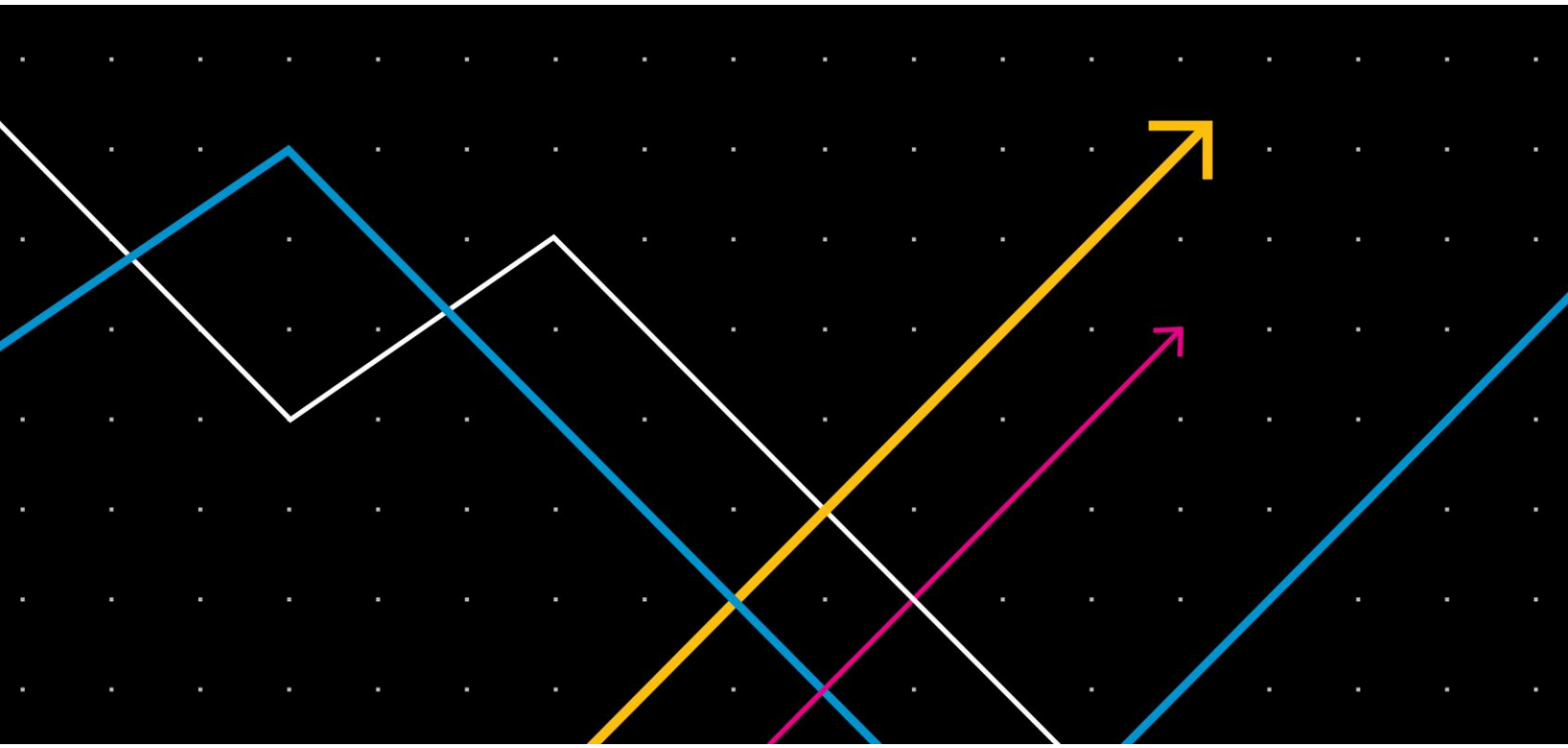


RESEARCH REPORT

Can We Measure School Quality Using Publicly Available Data?

Tomas Monarrez *Matthew Chingos*

July 2020

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments

This report was funded by the Bill & Melinda Gates Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

We also thank Grace Luetmer for excellent research assistance, the North Carolina Education Research Data Center for sharing the student-level data we used in this analysis, Kristin Blagg and Constance Lindsay for helpful conversations in the early stages of this project, and Erica Blom and Doug Wissoker for comments on an earlier draft of the report.

Executive Summary

Each year, state education agencies release school-level achievement results on assessments in mathematics and reading. Several actors use this information to compare school quality, including policymakers, parents, and the media. But school quality measures based on test scores are often biased because students that are likely to do well on tests, regardless of the school they attend, tend to cluster together in certain schools and avoid others altogether.

In this study, we assess whether publicly available data on school test scores and student characteristics can be used to generate high-quality measures of schools' effects on student achievement. We aim to determine whether adjusted test scores that account for correlations between school demographics and achievement still contain significant bias and whether they get us close enough to schools' true effectiveness to be useful for at least some purposes.

We begin by computing "north star" measures of school quality by applying best practices from the applied econometrics literature on school "value-added" estimation, using student-level administrative data from North Carolina schools. We propose an innovative method of adjusting publicly available data, leveraging the results of a student-level analysis based on restricted-use data from the National Assessment of Educational Progress (NAEP). We assess the validity of the results of our adjustment to publicly available data by comparing it with the north-star measures of school quality.

We find that our adjustment performs slightly better than a more traditional regression adjustment that uses only school-level data. Our approach produces estimates of school quality that are modestly correlated with the north-star value-added measures, with correlation coefficients of about 0.70 in math and 0.61 in reading. When we instead use proficiency data published by the US Department of Education's EDFacts, these correlations decrease to 0.63 in math and 0.50 in reading. This is at least partly the result of measurement error introduced into the EDFacts data to protect student privacy.

Next, we use EDFacts to apply our adjustment methodology to schools nationwide, applying state-specific estimates of the within-school correlation between student demographics and test performance generated from student-level NAEP data. We cannot compare these measures with value-added metrics outside North Carolina, but we find that the adjustment changes the ranking of schools relative to raw proficiency rates in the right direction, though the way this plays out varies by state and type of school. Along with this report, we have published code allowing other users to compute our adjusted measures.

Our results show that demographic adjustments to public data are an improvement on unadjusted proficiency rates in that they are more highly correlated with school value-added estimates. Additionally, using our proposed methodology produces marginal improvements over traditional regression methods. But considerable bias and noise likely remain in the adjusted test scores.

A key implication of these findings is that publishing school quality ratings or making policy decisions based on these measures is fraught. Adjusted test score measures may better inform efforts by researchers, policymakers, philanthropists, parents, and others seeking to understand the role that nonschool factors play in distorting comparisons of average test scores across schools, especially in states that publish only aggregate data.

Can We Measure School Quality Using Publicly Available Data?

Each year, state education agencies release school-level achievement results on assessments in mathematics and reading. Several actors use this information to compare school quality, including parents choosing schools for their children, policymakers implementing school accountability policies, and philanthropists seeking to understand whether their investments are improving student performance. One prominent example is the popular website GreatSchools.com, which generates public school ratings based on school test score data for almost every US school.

But the usefulness of test score data for the measurement of schools' impacts on students is limited by student sorting into schools. School quality is the impact a school has on student outcomes (relative to other schools they could have attended), net of all other factors that may influence student outcomes, including family background, innate scholastic ability, neighborhood effects, and other factors that determine both cognitive and noncognitive development. It is difficult to assess school quality using test score data because schools' enrollments are selected. The students in a given school are seldom representative of a school system's student body (Monarrez, Kisida, and Chingos 2019). Thus, a school may have high (low) test scores both because it is high (low) quality or because it enrolls students that are more advantaged (disadvantaged) in aspects relevant to achievement, and there is no way to know which explanation is correct.

For example, schools that largely serve students from groups that tend to score lower on tests, such as economically disadvantaged students, are likely to have lower average scores than schools that largely serve affluent students. Thus, average test scores are a biased measure of school quality because students that are likely to do well on tests, regardless of the school they attend, cluster together in certain schools and avoid others altogether. Rewarding or punishing schools on this basis tends to reward affluent schools while punishing schools that educate disadvantaged students.

Furthermore, theory suggests that rewarding schools based on test scores—say, by giving them a good rating in a popular website—creates perverse incentives for school equity. First, schools will have incentives to attract affluent students and limit the entry of disadvantaged students to remain competitive and attractive to parents (Abdulkidaroglu et al. 2019; Rothstein 2006). Second, parents seeking the best education for their children will be more likely to choose residences associated with highly rated schools, which will, perhaps inadvertently, point them to segregated neighborhoods.

Indeed, evidence has suggested that the expansion of GreatSchools.com may have increased neighborhood segregation in many cities (Hasan and Kumar 2019).

Some states address these concerns by publishing school quality metrics based on adjusted test score estimates that capture student growth over time, such as "value-added" measures (Angrist et al. 2017) or student growth percentiles (Betebenner 2011). These metrics control for past student achievement using regression methods and typically depend on access to longitudinal student records.

There is a large literature in applied econometrics on the estimation of educational value-add. Much of this research focuses on estimating teacher value-add within schools, but the machinery of the adjustment is the same when estimating school value-add. The key difference is that student selection is more pervasive and extensive along the school selection margin than the teacher selection margin within schools. Still, evidence shows that teacher value-added models can be unbiased if they use the correct controls (Chetty et al. 2014), though this is still a matter of debate because of counterevidence (Rothstein 2010, 2017).

Recent contributions to the applied literature have placed the merit of school value-added measures on stronger footing (Angrist et al. 2017). These contributions were made possible by school lottery data, which generates random variation in student assignment that can be leveraged to estimate schools' true average impact on student achievement. Angrist and coauthors show that, when comparing school value-added estimates based on lottery variation to conventional value-added estimates, the conventional measures get it right, on average, but they do not accurately estimate the value-add of individual schools.[1] An implication of these results is that even though the estimate for any one school is imprecise, estimates of average quality for large groups of schools may have better properties.

The applied academic literature has thus established that the conventional value-added measures we can construct with longitudinal student records (which we treat as the "truth" in this report) still suffer from selection bias. To be sure, conventional value-added estimates are highly correlated with school effectiveness, but there is still significant bias in the estimate for any one school. So even though value-added estimates of school quality that use best practices from the literature are better than quality metrics based on unadjusted test scores, their overall usefulness in informing policy is still a matter of academic debate.

Further, the data on school academic performance made available to the public are coarse and prone to measurement error. Many states publish only the share of students at a school that score at or above a state-defined proficiency threshold on each test. The US Department of Education gathers

these proficiency rates and publishes them through its EDFacts initiative. This is the only publicly available government source of data on school test scores on a national scale.[2] Still, there are important measurement error limitations in these data because EDFacts purposefully degrades the quality of proficiency data to protect student privacy in small schools. Nonetheless, these are often the best available data on school performance. Understanding how well they can be used to assess school quality is important for education policy.

