NATIONAL COMPREHENSIVE CENTER FOR TEACHER QUALITY

Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects

MARCH 2011

Research & Policy Brief

ACKNOWLEDGMENTS

The TQ Center would like to thank the following individuals, who were instrumental in the development of this brief: Bill Slotnik, Community Training and Assistance Center; Margaret Heritage, Ph.D., National Center for Research on Evaluation, Standards, & Student Testing; Megan Dolan, Mid-Atlantic Comprehensive Center; Bonnie Billingsley, Ed.D., Virginia Tech; and E. Caroline Wylie, Ph.D., ETS.

Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects

This **Research & Policy Brief** was developed to help states consider options for assessing student learning growth for the majority of teachers who teach content not assessed through standardized tests.

March 2011

Laura Goe, Ph.D., *ETS*
Lynn Holdheide, *Vanderbilt University*

NATIONAL COMPREHENSIVE CENTER FOR TEACHER QUALITY

CONTENTS

Introduction 1
Nontested Subjects and Grades 1

Measuring Growth 2
Why Measure Growth? 2
How Is Growth Measured? 3
Federal and State Priorities 4
Expectations for Teachers 5
Attribution and Student–Teacher Links 5

Factors for Consideration 7
Student Competencies in Specific Content Areas and Grade Levels 7
Identification of Reliable and Valid Assessments 7
Schoolwide Value-Added Models for Teachers of Nontested Subjects and Grades. 9
Measuring Student Learning Growth for Teachers in the Arts and Other Nontested Subjects. . . 13
Measuring Student Outcomes for “Caseload” Educators 13
Alignment With Federal Priorities 14
Application to All Grades and Student Populations 15

Standardized Evidence Collection 16
Measures That May Improve Teacher Performance 18

State Guidance to Districts 18
Comparability: Across or Within Districts? 18
Measures 19
Exceptions 19
Ongoing Research on Systems, Models, and Measures 21
Considerations for States: Moving Forward 21

Conclusion. 22

References 23

INTRODUCTION

The growing need for more information about measuring teachers' contributions to student learning growth, particularly in nontested subjects and grades, is the impetus for this Research & Policy Brief. Although the research base in this area is disappointingly limited, the brief includes considerations and suggestions based on current models and experiences from the field. Although the brief is intended for use by states in developing statewide systems and providing guidance to districts, it also may be helpful to districts charged with designing and implementing evaluation models that fit within state and federal guidelines.*

For many states, the need to implement comprehensive teacher evaluation systems that consider teachers' contributions to student learning growth is clear and immediate. But because there are no research-based models for incorporating this component into teacher evaluation systems, states are experimenting with a variety of strategies to move forward. In fact, even without research to support particular approaches to evaluating teachers' contributions to student learning growth, states are proceeding—sometimes on very short timelines—to collect such evidence and incorporate it into a system of multiple measures of teacher performance. This endeavor is challenging even when there are standardized test scores that can be used as evidence of students' achievement progress, but it is especially complicated when no standardized measures exist, which is the case for the substantial percentage of teachers of nontested subjects and grades.

This Research & Policy Brief provides information about options for states to explore as well as factors to consider when identifying and implementing measures. The brief also focuses specifically on federal priorities to help ensure that evaluation systems meet the high expectations set for teacher evaluation. Finally, the brief emphasizes the importance of fairly measuring *all* teachers, including them in the evaluation process, and ensuring validity in measurement.

Nontested Subjects and Grades

In *The Other 69 Percent: Fairly Rewarding the Performance of Teachers of Nontested Subjects and Grades* by Prince et al. (2009), "the other 69 percent" refers to the percentage of teachers whose contributions to student learning cannot be measured with test-based approaches (e.g., value-added models) because they teach subjects or grades that are not assessed with standardized tests.

