

## Interim Report from Research Committee on STAR rating

*This is an interim report of the Research Committee, primarily directed to the new members of the State Board joining in January 2021.*

*This report attempts to assemble the most central data, testimony, concerns, options, and pro/con arguments considered by the Research Committee. **But as an interim report, the concerns and arguments made below must therefore be considered “interim”.** Some of the arguments made in this report may be found wanting as all State Board members learn more and as new State Board members and the public weigh in. Moreover, the discussions summarized here are grist for ultimate judgments; as we move towards recommendations for revisions to the STAR Framework, different members, as well as the public, will weigh different values more or less heavily in considering the inherent trade-offs of different options.*

*We expect to continue this discussion on the STAR Framework going into the new year with Board members, experts, and members of the general public. We look forward to then developing draft recommendations for discussion.*

### Executive Summary

The Research Committee met on a roughly monthly basis, from early spring 2020 through December 2020, to discuss the STAR Rating and whether and how it should be changed. A key theme of the Committee discussions was a desire to ensure that our statewide accountability system required by the federal Every Student Succeeds Act (ESSA) evolves to reflect as accurately and fully as possible the quality of schools; leverages the system to improve educational outcomes for students; and accomplishes these goals in ways that don’t unintentionally create harmful consequences for schools and their students. Our accountability system must allow us to efficiently and reliably distinguish which schools the state and other schools can learn from and which schools need support or intervention. The purpose of this report is to summarize for the incoming Board members, as well as for the public, our findings and work to date.

In September, the Committee recommended, and the full Board voted on SR20-11, “On Improving the School Transparency and Reporting (STAR) Framework”, which is attached.<sup>1</sup> The resolution had four major points:

1. The rating, as currently formulated, produces “*ratings that are systematically lower for schools that enroll larger proportions of students designated at-risk — and that these lower ratings do not necessarily reflect lower quality education.*”
2. The way in which D.C. currently rates schools must be altered in some way “*so that the STAR rating minimizes undeserved bias against schools with large numbers of at-risk student groups.*” The resolution doesn’t propose a specific alternative, leaving that to subsequent exploration by the Committee.

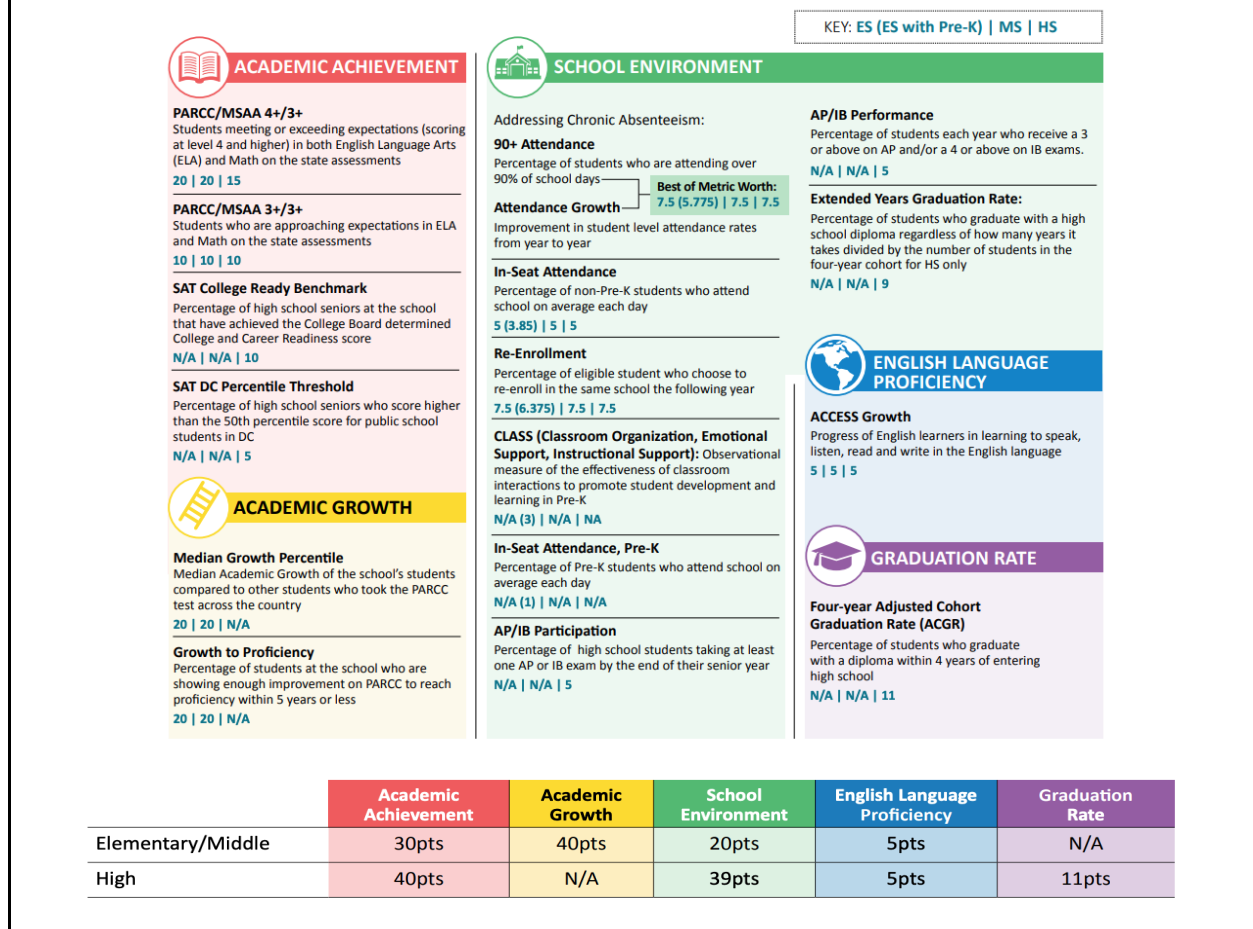
---

<sup>1</sup> [SR20-11 "On Improving the School Transparency and Rating \(STAR\) Framework"](#)

3. There are additional issues around the rating that raise concerns for some Board members and need further exploration. Specifically, the resolution notes that for many schools a low rating does not “necessarily lead to identification of a school’s areas of need or target assistance to it”; that the rating formula makes “it more difficult to identify, celebrate and learn from the schools that are genuinely improving student learning” and may cause some schools to be wrongly “perceived as low-quality”; and, that we may want to provide a broader view of school quality, possibly by expanding the range of “indicators.” All members do not share all of these concerns; the various views are explained below (Section II and III).
4. At the close of this exploration, the Board should recommend changes to the STAR Rating, and possibly how it is used, in an effort to address the concerns named in SR20-11 and others raised by the public.

