Data Explained: School Quality Report Comparison Group and Impact Score

Contents

Schools Have Different Performance Levels

Graduation Rate 81%
Average Math Score: 72

Graduation Rate 91%
Average Math Score: 84

Graduation Rate 98%
Average Math Score: 91

Based on the data above, you might think the school with the highest test scores is the best school. But this is not necessarily true! Researchers have found that only 10% - 20% of the differences in test scores are due to school quality while the rest of the difference is due to other factors, such as students' incoming preparation for school.[1] [2] The comparison group methodology in the School Quality Reports is designed to control for these factors, which are generally outside of the control of the school.

Comparison Groups

Below is a screenshot from a School Quality Snapshot.


Graduation

4-Year Graduation Rate
90% of the class of 2019 graduated within four years
Comparison Group*: 83%
Bronx: 73%
City: 78%

[1] https://journals.sagepub.com/doi/10.3102/0013189X18759524

[2] https://bpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/9/2108/files/2019/06/KaneStaiger_brookings2002.pdf

The Comparison Group helps answer an important question: How well is the school helping its students grow and improve? Just looking at how high or low a school's overall results – whether its graduation rate or test scores – are doesn't answer that question because it doesn't consider where the students started or the challenges they face.

To infer school quality, the comparison group estimates: How would students who attended this school have performed on average if they had instead enrolled at a random school in the NYC Public School system? The comparison group captures the average performance of students in NYC with the characteristics of the school's students.

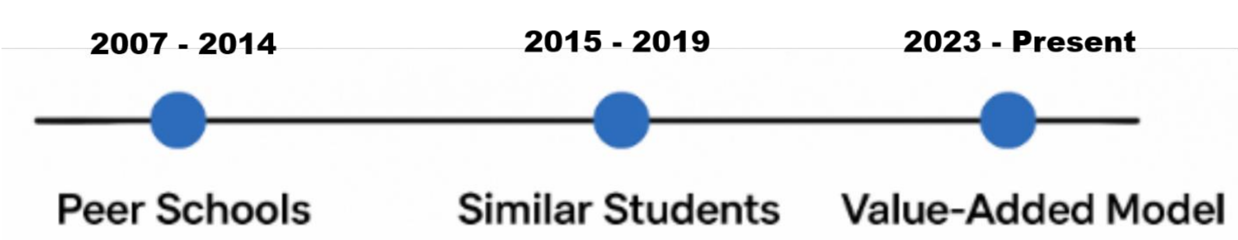
Evaluating our Evaluations

Our estimates of school quality are measured themselves in some important ways including:

- *Validity*: Higher quality schools are more likely to do well on non-test measures such as surveys and expert reviews. Students that attended different schools due to their random admission tiebreak numbers have better outcomes at higher quality schools.
- *Reliability*: Schools that are found to be highly effective one year are likely to be highly effective next year as well.
- *Fairness:* Our quality estimates should not be highly correlated with factors outside the school's control like incoming test scores.

Comparison Group Timeline

Our reports have used different methods at different points in time:



In the Peer Schools method, each school was compared to up to 40 other schools with similar students. We received a lot of feedback in these years that the 40 schools were not actually similar enough to the school being measured. In 2015– 2019 we were able to refine our methods and improve validity, reliability, and fairness by changing to a Similar Students methodology. In those years, each student was compared to a peer group of 50 similar students from around the city.

A few years ago, Dr. Joshua Angrist from Blueprint Labs, a non-partisan research lab based at MIT, asked to speak with NYCPS leadership about the Value-Added Model methods that his team developed. In this method, each school is compared to a "counterfactual," which is a hypothetical school of average quality. This counterfactual serves as a benchmark to gauge "if students at a given school, school ABC, had instead enrolled at a random school in the NYC Public School system, what performance level on average, would these students have achieved?" The answer to this question depends on the school's quality.

Our team examined the validity, reliability, and fairness of this methodology, and found that they were an improvement on our Similar Student models. We were also impressed with the strong reputation that this work has among researchers as evidenced by Dr. Angrist's Nobel Prize in Economics in 2021 for work identifying causal effects in the labor market and education.