Measuring effectiveness for the "other 69 percent" is probably the most challenging aspect of including student achievement growth as a component of teacher evaluation. According to Prince et al. (2009),

> Identifying highly effective teachers of subjects, grades, and students who are not tested with standardized achievement tests—such as teachers of art, music, physical education, foreign languages, K–2, high school, English language learners, and students with disabilities—necessitates a different approach. It is important that states and districts provide viable options for measuring the progress of these groups of students and the productivity of their teachers, both of which contribute to school performance. (p. 1)

Statewide standardized testing is typically conducted for reading/language arts and mathematics in Grades 4–8 as required by the Elementary and Secondary Education Act (ESEA), as reauthorized by the No Child Left Behind Act. Likewise, some states, albeit a smaller number, conduct such testing in certain grades for other subjects such as science

* See http://www.tqsource.org/webcasts/201012Workshop/Teacher_Effectiveness_Workshop_Glossary.pdf for a glossary of commonly used terms in current teacher evaluation reform efforts.

and social studies. Nontested subjects and grades in which standardized tests are not administered include the following:

- Subjects with standards that cannot be adequately or completely measured with a paper-and-pencil test (e.g., art, music, industrial arts, drama, dance)
- Subjects in lower elementary grades for which students cannot be reliably tested with paper-and-pencil or computerized tests (e.g., Grades K–2)
- Subjects/grades for which states have chosen not to test because of cost and priority relative to "core" academic subjects

In addition to nontested subjects and grades, there are certain student populations and/or situations for which standardized test scores are not available or utilized (e.g., students with cognitive disabilities). The Individuals with Disabilities Education Act of 2004 allows for the use of alternative assessments for students for whom the standardized assessment is inappropriate, even with reasonable accommodations. Moreover, smaller teacher caseloads for some student groups, such as students with disabilities and English learners, produce results that are statistically less reliable, often resulting in such groups being excluded in value-added or other growth models (Amrein-Beardsley, 2008; Feng & Sass, 2009).

Inclusion of teachers in nontested subjects and grades in an evaluation system that is based in part on teachers' contributions to student learning growth requires the identification or development of appropriate measures and methods to accurately determine students' growth toward grade-level and subject standards. Clearly, this task requires standards for every subject and/or grade level. If standards are nonexistent or poorly specified, it will be difficult to accurately determine teachers' contributions toward growth in those subjects and grades, so ensuring that academic standards exist for every subject and grade should be a priority.

MEASURING GROWTH

Why Measure Growth?

Teachers are the most influential school-based factor on student achievement (Rivkin, Hanushek, & Kain, 2005; Sanders & Horn, 1998; Sanders & Rivers, 1996). Although studies have shown that some teachers are more effective than others at helping their students achieve at high levels, most indicators of teacher quality (e.g., credentials, characteristics, and observable practices) are generally poor predictors of student learning growth (Goe, 2007; Rice, 2003; Wayne & Youngs, 2003). Teachers' scores on observation instruments have not been highly correlated with student learning growth (Weisberg, Sexton, Mulhern, & Keeling, 2009). However, it is not surprising that correlations are weak when the factors to be measured with observations are not well specified or when raters are poorly trained or inadequately monitored for scoring consistency after training.

Most of the indicators used in the past to determine teacher quality have been found to be inadequate, particularly when used in isolation, in differentiating between teachers whose students perform well and those whose students are not making adequate progress. Recent federal funding opportunities have emphasized teacher effectiveness and teacher evaluation based on teachers' contributions to student achievement. This focus on evaluating teachers by measuring *student growth* rather than attainment is fairer to teachers whose students enter classrooms well below grade level. Teachers should not be penalized for choosing to teach in schools in which students are considerably behind their peers in proficiency. This is not to say that students' mastery of appropriate grade-level standards is unimportant, but moving students as close as possible to proficiency, even if all students are not able to reach it, should be the focus of teachers' efforts. Teachers should be given

credit when these efforts succeed, and using multiple measures of student learning growth is essential to ensure that teachers in all subjects and grades are fairly credited.

How Is Growth Measured?

Since the initial passage of ESEA, standardized assessments have been used to determine student progress toward academic standards. Value-added models and other growth models have generated considerable interest for showing growth over time for students, and lately, for the teachers of those students. Recent efforts to create statewide longitudinal data systems that link teachers with their students' achievement have set the stage for states and districts to use student learning growth on standardized tests as part of determining teacher effectiveness. However, in most states, only reading/language arts and mathematics in Grades 4–8 are actually tested with state standardized assessments, meaning that teachers in most subjects and grades do not have state standardized test results that can be used as components of teacher evaluation.