### What is inside the STAR rating?

Below is a chart from the Office of the State Superintendent (OSSE) providing an overview of what the STAR Framework measures.<sup>2</sup>



<sup>2</sup> OSSE (2019) "School Transparency and Reporting (STAR) Framework Overview"

***What does the designation “at-risk” mean?<sup>3</sup>***

Students who are designated at risk are those who qualify for Temporary Assistance for Needy Families (TANF) or Supplemental Nutrition Assistance Program (SNAP), have been identified as homeless during the academic year, who are under the care of the Child and Family Services Agency (CFSA or “foster care”), and who are high school students at least one year older than the expected age for their grade.

***What encompasses socioeconomic status (SES)?***

The American Psychological Association defines socioeconomic status (SES) as “encompassing not just income but also educational attainment, financial security, and subjective perceptions of social status and social class. Socioeconomic status can encompass quality of life attributes as well as the opportunities and privileges afforded to people within society.”<sup>4</sup>

---

<sup>3</sup> [OSSE Data and Reports](#)

<sup>4</sup> American Psychological Association, [Education and Socioeconomic Status Factsheet](#)



## Table of Contents

<b>I.</b>	<b>Issues to explore around bias in the current rating design .....</b>	<b>5</b>
	A. The evidence	
	B. What STAR elements are most correlated to student income?	
	1. PARCC proficiency	
	2. Re-enrollment	
	3. All-students sub-group	
	4. Additional issues: attendance, HS graduation and other high school measures	
	C. What options exist for reducing the bias of these indicators?	
	1. Greater reliance on growth, less on proficiency	
	2. Alter weights of student subgroups	
	3. Modify the definition of reenrollment.	
	4. Use the “best of” a school’s proficiency or growth rate	
<b>II.</b>	<b>Other concerns with the STAR rating .....</b>	<b>12</b>
	A. Leading vs. lagging indicators	
	B. Perverse (and constructive) incentives	
	C. Single summative rating (vs. dashboard)	
	D. Connecting low-rated schools to support for school improvement	
	E. Effects of low ratings on schools	
	F. Capturing the full range of quality	
<b>III.</b>	<b>Options for addressing these concerns &amp; inherent tradeoffs .....</b>	<b>16</b>
	A. How to handle proficiency & growth	
	B. How to display accountability data (summative rating, data dashboard, quadrants)	
	C. How to handle school climate data	
	D. Other options	
<b>IV.</b>	<b>Attachments/Appendix .....</b>	<b>17</b>
	A. May 22, 2020 - State Board Staff Memo: Commissioner Blomstedt Conversation on AQuESSTT	
	B. June 16, 2020 - State Board Staff Memo: Non-Summative ESSA State Accountability Plans	
	C. July 6, 2020 - State Board Staff Memo: School Climate Research	
	D. August 5, 2020 State Board Staff Memo: Adapting School Climate for Distance Learning...	
	E. July 9, 2020 - State Board Staff Report: ESSA Accountability System & School Report Card FAQs	
	F. September 3, 2020 - State Board Staff Memo: Leading/Lagging Indicators Memo	
	G. September 9, 2020 - State Board Staff Memo: May-June 2020 Public Meeting School Quality Panel	
	H. September 16, 2020 - SR20-11 “On Improving the School Transparency and Reporting (STAR) Framework”	



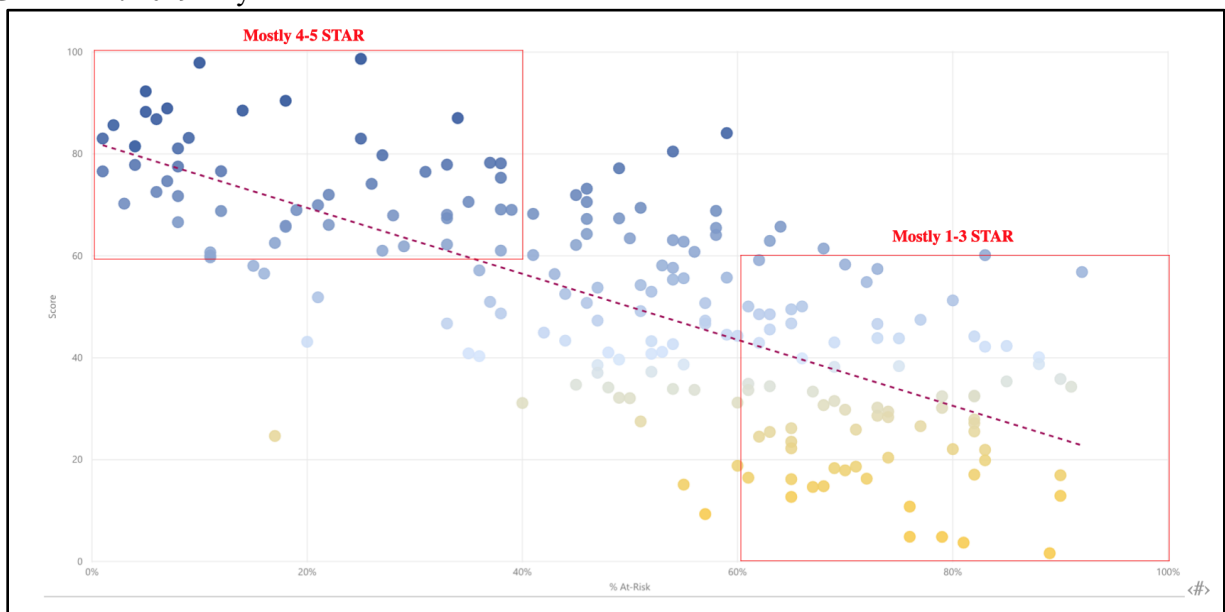
## I. Issues to explore around bias in the current STAR Rating design

### A. The Evidence

The Research Committee invited a variety of education experts to share their reflections on the STAR Rating. Many witnesses noted different ways in which the rating formula might disadvantage schools with large percentages of students designated at-risk, mobile, etc. Four graphs, presented to us by expert witnesses Aaron Cuny, Director of UpliftEd Services, and Dr. Betsy Wolf, Research Scientist at the Institute for Education Sciences, show how and why the current rating system conflates quality with students' household income and other student characteristics.<sup>5</sup>

There is virtually complete overlap between 4/5-star schools and low proportions of students designated at-risk.

