ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-83

Are Schools Deemed Effective Based on Overall Student Growth Also Closing Achievement Gaps? Examining the Black-White Gap in Schools

James Soland
University of Virginia

Research has begun to investigate whether teachers and schools are as effective with certain student subgroups as they are with the overall student population. Most of this research has examined the issue by trying to produce causal estimates of school contributions to short-term student growth (usually using value-added models) and has emphasized rank orderings of schools by subgroup. However, not much is known about whether schools contributing to long-term growth for all students are also contributing to student growth by subgroup in ways that might close achievement gaps. In this study, schools' contributions to student growth are estimated separately for Black versus White students. Results show that focusing on rank orderings of schools alone can mask troubling trends in relative achievement over time. Options for how policymakers can sensibly hold schools accountable for student growth, including under The Every Student Succeeds Act, are discussed.

VERSION: June 2019

Suggested citation: Soland, James. (2019). Are Schools Deemed Effective Based on Overall Student Growth Also Closing Achievement Gaps? Examining the Black-White Gap in Schools . (EdWorkingPaper: 19-83). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-83

Abstract

Research has begun to investigate whether teachers and schools are as effective with certain student subgroups as they are with the overall student population. Most of this research has examined the issue by trying to produce causal estimates of school contributions to short-term student growth (usually using value-added models) and has emphasized rank orderings of schools by subgroup. However, not much is known about whether schools contributing to long-term growth for all students are also contributing to student growth by subgroup in ways that might close achievement gaps. In this study, schools' contributions to student growth are estimated separately for Black versus White students. Results show that focusing on rank orderings of schools alone can mask troubling trends in relative achievement over time. Options for how policymakers can sensibly hold schools accountable for student growth, including under The Every Student Succeeds Act, are discussed.

Keywords: school effectiveness, growth modeling, accountability, program evaluation, achievement gaps.

Are Schools Deemed Effective Based on Overall Student Growth Also Closing Achievement Gaps? Examining the Black-White Gap in Schools

Research suggests that estimates of school and district quality based on achievement at a given point in time (status) are very different than when based on estimates of growth in achievement over time (most recently, Reardon, 2017). Based in part on this body of research, practitioners, policymakers, and researchers are paying more attention to growth (Betebenner, 2008; Castellano & Ho, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Thum, 2003). As a case in point, there is newfound interest in holding schools accountable on the basis of student growth rather than status. Under The Every Student Succeeds Act (ESSA) of 2015, 47 states plan to use student growth as an accountability indicator in elementary and middle school, and 33 states weight student growth the same or more than status (ESSA Plans, 2017).

One of the major underlying reasons that accountability policies might focus on growth is to improve achievement among low-performing students, improve achievement among racial minority students and, ideally, close gaps between racial minority and White students (Klein, 2016). In particular, the Black-White achievement gap has garnered much attention in policy, practice, and research. Studies show that, for instance, only 5% of Black students ages 13 and above have mathematics scores in the top quartile of the White math score distribution (Neal, 2006). These gaps likely have long-term consequences, including severe underrepresentation of Black students in selective universities (Reardon, Baker, & Klasik, 2012).

Based in part on these gaps and the intent of many policies to help close them, there is emerging evidence on teacher and school effectiveness by student subgroup (Aaronson et al., 2007; Lockwood and McCaffrey, 2009; Loeb et al., 2014; Fox, 2016). Most such research emphasizes causal estimates of school contributions to short-term student growth with particular emphasis on rank orderings of schools by subgroup. Thus, most analyses in the literature address a question like: do the schools that improve the standing of White students in the test score distribution also improve the standing of racial minority students? In some cases, these analyses have been performed by estimating value-added models (VAMs) separately by student subgroup to see if White and minority students are moving positively not in the overall distribution of scores, but of their respective distributions (e.g., Loeb, Soland, & Fox, 2014).

However, few studies consider whether the schools contributing to long-term student growth are the same contributing to long-term subgroup growth, and to closing gaps in particular. That is, questions about school contributions to the relative growth of different subgroups could be asked in a different way: are schools generating above-average growth for White students also generating growth among racial minority students sufficient to help close achievement gaps? One could imagine several scenarios for relative growth including (1) White and minority students both growing above average but gaps not closing, (2) White and minority students both growing below average and gaps not closing, (3) White students growing at or above the average but minority students growing faster such that gaps are closing, and (4) White students growing below the average and minority students growing faster such that gaps are closing. These examples are not exhaustive. For example, they do not include scenarios in which school-level gaps are widening, which is plausible (Downey & Broh, 2003; von Hippel, Workman, & Downey, 2018). These differences are often obscured by the focus on rank orderings and short-term causal estimates in the current literature on school effectiveness with student subgroups.

This study considers how much inferences about school contributions to long-term student growth differ when simple rank orderings are compared versus exploring more nuanced questions about relative student growth over time. The issue is examined by comparing school effectiveness estimates for White versus Black students in a large Southern state that administers a vertically scaled achievement test suitable for use in growth models. Using these data, three research questions are explored:

1. How different are rank orderings of schools for Black versus White students when based on estimates of school contributions to long-term student growth?
2. How many schools have growth estimates that are higher for Black students than White students (suggesting gaps are closing) or vice-versa?
3. Are gaps widening or narrowing in schools with high growth for White students?

Though all of these questions use a growth model rather than a traditional lag-score VAM, Question 1 focuses on the sensitivity of rank orderings of schools, which is common in the VAM literature. By contrast, Questions 2 and 3 rely more heavily on the test's vertical scale to better disaggregate variable patterns in growth by subgroup. The intent of this study is not to argue for one model or approach to comparing estimates over another. Rather, the intent is to determine whether inferences about school effectiveness shift at all when questions about gaps and gap closures are made more paramount. By examining these questions, this study brings a different angle to questions of what an "effective" school is in the context of equity concerns.

Background

In this section, evidence on the Black-White achievement gap is briefly discussed. Then, research on differential effectiveness of schools and teachers by subgroup is reviewed.

Black-White Achievement Gaps

Research shows that, on average, Black students score well below White students on standardized tests of math and reading (Fryer Jr & Levitt, 2004; Hedges & Nowell, 1999; Jencks

& Phillips, 2011; Neal, 2006). Studies demonstrate that these Black-White gaps begin very early in students' schooling and often persist through high school, though these patterns are not always consistent. Achievement gaps are present and sizeable in kindergarten, according to research using the Early Childhood Longitudinal Program or ECLS-K (Downey, Von Hippel, & Broh, 2004; Fryer Jr & Levitt, 2004). Specifically, the gap in reading is .32 standard deviations and the gap in mathematics is .54 standard deviations (Quinn, 2015). The mathematics gap appears to narrow between kindergarten and third grade, but widens in reading during that same period (Murnane et al., 2006). Another study suggests that mathematics and reading gaps grow between first and second grade, then increase idiosyncratically in subsequent elementary grades (Phillips, Crouse, & Ralph, 1998).