How results from standardized tests are actually used as part of teacher evaluation remains an open question because states and districts are just beginning to use linked student–teacher data and growth models, (e.g., value-added models). Tennessee is at the forefront of these efforts because it has been using the Tennessee Value-Added Assessment System (TVAAS) for more than a decade to provide individual teachers and their principals with the teachers' district rank based on value-added measures. Many more states are developing systems that will allow them to use growth models such as EVAAS (the version of TVAAS that is not state-specific) as well as the Colorado Growth Model, which focuses on students' growth toward proficiency (See "Different Approaches to Measuring Students' Growth"; Betebenner, 2008).

DIFFERENT APPROACHES TO MEASURING STUDENTS' GROWTH

Although most teachers currently cannot be evaluated with growth models based on standardized tests, it may be helpful to understand how growth models might fit within an evaluation system. A number of states are planning to implement (or already have implemented) value-added or other types of growth models. In its simplest form, the value-added measure as it is used for evaluating teachers is calculated as follows: Students' previous test scores are used to create predicted test scores for a given year. The difference between the predicted and actual test scores are growth scores. Teachers' contributions to student learning are determined by calculating the average of all of their students' growth scores. The teachers are then ranked with other teachers within a district (or other unit of interest) according to how much they contributed to student growth, and this ranking is their value-added "score."

In some value-added models, only students' prior achievement scores are used in the calculation; other models include students' gender, race, and socioeconomic background; still others include information about teachers' experience. With a value-added measure, teachers whose students performed as well as predicted are considered "average" teachers; those whose students performed better than predicted are considered "above average" or "highly effective"; and those whose students performed worse than expected are considered "below average."

The Colorado Growth Model focuses instead on student growth percentiles. Students are compared with their academic peers (i.e., students at the same starting point in achievement) to determine normative growth. The goal is to determine students' standing relative to their academic peers. Thus, if students' scores are better than those of their academic peers, they are performing well. All of a teacher's students can be scored in this way, resulting in an average growth for the class or the teacher's roster, which can then be attributed to the teacher's efforts in much the same way value-added scores are.

Whenever such models—whether value-added models, the Colorado Growth Model, or other models—are used, results should never be considered in isolation as the sole measure of a teacher's performance but rather included in a system of multiple measures that produces a comprehensive picture of a teacher's performance.

However, results obtained through such growth models have rarely—until now—been used as part of teacher evaluation. Even in those states that have the capacity to collect such information, questions remain about the accuracy of the information, given evidence of year-to-year fluctuation in teachers' scores (Braun, Chudowsky, & Koenig, 2010; Koedel & Betts, 2009; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Schochet & Chiang, 2010).

For teachers in nontested subjects and grades, there are few state models that demonstrate how contributions to student learning growth can be systematically measured and analyzed in ways that allow for differentiation among teachers. Some experiments are currently under way in collecting evidence of student learning growth for these teachers, but research has not yet been conducted on how such evidence is being used within evaluation systems.

Federal and State Priorities

To position themselves for a successful Race to the Top bid, many states passed new legislation mandating that student achievement growth be included as part of teacher evaluation. Federal priorities (Secretary's Priorities for Discretionary Grant Programs, 2010) specify that acceptable measures for determining teachers' contributions to student learning must meet the following requirements:

- Rigorous
- Between two points in time
- Comparable across classrooms

These terms are not explicitly defined in Race to the Top guidance. In fact, the federal government has declined to offer definitions for these terms, preferring instead to encourage states to define them locally. For federal purposes, Race to the Top winners must follow through with what they promised in their plans, which may include defining terms. The following considerations may provide some assistance in the development of state definitions:

- ***Rigorous*** measures may exhibit high expectations for student progress toward college- and career-readiness. In other words, an assessment that measures student progress in social studies would be designed to measure students' mastery of grade-level standards for that subject. Thus, a student who does well on such an assessment should be on track to successful, on-time promotion to the next grade and ultimately to graduation.
- ***Between two points in time*** may mean assessments that occur as close as possible to the beginning and end of a course so that the maximum growth toward subject/grade standards can be shown.
  - *Example:* An Advanced Placement (AP) test may serve as an end point, but another assessment (aligned with the state standards and focused on the specific knowledge and skills measured by the AP tests) will likely need to be administered at the beginning of the year to establish students' level of mastery of the standards when they begin the course to determine teachers' contributions to student growth. The process of collecting evidence of students' initial skills and knowledge should not be undertaken lightly. Ideally, an assessment that has been designed and created by experts specifically to serve as a pretest should be used.
  - *Example:* Student portfolios representing mastery of standards could be collected at the end of the year. However, at the beginning of the year, teachers would need to collect and score evidence (i.e., activities or assessments aligned with the state standards and focused on

the specific knowledge and skills needed for creating a successful portfolio) that would allow them to formulate an initial score point for each student. Through this process, increased knowledge and skills could be documented for individual students.

