.025	0.04	0.10	0.16	0.28	0.49	1.00
				White-Hispanic Gap, ELA			
Grade 3	-0.085	1.00					
Grade 4	-0.076	0.50	1.00				
Grade 5	-0.076	0.33	0.47	1.00			
Grade 6	-0.079	0.16	0.27	0.47	1.00		
Grade 7	-0.083	0.09	0.18	0.30	0.50	1.00	
Grade 8	-0.092	0.04	0.09	0.13	0.27	0.45	1.00
				White-Hispanic Gap, Math			
Grade 3	-0.032	1.00					
Grade 4	0.004	0.45	1.00				
Grade 5	-0.001	0.27	0.40	1.00			
Grade 6	0.003	0.17	0.26	0.46	1.00		
Grade 7	0.014	0.12	0.21	0.34	0.49	1.00	
Grade 8	-0.035	0.07	0.12	0.19	0.31	0.44	1.00

Footnote: Table presents within-district (as opposed to national) gap trend estimates; achievement outcomes are therefore Y09-scaled (standardized around national mean achievement in 2009). Analyses restricted to district-grades that meet the MDR (minimum data requirement). Estimates produced via OLS with precision weighted outcomes, and models include district fixed effects.

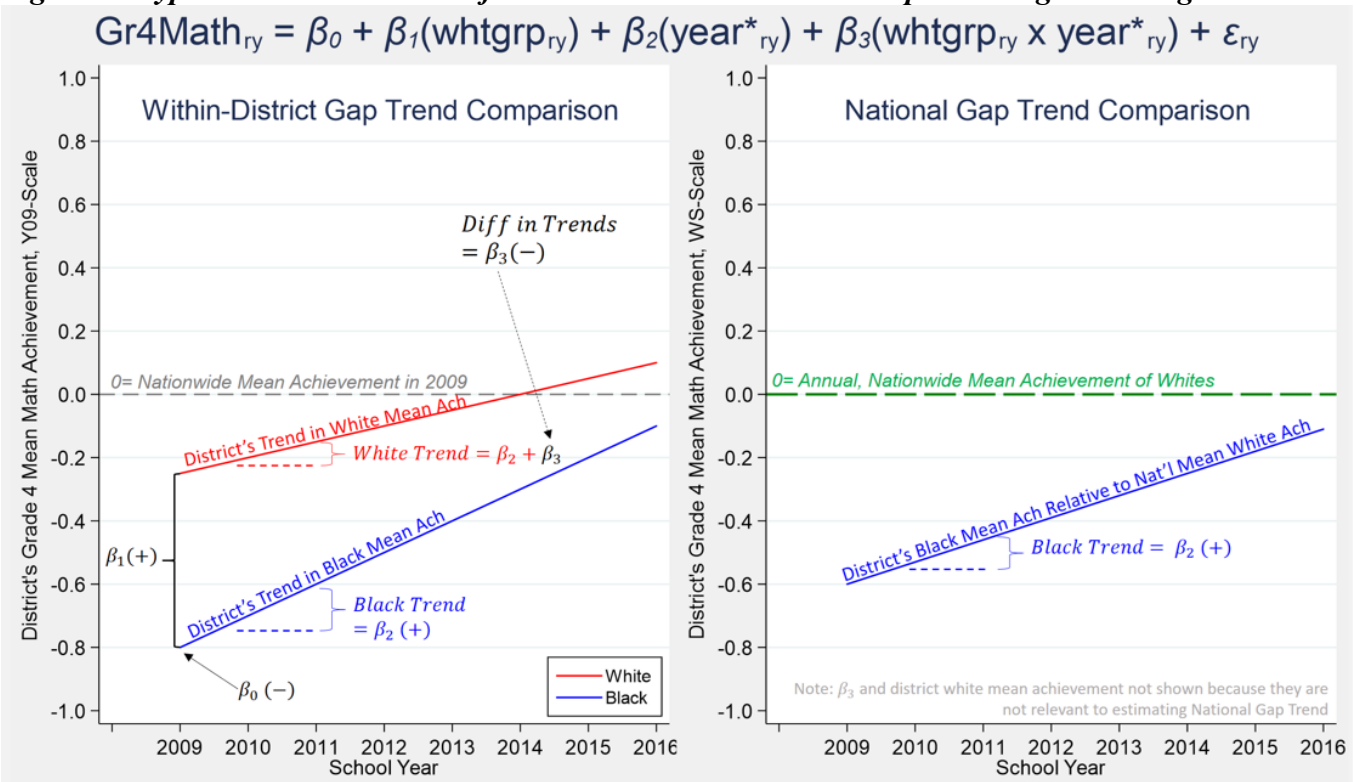
Table 6. Correspondence of Within-District Gap Trend Estimates across Subjects by Racial Dyad, Overall and Separately by Grade

	All Grades	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	Within-District Gap Trends, White-Black Gap						
Corr(ELA, Math)	0.61	0.62	0.62	0.62	0.58	0.61	0.56
% of Gaps that Widen in Both Subjects	39%	33%	36%	41%	43%	44%	39%
% of Gaps that Shrink in Both Subjects	32%	39%	34%	31%	28%	27%	31%
% of Gaps that Shrink in ELA Only	14%	11%	17%	13%	15%	15%	16%
% of Gaps that Shrink in Math Only	15%	18%	13%	15%	14%	13%	14%
Mean Diff in Gap Trend Estimate (ELA- Math)	-0.002	0.034	-0.030	0.005	-0.011	-0.011	-0.003
	Within-District Gap Trends, White-Hispanic Gap						
Corr(ELA, Math)	0.60	0.60	0.60	0.59	0.59	0.63	0.61
% of Gaps that Widen in Both Subjects	26%	26%	26%	26%	26%	28%	22%
% of Gaps that Shrink in Both Subjects	43%	48%	40%	40%	40%	42%	49%
% of Gaps that Shrink in ELA Only	22%	17%	24%	24%	25%	23%	20%
% of Gaps that Shrink in Math Only	9%	10%	9%	10%	9%	6%	9%
Mean Diff in Gap Trend Estimate (ELA- Math)	-0.075	-0.051	-0.082	-0.079	-0.083	-0.101	-0.064

Footnote: Table presents within-district (as opposed to national) gap trend estimates; achievement outcomes are therefore Y09-scaled (standardized around national mean achievement in 2009). Analyses restricted to district-grades that meet the MDR (minimum data requirement). Estimates produced via OLS with precision weighted outcomes.

Figures

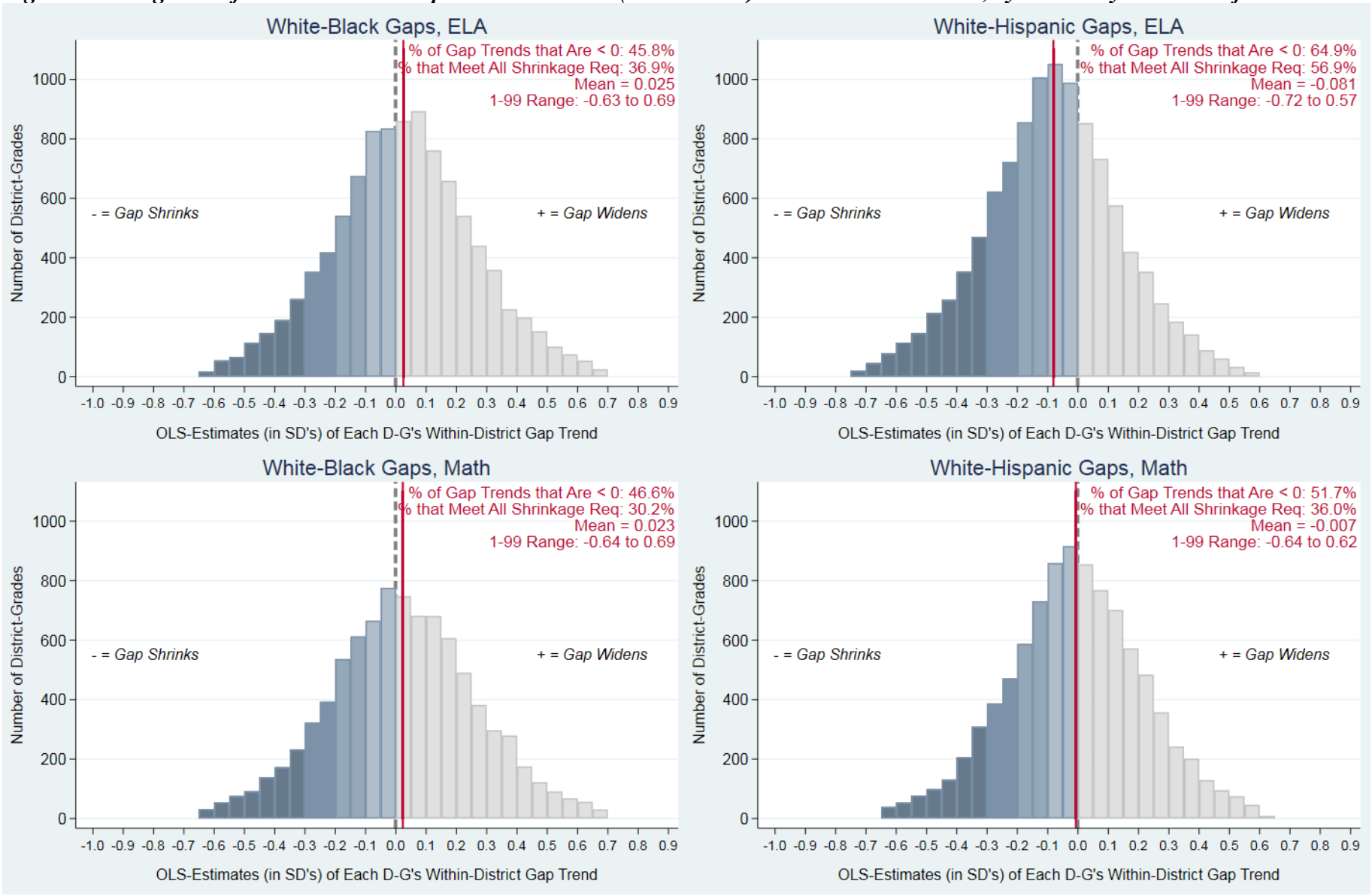
Figure 1. Hypothetical Illustration of Grade 4 White-Black Math Gap Shrinkage in a Single District



