Appendix C1:

Technical Appendix and Modeling

Overview and Update of Activity

In accordance with ESSA, Illinois State Board of Education (ISBE) has been working to develop a new accountability system that provides a more comprehensive approach to understanding student and school performance, with the ability to identify and support schools that might be excelling or underperforming in terms of academic and equity outcomes.

The legislation allows SEAs to develop a summative measure that aggregates several measures of school progress, including:

- Academic achievement (K-12)
- English language proficiency (K-12)
- Student growth or another valid and reliable statewide academic (K-8)
- Graduation (and extended graduation) rate (high school)
- At least one school quality or student success indicator (K-12)

As of August 25, 2016, ISBE has hosted three accountability work sessions, with a diverse group of stakeholders, to gather feedback and insight into the development of an accountability system that is both equitable and educative for schools, districts, and stakeholders. (See Appendix C for agendas and meeting minutes.)

Over the past several weeks, the Accountability Workgroup (and other similar committees, e.g. P20 Data, Assessment and Accountability Committee, and the Illinois Balanced Assessment Measure workgroup) have weighed in on several aspects of each of the accountability system, including possible indicators for academic, school climate, engagement, post-secondary readiness, access to advanced coursework, non-academic, and college-and-career domains (see pages 15-16 of ISBE Draft Plan 1).

With the recognition that Illinois does not currently have or collect standard, state-wide data collection for most of the suggested possible indicators above, the Accountability Workgroup was also interested in understanding the various options, approaches, and "behaviors" of the basic data that according to ESSA, would need to be part of the accountability system, specifically in terms of identifying schools in need of supports and intervention (e.g. K-12 academic achievement, K-12 EL proficiency, K-8 academic growth, HS graduation, and at least one school quality or student success indicator).

To that end, led by the Accountability Workgroup's Technical Steering Committee, ISBE has been working with staff from The Opportunity Institute, to run basic simulations to: (1) unpack and better understand each component or measure within the system; (2) see what an aggregate score might entail (as per the basic parameters of the legislation), and (3) model different approaches to each of those measures.

Specifically, ISBE and the Accountability Workgroup have requested analyses using modeling and existing data sets and patterns to supplement their discussion on:

- Methodologies and assumptions of **student academic growth measures**, including Student Growth Percentiles, Value Tables, Growth to Proficiency, and Hybrid models;

Illinois Student Academic Growth Calculation: Simulator

BACKGROUND

On October 28th, November 4th, and November 14th, as part of understanding and guiding each facet of the full accountability system, the Accountability Workgroup's Technical Steering Committee requested and reviewed models **of student academic growth calculations** as a first step to discussing how the growth component might be integrated into the larger aggregate school performance metric. The specific request was to model student growth percentiles (SGP), Value Tables and growth to proficiency, and hybrid models.

Current models indicate promise for each of the statistical approaches to measuring academic growth under discussion. In moving forward, ISBE and the Accountability Workgroup have been asked to consider the following:

1. Which approaches to student academic growth have appeal and which ones do not? Why or why not? Does the committee want to pursue all of the options outlined below, or a select few?
2. Are there additional approaches to student academic growth which are not included that the committee would like to see explored? Interest in a multiple measure approach to academic growth, to provide a more accurate (but more complicated) approach to growth?
3. Once the approach to student academic growth has been chosen, can the steering committee provide guidance on the multiple decision points that are part of each calculation type?

PROXY DATA

In order to preserve the reliability of any of the proposed modeling approaches, the first step focused on creating a simulator that generates sample school and student data in a way that matches Illinois school representation (see below for specifics) – and anticipates/moderates the data inconsistencies and gaps that are part of the "real" data set. This allows us to provide the requested simulations for discussion purposes.

The simulations were based a proxy data set of 100 schools that reflect Illinois demographic and enrollment patterns. The following patterns are integrated into the generation process:

1. Proportion of elementary, middle and high schools
2. Distribution pattern of student demographics, specifically:
    a. Individualized Education Plan status
    b. English Language Learner status
    c. Free/Reduced Lunch status
3. It then generates a unique set of up to 700 students based on a randomly assigned demographic profile of the school, and grades served & tested;
4. Test results are generated for those students, as a combination of:
    a. Randomness -- any possible outcome within the established range;
    b. Score distribution that is designed to mirror historical outcome patterns based on demographics;
    c. Inter-year dependencies to minimize outlier spikes and increase alignment to historical trends.

STUDENT ACADEMIC GROWTH MEASURE SIMULATIONS

Specifically, the Steering Committee expressed interest in the correlation of growth metrics to aggregated proficiency scores and the larger question of whether to pursue a multiple growth metric approach. As such, the simulations were designed to assist in answering two questions:

1. What is the correlation between aggregate academic performance scores and different growth approaches?
2. How highly correlated are different growth treatments to one another?

The simulations modeled three measures:

- Student Growth Percentile – originally requested
- Value Table – originally requested
- Growth-to-Proficiency (requested on October 28th)

The simulations were ***not*** intended for discussion of how specific categories of schools (e.g. schools with high ELL populations, schools with high poverty rates) are ranked under different growth calculations, but to simulate their interplay. Looking at correlations of growth measures with proficiency scores (of particular interest to the committee) and each other can inform this phase of the growth measure conversation.

Limitations of Simulations using Proxy Data

Some limitations of this stage include the following:

- As it relies on some random generation of data, the school-level indicators are not representative of specific schools. However, the overall trends are representative of the relationships between calculation treatments.
- The steering committee's discussion about the tension between the "purity" of a growth measures, versus building in demographic predictions are important to consider and can be addressed, but not at this stage.
- The behavior of the indicators is logical and statistically defensible, but because of the simulated data size limitations, the measure is not sensitive enough to capture a full array of scores.

Statistical Treatments

VALUE TABLE

Generation

The Value Table is calculated as follows:

- Students' scaled PARCC scores are assigned to Proficiency Levels using the published scales[1]
- Any student whose score goes up 1 level earns 25 points. Any student whose score declines 1 level loses 25 points[2]. Multiple levels earn or lose more points (2 levels = 50 points, 3 = 75, etc).
- Student point gains/losses are averaged. This is the school's Value Table score. It carries a range of -100 to +100, with a theoretical center at 0.

The school's Value Table score is aggregated as an average of all students, or all students in a chosen subset.