- ***Comparable across classrooms*** has two possible interpretations, both of which are useful to consider:
  - The measures used to show students' growth for a particular subject are the same or very similar across classrooms within a district or state.
  - The measures used in *nontested* subjects and grades are as rigorous as those in tested subjects and grades. In other words, measures used to document student learning growth in art, music, and social studies must be as rigorous as those for student learning growth in reading/language arts and mathematics.

Expectations for Teachers

Race to the Top defined an *effective teacher* as one whose students achieved at least one grade level of academic growth during the course of the year and a *highly effective teacher* as a teacher whose students achieved at least one and a half grade levels of academic growth during that time frame. Although not federally mandated, teachers are generally required to ensure that their students are on track to meet grade-level expectations. In addition, they are expected to regularly evaluate student progress and issue grades that reflect students' efforts and achievement in mastering the content. With new federal and state mandates calling for the inclusion of teachers' contributions to student learning in the evaluation process, growth must be documented in some way, which means that teachers in nontested subjects and grades need to focus on new approaches to measuring their students' progress—approaches that are rigorous, that provide data on growth between two points in time, and that are comparable across classrooms.

Attribution and Student–Teacher Links

Determining teacher attribution for particular students is challenging. What if a student receives services in a general education classroom in which coteaching occurs? Should both teachers be held accountable for student growth? How will paraprofessionals' contributions to student learning growth be sorted out from those of the content area or special education teachers?

In a recent TQ Center inquiry, 85 percent of the local and state special education administrators polled were of the opinion that both the general and special education teachers should be held accountable for all students in the class (Holdheide, Goe, Croft, & Reschly, 2010). However, there may not be widespread agreement for that approach. Linking student growth (or a portion thereof) to the appropriate teachers presents challenges.

One approach developed by the Ohio-based Battelle for Kids is the use of new linkage software that has the capacity to account for student mobility and shared instruction/coteaching in subject areas for which value-added data are available (See "Student–Teacher Linkage for Attribution"). This approach also may be viable using other types of student growth measures, as it facilitates a deeper and often necessary discussion regarding teacher roles and responsibilities. At this time, however, a research-based methodology for this type of teacher-led determination has yet to be established. In addition, its application in a non-value-added growth measure needs to be explored.

Teacher apprehension toward accountability systems including student growth measures can be minimized if teachers perceive the system to be fair and accurate. For example, failure to directly address which teachers are accountable for which students will likely result in pushback from teachers. In addition, teachers need to have an opportunity to verify their rosters of students and the length of time that students were on their rolls. This verification process is particularly important in schools with high rates of absenteeism or student mobility. Teacher involvement and support in this process is essential. Teachers must be involved in the processes of problem-solving, collecting data during implementation, and obtaining feedback on effectiveness. Teachers know their classrooms, their students, and the way in which they collaborate with other teachers.

STUDENT–TEACHER LINKAGE FOR ATTRIBUTION

Olentangy Local School District in Ohio and other districts across the country are taking value-added analysis to the classroom level with Battelle for Kids' innovative, Web-based BFK·Link™ solution to accurately "link" teachers to students. During the linkage process, teachers review and correct data used for teacher-level measures of effectiveness, including value-added analysis, by ensuring that all students taught are "claimed" by teachers for all subjects, accounting for student mobility and shared instruction/coteaching.

The BFK·Link process attempts to maximize correct matching of teacher effort to student outcomes through a transparent process. For example, for teachers working in a true coteaching situation, both teachers may each "claim" 50 percent of each student. Or, if students receive some support services in a resource room, the general educator may claim 70 percent while the special education teacher claims 30 percent. Student standardized test scores are then linked with teachers for the percentages specified.