	Within-District Gap Shrinkage	National Gap Shrinkage
Called:		
Compare:	Blue to Red	Blue to Green
Achievement Scale:	Y09-scale (0 always represents overall mean achievement nationally as of 2009)	WS-scale (0 always represents White student mean achievement nationally in each given year)
Shrinkage Coefficient:	β_3	β_2
Interpretation of Coefficient:	If β_3 is negative, then the gap between the district's White and Black students is shrinking from 2009 - 2016	If β_2 is positive, then the district's Black students' mean is getting closer to the mean of White students nationally from 2009 - 2016
Other Eligibility Requirements?	Yes	No
Describe Other Eligibility Requirements:	β_1 is positive (true > 99% of the time) + β_2 is positive*	N/A

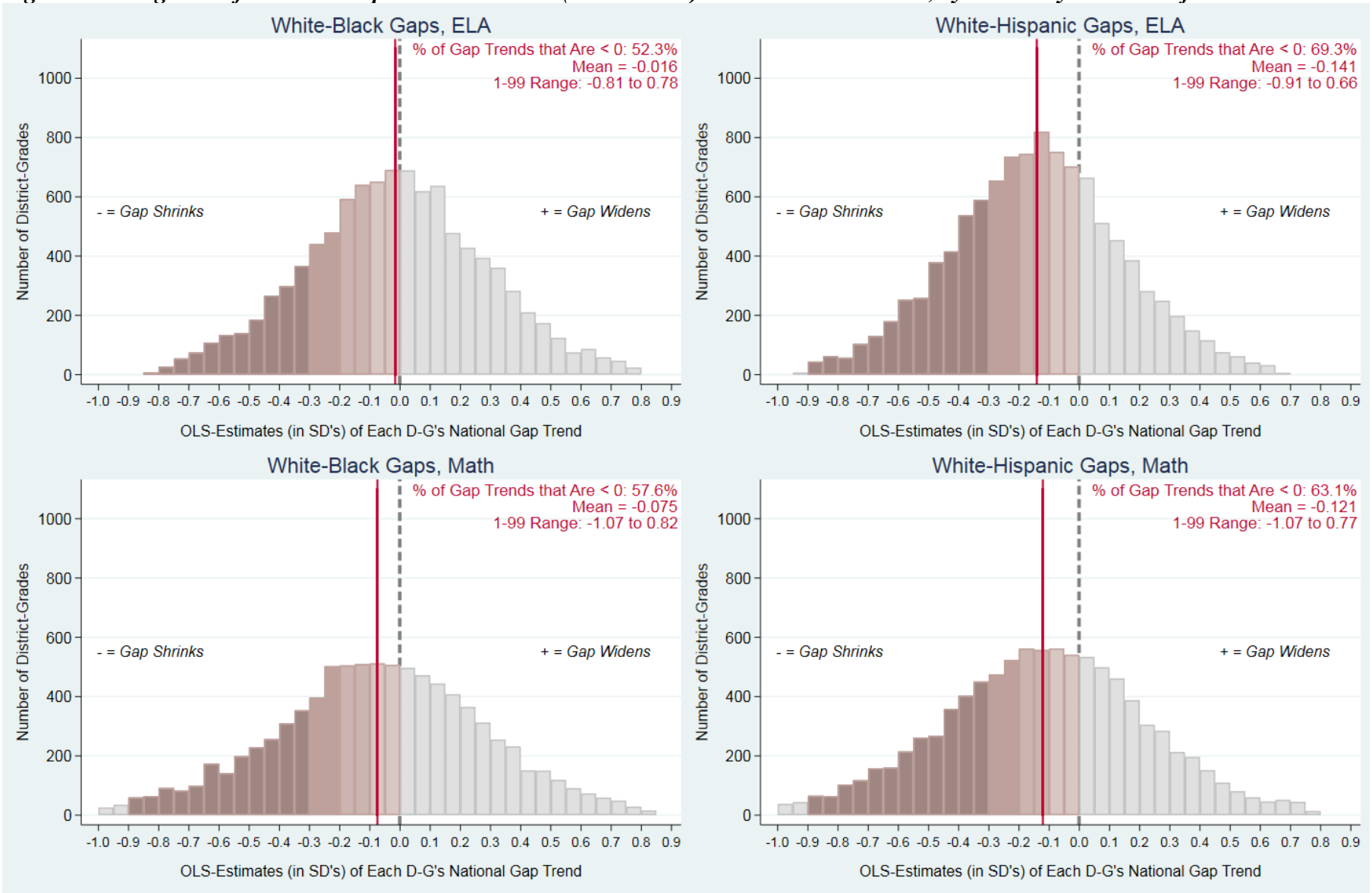
*Footnote: Figure 1 is introduced conceptually on pg. 10 and then again using the equation in the Methods section (pg. 13 of the narrative). Note that, on the right side, the national mean achievement of White students did not decrease during this period, which ensures that a positive β_2 indicates that Black mean achievement is increasing in an absolute sense.*By standardizing achievement outcomes relative to 2009 for within-district gap shrinkage, if β_2 is positive, then Black mean achievement is increasing in an absolute sense. By requiring β_2 to be positive for within-district gap shrinkage, we eliminate cases where gaps are only shrinking because the district's Black student performance is declining less precipitously than that of the district's White students.*

Figure 2. Histograms of Within-District Gap Trend Estimates (2009 – 2016) in each District-Grade, by Racial Dyad and Subject