Trends in later grades are less clear, though gaps persist. Research using the National Assessment of Education Progress (NAEP) long-term trends data finds that the mathematics gap widens between ages 9 and 13 (Ferguson, 2001; Neal, 2006; Phillips, Crouse, & Ralph, 1998). Meanwhile, research using the National Education Longitudinal Study showed that the mathematics gap increases by .07 standard deviations per year between grades 8 and 10, and by less than .01 standard deviation between grades 10 and 12 (LoGerfo, Nichols, & Chaplin, 2007). There is also some emerging evidence that, in later grades, a portion of the estimated gap may be due to differential engagement on the test between Black and White students rather than true differences in achievement (Soland, 2018a, 2018b).

These gaps are also associated with important academic and life outcomes, which makes addressing them more pressing. As Heckman and Vytlacil (2001) point out, returns on education tend to be greatest among students in the top quintile of the test score distribution. Thus, Neal's (2006) finding that only 5% of Black students ages 13 and above have mathematics scores in the top quartile of the White math score distribution could have implications for the earnings of Black students. Achievement gaps are also associated with lower college enrollment and persistence (Reardon, Baker, & Klasik, 2012), and differential participation in quantitative college majors and professions (Brown-Jeffy, 2009; Friedman, 1989).

School Effectiveness by Student Subgroup

Not much is known about how much school effectiveness estimates differ by student subgroup, at least not in a growth modeling context. Prior research on differential effects in growth by subgroup has generally found relatively small differences in the effects of teachers and schools across subgroups (Aaronson et al., 2007; Lockwood and McCaffrey, 2009; Loeb et al., 2014; Fox, 2016). For example, Thomas and Collier (2002) showed that schools can be differentially effective for English learners dependent on the instructional program used, but the differences were small in magnitude. Meanwhile, other research has shown that school-level estimates of student achievement and Black-White gaps in that achievement differ dependent on the proportion of racial minority students in the school (Brown-Jeffy, 2009), but that work was not primarily investigating school contributions to that achievement net of student background characteristics.

While there are not many studies that examine school rank orderings (or general effectiveness) by student subgroup, a vast literature considers whether inequality by subgroup grows during the school year. Perhaps most prominently, Downey and Broh (2003) found that schools serve as important equalizers with most gaps growing faster during summer than during the school year (the Black-White gap being the one exception). This issue has been revisited many times over the intervening years. von Hippel, Workman, and Downey (2018) similarly showed that socioeconomic gaps often shrink during the school year and expand during the summer, while the Black-White gap tends to grow during school. However, other work, including research by von Hippel and Hamrock (2019) finds that different gaps expand during different seasons and to varying degrees depending on data sources and measures of achievement. In short, while little is known about the relative effectiveness of schools based on contributions to student growth across racial/ethnic subgroups, considerable research shows that schools have some effect on relative achievement patterns over time.

Teacher Effectiveness by Student Subgroup

A wider range of work on this issue has been conducted at the teacher level. For instance, Loeb et al. (2014) showed that VAM-based rank orderings of teachers were fairly similar for English learner (EL) students versus non-ELs, but that there could be differences sufficient to impact which teachers were identified as extremely low performing. These analyses were conducted by estimating VAM separately for ELs and non-ELs, then comparing rank orderings of teachers. Similarly, Newton, Darling-Hammond, Haertel, and Thomas (2010) showed that teacher rank orderings can shift substantially dependent on the race/ethnicity of the students taught even when controls for student demographics are included in models. Such findings have been confirmed in other studies (Aaronson et al., 2007; Lockwood and McCaffrey, 2009).

Two other studies have examined whether teacher effectiveness differs for initially high- versus low-achieving students. Both found that teachers who are effective with one group generally tend to be effective with the other. In the first study, VAM estimates were produced separately for these two groups of students, and correlations were shown to be moderate (Aaronson, Barrow, & Sander, 2007). The correlations may have been even higher had they been corrected for attenuation (Aaronson et al., 2007). The second study showed that 10% of the variance in teachers' VAM estimates is associated with the interaction between teachers and individual students, suggesting that most of the variance is shared across the two groups (Lockwood & McCaffrey, 2009).

A few other studies examine how exposure to different types of teachers may affect achievement gaps. For instance, Clotfelter, Ladd, and Vigdor (2005) found that novice teachers are distributed among schools and classrooms in a way that disadvantages Black students. In a follow up study, Clotfelter, Ladd, and Vigdor (2006) showed that teacher experience is consistently associated with achievement, and that more experienced teachers are more likely to be sorted into high-income, less racially diverse schools. Both sets of findings could relate to how effective schools are in closing gaps.

There is also some evidence that teachers may bring race-based biases to their perceptions of students, which could affect whether a school is closing gaps dependent on how teachers are sorted into schools. For instance, Dee (2005) showed that racial dynamics between students and teachers have large effects on teacher perceptions of student performance. However, the effects tended to be concentrated among students of low socioeconomic status and in the South. Other work has shown similar results for teacher perceptions of college readiness (Fox, 2015; Soland, 2013) and whether the students are prepared for gifted and talented programs (Grissom & Redding, 2015).

Sample, Measures, and Methods

In this section, the sample, measures used, modeling strategies, and analytic approaches to comparing estimates of school effectiveness are discussed.

Analytic Sample

Student Growth, School Effectiveness, and Gaps

Data were obtained from a cohort of students in a Southern state that administers tests in math and reading during the spring each year. Table 1 provides descriptive statistics on the students in the sample, who ranged from roughly 92,000 to 139,000 in number depending on the year. Although descriptive tables include students of other races (i.e., not White or Black) to provide a better sense of the sample, all analyses used a sample limited only to Black and White students. To ensure that the sample of students matched those from the state in question, an entropy balancing technique was used to re-weight the sample such that it better matches a vector of school-level characteristics obtained from National Center on Educational Statistics (NCES) data. For more details on the method used, please see Hainmueller (2012).

Although a cohort design was employed, the cohort is not intact: students move in and out of the sample so long as they have at least one valid test score. As Table 1 shows, the sample size increased slightly with each year because more and more students in the state were tested using MAP Growth. To make sure results were not sensitive to shifts in the students included over time, the sample was also limited to (a) students who had test scores in both second and fifth grade and (b) students with test scores in every year. The latter is akin to the approach taken in many other studies on school effectiveness that limit the sample only to students with all available test scores (e.g., McEachin & Atteberry, 2017).

Students began in second grade and finished in fifth. The sample was limited to these grades in order to estimate the contributions of schools to students' growth during all of elementary school (excluding Kindergarten and first grade, which are infrequently tested by states). To that end, students were assigned to their modal elementary school, though results were consistent when students could move. For modeling purposes, schools serving fewer than 10 students at a given test administration were excluded (Loeb, Soland, & Fox, 2014). While the models can be estimated when enrollment is below 10 students, such schools are often anomalous in terms of their students and curricular model. For instance, some of these schools educate students with disciplinary infractions, and may use the test as a placement screener. With these restrictions, school contributions to student growth were estimated for 521 schools in reading and 529 in math.