How the Value-Added Model Works

The Value-Added Model uses a regression model to estimate the relationship between school enrollment and student outcomes, controlling for student background and prior achievement to isolate the school's impact on those outcomes.

- **What do you mean by “a regression model?”** A regression is a statistical technique that measures the relationship between a dependent variable – in this case, a student’s performance on a given metric, such as State test scores – and any independent variables, such as enrollment in a specific school. Using a regression model helps us estimate the extent to which enrollment in a specific school (independent variable) impacts a student’s performance on a given metric (dependent variable).
- **How do you account for the backgrounds of students at my school?** To isolate a school’s impact on student outcomes, we control for factors that shape students’ starting points and challenges—such as incoming test scores, socio-economic status, special education program recommendations, and English language proficiency. In our regression model, we hold these factors, our control variables, constant in order to isolate the estimated impact of a school on student outcomes. Holding a control variable constant means that results for a student with a certain characteristic, such as having an IEP, are compared to results for students with that same characteristic within the same school type. To better understand this, consider kids on a playground slide that has multiple lanes. If you were trying to see which lane is the fastest to slide down, you might think that the third lane, for example, is the fastest because the kids are going down the fastest in this lane. But, when you look closer, you notice that the kids in the third lane are older and taller than most of the kids in other lanes. To know which lane is really the fastest, then, you need to take the kids’ height and age into account since these factors affect their speed. To do so, you can compare speeds across all of the lanes by comparing only kids of the same age and height across the different lanes. Considering the age and height of kids in each lane ("controlling for age and height") allows you to isolate the speed of the lane from the ages and heights of kids who play at each lane. While the value-added model is more complex due to more variables, this is essentially what we do when we control for student factors to isolate school impact on student outcomes.
- **Why did you switch to this model of accounting for student backgrounds and prior achievement?** The similar students model matched students with 50 similar students based on certain shared characteristics. Using a regression allows us to consider even more students’ performance in our estimates and control for more factors at a time. In fact, we consider all similar students across all schools within the same school type or grade bands as your school to determine an impact score. Looking at so many more students increases the validity, reliability, and fairness of our model. "With this model, your school’s Comparison Group values are essentially a weighted average of the outcomes of the students enrolled in all other

> the schools in the City, where the weight given to each student depends on the characteristics of your student body. Your school impact score compares the average performance of students at your school with this comparison group value."

Running this regression model provides our estimate of the school's impact on student outcomes for each performance metric, holding other student factors constant. We use this estimate to form the Comparison Group values displayed in the School Quality Reports.

For example, take an elementary school with a mean ELA State test score of 3.1 for all students. Their school quality estimate from the value-added model was 0.3, which can be interpreted by saying that, all else equal, enrolling in this school increases a student's ELA State test score by 0.3. Their Comparison Group value would be 3.1 – 0.3 = 2.8. We estimate that had the students at this school randomly attended any other school in the city, their average ELA State test score would be 2.8. Attending this school, though, these students had a mean ELA State test score of 3.1. Thus, the school has outperformed expectations and can be said to have a positive impact on student test scores, holding all else constant.

Steps for the regression:
There are three steps to calculating our current iteration of the Comparison Group (Value-Added Model):

