A Guide to Understanding and Selecting Measures of Growth for Smarter Balanced Members

Prepared by Joseph Martineau
Senior Associate, Center for Assessment

May 22, 2016

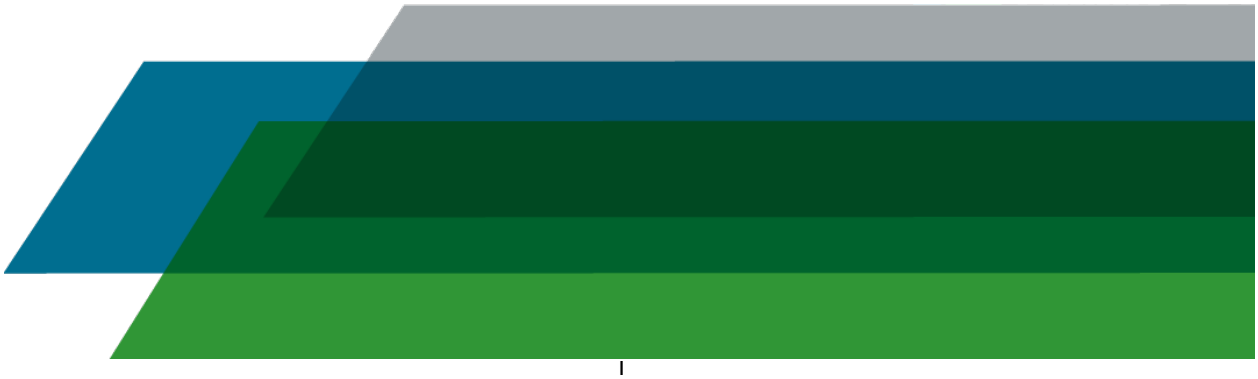

Table of Contents

A Guide to Understanding and Selecting Measures of Growth for Smarter Balanced Members

Smarter Balanced has performed many services on behalf of its members, one of which is the development of a computer adaptive test (CAT) that while constrained to meet blueprint requirements first, improves on the precision of student scores, particularly on the extremes. This is a considerable advantage in measuring student growth in that (1) floor and ceiling effects can bias measures of growth for very high and very low achieving students, and (2) the typically much higher measurement error on the extremes in fixed forms tests result in highly unreliable measures of growth for very high and very low achieving students.

Another service provided by the consortium is the development of a vertical scale. While there are scholarly disagreements about the use and qualities of vertical scales, the development of a vertical scale opens up many more options to Smarter Balanced members in measuring growth. In particular, a vertical scale makes available simple measures of growth that do not require highly complex statistical models to compute.

This white paper provides Smarter Balanced member states with guidance on the following:

- Important background steps in selecting one or more growth measures from the wide variety of available measures
- Understanding the various classes of interpretations that can be made from growth models
- Understanding the various analytical models that can produce growth scores
- Selecting from the available growth measures (e.g., the complete set of types of interpretations available from each type of analytical model) by reviewing important considerations for states to deliberate upon in selecting one or more growth measure appropriate to the states' intended interpretations of growth.

Important Background Steps

The Importance of a Theory of Action

States have implemented a variety of measures of student growth on their state assessments for a variety of reasons. The wide variety of growth measures available to states has a large array of characteristics, which make them appropriate for different purposes. Depending on the purposes a state has for a growth measure, a state may need to use more than one measure of growth. Successfully selecting and implementing one or more measures of growth depends on a sound theory of action detailing how measures of growth fit into an overall educational policy framework. Shakman and Rodriguez (2015) developed a toolkit for creating such a theory of action (or logical framework).

Identifying Intended Uses of Growth Measures

In developing a theory of action, a variety of intended uses of growth measures may be identified, as there are many uses to which growth measures can be put, both at the individual student and aggregate (e.g., for a classroom, school, district, or state) levels. Identifying the intended uses is an important step in defining the intended interpretations of growth scores. The most common uses of growth measures include the following:

- **Reporting**: including but not limited to presentation on parent reports, individual student reports, class roster reports, class summary reports, school summary reports, district summary reports, or state summary reports
- **Data files**: including but not limited to inclusion in individual (student-level) or aggregate (e.g., classroom, school, district, or state) data files
- **Accountability**: including, but not limited to use of individual growth scores to quality a student as proficient even if her status measure did not indicate proficiency, and inclusion of aggregate growth scores in accountability systems
- **Educator evaluation**: including, but not limited to use of individual growth scores to qualify a student as proficient even if her status measure did not indicate proficiency, and inclusion of aggregate growth measures in an educator's evaluation
- **Program evaluation**: including the use of growth scores to evaluate the effectiveness of a curriculum, instructional program, or intervention
- **School improvement**: including the use of growth scores to identify school improvement needs, to improve existing school improvement plans, and to evaluate the effectiveness of school improvement plan implementation
- **Policy development**: including the use of growth scores to identify policy needs, to improve existing policy, and to evaluate the effects of existing policy

Identifying Intended Interpretations of Growth Measures

Based on the theory of action and an understanding of the intended uses of growth measures, the intended interpretations of measures of growth can then be defined. It is likely to become clear that there are multiple intended interpretations (which may require multiple measures of growth). Without explicitly matching the selection of one of more growth measures to intended interpretations, it is unlikely that intended policy goals intended to be served by measures of growth will be realized.

This white paper describes various potential interpretations of growth measures as well as various types of analytical models used to produce growth measures, drawing in large part from the excellent work of Castellano & Ho (2013). This white paper recasts the classifications provided by Castellano and Ho in terms of intended interpretation first, since the match of the intended interpretation to the theory of action is the most important step. It also addresses other characteristics of growth models on the basis of intended interpretation and analytical model used to create the growth measures. In addition to the intended interpretations, certain characteristics of the various available growth measures may better fit with a theory of action than others.

Defining Growth

The typical psychometric definition of growth is "a simple or estimated difference on the same score scale from one point in time to another." Measures of growth matching this simple typical psychometric definition can be problematic for many reasons, so the psychometric community has developed additional measures of "growth" that serve as proxies for this simple definition. However, it is not clear that it is necessary to treat these simply as proxies for growth (or as "growth"). Two relevant dictionary definitions of growth (adapted slightly from those available at dictionary.com) are:

- An increase (or change) in magnitude
- Progression (or regression) from one stage to another

The typical psychometric definition of growth fits in the first definition. Measures typically labeled as proxies for growth generally fit into the second definition. The first definition requires measurement of status to be on the same scale over time. The second definition allows (but does not require) measurement of status to be on different scales at each measurement occasion.

Potential Interpretations of Growth

Interpretations of growth measures[1] can be classified as either retrospective or prospective, with the following basic definitions:

- *Retrospective*: interpretation is based only on already observed score data
- *Prospective*: interpretations are based on extrapolations from observed score data into the future (sometimes also called *growth-to-standard*)

Each of these two broad categories of interpretation has multiple subcategories, as explained below:

Interpretations Focused on Observed Achievement

There are three types of interpretations of growth that focus on observed achievement: scale-referenced; norm-referenced; and criterion-referenced judgment interpretations; with the following basic definitions:

- *Scale-referenced*: interpreted relative to the achievement scale.
- *Norm-referenced*: interpreted relative to other students/units[2].
- *Criterion-referenced judgment*: interpreted relative to judgmentally defined growth targets (or criteria).

One or more examples of each of these three types of interpretations are given below.

Scale-referenced

In this type of interpretation, the measure of growth is expressed as the number of points the student gained on the score scale. Because the growth measure is expressed in this way, understanding such growth scores requires a strong understanding of the score scale. An example of a scale-referenced retrospective interpretation is begun in Figure 1, where Cruz (a fictional student) scores 2500 on the Smarter Balanced mathematics scale in grade 3, and 2600 on the Smarter Balanced mathematics scale in grade 4 the next year. He gained 100 points on the Smarter Balanced mathematics scale.

[1] There are other interpretations of growth measures than those described here, but they have not been widely used.
[2] Such as classrooms, schools, districts, states, or other units of aggregation in the educational system.

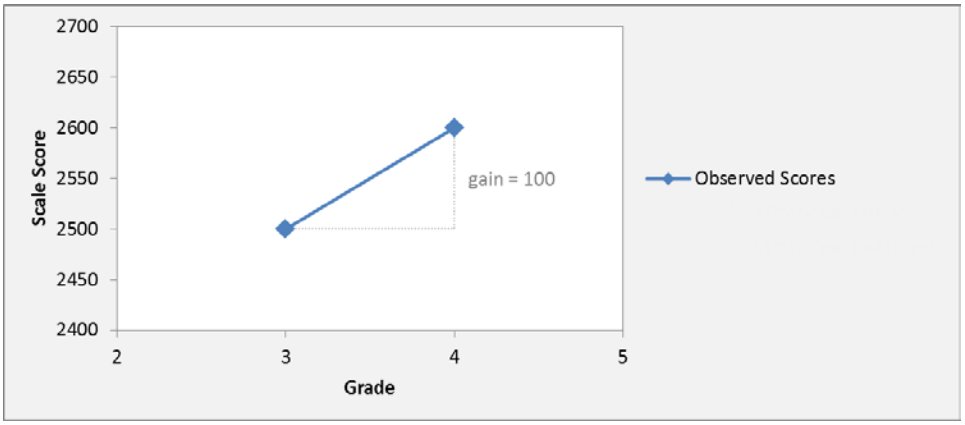

Figure 1. Two scores on the same scale.

Considerations with Scale-referenced Interpretations. Scale-referenced interpretations tend to be the simplest interpretations as long as the meaning of the scale is understood. However, it is important to intimately understand the score scale in order to understand what a gain of 100 means. Figure 2 demonstrates this importance with fictional *level 3* or *proficient* cut scores in the two grades. In this hypothetical case, Cruz scored above the cut score in grade 3 but below in the cut score in grade 4. This additional information about the score scale helps to understand that a gain score of 100 was, in this hypothetical case, not sufficient to keep pace with increasing expectations from grade 3 to grade 4.

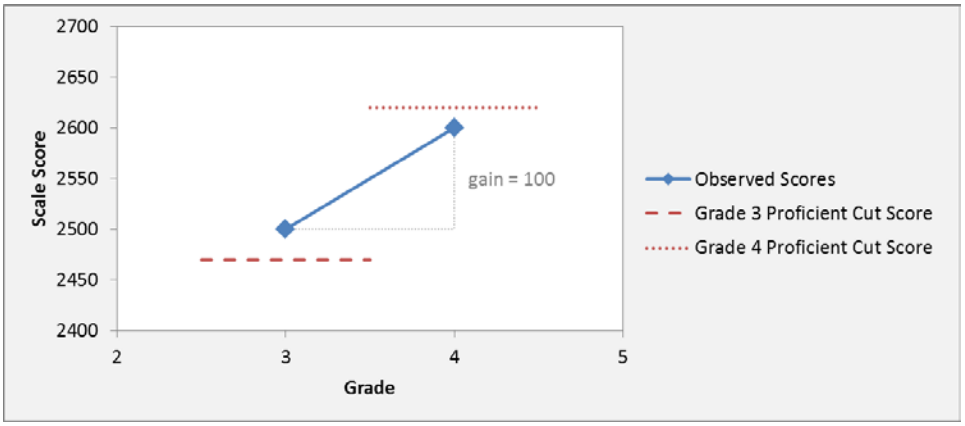

Figure 2. Two scores on the same scale with additional information about the scale.

Norm-referenced

With norm-referenced interpretations, growth is interpreted relative to a specific population. This type of interpretation requires an understanding of relative performing. Norm-referenced interpretations can be made for both individual students and for units[3]. For example:

[3] Classrooms, schools, districts, states, or any other unit of aggregation in the educational system.

- From the prior hypothetical example, Cruz's gain score of 100 points on the Smarter Balanced mathematics scale might be interpreted in a norm-referenced way by indicating that it was *below average*, or *greater than only 30 percent of public school fourth grade students in the state*.
- In another hypothetical example, Alina had a Student Growth Percentile of 78 on the grade 4 Smarter Balanced mathematics test, meaning that she scored better than 78 percent of her academic peers (those with the same score history on the grade 3 Smarter Balanced tests).
- In another hypothetical scenario, the average gain score from grade 7 to 8 of Carver Middle School students on the Smarter Balanced English language arts/literacy (ELA/literacy) tests was 52 points. *This was greater than 57 percent of schools in the state with 8th graders*.

Considerations with Norm-referenced Interpretations. Norm-referenced interpretations tend to be moderately simple, but they do require an understanding of relative standing. They also tend to introduce a zero-sum game in that because scores are reported relative to a population, some students will always have low growth scores, even if an entire state improves over time.

Criterion-referenced Judgment

With criterion-referenced judgment interpretations, growth is interpreted relative to one or more judgmentally-defined targets (or criteria) for adequate growth. The judgmentally-defined criterion for adequate growth is made by a policymaker or a policymaking body such as a standard setting panel, a board of education, a superintendent, legislators, a governor, or other state education agency official. The adequate growth criterion may be applicable for all students, or only in specific circumstances. For example:

- Cruz did not score at or above proficient, but his gain score of 100 points counts as adequate growth because *state law says that any non-proficient student with a gain score of 75 points or higher can count as proficient* for accountability purposes.
- Alina did not score at or above proficient, and Alina's SGP of 28 on Smarter Balanced grade 4 math test does not count as adequate growth because the *State Board of Education identified an SGP of 60 as adequate growth for students who are not yet proficient*.
- The state's accountability director has identified that any student whose growth score was higher than 40% of student's statewide can count as adequate growth in state accountability system.

Considerations with Criterion-referenced Judgment Interpretations. Criterion-referenced judgment interpretations provide a clear target for growth. However, they tend to be arbitrary. Depending on whether the policy body that developed the targets considered whether the targets they set will support the state's policy goals, the arbitrariness may result in incoherence with the state's policy goals.

Interpretations Focused on Future Achievement (Growth-to-Standard Interpretations)

There are two classes of interpretations of growth focused on future achievement: trajectories and projections, each of which have two sub-types. They have the following basic definitions:

- *Criterion-referenced Trajectory*: based on a hypothetical trajectory of scores into the future.
    - *Criterion-referenced trajectory continuation*: If the recent trajectory of test scores were continued into the future for a specified number of years, the student would/would not meet a target achievement level at the end of that time.