In this report, we ask, How can publicly available data on school proficiency rates and characteristics be best used to make inferences about the impact schools have on their students' test scores? We implement several statistical tests assessing whether regression-adjusted proficiency rates that use only publicly available data can get us closer to school value-added estimates based on restricted-access student records and best practices from the literature. The latter are constructed as "north star" estimates of how schools add value, using administrative student records from North Carolina middle schools, following the models specified in Angrist et al. (2017). We then construct proficiency rates for the same sample of schools and adjust them using the school-level characteristics that are commonly available in public datasets, such as racial and ethnic composition and the share of students who receive free lunch, who are in special education, and who have limited English proficiency.

These correlations are generally positive but modest in magnitude. There is considerable variation in school value-add that is not captured by the adjusted proficiency rates, implying that there is considerable bias and noise in them. The problem is that we cannot tell the bias and noise parts from each other with the available data.

In addition, we report the performance of a "pseudo-regression" adjustment to proficiency data that uses coefficients estimated with restricted-access student-level data from NAEP. The restricted NAEP data are a nationally representative sample of student achievement, but it does not include data on all schools. We use these data to estimate within-school relationships between student characteristics and achievement by state, which we use in our adjustment. Intuitively, we attempt to capture part of the achievement variation that cannot be explained by schools more accurately than is possible with public data. These pseudo-regression-adjusted measures have a higher correlation with our preferred estimates of value added than the standard regression adjustment, though the improvement is small. We conclude that this kind of adjustment may improve on conventional methods, but more work is needed to assess its value. These measures are still likely to be biased and noisy predictors of our best estimates of value added (which are biased themselves, but to a lesser extent).

Finally, we implement our adjustment methodology nationally using EDFacts data on proficiency rates for most US schools. We document how conclusions about the distribution of school quality would change depending on whether we use raw proficiency rates or our adjusted ones to define quality. Along with this report, we have published the necessary code and supplemental data necessary for other interested parties to implement these adjustments. Our hope is that making this technical material available will foster further work and discussion on the accurate school quality measurement. We conclude the report with a descriptive analysis of the type of schools that are most likely to change in relative quality rankings as a result of these adjustments.

Data

We use longitudinal student records from the North Carolina Education Research Data Center (NCERDC) to construct benchmark school value-added measures. These data contain the universe of students attending public schools in North Carolina and include annual information on test performance and demographics, including race or ethnicity, gender, special education status, socioeconomic status (free and reduced price lunch), and English language learner status. We use the NCERDC data to compute school value-added models of student test performance controlling for demographics and past performance from 2012 to 2017. We also generate school-level data that mimic public datasets by aggregating the NCERDC records to the school level and constructing adjusted proficiency rates. We focus on middle schools for this analysis.

In addition, we construct various school quality measures based on publicly available school-level proficiency data from the 2015–16 EDFacts. We make demographic adjustments to these data that are feasible using publicly available sources, drawing on demographic data from the Common Core of Data, which lets us measure school racial and ethnic composition and the share of students receiving free lunch, and the Civil Rights Data Collection, which allows us to capture the share of students who have limited English proficiency or are in special education. Finally, we use regression coefficients estimated from the restricted-use data at the student level from NAEP, which are representative both nationally and of each of the 50 states and the District of Columbia.

Choosing a North Star

To assess methods for adjusting outcomes data, we need to know what we are aiming for. This means we need to compute the best possible estimate of schools' causal effects on student achievement using

observations of student outcomes, demographics, and school assignments. This daunting task is the subject of a lengthy literature in the applied economics literature. We borrow ideas from the literature to construct "north star" estimates of school quality. Our work borrows heavily from Angrist and coauthors (2017), who evaluate different value-added model specifications.

Ideally, we would measure the causal effects of attending every school by randomly assigning students to schools. In such a world, all schools would enroll a random mix of students, such that average differences in achievement between schools could be directly interpreted as schools' causal impacts on students. But in the real world, we have to worry about student selection into schools. If students that are more likely to do well in assessments (regardless of the school they attend) are more likely to cluster in certain schools, average differences in school achievement provide biased inferences about schools' causal effects. Such naïve comparisons between schools conflate school quality with student characteristics. But we can partially adjust for differences in student attributes using multivariate regression methods. This is the key insight of the school value-added literature.

Because the NCERDC does not have lottery data to test our value-added estimates against, we follow the best practices Angrist and coauthors used, and we assume that their preferred models for Boston public schools have similar properties in North Carolina schools. We use student-level longitudinal data from North Carolina to construct different measures of school quality in 2017 and years prior. We focus on middle schools, defined as any schools serving grades 6 through 8, because we can measure student performance at the beginning and end of the middle school years (which is not the case for elementary schools because state testing begins in third grade in North Carolina and most other states).

Our middle school value-added models include a key control variable, fifth-grade test scores, which could not have been affected by the current school (which is not the case for sixth- and seventh-grade scores). Angrist and coauthors (2017) examine growth from fifth grade to sixth grade.[3] Intuitively, including this control variable gives our estimates of school quality their value-add. The idea is that we base our school comparisons within groups of students that had similar test scores at the end of elementary school. This follows the same intuition of student growth models that attempt to measure the change in (not the level of) student achievement attributable to a school. In addition, we control for the demographic variables available in these datasets: race or ethnicity, gender, free lunch status, limited English proficiency status, and special education status.

It is worth stressing that the principal limitation of school-level public datasets is their inability to include students' prior test scores. Without this control, we cannot make sure we are comparing

students with similar prior achievement, and our estimates using school-level data will be biased regardless of whether we include demographic controls. Even though certain variables are correlated with prior achievement, such as race or ethnicity and socioeconomic status, there is still substantial achievement variation among students with identical demographics. Adjusting for demographics does not help us account for this variation, making demographics imperfect proxies for student achievement, and the use of imperfect proxies can cause bias in school quality estimates. We direct readers to the appendix for a formal description of our model.

We first estimate value added using the NCERDC data for the 2012–16 middle school cohorts. If these estimates are sufficiently precise and we assume that school quality is stable from one year to the next, one way we can partially assess the stability of these estimates is to check whether they accurately predict next year's cohort's achievement (Chetty, Friedman, and Rockoff 2014). We make this test operational by fitting models of 2017 student test scores on our school value-added estimates and student controls and observing the regression coefficient on value added, commonly called "lambda" for these types of tests. A lambda coefficient of 1 would imply that school value-added estimates have a one-to-one relationship with future student achievement that cannot be explained by demographics. A lambda of 0 would imply that the value-added estimates have no predictive power for next year's cohort achievement.

Table 1 shows the results of our tests of how well value-added estimates predict future student performance. The table also shows the estimated standard error of lambda, its corresponding 95 percent confidence interval, the size of the 2017 cohort used in the estimation, and the R-squared of the forecast model (i.e., its "goodness of fit"). The results suggest that our preferred value-added measures are quite predictive of future student performance. The lambda coefficients are between 0.80 and 0.90, so even though we can reject that the coefficient is 1 (i.e., our predictions are still imperfect), the value-added estimates demonstrate substantial stability over time. But without imposing strong assumptions about how selection bias may evolve, these estimates tell us little of the degree of bias in the value-added measures.

Still, it is encouraging to see that value added is highly predictive of student success in the future. The rows in table 1 assess whether the number of cohorts used in the value-added estimation model affect the model's predictive power. Our estimates of lambda trend noticeably downward as we use more cohorts of students in our value-added estimation. This result leads to the unintuitive conclusion that more years of data actually hurts the predictive power of our value-added estimates. This downward trend likely reflects regression to the mean in school value-add (i.e., some schools look better

in some year by chance, and they return to their normal levels the next year) and the fact that school quality changes over time, so older data are less useful for predicting the future than more recent data.

Altogether, we find that the 2016 cohort estimates using a panel data structure and a full set of demographic and lagged outcome controls are quite predictive of future performance. But even the most predictive measures do not show a one-for-one relationship between value-added measures and future performance. This likely reflects the presence of selection bias, measurement error, and changes in school quality. Our north-star estimates are thus imperfect, even though they are the best we can do with the available data.

TABLE 1

How Well Value-Added Measured Predict Future Student Performance

Demographics + fifth-grade test score adjustment

			Confidence Interval			
	Lambda	**SE**	**Lower**	**Upper**	***N***	R^2
2016	0.887	0.016	0.856	0.918	599,772	0.751
2015–16	0.858	0.018	0.822	0.893	599,783	0.749
2014–16	0.835	0.023	0.790	0.879	599,783	0.748
2013–16	0.822	0.027	0.770	0.874	599,783	0.746
2012–16	0.808	0.027	0.755	0.860	599,783	0.745

Source: Authors' calculations using North Carolina Education Research Data Center data.
Notes: SE = standard error.

How to Adjust Data in the Real World

The analysis above shows that value-added estimates using best practices from the applied literature can have desirable properties (even though they are biased), but student-level longitudinal data systems are not publicly available, so we now explore how to use publicly available data to make accurate inferences about school quality. Publicly available data typically are averaged for schools and districts, with breakdowns by student subgroups (e.g., race or ethnicity), and are often reported in terms of the share of students who are proficient (not average scale scores), as we discuss above.

We can mimic the usual constraints of publicly available data by constructing school-level average outcomes and characteristics using the NCERDC data. We test four measures based on school-level test score data, by comparing them with our north-star estimates. There are two key limitations when using these data. First, we cannot control for students' prior test scores, so we cannot interpret these measures as student growth. Second, we lose information when we aggregate from the student level to the school level. Our first two candidate school quality metrics are as follows:

1. Unadjusted average school achievement or school proficiency rates.
2. Measures based on ordinary least squares (OLS) residuals constructed from school-level regressions that control for student demographics: the share of students at the school who are female, black, Hispanic, Asian, or another race; who receive free or reduced-price lunch; who have limited English proficiency; and who have a disability. Specifically, we estimate the equation

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where Y_i is the average test score or proficiency rate at school i, and X_i contains all our controls. Our estimate of school quality is the estimated OLS residual, $\hat{\epsilon}_i = Y_i - \hat{\beta} X_i$, from this regression.

We also test the performance of the pseudo-regression adjustment using the coefficients from a student-level regression. We do this because the coefficients from the school-level regression might be biased if the control variables are highly correlated with school causal effects. For instance, if schools that have high shares of students receiving free lunch also tend to be low-quality schools, the coefficient on the share of students receiving free lunch will capture the impact of structural inequities associated with economic need (assuming free lunch status captures this) and the fact that the schools that serve these students are not the best, perhaps because of underfunding, high teacher turnover, or other factors. Estimating the relationship between proficiency and free lunch status using only within-school variation takes care of this issue, providing "clean" estimates of the relationship between socioeconomic status and achievement, but this requires student-level data.