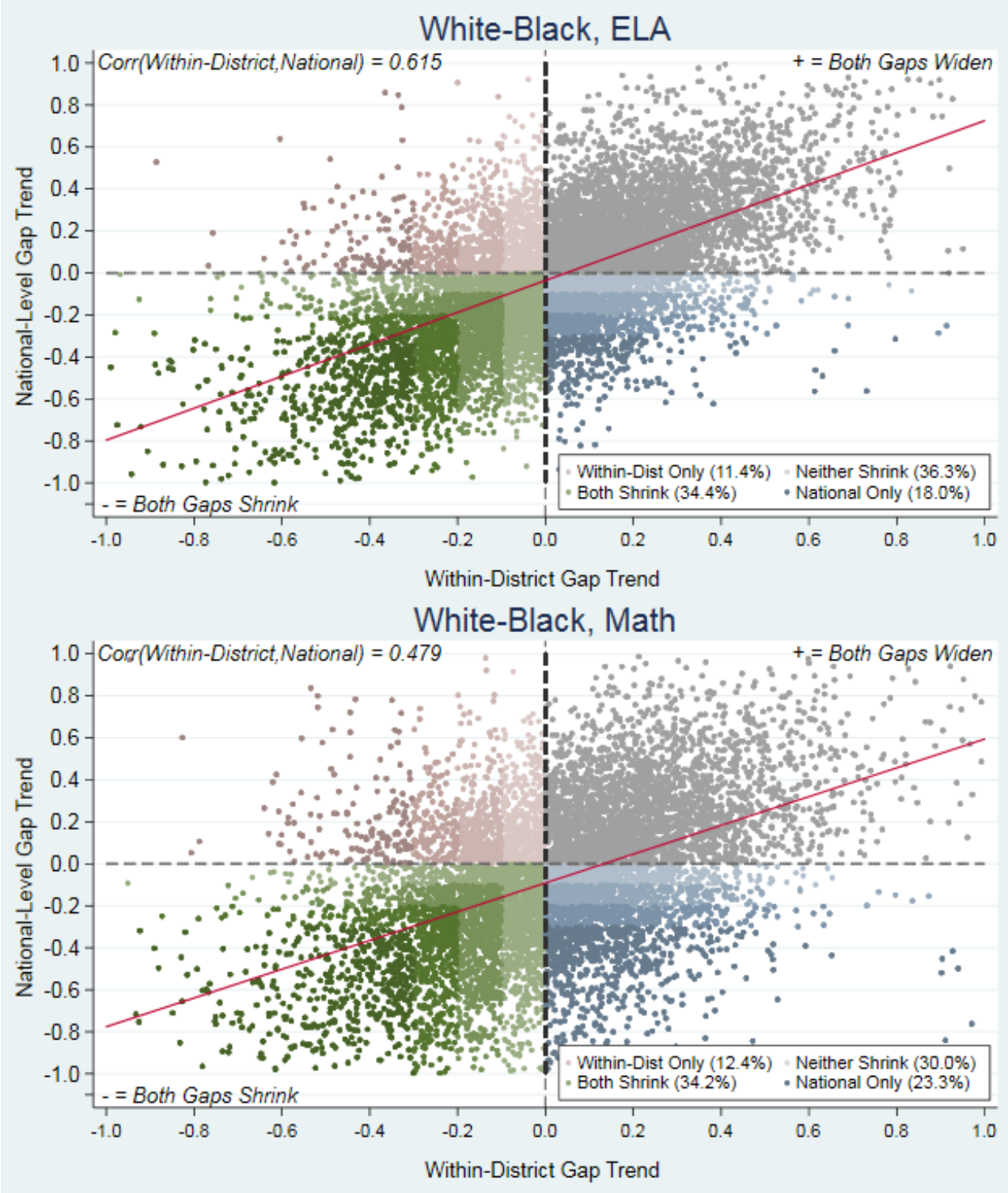
Footnote: Figure presents within-district (as opposed to national) gap trend estimates; achievement outcomes are therefore Y09-scaled (standardized around national mean achievement in 2009). Analyses restricted to district-grades that meet MDR. Estimates produced via OLS with precision weighted outcomes. Estimates in histograms limited to the 1st-99th percentile range for sake of visual clarity, however statistics in upper-right hand corners (red) are not.

Figure 3. Histograms of National Gap Trend Estimates (2009 – 2016) in each District-Grade, by Racial Dyad and Subject



Footnote: Figure presents national (as opposed to within-district) gap trend estimates; therefore WS-scaled (achievement outcomes standardized around the national mean for White students in the given year). Analyses restricted to district-grades that meet MDR. Estimates produced via OLS with precision weighted outcomes. Histograms limited to the 1st-99th percentile range for visual clarity, however statistics in upper-right hand corners (red) are not.

Figure 4. White-Black Gaps: Scatterplot of National vs. Within-District Gap Trends

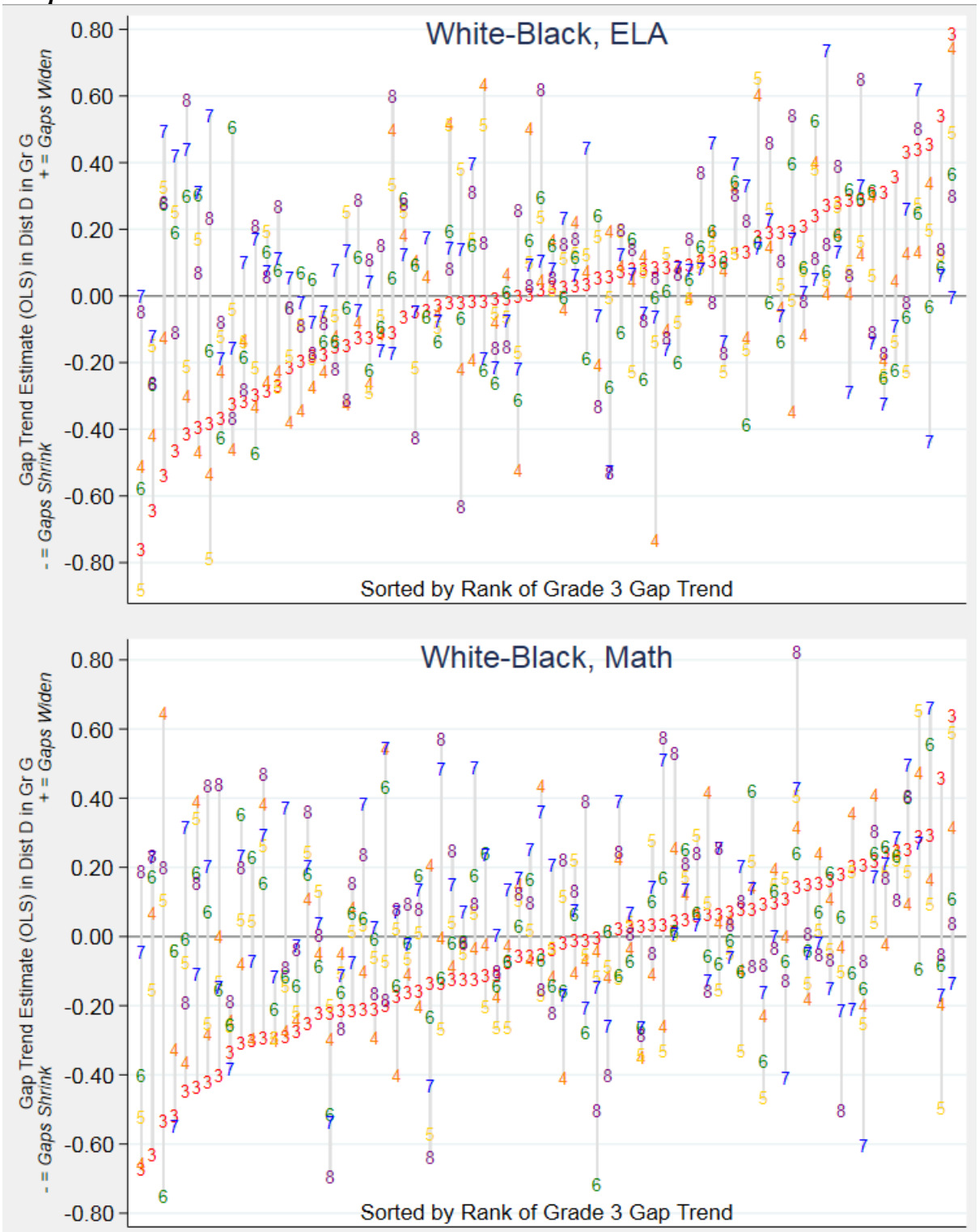


Footnote: Each dot represents a district-grade. For national gap trends (y-axis), achievement outcomes are always WS-scaled (standardized around the national mean achievement of White students in the given year). For within-district gap trends (x-axis), achievement outcomes are always Y09-scaled (standardized around national mean achievement in 2009). Estimates are produced via OLS with precision weighted outcomes. Estimates in scatterplot have been limited to the 1st-99th percentile range for sake of visual clarity, however reported statistics have not.

Figure 5. White-Hispanic Gaps: Scatterplot of National vs. Within-District Gap Trends