  - *Criterion-referenced trajectory target*: The trajectory of future scores that would be needed for the student to meet a target achievement level after a specified number of years.
- *Criterion-referenced Projection*: based on prediction of students future scores, using students' prior scores and the relationship among scores in lower grades and higher grades from one or more prior cohorts of students.
  - *Criterion-referenced projected category*: whether the student is predicted to score in a desirable performance level at some specified number of years in the future.
  - *Criterion-referenced projected probability*: the predicted probability that the student will score in a desirable performance level at some specified number of years in the future.

One or more examples of each of these four types of interpretations are given below.

Criterion-referenced Trajectory Continuation

In this type of interpretation, the measure of growth is expressed as whether the student would become proficient at some specified point in the future if her past trajectory of test scores is continued until that point. Figure 3 begins a hypothetical example with four important pieces of information:

- Cruz scored 2500 in grade 3
- He scored 2600 in grade 4
- His gain score from grade 3 to 4 was 100 points (a trajectory of a 100 point gain per year)
- The proficiency cut score in grade 7 is 3000

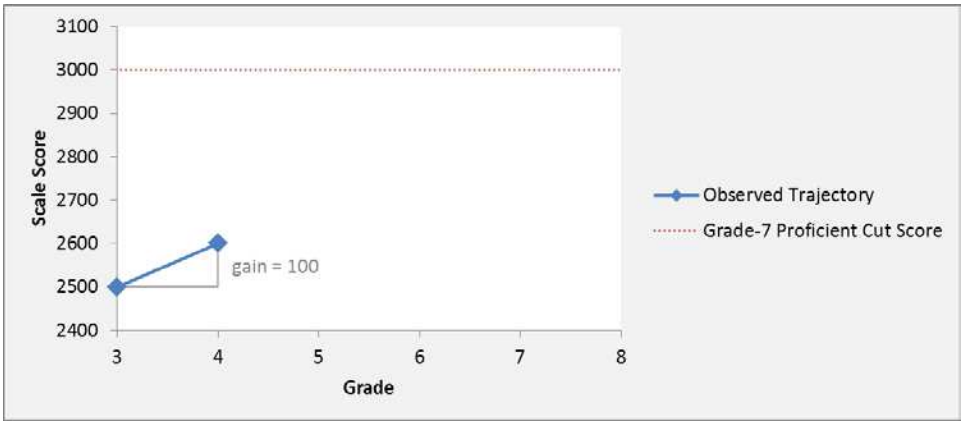

Figure 3. Observable portion of a criterion-referenced trajectory continuation interpretation using gain scores.

If Cruz's trajectory is continued through 7th grade, his scores would look like Figure 4, in which the same gain of 100 points is continued each year. In this scenario, his score in grade 7 would be 2900, 100 points shy of the *proficient* cut score. Therefore, based on continuing his trajectory until the end of 7th grade, his observed gain score of 100 points does not constitute adequate growth to become proficient by the end of grade 7.

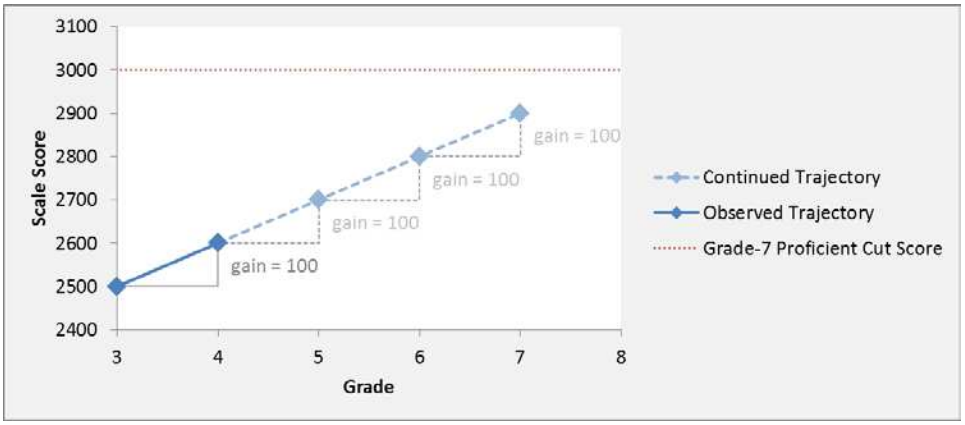

Figure 4. A criterion-referenced trajectory continuation interpretation using gain scores.

Another example of a trajectory-based continuation interpretation is provided below. In this case, it is done using student growth percentiles (SGPs) rather than gain scores beginning in Figure 5.

In this figure, there are again four important pieces of information:

- Alina scored 2500 in grade 3
- She scored 2680 in grade 4
- Her SGP was 65, meaning that her grade 4 score was higher than 65 percent of students with the same scores in grade 3. This gives her an observed trajectory of an SGP of 65 (her gain score is 180 points, but the interpretation of her growth is based on her SGP).
- The proficiency cut score in grade 7 is 3000

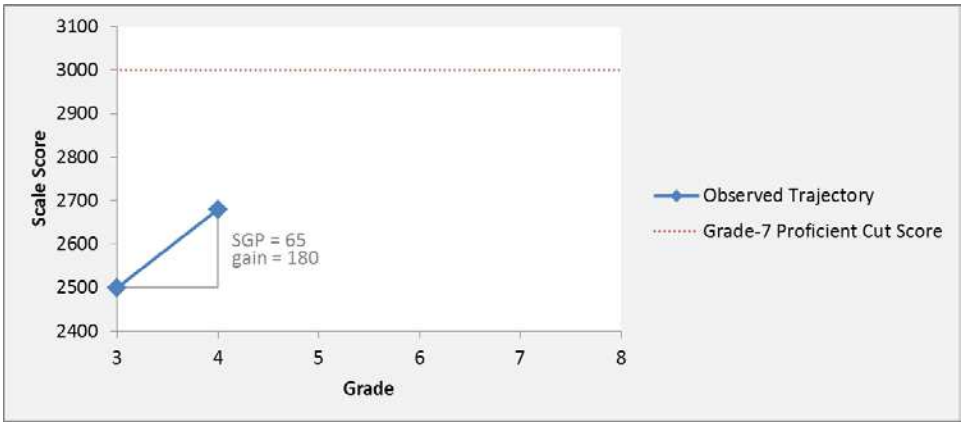

Figure 5. Observable portion of a criterion-referenced trajectory continuation interpretation using SGPs.

If Alina continues her trajectory by achieving an SGP of 65 until the end of 7th grade, her set of scores would look like Figure 6. Note that in each grade, an SGP of 65 could mean a different gain score because SGPs are norm-referenced (compared to other students) rather than scale-referenced (compared to the score scale). This can be helpful in that typical growth tends to be lower in higher grades than in lower grades. In this hypothetical case, by continuing her SGP of 65, Alina would become proficient by the end of grade 7, meaning that her observed SGP of 65 is adequate to become proficient by the end of grade 7.

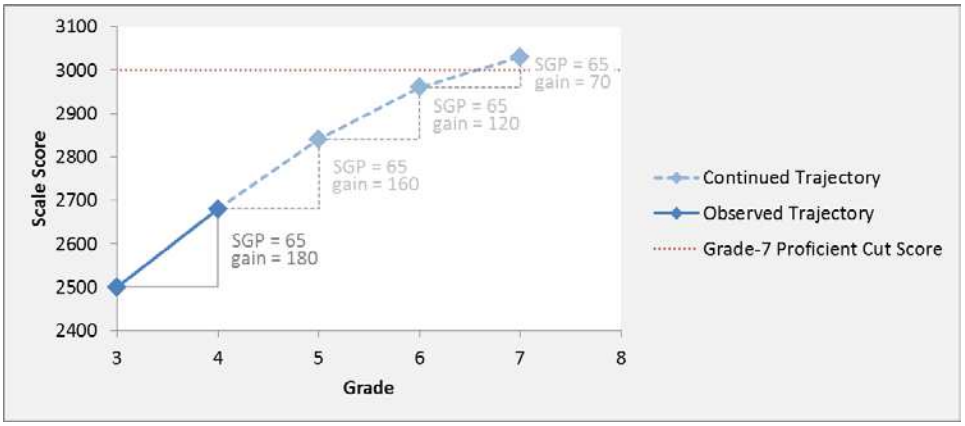

Figure 6. A criterion-referenced trajectory continuation interpretation using SGPs.

Considerations with Criterion-referenced Trajectory Continuation Interpretations. These interpretations have the benefit of evaluating observed growth against a future target level of achievement (meaning that they encourage closing achievement gaps). However, they do not evaluate how likely it is that students will achieve that future target level of achievement because they simply assume that past growth will continue into the future.

Criterion-referenced Trajectory Target

In this type of interpretation, based on observed achievement scores, a target is set for the growth needed to meet a desirable level of achievement at a defined point in the future. Future growth until the defined point in the future is evaluated against the target growth. Figure 7 shows a hypothetical example of a trajectory-based target interpretation of growth. In this example, Cruz scored 2500 in grade 3 and 2600 in grade 4, for a gain score of 100 points. However, with a cut score of 3000 in grade 7, he needs to gain 400 points in three years to become proficient in grade 7. That means he has a target gain score of 134 points per year, or a target achievement score of 2734 in grade 5, 2868 in grade 6, and 3000 in grade 7. His future scores will be compared to the target scores to determine if his growth is adequate.

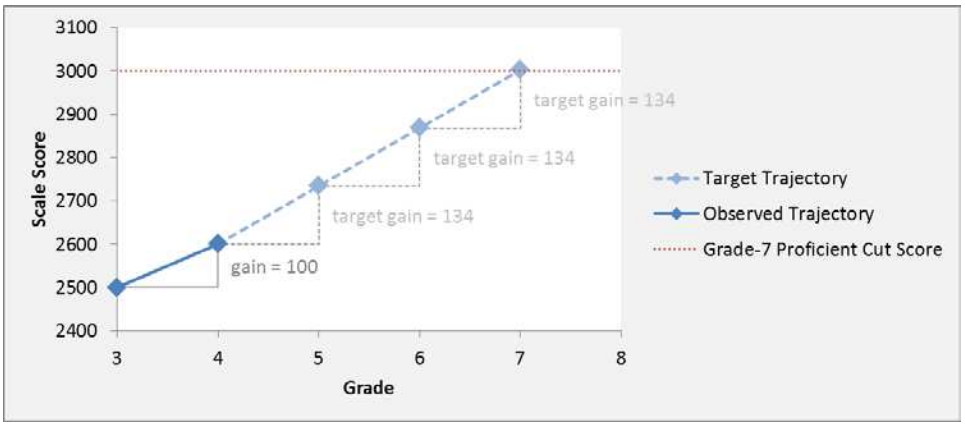

Figure 7. A criterion-referenced trajectory target interpretation using gain scores.

A similar example based on SGPs is presented in figure 8. In this hypothetical situation, Alina scores 2500 in grade 3, and 2680 in grade 4, for an SGP of 65. However, to achieve the cut score of 3000 in grade 7, she needs to achieve an SGP of only 50 through 7th grade. Therefore, her adequate growth percentile (AGP) is 50. Her next year's observed SGP is evaluated against her AGP. Each year a new AGP for achieving the grade 7 proficient cut score is calculated, against which her next year's observed SGP is evaluated.

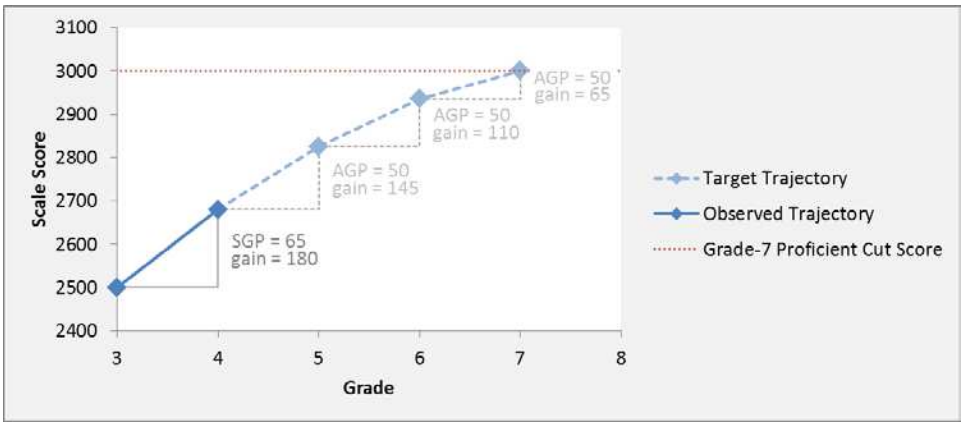

Figure 8. A criterion-referenced trajectory target interpretation using SGPs.

Considerations with Criterion-referenced Trajectory Target Interpretations. These interpretations have the benefit of setting a target for future growth that will lead to achieving a desirable future level of achievement (meaning that they encourage closing achievement gaps). However, they do not evaluate how likely it is that students will achieve that future target level of achievement because they simply create targets for future growth without regard to how likely they may be.

Criterion-referenced Projected Category

In this interpretation, existing data from previous cohorts are used to establish statistical relationships between scores in lower grades and scores in higher grades. Using those established statistical relationships, for example, the most likely score for a student in grade 8 can be predicted from her scores in grades 3-5. A hypothetical example begins in Figure 9. In Figure 9, Zaina's grade 3, 4, and 5 scores are presented. In grades 3-5 she had scores in both ELA/literacy and mathematics. In grade 4 she also had a score in science.