In typical classrooms, teachers claim 100 percent of most of their students, with reduced percentages for students with special needs who receive services from other teachers. The system verifies accuracy by marking cases in which a student has more or less than 100 percent for inspection (i.e., more than one teacher is contributing to that student's scores, but the teachers' combined percentages do not add up to 100), and the teachers are asked to reevaluate. When percentages add up to 100 percent, the BFK·Link solution calculates scores proportionally.

The use of value-added analysis to inform instruction and high-stakes decisions requires accurate linkage of teachers to students. For more information, see *The Importance of Accurately Linking Instruction to Students to Determine Teacher Effectiveness* (Battelle for Kids, 2009), a white paper commissioned by the Bill & Melinda Gates Foundation.

FACTORS FOR CONSIDERATION

States and districts attempting to incorporate student growth into their teacher evaluation systems are faced with the challenge of identifying other valid and reliable measures for teachers of nontested subjects and grades. Though the research base is still developing, the following questions may be useful to consider during the problem-solving process:

- Is there a consensus on the competencies students should achieve in this content area?
- What assessments/measurements can be used to reliably measure these competencies with validity?
- Should the use of schoolwide value-added models be considered as a means to measure student progress in nontested subjects and grades?
- How will growth in performance subjects (e.g., music, art, physical education) be determined?
- How will related personnel ("caseload" educators) be factored into the system?
- Do these measurements meet all of the federal requirements (i.e., rigorous, between two points in time, and comparable across classrooms)? Are measurements aligned with federal priorities?
- Can these measurements be applied to all grades and student populations?

Student Competencies in Specific Content Areas and Grade Levels

In most states, content standards are designed by a group of experts and practitioners to encourage proficiency for every student by defining the knowledge, concepts, and skills students should acquire for each subject. Each standard typically has clearly defined statements and examples of what all students should know and be able to do at the end of a particular grade. These standards often drive changes in certification, assessment, curriculum, instructional strategies, and teacher professional development. Therefore, a transparent alignment to these content standards offers guidance when identifying and/or designing assessments to measure student progress, which could be used to determine teachers' contributions for evaluation purposes. In states in which subject content standards exist, these standards provide a basis for the identification and development of assessments.

Identification of Reliable and Valid Assessments

States are struggling most with determining appropriate measures for evaluating teachers' contributions to student learning growth in the nontested subjects and grades. The challenge facing many states, including the Race to the Top award recipients, is to identify valid, reliable processes, tools, assessments, and measures that allow them to collect data to measure every teacher on his or her contributions to student learning growth. Many current approaches to measuring teachers' contributions to student learning in the nontested subjects and grades do not meet all of the federal criteria of rigor, comparability, and growth measured across two points in time.

Local and state education systems have taken various approaches, each of which has its own strengths and limitations as indicated in Table 1. None of these options is "perfect," and concerns about validity, reliability, and costs are associated with nearly all of them. The trade-offs involved with using these measures should be considered by stakeholder groups as well as state and district evaluation and assessment personnel.

Table 1. Options for Measuring Student Growth to Inform Teacher Evaluation in Nontested Subjects and Grades