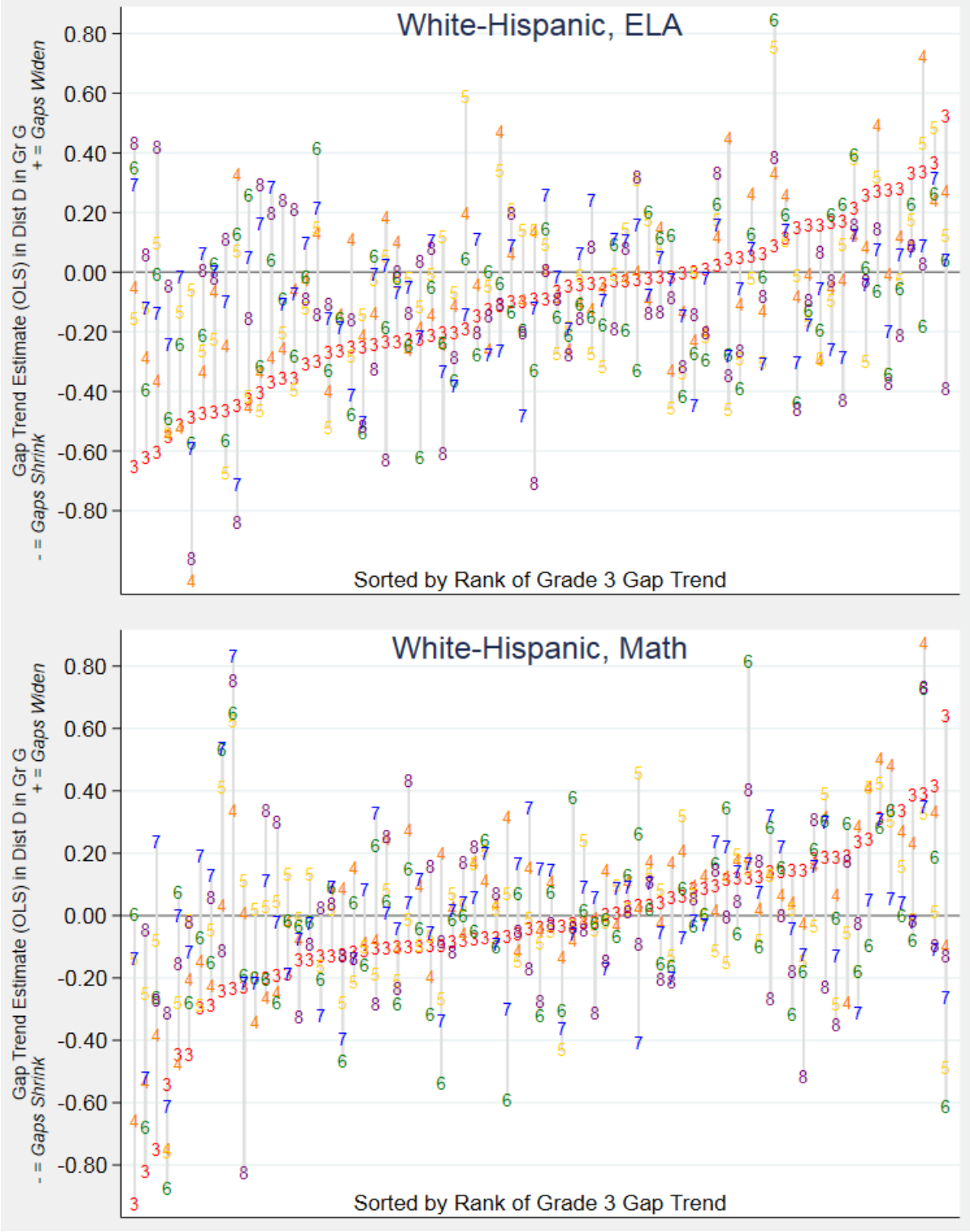
Footnote: Each dot represents a district-grade. For national gap trends (y-axis), achievement outcomes are always WS-scaled (standardized around the national mean achievement of White students in the given year). For within-district gap trends (x-axis), achievement outcomes are always Y09-scaled (standardized around national mean achievement in 2009). Estimates are produced via OLS with precision weighted outcomes. Estimates in scatterplot have been limited to the 1st-99th percentile range for sake of visual clarity, however reported statistics have not:

Figure 6. White-Black Within-District Gap Trends: Variation Across Grades in 75 Randomly Sampled Districts



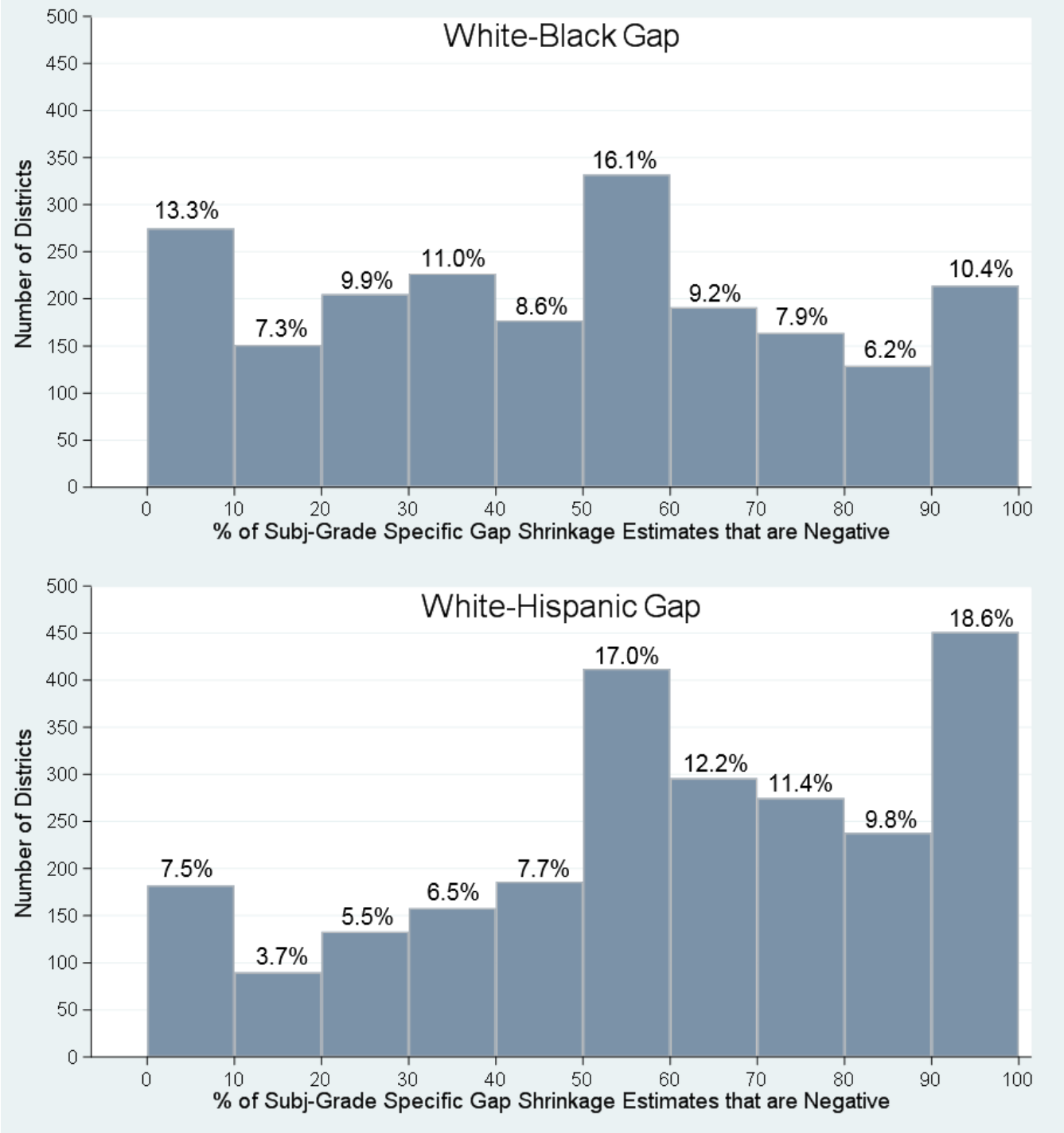
Footnote: Each plotted number represents a within-district gap trend estimate in corresponding grade of the number. Each of the 75 randomly sampled districts are represented by a vertical line connecting it's up to six grade-specific gap trend estimates. Achievement outcomes are Y09-scaled (standardized around national mean achievement in 2009). Estimates are produced via OLS with precision weighted outcomes.

Figure 7 White-Hispanic Within-District Gap Trends: Variation Across Grades within 75 Randomly Selected Districts



Footnote: Each plotted number represents a within-district gap trend estimate in corresponding grade of the number. Each of the 75 randomly sampled districts are represented by a vertical line connecting it's up to six grade-specific gap trend estimates. Achievement outcomes are Y09-scaled (standardized around national mean achievement in 2009). Estimates are produced via OLS with precision weighted outcomes.

Figure 8. Percent of Each District's (up to 12) Within-District Gap Trends that are Negative



Footnote: Figure presents within-district (as opposed to national) gap trend estimates; achievement outcomes are therefore Y09-scaled (standardized around national mean achievement in 2009). Analyses restricted to district-grades that meet the MDR (minimum data requirement). Estimates produced via OLS with precision weighted outcomes. If gap shrinkage is consistent across subjects and grades, we would expect to observe more districts in the tails of these histograms and fewer in the middle (i.e., districts would either shrink gaps in none or all grade-by-subject estimates).