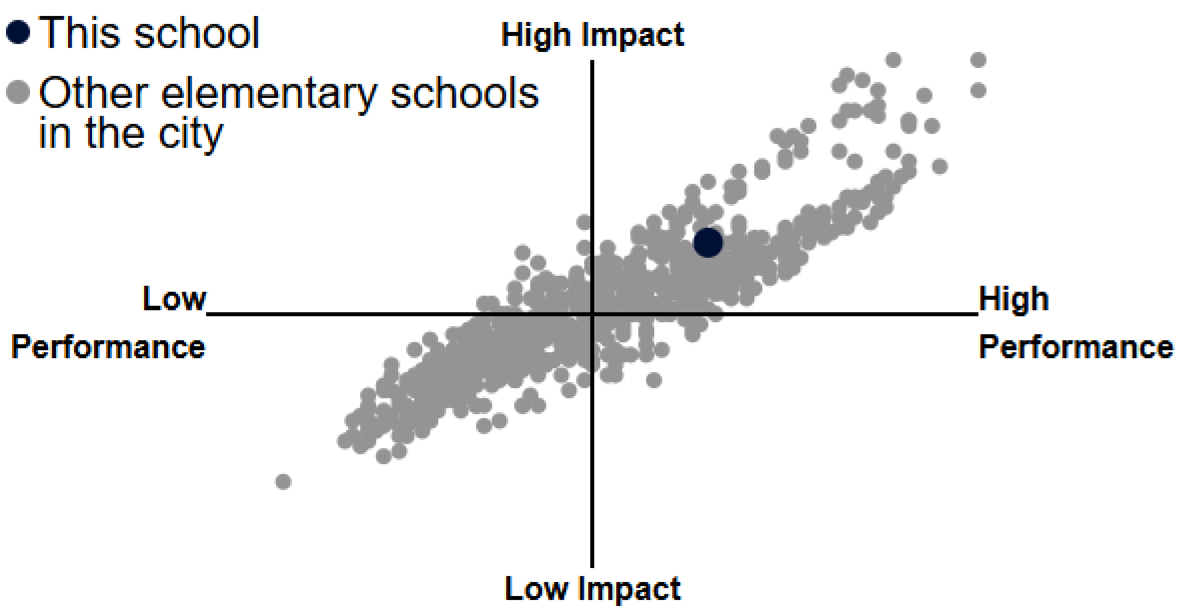
- **Step 1:** We use a regression model to estimate the relationship between school enrollment and student outcomes, controlling for student background and prior achievement to isolate the school's impact.

- **Step 2:** We adjust these impact estimates to reduce statistical noise and ensure that differences reflect true school quality rather than random variation. This adjustment is called "shrinking" as it reduces the variance of the estimates and improves their mean squared error, or the average squared distance between a school's estimated impact score and its true impact score. This is important for small schools where an impact estimate might be highly affected by the addition or removal of just one student. Shrinking helps make these estimates less extreme and more accurate. For larger schools, which have more observations due to higher enrollment, shrinkage has a small effect on impact estimates.

- **Step 3:** We use the adjusted estimates to calculate how each school's students would have performed at a hypothetical NYCPS school of average quality, enabling a fair comparison.

Impact Score

On the School Quality Dashboard we present the value-added model results on a standardized scale. These rescaled value-added scores are called the Impact Score.



The horizontal axis of this chart is based on performance for metrics such as the average test scores or graduation rate. This "performance score" only takes the school's raw performance relative to the citywide average, by school type, into account. The vertical axis is the "impact" which uses our value-added model. Both scores are standardized to be between 0.00 and 1.00 so the mean is 0.50.

The impact and performance scores on the SQR: Dashboard use limited metrics to be more predictive of student success.

Step 1: For each of the relevant metrics, listed below by school type, we calculate the difference between the school's value and their Comparison Group value. (This is, in fact, the regression model estimate.)

Step 2: We standardize these differences, translating the scores to a 0.00-1.100 scale. We use the range of differences for a particular metric and school type to rescale these differences, excluding outliers. If a standardized difference is outside of the 0.00-1.00 scale because it is an outlier, the value is capped at 0 (if it is negative) or 1 (if it is positive).

Step 3: We take a weighted average of the standardized scores for the relevant metrics to produce the impact score displayed in this chart.

Here's what those steps look like for an example elementary school with the following characteristics:

- ELA state test mean score of 3.32. The Comparison Group value for this school is a mean score of 2.86.
- Math state test mean score of 3.35. The Comparison Group value for this school is a mean score of 2.99.

Step 1: The difference between the school's value and their Comparison Group value is 0.46 (3.32 - 2.86) for ELA and 0.36 (3.35- 2.99) for Math.
Step 2: We find the highest and lowest differences for mean scores in ELA and Math among elementary schools. Let's take the following school as an example:.