---

[1] Published scores refers to Illinois' existing PARCC cut scores, available at www.parcconline.org

[2] 25 was chosen as a standard Value Table parameter

Assumptions

The Value Table was built with the following assumptions:

- Standard 100-point scale derived from changes in students' PARCC Proficiency Levels (PLs).
- The PLs are a 5-point scale.
- A linear scale that gives schools 25 points for each level gain experienced by a student and deducts 25 points for each level drop.

Note: Due to the limitations of generating sample data, Value Table calculations will tend towards 0 more than when we transition to live data.

Decision Points

The following decision points will need to be resolved in developing a Value Table algorithm:

- Are we going to link it to Proficiency Levels or some other subdivision of PARCC scores?
- Is the linear scale appropriate, or do we want to take another approach? For example, the scale could give greater weight to larger incremental gains or losses.

GROWTH-TO-PROFICIENCY (GTP)

Generation

The Growth-to-Proficiency measure is a calculation that matches student performance against growth expectations toward a defined standard (e.g. proficiency) at a particular point (e.g. 8th grade). Of the list provided, GTP includes the greatest number of decision points that need to be addressed, as outlined below.

The GTP modeled in the simulation works on a binary flag system, in which students are "flagged" as either meeting their growth-to-proficiency target or not. In this model, students can make GTP in one of two ways. Either they are currently testing at or above grade-level proficient (PL 4 or 5), or the change in their score from last year to this year is equal to or greater than the linearized incremental growth the student would need to make in order to test proficient by grade 8 or grade 10.

Alternatively, a continuous scale may be used. Instead of applying a binary yes/no flag, it would assign a score based on a student's progress toward the goal. The benefit of this approach is that it captures partial growth and growth above a single year's target.

The school's GTP score is aggregated as the total number of students, or students in a chosen subset, who met the growth target divided by the total number of students or students in the subset.

Assumptions

The GTP was built with the following assumption:

- Students are flagged as achieving the requirement if they are proficient or if the linearization of their score leads to proficiency by 8th grade.[3]

[3] This is based on the assumption that the next modeling tool will focus on grades 3-8 data

Decision Points

The following decision points will need to be resolved in developing a GTP algorithm:

- Do we use a continuous scale or binary measure?
- Do we include students who are already proficient as meeting growth, or do they have to at least maintain their level?
- What projection model should be used? Linear or non-linear growth? How will the growth benchmarks be generated?
- Is there a high school performance metric for which GTP would be an appropriate measure?

Note: If we include normalizations for other students in similar situations, the GTP measure will be acting similarly to a value-added measure.

STUDENT GROWTH PERCENTILE (SGP)

Generation

As the SGP cannot be calculated without broad population data for a comparison group, it was synthesized from the scaled scores and the Value Table. While it is not an accurate reflection of the SGP scores that schools might earn, it is a reasonable estimate of the calculation interplay ***behavior*** that can be expected (i.e., larger score gains tend to translate to higher SGP scores). This is a limitation of the tool that will be removed when we build the live data modeling version.

The SGP is aggregated as a median of all students, or all students in a chosen subset.

Assumptions

The SGP was built with the following assumptions:

- 2 years of data
- Students matched by exact scores in the prior year

Decision Points

The following decision points will need to be resolved in developing an SGP algorithm:

- How many years of data should be considered? While 2 is the most common number in state SGP systems, student scores could be matched for 1, 2, or 3 years.
- Are we matching students by exact scores, static score bands or dynamic score bands?
- If we are using score bands, how wide? Wider score bands mean larger data pools for student comparison groups, but they also mean lower precision of measure.

HYBRID MODELS

Whereas the approaches listed above are stand-alone growth calculations, hybrid models bring together multiple approaches to student growth to get a more comprehensive picture of student growth.

a. Advantages – Hybrid models allow for increased flexibility in looking at data. Based on the fact that two measures of growth may capture different aspects of growth, a hybrid model may better represent the reality of growth.
b. Disadvantages –When using a hybrid model, it is difficult to balance the use of multiple growth measures. Use of such a model increases the complexity overall and may be more difficult to communicate to stakeholders.

Best fit questions – Since this approach involves a hybrid of two student growth measures, the best fit questions from each individual approach may be answered by the modeling. Another question that may be addressed is, "How do we situate growth in a space between individual measures?"

Exploring Outcomes of Growth Calculations on Model Schools

In order to better understand the information provided through different growth models, the modeling team used simulations to explore growth outcomes of schools serving different populations. This was done in two steps:

1) Creating a set of 3 demographically diverse schools, simulating their outcomes, and seeing whether their proficiency and growth scores were correlated;
2) Looking for 3 demographically diverse schools within a set of simulated schools to see how well their growth metrics were aligned.

Step 1: Proficiency and Growth: Are we looking at the same thing?

The first question explored through this phase of modeling is "do the academic growth metrics tell us something significantly different from relative academic attainment?"

To explore this question, three model schools with very different student populations were selected to illustrate the relationship between proficiency scores and growth models under investigation. They were chosen by establishing demographic profiles, running them through a simulated growth model, and then looking at their results.

	School A -- Alexander School	School B – Bryant School	School C – Calvin School
Grades Served	K-8	K-8	K-8
English Learner	80%	5%	10%
Free/Reduced Lunch	75%	10%	70%
Special Ed/IEP	30%	25%	5%

KEY:	Alexander	Bryant	Calvin

Model School Performance, by calculation type

	School A -- Alexander School	School B – Bryant School	School C – Calvin School
PARCC Post Test Aggregate Score	734	742	750
Value Table Score	0	0	0
Student Growth Percentile (SGP) Score	50%	50%	50%
Growth to Proficiency (GTP) Score	85%	87%	85%

