AIR®
American Institutes for Research®

# Selecting an Approach to Measuring Student Learning Growth for Educator Evaluation

## Information for Florida School Districts

July 2014

# Selecting an Approach to Measuring Student Learning Growth for Educator Evaluation: Information for Florida School Districts

**July 2014**

AIR®

American Institutes for Research®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY 877.334.3499

**www.air.org**

07/2014

# Contents

**Page**

## Introduction

Under the Student Success Act of 2011 (Senate Bill 736) and Florida's successful Race to the Top application, districts around the state must develop new systems for educator evaluation. Educator evaluation systems are developed locally by each school district and in accordance with provision of Section 1012.34(3)(a)1.-4. must incorporate at least three measures of teacher performance: student outcomes, instructional practice, and professional and job responsibilities. This document is designed to provide districts with guidance and options for incorporating the performance of students into its teacher evaluation systems.

Section 1012.34(3)(a)1., F.S. requires that 50 percent of a teacher's evaluation be comprised of measures of student performance[1]. For grades and subjects with approved models that use statewide assessments, districts must use data from these models provided by the department. Currently, this includes the following grades and subjects:

- Reading (4th, 5th, 6th, 7th, 8th, 9th, 10th)
- Mathematics (4th, 5th, 6th, 7th, 8th)
- Algebra 1 (9th)

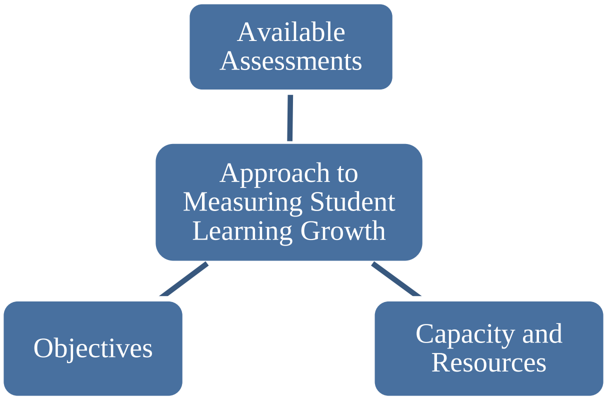
For teachers whose course assignments include subjects and grades that are not assessed with statewide assessments, or when an approved model using statewide assessment data is not in place, districts must develop their own approaches to measure student outcomes for teacher evaluation purposes. This document is a resource for districts in considering approaches to measure student performance for the purposes of educator evaluation. This document assumes that districts have already made decisions about key aspects of their educator evaluation systems, such as deciding which and how many of a teacher's courses are to be used for evaluation and how evaluation components will be weighted and used for decision-making.

As shown in Figure 1, in selecting an approach to measure the performance of students for teacher evaluation purposes, districts will need to consider three key factors:

- Objectives
- Available Assessments
- Capacity and Resources

[1] This amount can be reduced to 40% in cases where a teacher has fewer than 3 years' worth of student performance data available.

**Figure 1. Key Issues in Selecting an Approach to Measure Student Learning Growth**



Each of these three factors has implications for how best to approach measuring student learning for educator evaluation in different courses and settings.  Selecting a final approach may depend on an iterative process that includes examining and revisiting all three factors.

After examining these factors, this document provides descriptions of four approaches to measure the performance of students, illustrating how these factors relate to each model, including:

- Percent Proficient Models
- Simple Growth Models
- Advanced Statistical Models
- Student Learning Objectives

It is important to note that although we present each of the approaches individually, it may be possible to select a hybrid approach. Appendix C provides a full example of how a district might select an approach to measuring student learning growth based on the information in this document.

## Issue #1. Objectives

The first factor a district should consider in measuring student performance is the relevant objectives for each course. That is, for each course, districts must first identify the expectation for the teacher with respect to student learning. Making a clear statement about what a teacher is responsible for in a given course will help identify which models can most appropriately be used to measure how well teachers meet these objectives. Depending on the objective for the course, the question to be answered by an evaluation of student learning growth might be any of the following:

**How many students achieved a given level of proficiency by the end of the course?** For some courses, districts might determine that a teacher's primary responsibility is to help students attain a minimum, basic set of skills, regardless of students' knowledge and ability prior to the start of the course. This measure would be reasonable in cases where all students in the district or state start the course with similar levels of knowledge and ability prior to the start of the course, such as might be the case in a highly specialized curricular area. This measure might also be reasonable in cases where a student's attainment of proficiency is essential, such as a course centered on an industry certification, passing an AP or IB exam, or completing a graduation requirement. As an example, a district might set performance standards based on the percent of students who score a level 3 or higher on an AP exam.[2] Ideally, students should be equally well-prepared to attain the required proficiency level upon entering the course.