Metric	Minimum Difference between Value and Comparison Group for Elementary Schools	Maximum Difference between Value and Comparison Group for Elementary Schools
ELA mean score	-0.63	1.27
Math mean score	-0.70	1.28

We use these minimum and maximum differences to standardize every elementary school's scores. That gives our example school a score of 0.81 in ELA and 0.33 in Math. This makes sense, since they outperformed their Comparison Group (had a high impact) in ELA and underperformed the Comparison Group (had a lower impact) in Math.

Formula to standardize range to 0-1: $\frac{(value - \min\ )}{(max - \min\ )}$

ELA standardization: (0.46- -0.63)/(1.27- -0.63) = 0.57

Math standardization: (0.36 - -0.70)/(1.28- -0.70) = 0.54

Step 3: The average of the scores for the two metrics gives us the school's impact score reported on the SQR.

(0.57 + 0.54)/2 = 0.56

Our interpretation is that this school has a higher than average positive impact on its students in their ELA and Math state test scores as their impact score is above the mean of 0.5.

*Standardized scores do not inherently have a mean of 0.5, so we mechanically adjust the average on the SQR: Dashboard. This involves adding a constant to each score to bring the mean to 0.5, by school type. On the SQR: Dashboard chart, scores are capped at 1. However, some impact and performance scores might exceed 1 with the addition of this constant and these uncapped scores are available in the downloaded data file.

The metrics included in each school's impact score are determined by school type. The metrics are for all students; subgroup performance is not separated in the impact scores.

School Type	Metrics Included	Metric Weight
Elementary, Middle, and K-8 schools	Mean ELA State test score	50%
	Mean Math State test score	50%
High schools	Mean ELA Regents score	25%
	Mean Algebra I Regents score	25%
	4-year graduation rate	25%
	Postsecondary enrollment 6 months after graduation deadline	25%
Transfer schools	Mean ELA Regents score	25%
	Mean Algebra I Regents score	25%
	6-year graduation rate	50%

Comparison Group and Impact Score Frequently Asked Questions

Are value-added scores useful in improving student outcomes?

The Blueprint team recently released a paper,[3] using data from the School Quality Reports, that shows that sending your child to a school with high test-score value added and high survey results will benefit their chances of graduating from high school and college. They found picking a high school with high survey results was best at boosting the chance of high school graduation, while picking a high school with high test-score value added was best at boosting the chance of college success.

Why can't I see an individual student's estimated performance?

Using a regression model allows us to capture general patterns and average relationships between student factors (race, IEP status, socioeconomic status) and student outcomes (State test scores, graduation, credit accumulation) across the entire City. This model is, by nature, meant to provide predictions about the citywide data and not for the performance of individual students.

A prediction for an individual student's expected outcome is likely to be statistically imprecise because it's so specific. Individual estimates for students could not account for all of the circumstances of that exact student; by using citywide averages to calculate estimates and aggregating these estimates to the school level, we compensate for some of this variation and reduce statistical noise. If we were to model specific estimated outcomes for students, the standard error may be very high – for some individual students, we couldn't say with statistical certainty whether their expected chance of graduating should be 60% or 80%, for example, which makes estimates at this level inactionable.

Even averages can have statistical noise, though, which we address in our model through an adjustment called "shrinking" which we use to ensure that differences reflect true school quality rather than random variation.

[3] https://blueprintlabs.mit.edu/research/putting-school-surveys-to-the-test/

How does the value-added model control for factors outside of a school's control?

Regression estimates the relationship between a student's outcome (dependent variable, Y) and school enrollment (independent variable, D), controlling for student characteristics outside of a school's control (independent control variables, X).

$$Y_{ij} = \sum_{j=1}^{J} \beta_j D_{ij} + X_i' \gamma + \epsilon_{ij}$$

D_{ij} is an indicator variable that is equal to 1 if student *i* is enrolled in school *j* in a specific year

X_i' is a vector of controls for student *i*:

- Demographics: race, gender, English language learner status, special education program placement, and economic need (HRA assistance eligibility or temporary housing).
- Baseline scores: 3rd, 4th, and 5th grade scores for MS; 6th, 7th, and 8th grade scores for HS. Missing indicators replaced by city mean if needed.
- Grade fixed effects, baseline year fixed effects interacted with baselines and demographics.