Appendix A Tables and Figures

Appendix Table A1. Number of Years of Achievement Estimates per District-Grade Available in SEDA, by Subject and Subgroup

D-G's with Given Number of Annual Achievement Estimates *(of up to 8 yrs 2009 - 2016)*	ELA			Math		
	White	Black	Hispanic	White	Black	Hispanic
1	2,070 *(3.3)*	1,148 *(6.6)*	1,961 *(8.5)*	2,407 *(3.8)*	1,270 *(7.5)*	2,118 *(9.3)*
2	1,489 *(2.3)*	764 *(4.4)*	1,252 *(5.4)*	2,283 *(3.6)*	971 *(5.8)*	2,107 *(9.2)*
3	1,850 *(2.9)*	650 *(3.7)*	1,277 *(5.5)*	2,992 *(4.8)*	995 *(5.9)*	2,108 *(9.3)*
4	1,656 *(2.6)*	772 *(4.4)*	1,275 *(5.5)*	2,269 *(3.6)*	787 *(4.7)*	1,181 *(5.2)*
5	6,508 *(10.2)*	1,354 *(7.8)*	2,476 *(10.7)*	6,291 *(10.0)*	1,313 *(7.8)*	2,191 *(9.6)*
6	6,054 *(9.5)*	1,815 *(10.5)*	2,922 *(12.6)*	6,705 *(10.7)*	1,745 *(10.4)*	2,571 *(11.3)*
7	12,513 *(19.7)*	3,380 *(19.5)*	5,693 *(24.6)*	11,789 *(18.8)*	3,476 *(20.6)*	4,817 *(21.1)*
8	31,520 *(49.5)*	7,481 *(43.1)*	6,327 *(27.3)*	28,123 *(44.7)*	6,278 *(37.3)*	5,691 *(25.0)*
Total Number of D-G's	63,660 *(100.0)*	17,364 *(100.0)*	23,183 *(100.0)*	62,859 *(100.0)*	16,835 *(100.0)*	22,784 *(100.0)*

Footnote: D = District; G = Grade. MDR = Minimum Data Requirements. Table is limited to D-Gs with at least one annual achievement estimate in the given subject for the given subgroup.

Appendix Table A2. Seven Minimum Data Requirement (MDR) Definitions, and Percent of District-Grades that Meet Requirement, by Subject and Racial Dyad

			ELA		Math	
	Def #	Both Subgroups Must Have….	WBG	WHG	WBG	WHG
Least ← Most Restrictive	1	... (Both Anchor Years)	57%	47%	51%	42%
	2	... (Both Anchor Years) Or (6 of 8 Years)	63%	52%	58%	46%
	3	... (At Least 4 of 8 Years) & (5-Year Spread)	74%	67%	70%	61%
	4	... (At Least 5 of 8 Years)	78%	72%	74%	66%
	5	... (At Least 4 of 8 Years)	83%	79%	79%	71%
	6	... (At Least 2 of 8 Years)	93%	91%	92%	90%
	7	... No Minimum Data Requirement	100%	100%	100%	100%

Footnote: Denominator for percentages are the number of district-grades that have any possibility of estimating the given achievement gap (i.e., has at least one year with estimated achievement means in both subgroups).

Appendix Table A3. Comparison of Primary Within-District Gap Trend Results Using Seven Different MDR Definitions

	MDR Def. #1	MDR Def. #2	MDR Def. #3	MDR Def. #4	MDR Def. #5	MDR Def. #6	No MDR
				White-Black Gap, ELA			
(1A) D-G's w/ At Least 1 Year w/ Both Means	16,043 *(100%)*	16,043 *(100%)*	16,043 *(100%)*	16,043 *(100%)*	16,043 *(100%)*	16,043 *(100%)*	16,043 *(100%)*
(1B) Among (1A), D-G's that Meet this MDR	9,105 *(57%)*	10,046 *(63%)*	11,928 *(74%)*	12,563 *(78%)*	13,371 *(83%)*	14,872 *(93%)*	16,043 *(100%)*
(1C) % of Non-Wht Group Observed in D-G's in (1B)	67%	73%	82%	88%	90%	92%	100%
Within-District Gap Trend Estimates:							
N	9,105	10,046	11,928	12,563	13,371	14,874	14,874
Mean	0.03	0.02	0.02	0.02	0.02	0.03	0.03
SD	0.25	0.26	0.28	0.30	0.33	0.56	0.56
1st Percentile	-0.60	-0.63	-0.72	-0.77	-0.89	-1.53	-1.53
99th Percentile	0.68	0.69	0.75	0.79	0.91	1.68	1.68
Among (1A), D-G's where Gaps Shrink	3,277 *(36%)*	3,704 *(37%)*	4,517 *(38%)*	4,712 *(38%)*	5,066 *(38%)*	5,719 *(38%)*	5,719 *(38%)*
				White-Black Gap, Math			
(1A) D-G's w/ At Least 1 Year w/ Both Means	15,508 *(100%)*	15,508 *(100%)*	15,508 *(100%)*	15,508 *(100%)*	15,508 *(100%)*	15,508 *(100%)*	15,508 *(100%)*
(1B) Among (1A), D-G's that Meet this MDR	7,880 *(51%)*	8,987 *(58%)*	10,838 *(70%)*	11,428 *(74%)*	12,228 *(79%)*	14,218 *(92%)*	15,508 *(100%)*
(1C) % of Non-Wht Group Observed in D-G's in (1B)	63%	72%	80%	87%	88%	91%	100%
Within-District Gap Trend Estimates:							
N	7,880	8,987	10,838	11,428	12,228	14,222	14,222
Mean	0.03	0.02	0.02	0.02	0.02	0.02	0.02
SD	0.26	0.27	0.29	0.31	0.34	0.63	0.63
1st Percentile	-0.62	-0.64	-0.74	-0.78	-0.89	-1.90	-1.90
99th Percentile	0.68	0.70	0.77	0.80	0.92	1.94	1.94
Among (1A), D-G's where Gaps Shrink	2,264 *(29%)*	2,715 *(30%)*	3,379 *(31%)*	3,579 *(31%)*	3,874 *(32%)*	4,735 *(33%)*	4,735 *(33%)*

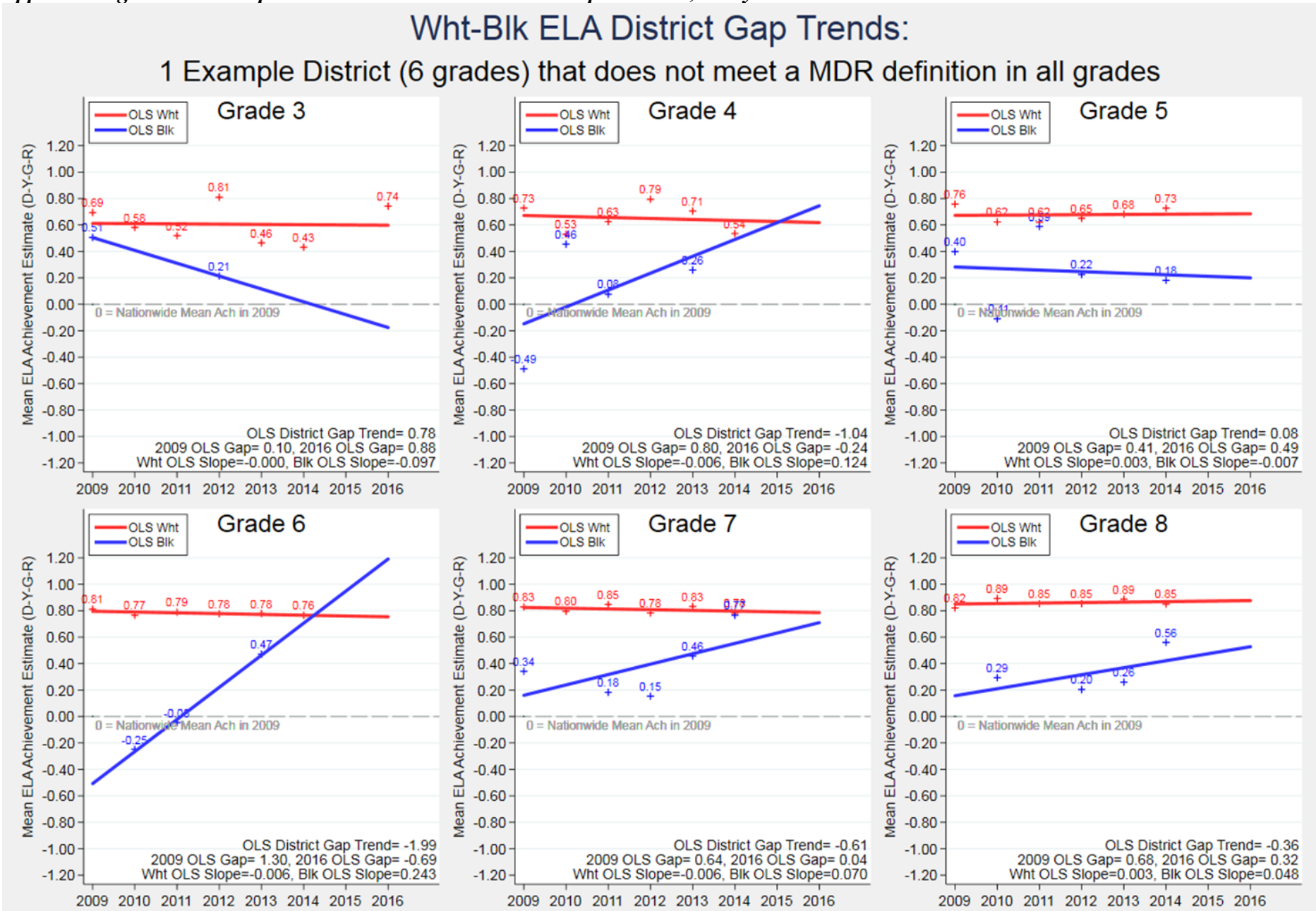
(continues onto next page…)