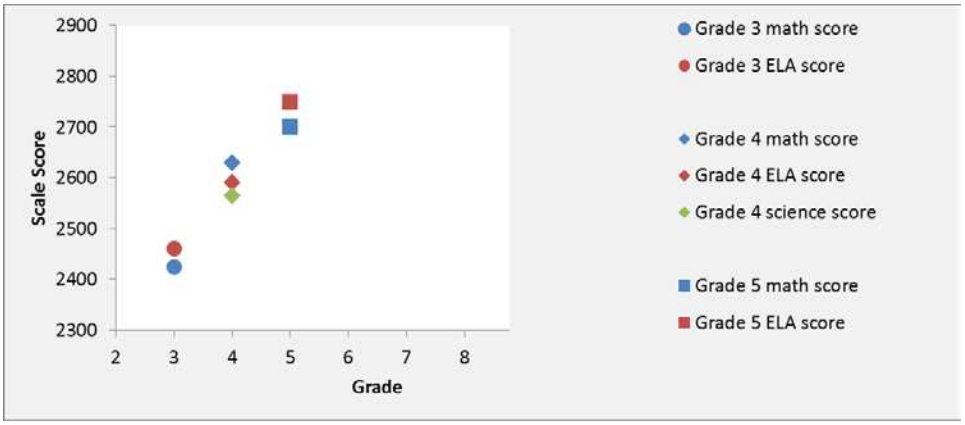

Figure 9. Observable portion of a criterion-referenced projected category interpretation.

Based on previous cohorts' data and the statistical relationships between scores in grades 3-5 and scores in grade 8, it is possible to predict her math score in grade 8, as shown in Figure 10. That predicted score is compared to the cut scores in grade 8 as shown in Figure 11, to identify the projected achievement level in grade 8. In this case, Zaina's predicted performance level in grade 8 is *proficient*.

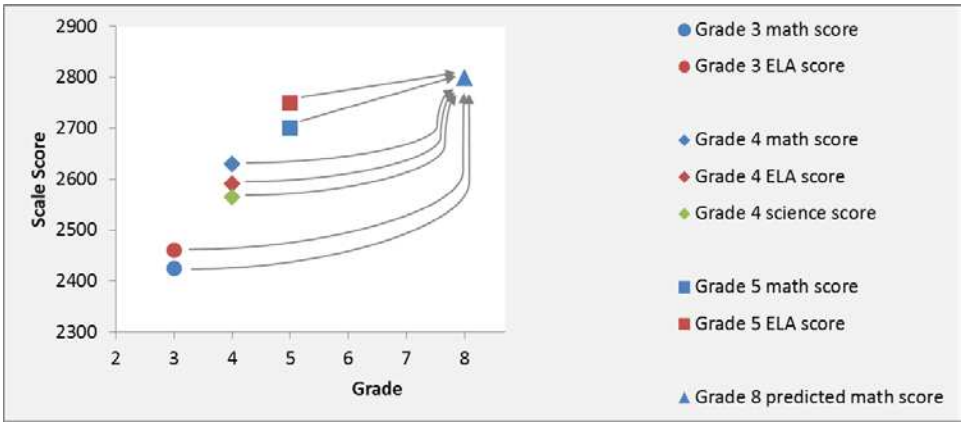

Figure 10. Predicted grade 8 score in a criterion-referenced projected category interpretation.

Figure 11. The predicted achievement category in a criterion-referenced projected category interpretation.

Considerations with Criterion-referenced Projected Category Interpretations. These interpretations have the benefit of evaluating the most likely level of future achievement (meaning that they encourage closing achievement gaps). However, they rely on stability of relationships between early grade scores and later grade scores over many years, meaning that because of scale drift and because of changes in the educational system, predictions may be inaccurate to a considerable degree. In addition, predicting a future achievement level is less stable than predicting the probability of future scores being in a specific achievement level (as shown below).

Criterion-referenced Projected Probability

The only difference between this interpretation and a criterion-referenced projected category interpretation is that rather than identifying what achievement category the predicted future score is in, the probability that the future score will be in each achievement category is instead reported. This is possible because there is error in the predicted score. An example is shown in Figure 12, where the exact predicted score is shown, but the range of possible scores is also projected using the red gradient, with higher probabilities depicted in darker red. In this case, the probability that Zaina's score will be at or

above *proficient* in grade 8 is 88 percent (because the probabilities that her score will be in the *proficient* or *advanced* category is 75 or 13 percent, respectively.)

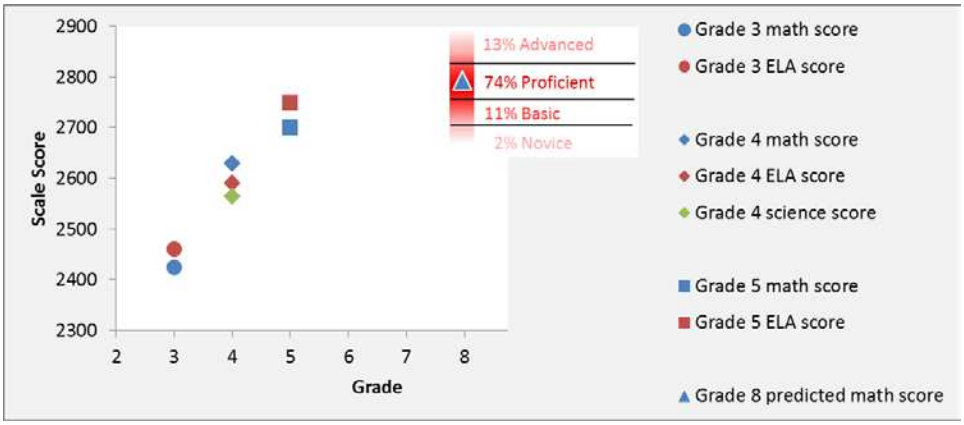

Figure 12. Predicted probabilities of a future score being in each achievement category in a criterion-referenced projected probability interpretation.

Considerations with Criterion-referenced Projected Probability Interpretations. These interpretations improve on the previous type by introducing more stability through predicting probabilities of multiple achievement levels rather than predicting a single achievement level.

Potential Analytical Models for Calculating Growth

There are eleven[4] types of analytical models used reasonably widely to calculate measures of growth. They include the following:

- Gain scores
- Z-score gains
- Baselined z-score gains
- Growth rates
- Student growth percentiles (SGPs)
- Baselined SGPs
- Transitions (e.g., from transition tables or value tables)
- Residual-based models, or value-added models (VAM), including
    - Status-based VAM
    - Gain-based VAM
    - Growth-rate-based VAM
- Future-status prediction models

Each type of model is described briefly below.

[4] Additional types of models for calculating growth are possible, but they have not been widely used.

Gain Score Model

A gain score model is the simplest of the analytical models for calculating measures of growth. It is simply the subtraction of an earlier score from a later score. Figure 13[5] provides a hypothetical example in which a gain score of 100 is observed (from a grade 4 score of 2600 and a grade 3 score of 2500). In order to understand what the gain score of 100 means, some form of additional information about the scale is provided (because the score is scale-referenced).

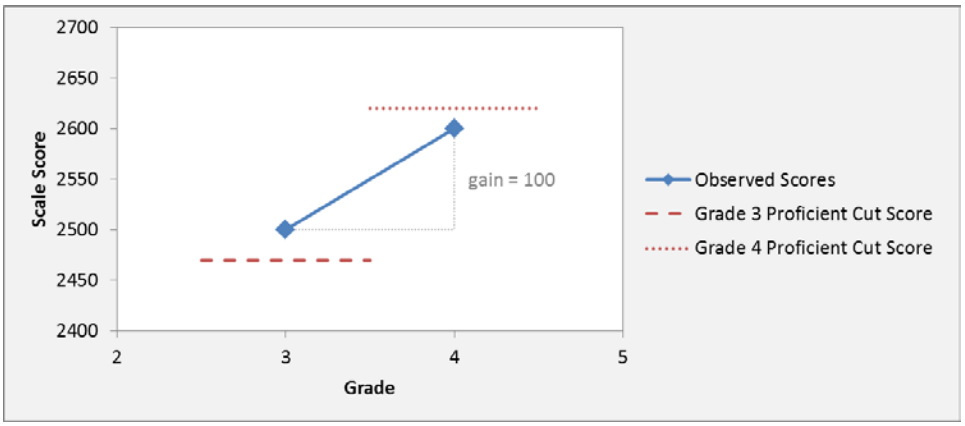

Figure 13. Hypothetical Example of a Gain Score Model.

For example, this gain score of 100 could be better understood by knowing either of the following:

- The grade 3 score was higher than the grade 3 proficient cut score but the grade 4 score was lower than the grade 4 proficient cut score. This indicates that a gain score of 100 points is insufficient to keep pace with increasing expectations from grade 3 to 4.
- A gain score of 100 from grade 3 to 4 is lower than 95 percent of students in the state. This indicates that a gain score of 100 points is a low growth score.

Gain scores can be aggregated across students in a unit (e.g., classroom, school, district, state) in at least the following ways:

- Averaging student gain scores to create a mean (or median) gain score.
- Calculating the percent of student gain scores in each of multiple ranges.
- Calculating the percent of students with gain scores at or above a specific threshold.
- Assigning a subjective value to each gain score (or range of gain scores) [6].

Considerations with Gain Scores

Gain scores require both scores to be on the same scale, meaning that for year-to-year growth, they require a vertical scale like the scale provided by Smarter Balanced. The benefits of a gain score model are that it is the simplest analytical model, stakeholders can calculate their own growth scores rather than

[5] Figure 2, repeated.
[6] Gives policymakers a way to explicitly value different degrees of growth according to their desirability.

needing to rely on experts to do so, and it strongly matches the common understanding of what it means to calculate a growth score. These are considerable advantages of a gain score model.

On the other hand, there are three considerable difficulties with understanding and communicating about gain scores. These are not intuitive, so they need some explanation. First, gain scores tend not to have the same meaning from one grade to the next (in terms of how relatively low or high a gain score is). For example, based on observed student data from five states with vertical scales, gain scores are generally lower for students who initially scored higher. To demonstrate this, the populations of students from the five states were divided into nine groups of equal size (deciles) based on their prior-grade scores. For each decile, the gain score from the prior grade was calculated. The results are displayed in Figure 14 for grade 4 mathematics and grade 8 reading.

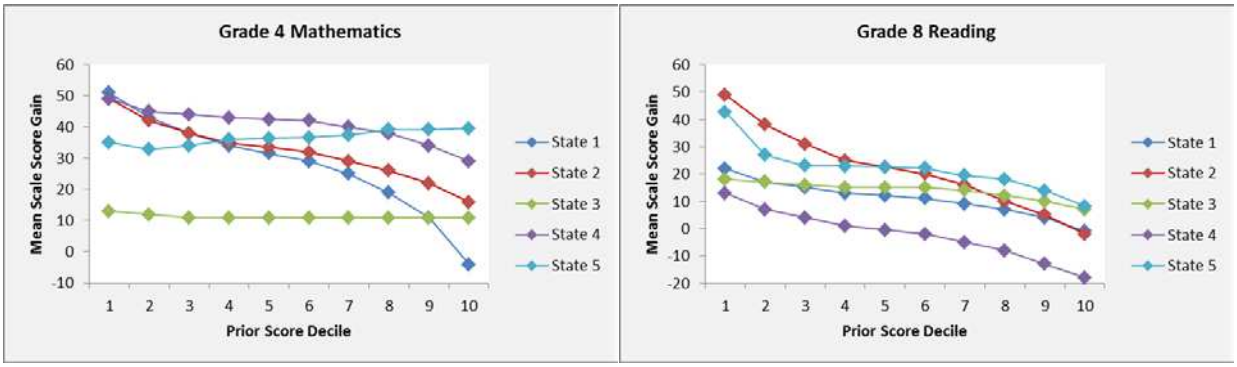

Figure 14. Gain scores by prior score decile for five states in two subject-area/grade combinations[7].

Second, a considerable proportion of gain scores tend to be negative. A few things combine to create this phenomenon:

- Score variance within grades tends to be large relative to average grade-to-grade gains.
- Correlations between pre-test scores and gain scores are negatively biased[8].
- Gain scores are affected by regression to the mean (students with scores further away from the mean, either above or below, are more likely than not to score closer to the mean in the next grade).
- Mean grade-to-grade growth tends to become smaller as grade level increases.
- Score variance tends to increase as grade level increases.

This phenomenon typical to vertical scales means that using a gain score analytical model to calculate growth scores presents a difficulty in interpretation, in that an intuitive interpretation of a negative gain

[7] Note that none of the states' score scales are equivalent. They are placed on the same graphic only for convenience.

[8] The observed gain score (g) is equal to the posttest observed score (x_1) minus the pretest observed score (x_1). In addition, observed scores are true scores (t_1 or t_2) plus measurement error (e_1 or e_2). Therefore, $g = t_2 + e_2 - t_1 - e_1$. Therefore, assuming that measurement errors are uncorrelated with each other and with true scores, $\rho_{x_1,g}$ incorporates the following correlations: ρ_{t_1,t_2-t_1}, $\rho_{e_1,-e_1}$, and ρ_{e_2,e_2}. If the only correlation incorporated into $\rho_{x_1,g}$ were ρ_{t_1,t_2-t_1}, it would be unbiased. However, because it incorporates $\rho_{e_1,-e_1}$ (-1) and ρ_{e_2,e_2} (1) the correlation of pre-test score and gain score is negatively biased and the correlation of post-test score and gain score is positively biased.

score is that a student did not learn over the last year. However, such an interpretation is likely not accurate. For example, Figure 15 shows, for an anonymous state with a vertical scale, the percentage of gain scores in each grade that were negative. The percentage of gain scores that are negative increases with grade level and is much higher in ELA/literacy than in mathematics. It is likely that a similar result will be the case for Smarter Balanced gain scores.

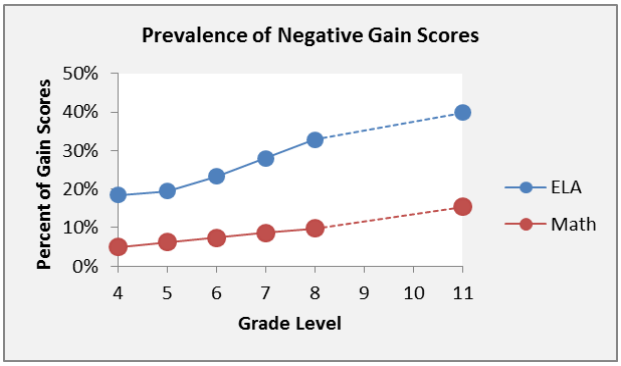

Figure 15. Prevalence of Negative Gains by Subject and Grade for an Anonymous State with a Vertical Scale.