**How many students attained a given level proficiency, given their beginning levels of performance?** In some courses the objective may be for students to attain a minimum, basic set of skills (to achieve a certain level of proficiency), but it may not be reasonable to expect that all students are equally likely to attain proficiency. For example, student knowledge and preparation prior to the start of the course may affect the likelihood that a student attains the minimum threshold. For example, the likelihood that a student attains proficiency on an Algebra II EOC exam may depend in part on the student's prior achievement in Algebra I and/or Geometry. Or there may be other factors outside of the teacher's control that affect the rate at which students learn. For example, some students with disabilities may progress toward proficiency at a different rate than other students. In such cases, it may be possible to statistically control for some of the factors that lead to these differences in performance.

**How much did students grow by the end of the course?** For some courses, a student's absolute level of achievement may be less important than whether or not the teacher helps all students grow equally during the course. This may be especially true in beginning skills- or performance-based courses, where students may have a long trajectory ahead of them and clear and consistent benchmarks for improvement can be identified for all students (e.g., all students improve speed in a given task by a certain amount). For example, a district might set performance standards that are based on how much students improve their baseline reading fluency as measured by the number of words per minute students are able to read.

**How much did students grow by the end of the course, given their beginning levels of performance?** With this type of objective, all students are expected to grow, but not all are

[2] It is important to note that the objective or question to be answered is distinct from the performance standard set for educators. Appendix A provides more information on performance standards.

expected to grow equally. In some courses, it is important that students build as much as possible upon what they have learned prior to the start of the course. Well-prepared students may be expected to grow more than others, or conversely, less well-prepared students may be expected to show very significant amounts of growth compared to better-prepared peers. For example, a district might establish performance thresholds based on the percent of students who meet or exceed average growth for students with similar beginning levels of performance, or the extent to which average growth in a teacher's classroom exceeds the average growth for students with similar beginning levels of performance across the district. At the state level, this is the perspective on learning that the VAM models take, taking prior performance and other factors into account to determine expected proficiency and incorporating differences of any magnitude above or below expectations into the final score.

## Issue #2. Available Assessments

After considering relevant objectives, the next factor a district should consider when determining how to incorporate student performance into educator evaluations is what assessment(s) are available (or can be made available), and what type of data these assessments produce. Changes introduced by SB 1642 now identify five different types of local assessments that districts may select for use in courses where there is not an approved model to measure student growth using a statewide, standardized assessment. These include:

1. Statewide assessments;
2. Other standardized assessments, including nationally recognized standardized assessments;
3. Industry certification assessments;
4. District-developed or district-selected end-of-course assessments; and
5. Teacher-selected or principal-selected assessments. Local school boards must adopt policies for selection, development, administration, and scoring of these local assessments. These assessments may take a variety of forms, including project-based assessments, adjudicated performances, and practical application assignments. The way the assessment is constructed will, in part, determine how it is used to evaluate teachers.

"Good" measures of student learning for use in educator evaluation are those that align closely with what students are expected to learn and teachers are expected to teach in a course, in terms of content, skills, and complexity. For some courses, an appropriate assessment may include multiple choice and short answer items; for other courses, longer writing assignments or performances may be needed to illustrate relevant student learning. Assessments should allow students at many levels of performance to demonstrate their learning, and administration and scoring procedures should be in place that can produce consistent results across students and classrooms. No assessment is a perfect measure of student knowledge and skills. To the extent possible, districts should consider methods to incorporate uncertainty into their measures of student learning growth. Appendix A provides additional details on uncertainty, and additional references on assessment design considerations can be found in Appendix B.

To identify teachers' contributions to student learning growth, the district will need assessments at two points in time: before learning begins and at the end of the course. These assessments can be thought of as pre-tests and post-tests. It is important to note that having information about student learning at two points in time **does not** necessarily require a pre-test and post-test that are the same, or that measure exactly the same content and skills. In fact, a pre-test that measures a student's preparedness to learn the material that will be assessed at the end of course can be preferable to a pre-test that is similar or identical to an end-of-course post-test. These types of pre-tests can be thought of as "proxy" pre-tests to assess students' baseline proficiency. One example of a "proxy" pre-test is the use of prior FCAT 2.0 Mathematics assessment data as a predictor in the Algebra 1 EOC VAM model. It is also important to note that in some courses, only one assessment may be available, and that a district's main objective may be to measure how many students achieve a given level of proficiency on that one assessment.

Examples of available course assessment structures may include:

**A single assessment that measures performance at the end of a course.** In some courses, no beginning assessment information may be available at all, or it may not be necessary or appropriate to administer a pre-test. For example, for students entering a foreign language course for the first time, it may not be appropriate to administer a beginning assessment that targets the same set of knowledge and skills that the ending assessment will, and students have no prior foreign language assessment data. This may also be the case in highly technical or specialized courses, such as, for example "Introduction to Photography."