Our student-level models include school fixed effects and generate the coefficients $\hat{\beta}_W$, where the W stands for "within school." The pseudo-regression-adjusted measures of school quality are defined as follows:

$$\hat{\epsilon}_{Wi} = Y_i - \hat{\beta}_W X_i.$$

We use "pseudo" because we are going beyond traditional econometric modeling by employing coefficient estimates generated from one dataset to make adjustments to a different dataset with a different sample and level of aggregation.

We perform the pseudo-regression adjustments using two data sources to estimate the $\hat{\beta}_W$:

1. the pseudo adjustment estimated with coefficients computed using NCERDC student-level records on the same demographic characteristics as above

2. the pseudo adjustment estimated with coefficients computed using state-representative student-level data from the restricted-access NAEP files

We summarize comparisons between our preferred value-added measures and these four candidate measures using scatterplots and the correlation coefficient, which theoretically ranges from −1 (perfect inverse correlation) to 1 (perfect correlation). The square of this coefficient is equivalent to R-squared, which tells us how much of the variation in the north-star measures is explained by a given measure.

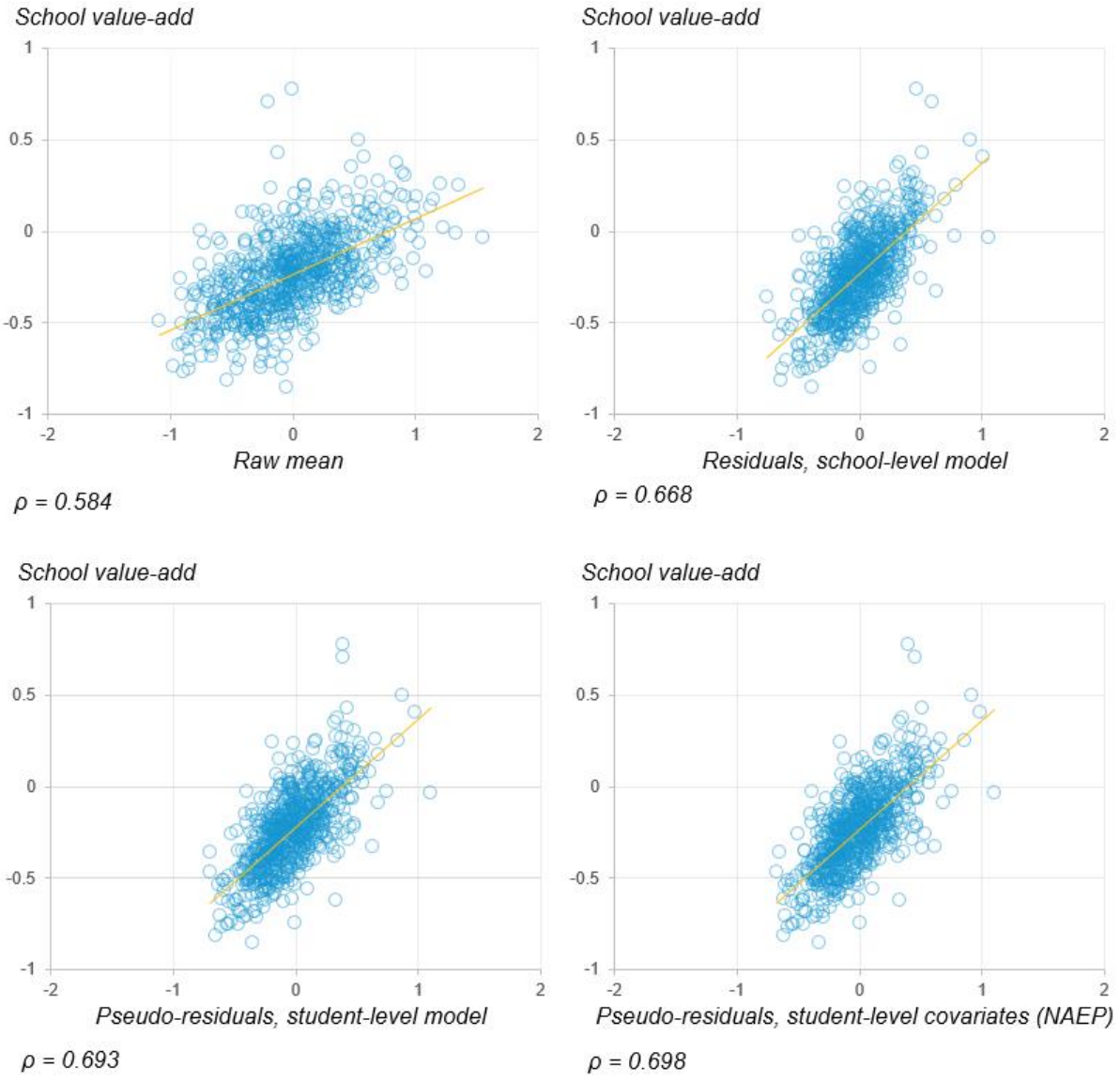
Figure 1 plots our north-star value-added estimates against various measures of school quality based on school-average math achievement data (averaged across sixth grade to eighth grade) and adjustments using school-level demographic data. Panel 1 shows that the correlation between value added and raw mean school achievement is 0.58, which means that average scores explain about 34 percent of the total variation in school value-add. Also notable is that the distribution of value-add is compressed relative to the distribution of mean scores. This implies that there is much less variability in quality between schools than one would expect observing average school scores alone.

Panel 2 plots value-add against residual mean scores using an OLS adjustment based on school-level covariates (measure 2, above). The correlation improves somewhat, to 0.67, suggesting that the traditional adjustment increases the correlation between value added relative to unadjusted mean scores. Panel 3 shows the pseudo-adjusted measure using coefficients estimated in a student-level model of NCERDC data. The correlation coefficient increases slightly to 0.69, implying that this adjusted measure explains 48 percent of the variation in school value-add. Panel 4 shows that using coefficients estimated from the sample of North Carolina students who took the eighth-grade math NAEP test generate similar results. This is encouraging, as it suggests that similar properties may hold for NAEP pseudo-regression adjustments applied to other states, which we explore in the next section.

FIGURE 1

Predicting School Value-Added Estimates Using Aggregate Proxies of School Quality

Standardized mathematics achievement in grades 6 through 8 in North Carolina schools



URBAN INSTITUTE

Source: Authors' calculations using NCERDC data.

Notes: NAEP = National Assessment of Educational Progress; NCERDC = North Carolina Education Research Data Center. This figure presents scatterplots of "north star" school value-add on the y-axis and various measures of school quality computed with school-level data based on standardized eighth-grade math scores on the x-axis. School value-added estimates are school fixed effect estimates in a regression of 2016 middle school student achievement controlling for student demographics and lagged student achievement measures in fifth grade. Aggregate (school-level) covariates denote an ordinary least squares adjustment for school-level covariates that are computed using NCERDC data. Student-level covariates denote an adjustment using coefficient estimates from a student-level regression that includes school fixed effects, using NCERDC data. Student-level covariates (NAEP) correspond to adjustments based on the coefficients of a student-level regression using NAEP samples from North Carolina.

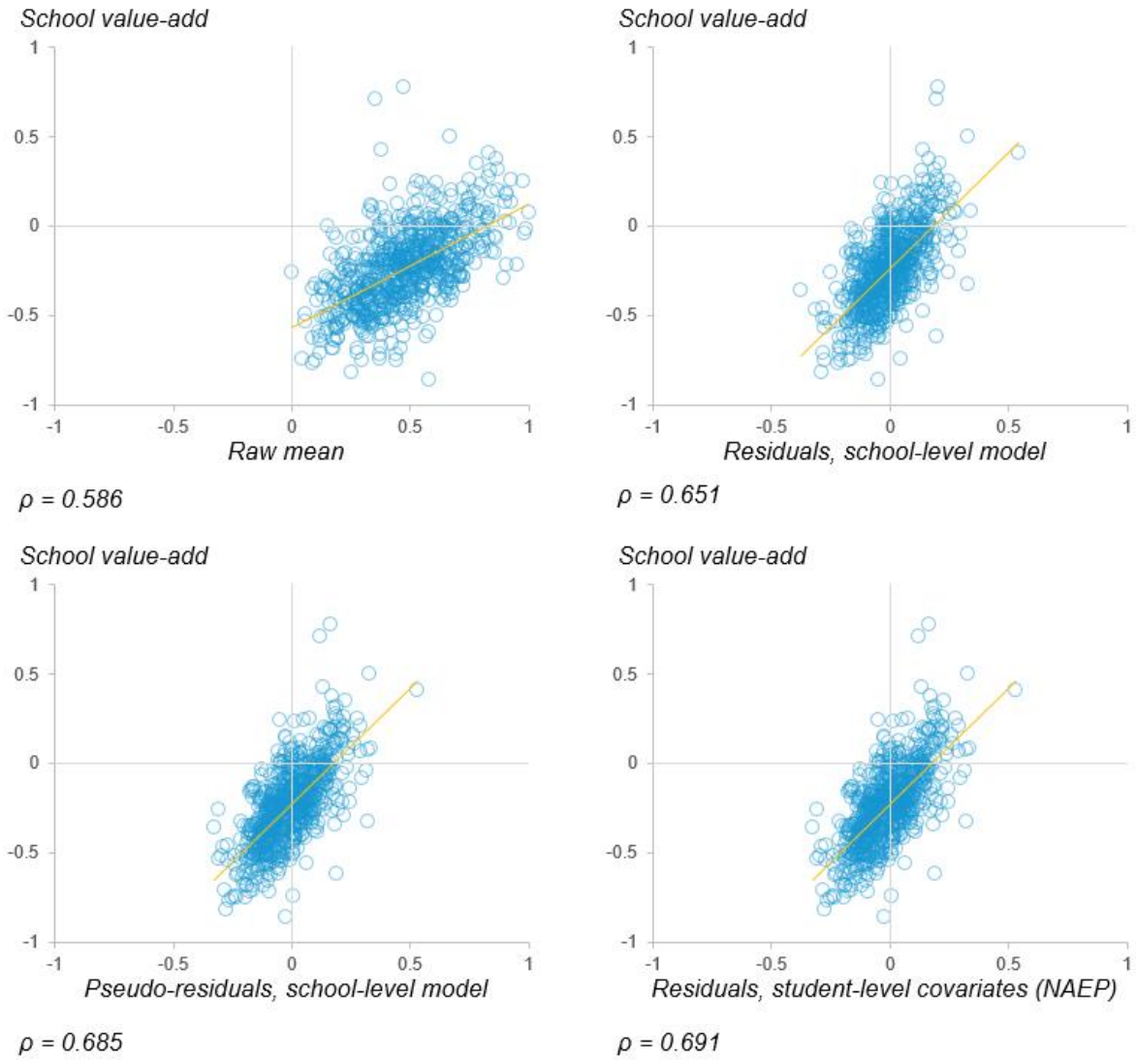
Figure 2 uses student proficiency rates in math to predict school value-add. For the NAEP data, we set a proficiency threshold that produces the same overall proficiency rate as on the state test. Our estimates suggest a similar pattern of results. School-level adjustments based on parameter estimates from student-level regressions perform better than traditional adjustments using school-level regressions. Notably, the NAEP student-sample pseudo adjustment produces similar results to traditional adjustment using the NCERDC student population data.