**GRAPH 1: 2019 Citywide STAR Performance**



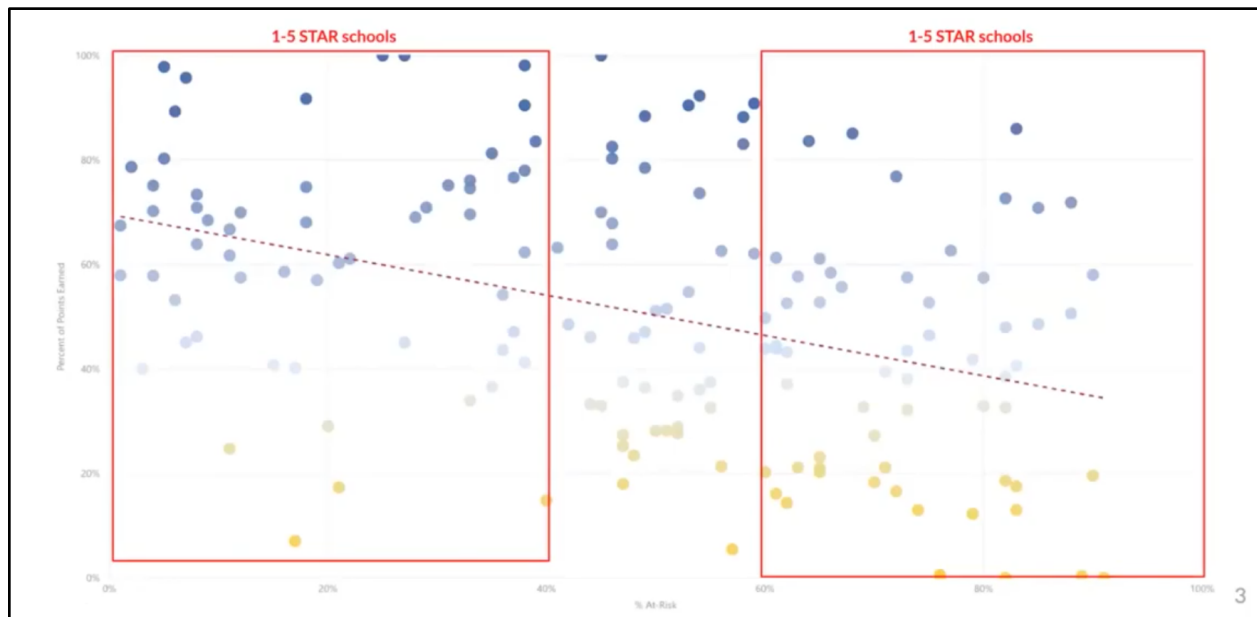
Graph 1, above, shows that virtually all of the schools with 4/5 stars have very low proportions of students designated at-risk and virtually all of the 1/2 schools have high proportions of students designated at-risk. As Cuny pointed out, these data tell us one of two stories: *Either* that virtually all of our schools that enroll more affluent students are high quality and virtually all that enroll the most students designated at-risk are low-quality *or* our rating system is not accurately measuring the quality of our schools. So, which is it? That the rating design penalizes schools with high percentages of students designated at-risk? *Or* that these schools are disproportionately of low-quality?

<sup>5</sup> [SBOE Research Committee Meeting March 30, 2020](#) Recording

To answer that question, one first must ask, “what is the definition of a high-quality school?” This is a more fraught question than it might seem at first glance. Cuny and Wolf both argued that a key definition (though not the only definition<sup>6</sup>) of a “high-quality” school is the extent to which, during the previous school year, the school’s students have learned the material set forth in the District’s academic standards. This element of quality can be measured (at least in reading and math) via “growth metrics” on the District’s annual standardized tests.<sup>7</sup>

But student growth, including students designated at-risk, is more evenly spread across schools.

**GRAPH 2: 2019 All-Student STAR Performance on Growth**



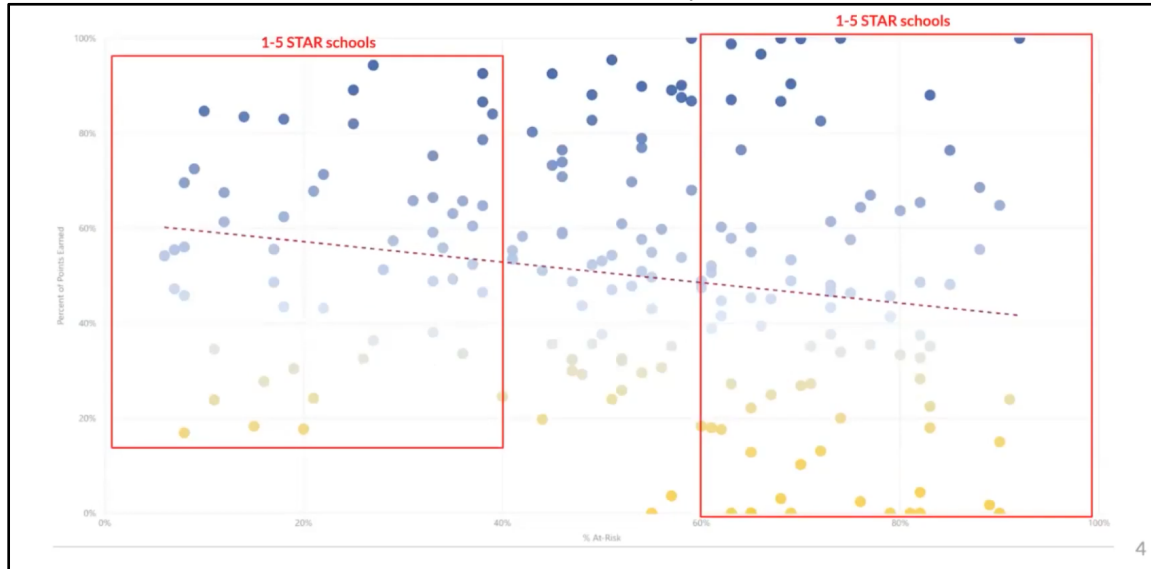
Graph 2, above, uses test score growth data to show how much students grew academically (and, to a lesser extent in their attendance) at schools with student populations of different incomes. We can see that varied growth rates exist across our schools, both those with low- and high-proportions of students designated at-risk. As Wolf testified, “our accountability system measures family income more than it measures school quality.”