Measures Used

In the state used, virtually all of the students take the Measures of Academic Progress (MAP) Growth, an assessment of math and reading. Scores are reported on the RIT scale, which ranges from approximately 120 to 290 and is a transformation of the logit-based Rasch model estimates of student achievement. The tests are vertically scaled, allowing for certain types of growth models to be estimated. MAP Growth is computer-adaptive, which means students should receive content matched to their estimated achievement regardless of grade level, helping avoid situations where students receive content that is too difficult or easy for them (and, thereby, floor and ceiling effects on the test). Altogether, these attributes of MAP Growth mean one can estimate student growth on a consistent and comparable scale for all time periods in the study.

Methods

Modeling Student Growth. The primary model used in this study was a standard three-level growth model with test scores (y_{tij}) at time (t) nested within students (i) nested within schools (j) of the form:

Level 1 (Test Scores) Eq. 1

$$y_{tij} = \pi_{0ij} + \pi_{1ij}(T_{tij}) + e_{tij}$$

Level 2 (Students)

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(BLACK_{ij}) + r_{0ij}$$
$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(BLACK_{ij}) + r_{1ij}$$

Level 3 (Schools)

$$\beta_{00j} = \gamma_{000} + u_{00j}$$
$$\beta_{01j} = \gamma_{010} + u_{01j}$$
$$\beta_{10j} = \gamma_{100} + u_{10j}$$
$$\beta_{11j} = \gamma_{110} + u_{11j}$$

Variance components of the model are as follows:

$$e_{tij} \sim \mathrm{N}\left(0, \sigma^2_{tij}\right) \qquad \text{Eq. 2}$$
$$\boldsymbol{r}_{ij} \sim \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{T_{\pi}})$$
$$u_j \sim \mathrm{MVN}\left(0, \boldsymbol{T_{\beta}}\right)$$

All models were fit in HLM 7 using a full maximum likelihood estimator.

Using the model, empirical Bayes (EB) estimates of school contributions to student test scores were produced. Per Equation 1, u_{10j} is the school level growth estimate for White students, and u_{11j} is the change in that slope for Black students (i.e., the Black-White gap in growth). Further, EB estimates of school contributions to students' linear growth for White students can be obtained as follows:

$$\hat{\beta}_{10j} = \hat{\gamma}_{100} + \hat{u}_{10j} \qquad \text{Eq. 3}$$

And then compared to the EB estimates of school contributions to linear student growth for Black students, which equals:

$$\hat{\beta}_{10j} + \hat{\beta}_{11j} = \hat{\gamma}_{100} + \hat{u}_{10j} + \hat{\gamma}_{110} + \hat{u}_{11j} \qquad \text{Eq. 4}$$

Finally, Equation 1 was fit also including polynomial terms for student growth (squared and cubic terms for time). While the polynomial model fit the data better, analyses were also conducted using the linear model because EB estimates and random effects correlations are more straightforward to interpret. Thus, both sets of estimates (linear and quadratic) were produced and will be discussed.

Question 1. How different are rank orderings of schools when based on estimates of school contributions to long-term student growth? For this question, estimates were compared in two ways. First, cross-tabulations of the bottom 5% of schools in the distribution of EB growth estimates were generated for Black-versus White estimates, as well as Black versus overall estimates (the model in Equation 1 but with no level-2 covariate for race). Second, these same three sets of estimates were correlated to examine rank orderings more generally. While the growth model used herein is very different than lag-score VAM models used in most of this literature, the approach to comparing estimates by focusing on rank orderings is common in the current literature on teacher and school effectiveness with subgroups, which relies on VAMs.

In addition, a scatterplot that standardizes the EB estimates separately by race was produced. That is, for White students, the standardized estimate equals

$$Z = \frac{\hat{\beta}_{10j} - \bar{\hat{\beta}}_{10j}}{SD(\hat{\beta}_{10j})} \qquad \text{Eq. 5}$$

A comparable effect size is produced for Black students. Thus, the scatterplot of those two race-specific standardized scores shows whether a school's growth for White students relative to the overall school-level distribution of growth for White students is comparable to the growth for Black students in that school relative to the overall school-level distribution of growth for Black students.

Question 2. How many schools have growth estimates that are higher for Black students than White students (suggesting gaps are closing) or vice-versa? For this question, EB estimates for school contributions to student growth were also used, but the means of comparing those estimates was different, and relied in particular on the original scale to better understand growth trends by subgroup. Specifically, two scatterplots of the EB estimates were produced. The first scatterplot used scores on the original RIT scale to show how many RIT a Black student grew on average relative to a White student who grew, say, 5 RIT per year, on average. Obviously, if the Black and White estimates are the same, then they are growing at the same rate, and achievement gaps present in 2nd grade will persist undiminished. Higher growth for Black students suggests the gap in that school is closing.

The second scatterplot is comparable to the one used in Question 1, but standardizes all estimates without regard to race. That is, while each school has a separate Black and White estimate, those estimates are considered part of a single distribution and standardized relative to the standard deviation of all the EB estimates for Black and White students pooled. Such an approach essentially standardizes all estimates, Black and White, based on a common metric. Thus, presenting results in this way shows how far above/below the mean of all school-level EB estimates White students in the school are compared to how far above/below that mean Black students are in standard deviation units.

Question 3. Are gaps widening or narrowing in schools with high growth for White students? In Equation 1, u_{10j} is a parameter capturing school-level growth for White students. In that same model, u_{11j} is a parameter for the Black-White gap in that growth slope. Thus, the correlation between u_{10j} and u_{11j} provides an estimate of the association between growth for White students and growth in the gap. For example, a high negative correlation would mean that, as the growth slope for White students increases, the gap between Black and White students is widening significantly. Reliability adjusted correlations of these parameter estimates can be taken directly from the variance-covariance matrix of the random effects, $\boldsymbol{T_\beta}$.

Results

Student Growth, School Effectiveness, and Gaps

Findings are presented question by question. Results did not shift substantively when the sample included students with scores in both second and fifth grade, nor when using students with at least one test score in each year. The insensitivity may be due to the fact that the test is used so consistently statewide. Thus, results for a sample of students that included anyone with at least one test score are reported.

Question 1. How different are rank orderings of schools when based on estimates of school contributions to long-term student growth?

Tables 2(a) and 2(b) show the cross-classifications of schools in the bottom 5% of the school effectiveness distribution (EB estimates) in math. Table 2(a) compares the estimate for Black students to that of White students, and Table 2(b) compares the estimate for Black students to the overall (unconditional estimate). Generally, there is a fair amount of consistency in classification. When comparing Black and White growth estimates, 17 schools would be in the bottom 5% for both sets of students, but 10 would be in the bottom 5% for Black students but not White students. Differences in classification rates are similar for Black students relative to the overall (Black and White unconditional) sample estimates. Estimates for Black versus White students are generally highly correlated (~.94 in math).

Tables 3(a) and 3(b) show the same results in reading. Here, results are much more consistent by subgroup. For example, in 3(a), 25 schools are in the bottom 5% based on both sets of estimates, and only 4 schools total move out of the bottom 5% dependent on race. Further, Black and White estimates are correlated .99 in reading.

As discussed in the methods section, models were also fit that included a polynomial growth term. However, rank orderings and schools in the Bottom 5% did not change substantively. Therefore, only linear results are reported.