In addition, even though proficiency rates throw out a lot of information about student performance above and below the proficiency cutoff, the correlations of value added with measures based on proficiency rates are only slightly lower than those with measures based on test scores. When we compute similar scatterplots for reading scores and proficiency rates (appendix figures A.1 and A.2), results are qualitatively similar in direction but weaker in magnitude, with correlations that are generally lower and much less of an improvement from panels 1 to 4 (0.53 to 0.55 for scores and 0.52 to 0.53 for proficiency rates).

FIGURE 2

Predicting Preferred Estimates of School Value-Add Using Aggregate School Proficiency Rates

Standardized mathematics proficiency rates in grades 6 through 8 in North Carolina schools



URBAN INSTITUTE

Source: Authors' calculations using North Carolina Education Research Data Center data.

Notes: NAEP = National Assessment of Educational Progress. This figure presents scatterplots of "north star" school value-add on the y-axis and various measures of school quality computed with school-level data based on eighth-grade proficiency rates on the x-axis. School value-added estimates are school fixed effect estimates in a regression of 2016 middle school student achievement controlling for student demographics and lagged student achievement measures in fifth grade. Aggregate (school-level) covariates denote an ordinary least squares adjustment for school-level covariates that are computed using North Carolina administrative data. Student-level covariates denote an adjustment using coefficient estimates from a student-level regression using North Carolina administrative data.

In summary, evidence suggests that the school-level NAEP pseudo-regression adjustment can produce estimates that match our preferred estimates of quality slightly more closely. More specifically, although we are aware that even our best value-added measures may be biased, using them as a benchmark estimate of school quality has allowed us to establish that school-level achievement data that are "pseudo-residualized" from school-level demographics using coefficients from a NAEP student-level model is slightly more highly correlated with quality than traditional OLS-adjusted measures. But this improvement is marginal. The result is useful, though, because we can easily apply the pseudo-regression adjustment based on NAEP to achievement data from other states as well, using the publicly available EDFacts data on school proficiency rates. But moving to the public data comes with its own limitations.

Most prominent among these limitations is that we can seldom compare estimates of school quality coming from EDFacts to higher-quality value-added estimates stemming from student-level longitudinal databases. We thus set out to confirm that the results above hold when using school-level North Carolina proficiency data from EDFacts. Another limitation of the public data is that EDFacts data are censored to protect privacy and often report proficiency rates in "bins" (e.g., 57 percent would be reported as 55 to 60 percent or 40 to 60 percent, depending on school size). For schoolwide performance of schools with substantial enrollment, these bins are generally narrow, but they are wider in small schools. This introduces an additional measurement error component to our analysis, urging further caution in interpreting our results out of context. Furthermore, some schools have missing proficiency information, meaning that the sample of schools differ between the NCERDC and EdFacts datasets. We enforce equality between the sample in the results shown below by making our NCERDC sample of schools match the schools in EdFacts reporting proficiency data.

Table 2 presents correlation estimates between our preferred school value-added measures and adjusted measures based on EdFacts. We also show the correlation estimates using school-level data from the NCERDC that we presented above, for comparison purposes. The results show that the correlation between the pseudo-regression estimates of quality using EdFacts and our preferred school value-added estimates from NCERDC is somewhat attenuated compared with estimates that use NCERDC school-level data. This is to be expected, given that EdFacts data contain considerable measurement error.

Nonetheless, it is notable that the qualitative pattern of the EdFacts results is similar to what we reported above using NCERDC school-level data. Adjustments of the EDFacts school-level proficiency data based on the NAEP microdata are correlated 0.63 with value-add in math, compared with 0.57 for raw proficiency rates. The results also suggest that the pseudo-regression adjustment to EdFacts data

provides a marginal improvement to the traditional school-level adjustment, moving the correlation from 0.59 to 0.63. This is true for both math and reading scores, although correlations are more attenuated for reading.

TABLE 2

Correlations of School Value-Add and Measures of School Quality in North Carolina

Based on various adjustments of proficiency rates and data sources

	Math	Reading
NCERDC (confidential data)		
Raw proficiency	0.586	0.510
School-level OLS adjustment	0.651	0.563
Student-level OLS adjustment	0.685	0.624
School-level pseudo-OLS adjustment (NAEP)	0.691	0.614
EdFacts (public or censored data)		
Raw proficiency	0.566	0.472
School-level OLS adjustment	0.590	0.459
School-level pseudo-OLS adjustment (NAEP)	0.634	0.505

Source: Authors' calculations using NCERDC and EdFacts data.

Notes: NAEP = National Assessment of Educational Progress; NCERDC = North Carolina Education Research Data Center; OLS = ordinary least squares. School value-added estimates are school fixed effect estimates in a regression of 2016 middle school student achievement controlling for student demographics and lagged student achievement measures in fifth grade. Aggregate (school-level) covariates denote an OLS adjustment for school-level covariates that are computed using NCERDC data. Student-level covariates denote an adjustment using coefficient estimates from a student-level regression that includes school fixed effects, using NCERDC data. Student-level NAEP pseudo-regression estimates correspond to adjustments based on the coefficients of a student-level regression using NAEP samples from North Carolina.

We next consider whether our adjustments have meaningful impacts on relative school rankings. We first compare how rankings change as we move from a metric based on raw proficiency rates to a metric based on our preferred value-added estimates. We then see how things change when we go from raw proficiency rates to the NAEP pseudo-residual measure. Table 3 presents the results. We use quintiles of the distribution of each of these variables to approximate rankings without having to list all schools in the data. Schools in low quintiles are low performers, while schools in high quintiles are high performers, according to each metric. Table 3 summarizes the joint distribution between both measures and provides more information than the correlation coefficient.

The cross-tabulation in the upper panel of table 3 shows that even though raw proficiency quintiles are positively correlated with school value-added quintiles, there is considerable dispersion. Schools in the lowest raw proficiency quintile tend to be clustered in quintiles with lower value-add, but 8 percent of the schools are above the median, and 11 percent are around the median. On the other end of the spectrum, schools with high proficiency rates are not necessarily those in the highest quintile of value added. In fact, 15 percent of these high-achieving schools are near the median, while 11 percent are in

the two lowest quintiles. Finally, schools with median proficiency rates are equally likely to be considered in the highest or lowest quintile of value added.

The lower panel of table 3 shows a parallel summary of the joint distribution of two school quality metrics, comparing raw proficiency rates with the pseudo-regression-adjusted measures using NAEP coefficients. We focus on the NCERDC data to limit the role of measurement error, though results using EdFacts look similar (available upon request). Although the pseudo-regression adjustment also changes rankings relative to raw proficiency rates, it does so to a lesser extent than the value-added estimates. For instance, although 8 percent of schools in the lowest quintile of raw proficiency are moved to highest two quintiles of value-add, no schools are moved this high up by the pseudo-regression-adjusted measure. Symmetrically, no schools in the highest quintile of raw proficiency are moved down to the lowest quintile of the pseudo-regression adjustment, but for value-added-based quintiles, this happened for about 11 percent of schools.

Because we have a good sense that the degree of selection bias in the value-added estimates is lower than in the pseudo-regression estimates, the results in this section suggest two main takeaways. First, the traditional OLS and the pseudo-regression adjustment push school quality estimates in the same direction as value added. Quality metrics based on the pseudo-regression adjustments are positively correlated with our best value-added estimates to a slightly higher degree than the traditional OLS adjustment. These adjustments influence school rank changes in a similar way and direction as the value-added measures and may significantly alter the relative rankings of schools.

Second, the adjustment is not particularly powerful, statistically. Though positively correlated with value added, these correlations are of modest magnitude, never exceeding 0.7. Rank changes based on the pseudo-adjusted measures are attenuated relative to those of the value-added measure. But the available data and econometric methods leave us with little to say about how much the weakness of the adjusted measures is because of noise or bias. Thus, we advise caution in overinterpreting these types of estimates for the purpose of policy or rating schools.

TABLE 3

Quintile Rank Changes from Raw Proficiency Rates to Quintiles of Adjusted School Quality Measures

Based on various adjustments of proficiency rates and data sources

	Quintile of School Value-Add					
	1	**2**	**3**	**4**	**5**	***N***
Quintile of raw proficiency rate						
1	78	42	17	9	3	149
	52.3%	28.2%	11.4%	6.0%	2.0%	
2	45	37	27	22	18	149
	30.2%	24.8%	18.1%	14.8%	12.1%	
3	18	36	39	37	19	149
	12.1%	24.2%	26.2%	24.8%	12.8%	
4	7	19	43	47	33	149
	4.7%	12.8%	28.9%	31.5%	22.1%	
5	1	15	23	34	75	148
	0.7%	10.1%	15.5%	23.0%	50.7%	
N	149	149	149	149	148	744

	Quintile of Pseudo-OLS Adjusted Measure					
	1	**2**	**3**	**4**	**5**	***N***
Quintile of raw proficiency rate						
1	101	39	9	0	0	149
	67.8%	26.2%	6.0%	0.0%	0.0%	
2	39	47	35	27	1	149
	26.2%	31.5%	23.5%	18.1%	0.7%	
3	8	51	54	23	13	149
	5.4%	34.2%	36.2%	15.4%	8.7%	
4	1	12	46	63	27	149
	0.7%	8.1%	30.9%	42.3%	18.1%	
5	0	0	5	36	107	148
	0.0%	0.0%	3.4%	24.3%	72.3%	
N	149	149	149	149	148	744

Source: Authors' calculations using North Carolina Education Research Data Center and NAEP data.

Note: NAEP = National Assessment of Educational Progress; OLS = ordinary least squares. Student-level NAEP pseudo-OLS estimates correspond to adjustments based on the coefficients of a student-level regression using NAEP samples from North Carolina.

Adjusting Performance Data beyond North Carolina

We next perform the pseudo-regression adjustment to public EDFacts data for all states, using within-school coefficients from a set of state-specific models based on restricted-access NAEP data. Specifically, for each state separately, we estimate a student-level OLS model of proficiency indicators on school fixed effects and the following student demographics: free lunch status, limited English proficiency status, participation in special education programs, female indicator, Hispanic indicator, black indicator, Asian indicator, and an indicator for not being in any of these racial or ethnic groups

(white is the base category). The presence of school fixed effects implies that the estimated relationships come from within-school comparisons, ensuring that our coefficients are not biased by the potential correlation between student characteristics and latent school quality.