**Multiple assessments that measure performance prior to and at the end of a course, are directly comparable in terms of content, and have the same measurement scale.** In some courses, districts may administer assessments at the beginning and end of a course that are directly comparable in terms of content and skill – for example, an assessment of reading for a second grade reading course that is given at the beginning and end of the grade level, or in an advanced foreign language course where students are focused on refining skills. In cases like these, it may be appropriate to assess students at the beginning of the year on exactly the same knowledge and skills on which they will be assessed at the end of the course. In addition, these assessments may be directly comparable in terms of measurement scales – that is, the tests are constructed in such a way that it would be possible to simply subtract one score from another to see how much a student had grown. This type of comparable score is generated from "equated" test forms (i.e. tests that are not identical but were designed and have been shown to be comparable through field testing and statistical analysis) or identical test forms.

**Multiple assessments that measure performance prior to and at the end of a course and are directly comparable in terms of content, but do not have the same measurement scale.** In some courses, districts may administer assessments at the beginning and end of a course that are directly comparable in terms of the content and skills assessed, but the scores produced are not directly comparable because the district does not want to give identical tests and has not statistically established comparable scales.

**Multiple assessments that measure performance prior to and at the end of a course, and are not directly related in terms of content, and do not have the same measurement scale.** In some courses, beginning assessment information may be available for subjects that are related, but that do not directly provide information on beginning performance in that subject. For example, students in a high school chemistry course may have previously taken a biology or earth science course assessment. While the biology and earth science assessments may not provide direct information about students' readiness to learn chemistry, scores from these assessments may serve as a reasonable proxy measure of students' readiness to learn chemistry.

Finally, districts should also consider whether or not they have historical assessment data, because expectations for student achievement could be based on past achievement of similar students. For example, in some courses, districts may have a long history of collecting data about student performance, while in other courses formal end-of-course assessments may not have been typically administered, or the results from such end-of-course assessments may not have been maintained after the conclusion of the course. It may be known, for instance, that each year students typically grow by 25 points on a given set of pre- and post-tests. If such information is not available, the standards by which teachers are evaluated may need to evolve as more data become available.

## Issue #3. Capacity and Resources Needed

The third factor a district will need to consider in selecting approaches to measuring student performance is the capacity and resources that are required to implement each approach. Here we outline several elements to consider, and how they might influence the evaluation option(s) selected by a district.

### Amount and Type of Data Required

Some of the more complex approaches to measuring student performance rely on statistical methods, which typically require larger amounts of data in order to generate teacher-level scores. It may be possible to generate scores with information from just a few courses in a district, but more information will produce more precise scores. Different approaches to measuring student performance can be thought of as falling into two levels with respect to amount of data required:

- High: Information from a relatively large number of students (ideally, at least several hundred) with the same beginning and ending assessment scores is required, with more precise estimates generated as sample sizes increase. Other student data (such as background characteristics) may also be needed.
- Low: Information from a single classroom of students may be sufficient to produce scores.

### Statistical/Technical Capacity

Some approaches to measuring student performance rely on advanced statistical methods, which require statistical knowledge, potentially expensive software licenses, and computing capacity to be available in a district. Different approaches to measuring student performance can be thought of as falling into two levels with respect to statistical and technical capacity required:

- High: A relatively high level of statistical knowledge is needed to analyze and interpret data with this approach; specialized statistical software may also be needed. This might include hierarchical models like a value-added model (VAM).
- Medium: Some statistical training is necessary, but not necessarily graduate level. This might include simple regression models, z-score transformations, percentile rankings or other relatively basic statistical processes.
- Low: No statistical knowledge or specialized software is needed; a district administrator or educator with training and simple spreadsheet tools should be able to implement the model.

### Teacher and Principal Time Requirements

Some approaches to measuring student performance are highly flexible and customizable at the individual teacher or school level, which may require a greater investment of time both to compute and explain for districts implementing these. Other approaches rely on district-level computations.  Different approaches to measuring student performance can be thought of as falling into two levels with respect to teacher and principal/school time requirements:

- High:  Teachers and principals would be likely to have to establish parameters for the model at the school level; significant training might be required; significant data collection and analysis would be required at the school level.
- Low:  Teachers and principals would need to provide assessment data to the district, but all analysis and computation would be carried out at a district level.

## Four Approaches to Measuring Student Learning Growth

This section provides a short description of four approaches to measuring student learning growth, and shows how they relate to the three issues discussed in the previous section. The approaches discussed are:

- Percent Proficient Models
- Simple Growth Models
- Advanced Statistical Models
- Student Learning Objectives

### Percent Proficient Model

Under the Percent Proficient model, a teacher or school is evaluated based on the share of the students who attain a certain proficiency (or other) threshold. No Child Left Behind's adequate yearly progress requirements for schools are well-known Percent Proficient models.

### Example

To receive college credit on the Advanced Placement United States History Exam, a student must earn a score of 3 or higher. In addition, the district carefully selects students for the course based on their performance in prior social science courses, and so all students in the course can be considered to have similar levels of preparation. Based on these criteria, the district has decided that teachers of A.P. U.S. History will be evaluated based on the percent of students who achieve a score of 3 or higher on the test.