<sup>6</sup> For example, mastery of grade-level material in District academic standards, as measured by “proficiency” on annual standardized tests, is also regarded by many as a key, even required, element of quality.

<sup>7</sup> See OSSE’s [2020 D.C. School Report Card & Technical STAR Framework Technical Guide](#), pages 20, 54-58, and 107–108 for explanations of the Districts growth metrics.

Graph 3, below, shows that the growth of students designated at-risk—a critical aspect of school quality—is spread fairly equally across the city’s schools.

**GRAPH 3: 2019 At-Risk STAR Performance on Growth-Only Metrics**



These above three graphs indicate that the current rating system gives lower ratings to schools with high percentages of students designated at-risk *even when those schools led their students to grow in the same range as students at some of the higher-rated schools*. These data are the basis for the board resolution committing to “*adjusting and/or enhancing how the District measures and or reports school quality so that the STAR rating minimizes bias against schools with large numbers of at-risk student groups.*”

After reviewing data and testimony from experts like Cuny and Wolf showing the conflation of quality, income and other student characteristics, the Committee asked two key questions:

- What elements are most correlated to student income?
- Are there ways to adjust how we define or measure these indicators to reduce correlations between students designated at-risk (and other student populations) and measures of quality)?

***B. What STAR elements are most correlated to student income?***

Data from the Office of the State Superintendent of Education (OSSE) itself shows the indicators that make up the STAR rating at each school level and the correlations between each STAR metric and a school’s population of students designated at-risk. Based on Table 1, below, the greatest correlations across all grades are with PARCC proficiency scores, attendance, and re-enrollment rate. In high school, graduation rates and indicators around SAT and AP/IB scores are also highly correlated.

**TABLE 1: Linear Regression Analysis of Percentage of Students who are At-Risk on Metric Score, by Framework**

2019 STAR Brief: Appendices

*Table 5: Linear Regression Analysis of Percentage of students who are At-risk on Metric Score, by Framework*

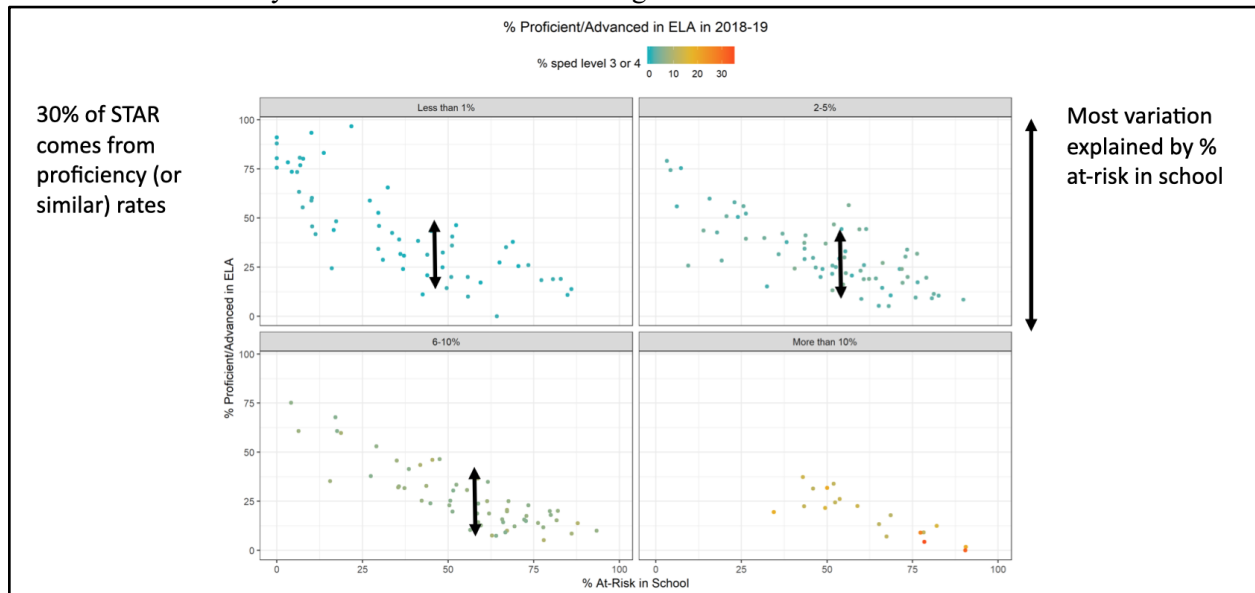
Metric	Elementary School	Middle School	High School
CLASS - Pre-K Classroom Organization	0.033		
CLASS - Pre-K Emotional Support	0.039		
CLASS - Pre-K Instructional Support	-0.006		
Pre-K In-Seat Attendance	0.557		
Growth to Proficiency – ELA	0.025	0.073	
Growth to Proficiency – Math	0.014	0.090	
Median Growth Percentile ELA	0.219	0.025	
Median Growth Percentile Math	0.101	0.064	
90% Attendance	0.653	0.339	0.242
ACCESS Growth	-0.016	0.087	0.258
Attendance Growth	0.006	0.084	0.098
In-Seat Attendance	0.595	0.290	0.239
PARCC 3+/MSAA3+ ELA	0.656	0.547	0.818
PARCC 3+/MSAA3+ Math	0.477	0.545	0.773
PARCC 4+/MSAA3+ ELA	0.641	0.605	0.794
PARCC 4+/MSAA3+ Math	0.553	0.527	0.679
Re-enrollment	0.295	0.271	0.561
AP/IB Participation			0.579
AP/IB Performance			0.721
Extended Years Graduation Rate			0.418
Four-Year Graduation Rate			0.510
SAT College and Career Ready Benchmark			0.824
SAT DC Percentile			0.743