β_j is the causal effect of attending school *j* on student outcome *Y*, our measure of quality. We recenter the coefficients so that they correspond to the school effects for the "average" NYC student.

Regression controls for factors outside of a school's control by adjusting a student's outcome at a school for that student's expected performance had they attended a different school, based on the average performance of similar students citywide. The set of similar students citywide is determined by the control variables.

For example, suppose the control variable, X, was the Economic Need Index (ENI) and the dependent variable, Y, was the score on a state standardized math test. Regression can be interpreted in three steps. First, regression calculates the citywide average math score among students identified as having economic need (ENI=1) and not (ENI=0). The coefficient on the ENI variable, X, measures the difference in average math scores among these two groups of students. This generates a predicted outcome for each student based on the average outcome of students in their ENI group. Second, each student's adjusted performance is calculated as the difference between their actual outcome and their predicted outcome. A school's Comparison Group is the average predicted outcome of its students. Third, the regression calculates a school's value-added as the average of its students' adjusted performance. In other words, this value-added is the difference between the average performance of its students and the school's Comparison Group performance. The coefficient on the school enrollment variables, D, measures the school's overall value-added for a metric. In schools with high average adjusted outcomes, students perform better than similar students at other schools. In this way, regression controls for the ENI of a school's student body.

The regression controls for many student characteristics in addition to ENI in the control independent variables, X. By including a detailed set of control variables that are associated with differences in outcomes and enrollment patterns, the regression compares students to an appropriate set of similar students citywide. When regression determines a given student's set of similar students, it also balances the number of students (which makes more statistically precise predictions) with how similar those students are (which makes the set of students more representative). For example, consider the set of similar students for a student scoring 2.38 on a prior state standardized math test. Rather than comparing this student only to those scoring exactly 2.38, the regression adjusts for a cubic polynomial relationship between prior scores and outcomes, so that students scoring 2.39 also contribute to the Comparison Group.

Why do school factors such as admission method, budget, facilities, and mergers not factor into the value-added model?

Our impact scores are primarily intended to measure school quality from the perspective of a family choosing a school. For example, if a parent chooses a school based on our impact estimate, then we think that school should create the largest probability of success in career and college for each student.

School quality is a complicated issue and, while the pedagogical effectiveness of the staff plays an important role, it is not the only factor. The quality of school facilities, transportation, neighborhood factors, admission methods, budget, and building changes like mergers and co-locations can all impact school quality in ways beyond the control of the school staff.

However, unlike incoming student characteristics like prior test scores or demographics, these factors are not given for a family choosing a school. So, even though those factors are not within the school staff control, they might impact the chance of student success in career and college and are therefore not things we would try to control for directly in our methods.

Some of these school factors are controlled for indirectly as student factors, to the extent they show up in student data. For example, a school with screened admissions will have students with higher incoming test scores, which is accounted for in the model that calculates impact scores.

How can we be confident that the value-added model truly represents school quality given all the possible factors outside the control of school staff that are not included in the control variables?

While no measure can be perfect, we use a variety of checks and measurements to maximize the validity, reliability, and fairness of our school quality estimates. Our reliability measures ensure that our estimates are relatively stable from year year-to-year. We ensure fairness by minimizing correlations between our quality measures and factors outside of schools' control like incoming test scores. We compare our results to school survey and expert review results to measure validity. Another way we ensure validity of our school quality estimates is to take advantage of the fact that some students are assigned to schools based on lottery numbers – our admissions tiebreaker values.

This quasi-experimental forecast validity confirmation is presented in detail in a recent paper from Blueprint Labs[4]. This method identifies groups of students who applied to similar schools but were assigned differently due to randomness in admissions (e.g., lottery numbers or falling just above or below the admissions cutoffs for screened programs). Since these students are similar and quasi-randomly assigned, we can compare their outcomes based on the impact score of the school they attended to assess how well impact scores predict of student outcomes.