Table 1 summarizes how a percent proficiency model relates to the issues described in the previous sections.

**Table 1. Percent Proficient Model**

| | |
|---|---|
| Question(s) this model can answer: | ▪ How many of my students attained a given level of proficiency? |
| When is this model/question appropriate? | ▪ When a demonstration of a certain level of proficiency is necessary, such as industry certification or a graduation requirement, AND<br>▪ When students enter the course equally well prepared to learn the material |
| Assessment(s) needed: | ▪ Single end-of-course assessment that appropriately measures content and skill taught in the course, along with established proficiency threshold(s). |
| Resource requirements: | ▪ Low. Percent proficiency can be easily computed and does not require large amounts of data or time. |

| | |
|---|---|
| Cautions: | ▪ If students are not equally well-prepared at the start of the course or other factors outside of a teacher's control not addressed, then differences in the percent of students attaining proficiency may not accurately reflect teacher contributions to student learning. |

**Simple Growth Model**

A Simple Growth model compares each student's test scores at two points in time: one score prior to the start of the course or early in the course term, and one score near the end of the course. For example, students' Spring grade 7 math scores might be compared to their Spring grade 6 math scores, or students' Fall grade 4 reading scores might be compared to their Spring grade 4 reading scores. Each student has a certain amount of growth from pre- to post-test, and this information is then summarized for each educator. Each student's growth or the average growth for the teacher's class is then compared to a performance standard, such as the percent of students who attained or exceeded average or median growth.

**Examples**

The district has decided that grade 2 reading teachers will be evaluated using a Simple Growth model based on students' grade 1 and grade 2 Stanford Achievement Test, 10th Edition (SAT-10) reading scores. The scales of the tests are the same, so each student's grade 1 and grade 2 reading scores are directly comparable, and the tests measure a progression of reading skill from grade to grade. Teachers are evaluated based on the share of their students who meet or exceed the average growth within the district. Using historical data for 5 cohorts of students, the district has determined that average growth is 30 points. Each teacher in the district will then be evaluated by the percent of students who grow at least 30 points.

The district has also decided that high school Art Exploration 2 teachers will be evaluated using a Simple Growth model based on students' scores from their Art Exploration 1 and 2 courses. Both courses use performance-based assessments in which students receive rubric-based scores of 1, 2, 3, and 4 (with 4 being the highest). The course content is related, and the assessments and rubrics were developed at the same time by a committee of art teachers. Teachers will be evaluated based on the share of their students who increase at least one performance level, or for students whose Art Exploration 1 scores were 4, who maintain that level.

Table 2 summarizes how a simple growth model relates to the issues described in the previous sections.

**Table 2. Simple Growth Model**

| | |
|---|---|
| Question(s) the model can answer: | ▪ How many students grew by a given amount by the end of the course? |
| When is the question/model appropriate? | ▪ When all students are expected to learn at the same rate, regardless of their initial level of achievement, disability status, English proficiency level, or other student characteristics (e.g. students with high levels of starting performance are expected to grow as much as students with lower levels of starting performance, and vice versa).<br>▪ When the top priority is that all students achieve a minimum amount of growth (and it does not matter if some students far exceed the expected growth, while some barely do). |
| Assessment(s) needed | ▪ Two assessments that measure the same or related content and skills<br>▪ Two assessments that have the same measurement scale (i.e. through equated test forms or a vertical scale)<br>▪ Assessments where students can equally easily demonstrate growth no matter whether they are low- or high-performing (e.g. where a student with a starting score of 20 can grow by 20 points as easily as a student with a starting score of 60, or a student with a starting level of 1 can improve one level as easily as a student with a starting level of 3). |
| Resource requirements: | ▪ Low/Medium. Simple growth can generally be relatively easily computed and does not require large amounts of data or time.<br>▪ Higher data requirements if performance standards will be set using district data only (i.e. if averages based on national or state data are not available). |
| Cautions: | ▪ Simple growth measures require assessment scores that can be directly compared, and where all students can demonstrate similar amounts of growth. If these conditions are not met, then differences in student growth may not accurately reflect teacher contributions to student growth. |

### Advanced Statistical Models

In addition to the Percent Proficient and Simple Growth models, there are a number of more advanced statistical models that can be used to analyze data. These methods can involve statistical techniques such as multi-level modeling and multivariate regression. Examples include:

- Covariate Adjustment Models
    - Student Growth Percentile Models
- Logit Models
    - Ordered Logit Models

Although a detailed description of each of these models is beyond the scope of this document, a brief description of each is provided below.