Finally, the content measured on each grade-level test changes qualitatively from grade to grade. This is intentional in that the content taught in each grade builds on the content learned in the previous grade. Sometimes the content changes considerably (for example, when algebraic concepts switch from light to heavy representation in mathematics). Each grade's scores may be identified through various analyses as being unidimensional, but this does not mean that the same construct is being measured across grades. The assessment industry deliberately attempts to create unidimensional (or essentially unidimensional) scores by selecting items with scores that have high correlations with overall test scores. However, this only applies within grades, not across grades. Scores are likely to have a qualitatively different meaning in one grade than in another, complicating the interpretation of gain scores.

Z-score Gain Model

The only difference between a z-score gain model and a gain score model is that the scores are standardized so that the mean of scores in each grade is zero and the standard deviation of scores in each grade is one. The scores may or may not also be normalized to account for differences in the shape of the distributions of scores across grades (e.g., skewness). What this means is that the meaning of the original scale is lost in exchange for making the distributions of scales equivalent across grades or across subjects. An example is given in Figure 16. In this situation, Marta scored a 0.2 in grade 3 and a -0.3 in grade 4, meaning that she scored 2/10 of a standard deviation *above* the statewide average in grade 3, but 3/10 of a standard deviation *below* the statewide average in grade 4. Her z-score gain was -0.5, meaning that her relative standing compared to all other students in the state declined by 1/2 standard deviation.

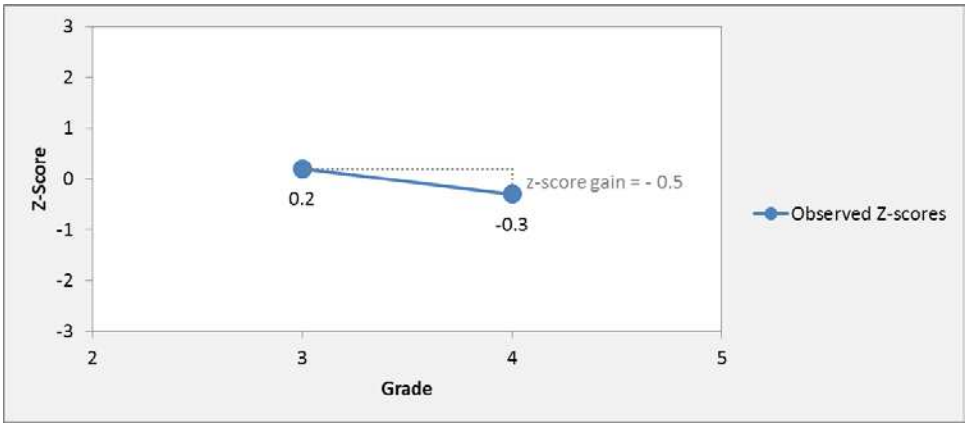

Figure 16. Hypothetical Example of a Z-score Gain Model.

Z-score gains can be aggregated across students in a unit (e.g., classroom, school, district, state) in at least the following ways:

- Averaging student z-score gains to create a mean (or median) z-score gain.
- Calculating the percent of student z-score gains in each of multiple ranges.
- Calculating the percent of students with z-score gains at or above a specific threshold.
- Assigning a subjective value to each z-score gain (or range of z-score gains)[9].

Considerations with Z-score Gains

Z-score gains have the advantage that they allow for comparing growth on different subjects, for aggregating growth across subjects, and for calculating a measure of growth even when a test in a subject area changes. For example, an accountability model may require maintaining growth trends even though a test has changed. Using z-score gains does not make the tests comparable, but it does provide a next-best solution by examining relative standing on the two different measures. This can be a considerable advantage.

However, z-score gains also come with some considerable challenges, such as explaining what a z-score means, a modest level of complexity in calculation, and the same difficulties that come with gain scores.

Baselined Z-score Gain Model

One problem with a z-score gain model is that it introduces a zero-sum game. When scores are expressed as the number of standard deviations above or below the statewide average score, it means that it is not possible for every student to improve his or her relative standing. There will always be some students who improve their standing and some that decline in standing. One way that this issue has been addressed is by setting a baseline for calculating z-scores. This is done by using the mean and standard deviation for each grade level from the baseline year rather than from the current year. By using baseline-year statistics, if a statewide improvement in growth/achievement occurs, it is theoretically possible for every student to improve his or her standing relative to the students from the baseline year. Baselined z-score gains can be aggregated in the same way as z-score gains.

[9] Gives policymakers a way to explicitly value different degrees of growth according to their desirability.

Considerations with Baselined Z-score Gains

The benefit of baselined z-score gains over regular z-score gains is that they eliminate a zero-sum game. The drawbacks of baselined z-score gains are that they add another level of complexity in both calculation and communication (by having to do calculations as if the scores had occurred in the baseline year).

An additional drawback of baselined z-score gains is that they are susceptible to scale drift. Because the scores are always treated as if they had occurred in the baseline year, any drift in the scale will cause baselined z-scores to be inaccurate to the degree that the scale has drifted.

Growth Rate Model

A growth rate model uses more than two scores in the same subject area to estimate a student's growth rate. This addresses some of the problems with gain scores (instability because of measurement error and regression to the mean). An example begins in Figure 17. In this example, Marc has taken the Smarter Balanced ELA/literacy test in grades 3-6. To estimate the student's annual growth rate, a linear regression of scores on grade level is fitted to Marc's four scale scores. The slope of the linear regression is the estimated annual growth rate. In this case, it results in an annual growth rate of 153 points per year.

In addition to interpreting a growth rate in a scale-referenced way (in terms of the number of scale points), a growth rate can also be interpreted in other ways. For example, in Figure 18, a growth rate model is used to make a trajectory-based continuation interpretation. In this example, Marc's growth trajectory is continued through grade 8. Because the continued trajectory crosses the grade 8 proficient cut score before grade 8, his growth rate counts as adequate growth to become proficient by grade 8.

In addition to estimating a linear growth rate, it is also possible to estimate a non-linear growth rate. However, it is important to note that a growth rate model does not improve on a gain score model until there are at least four data points (for a linear growth rate[10]) or more than four data points (for a non-linear growth rate).

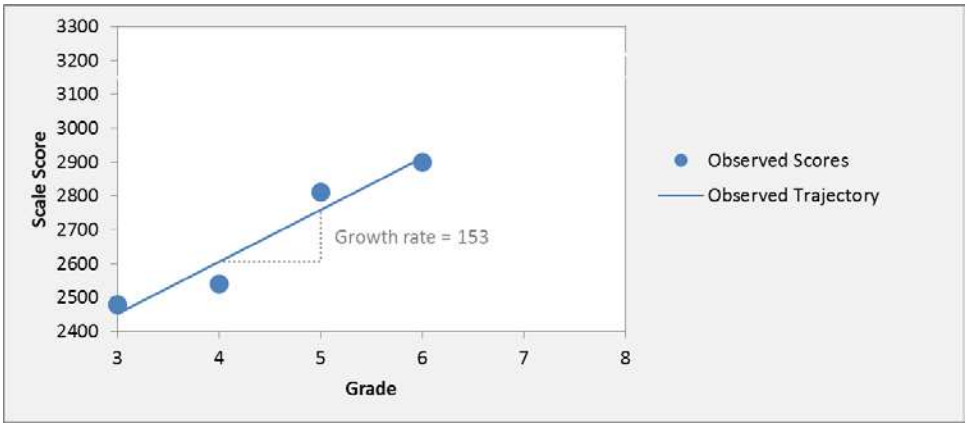

Figure 17. Four Years of Scale Scores.

[10] With only three data points, the estimate of a growth rate depends only on the first and third data points. The intercept of the linear regression depends on the middle data point.

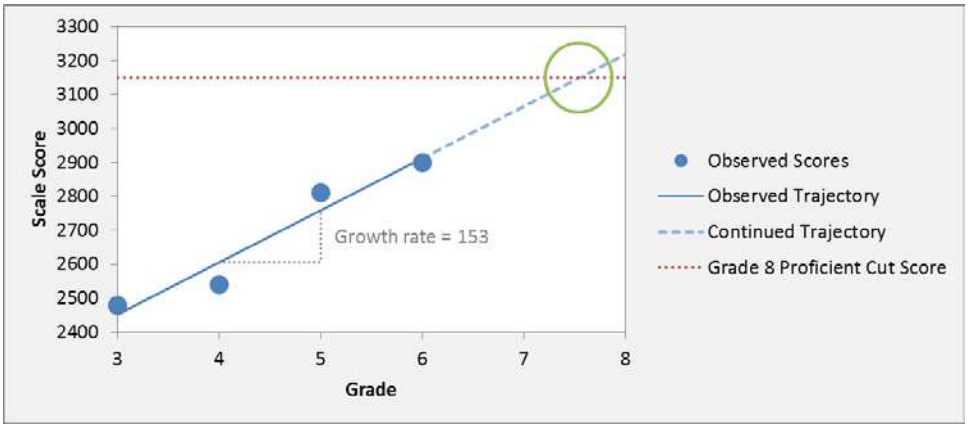

Figure 18. A Growth Rate Model Used to Make a Trajectory-based Continuation Interpretation.

Growth rates can be aggregated across students in a unit (e.g., classroom, school, district, state) in at least the following ways:

- Averaging student growth rates to create a mean (or median) growth rate.
- Calculating the percent of student growth rates in each of multiple ranges.
- Calculating the percent of students with growth rates at or above a specific threshold.
- Assigning a subjective value to each growth rate (or range of growth rates) [11].

Considerations with Growth Rates

Growth rates reduce the instability attendant to gain scores, z-score gains, and baselined z-score gain that comes from using only two scores to calculated a measure of growth. By using at least four scores in the calculation of growth, growth rates accounts for measurement error to some degree. In addition, growth rates can be interpreted as annual rate of gain and are much more consistent with the common understanding of what it means to measure growth.

The drawbacks of baselined z-score gains include that (1) they add complexity in both calculation and communication by employing regression, (2) they assume a linear rate of growth when growth may not be linear, (3) they assume that the meaning of scores do not change from grade to grade, and (4) they require at least four years of data to improve on data obtained from gain scores.

Student Growth Percentile (SGP) Model

An SGP model uses a complex regression technique to estimate the percent of her peers a student outscores, where “peer” is defined as students with the same score history. A conceptual example is given below (the computation is considerably more complex than described in the example, but the

[11] Gives policymakers a way to explicitly value different degrees of growth according to their desirability.

concept is solid). For example, if Katya is in fifth grade and scored 2420 in grade 3 and 2600 in grade 4, her peer group is the set of all fifth grade students in the state who had the same scores in those grades (shown in the left panel of Figure 19). To obtain a grade 5 SGP for Katya, we also need to gather the current (grade 5) scores of her peer group (shown in the left panel of Figure 19). Including Katya, there are 10 students in her peer group.

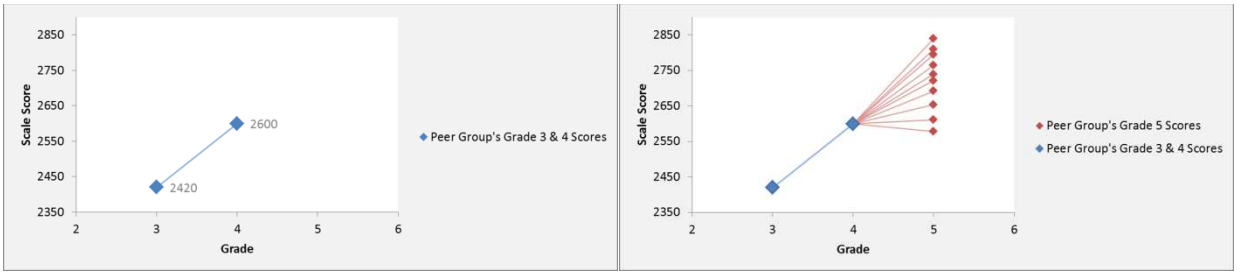

Figure 19. SGP peer group prior test scores and current test score.

Next, we locate Katya's score among her peer group (left panel of Figure 20). We see that she scored higher than 6 of the 10 students in her peer group, or 60% of her peer group. Therefore, her SGP is 60 (right panel of Figure 20).

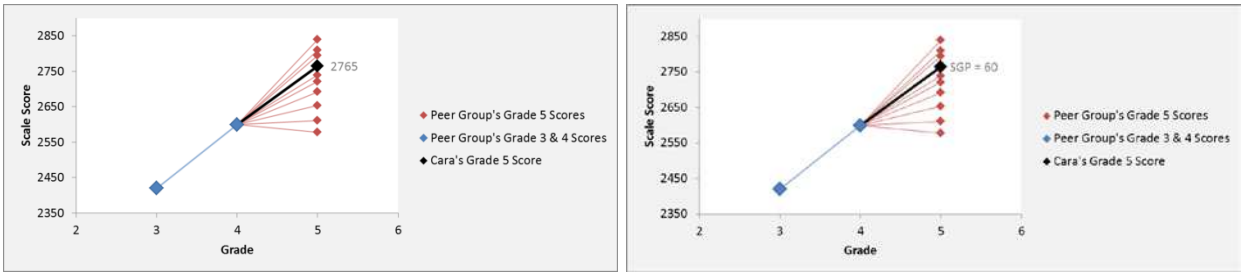

Figure 20. Katya's current score among her peer group and her SGP.

SGPs are inherently norm referenced because they represent a student's relative standing compared to her peers. However, as shown above in Figures 5-8 (and the associated text), SGPs can also be used to make *Trajectory-based Continuation* and *Trajectory-based Target* interpretations.

Growth rates can be aggregated across students in a unit (e.g., classroom, school, district, state) in at least the following ways:

- Averaging SGPs to create a mean (or median) SGP (or MGP).
- Calculating the percent of SGPs in each of multiple ranges.
- Calculating the percent of SGPs at or above a specific threshold.
- Assigning a subjective value to each SGP (or range of SGPs) [12].

[12] Gives policymakers a way to explicitly value different degrees of growth according to their desirability.

Considerations with SGPs

The benefits offered by SGPs include that (1) SGPs are conceptually simple, (2) they can be used when tests change from one year to the next, (3) they can easily be compared and aggregated across content areas, and (4) they can be made highly stable by including many past scores in all content areas to form peer groups.

The drawbacks of SGPs include that (1) estimation is highly complex and (2) they introduce a zero-sum gain because they are inherently norm-referenced. When SGPs are calculated normally every year, there will be approximately one percent of students in each SGP from 0-99. When a target is set for an acceptable SGP, if an entire state improves, it will not be reflected in SGPs.