Student Growth, School Effectiveness, and Gaps

Figure 1 shows a scatterplot of EB growth estimates standardized separately by race. The figure also uses unique markers by quadrant of the graph. For instance, schools in the lower left and upper right quadrants are producing below-average or above-average growth (respectively) for both White and Black students. According to this standardization, most of the schools are consistently above or below average for both races. In math, there are 26 schools out of 529 that had below-average growth for Black students but above-average for White students. In reading, there are 7 schools out of 521 that had below-average growth for Black students but above-average for White students.

Question 2. How many schools have growth estimates that are higher for Black students than White students (suggesting gaps are closing) or vice-versa?

Table 4 shows school-level fixed effects estimates and variance components from the model in Equation 1. As the table shows, mean achievement for White students in 2nd grade was roughly 197 RIT, and Black students had an average RIT score that was 7.4 points lower. Meanwhile, White students grew linearly at about 7.8 RIT per year between second and fifth grade, whereas Black students grew about .78 RIT slower. Thus, on average, gaps tend to be increasing over time. Table 5 shows the same results in reading. Here, White students grew about 6.4 RIT per year with Black students rowing roughly .4 RIT per year slower. Once again, gaps tend to be increasing over time, though not as substantially.

Figure 2 shows a scatterplot of EB estimates in math using the original RIT scale, and includes both an identity line and a regression line fit to the data. As the figure shows, school-level estimates of student growth are higher for White students than Black students. For example, in schools with a mean slope of 7 RIT for White students, Black students are growing just over 6 RIT. One should note that only two schools are above the identity line. In other words, only two schools appear to be closing the gap in math achievement consistently during elementary school.

Figure 3 presents the same plot in reading. Here, the gap between the regression and identity lines is much lower and appears to decrease as school-level growth for White students increases. Once again, no schools have estimates above the identity line. Therefore, no schools appear to be closing the gap in math achievement consistently during elementary school.

To help illustrate this point, Figure 4 includes histograms of EB estimates for the school-level gap in the growth slope between Black and White students (math and reading). As the figure shows, virtually all schools have Black students who are growing slower, on average, compared to White students. In math, many Black students are growing roughly one RIT per year slower, on average, compared to White students. To put this result in context, in fifth grade mathematics nationally, schools growing at the 50^{th} percentile have a mean gain of 8.6 RIT and schools at the 40^{th} percentile have a mean gain that is roughly one RIT lower. Thus, one RIT worth of growth per year can mean a substantive difference in where a school's growth falls in the national rank orderings based on MAP Growth achievement and growth norms (Thum & Hauser, 2015).

Figures 5 and 6 replicate the scatterplots from Question 1, but standardize EB estimates without regard to race. In math, 234 schools show above-average growth for White students, but below-average for Black students. In reading, that number is much lower, but still high at 161 schools. For both subjects, no schools produce above-average growth for Black students with below-average growth for White students, perhaps the most likely scenario in which the Black-White gap might close.

Question 3. Are gaps widening or narrowing in schools with high growth for White students?

In this question, model-based correlations among random effects estimates are discussed. In reading, the random effects for the White slope and the Black-White gap in that slope are correlated .17. This suggests that, the larger the growth for White students in a given school relative to the mean growth for White students, the smaller the gap in growth between White and Black students. Therefore, in reading, there is a "rising tide lifts all boats" phenomenon when it comes to growth by race. By contrast, the correlation in math is -.38. In plain terms, the larger the growth for White students in a given school relative to the mean growth for White students, the larger the gap in growth between White and Black students. Therefore, in math, the schools producing the fastest growth in math tend to be leaving Black students behind even further.

Discussion

Estimating school contributions to student growth is gaining emphasis among policymakers, practitioners, and researchers. In particular, ESSA has paved the way for most states to weight school contributions to student growth heavily in their implementation plans (CCSSO, 2017). Such laws place great emphasis on rank orderings of schools, including identifying those as extremely low performing. One primary motivation for emphasizing growth is to help low-achieving students catch up and, thereby, close achievement gaps by race and biological sex. Current literature is somewhat mixed on whether teacher and school effectiveness differs by student subgroup (Loeb et al., 2014; Newton et al., 2010). Perhaps based on the nature of accountability policy, virtually all related research focuses on causal estimates of school contributions to student growth in the short term, and compares estimates based primarily on rank orderings of schools.

Student Growth, School Effectiveness, and Gaps

In this study, when purely examining rank orderings, one could argue that school effectiveness is not all that different for Black versus White students. For example, EB estimates of school contributions to student growth are all correlated well above .9 and, while the schools deemed in the bottom 5% do shift by race, the majority of schools are consistently classified as being extremely low-performing. If one standardizes the estimates separately by race, the majority of schools showing above-average growth for White students are also showing above-average growth for Black students. Thus, when we talk about rank orderings of schools as has been done in the literature—standardizing subgroup estimates separately and comparing rank orderings (e.g., Fox, 2016; Loeb et al., 2014)—not much is amiss.

Yet, these findings mask a more insidious story. Black students begin 2nd grade with much lower achievement in math and reading than their White counterparts. Further, whereas White students are growing roughly 7.75 RIT per year in math, on average, Black students are growing about 7 RIT per year, or about 10% lower than White students. Thus, not only are achievement gaps unlikely to close over time; they are, by all indications, growing during elementary school. When EB estimates were produced separately by race, but those estimates were pooled then standardized, no schools are providing above-average growth for Black students but below-average growth for White students. While high growth for Black students and lower growth for White students is not necessarily a desirable outcome, it is the most likely scenario under which achievement gaps will close. By contrast, regardless of how school-level estimates are standardized, many schools produce above-average growth for White students, but below-average for Black students. When EB estimates are standardized without regard to race, 234 schools in math and 161 schools in reading (out of roughly 530 schools) fall into this category. In sum, rank orderings and classifications of low-performing schools tell an incomplete story. Schools can be ranked similarly across races, and yet be contributing to larger achievement gaps over time.

Limitations and Future Research

This study has limitations worth mentioning. An obvious one is generalizability. Data are from one state. Therefore findings may not generalize beyond the state, nor perfectly to the population of students within that state (though the entropy balancing approach used should mitigate the within-state issue). Replication is warranted.

Another disadvantage of the data is that I do not have student covariates often used in the VAM literature. In particular, while the data include each student's race, biological sex, and achievement scores, the data do not have socioeconomic, special education, or English learner status. School-level covariates were more complete because the data could be merged with those form the National Center for Education Statistics (NCES). Thus, models included the same covariates as those used by McEachin and Atteberry (2017), including school proportions of White, Black, Hispanic, and free or reduced price lunch students. The models also controlled for total enrollment and whether the school is urban or rural.

Implications for Policy

What, then, are the implications for policy, and school accountability in particular? Under NCLB, subgroup performance was compared based on percent of students at proficient on the state test. These percent proficient metrics often showed gaps by subgroup, and could certainly be adapted to growth metrics for ESSA. However, as others have pointed out (Ho, 2008, 2009), proficiency-based metrics are hugely problematic given their sensitivity to scaling and proficiency thresholds. This body of research strongly suggests that proficiency or cut-score based metrics should be avoided. Thus, policymakers could benefit from using a metric that in some way quantifies that gap, but is less scale-dependent than a simple proficiency cut score.