**Covariate Adjustment Model**

A covariate adjustment model generates teacher scores based on how well a teacher's students perform relative to otherwise similar students. The model predicts each student's outcome test score conditional on variables included in the model, which may include students' prior test score(s) and other factors. Teachers whose students typically score higher than otherwise similar students receive above-average scores, while teachers whose students score lower than otherwise similar students receive below-average scores. For more information on the Covariate Adjustment Model used in Florida for statewide assessments, see the *2012-13 FCAT 2.0 Value-Added Model Technical Report.*[3]

**Student Growth Percentile Models**

A student growth percentile (SGP) model also generates teacher scores based on how well a teacher's students perform relative to otherwise similar students, but uses a different reporting metric. Student performance is typically reported as a percentile rank, and teacher scores as the median of student growth percentiles (SGPs) in his or her class. For example, if a student scores higher than 73 percent of otherwise similar students, the student's SGP is 73. Typically student growth percentile models use a particular type of regression known as quantile regression. For more information on SGP models, see the *Colorado Growth Model Proposal.*[4]

**Logit Models**

Logit models estimate the probability a student will attain a certain threshold or performance level, taking into account student academic history and often other characteristics of the student, the student's peers, and the school. Logit models can be useful when assessment outcomes are expressed as binary (e.g. pass/fail). For more information on logit models, see Guo & Zhao, "Multilevel Modeling for Binary Data," *Annual Review of Sociology, Vol. 26* (2000).

**Ordered Logit Models**

Ordered logit models estimate the probability a student will attain each of a small number of ordered performance levels, taking into account student academic history and often other student characteristics. Where a logit model examines the probability of attaining one outcome, the ordered logit examines the probability of attaining each of multiple outcomes (e.g. a 1, 2, 3, 4, or

[3] http://www.fldoe.org/committees/doc/FloridaComprehensiveAssessmentTestValue-AddedModelTechnicalReport1213.doc
[4] http://www2.cde.state.co.us/schoolview/documents/index_coaypgrowpro.pdf

5 on an AP Exam). For more information on ordered logit models, see Agresti et al., “Random-Effects Modeling of Categorical Response Data,” Sociological Methodology, Vol. 30 (2000).

### Example (Covariate Adjustment Model)

The district has decided that Chemistry 1 teachers will be evaluated using a covariate adjustment model based on students’ earth science and Chemistry 1 end-of-course scores. The content is related, although not directly comparable, nor are the measurement scales directly comparable. At the end of the course, the district uses a covariate adjustment regression model to estimate a predicted Chemistry score for each student in the district based on their prior earth science scores and disability status. Teachers are then evaluated based on the extent to which students in their course outperform otherwise similar students.

Table 3 summarizes how advanced statistical models relate to the issues described in the previous sections.

**Table 3. Advanced Statistical Models**

| | |
|---|---|
| Question(s) the models can answer: | ▪ On average, how much did my students grow, given their starting levels of performance? Did my students grow more than otherwise similar students?<br>▪ On average, how much did my students achieve compared to similar students (with similar beginning performance)? |
| When are these questions/models appropriate? | ▪ When it is important to measure growth or attainment taking into account students’ starting levels of performance and/or other factors relevant to learning |
| Assessment(s) needed: | ▪ Two assessments that measure the same or related content and skills<br>▪ Assessments do not need to have the same measurement scale (i.e. through equated test forms or a vertical scale) |
| Resource requirements: | ▪ High for data and technical capacity. Advanced statistical models generally require significant amounts of data and cannot be easily computed without technical expertise and sometimes specialized software.<br>▪ Low time requirements for educators since all computations are done centrally. |
| Cautions: | ▪ Advanced statistical models require significant amounts of data and statistical expertise to compute scores. If these conditions are not met, then teacher contributions to student growth may not be accurately estimated. |

## Student Learning Objectives

Student Learning Objectives (SLOs) are a process in which educators set growth targets for individual students or groups of students at the beginning of the year, and then assess whether or not students met these targets at the end of the year. The SLO development process typically involves several steps:

- Content standards are reviewed and core concepts are identified for each course.
- A valid and reliable assessment or set of assessments is chosen. Assessments used for SLOs can be fairly standardized – i.e. the same assessment can be used for all students in that course to ensure that achievement is measured uniformly across all students taking the course – or can be more customized, such as at the school or even classroom level.
- The students for whom the SLOs apply are identified and the interval of instruction, including the point at which data on student growth is collected, is established.
- Working together, teachers and administrators set growth targets for individual students or groups of students and document the rationale for these targets. Again, the method to select growth targets can be fairly standardized – i.e. the same process or targets can be used for all students in a course – or it can be more locally driven. Targets should reflect content standards, and rationale for targets might include measures of student knowledge and ability prior to the start of the course, and can include independent factors—such as language fluency or disabilities—affecting the rate at which students learn, and achievement of similar students in prior years.
- Teachers and administrators develop performance standards for evaluating teachers based on student success at meeting growth targets. As with the other steps in the process, this step, too can be standardized at the district level or carried out at the school or even classroom level.
- At the end of the relevant interval of instruction, assessment data are collected and scored and teacher ratings based on SLOs are prepared. This information is then combined with other information for a teacher's overall rating.