Step 1 Findings

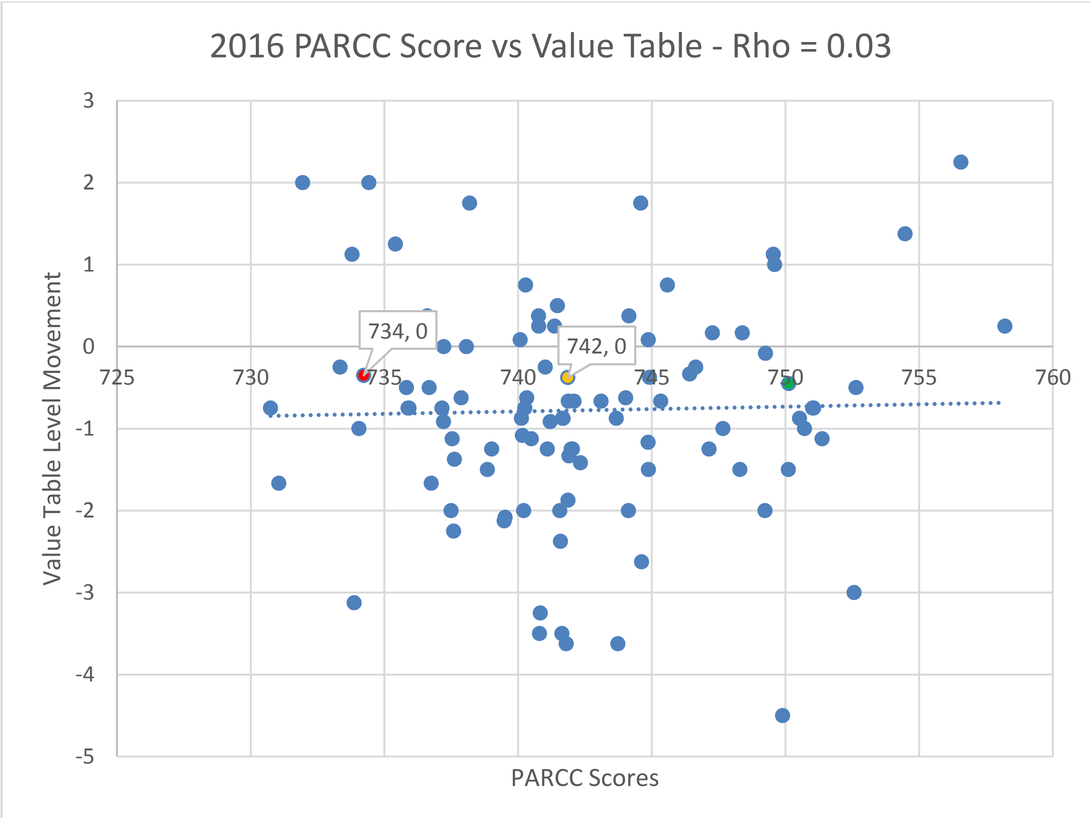
In each case, schools serving higher proportions of low-socio-economic status (SES, using Free and Reduced Price Lunch as proxy), English Learner (EL) and/or Special Education students have scored lower aggregate scores on the PARCC post-test. However, their growth scores on all three models are very similar.

This simulation suggests that the growth measures are not calculated in such a way as to penalize lower-performing schools or those which serve higher proportions of low-SES, EL and/or Special Education students. It further suggests that the growth metrics are generally measuring growth, independent of strict, single-year performance.

This does not imply that the actual data will show similar trends (i.e. that growth scores will be evenly distributed across demographics). This only shows that the growth scores, as calculated, are not *necessarily* linked to student demographics. Correlations between schools' demographic profiles and growth scores are not a consequence of the growth algorithms. Schools serving low-SES, high-EL and/or high Special Education populations may, in fact, have lower aggregate growth scores than schools that serve smaller proportions of these populations. However, if this is the case, it will not be because the measures themselves are biased.

We will know more once we complete the live data modeling.

Comparing PARCC Scores to growth as calculated through the Value Table method



	School A -- Alexander School	School B – Bryant School	School C – Calvin School
Grades Served	K-8	K-8	K-8
English Learner	80%	5%	10%
Free/Reduced Lunch	75%	10%	70%
Special Ed/IEP	30%	25%	5%

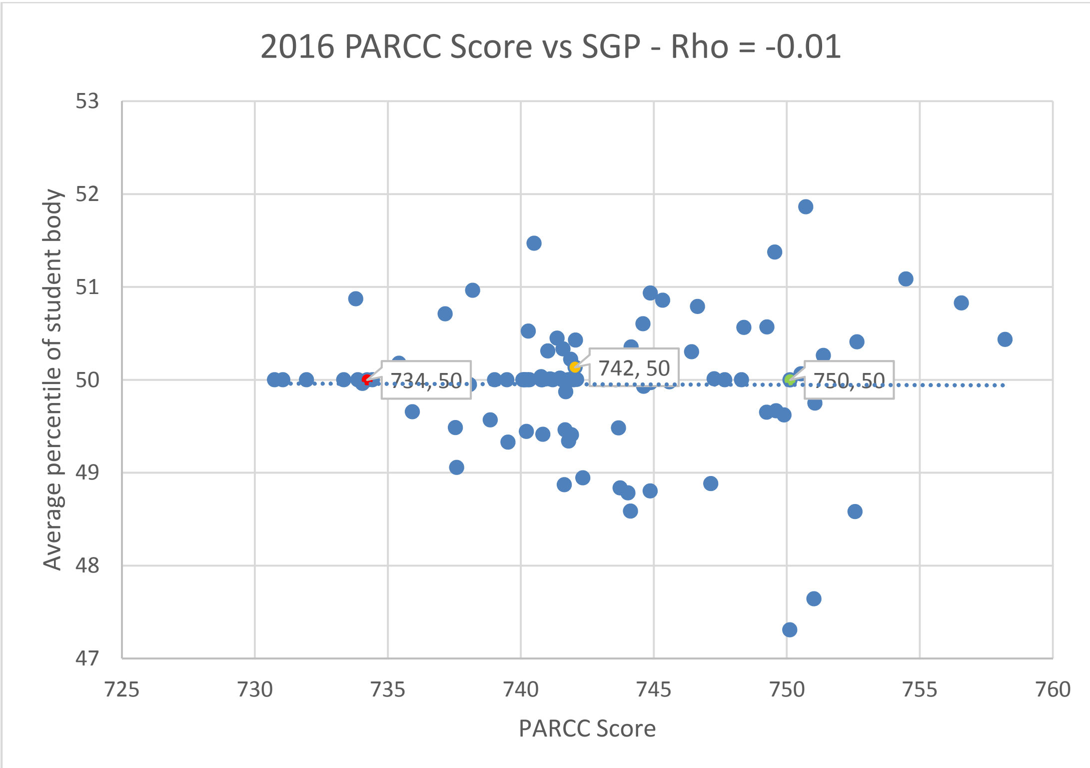
KEY: Alexander Bryant Calvin

	School A -- Alexander School	School B – Bryant School	School C – Calvin School
PARCC Post Test Aggregate Score	734	742	750
Value Table Score	0	0	0

Value Table scores indicate the "level" of movement of a school in the table. A score of 25 would indicate that the average student went up a whole proficiency level (i.e. from nearly proficient to proficient). A score of -25 would indicate that the average student in the school went down one proficiency level. A score of 0 would indicate that the average student remained at his or her current proficiency level.