One option might be to use estimates not unlike the ones in this study (even if not based on as many years of test scores). That is, schools could be ranked on their overall contributions to student growth, as well as the gap in the growth slope between, say, White and Black students. Implementing such a model in practice would not be overly complicated relative to some of the lag-score VAMs currently in use. That said, a growth modeling approach benefits from (though is not necessarily dependent on) having a vertically scaled test. Not all states currently use such a test (ESSA Plans, 2017). Further, fitting a growth model like the one in this study requires multiple years of data, and there are compelling practical and philosophical reasons to hold schools accountable in the short-term.

Whatever the specific approach, policymakers may wish to consider what the primary goals of the education system are, and for schools in particular. Is an ideal school one that is growing all of its students equally? Or one that is growing its high-achieving White students sufficiently to master grade-level content, but is also growing those Black students who are lower achieving even faster? The latter is a situation in which gaps would close without any serious detriment to White students. In order to identify such a situation—reasonable growth for all students, but higher growth for initially low-achieving subgroups—policymakers would need to move beyond rank orderings and rely more heavily on test scales. That is, by definition, schools cannot grow all students above average (even in Lake Wobegon). However, states can use tests that are scaled such that growth at each point on the scale can be tied to specific academic content. Such a scale then creates an opportunity to see if White students are learning what they should, and lower achieving Black students are still catching up. In short, when it comes to growth, determining school effectiveness across subgroups should probably not simply be comparative.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. *The Future of Test-Based Educational Accountability*, 155–170.

Brown-Jeffy, S. (2009). School effects: Examining the race gap in mathematics achievement. *Journal of African American Studies*, *13*(4), 388.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*(2), 190–215.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review*, *95*(2), 158–165.

Downey, D. B., Von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, *69*(5), 613–635.

Downey, D., & Broh, B. (2003). *Are schools the great equalizer?: School and non-school influences on socioeconomic and Black/White gaps in reading skills*. Columbus, OH: The Ohio State University, Columbus.

Student Growth, School Effectiveness, and Gaps

Ferguson, R. (2001). Test Score Trends Along Racial Lines, 1971-1996: Popular Culture and Community Academic Standards. *America Becoming: Racial Trends and Their Consequences*, *1*, 348–90.

Fox, L. (2015). Seeing potential: The effects of student–teacher demographic congruence on teacher expectations and recommendations. *AERA Open*, *2*(1), 2332858415623758.

Fox, L. (2016). Playing to teachers' strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy, 11*(1), 70–96.

Friedman, L. (1989). Mathematics and the gender gap: A met-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, *59*(2), 185–213.

Fryer Jr, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447–464.

Grissom, J. A., & Redding, C. (2015). Discretion and disproportionality: Explaining the underrepresentation of high-achieving students of color in gifted programs. *Aera Open*, *2*(1), 2332858415622175.

Heckman, J., & Vytlacil, E. (2001). Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. *Review of Economics and Statistics*, *83*(1), 1–12.

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

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351–360.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201–228.

Student Growth, School Effectiveness, and Gaps

Jencks, C., & Phillips, M. (2011). *The Black-White test score gap*. Washington, DC: Brookings Institute Press.

Klein, A. (2016). The every student succeeds act: An ESSA overview. *Education Week*, 114–95.

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439–467.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, *36*(4), 457–475.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education*, *4*(4), 572–606.

McEachin, A., & Atteberry, A. (2017). The impact of summer learning loss on measures of school performance. *Education Finance and Policy*, *12*(4), 468–491.

Murnane, R. J., Willett, J. B., Bub, K. L., McCartney, K., Hanushek, E., & Maynard, R. (2006). Understanding trends in the black-white achievement gaps during the first years of school [with comments]. *Brookings-Wharton Papers on Urban Affairs*, 97–135.

Neal, D. (2006). Why has black–white skill convergence stopped? *Handbook of the Economics of Education*, *1*, 511–576.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*, 23.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the Black-White test score gap widen after children enter school. *The Black-White Test Score Gap*, 229–272.

Student Growth, School Effectiveness, and Gaps

Quinn, D. M. (2015). Kindergarten Black–White test score gaps: Re-examining the roles of socioeconomic status and school quality with new data. *Sociology of Education*, *88*(2), 120-139.

Reardon, S. F., Baker, R., & Klasik, D. (2012). Race, income, and enrollment patterns in highly selective colleges, 1982-2004. Stanford, CA: Center for Education Policy Analysis at Stanford University.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*(3), 853–891.

Soland, J. (2013). Predicting high school graduation and college enrollment: Comparing early warning indicator data and teacher intuition. *Journal of Education for Students Placed at Risk (JESPAR)*, *18*(3–4), 233–262.

Soland, J. (2018a). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12).

Soland, J. (2018b). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323.

Thomas, W. P., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement.*

Thum, Y. M. (2003). Measuring progress toward a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research*, *32*(2), 153–207.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "are schools the great equalizer?" *Sociology of Education*, *91*(4), 323–357.

Table 1

Analytic Sample Descriptive Statistics

Year	Term	Grade	Test Admin.	Student Count	Race & Gender Proportions				Mean Achievement (RIT Scale)	
					Black	Latinx	White	Female	Math	Reading
2011	Spring	2	1	92,122	0.331	0.075	0.512	0.511	191.990	189.367
2012	Spring	3	2	104,614	0.331	0.073	0.514	0.510	204.730	200.479
2013	Spring	4	3	132,748	0.328	0.075	0.515	0.510	213.651	207.436
2014	Spring	5	4	138,896	0.326	0.076	0.516	0.508	220.796	212.596

Table 2(a) - Math

Cross-tabulations of Schools in the Bottom 5% Based on White Students Versus Black Students

		Black Bottom 5%		
		No	Yes	Total
White Bottom 5%	No	492	10	502
	Yes	10	17	27
	Total	502	27	529

Table 2(b) - Math

Cross-tabulations of Schools in the Bottom 5% Based on All Students (Black and White) Versus Black Students

		Black Bottom 5%		
		No	Yes	Total
Overall Bottom 5%	No	494	8	502
	Yes	8	19	27
	Total	502	27	529

Table 3(a) - Read

Cross-tabulations of Schools in the Bottom 5% Based on White Students Versus Black Students

		Black Bottom 5%		
		No	Yes	Total
White Bottom 5%	No	492	2	494
	Yes	2	25	27
	Total	494	27	521

Table 3(b) - Read

Cross-tabulations of Schools in the Bottom 5% Based on All Students (Black and White) Versus Black Students

		Black Bottom 5%		
		No	Yes	Total
Overall Bottom 5%	No	491	3	494
	Yes	3	24	27
	Total	494	27	521

Table 4

Estimates from Linear Growth Models Conditional on Race - Math

Fixed Effects	Coeff.	S.E.	t-ratio	p-value
Intercept (White)	197.139	0.251	785.470	<.001
Mean achievement in 2nd grade - White students				
Intercept (Black)	-7.376	0.182	-40.278	<.001
Mean difference in 2nd grade achievement - Black students				
Linear growth (White)	7.753	0.044	175.551	<.001
Mean difference in linear growth (Black)	-0.782	0.040	-19.664	<.001

