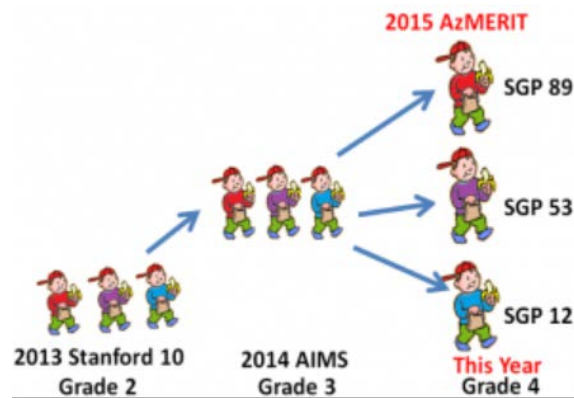


Ways To Calculate Growth or Change

The Accountability Advisory group considered a number of approaches to measuring growth. Below is a brief description of different ways to calculate growth.

Student Growth Percentiles – Median Growth Percentile Approach

Student growth percentiles compare ‘academic peers’ and rank them from the highest scoring to the lowest scoring by assigning percentile ranks called student growth percentiles. Academic peers are students who performed similarly on prior tests. It is not required to be the same test, it just should be a test of similar content. Students can be grouped based on one prior year, but often the student groups are based on as many years data as are available. Once the students are grouped, their performance on the current test determines the growth percentile.



Used to group similar students

Used to rank students

Pros:

- Schools have a history of using SGPs
- It appears easy to understand
- No matter where one falls on performance, one can still have a high SGP if the student does better than similarly scoring students

Cons:

- By definition, half the students are above 50 and half below. This means, for example, that the highest performing peer group will have students with high scale scores and lower growth percentiles.
- At a school or LEA level, the larger the population the more likely that the SGP will be at, or very near, 50.
- You cannot predict a student’s relative standing (SGP) in advance because it depends on how others do
- Instruction beyond what is measured on the test will not contribute to higher SGPs
- If everyone’s scores decline the student with the least decline gets the highest SGP.
- It does not necessarily drive students to higher levels of proficiency.

What behavior it drives:

- All students count, from the highest to the lowest
- Depth of instruction on tested material that exceeds what students typically get at the grade level is encouraged so students can be moved faster than comparable students.
- It allows schools to compare their student growth to that of other similar students in the state. It gives schools/district a measure of student growth that they cannot determine on their own as they don't have state level data.

Student Growth Percentiles (SGPs) – Categorical Weights

One problem with the median SGP is that the SGPs are normally distributed on a bell curve and the larger the school or district the more likely the median will be at or near the apex of the bell curve – the 50th SGP. This results in fewer points than proficiency for many schools. It also does not distinguish the performance of schools very well – a school with all average students would have the same median as a school that was doing fabulously with half the students and very poorly with the other half. Using categories of SGP performance and weighting those categories addresses these problems.

The first step is to categorize the Student Growth Percentiles into groups. The software used to calculate the SGPs provides the following categories, but others could be chosen:

VL= Very Low (SGP 1-19)
L= Low (SGP 20-39)
T= Typical (SGP 40-60)
H= High (SGP 61-80)
VH= Very High (SGP 81-90)

The Board would have to determine how many points to give each of the categories. For example, it could be decided to give the lowest category a weight of zero, average growth a one and very high growth a weight of 1.5. This structure could mirror the weights given to proficiency, so schools would be treated the same way for growth and proficiency (although it would be easier to get more students in the high proficiency category than very high growth because of the norm-referenced nature of growth and the cut point representing the tail of the normal curve).

Pros:

- Schools have a history of using SGPs
- It gives the most incentive for high growth
- The weights can be adjusted to reflect proficiency weights or creating a distribution that reflects the proficiency distribution
- Schools could get all their growth points by having all their students at expected growth.