In this case, each school has a different PARCC score, ranging from 734 (Alexander) to 750 (Bryant). However, none has made a significant shift in their Value Table score, with an effective score of 0. It is important to note that there can be a lot of difference in two schools achieving the same Value Table score, as it is calculated as an aggregate. For example, one of these schools might have all of its students remaining at exactly the same proficiency level, while another could have half of its students making significant gains while the other have backslides significantly. In both of those cases, the institutional Value Table score would be at or near 0, but the internal variability would be very different.

Comparing PARCC Scores to growth as calculated through the Student Growth Percentile method



	School A -- Alexander School	School B – Bryant School	School C – Calvin School
Grades Served	K-8	K-8	K-8
English Learner	80%	5%	10%
Free/Reduced Lunch	75%	10%	70%
Special Ed/IEP	30%	25%	5%

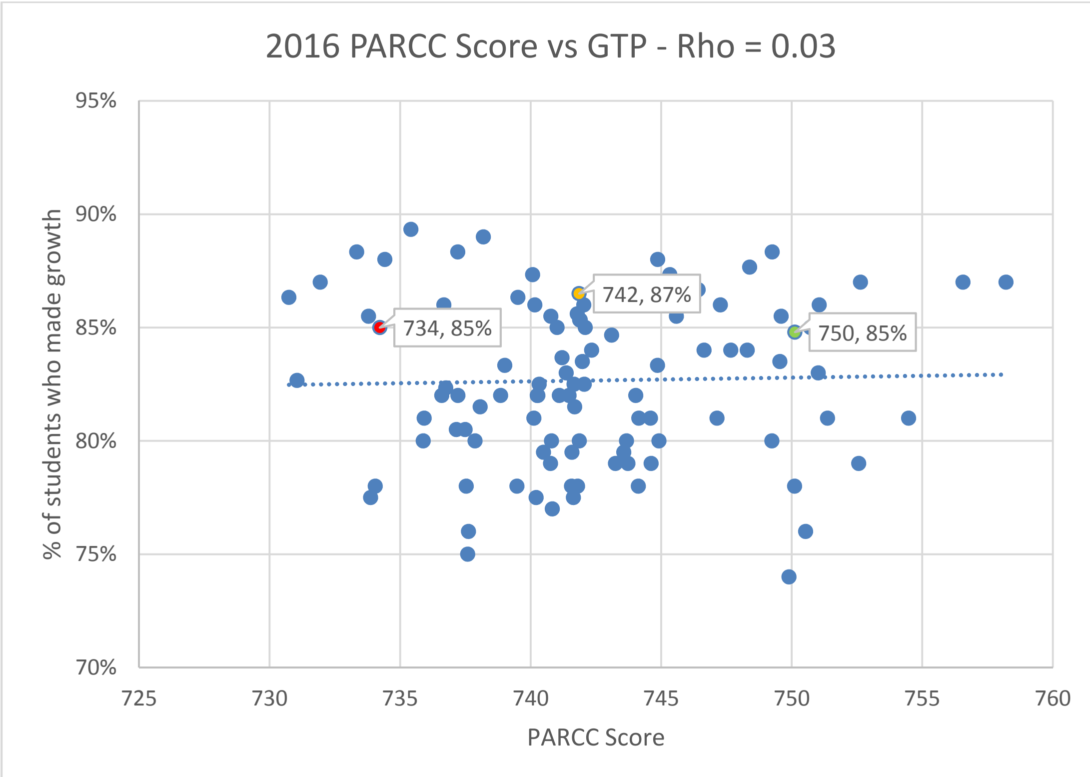
KEY: Alexander Bryant Calvin

	School A -- Alexander School	School B – Bryant School	School C – Calvin School
PARCC Post Test Aggregate Score	734	742	750
Student Growth Percentile (SGP) Score	50%	50%	50%

The Student Growth Percentile score of a school indicates the median movement of students in the school, as compared to their score peers throughout the state. Each student's current score is ranked, based on the scores of the students who had similar or identical scores in the last 1, 2 or 3 years. A student's score of 50 indicates that the student's current score was higher than half of the students with similar scores and lower than half of them. A school's score of 50 indicates that the median student in the school (the student for whom half of the students scored higher and half scored lower, on the SGP measure), ranked in the 50^{th} percentile on score in his or her own exam, based on the cohort of similar-scoring peers on prior exams.

As noted above, each school has a different PARCC score, ranging from 734 (Alexander) to 750 (Bryant). However, all received an SGP of 50, suggesting there's little to no relationship between these schools' PARCC scores and their growth as calculated through the SGP.

Comparing PARCC Scores to growth as calculated through the Growth to Proficiency method



	School A -- Alexander School	School B – Bryant School	School C – Calvin School
Grades Served	K-8	K-8	K-8
English Learner	80%	5%	10%
Free/Reduced Lunch	75%	10%	70%
Special Ed/IEP	30%	25%	5%

KEY: Alexander Bryant Calvin

	School A -- Alexander School	School B – Bryant School	School C – Calvin School
PARCC Post Test Aggregate Score	734	742	750
Growth to Proficiency (GTP) Score	85%	87%	85%

The Growth to Proficiency score of a school indicates the percentage of students in the school who met their annual proficiency goal. While there are many GTP models, this simulation credits students with meeting growth if they satisfy one of two criteria: (a) they achieve proficiency on the current-year exam, or (b) the delta between this year's exam and last year's exam, extrapolated to the 8th grade or high school exam year, would make them proficient by the terminal exam.

As noted above, each school has a different PARCC score, ranging from 734 (Alexander) to 750 (Bryant). However, all received close GTP scores, indicating proficiency scores were not closely linked to the growth of the student body as measured by GTP.

Step 2: Different Growth Calculations: Do they tell us different things or the same thing?

The second question explored through this phase of modeling is "do the different growth metrics tell us the same thing, or do we get a different picture by using different calculations?"