Baselined SGP Model

Because SGPs are inherently norm referenced, they also set up a zero-sum game in that approximately half of students will get an SGP below 50 and approximately one percent of students will be in each percentile from 0-99 every year, even when statewide improvements are made. In order to address this issue, it is possible to set a baseline by calculating SGPs for the current year as if the scores had been observed in a baseline year. By using baseline-year statistics, if a statewide improvement in growth/achievement occurs, it is theoretically possible for every student to achieve an SGP above 50. Baselined SGPs can be aggregated in the same way as SGPs.

Considerations with Baselined SGPs

The benefit offered by baselined SGPs over SGPs is that it eliminates the zero-sum game. The additional drawback of baselined SGPs over SGPs is that they are susceptible to scale drift. Because the scores are always treated as if they had occurred in the baseline year, any drift in the scale will cause baselined SGPs to be inaccurate to the degree that the scale has drifted.

Transition Model (Transition/Value Table)

Transition models require that cut scores be vertically articulated—as are the Smarter Balanced cut scores—so that tracking changes (or transitions) in students' achievement levels makes sense. A transition model may label each unique type of transition in a descriptive manner or with a value. For example, a transition from *basic* to *proficient* from grade 4 to grade 5 could be labeled an *improvement* (descriptively) or could be valued at 100 points as compared with a value of 50 for a transition from *proficient* to *proficient* for the same set of grades. Transition models often also subdivide achievement levels to capture smaller degrees of growth than might be captured with only a few achievement levels. For example, each achievement level might be divided into three sublevels. A hypothetical example dividing four achievement levels into low (L), mid (H), and high (H) sublevels is given in Figure 21. The left panel shows a basic transition table that shows the grade 3 performance sublevels down the left side and the grade 4 performance sublevels across the top. The cells of the table indicate students' achievement levels in both grades. For example, the right panel shows a hypothetical student (Raymond) who scored in the middle of the *Basic* level in grade 3 and in the middle of the *Proficient* level in grade 4.

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L												
	M												
	H												
Basic	L												
	M												
	H												
Proficient	L												
	M												
	H												
Advanced	L												
	M												
	H												

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L												
	M												
	H												
Basic	L												
	M								X				
	H												
Proficient	L												
	M												
	H												
Advanced	L												
	M												
	H												

Figure 21. Sample Achievement Level Transition Table from Grade 3 to Grade 4.

The various transitions (the cells that students' scores can fall into) can be labeled in two ways, as shown in Figure 22. In the left panel, each transition is labeled descriptively in the following manner:

- Transitions showing a decline by 3+ sublevels are labeled *Significant Decline* (SD).
- Transitions showing a decline by 1-2 sublevels are labeled *Decline* (D).
- Transitions showing the same sublevel in both grades are labeled *Maintaining* (M).
- Transitions showing an improvement by 1-2 sublevels are labeled *Improvement* (I).
- Transitions showing an improvement by 3+ sub-levels are labeled *Significant Improvement* (SI).

In the right panel, numeric values are used as the labels for each possible transition. The use of numeric values allows a state to specify which transitions it values most and which are the most problematic. The values are derived by a careful deliberation about which types of transitions the state desires to reward most.

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L	M	I	I	SI	SI	SI	SI	SI	SI	SI	SI	SI
	M	D	M	I	I	SI	SI	SI	SI	SI	SI	SI	SI
	H	D	D	M	I	I	SI	SI	SI	SI	SI	SI	SI
Basic	L	SD	D	D	M	I	I	SI	SI	SI	SI	SI	SI
	M	SD	SD	D	D	M	I	I	SI	SI	SI	SI	SI
	H	SD	SD	SD	D	D	M	I	I	SI	SI	SI	SI
Proficient	L	SD	SD	SD	SD	D	D	M	I	I	SI	SI	SI
	M	SD	SD	SD	SD	SD	D	D	M	I	I	SI	SI
	H	SD	SD	SD	SD	SD	SD	D	D	M	I	I	SI
Advanced	L	SD	SD	SD	SD	SD	SD	SD	D	D	M	I	I
	M	SD	SD	SD	SD	SD	SD	SD	SD	D	D	M	I
	H	SD	SD	SD	SD	SD	SD	SD	SD	SD	D	D	M

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L	3	5	10	15	20	26	29	30	30	31	31	32
	M	2	4	7	11	15	21	24	25	26	27	28	30
	H	1	3	5	8	12	17	20	21	22	24	26	28
Basic	L	0	3	6	7	10	13	16	18	20	22	24	26
	M	0	1	3	5	7	10	13	15	18	20	22	24
	H	0	1	2	4	6	8	11	14	16	18	20	22
Proficient	L	0	1	2	3	5	7	10	12	14	16	18	20
	M	0	1	2	3	4	6	9	10	12	14	16	18
	H	0	1	2	3	4	5	8	9	10	12	14	16
Advanced	L	0	1	2	3	4	5	7	8	9	10	12	14
	M	0	1	2	3	4	5	6	7	8	9	10	12
	H	0	1	2	3	4	5	6	6	7	8	9	12

Figure 22. Hypothetical Examples of a Transition Model.

Transitions can be aggregated across students in a unit (e.g., classroom, school, district, state) in at least the following ways:

- Calculating the percent of transitions with the same descriptive label (e.g., percentage of transitions in each of the SD, D, M, I, SI categories in the left panel of Figure 22).
- Calculating the percent of transitions deemed acceptable (e.g., the percentage of transitions that were in the M, I, and SI categories in the left panel of Figure 22).
- Averaging the values assigned to each transition (e.g., the values assigned in the right panel of Figure 22)[13].

In this example, Raymond's achievement level transition from grade 3 to grade 4 would be categorized as a *Significant Improvement* (SI) using descriptive labels. With numeric value labels, his transition would be assigned a score of 15. This can be seen in the top two panels of Figure 23. This can be compared to the transition of another hypothetical student (Marta, highlighted in blue), whose achievement level transitioned from *Low Proficient* in grade 3 to *Mid Basic* in grade 4. As seen in the bottom two panels, her transition is categorized as a *Decline* (D) using descriptive labels and as a score of 5 using numeric value labels.

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L	M	I	I	SI	SI	SI	SI	SI	SI	SI	SI	SI
	M	D	M	I	I	SI	SI	SI	SI	SI	SI	SI	SI
	H	D	D	M	I	I	SI	SI	SI	SI	SI	SI	SI
Basic	L	SD	D	D	M	I	I	SI	SI	SI	SI	SI	SI
	M	SD	SD	D	D	M	I	I	SI	SI	SI	SI	SI
	H	SD	SD	SD	D	D	M	I	I	SI	SI	SI	SI
Proficient	L	SD	SD	SD	SD	D	D	M	I	I	SI	SI	SI
	M	SD	SD	SD	SD	SD	D	D	M	I	I	SI	SI
	H	SD	SD	SD	SD	SD	SD	D	D	M	I	I	SI
Advanced	L	SD	SD	SD	SD	SD	SD	SD	D	D	M	I	I
	M	SD	SD	SD	SD	SD	SD	SD	SD	D	D	M	I
	H	SD	SD	SD	SD	SD	SD	SD	SD	SD	D	D	M

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L	3	5	10	15	20	26	29	30	30	31	31	32
	M	2	4	7	11	15	21	24	25	26	27	28	30
	H	1	3	5	8	12	17	20	21	22	24	26	28
Basic	L	0	3	6	7	10	13	16	18	20	22	24	26
	M	0	1	3	5	7	10	13	15	18	20	22	24
	H	0	1	2	4	6	8	11	14	16	18	20	22
Proficient	L	0	1	2	3	5	7	10	12	14	16	18	20
	M	0	1	2	3	4	6	9	10	12	14	16	18
	H	0	1	2	3	4	5	8	9	10	12	14	16
Advanced	L	0	1	2	3	4	5	7	8	9	10	12	14
	M	0	1	2	3	4	5	6	7	8	9	10	12
	H	0	1	2	3	4	5	6	6	7	8	9	12

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L	M	I	I	SI	SI	SI	SI	SI	SI	SI	SI	SI
	M	D	M	I	I	SI	SI	SI	SI	SI	SI	SI	SI
	H	D	D	M	I	I	SI	SI	SI	SI	SI	SI	SI
Basic	L	SD	D	D	M	I	I	SI	SI	SI	SI	SI	SI
	M	SD	SD	D	D	M	I	I	SI	SI	SI	SI	SI
	H	SD	SD	SD	D	D	M	I	I	SI	SI	SI	SI
Proficient	L	SD	SD	SD	SD	D	D	M	I	I	SI	SI	SI
	M	SD	SD	SD	SD	SD	D	D	M	I	I	SI	SI
	H	SD	SD	SD	SD	SD	SD	D	D	M	I	I	SI
Advanced	L	SD	SD	SD	SD	SD	SD	SD	D	D	M	I	I
	M	SD	SD	SD	SD	SD	SD	SD	SD	D	D	M	I
	H	SD	SD	SD	SD	SD	SD	SD	SD	SD	D	D	M

Grade 3 Achievement Level		Grade 4 Achievement Level											
		Novice			Basic			Proficient			Advanced		
		L	M	H	L	M	H	L	M	H	L	M	H
Novice	L	3	5	10	15	20	26	29	30	30	31	31	32
	M	2	4	7	11	15	21	24	25	26	27	28	30
	H	1	3	5	8	12	17	20	21	22	24	26	28
Basic	L	0	3	6	7	10	13	16	18	20	22	24	26
	M	0	1	3	5	7	10	13	15	18	20	22	24
	H	0	1	2	4	6	8	11	14	16	18	20	22
Proficient	L	0	1	2	3	5	7	10	12	14	16	18	20
	M	0	1	2	3	4	6	9	10	12	14	16	18
	H	0	1	2	3	4	5	8	9	10	12	14	16
Advanced	L	0	1	2	3	4	5	7	8	9	10	12	14
	M	0	1	2	3	4	5	6	7	8	9	10	12
	H	0	1	2	3	4	5	6	6	7	8	9	12

Figure 23. Student transition labels in transition models.

[13] Gives policymakers a way to explicitly value different degrees of growth according to their desirability.

Considerations with Transitions

The benefits offered by transitions include that (1) they tend to have familiar meaning because they tend to be criterion-referenced[14], (2) they are moderately simple to calculate and explain, (3) they can be calculated by stakeholders, and (4) they are relatively simple to aggregate and compare across grades and content areas.

The drawbacks of transitions include that (1) they tend to be relatively unstable, and (2) they introduce the nuance of growth being interpreted relative to increasing expectations.

Residual-based Models, or Value Added Models (VAMs)

With residual-based models, students' achievement scores, gain scores, or growth rates are modeled statistically using an approach that is intended to isolate the contributions of educators (e.g., teachers) or educational institutions (e.g., schools, districts) to students' achievement/gains/growth. Whether they are successful in these attempts is a matter of some controversy. It may be more accurate to say that they identify the degree to which scores/gains/growth rates in a given classroom, school, or district deviate on average from what would be expected in an average classroom, district, or school.

These models can vary from relatively simple[15] to moderately complex[16], to highly complex[17]. The purpose of such models is to estimate the degree to which each educator/institution detracts from or improves on the outcome for its students compared to the expected outcome if those students were being educated by an average educator/institution. In other words, these models attempt to estimate the value added by an educator/institution to its students' outcomes. For various reasons, the "value added" to an individual student's score/gain/growth tends to be relatively unstable[18]. Therefore, VAM scores tend to be unit-aggregate measures rather than individual-student measures.

Status-based VAM

With status-based VAM, the student outcome is the achievement score (or status). A hypothetical example begins in Figure 24. In this example, the grade 3 and grade 4 test scores of all students are statistically analyzed, and what each student's grade 4 test score would be if she were in an average school is estimated from her grade 3 scores. In this case, Hong's grade 3 test scores in ELA/literacy, math, and science are used to predict what her grade 4 math score would be if she were in an average school. Each student's observed grade 4 scores may be different than her predicted scores. For example, in Figure 25, Hong's observed grade 4 math score is higher than her predicted score by 70 points. The

[14] If the categories are developed based on performance levels. It is possible to create categories in a norm-referenced manner, such as using deciles as the categories.
[15] For example, a linear regression with fixed effects (dummy variables) for each educator or institution.
[16] For example, a cross-classified hierarchical linear model with testing occasions nested within both students and educators/institutions.
[17] For example, a layered model with a persistence parameter estimating how value added decays over time.
[18] Value-added for individual students tend to be strongly affected by multiple measurement errors, regression to the mean, scale drift, and by attributing all otherwise unaccounted-for variance to educators/institutions.

value added to Hong's grade 4 test score is therefore 70 points[19]. The value-added by the school is the average value added for the students in the school[20].

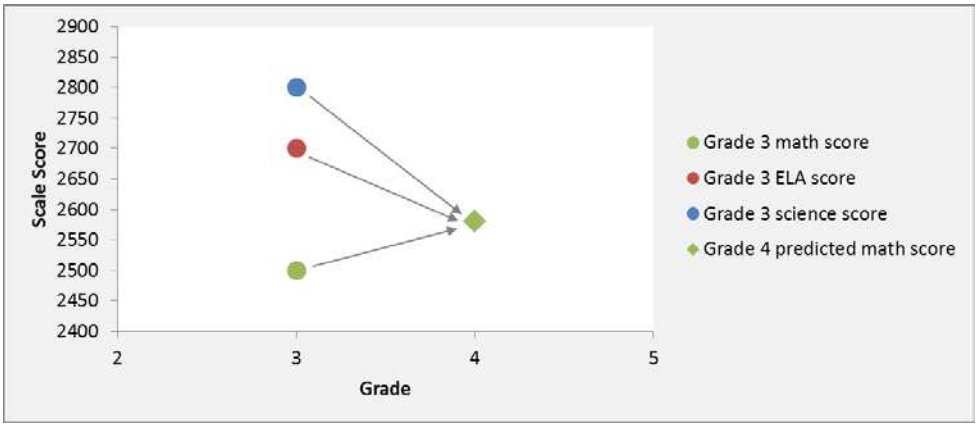

Figure 24. Observed Grade 3 and Predicted Grade 4 Test Scores in a Status-Based VAM.