It is important to note that SLOs can make use of any of the approaches to measuring student learning growth already discussed (percent proficient models, simple growth models, and advanced statistical models), so long as that information is used to set targets for individual students at the beginning of the course. In some of the other approaches, determinations about student performance standards or targets (or educator performance standards) might not be made until the end of the course, when data for all assessments was available.

### Example

Ms. Brown teachers AP Computer Science A at Washington High School. Over the past 10 years, Ms. Brown has noted that students who do very well in Algebra 2 typically do well in her Computer Science A course. All of the students in her course this year took the district's Algebra 2 end-of-course exam last year, which is scored with a range of 0-300. Students entering the Computer Science course must have at least a score of 200. Using this information, Ms. Brown and Washington High's principal agree that Ms. Brown will be evaluated based on the share of students who score at different levels on the AP Computer Science A exam. Specifically, Ms. Brown and her principal set the following targets:

| **Algebra 2 Score** | **AP Computer Science Exam Target Score** |
|---|---|
| Above 250 | 4 or 5 |
| 200-250 | 3 |

Her growth rating will be based on the percent of students who meet their targets. Table 4 summarizes how SLOs relate to the issues described in the previous sections.

**Table 4. Student Learning Objectives**

| | |
|---|---|
| Question(s) the models can answer: | ▪ How many students met the growth or achievement targets set at the beginning of the year? |
| When is the question/model appropriate? | ▪ When it may be important to allow flexibility in the approach to measuring growth |
| Assessment(s) needed: | ▪ Depends based on approach to measuring growth used |
| Resource requirements: | ▪ Typically low for data and technical capacity.<br>▪ Typically high time requirements for educators since they are often involved in gathering data to develop targets, developing targets and documenting their rationales, and assessing if targets are met, in addition to possibly designing and administering assessments. |
| Cautions: | ▪ To be comparable across a district, SLOs require significant investment in developing common resources and processes. |

## Conclusion

Table 5 summarizes considerations for districts in selecting an approach to measuring student learning growth.

**Table 5. Summary of Considerations for Selecting an Approach to Measuring Student Learning Growth**

| *Choose This Approach When…* | **The Question to Be Answered Is….** | **You Have Assessments With These Characteristics….** | **The Amount of Data Available Is…** | **Your Statistical and Technical Capacity Is…** | **The Burden and Autonomy to Be Placed on Schools/Teachers Can Be….** |
|---|---|---|---|---|---|
| **Percent Proficient** | How many students met minimum proficiency at the end of the course? | One assessment that can measure proficiency at end of the course | Low | Low | Low |
| **Simple Growth** | How much did students grow by the end of the course? | Multiple assessments that are directly comparable in terms of content/skills and scales | Low | Low/Medium | Low |
| **Advanced Statistical Models** | How much did students improve or grow by the end of the course, given where they started? How much did students achieve by the end of the course, given where they started? | Multiple assessments that may not be directly comparable (but are related) in terms of content/skills and scales | High | High | Low |
| **SLOs** | How many students achieved a growth or achievement target *set at the beginning of the year?* | Any of the above, depending on the method used to set growth targets | Low | Low | High |

For each course, districts must first identify what the expectation is for the teacher with respect to student learning. Is the teacher responsible for helping all students attain proficiency, regardless of their level of preparation? Or is the teacher responsible for helping all students learn as much as possible? Making a clear statement about what a teacher is responsible for in a given course will help identify which models can most appropriately be used to measure how well teachers meet this responsibility.

To identify teachers' contributions to student learning growth, good measures of student learning are needed. The second issue a district will need to consider in measuring student learning growth is what assessment(s) are available (or can be made available) at each various points in time, and what type of data these assessments produce.

Finally, the district will need to consider the capacity and resources that are required to implement each of the models described above. There are several elements to consider with respect to capacity and resources, including the amount of data required, the statistical/technical capacity of the district, and the burden implementing the model places on teachers and principals.

The objectives of the course, availability of assessments, and the capacity and resources of the district may have implications for how best to approach measuring student learning for educator evaluation in different courses. Selecting a final model for each course may depend on an iterative process of examining and revisiting all three issues.

## Appendix A. Considering Educator Performance Standards

There are two broad categories of performance standards that districts can consider when thinking about how to classify teacher performance – relative and absolute.

A *relative* performance standard involves determining teachers' performance levels on the basis of their relative position in the overall distribution of teachers.

An *absolute* performance standard for teachers involves comparing a teacher's student's performance with an established policy-relevant benchmark and classifies teachers into performance levels on the basis of that comparison.

A relative standard might compare each teacher to all other teachers in the district: "above average", "average" or "below average. An absolute standard might compare the performance of each teacher's students to an absolute standard: "greater than 100 points' growth, on average," "about a 100 points' growth, on average," or "less than 100 points growth, on average," for example.