### 1. PARCC proficiency

Graph 4, below, suggests the strong correlation between proficiency and a school’s enrollment of students designated at-risk. This graphic illustration is consistent with OSSE’s correlation data, which correlates PARCC proficiency and a school’s at-risk enrollment, with the correlation for high school English—a stunning .818.



**GRAPH 4: Proficiency and Percent of Students Designated At-Risk**



In her testimony, Wolf drew on the work of Dr. Sean Reardon, the Director of the Educational Opportunity Project<sup>8</sup>, to distinguish between test score “proficiency” and test score “growth.” The case made by Reardon and others is that, on average, students whose families earn relatively lower incomes enter school further behind than others. Insofar as all students make equivalent and strong progress in school, students from relatively lower-income families will still be farther behind and earn lower test scores. That proficiency gap does not mean that the school attended by such students is lower quality; it means that the students entered the school further behind.

Thus, as Wolf explains it, judging schools based on students’ absolute test scores (including whether students have reached a particular proficiency threshold) does not actually capture school quality because the final score is so related to where students began. Further, if these proficiency test scores are used to rate schools, they will result in ratings that are systematically lower for schools that enroll larger proportions of students from lower-income families. Additional testimony and slides from Wolf showed a similar disconnect between growth and proficiency in schools with relatively large enrollments of students with disabilities and English language learners (ELLs). This leads to an initial key consideration for the Committee: *Should we reduce or eliminate the proficiency indicator?*

## 2. Re-enrollment

With regard to the high correlation between student re-enrollment and at-risk designation, Wolf cites a report from the D.C. Auditor’s report<sup>9</sup> on enrollment, which found that mobility is higher for students designated at-risk and students from Wards 7 and 8; and that a student’s change in school often (1/3 of the time) coincides with the student’s residence having changed. Another witness, Erin Roth, Director of Education Research at the Office of the D.C. Auditor, testified that re-enrollment is often connected to student homelessness. Dr. Constance Lindsay of the UNC School of Education testified that high levels of student mobility in D.C. affects the accuracy of various indicators, including re-enrollment.

<sup>8</sup> Stanford University, [Educational Opportunity Project](#)

<sup>9</sup> D.C. Auditor (2018) “[A Study of Enrollment Projections for D.C.’s Public Schools: Assuring Accuracy and Transparency](#)”

In short, re-enrollment is substantially connected to housing stability and, therefore (like proficiency), may say more about the socioeconomic status of a school's students than about its quality. For this reason, Wolf, Roth, and Lindsay all noted that re-enrollment may be an inappropriate indicator for judging school quality and/or it may need to be redefined.

### 3. All-students sub-group

A school's STAR rating is based in part on the weighted scores of *all* students (75 percent) in the school and, in part, on the weighted scores of the school's most vulnerable student groups (25 percent). The rating is intentionally formulated this way so that a school is unlikely to get the highest rating if it focuses on its high-achieving students at the expense of attention to students designated at-risk. Cuny makes the case that insofar as the rating relies so heavily on a comparison of the performance of "all students," as opposed to a comparison of the performance of specific groups of students (e.g. designated at-risk, ELL status, special education (SPED) status, etc.), the ratings will give higher scores to schools with low proportions of percentages of students designated at-risk. This happens because, as noted above, higher income students will, on average, enter school and each subsequent grade at a higher score level than their less affluent peers.

### 4. Additional issues:

Other indicators with a high correlation to enrollment of students designated at-risk include attendance, high school graduation, and other elements used to rate high schools. The Committee has noted but not focused discussion on these indicators.

## ***C. What options exist for reducing the bias of these indicators?***

One key goal of the Committee is to ensure that our statewide accountability system evolves to most accurately reflect the quality of schools. *Are there ways to adjust the elements of the STAR Framework to reduce the conflation of quality with student characteristics?*

### 1. Greater reliance on growth, less on proficiency

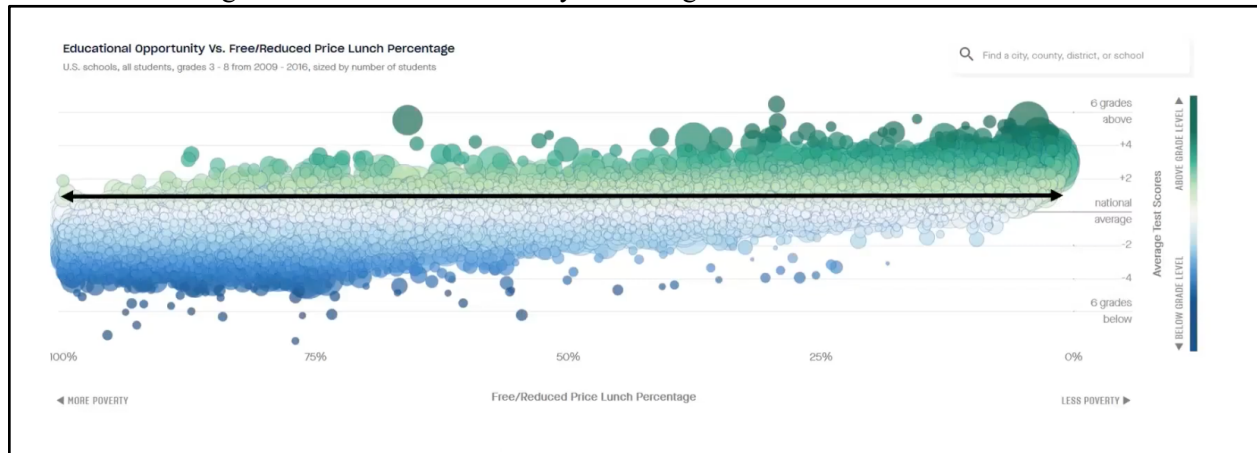
Standardized tests tell us important information about what our students know. We suspect that most families would not want to eliminate test scores from any picture of school quality. But, is there a way to harness their power for providing that information in a way that is not afflicted by the bias of proficiency rates?