Option for Measuring Student Growth for Teacher Evaluation	Strengths of This Measure	Limitations of This Measure
Use existing tests designed for other purposes, such as end-of-course tests that may be included with some curriculum packages.	Tests developed by the creators of the curriculum are likely to be aligned well with the content of the course. It may be possible for the creators of the curriculum to develop appropriate pretests if they are not included in the package.	Validity is a concern whenever a measure is used in a way that was not intended by the maker of the assessment (e.g., turning end-of-course assessments into pretests). Discussions with the test maker about using tests for other purposes may provide insight into how validity may be affected.
Create new tests for areas in which few assessments exist.	Tests can be developed in alignment with specific grade/subject standards.	This option is a costly undertaking, given how much effort goes into developing valid and reliable tests that can accurately measure students' knowledge and skills based on a set of subject/grade standards. Paper-and-pencil tests may not be appropriate as the sole measure of student growth, particularly in subjects requiring students to demonstrate knowledge and skills (e.g., art, music).
Use the four Ps—portfolios, products, performances, or projects—to measure student growth over time for subjects in which standards require students' to demonstrate mastery.	Evidence about student growth in particular knowledge and skills can be documented over time using performance rubrics. Portfolios and projects reflect skills and knowledge that are not readily measured by paper-and-pencil tests.	Training would be required for everyone involved in using rubrics to ensure reliability (i.e., all raters agree on how the evidence reflects different levels of achievement). Performance ratings are best conducted by groups of raters rather than individual teachers; bringing raters together to examine student work may be a logistical challenge.
Give teachers in nontested subjects and grades a "prorated" score for collaboration with a teacher in a tested subject (i.e., an art teacher collaborating with a mathematics teacher).	No additional resources are required. This option is similar to the Teacher Advancement Program (TAP) model.	Determining prorated scores would be problematic, threatening the validity of the information. Differences among methods of determining contributions of these collaborating teachers may make it difficult to ensure comparability.
Use other measures (e.g., classroom observations) for these teachers.	No additional resources are required.	No information about student achievement is obtained, meaning that this option will not meet federal priorities and many state requirements. Observations and other measures focused on teacher practice offer little information about students' actual achievement in a teacher's classroom.
Use student learning objectives (i.e., the teacher selects objectives and determines how to assess student growth toward meeting objectives).	Teachers benefit from being directly involved in assessing students' knowledge and skills. Teachers can set learning objectives based on students' special needs (e.g., students with disabilities or English learners). This option is applicable to all teachers and subjects.	Comparability across classrooms will be problematic because of teachers' selection of assessments and objectives. This option is very resource-intensive for principals or district personnel who approve objectives, provide teachers with guidance, verify outcomes, and so on.

Schoolwide Value-Added Models for Teachers of Nontested Subjects and Grades

The use of schoolwide value-added scores has been suggested as a way to evaluate teachers in nontested subjects and grades to remedy the lack of available measures. Similar to the Teacher Advancement Program (TAP) model, it is perhaps the least expensive method of including these teachers in a test-based evaluation system because new measures and teacher training are not required. In this scenario, teachers of nontested subjects would be given the schoolwide value-added average in place of individual growth results.

This approach presents some additional challenges for a number of reasons, including questions about rigor and comparability when judgments are made about individual teacher performance based on students they never taught. Furthermore, it is much more difficult to learn about teachers' contributions to student achievement if they are assigned scores based on other teachers' efforts. Mathematics and reading/language arts value-added information will not be useful to teachers in improving their performance in subjects such as art, social studies, and science. In addition, failing to measure progress in these subjects and for certain students devalues the contributions those teachers make to student learning and provides no information about their effectiveness in teaching their subject matter.

Using Existing Assessments

In the search for measures to determine teachers' contributions to student learning growth, it is likely that an iterative process will be needed. After a potential instrument is identified, it is necessary to demonstrate that the measure is valid for the intended purpose (i.e., that the measure does, in fact, differentiate among teachers whose students have high levels of learning growth and teachers whose students' learning did not increase at acceptable levels). Because the measures that might be used for teacher evaluation have not been validated for this purpose, it is important to analyze data collected by using these measures and determine whether the data show differences among teachers and whether results from using these measures correlate with other measures in the evaluation system.

The validation process generally starts with determining the factors that need to be measured and for what purpose. As part of this process, it is important to consider the evidence needed to measure teachers' contributions to student learning growth. Evidence will have been gathered to build a case for using a particular measure as part of the evaluation system (Herman, Heritage, & Goldschmidt, in press). After the types of necessary evidence are determined, measures and instruments that can be used to collect such evidence must be identified. Then, results from using measures must be analyzed to determine how the measures performed in practice.

For example, if the district wanted all Grade 8 reading/language arts teachers to administer an essay to students at the beginning and end of the year to establish student growth, the district would need to score (or preferably have teachers score together) the essays and determine whether they show student learning growth. A distribution of scores would need to be made and cross-referenced with teachers to determine whether more or less growth occurred in particular teachers' classrooms or the pattern of growth is random. A random pattern would suggest that the growth students made was not necessarily attributable to a particular teacher's efforts, whereas a pattern of higher or lower growth associated with a particular teacher may be an indicator of his or her efforts. Comparing these results with results from additional measures (e.g., other assessments, projects, portfolios) should then be helpful in validating the usefulness of the essays in showing teachers' contributions to student growth.