To explore this question, 3 demographically diverse schools were chosen from within a set of simulated schools. We then looked at their performance on the three growth metrics under consideration to illustrate their relative alignment.

	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

PARCC Scores, 2015 and 2016

Before exploring the different growth metrics, it's useful to take a look at the performance data, both the scores and the 2-year difference, or "Delta PARCC," in scores. This provides a baseline for considering whether and how a growth metric may be providing different information than simply asking "did the number go up?"

School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

Model School Performance, by calculation type

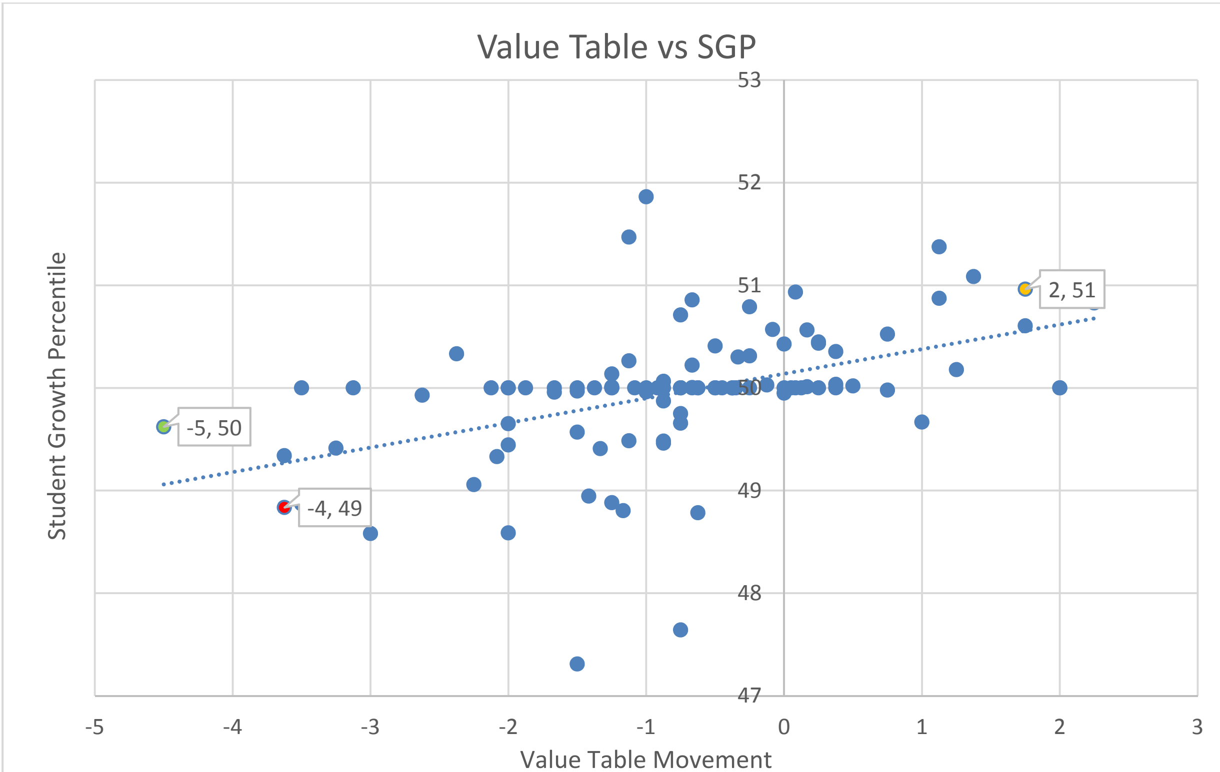
School	Delta PARCC	SGP	Value Table	GTP
Maple	-4	49	-4	78%
Elm	-2	50	-5	74%
Aspen	+2	51	+2	81%

Step 2 Findings

While the general trend shows that the measures report relatively consistently, there are some differences in the measures. For example, Elm is higher than Maple in Delta PARCC and SGP, but below in GTP and Value Table (though they are close in all measures). This tells us that the growth measures are capturing different phenomena. For example, it may show that one school is moving more students incrementally, while another school has larger gains by a smaller number of students.

The next step is to determine which metric(s) to choose, whether a blended approach is appropriate, and how to pursue an aggregate growth metric.

Comparing Value Table Scores to Student Growth Percentile Scores



KEY: Maple Elm Aspen

	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

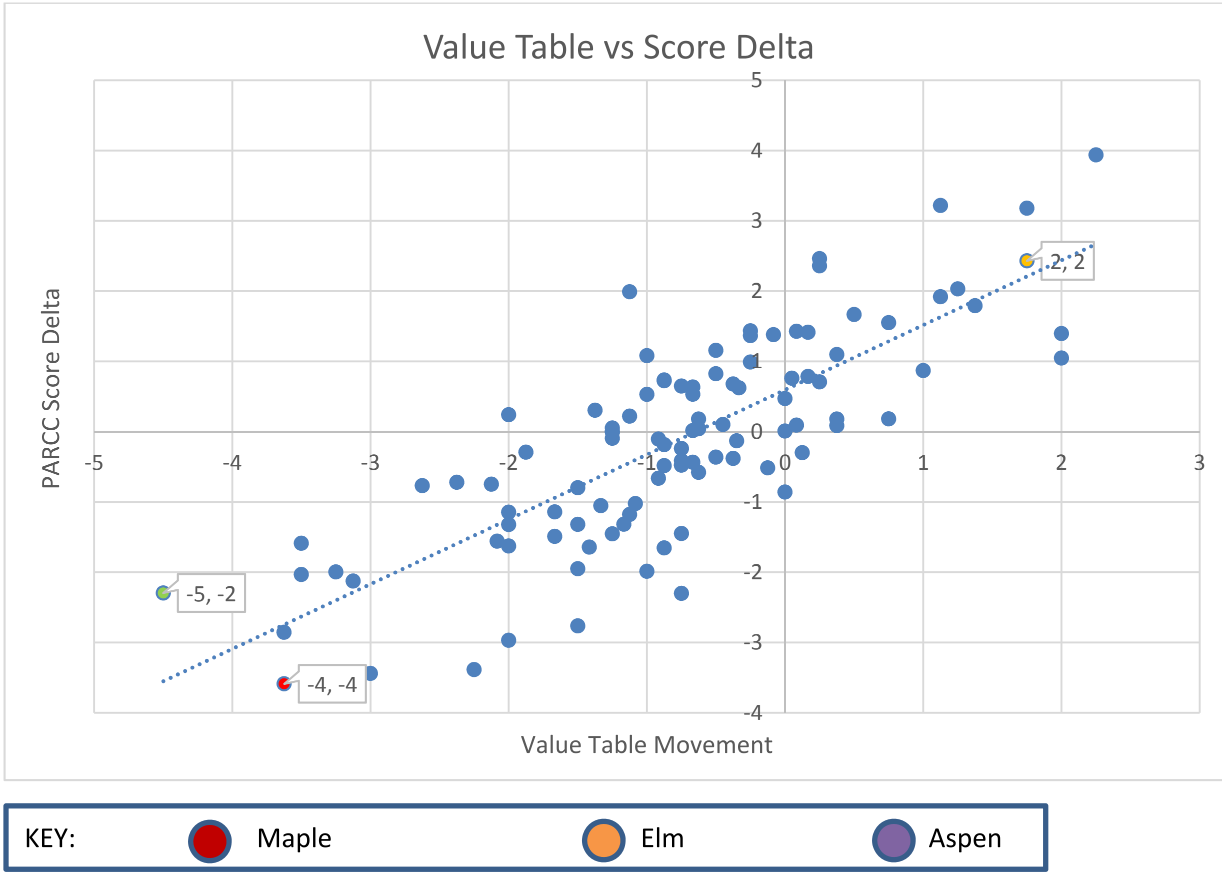
School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