Cons:

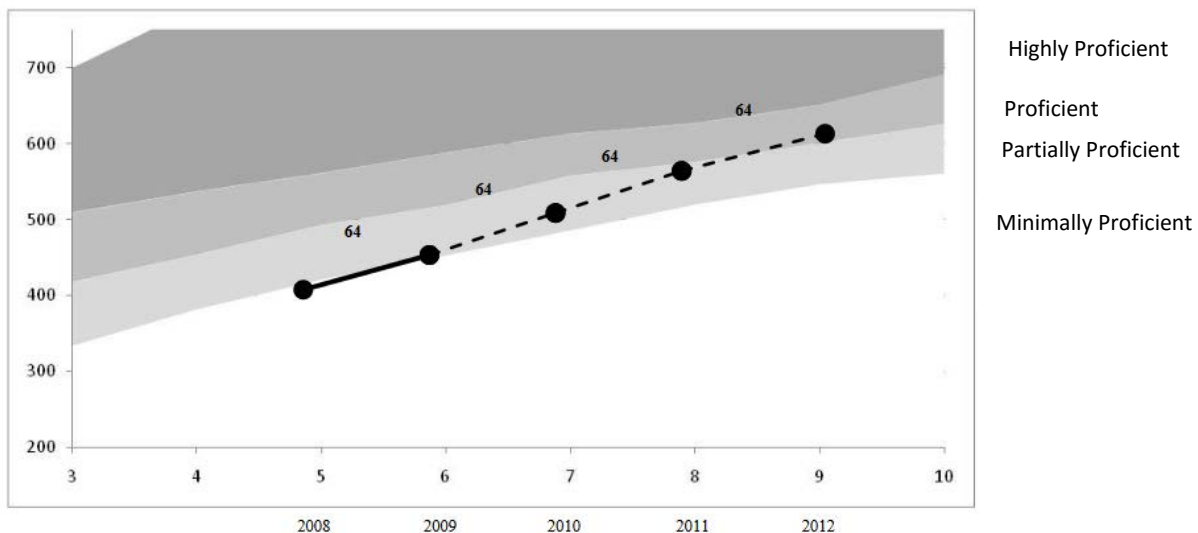
- The schools could get more than 100 points unless the total points were capped at a maximum of 100.
- If everyone's scores decline the student with the least decline gets the highest SGP category.
- It does not necessarily drive students to higher levels of proficiency.
- A school does not know what performance is specifically required to get a higher SGP because it depends on how the comparison group performs.

What behavior it drives:

- The highest students count more, so can compensate for lower growth students.
- Depth of instruction on tested material that exceeds what students typically get at the grade level is encouraged so students can be moved faster than comparable students.
- This calculation focuses schools on students with the highest potential for growth.

Growth to Target

The Growth to Standard or Growth to Target model typically defines a baseline year, such as third grade or the first grade tested. The student then has a growth trajectory calculated so that the student is at proficient within three to four years or by eighth grade. For accountability purposes, the student who hits the target for that grade or who is at proficiency is counted as making adequate growth. In the case of proficient students, a growth to target of highly proficient can be calculated and that progress can be rewarded.



Pros:

- It focuses on getting students to proficient and/or highly proficient.
- Each student has an individually calculated growth trajectory.
- Arizona's tests are vertically scaled so every student is on the same scale from three through high school, which makes tracking students on that scale a logical thing to do.

- You can set a growth trajectory for highly proficient so that almost all students have a goal.
- It is based on scale scores achieved, not relative standing.

Cons:

- Change in individual scale scores is not always, or usually, linear from one year to the next. For example, this year there was a big spike in fifth grade scores. If this is due to a test effect, we give a lot of credit to the 4th to 5th growth, but we can expect a lower increase from fifth to sixth this year because the spike may not have been due to student learning.
- The later a student starts the steeper the expectation may be, which may be unrealistic.
- It may not make sense with high school math because of different content, even if the scores are on the same scale.
- With the large number of partially and minimally proficient students and the higher expectations that rise at each grade, it is not clear that large numbers of students can be moved to proficiency within three years or whatever the time frame will be.

What it drives:

- Moving students to proficient or highly proficient over time.
-

Source: <http://www2.cde.state.co.us/schoolview/documents/CUKUMU.pdf>

<http://dpi.wi.gov/sites/default/files/imce/accountability/pdf/growthwhitepaper.pdf>

Scale Score Change

The early proposal was to measure change in scale score points with an adjustment to eliminate negative change. The recent presentation to the A-F Committee compared scale scores in one year to scale scores in the next year. Every student that increases by one or more counts as growing or making gain.

Pros:

- It will advantage most of our schools because by design scale scores should increase from grade to grade on the vertical scale unless the increase in scale scores from one grade to the next is taken into account.
- It is somewhat simple to understand because one can see scale score to scale score change. However, the addition of an adjustment masks this transparency somewhat because it is not the difference between scale score time 1 and scale score time 2.