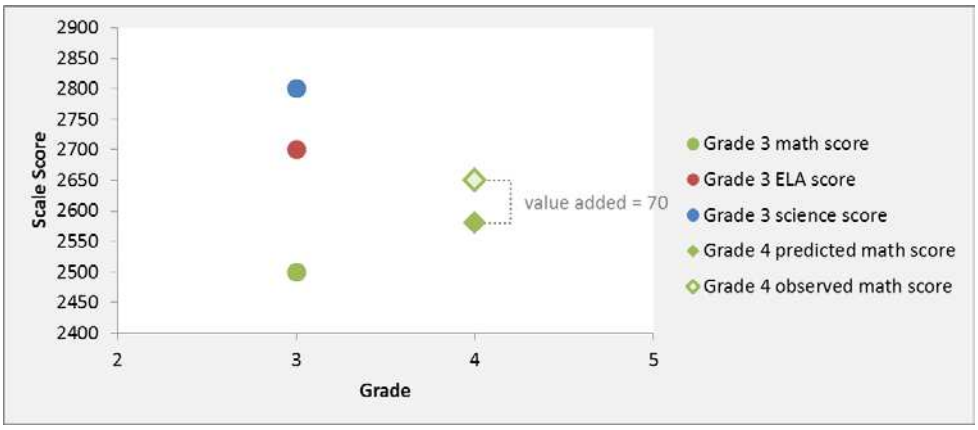

Figure 25. Observed and Predicted Test Scores in a Status-based VAM.

Gain-based VAM

With gain-based VAM, the student outcome is a gain score. A hypothetical example begins in Figure 26. In this example, the grade 3 test scores and grade 4 gain scores of all students are statistically analyzed. What each student's grade 4 gain score would be if her were in an average school is estimated from his grade 3 scores. In this case, Jerry's grade 3 *test scores* in ELA/literacy, math, and science are used to predict what her grade 4 math *gain score* would be if she were in an average school.

[19] The term "value added" is used here for convenience in explanation (it typically only applies to the average difference between predicted and observed scores across students of an educator/institution).
[20] Approximately. Some VAMs adjust for the size of the sample of students served by an educator/institution or account for decay in the value added by an educator/institution over time.

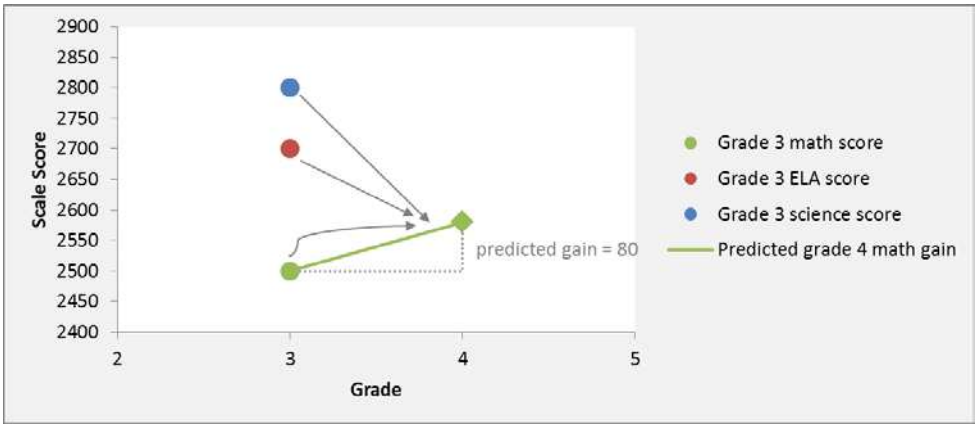

Figure 26. Observed Grade 3 Test Scores and Predicted Grade 4 Gain Score.

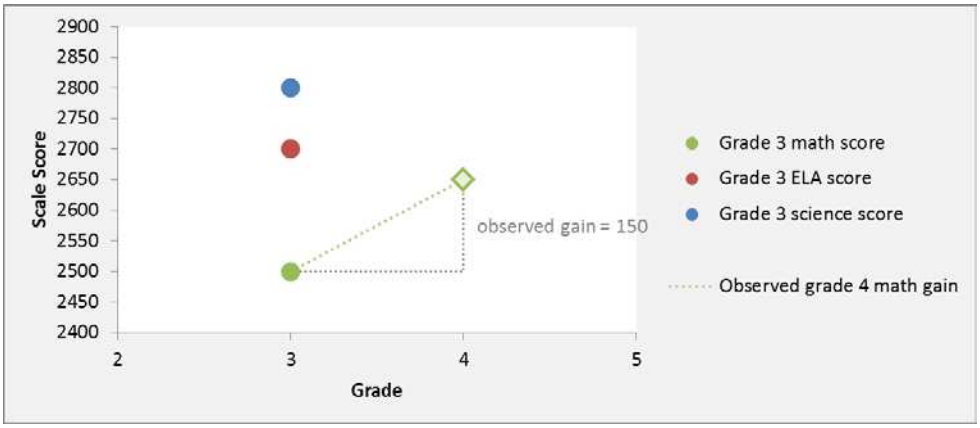

Figure 27. Observed Grade 3 Test Scores and Grade 4 Gain Score.

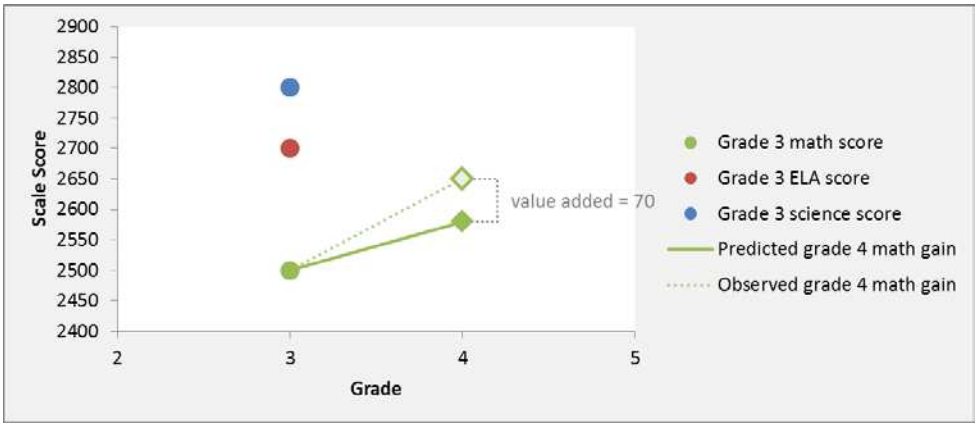

Figure 28. Value-added in a Gain-based VAM.

Each student's observed grade 4 scores may be different than her predicted scores. For example, in Figure 27, Jerry's observed grade 4 math gain score is higher than her predicted gain score. The difference (or residual) between the *predicted gain score* in an average school and the *observed gain score* is the value added to Jerry's gain score, as shown in Figure 28. The value-added for a school is the average value added to the gain scores of all students in the school[21].

Growth-rate-based VAM

With growth-rate-based VAM, the student outcome is a growth rate. A hypothetical example begins in Figure 29. In this example, an observed growth rate is estimated from Liana's math scores in grades 3–6 by fitting a linear regression of her test scores on grade level.

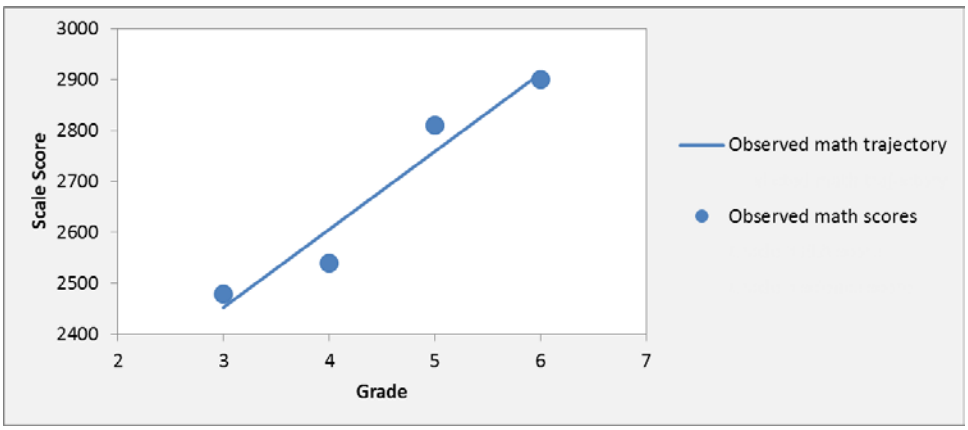

Figure 29. Linear Regression of Achievement on Grade Level in a Growth-rate-based VAM.

In addition to an observed estimated growth rate for Liana, what her growth rate would have been if she had been in an average school from grade 3-6 is also estimated, as shown in Figure 30.

As shown in Figure 31, the predicted growth rate (125 in this example) is subtracted from the observed growth rate (153 in this example) to calculate the value added to Liana's growth rate (28 in this example).

The value added score for the school is calculated as the average of the value added to the growth rate of all students in the school.

[21] It should be noted that the difference between a status-based VAM and a gain-based VAM is negligible when there is only one year of pre-test data.

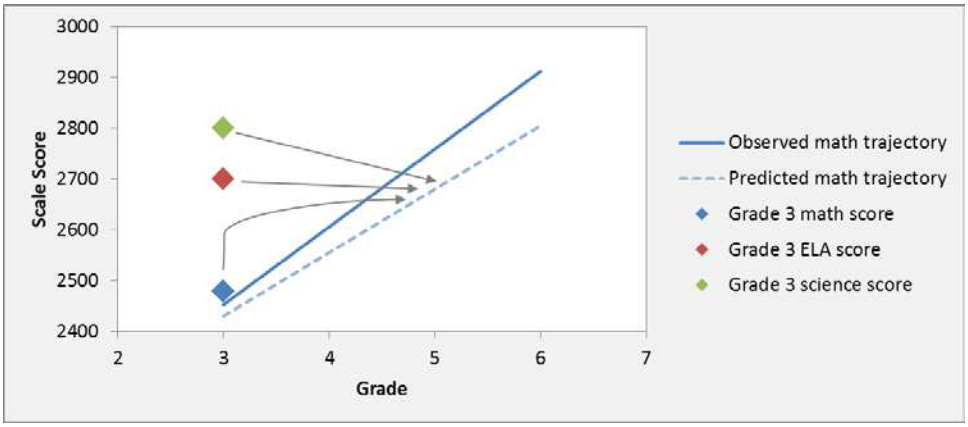

Figure 30. Estimated Growth Rate in an Average School in a Growth-rate-based VAM.

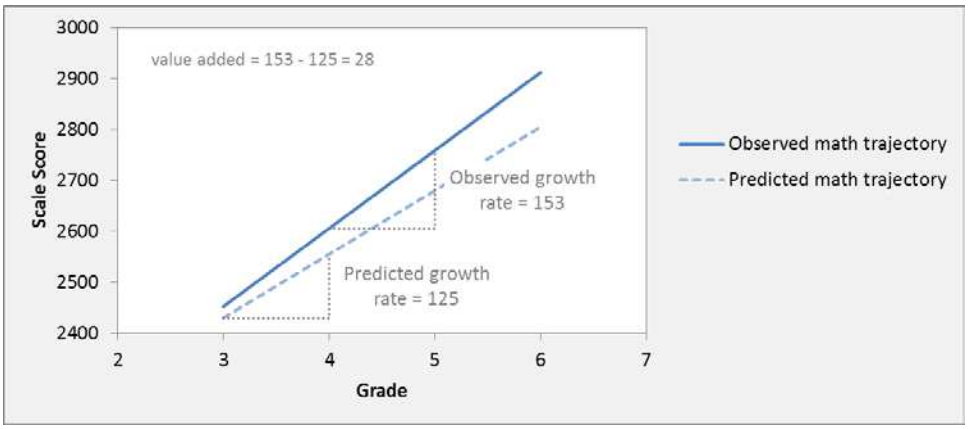

Figure 31. Difference between Observed and Predicted Growth Rate in Growth-rate-based VAM.

Considerations with VAMs

The benefits offered by VAMs include that (1) they attempt to isolate the effects of classrooms, schools, and districts on student status, gains, or growth[22], (2) they can be made relatively stable by conditioning on many past scores in multiple content areas, and (3) if they use test scores that have been standardized, they are relatively simple to aggregate and compare across grades and content areas.

The drawbacks of VAMs include that (1) they tend to introduce a zero-sum game because resulting scores are reported relative to the average classroom, school, or district; (2) they are highly complex to calculate; (3) they introduce considerable difficulties in communication because they simultaneously require understanding of the VAM score scale, norm-referencing, and conditioning on various pre-test

[22] If the categories are developed based on performance levels. It is possible to create categories in a norm-referenced manner, such as using deciles as the categories.

scores; and (4) they can easily be over-interpreted as identifying a causal effect on student scores/gains/growth when doing so would generally require random assignment of students to teachers.

Prediction (Projection) Models

With prediction models, two sets of data are required. First, a set of data from at least one previous cohort is required to establish statewide statistical relationships between scores in earlier grades and scores in later grades. Those statistical relationships are then assumed to remain stable over time. The second set of data consists of scores in earlier grades for the current cohort of students. Future scores of the current cohort of students can then be predicted using the statistical relationships established via analysis of prior cohorts' data. Examples are given in Figures 9-12 and the associated narrative.

Prediction model results can be aggregated across students in a unit (e.g., classroom, school, district, state) in at least the following ways:

- Using a projection-based category interpretation by…
    - Calculating the percent of predicted/projected future scores in each of multiple ranges such as achievement levels.
    - Calculating the percent of predicted/projected future scores at or above a specific threshold such as a *proficient* cut score.
    - Assigning a subjective value to each predicted score (or predicted achievement level)[23].
- Using a projection-based probability interpretation by…
    - Calculating the average probability of predicted scores being in each of multiple ranges such as achievement levels.
    - Calculating the average probability of predicted scores being at or above a specific threshold such as a *proficient* cut score.