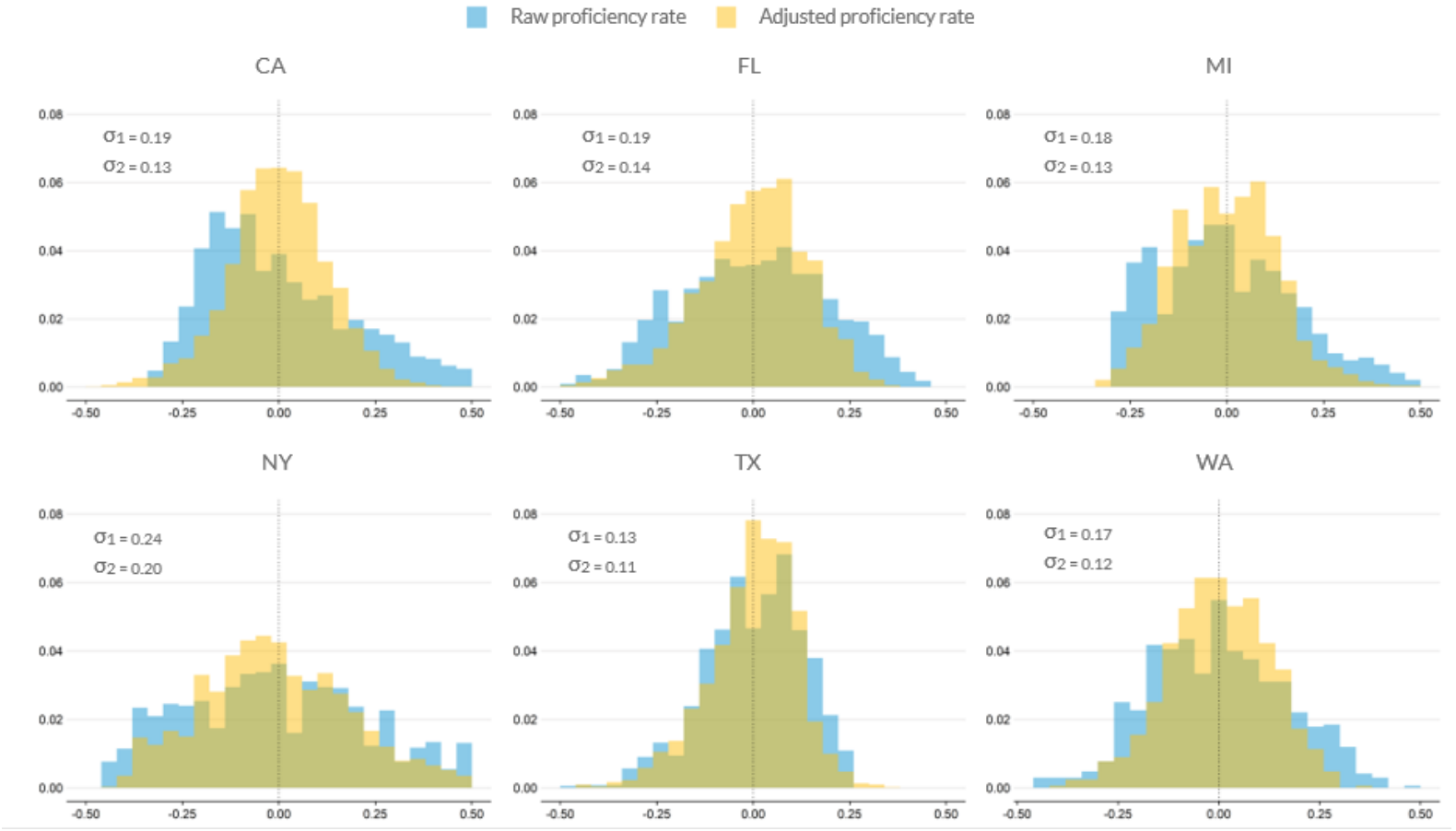
We cannot validate the analysis using high-quality value-added measures in states other than North Carolina, but we can compare NAEP-adjusted measures with raw proficiency rates. Figure 3 explores differences between the raw proficiency rate and the NAEP pseudo-regression-adjusted rate in California, Florida, Michigan, New York, Texas, and Washington. These plots are recentered so that the distribution of both measures has a mean of zero. Zero, therefore, represents the average school in the state according to each metric. The histograms highlight differences in the overall spread of these measures. As expected, the regression adjustment tightens the variability in the distribution of the school quality measure. This has the implication that the overall "quality difference" between the best and worst schools in the state is smaller when we adjust for student demographics. In the econometrics literature focusing on the causal identification of school value-add, such tightening of the distribution after adjusting for observable characteristics is treated as an indication of the extent of selection bias in the raw measures of student achievement (Angrist et al. 2017). We therefore report the standard deviation of both of these distributions for each state in the figure.

The distribution of school quality is compressed, albeit only slightly, by our regression adjustments, suggesting that our adjusted measures clean up some of the selection bias in the raw proficiency measures. The degree of compression varies by state. California's estimated school proficiency distribution is compressed the most (of the six states presented here) by the demographic adjustment, going from a standard deviation of 0.19 to 0.13. In other states, the adjustment leads to a smaller compression of the distribution, ranging from 1 to 5 percentage points, or at most 70 percent of the spread of the raw proficiency rates. For context, Angrist and coauthors (2017) report that the standard deviation of average school achievement is double the size of the standard deviation of their school value-added estimates. By this benchmark, we can conclude that our adjustment is not entirely ineffectual at eliminating selection bias, though this varies considerably by state.

FIGURE 3

Distribution of Unadjusted versus Adjusted Proficiency Rates in Six States

Mathematics proficiency in public schools serving eighth grade



URBAN INSTITUTE

Source: Authors' calculations using 2015–16 data from EdFacts, the Common Core of Data, and the Office for Civil Rights.

Note: NAEP = National Assessment for Educational Progress. NAEP pseudo-regression adjustments are computed using coefficients from a student-level regression using student-level NAEP.

We also look at the statistical relationship between unadjusted and adjusted school quality measures for the same six states (figure 4). In every state, the two measures are positively correlated, as we would expect, given that we are plotting the dependent variable in a regression on its estimated residuals. But there is considerable variation in adjusted rates among schools with similar raw proficiency rates.

This variation is informative of how much the adjustment matters across different levels of raw proficiency. For example, among the California middle schools in which 40 percent of students score proficient in math, the adjusted rate ranges by about 40 percentage points, from 10 percentage points below average to more than 30 percentage points above average. In contrast, for similar levels of raw proficiency (about 40 percent) in Texas, the regression adjustment varies in a tighter range, less than 10 percentage points. If we take these adjusted measures of school quality as unbiased, schools with 40 percent raw proficiency vary in quality to a greater degree in California than in Texas.

Thus, the "shape of the cloud" in the scatterplots presented in figure 4 is informative of the range in the distribution of the raw proficiency rate for which the regression produces the largest adjustments. This can be used to assess, say, whether schools with low proficiency rates versus those with high proficiency rates are more likely to change in relative rank once the regression adjustment is implemented. For Texas, the scatterplot has a tight range for low-proficiency schools and a wide range for high-proficiency schools, meaning that schools with low proficiency rates have a small range of adjustment compared with schools with high proficiency rates. In Michigan, the opposite is true: schools with low proficiency rates have larger adjustment ranges than those with high proficiency rates (the scatterplot range thins from left to right on the horizontal axis). Both Texas and Michigan differ markedly from California and New York, for which the regression adjustment generates about the same variability in our measure across the range of raw proficiency.

What can we gather from knowing the adjustment matters more at different points in the distribution of raw proficiency? These patterns are suggestive of the degree of selection at different levels of proficiency, where "selection" refers to the set of covariates used in the adjustment (e.g., race or ethnicity, free and reduced-price lunch status, and gender). For instance, if high-proficiency schools have large adjustments relative to low-proficiency schools, certain covariates likely have a strong correlation with proficiency and high-proficiency schools are unrepresentative of these strong covariates relative to the state average. An example could be that schools with high proficiency rates have a smaller share of students receiving free and reduced-price lunch than the average school, leading to large downward adjustments for them. Intuitively, these plots highlight the areas of the distribution

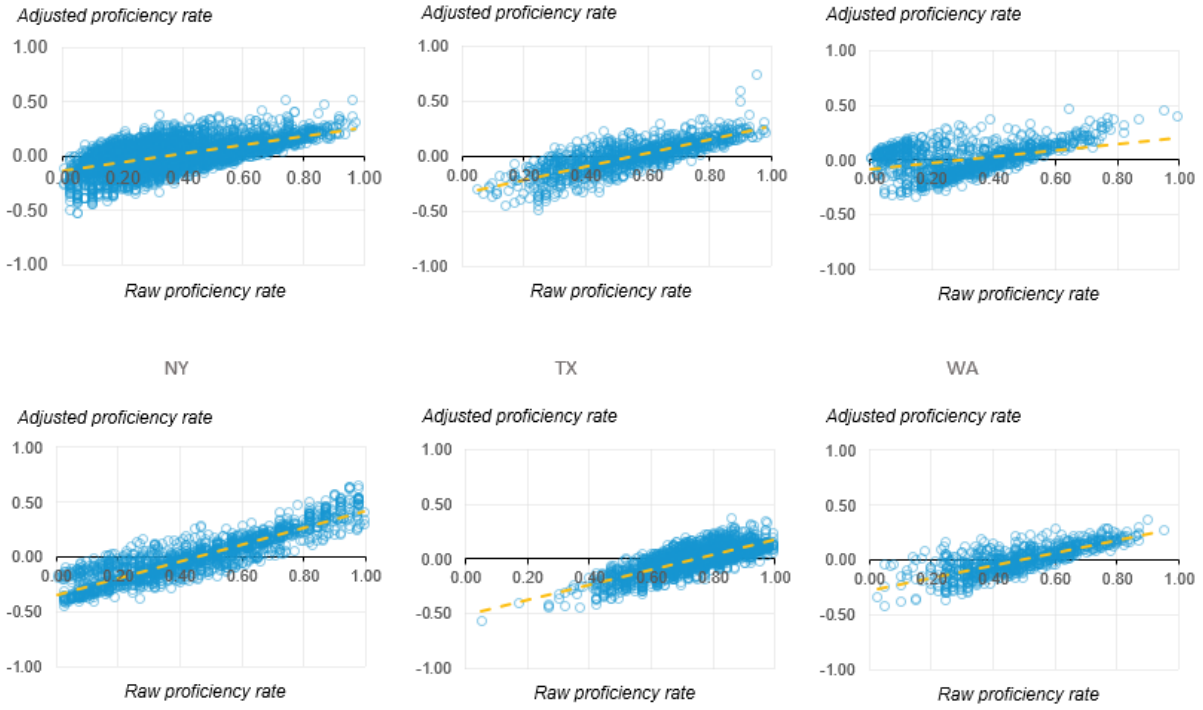
of school proficiency where the covariates used in the adjustment have the greatest effect and where stakeholders need to be more careful when using proficiency rates to measure school quality.