If the model is accurate in forecasting outcomes, we expect, for instance, that students randomly assigned to a high school with a 1 percentage point higher impact score for college enrollment would be approximately 1 percentage point more likely to enroll in college. We conducted this validation across multiple outcomes and found strong predictive performance. Typically, a one-unit increase in a school's impact score corresponded to an average increase of 0.8 to 1 unit in student outcomes. The close correspondence between a school's impact score and the performance of students who attend it by virtue of a lottery demonstrates the predictive accuracy of impact scores. In contrast, these lottery-based tests of predictive accuracy demonstrate that measures of school performance calculated using unadjusted average student outcomes weakly predict school effects. For example, a one-unit increase in a high school's unadjusted average test score corresponds to a roughly 0.2 unit increase in student outcomes for randomly assigned students.

[4] https://direct.mit.edu/rest/article/106/1/1/108832/Credible-School-Value-Added-with-Undersubscribed

How are 6-12 and K-12 schools given impact scores?

The measures used in the impact score for high school are different than the measures used for middle and K-8 schools. Because of this, schools that serve grades 6-12 or K-12 receive two separate School Quality Reports and two separate impact scores. One impact score covers the results of students in grades 9-12 and the other covers the results of students in the other grade levels. The 9-12 results are compared to all other schools that serve those grades. The K-8 or 6-8 results are compared to schools that serve those grade levels.

How do students taking Regents exams in grades 7 and 8 affect impact scores?

Students who take the Regents in lieu of the regular Math exam are still included in the Impact Score calculation. For each student who has a Regents score, we impute what their most likely regular exam score would have been using the table in Appendix A of the Educator Guide to the Elementary, Middle, and K-8 School Quality Reports.[5] If a student takes both the regular exam and the Regents, *only* the Regents score is used so as to avoid an unintended incentive for unnecessary double testing. The Regents scores also contribute to the School Quality Report metric “Percent of 8th Grade Students Who Earned High School Credit.” That metric is a part of the SQR Instruction and Performance Rating but not a part of the impact score.

How are students with disabilities categorized in our model?

Students with different disabilities can have very different patterns of exam performance and progress. Our research has shown that program placement is the closest data we have to use as a proxy for severity of disability. For example, we have found that students with IEPs for related services (RS) only have state test progress as high or higher as students without IEPs while students with special class (SC) placements have much lower progress than students with integrated co-teaching (ICT) or special education teacher support services (SETSS) placements. In our value-added model, RS, ICT, SETSS, and SC students are in separate groups.

[5] https://infohub.nyced.org/docs/default-source/default-document-library/2023-24-educator-guide-ems.pdf

Which schools have impact scores?

Impact scores are provided for most elementary schools, middle schools, high schools, and transfer high schools. Schools may be missing impact scores if most of their students do not take the standard state tests (e.g. District 75, K-2 schools, Pre-K, HSE programs).

Which students should I focus on in order to improve my impact score?

Impact scores are, by design, representative of your school's unique student population. Raising student achievement for all students will likely result in a higher impact score, especially because the metrics used to calculate a school's score are based on all students' performance. Additionally, impact scores are calculated using last year's data. They cannot tell you what a certain student should achieve in the coming school year, as they use the average performance of students across the city to provide the estimated scores we use in your impact score. The regression model's validity comes from using these averages; at the individual student level, estimates are too imprecise to be reliable.

That said, you can use the School Quality Report (SQR) data that we share at multiple points in the fall to a) validate that the underlying data used in these calculations is correct and b) understand which groups of students are performing above or below their Comparison Group values. How much your students are performing above or below their Comparison Group is ultimately what determines your impact score.

Here are a couple of ways in which an elementary school's principal could use the SQR data to focus their efforts on student achievement:

Use the **SQR: Snapshot, SQR: Dashboard,** or **SQR: Student Workbooks** to see the demographic breakdown of your school. This school has a high percentage of students who are Hispanic/Latinx. This group of students will contribute heavily to the school's impact score.