The choice of absolute or relative standards should reflect the mechanisms by which the district's evaluation system is intended to function. That is, districts should clearly define the intended outcomes of setting performance standards so that they can determine whether relative or absolute performance standards will more easily lead to those outcomes. For example, if a goal is to recognize a certain proportion of high- or low-performing teachers each year, a relative approach might be best. If, on the other hand, the district wants to recognize only teachers whose students are making sufficient progress to meet or exceed a given standard (and the number of teachers identified does not matter), an absolute approach might be preferable. However, it is important to keep in mind that since student performance data will likely be combined with other data, the performance standard used for student performance data will not be the only factor which must link to the desired outcomes.

### Taking measures of uncertainty into account in setting standards

All student performance measures contain some degree of uncertainty. Uncertainty can arise from chance factors influencing student performance not accounted for by the model, chance factors related to the set of students the teacher is serving, or important changes in teacher performance. When developing performance standards, it is important to incorporate this uncertainty into the design and implementation of the standards so that the classification of teachers into performance categories isn't random, unfair or arbitrary.

Districts can manage uncertainty in several ways:

a) Through the choice of numbers of performance levels (with fewer performance levels, there may be fewer chances for misclassification). Although evaluation systems are required by law to differentiate among four different levels of performance, the student performance component could have more, the same, or fewer performance levels.

b) By using certainty criteria to establish performance standards

c) By considering the use of multiple estimates of teacher performance over time to set performance standards

### *Confidence Intervals*

Estimates of teacher effectiveness are often imprecise. A teacher's score might be about average, but because the score is only an imprecise estimate of the teacher's true contribution to student learning, a teacher who appears average might in truth be well above or below average.

Ideally, student growth models provide an estimate of each teacher's student growth together with an estimate of the level of uncertainty in this measure. Such an estimate of uncertainty is typically presented as a *standard error*. These standard errors can be used to create confidence intervals around a teacher's score, so we can express a level of certainty in the teacher's score. For example, if we set a 95 percent confidence interval, we can be 95 percent certain that a teacher's true score falls within that confidence interval.

Districts can use standard errors in setting performance standards by establishing a confidence interval, so that a teacher would only be classified as above or below a given score only if that conclusion were supported with a given level of confidence. For example, a district could set a 75 percent confidence interval, and even a teacher whose observed score exceeded the established criterion score—above average, for example—would not qualify for the performance level unless the confidence interval implied that we were 75 percent certain that the teacher's true score was above the criterion score.

Similarly, one might identify a second group of teachers whose performance is clearly below the criterion, with the established confidence interval. Teachers for whom a judgment about whether or not their score was truly above or below the criterion score could not be made with the desired level of certainty would be considered neither above nor below the standard (or "similar" to the criterion score).

Confidence intervals can be used with relative or absolute performance standards. Districts could also use multiple confidence intervals—e.g. 70 percent and 95 percent, so that a teacher whose score was above a particular cut score with more than 70 percent confidence but less than 95 percent confidence could fall into one performance level, while those above 95 percent could be in another.

The larger the confidence interval, the more certain we are that a teacher's true score falls within the confidence interval. The larger the confidence interval, the more likely we are to label as average a teacher who is truly well below or well above average. The smaller the confidence interval, the more likely we are to label as well above or well below average a teacher who is truly average.

### *Using estimates of teacher performance for multiple years*

Estimates of teacher effects become more stable and possibly more reflective of true teacher performance as the sample size upon which they are based increases. Knowing this, and in cases where standard errors may not be available, districts can consider using estimates from multiple years to increase precision in classification. This could be done in two ways: by setting a performance standard that is based on multiple years of data, or by creating an overall performance standard over a given period of time that is based on a sequence of yearly performance standards.

Setting a performance standard based on multiple years of data means scores must be aggregated over time in some way.  For example, assuming the scores were comparable or had been converted to a common metric, a district could average value-added scores across three years. The district could use an average with equal weights, an "information weighted" average of the scores (which gives the most weight to the scores with the smallest standard errors - the most precise estimates), or a student-weighted average (which gives more weight to the scores with higher numbers of students included).

Districts could also consider two-year or three-year rolling averages (so that a teacher's score is continually "refreshed" as new data become available) as the basis for setting or applying a performance standard.  The clear disadvantage of using multiple years of data to classify teachers according to a performance standard, or to set a performance standard, is that there will be many teachers do not have the requisite years of data needed (teachers new to the classroom or district, for instance).

An alternative way districts can take multiple years of data into account is to set a performance standard for each year, classify teachers based on each year's score, and  then use the sequence of scores to determine a final performance level. For example, a district could determine that for a teacher to be considered exemplary, she must be in the top quartile of teachers for three years. Or a district could determine that a teacher whose score exceeds the average with 95 percent confidence for two consecutive years is highly effective.