One way is through using test score "growth" as an indicator of school quality. Basing ratings on test score "growth"—that is, how much students have grown or progressed in a given year based on their test scores—rather than on proficiency rates, allows us to use test scores, but in a way that does not discriminate against schools whose students start out behind. In fact, the current rating formula uses growth metrics as 40 points out of a total score of 95 at the elementary and total score of 100 points at the middle school levels; proficiency metrics are weighted at 30 points of the total scores for elementary and middle school students.

According to Reardon, proficiency reveals the "educational opportunity present in children's *lives*;" growth reveals the educational opportunity present in children's *schools*." By this logic: If we want to evaluate the

quality of a child’s *school*, we should be looking at *growth*, not proficiency. The difference between the two can be seen in Graph 5.<sup>10</sup>

**GRAPH 5: Average Test Scores Nationwide by Percentage Free/Reduced Lunch-School Level**



Wolf likewise argued that growth provided a useful window on school quality. She also noted that various growth metrics limited bias to a greater or lesser degree. She suggested that the Median Growth Percentile (based on a formula in which students’ growth is compared to the growth of “comparable” students) used in the STAR rating may contain unnecessary bias and encouraged the State Board to consider alternative growth-metrics. Roth argued for a growth metric based on PARCC scale scores.

Options raised by Committee members include increasing the emphasis on test score growth while decreasing or eliminating reliance on proficiency, while others believe proficiency deserves its prominence. This is further discussed under trade-offs. All Committee members look forward to the addition of a growth measure to the high school accountability framework, something OSSE has committed to but which has been stalled by the pandemic.

**2. Alter weights of student sub-group: Reduce the weight of the all-students group; increase weight of group of students designated at-risk; add student groups based on longevity at the school.**

Cuny believes that proficiency is an important indicator. He has proposed a way to retain proficiency as an indicator while reducing its biasing effect. He proposes rethinking how student groups get weighted in the rating, such that the rating makes use of “equitable student group weights.” One option he proposes calls for creating a new “not designated at-risk” student group and then equitably weighting all of the student groups (others are: designated at-risk, ELL status, and SPED status) in the rating. Another option would keep the “all-students” group (now weighted at 75 percent) but weight it the same as other student groups. Each of these options would create a more level playing field across schools that educate vastly different proportions of students who enter school and each grade at lower- and higher-levels, have disabilities, or are ELL. It would also minimize the current biasing effect of re-enrollment, as discussed above, and of attendance.

<sup>10</sup> For more on the work and findings of the Educational Opportunity Project, please see <https://edopportunity.org/explorer/#/split/us/schools/avg/ses/all/3.5/38/-97>.

Another option, proposed by Josh Boots, Director of EmpowerK12, would further break down the "All Students" group into additional student groups: new students, second year students, and students with three or more years at the same school. Research shows students who are new to school, regardless of socioeconomic background, struggle during their first year. Such a breakout could encourage more attention to these new students. It would also make it possible to view a school's success with longer-term students who have been at the school for multiple years, enough time, perhaps, for the school to have helped the student to overcome the initial educational disadvantage had experienced the school over a longer time.

### 3. Modify the definition of reenrollment

Wolf proposed redefining reenrollment so that it only counted students who had not changed residence in the past year. This would minimize (or even eliminate) the biasing effect of student transience and homelessness on the reenrollment indicator.

### 4. Use the "best of" a school's proficiency or growth rate

Under the proposal from Boots, where the rating would count the higher of a school's growth or proficiency rates, such an approach echoes how the current rating counts either a school's attendance rate or its attendance growth, whichever is better.

This is not an exhaustive list of possible options. It does suggest the variety of ways in which the current conflation of school quality and student characteristics can be addressed.

## **II. Other concerns with the STAR rating**

Committee members raised and heard from a number of experts about other concerns with the STAR rating. Among the key ones:

### ***A. Leading vs. lagging indicators***

Under the current system, a school is rated largely based on "lagging" indicators—that is, the school's scores are the result of the school's (and students') practices up until the point that the indicator is measured, but shed little light on whether a school is improving or not. This is especially true of test scores (proficiency or growth), which show what a student has learned until that point "but do not necessarily inform how well practices, people, strategies, materials, or technologies schools are investing in are leading to high student academic performance."<sup>11</sup>

Some Committee members believe this reliance on lagging vs. leading indicators can be extremely damaging, especially to a school that is on the upswing. For example: Suppose that a high poverty school, where students on average entered school with kindergarten skills that were substantially below-average. Each year, they learned at a normal, average rate, and thus across the grades of this school, students were, as they were when they entered school, substantially below average. Over the past year, changes were put in place in the school. The test scores at the end of this school were just as they had always been, since the changes had not yet had time to affect test scores. But thanks to these changes, the school's scores would rise in the next year and the next, and so on. However, if the effect of these changes is ignored in the rating, this school-on-the-rise would earn, under our current system, only one of two stars. As a result, teachers are demoralized and leaving, parents may find other places to send their children, the budget shrinks due to

---

<sup>11</sup> SBOE staff memo, "[Using Leading Indicators to Assess School Performance](#)" (September 3, 2020), page 3

enrollment loss—and, soon, this school-on-the-rise will be the failure it was labeled. But what if we could identify the new processes that were underway, that were ready to propel this school’s students to new levels of learning? We would surely want to include an “indicator” of this new process in the rating or report in an important way.