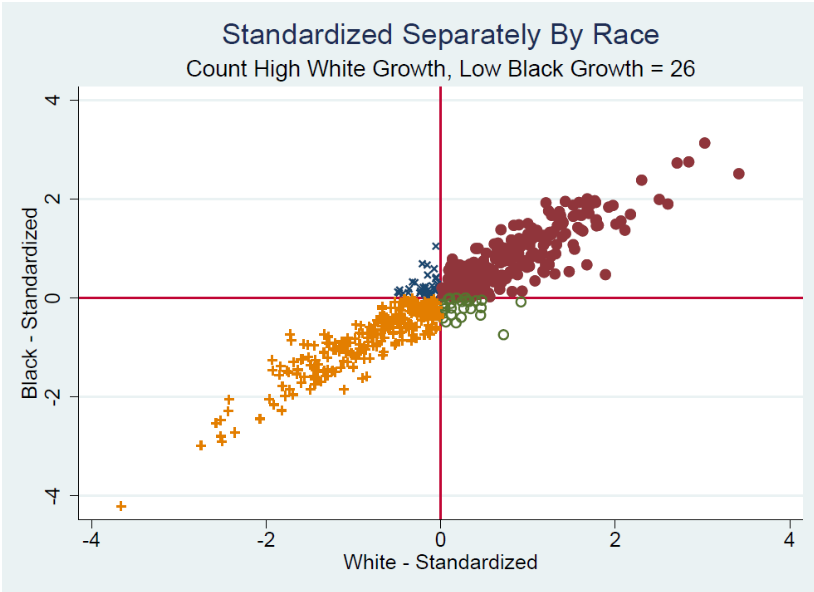
Variance Components - Level-3	SDs	Variances		p-value
Intercept (White)	5.305	28.150		<.001
Mean achievement in 2nd grade - White students				
Intercept (Black)	2.565	6.582		<.001
Mean difference in 2nd grade achievement - Black students				
Linear growth (White)	0.864	0.747		<.001
Mean difference in linear growth (Black)	0.426	0.181		<.001

Table 5

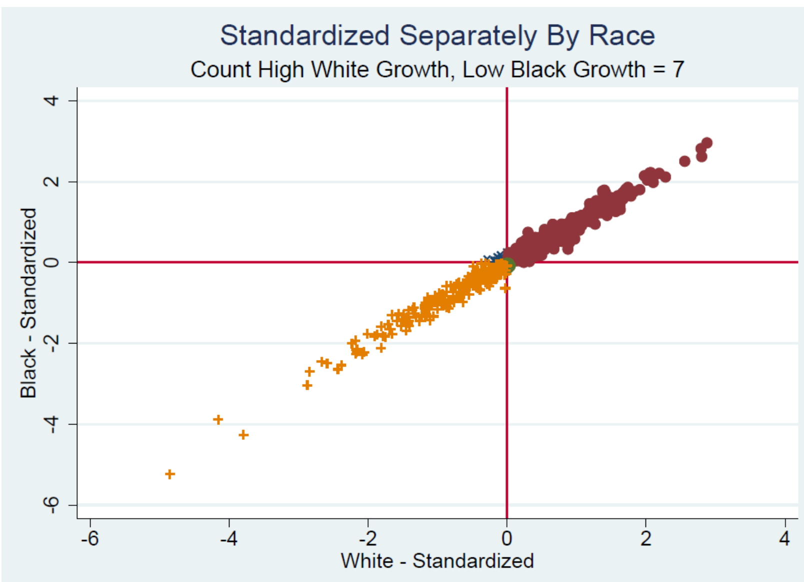
Estimates from Linear Growth Models Conditional on Race - Reading

Fixed Effects	Coeff.	S.E.	t-ratio	p-value
Intercept (White)	194.523	0.262	743.239	<.001
Mean achievement in 2nd grade - White students				
Intercept (Black)	-7.642	0.233	-32.791	<.001
Mean difference in 2nd grade achievement - Black students				
Linear growth (White)	6.396	0.036	178.509	<.001
Mean difference in linear growth (Black)	-0.404	0.037	-10.785	<.001

Variance Components - Level-3	SDs	Variances		p-value
Intercept (White)	5.277	27.842		<.001
Mean achievement in 2nd grade - White students				
Intercept (Black)	3.242	10.509		<.001
Mean difference in 2nd grade achievement - Black students				
Linear growth (White)	0.642	0.413		<.001
Mean difference in linear growth (Black)	0.271	0.074		<.001



Math



Reading

Figure 1. Scatterplots of Black versus White EB estimates of school effectiveness standardized by subgroup

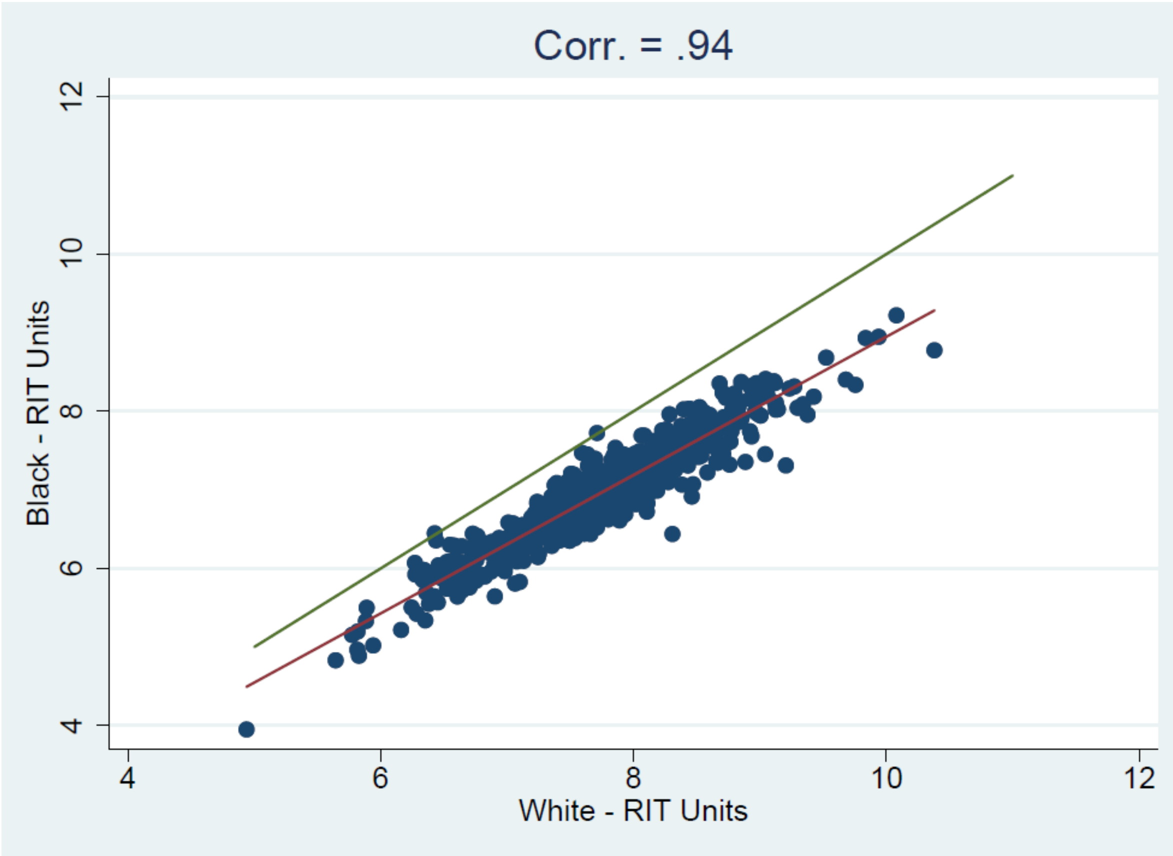


Figure 2. Scatterplot of school-level linear growth estimates in math for White versus Black students using the original RIT scale with a regression line (red) and identity line (y = x, in green).