Considerations with Prediction

The benefits offered by prediction models include that (1) they have a relatively familiar meaning because they are interpreted relative to commonly understood and important levels of future achievement (e.g., achieving proficiency at some point in the future), (2) they can be stabilized by using many prior test scores in multiple content areas in the prediction, and (3) they can be aggregated across subjects and grades with only moderate complexity.

The drawbacks of prediction include that (1) they introduce the concept of statistical prediction which is both complex to calculate, (2) they can be affected by scale drift because they tend to predict at least three years into the future, and (3) they can be affected by changes in the relationships between test scores in different grades because predicted scores are predicted by treating early-grade score as if they had been observed in cohort of students used to set up the prediction equations.

[23] Gives policymakers a way to explicitly value different degrees of growth according to their desirability.

Potential Growth Measures at the Intersection of Interpretation and Analytical Model

Specific measures of growth exist at the intersection of interpretation and analytical model. Most analytical models are capable of producing growth measures for multiple interpretations, but not every analytical model can produce growth measures with every interpretation. Figure 32 shows which interpretations are supported by which analytical models.

Z-score gain models, baselined z-score gain models, SGP models, and baselined SGP models do not support scale-referenced interpretations because they are inherently norm-referenced (they transform the score scale into a different scale). Prediction models are only used to project a student's future score, so they do not support any interpretations other than projection-based interpretations. VAM models do not support trajectories because they explicitly account for the educator/institution a student is served by, but future educators/institutions are unknown. Z-score gain models and baselined z-score gain models do not support trajectory-based interpretations because such interpretations require referring to the original score scale, which z-scores eliminate. Finally, no models other than prediction models support projection-based interpretations because only prediction models statistically project student scores into the future.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	Y	N	N	Y	N	N	Y	Y	Y	Y	N
	Norm-referenced	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
	Criterion-referenced judgment	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
Future Scores	Criterion-referenced trajectory continuation	Y	N	N	Y	Y	Y	Y	N	N	N	N
	Criterion-referenced trajectory target	Y	N	N	Y	Y	Y	Y	N	N	N	N
	Criterion-referenced category projection	N	N	N	N	N	N	N	N	N	N	Y
	Criterion-referenced probability projection	N	N	N	N	N	N	N	N	N	N	Y

Figure 32. Interpretations of Growth Measures Supported by the Various Analytical Models.

Additional Characteristics of Potential Growth Measures to Consider

While the most important characteristics to consider in selecting a measure of growth is the match of the interpretation and analytical model to the intended use, there are also other important characteristics to consider.

Ability to Improve Stability by Conditioning on Other-subject Prior Test Scores[24]

This characteristic may be important because growth measures tend to be unstable compared to achievement scores. Using information about the relationship between scores in one subject and prior scores in other subjects can help to reduce the instability. The ability of each potential measure of growth

[24] Without changing the nature of the growth measure. For example, accounting for other-subject prior scores in calculating a gain score would change the nature of the score from a gain score in that subject into a gain score in that subject adjusted from prior achievement in other subjects.

to accommodate this ability is shown in Figure 33. Gain scores, z-score gains, baselined z-score gains, growth rates, and transitions tend not to have this capacity because they tend to be relatively simple models.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	N	-	-	N	-	-	N	Y	Y	Y	-
	Norm-referenced	N	N	N	N	Y	Y	N	Y	Y	Y	-
	Criterion-referenced judgment	N	N	N	N	Y	Y	N	Y	Y	Y	-
Future Scores	Criterion-referenced trajectory continuation	N	-	-	N	Y	Y	N	-	-	-	-
	Criterion-referenced trajectory target	N	-	-	N	Y	Y	N	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	Y
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	Y

Figure 33. Ability of Potential Growth Measures to Improve Stability by Conditioning on Other-subject Scores.

Degree of Mismatch to Common Understanding of Calculating Growth

This characteristic may be important because the common understanding of growth is that it represents the amount of change from one point in time to another. Without a successful strategy for explaining the problems with such simple measures of growth, other measures may face opposition on the basis that they can't be used in the same way a tape measure can be used to measure growth in height. The degree of mismatch of each potential type of growth measure to the common understanding of growth is displayed in Figure 34. An explanation of the ratings in Figure 34 is given below:

- Gain scores and growth rates have a low degree of mismatch because they can be interpreted as the *difference in scale scores from one grade to the next*.
- Z-score gains (and their baselined versions) have a moderate degree of mismatch because scores are *transformed to a z-score metric before subtracting prior scores from current scores*.
- Transitions have a moderate degree of mismatch because they are based on ranges of scale scores (such as achievement levels in a criterion-referenced calculation, or deciles in a norm-referenced calculation). *Comparing ranges of performance from one grade to the next* is a form of subtraction that may be difficult for stakeholders to understand.
- Gain-based and growth-rate-based VAMs have a moderate degree of mismatch because they report average differences between predicted gain scores/growth rates (given an average educator/institution) and observed gain scores/growth rates, though this may be reduced by conditioning on many prior test scores.
- SGPs (and their baselined versions) have a high degree of mismatch because they report on *current status relative to peers with the same score history*.
- Status-based VAMs have a high degree of mismatch because they report average differences between predicted achievement scores (given an average educator/institution) and observed achievement scores.
- Prediction-based scores have a high degree of mismatch because they report *predicted achievement levels*.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	L	-	-	L	-	-	M	H	M	M	-
	Norm-referenced	L	M	M	L	H	H	M	H	M	M	-
	Criterion-referenced judgment	L	M	M	L	H	H	M	H	M	M	-
Future Scores	Criterion-referenced trajectory continuation	L	-	-	L	H	H	M	-	-	-	-
	Criterion-referenced trajectory target	L	-	-	L	H	H	M	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	H
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	H

Figure 34. Degree of Match to Common Understanding of Calculating Growth.

Whether the Measures Creates a Zero-Sum Game

This characteristic may be important because a common concern of stakeholders is that when a zero-sum (or fixed-pie) game is created, system-wide improvements are not rewarded. This happens because if norms are reset every year, there will always be approximately the same proportion of student below and above average (or in each percentile). However, if a baseline set of norms is put in place and subsequent data are compared to the baseline, it is theoretically possible for all students to be above average if the entire system improves. The degree to which each measure creates a zero-sum game is displayed in Figure 35. An explanation of the ratings in Figure 35 is given below:

- Z-scores gains are created by norming scores against the distribution of scores they came from, creating a zero-sum game.
- SGPs are created by comparing each student's scores to a norming group that took the same set of prior-grade tests, creating a zero-sum game.
- In VAM, value added is compared to the classroom, school, or district of average effectiveness. Therefore, some will always be below average.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	N	-	-	N	-	-	N	Y	Y	Y	-
	Norm-referenced	N	Y	N	N	Y	N	N	Y	Y	Y	-
	Criterion-referenced judgment	N	Y	N	N	Y	N	N	Y	Y	Y	-
Future Scores	Criterion-referenced trajectory continuation	N	-	-	N	Y	N	N	-	-	-	-
	Criterion-referenced trajectory target	N	-	-	N	Y	N	N	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	N
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	N

Figure 35. Whether Growth Measures Create a Zero-Sum Game.

Relative Degree of Difficulty in Aggregating Across Subjects and Scales

This characteristic may be important because it may be necessary to calculate growth scores in the same subject area even when a test has changed, and because it may be necessary to aggregate growth scores across subjects (for example in an accountability system). The degree of difficulty is presented in Figure 36, followed by an explanation of the ratings.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	H	-	-	H	-	-	M	H	H	H	-
	Norm-referenced	L	L	L	L	L	L	L	L	L	L	-
	Criterion-referenced judgment	L	L	L	L	L	L	L	L	L	L	-
Future Scores	Criterion-referenced trajectory continuation	M	-	-	M	M	M	M	-	-	-	-
	Criterion-referenced trajectory target	M	-	-	M	M	M	M	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	M
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	M

Figure 36. Degree of Difficulty in Aggregating Across Subjects and Scales.

- Scale-referenced gain scores, growth rates, and VAMs present a high degree of difficulty (H) in combining across different versions of a test with different scales or across subjects. This is because they are reported relative to achievement score scales, which are unique to each test.
- Scale-referenced transitions, trajectory-based interpretations, and projection-based interpretations present a moderate degree of difficulty (M) because in order for comparisons across tests to be meaningful, the tests must share a common set of approximately comparable achievement levels.
- With norm-referenced interpretations such as *above average* or *percentile rank*, it is easy (L, or low degree of difficulty) to compare scores on a previous version of a test with a new version of a test, or in one subject versus another.

Relative Degree of Complexity of Growth Score Calculations

This characteristic may be important because the more complex the calculation of growth measures is, the more likely it is that they will be criticized for coming from a “black box.” Stakeholders may desire to be able to calculate their own growth scores. On the other hand, more complex models are more likely to be able to account for important nuances in order to make growth measures more valid for their intended uses and interpretations. The degree of complexity in calculating the various growth measures is presented in Figure 37, followed by an explanation of the ratings.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	L	-	-	M	-	-	M	H	H	H	-
	Norm-referenced	M	M	H	M	H	H	M	H	H	H	-
	Criterion-referenced judgment	L	M	H	M	H	H	M	H	H	H	-
Future Scores	Criterion-referenced trajectory continuation	M	-	-	M	H	H	M	-	-	-	-
	Criterion-referenced trajectory target	M	-	-	M	H	H	M	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	H
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	H

Figure 37. Degree of Complexity in Calculating Growth Measures.

The following factors introduce moderate complexity (M) into growth score calculations:

- Comparing scores to a norming group as in z-scoring, SGPs, or other norm-referencing.
- Applying a prior year's norms to a current year's scores as in baselining.
- Comparing categories across years as in transitions.

Where more than one factor introducing moderate complexity applies to a measure, it is rated as highly complex (H). Finally, sophisticated statistical models (as in SGP, VAM, and prediction models) introduce a high degree of complexity in growth score calculation. Measures without any complicating factors are rated as low (L).

Relative Degree of Difficulty in Communication

This characteristic may be important because difficult-to-communicate measures of growth require considerable effort in strategic communication and professional development. The degree of difficulty in communication about the various potential growth measures is displayed in Figure 38, followed by an explanation of the ratings.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	H	-	-	H	-	-	H	H	H	H	-
	Norm-referenced	L	H	H	L	L	M	H	H	H	H	-
	Criterion-referenced judgment	L	H	H	L	L	M	H	H	H	H	-
Future Scores	Criterion-referenced trajectory continuation	M	-	-	M	M	M	M	-	-	-	-
	Criterion-referenced trajectory target	M	-	-	M	M	M	M	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	M
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	H

Figure 38. Degree of Difficulty in Communication.

The following factors introduce moderate difficulty (M) into communication about growth measures:

- Mathematical scale transformations, as in z-score gains.
- Applying a prior year's norms to a current year's scores as in baselining.
- Comparison of performance categories from one year to the next when expectations increase over years (e.g., is a student keeping pace with increasing expectations?) as in transitions.
- Following an observed trajectory into the future.
- Understanding prediction of future scores based on relationships of scores in previous cohorts.
- Understanding probabilities.

Where more than one factor introducing moderate difficulty applies to a measure, communication becomes hard (H). Finally, a large number of negative growth scores makes communication hard (H), as with gain scores, z-score gains (including baselined versions), transitions, and VAMs.

Relative Degree of Correlation with Status Scores

This characteristic is important because if growth scores are highly correlated with status scores, the purpose of using growth scores has likely been thwarted because the growth scores offer little new information. The relative degree of correlation of growth measures with status scores is displayed in Figure 39, followed by an explanation of the ratings.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	M	-	-	L	-	-	M	L	M	L	-
	Norm-referenced	M	M	M	L	L	L	M	L	M	L	-
	Criterion-referenced judgment	M	M	M	M	M	M	M	M	M	M	-
Future Scores	Criterion-referenced trajectory continuation	H	-	-	H	H	H	H	-	-	-	-
	Criterion-referenced trajectory target	H	-	-	H	H	H	H	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	H
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	H

Figure 39. Relative Degree of Correlation with Status Scores.

The following factors tend to introduce a moderate (M) degree of correlation of growth scores with status scores:

- The use of only two test scores, as in gain scores, z-score gains (including baselined versions), transitions, and gain-based VAM (because of regression to the mean).
- Judgmental definition of adequate growth, as in judgment-defined growth-criterion interpretations (because policymakers tend to identify adequate growth as a level of growth likely to lead to a positive achievement outcome, though this may not always be the case).

Where more than one factor tending to introduce moderate correlation applies to a measure, the degree of correlation is rated as high (H). Finally, any future-focused interpretation based on meeting a desirable level of achievement in the future introduces a high degree of correlation with status measures. The fewer the number of years given to achieve the desired level of achievement, the higher the correlation will be.

Relative Degree of Unreliability/Instability of Individual Student Growth Scores

This characteristic is important because relatively high levels of error in individual student growth scores increases the likelihood that poor educational decisions will be made on the basis of those scores. The relative degree of stability/reliability of growth measures is displayed in Figure 40, followed by an explanation of the ratings.

Because VAM models are typically not used to produce student-level scores, this characteristic is not applicable to VAMs. For other measures of growth, the following factors tend to moderately (M) reduce stability/reliability:

- Projection into the future.
- Creating hypothetical trajectories into the future.
- Discarding information by dividing a scale into categories.
- Discarding information by not using off-subject scores to improve stability.
- Increasing susceptibility to scale drift by baselining.

Where more than one factor tending to reduce stability applies, the degree of instability/unreliability is rates as (H). Finally, any measure of growth based only on two scores (as in gain scores, z-score gains, and transitions) tend to have high levels of instability/unreliability.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	H	-	-	M	-	-	H	-	-	-	-
	Norm-referenced	H	H	H	M	L	M	H	-	-	-	-
	Criterion-referenced judgment	H	H	H	M	L	M	H	-	-	-	-
Future Scores	Criterion-referenced trajectory continuation	H	-	-	H	M	H	H	-	-	-	-
	Criterion-referenced trajectory target	H	-	-	H	M	H	H	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	H
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	M

Figure 40. Relative Instability/Unreliability of Growth Measures.