(continued from previous page…)

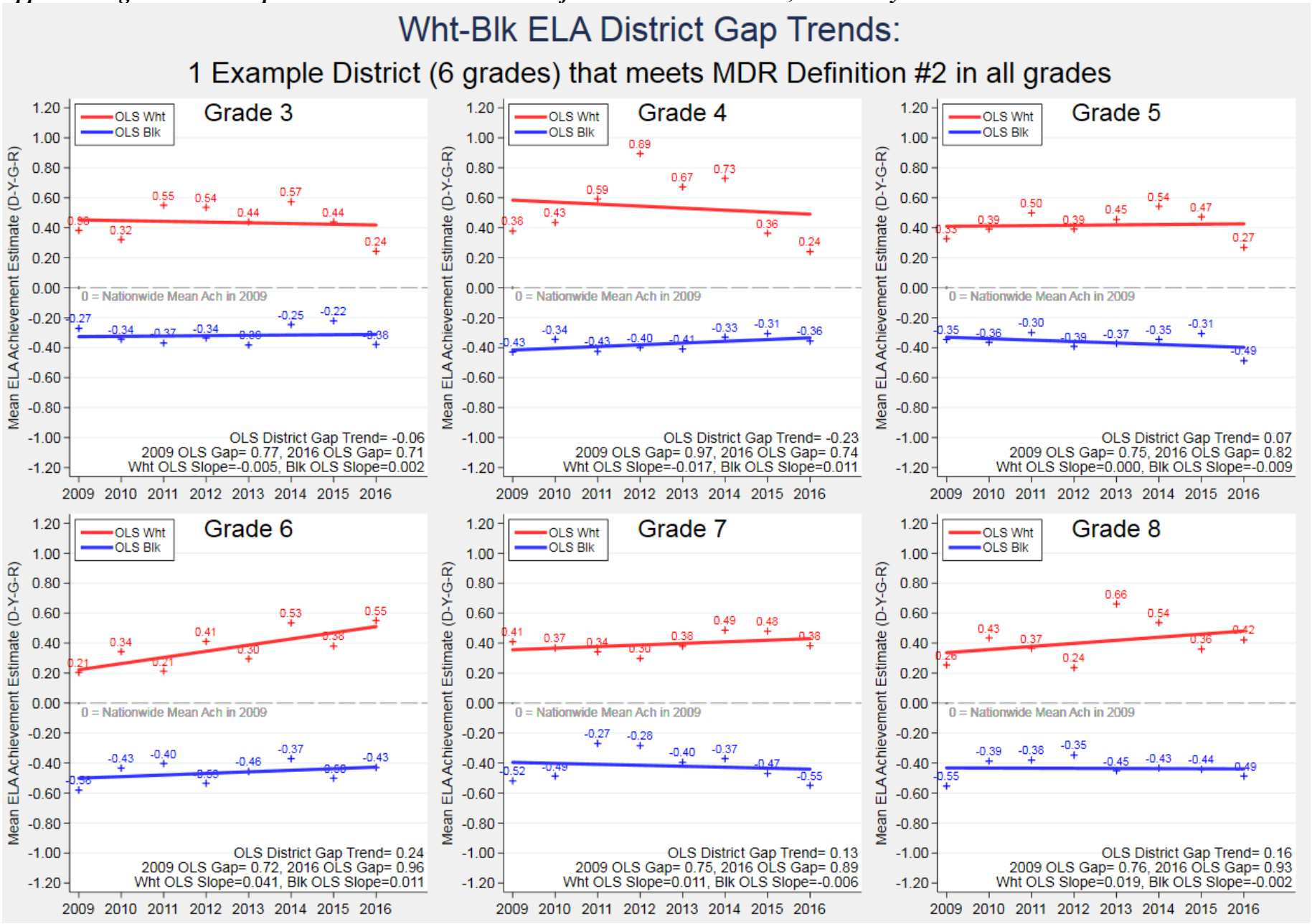
	MDR Def. #1	MDR Def. #2	MDR Def. #3	MDR Def. #4	MDR Def. #5	MDR Def. #6	No MDR
				White-Hispanic Gap, ELA			
(1A) D-G's w/ At Least 1 Year w/ Both Means	21,088 *(100%)*	21,088 *(100%)*	21,088 *(100%)*	21,088 *(100%)*	21,088 *(100%)*	21,088 *(100%)*	21,088 *(100%)*
(1B) Among (1A), D-G's that Meet this MDR	9,941 *(47%)*	10,872 *(52%)*	14,198 *(67%)*	15,274 *(72%)*	16,557 *(79%)*	19,201 *(91%)*	21,088 *(100%)*
(1C) % of Non-Wht Group Observed in D-G's in (1B)	65%	69%	80%	86%	88%	91%	100%
Within-District Gap Trend Estimates:							
N	9,941	10,872	14,198	15,274	16,557	19,223	19,223
Mean	-0.08	-0.08	-0.08	-0.08	-0.08	-0.07	-0.07
SD	0.25	0.25	0.29	0.31	0.36	0.64	0.64
1st Percentile	-0.70	-0.72	-0.84	-0.91	-1.06	-1.96	-1.96
99th Percentile	0.57	0.58	0.71	0.77	0.93	1.90	1.90
Among (1A), D-G's where Gaps Shrink	5,641 *(57%)*	6,184 *(57%)*	7,887 *(56%)*	8,388 *(55%)*	9,009 *(54%)*	10,261 *(53%)*	10,261 *(53%)*
				White-Hispanic Gap, Math			
(1A) D-G's w/ At Least 1 Year w/ Both Means	20,654 *(100%)*	20,654 *(100%)*	20,654 *(100%)*	20,654 *(100%)*	20,654 *(100%)*	20,654 *(100%)*	20,654 *(100%)*
(1B) Among (1A), D-G's that Meet this MDR	8,671 *(42%)*	9,589 *(46%)*	12,535 *(61%)*	13,559 *(66%)*	14,755 *(71%)*	18,570 *(90%)*	20,654 *(100%)*
(1C) % of Non-Wht Group Observed in D-G's in (1B)	62%	65%	75%	81%	83%	90%	100%
Within-District Gap Trend Estimates:							
N	8,671	9,589	12,535	13,559	14,755	18,586	18,586
Mean	-0.01	-0.01	0.00	-0.01	-0.01	0.00	0.00
SD	0.24	0.25	0.30	0.32	0.36	0.73	0.73
1st Percentile	-0.62	-0.64	-0.79	-0.86	-0.98	-2.34	-2.34
99th Percentile	0.59	0.62	0.81	0.84	1.00	2.40	2.40
Among (1A), D-G's where Gaps Shrink	3,060 *(35%)*	3,451 *(36%)*	4,544 *(36%)*	5,011 *(37%)*	5,468 *(37%)*	7,013 *(38%)*	7,013 *(38%)*

Footnote: D = District; G = Grade. MDR = Minimum Data Requirements. Primary results reported in narrative come from MDR Def. #2.

Appendix Figure A1. Example District with Grades that have Sparse Data, Noisy Trend Estimates



Appendix Figure A2. Example District that Meets MDR Definition #2 in all Grades, Less Noisy Trend Estimates



Appendix B. Tables and Figures

Appendix Table B1. 2009 National Mean/SD NAEP Scores for All Students, by Subject and Grade Level.

Grade	ELA		Math	
Level	Mean	SD	Mean	SD
3	206.2	38.1	227.6	27.8
4	217.0	37.7	238.1	29.8
5	227.8	37.4	248.6	31.7
6	238.6	37	259.1	33.7
7	249.3	36.6	269.6	35.6
8	260.1	36.3	280.1	37.6

Footnote: Source: Fahle et al., (2019), Table 6. NAEP Means and Standard Deviations by Year and Grade. Their Table 6 shows the interpolated national NAEP estimates (estimates shown in red are not interpolated). Fahle et al. used the expanded population estimates, which may differ slightly from those reported publicly on the website.