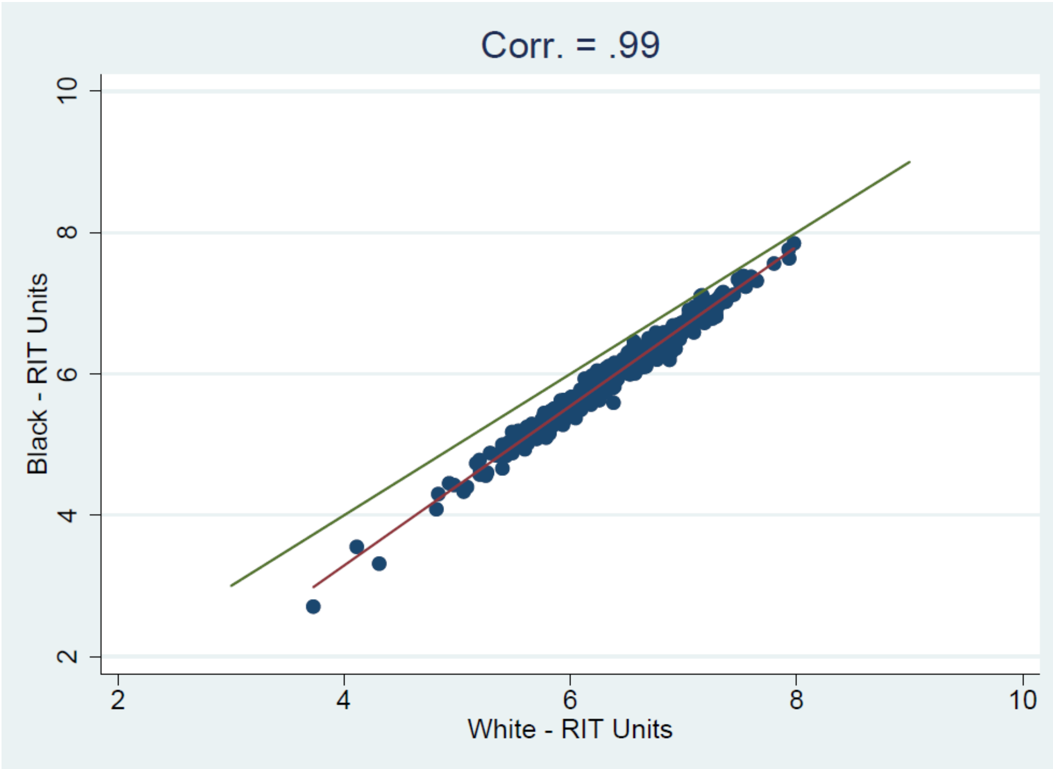
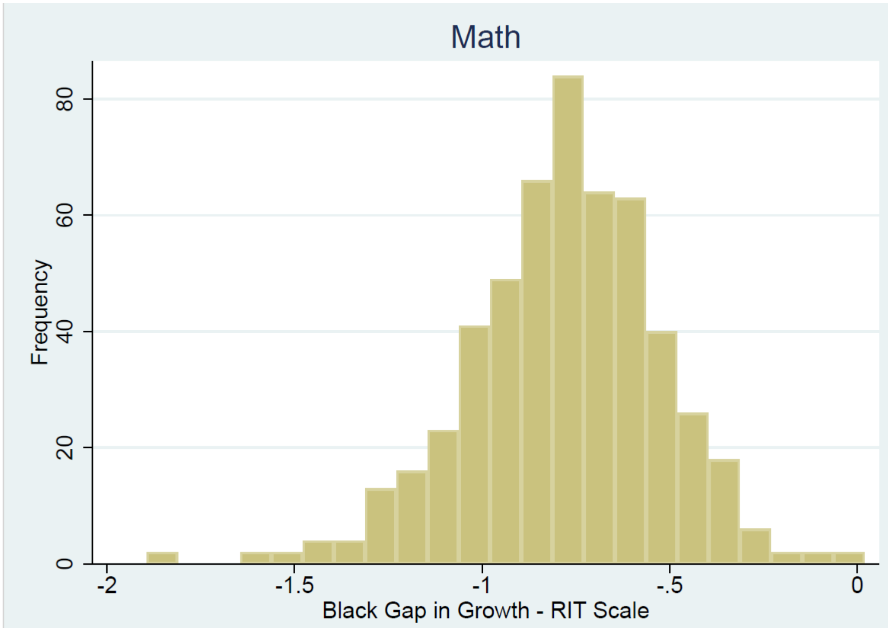


Figure 3. Scatterplot of school-level linear growth estimates in reading for White versus Black students using the original RIT scale with a regression line (red) and identity line (y = x, in green).



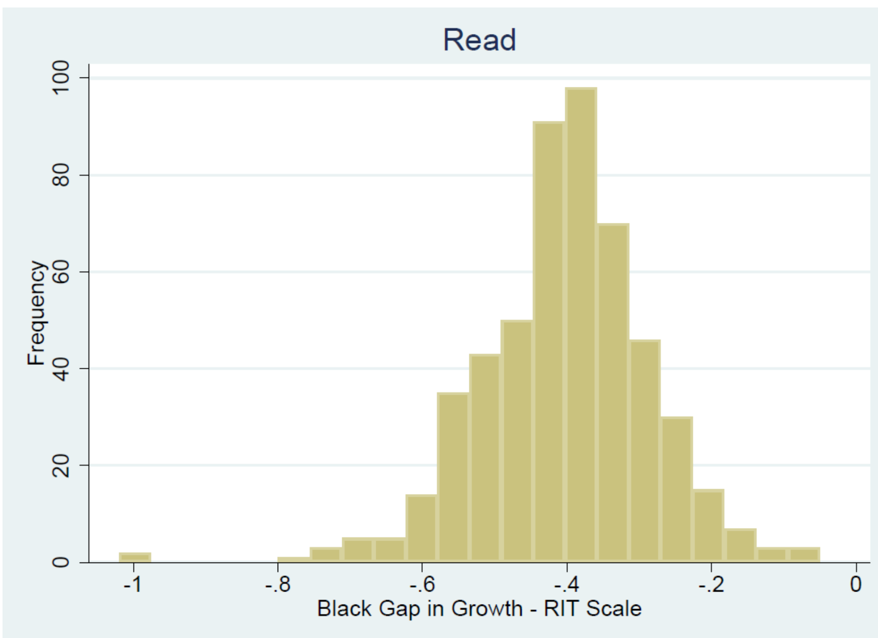


Figure 4. Histograms of EB estimates of the Black-White gap in growth by school.