As such, the scatterplots in figure 4 can help us pinpoint the range of proficiency rates for which they may or may not be better approximations of school quality. To be sure, having a tight adjustment range does not imply that proficiency is a strong measure of quality, but a large range does imply that raw proficiency may be a particularly bad school quality proxy.

FIGURE 4

Relationship between Unadjusted and Adjusted Proficiency Rates in Six States

Mathematics proficiency rate in public schools serving eighth grade



URBAN INSTITUTE

Source: Authors' calculations using 2015–16 data from EdFacts, the Common Core of Data, and the Office for Civil Rights.
Notes: NAEP = National Assessment of Educational Progress. Student-level NAEP pseudo-regression estimates correspond to adjustments based on the coefficients of a student-level regression using NAEP samples from North Carolina.

Readers who have a geographic intuition of the school system in question might find it useful to look at a choropleth map of the changes in school quality quintiles generated by our adjustment. Figure 5 is a map of the New York City Department of Education school attendance boundaries (as reported by the National Center for Education Statistics' 2015–16 School Attendance Boundary Survey). Schools are placed in five quintiles of the raw (left panel) distribution of proficiency rates (defined within New York

City) and five quintiles of the NAEP-adjusted proficiency rate distribution (right panel). Dark colors correspond to low-proficiency schools, and light colors correspond to high-proficiency schools.

The raw proficiency rate in New York City paints a picture familiar to most New Yorkers.

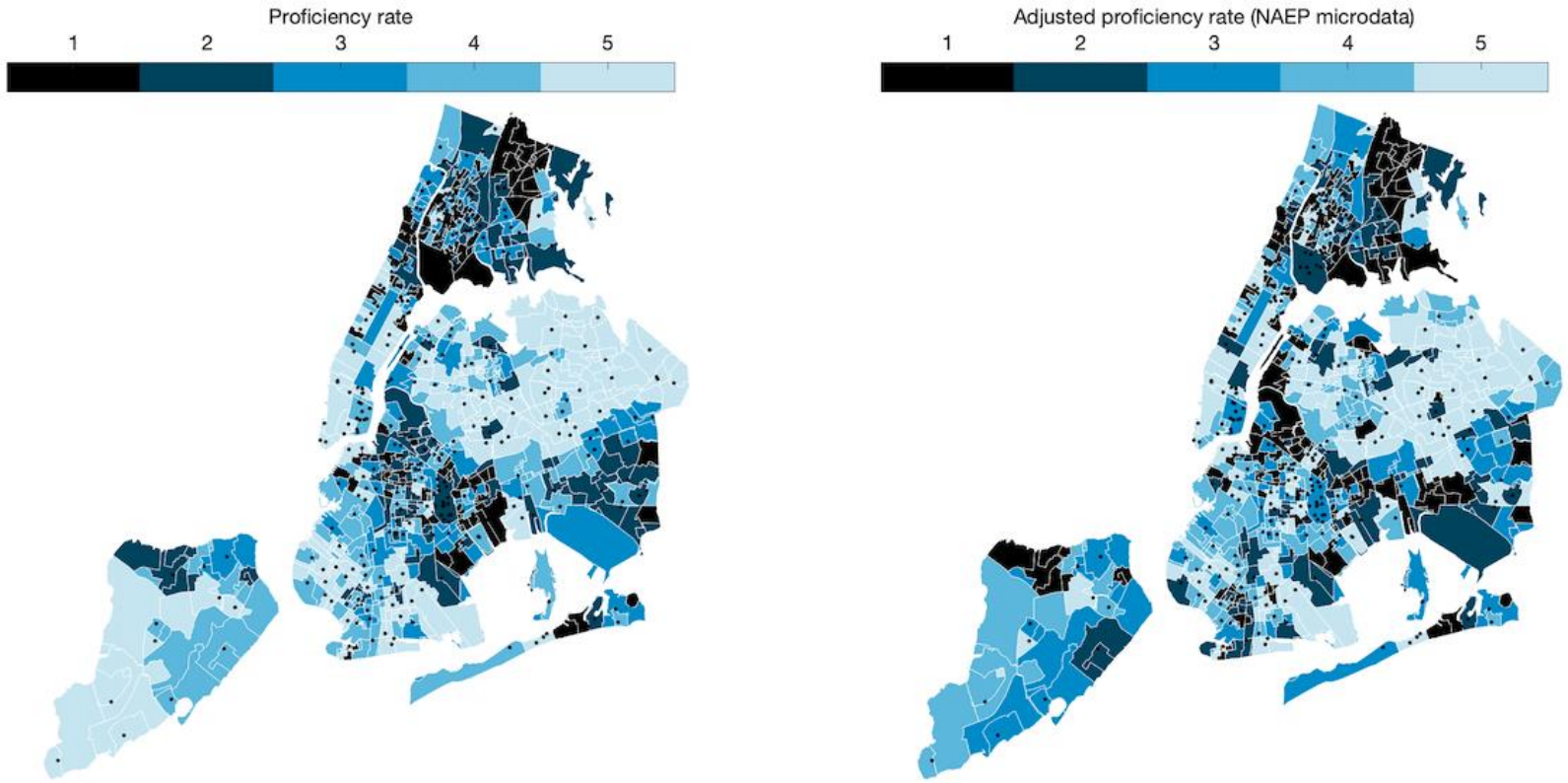
Schools in Manhattan, Queens, and Staten Island tend to be in the top quintiles of the proficiency distribution. In contrast, schools in the Bronx and Brooklyn are overrepresented in the low quintiles of the distribution. What changes when we perform our pseudo-regression adjustment (right panel)? The most notable finding is that the map is not dramatically different from the one on the left. This simple visual test indicates that (1) raw proficiency rates tend to be a decent proxy for school quality in a city like New York or (2) our covariates do little to eliminate selection bias in our adjusted measures of school quality. With the data at hand, it is impossible to distinguish between these two scenarios.

Still, there are some, albeit not drastic, differences between the two maps suggesting that in some schools, our demographic adjustment generates differences in the quintile ranking of schools when comparing the two measures. Schools in Staten Island tend to be adjusted downward by our estimator. In contrast, a few schools in the Bronx are adjusted upward. These patterns mean that some schools in Staten Island have the observable characteristics of schools that tend to perform well, but they do not reach those expectations in reality. The handful of schools in the Bronx have the opposite pattern. Their observable characteristics resemble those of schools that tend to perform poorly, but in practice, those schools actually exceed these low expectations. Thus, one use of our measures is to potentially investigate the institutional details of high-performing schools (relative to low-performing schools) to reveal best practices.

FIGURE 5

Unadjusted versus Adjusted Proficiency Rates in New York City

Mathematics proficiency rates in public schools serving eighth grade



URBAN INSTITUTE

Source: Authors' calculations using 2015–16 data from EdFacts, the Common Core of Data, and the Office for Civil Rights. Map from the 2015–16 School Attendance Boundary Survey.

Notes: NAEP = National Assessment of Educational Progress. Preliminary NAEP adjustments are computed using coefficients from a student-level regression using North Carolina data. Dots indicate schools.

Table 4 provides further evidence that our adjustment substantially changes the relative ranking of schools, even though the overall variability in the distribution of estimated school quality does not change dramatically relative to the raw distribution of proficiency rates. This analysis is useful for exploring the properties of the adjusted measures—in that it is parallel to the comparisons in table 3—but should be interpreted with caution, given that none of these measures are close to being unbiased measures of school quality.

Mirroring the previous results, we place California schools into five categories defined by quintiles of the distribution of raw and adjusted proficiency rates. The first quintile corresponds to the lowest-performing 20 percent of schools, and the fifth one corresponds to the highest-performing 20 percent of schools. Table 4 shows how schools move among these buckets when we execute the pseudo-regression adjustment to the California EDFacts school achievement data, such that the quintiles correspond to the statewide distribution. Nearly 43 percent of schools in the lowest quintile of the raw proficiency rate remain there when adjusted, while 34.8 percent move up to the second quintile and 17.6 percent move up to the third quintile. About 5 percent move from the lowest quintile to the highest and second-highest quintiles. Similar movement between quintiles takes place as we move up to higher quintiles of the raw distribution of proficiency. At the second quintile of raw proficiency, 19 percent of schools move to the highest two quintiles. Eleven percent of schools in the middle quintile and 24 percent of schools in the second-highest quintile move to the highest quintile.

There is also downward movement between quintiles. At the fourth quintile of raw proficiency (i.e., schools that have better than median proficiency but are below the top 20 percent of schools), there is considerable movement to the bottom quintiles of adjusted proficiency. Two percent of schools in this group move to the bottom quintile, while 15.5 percent shift to the second-lowest quintile. No school in the top 20 percent of schools shifts to the bottom quintile, and only one moves to the second quintile. This may suggest that schools with the highest proficiency are indeed the best schools in the state, provided all the above caveats about the accuracy of these quality measures, but this seems to be a pattern unique to California that does not hold for our data on other states (tabulations for other states are available upon request).

The findings in table 4 exemplify that our pseudo-regression adjustment of EDFacts data using regression coefficients from NAEP generates significant changes in school rankings statewide. This confirms that observable school characteristics are indeed highly correlated with proficiency rates. When we define school quality measures to adjust for school characteristics, the measured distribution of quality shifts considerably. But we can say little about the improved accuracy of these measures,

which limits the direct implications of our adjustment for policymakers. All we can say is that the adjusted measures are arguably better, as they at least attempt to account for selection bias.

TABLE 4

Switches between Raw and Adjusted Proficiency Rate Quintiles in California

Mathematics proficiency rates in public schools serving eighth grade

	Quintile of School Value-Add					
	1	**2**	**3**	**4**	**5**	***N***
Quintile of raw proficiency rate						
1	215	176	89	23	3	506
	42.5%	34.8%	17.6%	4.5%	0.6%	
2	192	128	153	90	21	584
	32.9%	21.9%	26.2%	15.4%	3.6%	
3	81	117	80	104	47	429
	18.9%	27.3%	18.6%	24.2%	11.0%	
4	12	78	150	143	121	504
	2.4%	15.5%	29.8%	28.4%	24.0%	
5	0	1	28	140	308	477
	0.0%	0.2%	5.9%	29.4%	64.6%	
N	500	500	500	500	500	2,500

Source: Authors' calculations using 2015–16 data from EdFacts, the Common Core of Data, and the Office for Civil Rights.
Note: Preliminary National Assessment for Educational Progress adjustments are computed using coefficients from a student-level regression using North Carolina data.

The nature of the findings above raise an additional question that we can investigate empirically: Which schools are shifted upward in ranking by our pseudo-regression adjustment? Table 5 provides an answer. For succinctness, we estimate a univariate OLS regression with varying school characteristics as the outcome and an indicator variable of whether a school is adjusted upward as the explanatory variable. This allows us to understand the statistical precision of our estimates. The slope coefficient from this regression is interpreted as the mean difference in school characteristics between schools that are adjusted upward and those that remain in the same quintile or are adjusted downward.