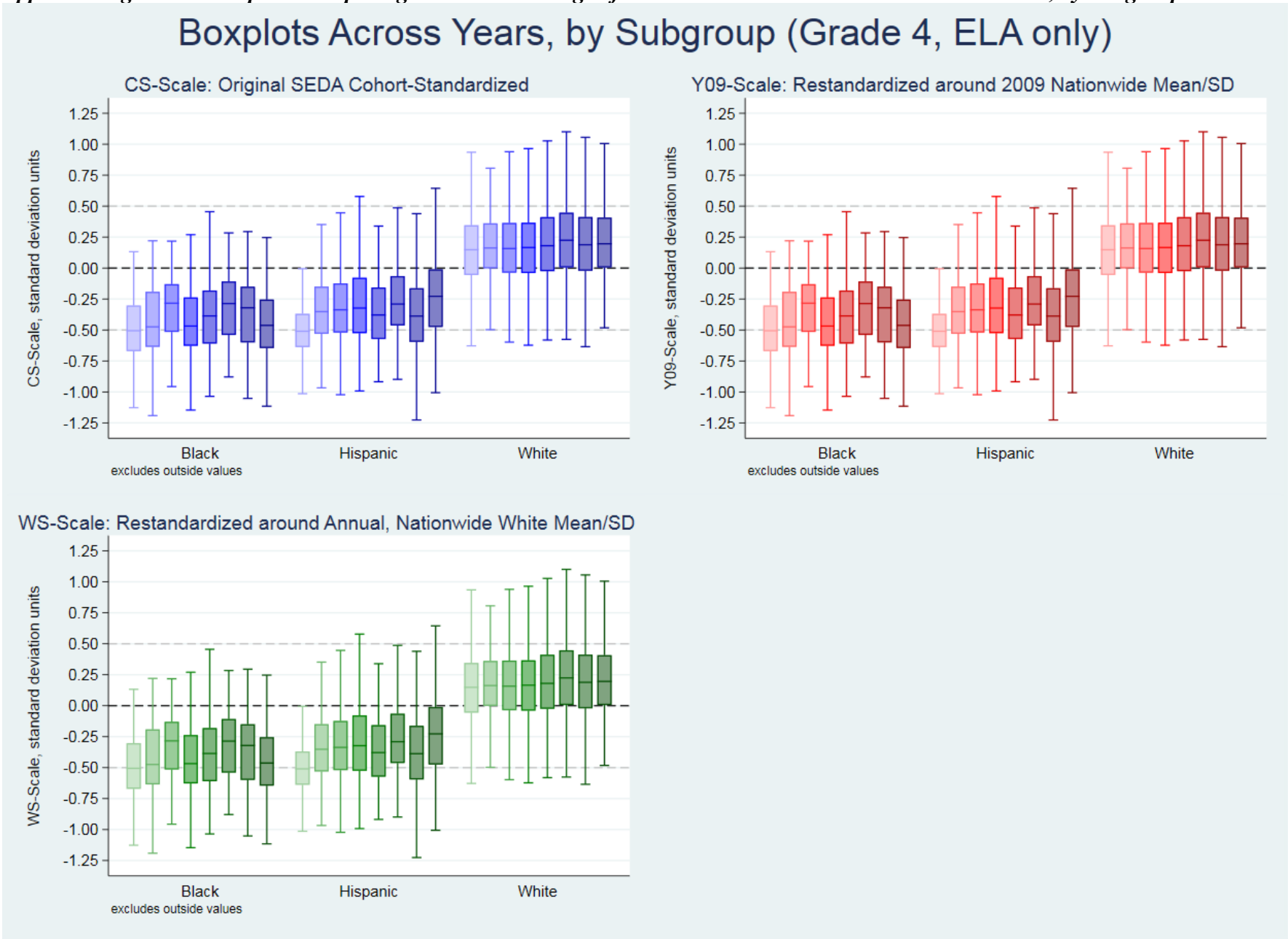
Appendix Table B2. National Mean/SD of NAEP Scores for White Students, by Year, Grade, and Subject

Subject	Year	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
ELA	2009	218.5	229.0	239.5	250.0	260.5	271.0
		(32.3)	*(32.0)*	*(31.8)*	*(31.5)*	*(31.3)*	*(31.0)*
	2010	219.0	229.5	240.0	250.5	261.0	271.5
		(32.3)	*(32.0)*	*(31.8)*	*(31.5)*	*(31.3)*	*(31.0)*
	2011	219.5	230.0	240.5	251.0	261.5	272.0
		(32.3)	*(32.0)*	*(31.8)*	*(31.5)*	*(31.3)*	*(31.0)*
	2012	219.8	230.5	241.3	252.0	262.8	273.5
		(32.8)	*(32.5)*	*(32.3)*	*(32.0)*	*(31.8)*	*(31.5)*
	2013	220.0	231.0	242.0	253.0	264.0	275.0
		(33.3)	*(33.0)*	*(32.8)*	*(32.5)*	*(32.3)*	*(32.0)*
	2014	220.9	231.5	242.1	252.8	263.4	274.0
		(33.3)	*(33.0)*	*(32.8)*	*(32.5)*	*(32.3)*	*(32.0)*
	2015	221.8	232.0	242.3	252.5	262.8	273.0
		(33.3)	*(33.0)*	*(32.8)*	*(32.5)*	*(32.3)*	*(32.0)*
	2016	221.0	231.5	242.0	252.5	263.0	273.5
		(34.4)	*(34.0)*	*(33.6)*	*(33.3)*	*(32.9)*	*(32.5)*
	2017	220.3	231.0	241.8	252.5	263.3	274.0
		(35.5)	*(35.0)*	*(34.5)*	*(34.0)*	*(33.5)*	*(33.0)*

Subject	Year	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Math	2009	237.0	248.0	259.0	270.0	281.0	292.0
		(24.3)	*(26.0)*	*(27.8)*	*(29.5)*	*(31.3)*	*(33.0)*
	2010	237.5	248.5	259.5	270.5	281.5	292.5
		(24.3)	*(26.0)*	*(27.8)*	*(29.5)*	*(31.3)*	*(33.0)*
	2011	238.0	249.0	260.0	271.0	282.0	293.0
		(24.3)	*(26.0)*	*(27.8)*	*(29.5)*	*(31.3)*	*(33.0)*
	2012	238.6	249.5	260.4	271.3	282.1	293.0
		(24.9)	*(26.5)*	*(28.1)*	*(29.8)*	*(31.4)*	*(33.0)*
	2013	239.3	250.0	260.8	271.5	282.3	293.0
		(25.5)	*(27.0)*	*(28.5)*	*(30.0)*	*(31.5)*	*(33.0)*
	2014	238.3	249.0	259.8	270.5	281.3	292.0
		(25.4)	*(27.0)*	*(28.6)*	*(30.3)*	*(31.9)*	*(33.5)*
	2015	237.3	248.0	258.8	269.5	280.3	291.0
		(25.3)	*(27.0)*	*(28.8)*	*(30.5)*	*(32.3)*	*(34.0)*
	2016	237.1	248.0	258.9	269.8	280.6	291.5
		(26.3)	*(28.0)*	*(29.8)*	*(31.5)*	*(33.3)*	*(35.0)*
	2017	237.0	248.0	259.0	270.0	281.0	292.0
		(27.3)	*(29.0)*	*(30.8)*	*(32.5)*	*(34.3)*	*(36.0)*

Footnote: We retrieve national average scale scores and standard deviations for grade 4 and 8 mathematics and reading for White, students in public schools in 2009, 2011, 2013, 2015, and 2017 from the NAEP Data Explorer (https://www.nationsreportcard.gov/ndecore/xplore/NDE). We follow the procedures outlined by Reardon et al. (2019) on pp. 20-21 to interpolate and extrapolate linearly to obtain mean achievement estimates for White students in grades 3, 5, 6, and 7 and even school years, in which NAEP is not administered. NAEP Data Explorer Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009, 2011, 2013, 2015, and 2017 Mathematics and Reading Assessments. NOTE: Black includes African American, Hispanic includes Latino, and Pacific Islander includes Native Hawaiian. Race categories exclude Hispanic origin. Prior to 2011, students in the "two or more races" category were categorized as "unclassified."

Appendix Figure B1. Boxplots Comparing the Three Scalings of SEDA Grade 4 ELA Achievement Estimates, by Subgroup and Year



Appendix Figure B2. Full HLM Model

The multilevel random effects model is illustrated for math achievement for Black and White subgroups in fourth grade.

Level 1: Subgroup-Specific (r) Annual (y) Achievement Scores, Nested within Districts [where $\epsilon_{\mathrm{ryds}} \sim \mathrm{iid}, \mathrm{N}(0, \sigma)$]

$$Gr4Math_{ryds}^{Y09-scale} = \pi_{0ds} + \pi_{1ds}(whtgrp_{ryds}) + \pi_2(year^*_{ryds}) + \pi_3(whtgrp_{ryds} \times year^*_{ryds}) + \epsilon_{ryds}$$