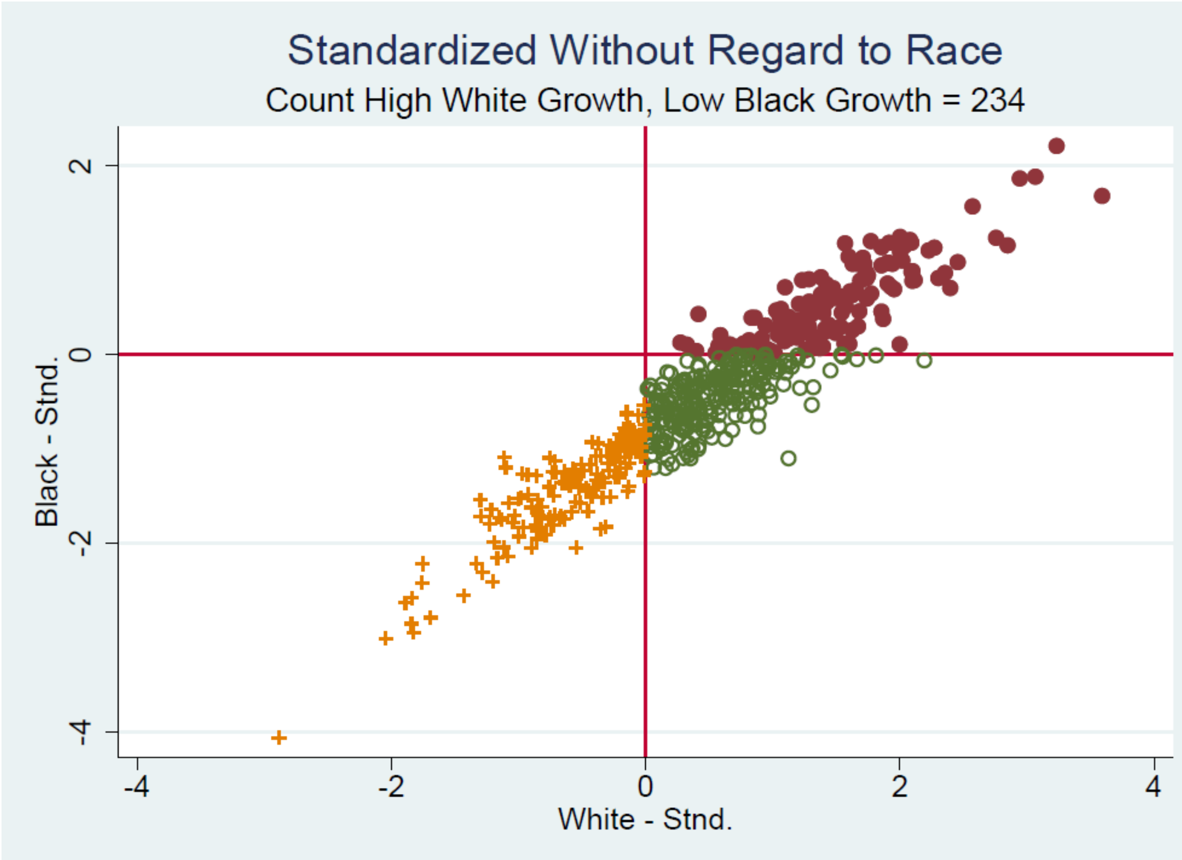


Figure 5. Scatterplot of EB growth estimates standardized without regard to race in math.

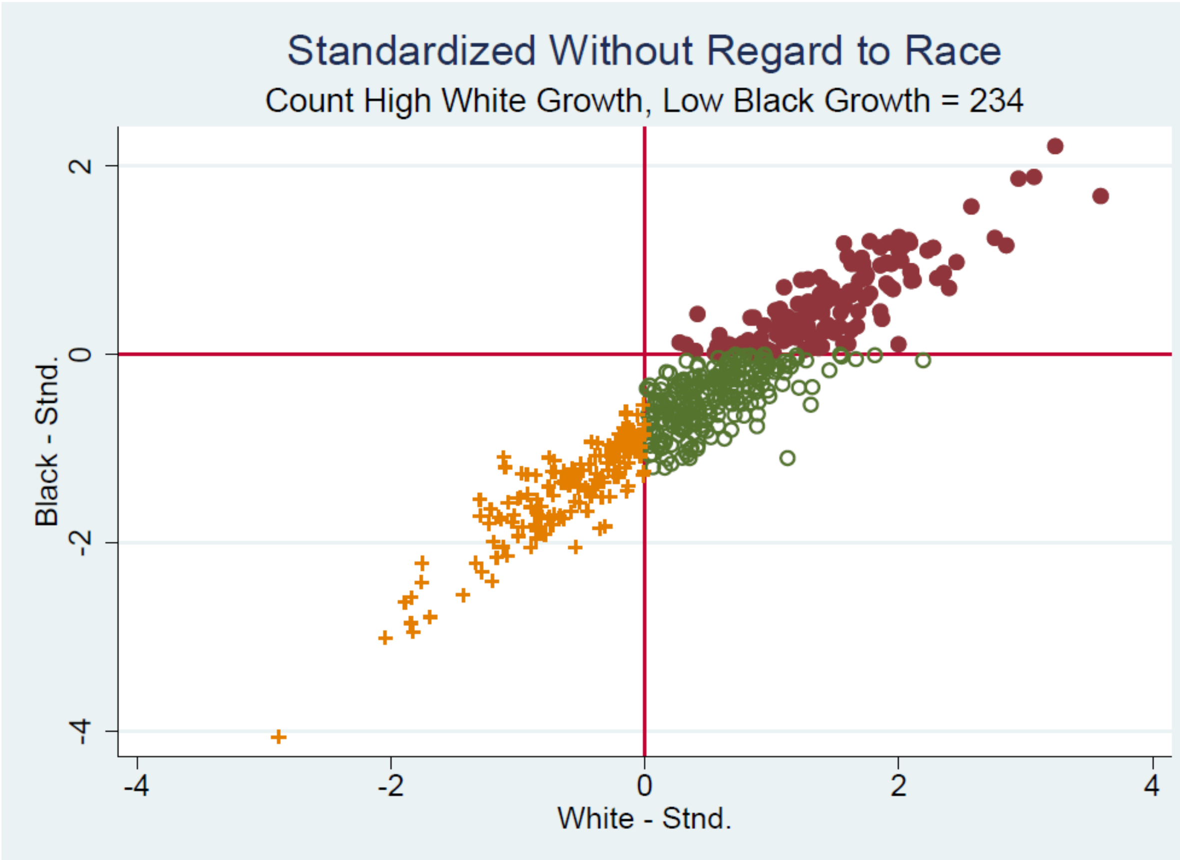


Figure 6. Scatterplot of EB growth estimates standardized without regard to race in reading.