School	SGP	Value Table
Maple	49	-4
Elm	50	-5
Aspen	51	+2

In this example, you can see that Elm moves further down the value table than Maple, despite their better SGP score. You can also see that Aspen receives a +2 value table score for a 1 point SGP improvement, while an equivalent negative SGP score results in a Value Table loss 2.5 times greater.

This suggests that the Value Table calculation reflects a different growth phenomenon than SGP. Given that they are calculating different things (magnitude of proficiency-level gain/loss vs aggregated score gain, relative to one's prior score peers), this is not surprising.

Comparing Value Table Scores to PARCC Score Delta



	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

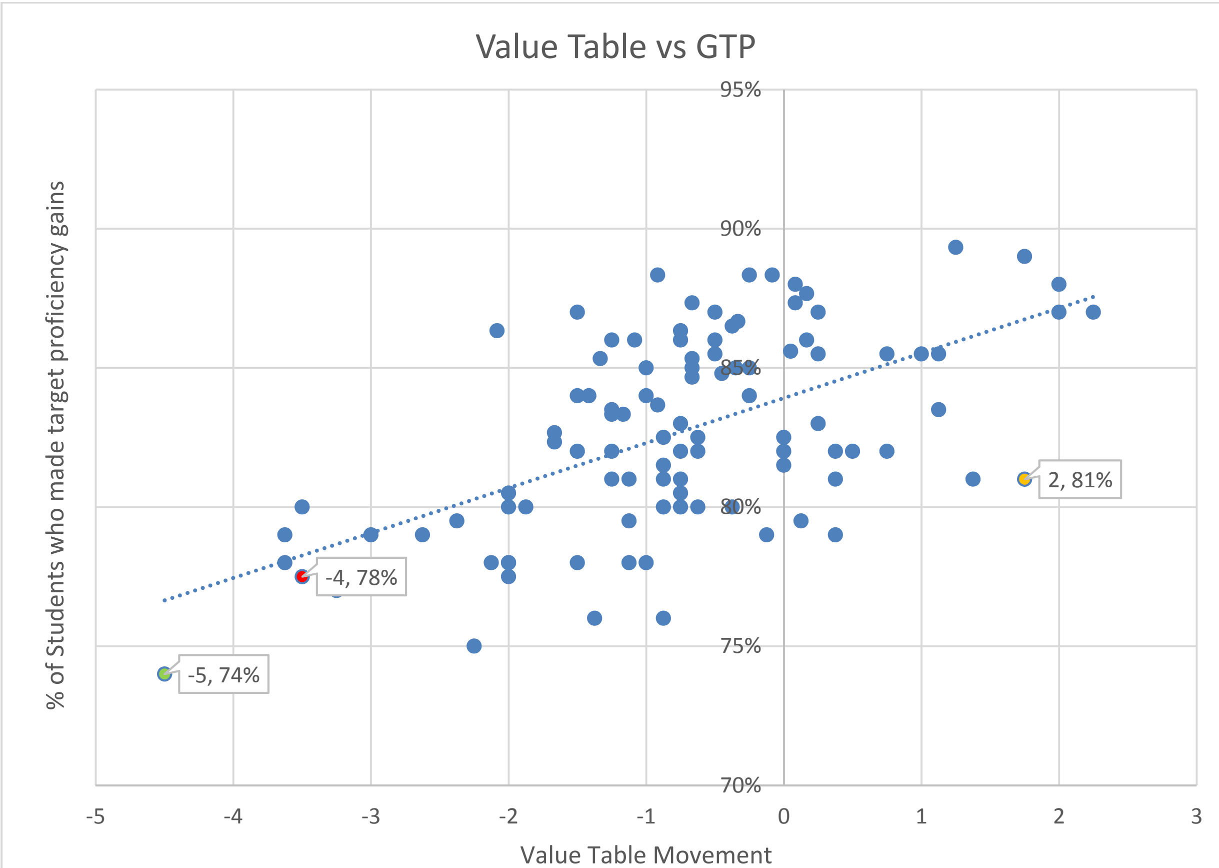
School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

Model School Performance, by calculation type

School	Delta PARCC	Value Table
Maple	-4	-4
Elm	-2	-5
Aspen	+2	+2

In this example, you can see that Elm moves further down the value table than Maple, despite their better PARCC Score Delta. You can also see that Aspen receives a +2 value table score for a 2 point PARCC improvement, while an equivalent negative PARCC Score Delta results in a Value Table loss 2.5 times greater. This suggests that the Value Table calculation reflects a different growth phenomenon than simply improving or declining proficiency scores. When one considers the relative sizes of the proficiency level “buckets” on the PARCC, it is not surprising that these measures would have differences.