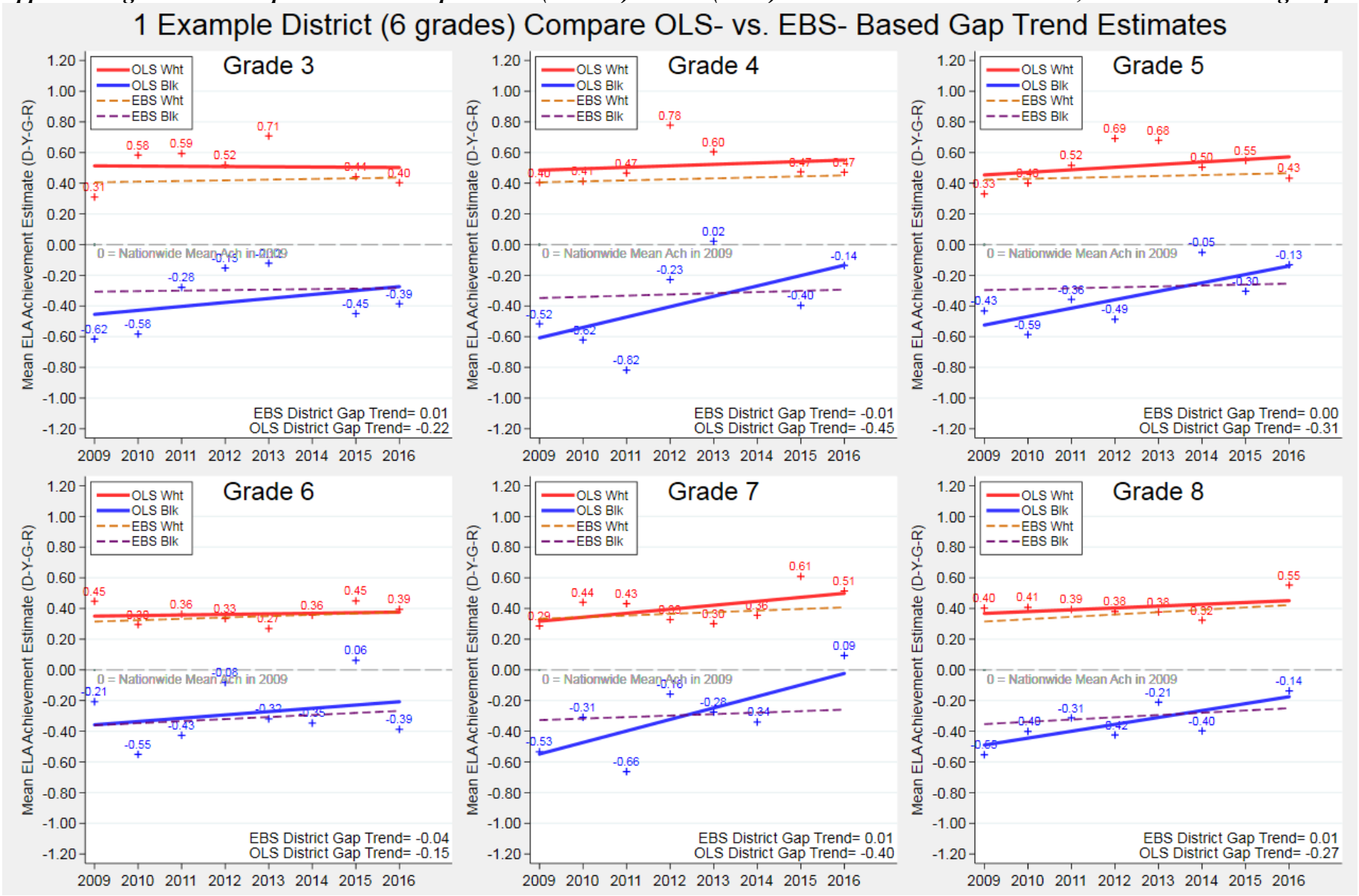
Level 2: Districts, Nested within States [where $r_{0ds}, r_{1ds}, r_{2ds}, r_{3ds} \sim iid, N(0, \boldsymbol{\tau})$]

$\pi_{0ds} = \beta_{00s} + r_{0ds}$ ← π_{0ds} = *avg. Black subgroup starting point (2009) in district d (in state s)*
$\pi_{1ds} = \beta_{10s} + r_{1ds}$ ← π_{1ds} = *avg. diff in White/Black starting points (2009) in district d (in state s)*
$\pi_{2ds} = \beta_{20s} + r_{2ds}$ ← π_{2ds} = *avg. Black subgroup trend in district d (in state s)*
$\pi_{3ds} = \beta_{30s} + r_{3ds}$ ← π_{3ds} = *avg. diff in White/Black trends (gap shrinks/expands) in district d (in state s)*

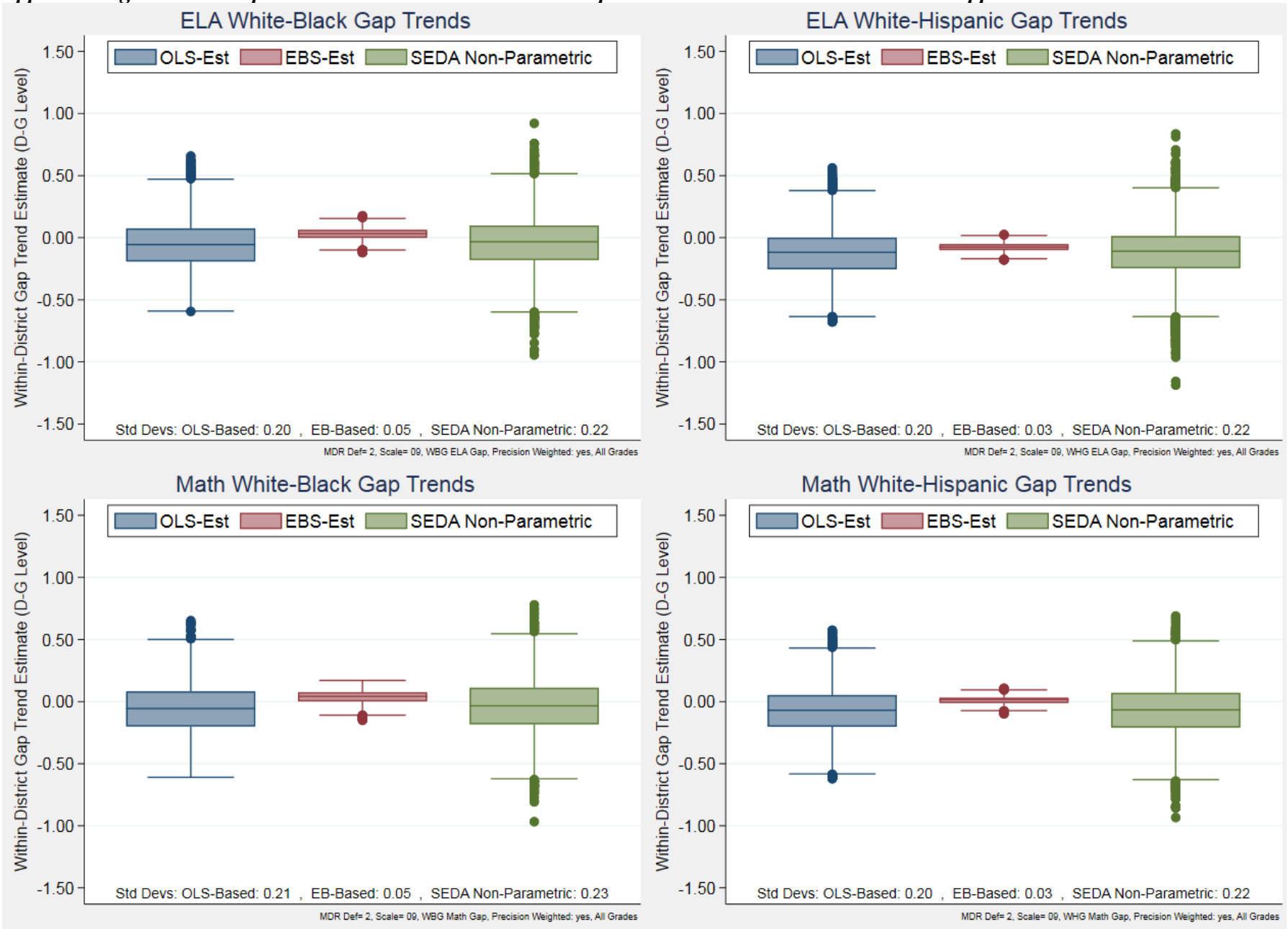
Level 3: Across States [where $\mu_{00s}, \mu_{10s}, \mu_{20s}, \mu_{30s} \sim iid, N(0, \mathbf{T})$]

$\beta_{00s} = \delta_{000} + \mu_{00s}$ ← β_{00s} = *avg. Black subgroup starting point (2009) in state s*
$\beta_{10s} = \delta_{100} + \mu_{10s}$ ← β_{10s} = *avg. diff in White/Black starting points (2009) in state s*
$\beta_{20s} = \delta_{200} + \mu_{20s}$ ← β_{20s} = *avg. Black subgroup trend in state s*
$\beta_{30s} = \delta_{300} + \mu_{30s}$ ← β_{30s} = *avg. diff in White/Black trends (gap shrinks/expands) in state s)*

Appendix Figure B3. Example District: Compare EBS (Dashed) to OLS (Solid) Achievement Trend Estimates, White & Black Subgroups



Appendix Figure B4. Compare Variation in Within-District Gap Trend Estimates across Estimation Approaches



Appendix Figure B5. Comparison of Within-District Gap Trend Estimates, With and Without Precision-Weighting