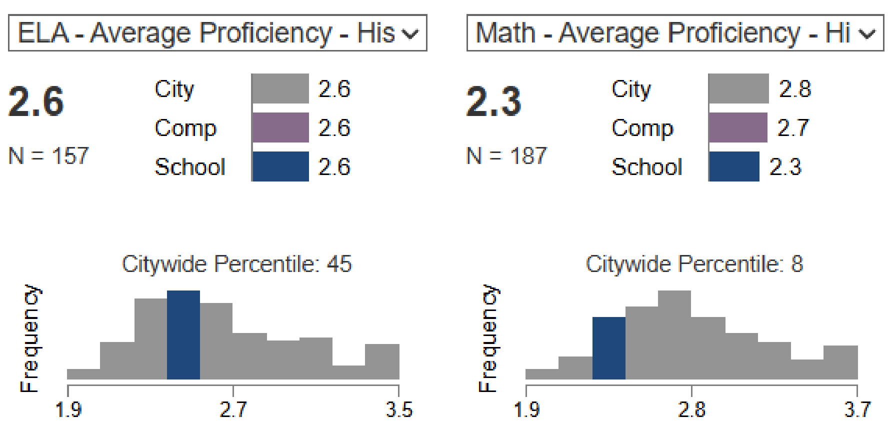
School Description | Instruction and Performance

Student Body

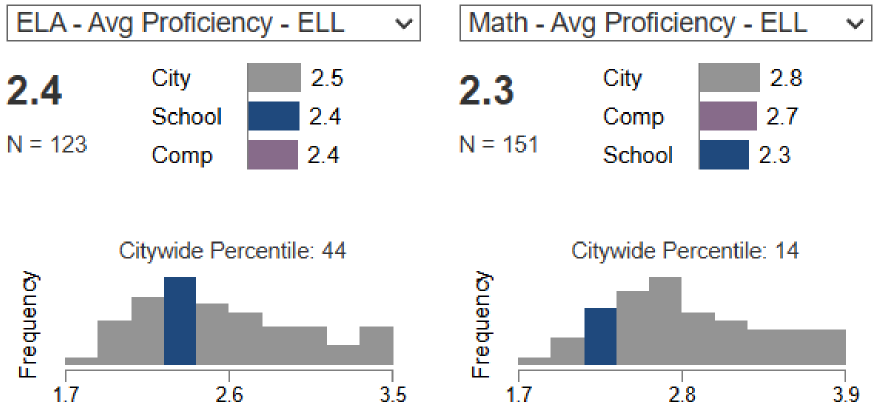
Student Demographics

Asian: 22%
Black: 13%
Hispanic or Latinx: 59%
Native American: 3%
Native Hawaiian/Pacific Islander: 1%
White: <1%
English language learners: 48%
Students with IEPs: 18%
Economic Need Index: 89%
Female: 49%
Male: 51%
Neither female nor male: 0%

Under the Student Achievement Metrics section on the **SQR: Dashboard**, use the dropdowns to select "ELA - Average Proficiency – Hispanic/Latinx" and "Math - Average Proficiency – Hispanic/Latinx." You can see that in ELA, this group of students performed, on average, at the same level as their Comparison Group. In Math, though, they performed, on average, under the Comparison Group estimate by 0.4 points. Since there is room for growth relative to the Comparison Group on Math for the Hispanic students at this school, focus on improving these students' scores would also likely result in a higher school impact score.



Since the school also has a high percentage of students who are English language learners, they can use the **SQR: Dashboard** to see how this group of students did relative to their Comparison Group. Select "ELA - Avg Proficiency – ELL" and "Math - Avg Proficiency – ELL." The students who are English language learners at this school performed, on average, similarly to their estimate in ELA, but under this estimate in Math. This school could support ELL students in Math and, if students begin performing similarly to like students on State tests on average, the school's impact score would likely improve.



For student-level data, principals can consult the **SQR: Student Workbooks** at multiple points when they are provided in the fall. The "Student Characteristics" tab of the **SQR: Student Workbooks** includes demographic and baseline achievement data for each student in your school which are used in the impact score calculation to ensure we control for factors outside of the school's control.