This kind of indicator can be described as a “leading” indicator. As an example, Chicago Public Schools uses the 5Essentials Survey to monitor school performance using five indicators, including “Supportive Environments,” “Collaborative Teachers,” and “Ambitious Instruction”. Research on the 5Essentials indicates that “Schools strong on at least three of the 5Essentials are 10 times more likely to improve student outcomes. A low score in just one of the 5Essentials reduced the likelihood of improvement to less than 10 percent.”<sup>12</sup>

A number of Committee members and various witnesses have argued that the rating should include more “leading indicators.” Climate surveys like those used in the Chicago example above are one example of such a leading indicator. Other leading indicators might include teacher retention rates, which research shows is highly connected to climate and school quality. (See Section III, C (below) for concerns that were raised about using climate surveys in school ratings.)

### ***B. Perverse (and constructive) incentives***

Education researchers have written a great deal about the perverse incentives that can lead schools to overly emphasize one important educational focus over another or to engage in “gaming” when specific indicators are attached to very high stakes. A classic example: If a school’s rating is heavily based on reading and math scores, the school may de-emphasize social studies, science, and the arts in a way that weakens students’ overall education. Another classic: If schools are judged based on how many students reach a particular test threshold—as opposed to their overall progress—there is a “perverse” incentive for the school to focus its educational efforts on those students who are the closest to the threshold (“bubble kids”) and thus the most likely to be pushed over the important score threshold and to generate an improved evaluation for the school. But the flip side of incentives is also consequential; if the incentive causes resources to be focused on improving a practice that will generally improve education, the incentive created by including that indicator in the rating can be positive. One example from our current rating system: 25 percent of the rating score is based on the performance of vulnerable student populations (i.e., designated at-risk, ELL status, and SPED status), ideally incentivizing schools to put extra resources/effort towards the education of students in these groups.

One specific perverse incentive in our system has to do with how our PARCC scores are reported and was raised by Roth. Her research indicates that the way the way 8<sup>th</sup> grade math scores are reported (in which scores on different levels of math tests are aggregated) may disincentivize schools from offering higher level math courses. Roth wondered whether our current reporting/rating system disincentivizes middle schools from offering algebra, as these students would then be taking a harder test and possibly lowering the school’s rating.

All Committee members have expressed concern that the current system de-emphasizes the importance of a well-rounded education, emphasizing reading and math scores above social studies, science, and the arts. Some have raised concerns about the over-emphasis on testing itself, which can spawn the use of too many

---

<sup>12</sup> UChicago Impact “[Illinois Uses the 5Essentials to Promote Strong Education Outcomes: How Legislative Action Catalyzed Statewide School Improvement in Illinois](#)” (2019)

practice and interim tests, gobbling up classroom learning time. A number of witnesses and members also noted that the focus on high proficiency scores may lead schools (despite rules that prohibit this) to prefer to enroll high-scoring students over other students; and to encourage, even push, current lower-scoring students to withdraw from the school.

### ***C. A single summative rating (vs. a dashboard)***

Under the current rating system, a school gets a single “summative” rating, specifically the school is awarded between one and five stars, supposedly indicating the school’s quality. Some members and witnesses have raised several concerns about the use of such a rating, including that:

1. By its very nature, it fails to capture the complexity and breadth of quality because it is based on a narrow set of indicators and so is, in an important way, inaccurate.
2. It can be demoralizing to staff and students, making it harder to improve, and—because of the potentially spiraling effects of a lower reputation on enrollment and then budget—hard to recover from.
3. As a composite of multiple indicators of different weights, what it actually measures is not immediately transparent and, ultimately--and necessarily--reflects debatable choices about the weights of different indicators.

Instead of the single summative rating, some members suggest that a dashboard could provide a fuller picture of a school’s quality. Dashboard supporters argue that a dashboard would also take some of the pressure off of particular indicators that might otherwise exert a “perverse” influence on school practices; it does not “brand” some schools as summarily worse than others; and, advocates argue, because it doesn’t mash up the measures of different indicators and weights, it provides a more transparent view of a school’s performance.

Some members noted that the single summative rating provided by the STAR rating is an important tool of empowerment for families and that a summative rating is vital to providing families the clear, accessible information they need. These important trade-offs are further discussed below.

### ***D. Connecting low-rated schools to support for school improvement***

Some members lamented that regardless of the pros and cons of the summative rating or specific indicators, a low rating (overall or in one area) does not clearly lead to financial, diagnostic, or technical support being provided to schools that are shown to be in need of improvement. (The exception: Title I schools rated in the bottom 5 percent of all schools in the state get extra financial support, mainly from the federal government as part of ESSA school improvement efforts.<sup>13</sup> For some members, this connection between identified weakness and needed, appropriate assistance should be among the top purposes of the rating system. How can we strengthen that connection?

Some Committee members are interested in looking at school inspections as a vehicle for providing clearer information about the changes schools need to make and the kinds of support they need. In the District, the charter sector uses school inspections to help rate school quality while also providing useful information to school staff (and the Public Charter School Board (PCSB) itself) to aid in improvement. The Committee heard from Dr. Matthew L. Blomstedt, Nebraska Commissioner of Education, about his state’s Accountability for a Quality Education System, Today and Tomorrow (AQuESTT) framework, which is

---

<sup>13</sup> From page 30–31 of the District’s [March 2017 State Plan for ESSA](#)



the basis for a diagnostic/improvement process that includes the school's rating themselves, reviews of the ratings and walk-throughs/inspections by independent third-party contractors and state officials, evidence-based analysis, providing diagnostic information to schools (and districts), and, for selected schools, a quality improvement process involving the state or districts.

Committee members also raised the possibility of bifurcating the rating in order to distinguish between those factors over which the school had control (e.g. student growth, teacher retention, school climate) and those that were more related to resources and student demographics, over which the city and/or LEA had greater responsibility and ability to address.