Relative Degree of Susceptibility to Equating/Scale Drift

This characteristic may be important because increasing the number of years of observed data used for calculating measures of growth can increase the stability of growth measures, but that has to be balanced with increasing susceptibility to scale drift, which can decrease scale stability. The relative degree of susceptibility of growth measures to scale drift is displayed in Figure 41, followed by an explanation of the ratings.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	M	-	-	M	-	-	M	L	M	M	-
	Norm-referenced	M	M	H	M	L	H	M	L	M	M	-
	Criterion-referenced judgment	M	M	H	M	L	H	M	L	M	M	-
Future Scores	Criterion-referenced trajectory continuation	H	-	-	H	H	H	H	-	-	-	-
	Criterion-referenced trajectory target	H	-	-	H	H	H	H	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	H
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	H

Figure 41. Relative Susceptibility to Scale Drift.

The following factors tend to increase susceptibility of growth measures to scale drift to a moderate degree:

- Assuming a vertical scale (as in gain scores, growth rates, and gain- or growth-rate-based VAM).
- Assuming vertical articulation of cut scores (as in transitions).

The following factors tend to increase susceptibility to scale drift to a high degree:

- Baselining (referencing the growth to a scale multiple years in the past).
- Prediction (referencing the growth to a scale multiple years into the future).

Number of Years of Data Required from the Same Assessment to Create Growth Measures

This characteristic may be important because it may be necessary for a state to calculate growth scores in the early years of implementing a new assessment (sometimes in the first year, usually by the second year). The number of years of data from the same assessment required to calculate growth scores is displayed in Figure 42, followed by an explanation of the ratings.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	2	-	-	4	-	-	2	1	2	4	-
	Norm-referenced	2	2	3	4	1	3	2	1	2	4	-
	Criterion-referenced judgment	2	2	3	4	1	3	2	1	2	4	-
Future Scores	Criterion-referenced trajectory continuation	2	-	-	4	2	3	2	-	-	-	-
	Criterion-referenced trajectory target	2	-	-	4	2	3	2	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	2
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	2

Figure 42. Number of Years of Data from the Same Assessment Required to Calculate Growth Scores.

- SGPs and status-based VAMs only require one year of data (they can use data from previous assessment program as pre-test scores).
- Gain scores, z-score gains, transitions, and gain-based VAM require two years of data from the same assessment because they assume either vertically aligned cut scores (transitions) or vertical scales (all others).
- Baselining requires at least three years of data (two years of data for a cohort of students to set the baseline, and two years of data from the next cohort to compare to the baseline).
- Prediction requires at least two years of data (two years of data for a cohort of students to establish the statistical relationships between scores in one grade and the next, the lower-grade data for the next cohort from which to predict their next-grade scores)[25].
- Growth rates and growth-rate-based VAM require at least four years of data from the same assessment because the calculation of growth rates does not improve measurement of growth over gain scores until four years of data are available[26].

Ability to Produce Individual Student Growth Scores

This characteristic may be important for specific uses. Value added models (VAMs) tend to be used to produce only aggregate growth scores, as shown in Figure 43.

Interpretation		Analytical Model										
Focus	Description	Gain Score	Z-score Gain	Baselined Z	Growth Rate	SGP	Baselined SGP	Transition	Status VAM	Gain VAM	Rate VAM	Prediction
Observed Scores	Scale-referenced	Y	-	-	Y	-	-	Y	N	N	N	-
	Norm-referenced	Y	Y	Y	Y	Y	Y	Y	N	N	N	-
	Criterion-referenced judgment	Y	Y	Y	Y	Y	Y	Y	N	N	N	-
Future Scores	Criterion-referenced trajectory continuation	Y	-	-	Y	Y	Y	Y	-	-	-	-
	Criterion-referenced trajectory target	Y	-	-	Y	Y	Y	Y	-	-	-	-
	Criterion-referenced category projection	-	-	-	-	-	-	-	-	-	-	Y
	Criterion-referenced probability projection	-	-	-	-	-	-	-	-	-	-	Y

Figure 43. Number of Years of Data from the Same Assessment Required to Calculate Growth Scores.

[25] Two years of data only works when tests are administered in adjacent grades. For states with a gap between 8th grade testing and high school testing (e.g., 10th or 11th grade), the number of years of data from the same assessment is 2 plus the number of grades skipped between grade 8 and high school assessment.

[26] Because if a linear regression is calculated on the basis of three score points, the slope (or growth rate) depends only on the first and last scores, with the intercept depending on all three. In the figures below (Figure A and Figure B), the first and third scores in each of the three panels in each figure are the same. Only the middle score differs. The slopes are the same in all three panels in each figure. Only the intercept differs across panels.

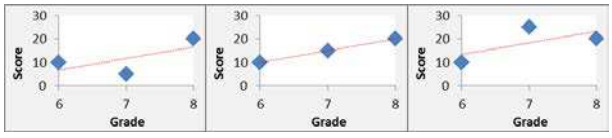

Figure A. Three panels with the same slope.

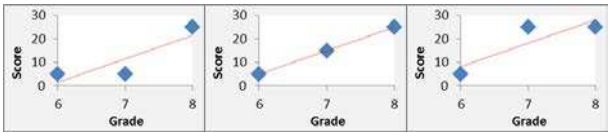

Figure B. Three panels with the same slope.

Consideration of Select Tensions in Measuring Student Growth

After developing a theory of action, states should identify any intended uses in each of these categories. Based on both the theory of action and the intended uses, the intended interpretations of growth scores can be more clearly defined. In identifying the intended uses and interpretations, states will almost assuredly need to deal with some tensions that are introduced by the use of growth measures. A few such tensions are discussed below.

Simplicity vs. Validity

Some interpretations of growth are more complex than others, and some analytical models are more complex than others. Though it is not always true that a simple interpretation or analytical model is less valid for an intended use, this is often the case. More complex interpretations and analytical models tend to be more capable of supporting important nuances of intended uses. States will need to evaluate the degree to which an increase in validity for a particular interpretation outweighs the benefit of a simple growth measure.

Expectations for Student Achievement vs. Expectations for Educational Effectiveness

One reason that growth measures are seen as desirable is that they tend to account for incoming achievement by measuring growth instead of just measuring achievement level. This is seen as leveling the playing field for educators and educational institutions such as schools and districts in that they are not held accountable for something they cannot control (students' incoming level of achievement), instead being held accountable for the degree of learning the student experiences while the educator/institution is responsible for teaching the student.

However, there are other things which teachers cannot control that affect both incoming achievement and growth. It is well documented that various demographics (such as economic disadvantage, race/ethnicity, having a disability, being mobile, or being limited English proficient) are strongly predictive of student achievement scores. It is also well documented that the same demographics are only weakly predictive of student growth scores, yet they remain somewhat predictive of student growth scores.

To further level the playing field for educators/institutions, any measure of growth can also be adjusted for the demographics that educators/institutions cannot control. However, states should be aware of the implications of doing so. The implications of holding educators accountable for student achievement scores, student growth scores, and student growth scores adjusted for demographics are summarized below:

- **Holding educators/institutions accountable for achievement scores**: fully privileges equal expectations for student achievement over equal expectations for educational effectiveness by implicitly stating the following:
    - We hold the same achievement expectations for students without regard to factors they cannot control (e.g., socioeconomic status, English proficiency, disability status, mobility, race/ethnicity, home environment).
    - In order to meet the equal achievement expectation, educators/institutions serving typically low-achieving student populations are expected to be much more effective at eliciting student learning than educators/institutions serving other student populations.
- **Holding educators/institutions accountable for unadjusted growth scores**: mostly privileges equal expectations for educational effectiveness over equal expectations of student achievement by implicitly stating the following:
    - We hold the same expectation for educational effectiveness for all teachers/institutions, and therefore equal growth expectations for students.

  - Existing achievement gaps are acceptable because if all students experience the same level of growth, existing achievement gaps will remain unchanged.
  - In order to meet the equal growth expectation, educators/institutions serving populations of students that typically grow at a slower rate are expected to be slightly more effective at eliciting student learning than educators/institutions serving other student populations.
- **Holding educators/institutions accountable for demographic-adjusted growth rates**: fully privileges equal expectations for educational effectiveness over equal expectations of student achievement by implicitly stating the following:
  - We hold the same growth expectations only for students who share the same demographic background because not to do so would be to expect some teachers/institutions to be more effective than others.
  - It is acceptable for existing achievement gaps to widen if the lower-achieving demographic group also tends to grow at a slower rate.
  - In order to meet the equal educator/institutional effectiveness expectation, it is acceptable if students from whom it is more difficult to elicit growth exhibit lower growth than other students.

Both the U.S. Department of Education (USED) and some states have attempted to balance equal expectations of student achievement with equal expectations of educational effectiveness. Under the No Child Left Behind (NCLB) growth model pilot program, USED implemented such an attempt by requiring states to use prospective interpretations of growth (or growth-to-standard interpretations) in one of the three following ways:

- Require that students observed growth be sufficient to become proficient within a maximum of four years (*trajectory-based continuation interpretation*).
- Require states to set a target growth trajectory sufficient to become proficient within a maximum of four years (*trajectory-based target interpretation*).
- Require states to predict a student's future achievement level no more than four years into the future (*projection-based category interpretation*).

As they select measures of growth, states will need to determine how important leveling the playing field for educators (i.e., allowing existing achievement gaps to remain or increase) is compared to leveling the playing field for students (i.e., narrowing or closing existing achievement gaps).

Equal Expectations of Student Achievement vs. Gaining New Information from Growth Measures

As described above, in order to maintain equal expectations for student achievement when using growth measures, it is important to use a growth-to-standard approach (e.g., future-focused interpretations using either trajectories or projections). However, the degree to which growth measures are correlated with status measures depends on how short the window for achieving a desirable status is. NCLB growth pilots allowed only 3-4 years to become proficient, resulting in growth scores being relatively strongly correlated with status scores, calling into question how much new information is included in growth scores. Allowing a longer window may improve the amount of new information available in growth scores. However, states will need to determine how long is reasonable to allow, and whether it is reasonable to provide a different time limit for students in elementary school versus students in high school.

Guarding against Measurement/Sampling Error vs. Guarding against Scale Drift.

It is possible to guard against growth scores being affected by measurement and sampling error by using more years of data to calculate growth scores and/or by averaging across multiple years of growth scores. However, the more years that are involved in growth calculations, the more susceptible the measures will be to scale drift. Drift across a small number of years is likely to be relatively small, but

small degrees of drift can accumulate over time to become important to consider. States will need to consider how to balance instability attributable to measurement and sampling error (which can be reduced by including more years of data) with instability attributable to scale drift (which can be increased by including more years of data).

Conclusion

States can use this guide to both understand and help them select one or more measures of growth. The recommended approach is to take the steps described in the figure below. By following a disciplined process for understanding and selecting a growth measure, states are much more likely to be successful in both implementing one or more growth measures and in achieving the goals driving the implementation of a growth measure.

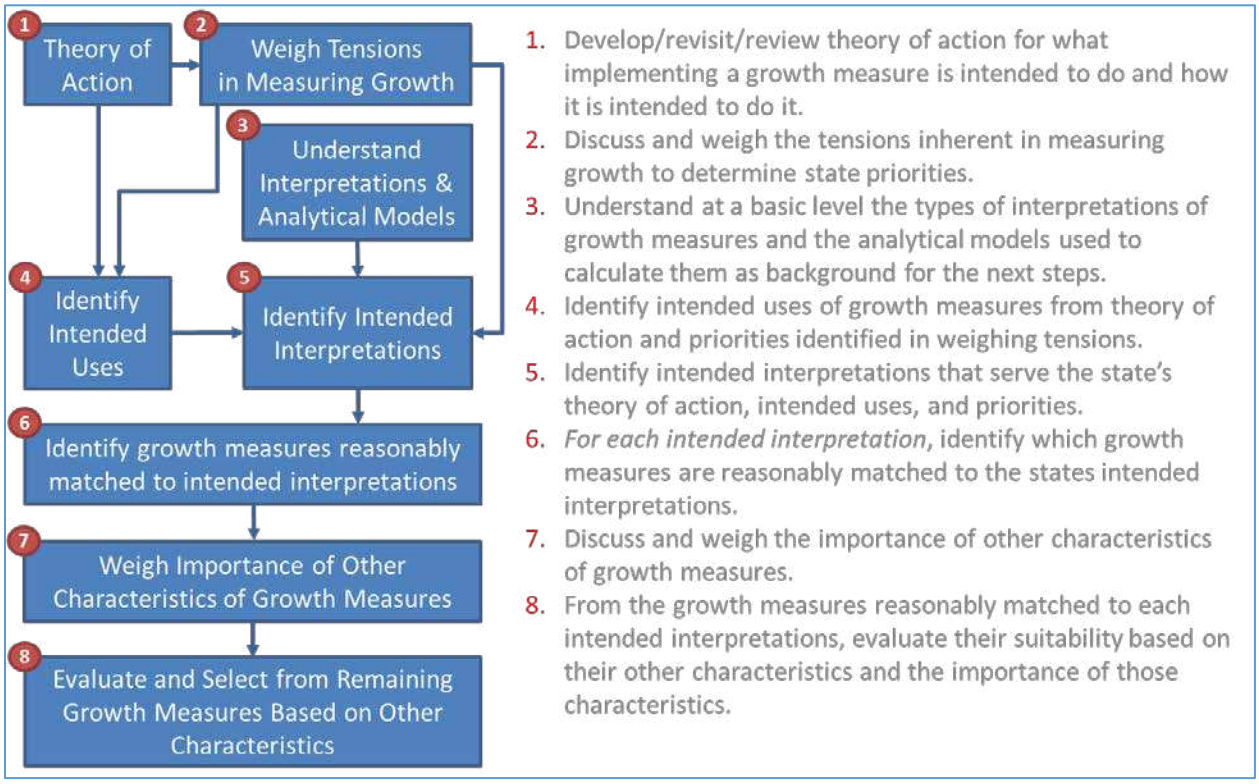

Figure 44. Suggested Process for Selecting One or More Measures of Growth.

References

Castellano, K. E., & Ho, A. D. (2013). A Practitioner's Guide to Growth Models. Washington, DC: CCSSO. Retrieved April 10, 2016 from http://www.ccsso.org/Documents/2013GrowthModels.pdf

Shakman, K., & Rodriguez, S. M. (2015). Logic Models for Program Design, Implementation, and Evaluation: Workshop Toolkit. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Education Laboratory Northeast & Islands. Retrieved April 20, 2016 from https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=401