Cons:

- Scale score change is typically not consistent within a test. Students with low performance typically have a much higher scale score gain than top performing students. In schools with a large number of students at either of these extremes the gain calculation will be skewed as a result of this psychometric characteristic and not true student performance. (This is true of

AIMS and other tests, statewide AzMERIT data with scale scores and number correct are not available for analysis to verify this is also true for AzMERIT.)

- Scale score gains are not consistent across tests. The average scale score gains in Mesa ranged from +24.6 in fifth grade ELA to -5.42 in eleventh grade ELA and 33.7 in fifth grade math to 7.4 in tenth grade math. A one scale score increase will not mean the same thing for each test.
- In practice, giving schools credit for a one scale score point improvement will err in identifying schools as having students that are improving when in fact not that many are truly improving. In Mesa 70% of the students had a one point or greater scale score increase in ELA and 83% had a one point gain or greater in math. Despite this, many are not improving in proficiency category ratings or on a trajectory to get to proficient in the immediate future.
- The tests are vertically scaled, so that every test within a subject are on the same scale with the scale scores increasing as you go up in grade. In the chart below, for example, if you got the lowest scale score each year you would show a significant gain in scale scores but still continue to be the lowest performing student. This pattern is also seen in math and the 2016 scale score ranges.

AzMERIT Scale Score Performance Level Table Spring 2015

Test Grade Level	ELA	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
3		2395-2496	2497-2508	2509-2540	2541-2605
4		2400-2509	2510-2522	2523-2558	2559-2610
5		2419-2519	2520-2542	2543-2577	2578-2629
6		2431-2531	2532-2552	2553-2596	2597-2641
7		2438-2542	2543-2560	2561-2599	2600-2648
8		2448-2550	2551-2571	2572-2603	2604-2658
9		2454-2554	2555-2576	2577-2605	2606-2664
10		2458-2566	2567-2580	2581-2605	2606-2668
11		2465-2568	2569-2584	2585-2607	2608-2675

- Schools are not incentivized to move students into higher performance categories since a minimal scale score increase will result in credit for growth.

What behaviors it drives:

- If a significant majority count as making growth without special intervention, the impact of this measure on school actions will be minimized.
- To the extent that scale scores automatically increase, the students needed to be targeted are just those who are declining in performance.

Performance Level Change

Another measure is change in proficiency category – minimally proficient to partially, partially proficient to proficient, proficient to highly proficient. This is a very gross measure of change, but one that is clear and obvious.



Pros:

- Change in performance level is easy to understand
- It can reward change at all levels – minimal to partial, proficient to highly -- not just getting to proficient.
- Arizona’s proficiency rate is low, so increasing proficiency levels is something that is a long-term goal.

Cons:

- The number of points/items needed to increase one performance level, or even a half level, is not consistent across levels. For example, at third grade ELA it appears AzMERIT has 77 scale score points in minimally proficient, 10 in partially proficient, 31 points in proficient and 33 in highly proficient (2016 assessment).
- If the movement is across proficiency levels, it will not reward large improvement within levels.

What behavior it drives:

- All changes in performance level count, so it rewards focusing on students on the bubble at each of the performance level change points. Students far from changing levels will not likely result in showing growth (although they could fall backward which may detract from the calculation if it is structured in that way).

Scale Score Band Change

Another approach that has been described is a variation of the two methods described above. The new ‘Florida model’ rewards movement across scale score bands, with more reward at the lower end of the scale than at the high end.



Rewards Movement Within Bands of Scale Scores As Well As Across Categories

Pros:

- Change in performance is described in scale score expectations
- Comparison of progress, year over year of a student against themselves.

- Scale score band change is easier to understand than a complex SGP/SGT calculation.
- Does not use quantile regression so it is not dividing the state data set in half, deciles, quintiles, quantiles, etc.
- It can reward change at all levels and growth within levels.

Cons:

- Creating bands is somewhat arbitrary and what students need to do to improve is less transparent than simple scale score change to teachers.
- Like SGP movement that is not statistically significant could be rewarded.
- Covariates that are not accounted for as explanatory reasons for current year performance.
- Regression model is not capturing growth relative to peer groups or relative to growth performance based on a set of covariates.

What behavior it drives:

- Change in performance for all students has the potential to be rewarded depending on the width of the bands.