Comparing Value Table Scores to Growth to Proficiency Scores



KEY: Maple Elm Aspen

	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

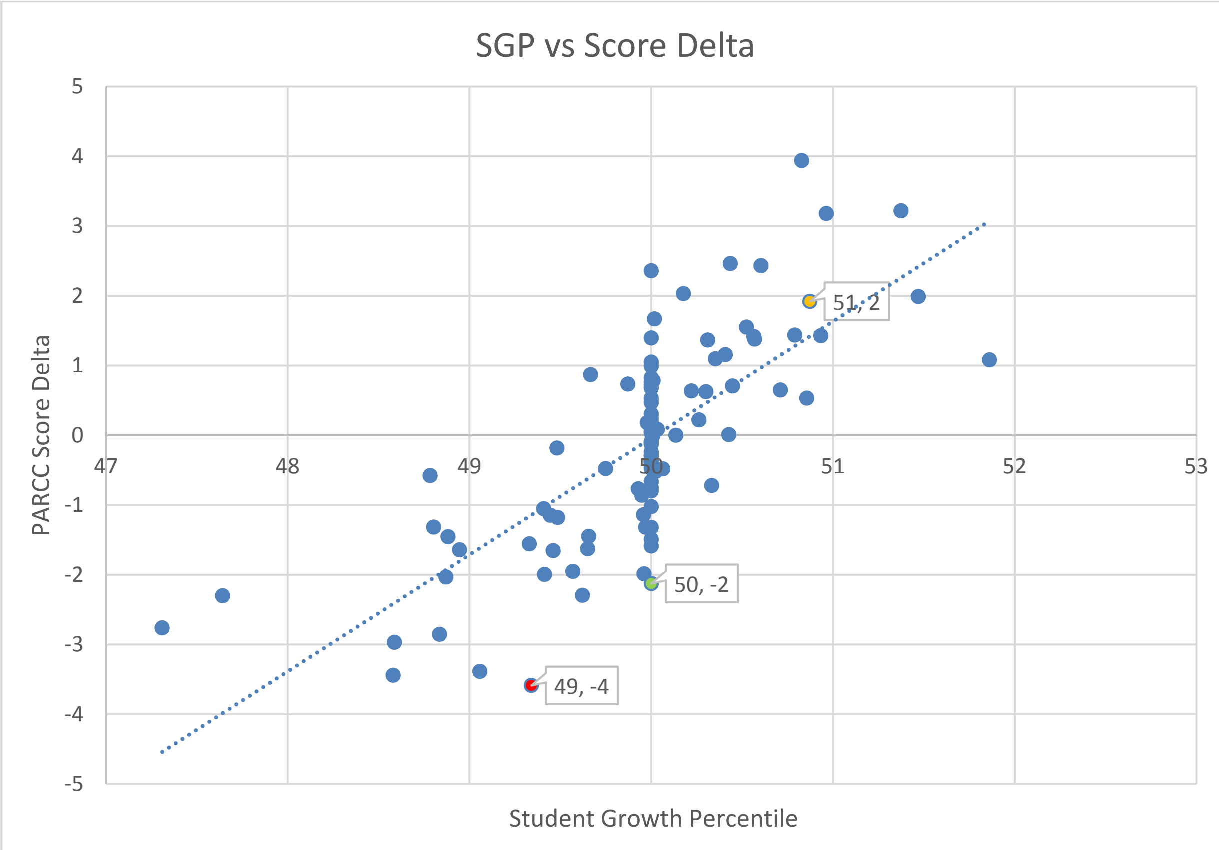
School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

Model School Performance, by calculation type

School	Value Table	GTP
Maple	-4	78%
Elm	-5	74%
Aspen	+2	81%

In this example, you can see that Aspen is well below the trend line, whereas it was above the trend line in the previous two graphs. This suggests that whatever was driving Aspen's high performance recognition in the Value Table, SGP, and PARCC Score Delta calculations is not reflected in the Growth to Proficiency model. In this case, it is likely that Aspen's average student is still below proficient, meaning that it is not benefitting from students who are already proficient being counted as "making growth." Similarly, whatever was driving Elm to be above the trend line in the PARCC Score Delta and SGP calculations has disappeared here, where they are underperforming peers. However, Maple's relative performance has stayed stable.

Comparing Student Growth Percentile Scores to PARCC Score Deltas



KEY: Maple Elm Aspen

	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

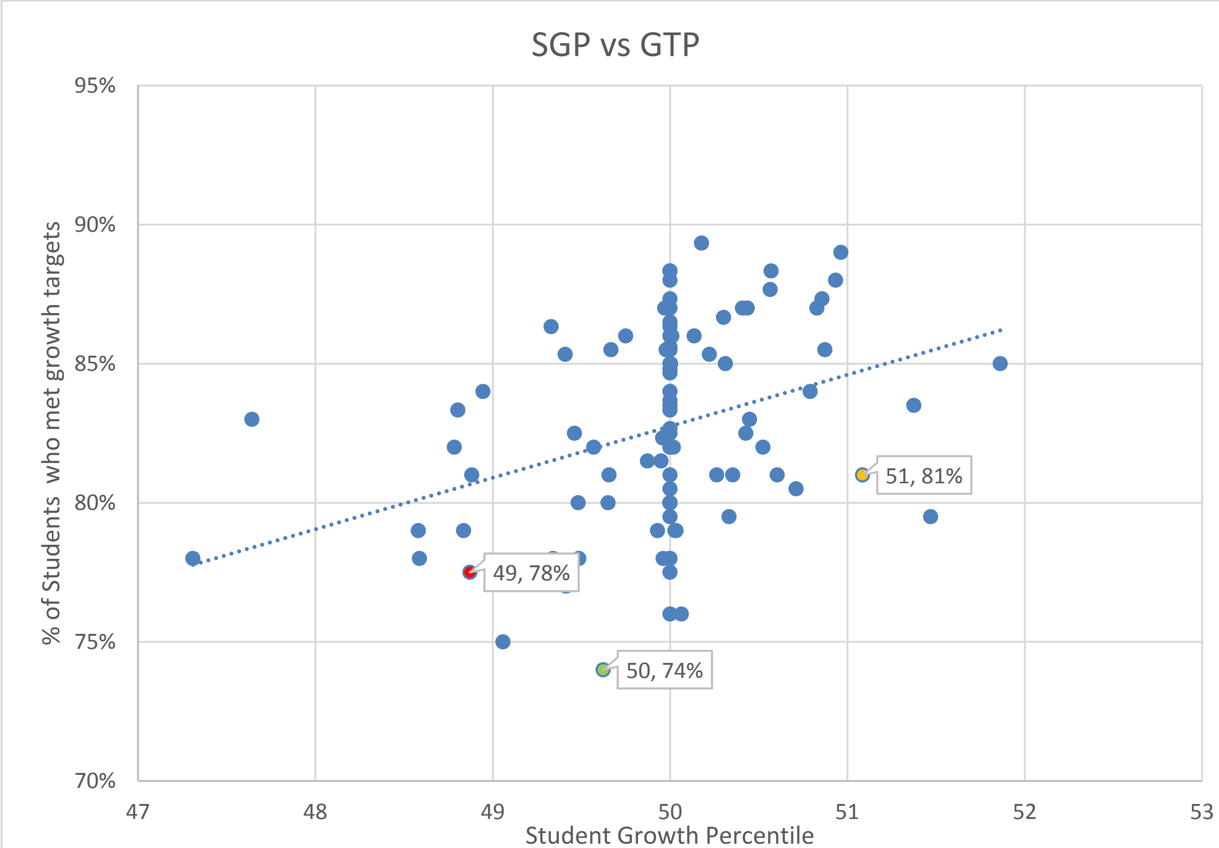
School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