Column 1 of table 5 uses the white share of enrollment at schools as the outcome. Schools that are adjusted upward are less likely to have a high white share of enrollment, about 33 percentage points less than schools that were adjusted downward. Put simply, this means schools with higher shares of white students are more likely to be adjusted down. The opposite holds when we use the Black or Hispanic share as the outcome (columns 2 and 3). The Black share of students at schools adjusted upward is 2.8 percentage points higher than at schools that we adjusted downward, and the Hispanic share of students is 41.9 percentage points higher. The Asian share of enrollment mimics the patterns we observe for the white share: an upward adjustment in school ranking is associated with a lower Asian share of enrollment (column 4). Unsurprisingly, changes in school percentiles predict little or no changes

in the schools' gender shares (column 5). But in line with our findings on racial and ethnic composition, schools adjusted upward are more likely to be schools with high poverty rates (column 6). The share of students at these schools receiving free and reduced-price lunch is 37 percentage points higher than at schools that were adjusted down.

The findings in table 5 elucidate the types of schools that change ranking with our pseudo-regression adjustment, but these differences largely reflect the underlying adjustment. Schools with high white shares of enrollment or low poverty rates tend to have higher proficiency rates, at least partly because the students at these schools tend to be from more privileged backgrounds, which motivated the entire regression adjustment exercise presented in this study. When we implement the regression (or pseudo-regression) adjustment for this correlation, we obtain a set of adjusted proficiency rates that account for this link. It is thus not surprising that schools with more white students and in more affluent areas are more likely to be adjusted down. This finding does not tell us that these schools are worse but merely shows that they tend to have high proficiency rates. Still, the findings in table 5 are useful to explain to lay audiences the nature of our regression adjustment, although caution is imperative in their interpretation.

TABLE 5

Predicting School Characteristics Based on Upward Quintile Adjustments in California

	(1) White	(2) Black	(3) Hispanic	(4) Asian	(5) Female	(6) FRPL
1(Adj. percentile > raw percentile)	-0.330*** (0.007)	0.028*** (0.005)	0.419*** (0.008)	-0.071*** (0.004)	-0.006*** (0.002)	0.370*** (0.007)
Constant	0.402*** (0.006)	0.050*** (0.002)	0.365*** (0.005)	0.111*** (0.004)	0.493*** (0.001)	0.477*** (0.006)
Share adjusted up	0.330					
R^2	0.364	0.017	0.474	0.061	0.005	0.406
Total obs.	2,500	2,500	2,500	2,500	2,500	2,500

Source: Authors' calculations using 2015–16 data from EdFacts, the Common Core of Data, and the Office for Civil Rights.
Notes: FRPL = free and reduced-price lunch. Preliminary National Assessment for Educational Progress adjustments are computed using coefficients from a student-level regression using North Carolina data.
*** $p < 0.01$.

Conclusion

This report has sought to assess whether publicly available data on school test scores can be used effectively to measure school quality in a way that is relevant to determine schools' impacts on student outcomes. This task is important for education policy given that popular school ratings based on these public data are used to make important decisions, not to mention that some states use similar data to

create policy. Bringing together insights from the academic literature as well as our own empirical evaluation of multiple data sources, we have established that the accurate measurement of school quality is difficult.

In many cases, schools are not comparable with each other. Because American public education is pervasively and enduringly segregated on the basis of race, ethnicity, and socioeconomic status, it is difficult to accurately know what would happen if we took a randomly chosen child from her school and moved her to a different school. In the best-case scenario, we can use school lotteries to mimic an experiment that would answer this question, but these scenarios are an exception. The second-best alternative is to leverage states' administrative student records to make school comparisons among students with similar prior academic performance. Still, the evidence suggests that even these student growth metrics are not always reliable estimators of individual schools' impacts. This means that there is not only selection in levels of student achievement but also in changes in student achievement, which is difficult to control for accurately.

A third-best option, which we focus on in this report given its feasibility with publicly available data, is to use school-level data on proficiency rates and adjust them for school demographics such as free and reduced-price lunch rates, racial and ethnic composition, and other commonly measured school characteristics. The idea here is that these characteristics are proxies for student background, such that the approach still provides a useful adjustment that can reduce selection bias in school test scores. But significant bias still likely remains, in part because this approach does not capture students' prior test scores.[4]

Our analysis of NCERDC and EDFacts data points to two broad conclusions about measuring middle school quality. First, on average, aggregate measures of school quality are predictive of our best estimates of school value-add. Adjusting these aggregate measures substantially improves average performance of these predictions. Second, aggregate measures of school quality, regardless of demographic adjustments, are noisy and likely biased predictors of school value-add. This means that although aggregate measures perform well on average, school value-add varies a lot within groups of schools that perform similarly on these measures.

These results imply that one must use caution when using these measures to make decisions that affect real-world decisions, such as families' choices of or policymakers' actions toward individual schools. We find it worrisome that popular school rating websites used these metrics to provide misleading information, especially because the important caveats regarding these measurement issues are often buried in the small print of a technical appendix, if they are mentioned at all. The fact that

these measures are so noisy also implies that policies indexed by measures of school quality based on publicly available data would get it wrong a lot. These limitations are in addition to any inherent limitations in the underlying assessment used, as standardized tests may miss important impacts that schools have on their students.

This means that school quality measures based on aggregate data should not be used to make high-stakes decisions about individual schools, such as evaluating the principal's performance. But they can still be useful for understanding performance levels or trends in groups of schools. Researchers, policymakers, or philanthropists could use the measures to identify groups of schools that are "beating the odds" for further study in a state that publishes only limited information on school performance. More broadly, comparing the adjusted and unadjusted measures can be helpful in approximating the role that nonschool factors play in distorting comparisons of schools.

In sum, our analysis raises questions about how best to use test score data to measure school performance, not just at the middle schools we study but at elementary and high schools where this task is harder (federal law requires testing in grades 3–8 and once in high school). We need more evidence to understand how to appropriately use test score data to measure school performance and hold schools accountable for achieving performance goals.

Appendix

Consider the following *constant-effects model* of the effects of schools on student achievement. Student i's potential outcome at school j, denoted Y_{ij}, is written as the sum of two noninteracting components:

$$Y_{ij} = \mu_j + a_i$$

where μ_j is the mean potential outcome at school j and a_i is student i's "ability," or achievement potential. This additively separable model implies that causal effects are the same for all students. The constant-effects framework focuses on the possibility of selection bias in value-added estimates rather than treatment effect heterogeneity (though we explore heterogeneity as well).

We use the indicator variable D_{ij} to indicate whether student i attended school j in sixth grade. The observed sixth-grade outcome for student i can therefore be written as

$$Y_i = Y_{i0} + \sum_{j=1}^{J}(Y_{ij} - Y_{i0})D_{ij} = \mu_0 + \sum_{j=1}^{J}\beta_j D_{ij} + a_i$$

The parameter $\beta_j \equiv \mu_j - \mu_0$ measures the causal effect of school j relative to an omitted reference school with index value 0. In other words, β_j is school j's value-add.

Conventional value-added models use regression methods to mitigate selection bias. Write

$$a_i = X_i'\gamma + \epsilon_i$$

for the regression of a_i on a vector of controls, X_i, which includes student characteristics and a flexible polynomial in fifth-grade test scores. Note that $E[X_i\epsilon_i] = 0$, by definition of γ. This decomposition implies that observed outcomes can be written as

$$Y_i = \mu_0 + \sum_{j=1}^{J}\beta_j D_{ij} + X_i'\gamma + \epsilon_i$$

Given the assumption that $E[D_{ij}\,\epsilon_i|X_i] = 0$ for all j, when we estimate the above equation using OLS, we obtain unbiased estimates of the causal estimates β_j. If these assumptions fail, OLS will produce biased estimates of the causal effect $\hat{\beta}_j = \beta_j + \pi_j$, where π_j captures the correlation between D_{ij} and ϵ_i that cannot be captured by including X_i in the regression.

TABLE A.1

How Well Value-Added Measures Predict Future Student Performance

By controls and number of cohorts used in value-added estimation

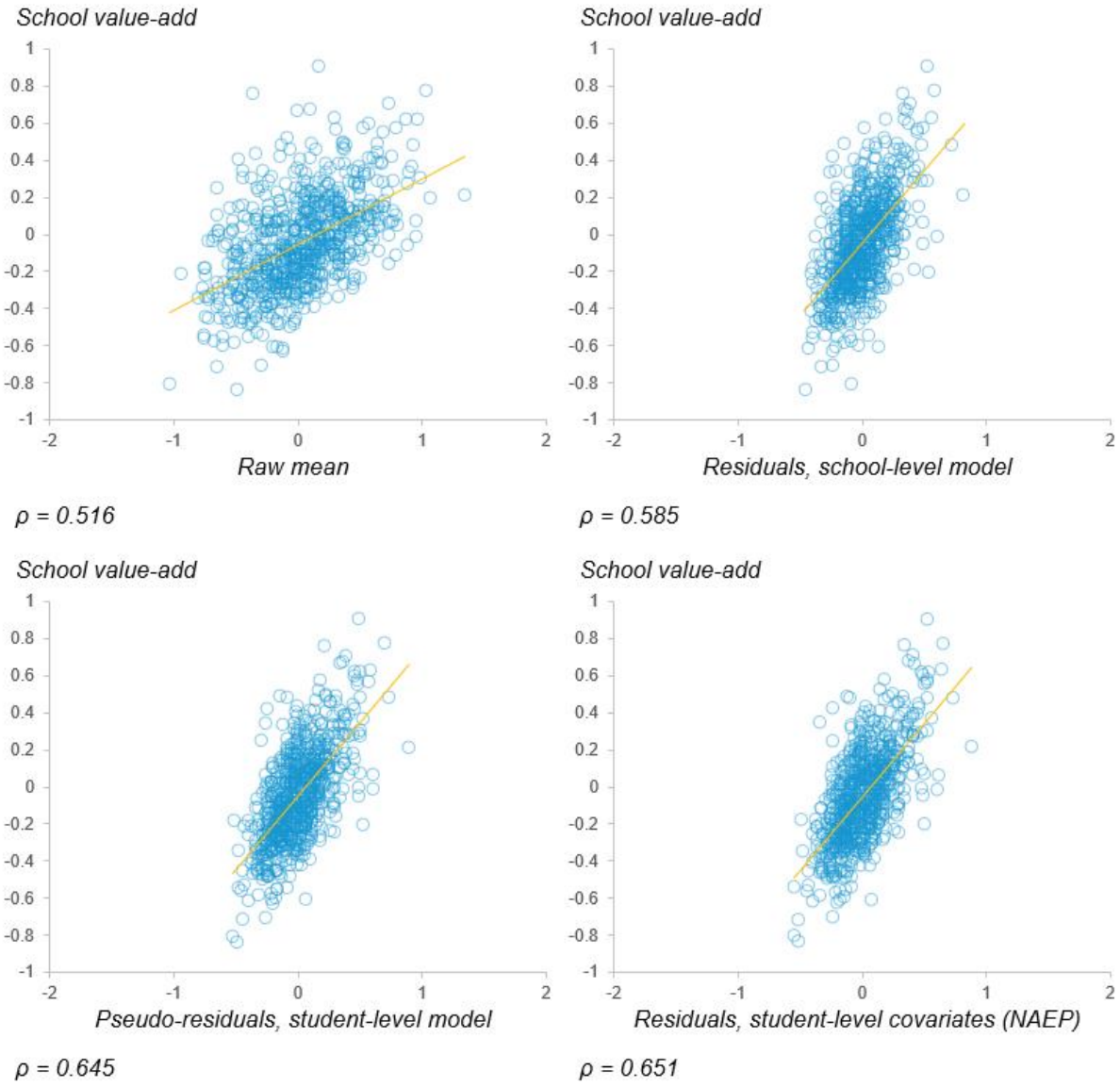
			Confidence Interval			
	Lambda	**SE**	**Lower**	**Upper**	***N***	***R^2***
Raw mean						
2016	0.140	0.015	0.110	0.170	599,772	0.706
2015–16	0.138	0.015	0.108	0.168	599,783	0.706
2014–16	0.134	0.015	0.104	0.164	599,783	0.705
2013–16	0.132	0.015	0.102	0.162	599,783	0.705
2012–16	0.103	0.015	0.100	0.160	599,783	0.705
Demographic adjustment						
2016	0.328	0.022	0.285	0.370	599,772	0.708
2015–16	0.319	0.023	0.275	0.364	599,783	0.707
2014–16	0.309	0.024	0.263	0.356	599,783	0.707
2013–16	0.302	0.025	0.253	0.351	599,783	0.707
2012–16	0.299	0.024	0.251	0.347	599,783	0.707
Demographics + fifth-grade test scores adjustment						
2016	0.849	0.023	0.803	0.895	599,772	0.710
2015–16	0.851	0.025	0.802	0.900	599,783	0.701
2014–16	0.835	0.030	0.776	0.894	599,783	0.709
2013–16	0.836	0.036	0.766	0.906	599,783	0.709
2012–16	0.834	0.034	0.767	0.902	599,783	0.709