In addition, validity is a matter of degree—it is seldom perfect, but a high degree of validity must be achieved when results will be used for high-stakes purposes such as teacher tenure, performance pay, and dismissal. Clearly, the higher the stakes, the greater validity is needed in terms of the evidence. In addition, validity can be improved over time by identifying which measures are and are not working to provide evidence to make decisions about teacher performance.

For most states and districts, waiting until the measures are perfected may be impractical, given the timelines to implement new teacher evaluation systems. So even though the measures may have weak evidence of validity in the first attempts at implementation, states and districts will benefit from creating a process to continually evaluate and strengthen the measures or eliminate those that continue to show weak evidence of validity. Over time, a collection of measures with strong evidence of validity will be created. Obviously, this process is neither quick nor easy, and it requires some expertise. Districts and states with limited capacity may consider joining forces with others in the region to share resources rather than "reinventing the wheel" in each district or state.

Utilizing existing assessments and avoiding the development of new assessments certainly holds appeal for implementation ease. Interim or benchmark assessments are already widely used in schools as a means to provide assessment of student progress toward content standards. In fact, schools that implement response to intervention (RTI) have likely identified measures for the progress monitoring component of implementation. These assessments are often embedded into the instructional cycle and are used to make the necessary instructional adjustments to facilitate student mastery. Working collaboratively with state and district RTI initiatives to identify potential sources of evidence for evaluation purposes may facilitate a combined effort to address the persistent achievement gaps in schools (See "National Center on Response to Intervention Progress Monitoring Tools Chart").

NATIONAL CENTER ON RESPONSE TO INTERVENTION PROGRESS MONITORING TOOLS CHART

The National Center on Response to Intervention annually publishes a progress monitoring tools chart to assist educators in identifying tools that best meet their needs. The Center's Technical Review Committee (TRC) independently established a set of criteria for evaluating the scientific rigor of progress monitoring tools.

Included in this chart are ratings for instrument reliability of the performance-level score, reliability of the slope, validity of the performance-level score, predictive validity of the slope of improvement, and disaggregated reliability and validity data. In addition, the charts include the standards by which the TRC reviewed each tool (e.g., whether the tool is available in alternative forms, its sensitivity to student improvement, and its ability to measure end-of-year benchmarks).

This chart can be accessed at **http://www.rti4success.org/tools_charts/progress.php**.

Although these existing assessments were not designed specifically to inform teacher evaluation, they may have merit for that purpose. However, it is not as simple as adopting existing assessments. A thorough review of each assessment should be conducted, including its validity in measuring progress on the specific content standards and its measurement reliability across students and teachers. Moreover, assurance that these assessments measure what is valued is essential if evaluation results will be used to make personnel and compensation decisions.

Examples of Approaches to Assessment

Hillsborough County, Florida. Hillsborough County, Florida, a recent Race to the Top award recipient, has taken the approach of developing new assessments specifically designed to assess content mastery and plans to use data to inform teacher evaluation. Each nontested subject will have a pretest and posttest in which student scores are averaged over a three-year period to determine teacher effectiveness. As indicated in Table 1, this approach is fairly time and cost intensive; however, newly developed end-of-the-course assessments are more likely to be readily aligned with the content standards and have the potential to meet two of the federal requirements: *comparability* and *across two points in time*. Compliance with rigor would be dependent on how the data are used to determine acceptable student growth, and therefore, teacher proficiency.

Delaware. The state of Delaware uses a combination of approaches in which existing and new measurements are identified, assessed, and determined to be acceptable by experts at the state level. With the assistance of trained facilitators, Delaware assembled a group of local practitioners, arranged by content area expertise, to conduct a thorough review of existing measurements. After consensus was reached, the group submitted to the state a listing of recommended assessments and/or methods to assess student growth toward the content standards. This listing is updated and shared regularly (after approval from an independent panel of experts).