Model School Performance, by calculation type

School	Delta PARCC	SGP
Maple	-4	49
Elm	-2	50
Aspen	+2	51

In this example, you can see that Elm is average in performance at 50%, which is a significant departure from its relative performance in all previous graphs. You can also see that Maple is receiving a marginally lower SGP score than its PARCC Score Delta peers. This suggests that the SGP scores reflect a different growth phenomenon than simply improving or declining proficiency scores.

Comparing Student Growth Percentile Scores to Growth to Proficiency Scores



KEY: Maple Elm Aspen

	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

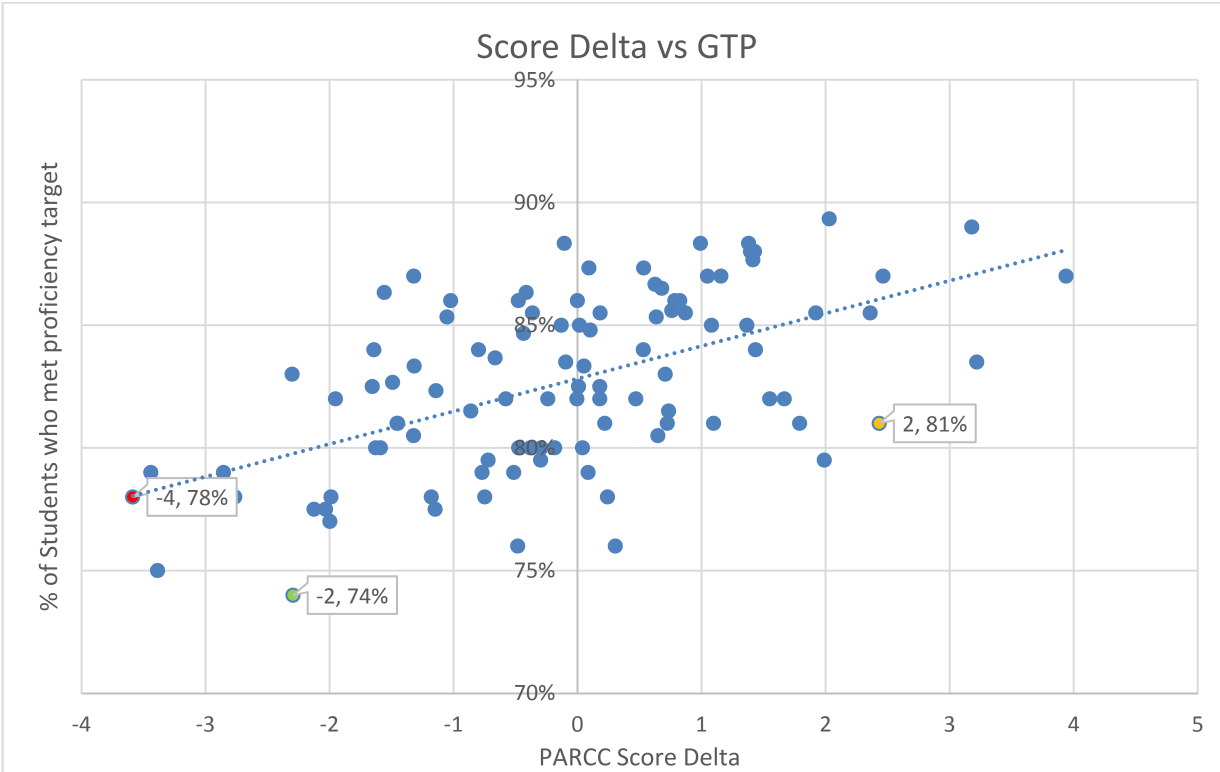
School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

Model School Performance, by calculation type

School	SGP	GTP
Maple	49	78%
Elm	50	74%
Aspen	51	81%

In this example, you can see that all three schools are well below the trend line, despite sitting on either side of it in previous graphs. This suggests that SGP and GTP are both capturing different growth phenomenon than PARCC Score Deltas and Value Tables.

Comparing PARCC Score Deltas to Growth to Proficiency Scores



	School A – Maple School	School B – Elm School	School C – Aspen School
Grades Served	K-8	K-8	K-8
English Learner	5%	11%	6%
Free/Reduced Lunch	26%	16%	74%
Special Ed/IEP	18%	0%	9%

School	2015 PARCC	2016 PARCC	Delta PARCC	Growth
Maple	746	742	-4	Larger decline
Elm	752	750	-2	Small decline
Aspen	743	745	+2	Small gain

Model School Performance, by calculation type

School	Delta PARCC	GTP
Maple	-4	78%
Elm	-2	74%
Aspen	+2	81%

In this example, you can see that the relationship between PARCC Score Deltas and GTP scores is giving Maple a similar placement relative to its peers, whereas both Aspen and Elm are underperforming expectations. This is a significant change from its relative placement when equally weighting PARCC and SGP, where it received a score well below PARCC Score Delta peers. This suggests GTP is positively recognizing Maple in this scenario, as compared to Elm, which moved in the opposite direction relative to its peers.

As our GTP measure rewards students currently at proficiency in the same way it rewards those making growth, a slightly different GTP measure may result in a different graph, entirely. For this reason, it is important to articulate what we want to capture in the growth calculation <u>before</u> settling on a statistical model.