Source: Authors' calculations using the North Carolina Education Research Data Center data.

Note: SE = standard error.

FIGURE A.1

Predicting School Value-Add Using School Quality Measures Based on Various Adjustments of Average Reading Achievement

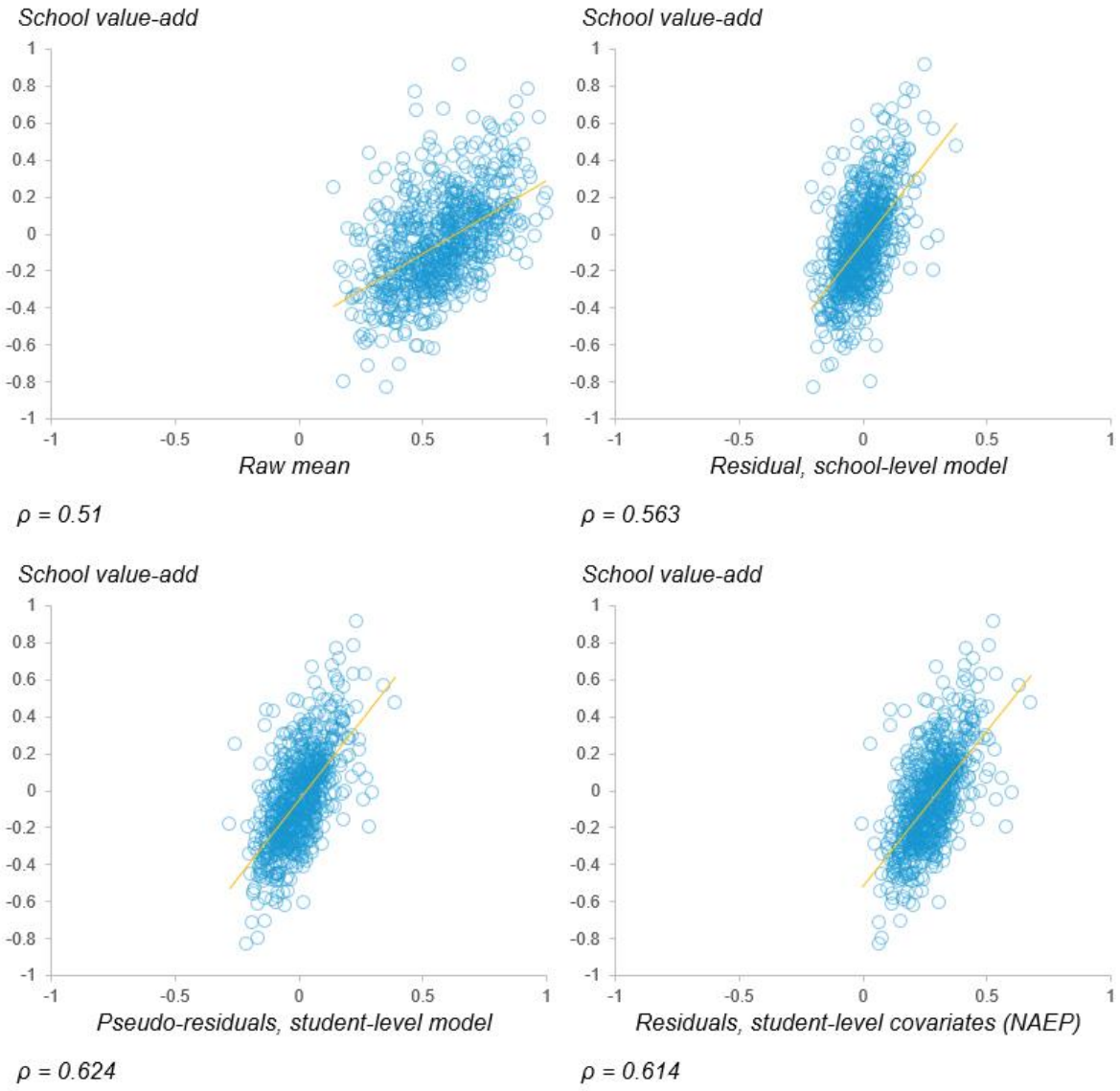


URBAN INSTITUTE

Source: Authors' calculations using data from the North Carolina Education Research Data Center, EDFacts, and NAEP.
Notes: NAEP = National Assessment of Educational Progress. This figure presents scatterplots of "north star" school value-add on the y-axis and various measures of school quality computes with school-level data based on standardized eighth-grade reading scores on the x-axis. School value-added estimates are school fixed effect estimates in a regression of 2016 middle school student achievement, controlling for student demographics and lagged student achievement measure in fifth grade. Aggregate (school-level) covariates denote an ordinary least squares adjustment for school-level covariates that are computed using North Carolina administrative data. Student-level covariates denote an adjustment using coefficient estimates from a student-level regression using North Carolina administrative data.

FIGURE A.2

Predicting School Value-Add Using School Quality Measures Based on Various Adjustments of Reading Proficiency Rates



URBAN INSTITUTE

Source: Authors' calculations using data from the North Carolina Education Research Data Center, EDFacts, and NAEP.
Notes: NAEP = National Assessment of Educational Progress. This figure presents scatterplots of "north star" school value-add on the y-axis and various measures of school quality computed with school-level data based on eighth-grade proficiency rates on the x-axis. School value-added estimates are school fixed effect estimates in a regression of 2016 middle school student achievement, controlling for student demographics and lagged student achievement measure in fifth grade. Aggregate (school-level) covariates denote an ordinary least squares adjustment for school-level covariates that are computed using North Carolina administrative data. Student-level covariates denote an adjustment using coefficient estimates from a student-level regression using North Carolina administrative data.

Notes

[1] More precisely, Angrist and coauthors formulate a new lottery-based test of conventional value-added models. In contrast with earlier studies, which implicitly look at average-across-schools validity in a test with one degree of freedom, Angrist and coauthors' overidentification test looks at each of the orthogonality restrictions generated by a set of lottery instruments. Intuitively, lotteries generate random variation in assignment at every school, allowing the authors to check whether each estimate of school value-add is accurate. Thus, the test asks whether conventional value-added estimates correctly predict the effects of randomized admission at every school that has a lottery, as well as predict an overall average effect. They apply the test to administrative records from Boston Public Schools.

[2] "The EDFacts Initiative," US Department of Education, accessed April 8, 2020, https://www2.ed.gov/about/inits/ed/edfacts/index.html.

[3] We also find that value-added measures controlling for fifth-grade scores are more predictive of eight-grade achievement (out of sample) than are value-added measures controlling for prior-year scores.

[4] The problem is that public datasets do not include the key control variable in conventional value-added models—students' prior test scores—and that important information is lost when we do not have student-level data. This means that these metrics cannot be interpreted as "student growth," which is much of the appeal of the value-added framework. Instead, these metrics are to be interpreted as variation in school test scores that cannot be explained by the characteristics used in the adjustment. In a sense, these measures "remove the influence" of school characteristics, making comparisons more fair or palatable, but it is unlikely that they completely eliminate selection bias.

References

Abdulkadriglu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters. 2020. "Do Parents Value School Effectiveness?" *American Economic Review* 110 (5): 1502–39.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics* 132 (2): 871–919.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2017. *An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys*. Working Paper 23478. Cambridge, MA: National Bureau of Economic Research.

Betebenner, Damian W. 2011. "A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories." Working paper. Dover, NH: National Center for the Improvement of Educational Assessment.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–632.

Hasan, Sharique, and Anuj Kumar. 2019. "Digitization and Divergence: Online School Ratings and Segregation in America." Working paper. Durham, NC: Duke University.

Monarrez, Tomas, Brian Kisida, and Matthew Chingos. 2019. *Charter School Effects on School Segregation*. Washington, DC: Urban Institute.

Rothstein, Jesse M. 2006. "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions." *American Economic Review* 96 (4): 1333–350.

———. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175–214.

———. 2017. "Measuring the Impacts of Teachers: Comment." *American Economic Review* 107 (6): 1656–84.

About the Authors

Tomas Monarrez is a research associate in the Center on Education Data and Policy at the Urban Institute. His research focuses on education policy topics as they relate to economic and racial inequality. Monarrez received bachelor's degrees in economics and mathematics from the University of Texas at Austin and earned his doctoral degree in economics from the University of California, Berkeley.

Matthew Chingos directs the Center on Education Data and Policy. He leads a team of scholars who undertake policy-relevant research on issues from prekindergarten through postsecondary education and create tools such as Urban's Education Data Portal. He received a BA in government and economics and a PhD in government from Harvard University.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

URBAN
INSTITUTE · ELEVATE · THE · DEBATE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org