#### ***E. Effects of low ratings on schools***

Committee members and witnesses noted a number of ways in which low-rated schools were hurt by the low rating. Robert Simmons, Executive Director of the Black Educators Initiative, talked about its demoralizing effect on staff and students and the increased pressure it placed on staff. Members spoke of how it spurred enrollment declines at some schools, which then triggered a vicious, continuing cycle of budget cuts, quality declines, and further enrollment losses. Other members noted that while the content of the STAR system merits redesign, parents deserve to know how schools perform as compared to state expectations and that right to know outweighs effects on school staff morale or feeling of pressure.

#### ***F. Capturing the full range of quality***

As discussed above, there was much discussion about whether proficiency rates (and to a lesser extent, re-enrollment) actually reflect school quality—or whether they more reflect the socio-economic status of the school's students. But what about the other indicators with high correlations to enrollment of students designated at-risk—attendance, chronic absenteeism, high school participation and achievement in AP/IB, and SAT scores? *Are these reasonable indicators of quality? Do they reflect an adequately broad range of what we regard as school quality? An adequately broad range of practices and values that we most want schools to embody? Are there other indicators that might do a better job? Or do we want to make use of a small set of key indicators?*

Our conversations on this topic are far from complete, but among the aspects of quality or possible additional indicators raised by members or witnesses are those that would measure: student engagement, teacher retention (including special attention to retention of non-white teachers), career readiness (including provision of career credentials, development of self-management skills and growth mindset), school safety, suspension rates, diversity of staff and students, and disproportionality in special education. Many witnesses and members were interested in school climate (including student satisfaction, sense of belonging and socio-emotional well-being), with specific interest raised in using qualitative measures (including interviews and focus groups) as well as quantitative measures and for paying special attention to non-white students' and teachers' sense of belonging or feeling marginalized in majority-white schools. There is a good deal of support from Committee members for an indicator around well-rounded education.

All members believe school climate is an important aspect of school quality. As noted earlier, certain aspects of school climate are predictive of future progress; including it as part of any rating can help identify schools that are on the upswing. Further, including climate in the rating signals the importance of it in school quality and may incentivize schools to focus on it as part of their improvement work. For these reasons, a number of members believe it needs to be a core part of either a rating or a dashboard.

But including climate in the rating could subject the rating to “gaming,” potentially hurting the reliability of the measure. For this and other reasons, some members believe that a measure of climate should not be included in a rating. This is further discussed in Section III, C (below).

### III. Options for addressing these concerns and inherent tradeoffs

#### *A. How to handle proficiency and growth*

For some Committee members, the inherent bias of proficiency, as explained by Reardon, is ample reason for eliminating proficiency from the rating formula or at least minimizing it; in the view of these members, the rating should reflect the quality of the school, not the proportion of lower income students the school enrolls, which proficiency inherently does. The rating should, in other words, rate only what’s in the school’s control.

**But there are counterarguments:** Our overall educational goal is, in fact, to bring *all* students to the proficiency level, regardless of where they started; bringing all students to that goal is the job of the school. Insofar as schools do not meet that goal, they cannot, by definition, be regarded as high-quality. A slightly different argument is that proficiency raises expectations that have for too long been too low; the goal of proficiency is so important that we must create a social pressure for schools to help students reach it, even if it’s much harder for some schools than others. In this view, the greater challenge faced by the lower income school to get a high rating is offset by the importance of sending the right signal: Students, families, and schools should be aiming for proficiency; that’s the definition of success and should be a key part of what earns a high rating. There may also be mid-way options, such as those proposed by Cuny and Boots to modify how subgroups get weighted and/or how indicators get defined.

*Is there a way to adequately maintain the important “signal” carried by proficiency without contributing to a bias in the rating? That’s a vital question that Board members have discussed and expect to return to.*

#### *B. How to display accountability data (summative rating, data dashboard, quadrants)*

As discussed above, some members are wary of a single summative score, believing that it cannot capture an adequate range of the factors that ultimately determine “quality” and that its composite character reflects a singular and non-transparent view of what defines quality. For these reasons, some Committee members want to consider alternatives to a single rating, such as a “dashboard,” which is used by eight states instead of a summative rating and by five others that used a dashboard in addition to a summative rating.<sup>14</sup>

**Another option:** Instead of rating schools with stars, schools would be placed in one of four quadrants, indicating the school’s combination of proficiency and growth. A school’s placement in a given quadrant would also provide a clear signal about the schools’ needs and the intensity of those needs.

**Other members believe the single summative rating is critical.** In this view, the summative score provides parents with a clear way to understand how students are performing academically compared with statewide expectations for students at their grade level. Parents can use that information as part of their choice process in identifying a school that best meets their learning needs. These members argued that while the content of the STAR system merits redesign, parents deserve to know how schools perform as compared

---

<sup>14</sup> See the State Board staff’s June 16, 2020 Memo: Non-Summative ESSA State Accountability Plans; See [https://nces.ed.gov/programs/statereform/tab1\\_13.asp](https://nces.ed.gov/programs/statereform/tab1_13.asp) for report on states’ approach to school ratings in their ESSA plan. States reporting no summative rating in their school ratings are California, Idaho, North Dakota, Oregon, and Pennsylvania.



to state expectations and that right to know outweighs effects on staff morale. In contrast to the summative score, they believe that a dashboard is overly dense and hard to understand. Further, advocates of a single score worry that the signal sent to schools by a dashboard will be unhelpfully diffuse, whereas the signal conveyed by the summative score is much clearer.

### ***C. How to handle school climate data***

While some members believe school climate is crucial to understanding school quality (see above) and, therefore, must be a key factor in any judgment or portrayal of school quality, others have raised substantial concerns with including climate indicators in any summative rating. They have noted that some climate measures were designed as research instruments and may not have been validated for educational purposes; they might not stand up to the complexities of practice. Further, if the survey results have an impact on accountability, there may be a tendency to over- or under-rate a school in order to boost (or damage) a school's reputation, thus rendering the data invalid. Finally, depending on how the climate data was used to report or rate a school, it was noted that a dashboard might not meet the ESSA disaggregation requirements for accountability indicators.

### ***D. Other options***

- *Customizable rating:* Provide parents and others the ability to create their own customized ratings by weighing various factors at the levels that reflect their values.
- *More indicators:* Including more indicators in the summative rating so that there is less incentive to “game” the system.