## Appendix B. Possible Resources on Assessment Quality

Rabinowitz, et al. (2013). *Choosing Assessments for Measuring Growth.* CSAI-WestEd. Available at: http://scee.groupsite.com/page/webinars#nov1web

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems* (extended version). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Available at http://www.cse.ucla.edu/products/policy/shortTermGrowthMeasures_v6.pdf

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association. Retrieved from http://teststandards.org

## Appendix C. Sample District Scenario

### Flower District: An Example

Flower District has convened a task force of social science teachers in the district to consider how to measure student learning growth for educator evaluation. The district has already determined that student growth will constitute 50 percent of an educator's evaluation, and that educators should be evaluated on student growth in courses which constitute the majority of their teaching assignment. Looking across teaching assignments in the district, the task force determines that they will need to develop an approach to measuring student growth for the following courses:

- 7th grade social studies
- 8th grade social studies
- U.S. History
- World History

The task force decides to begin with 7th grade social studies.

### *Objectives*

The task force first considers the course objectives. Some in the district argue that the primary goal of the course should be to help as many students as possible to attain proficiency. Others point out that achievement levels in the district are already very high, and that the primary goal of the course is for all students to learn as much as possible, regardless of their starting point. The task force realizes that by setting the goal to maximize the number of students who attain proficiency, the district will give teachers an incentive to direct more effort toward helping students near the margin of proficiency. The task force worries that if the goal is set to maximize the number of students who attain proficiency, teachers may overlook students whose chances of attaining proficiency is very low. The task force decides that their main objective is to maximize student achievement. That is, they want educators to focus on moving each student to a higher level of achievement. This decision eliminates Percent Proficient models from consideration, leaving the district to focus on Advanced Statistical Models and Simple Growth models (where the focus is on changes in achievement levels, not on overall score improvements).

### *Available Assessments*

The next issue the task force discusses is available assessments. There is currently no single end-of-course exam for 7th grade social studies that is implemented across the district. At some school sites, teachers require students to produce a capstone project based on guidelines she has developed for her course; at other sites, all 7th grade social studies teachers collaborate to create a standard 7th grade social studies end-of course exam for all students in the school. Just as there is no existing district-wide end-of-course exam for grade 7 social studies, there is similarly no end-of-course exam for grade 6 social studies. The task force believes that grade 6 social studies content is related, but not directly comparable to the content and skills taught in grade 7 social studies. They like the idea of a capstone project that includes a demonstration of knowledge and communication about a social science topic. They think it may be possible to develop a common rubric for 6th, 7th, and 8th grade that describes student performance levels (based on the knowledge, reasoning, and communication skills). The task force decides to assemble teacher-leaders from each school in the district to develop a such a rubric and guidelines for a major end-of-year project for middle grades social studies.

### *Capacity and Resources*

Having eliminated the Percent Proficient Model, the task force compares the resource requirements of the 3 remaining models. Advanced Statistical Models require large amounts of data and statistical expertise. The task force is not concerned about the data requirements, as Flower District is large and all students in the district are required to take social studies in 6th and 7th grade. However, Flower District does not currently have the statistical expertise necessary to develop and implement some of the Advanced Statistical Models, and the district would prefer not to devote resources to hiring an outside contractor to develop and implement the models.

The task force then considers a Simple Growth approach. The task force knows it does not have the skills or resources necessary to build assessments whose scales are directly comparable. But, as discussed, the members do believe that a major research project would be a good assessment of student knowledge and skills in middle grades social studies, and they believe can develop a rubric that can be used consistently to score these projects. The task force realizes they will need to gather examples of previous student projects and have multiple teachers score them as a starting point, and members also discuss ways to have teachers come together at the end of the year and score projects together. Teachers will then be evaluated based on the share of their students who increase their performance levels in each rubric area from year to year. The task force realizes this will impose a burden on teachers to develop standard guidelines and rubrics and devote time to scoring, but members believe this is also a good professional development opportunity and they are willing to reallocate some professional development funding for this purpose.

### *Decision*

Ultimately, the task force decides on a simple growth approach using changes in performance levels. A committee of grade 6, 7, and 8 social studies teachers representing schools across Flower District will be assembled to develop project guidelines, scoring rubric, and annotated samples of previous student projects. The task force recognizes that they are asking a lot of their participants and of teachers, but believes this approach is the best choice for Flower District.

## ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

## LOCATIONS

### Domestic

Washington, D.C.

Atlanta, GA

Baltimore, MD

Chapel Hill, NC

Chicago, IL

Columbus, OH

Frederick, MD

Honolulu, HI

Indianapolis, IN

Naperville, IL

New York, NY

Sacramento, CA

San Mateo, CA

Silver Spring, MD

Waltham, MA

### International

Egypt

Honduras

Ivory Coast

Kyrgyzstan

Liberia

Tajikistan

Zambia

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY 877.334.3499

**www.air.org**

*Making Research Relevant*

07/2014