Austin, Texas. States also may identify specific criteria required for assessments to be considered valid measures of student growth. In Austin, Texas, teachers participating in a pay-for-performance pilot are involved in determining student achievement growth through the development of student learning objectives (SLOs). SLOs are classroom, grouping, or skill-based objectives, and teachers' ability to meet the SLOs determines their level of effectiveness. The quality of SLOs in measuring student growth is established by a rubric that determines whether the objectives and associated assessments are rigorous, measureable, reliable, and valid and whether the projected growth trajectory is considered rigorous. Although this approach facilitates teacher investment in the process, which is a definite strength, maintaining rigor is dependent on the rubric's implementation fidelity among administrators and teachers. In addition, SLO results may not be comparable across classrooms because various assessments are used to establish student growth. Moreover, if the evaluation system includes observations conducted by administrators, the burden on the administrators may be substantial.

For more information about these assessment approaches, see "Practical Examples of State Evaluation Systems."

PRACTICAL EXAMPLES OF STATE EVALUATION SYSTEMS

Hillsborough County Public Schools, Florida

Hillsborough County is the recipient of a seven-year, $100 million Bill & Melinda Gates Foundation grant and has recently been awarded Race to the Top dollars to continue its efforts to improve results through the Empower Effective Teachers (EET) program.

The goals of EET are to:

- Develop a quality induction program for new teachers.
- Improve the teacher and principal evaluation system.
- Enhance the system of professional development.
- Provide effective incentives for teachers and improve the compensation plan.

Hillsborough County uses multiple measures to determine teacher effectiveness including peer and principal ratings using a modified version of Charlotte Danielson's *Framework for Teaching*. Those ratings make up 60 percent of teacher evaluations, with student performance on the Florida Comprehensive Assessment Test or end-of-course examinations making up the remainder.

Hillsborough County's stated commitment is to evaluate every teacher's effectiveness with student achievement growth, even teachers in nontested subjects and grades. To do so, Hillsborough County is in the process of creating pretests and posttests for all subjects and grades, expanding state standardized tests, and using value-added measures to evaluate more teachers.

In the 2010–11 school year, the statewide assessment program began transitioning to assessing student understanding of the Next Generation Sunshine State Standards through the implementation of the Florida Comprehensive Assessment Test® 2.0 (FCAT 2.0) and Florida End-of-Course Assessments.

Information on Hillsborough County's EET program can be accessed at **http://communication.sdhc.k12.fl.us/empoweringteachers/?p=611**.

Delaware

Delaware already had an excellent statewide evaluation system, which required classroom observations and encouraged teachers to focus on school, district, and state goals as well as their own professional growth. Delaware conducted a yearly external evaluation of the system, soliciting feedback from teachers and administrators through surveys, interviews, and focus groups. Revisions were made to the system yearly based on these results. The state also collaborated with the teachers union to ensure that evaluations were fair and responsive to the needs of the teachers and administrators. However, Delaware's system was lacking a mechanism to evaluate teacher contributions to student learning growth.

One reason that the state was awarded Race to the Top funds was the collaborative nature of the proposal, bringing stakeholders to the table at every step. As state staff focused on implementation, they continued to involve stakeholders in each step of the discussions. They valued teacher and administrator input, which was reflected in the steps they took to identify appropriate measures for the nontested subjects and grades as well as additional measures for teachers whose students took the state standardized test. A team of trained facilitators led groups of teachers as they met to discuss measures they currently used to evaluate their students' growth toward grade/subject standards. After discussing the merits of the measures and how they could be used, teachers made recommendations to the state about which measures to include.

The TQ Center and the Mid-Atlantic Comprehensive Center have been partners with Delaware during the implementation of its Race to the Top plans. In addition, Delaware has sought assistance from the Assessment and Accountability Comprehensive Center in convening a panel of experts to evaluate the potential measures for statewide use to show teachers' contributions to student growth in various grades and subjects. This process is ongoing.

Austin Independent School District Reach Compensation and Retention System, Texas

The Austin Independent School District Reach Compensation and Retention System is a four-year pilot incentive pay program for teachers and principals initiated in 2007–08. The program goals are to:

- Ensure quality teachers in every classroom.
- Provide professional growth opportunities.
- Increase retention.

The program focuses on student growth, professional growth, and schools with the highest need. Student growth is measured by student learning objectives (SLOs). Each teacher develops two SLOs—one that targets classroom performance and the other focused on a particular skill or subgroup of students (e.g., students with special needs). Each SLO must be a measureable objective that is approved by the principal. Teachers and principals undergo a series of trainings on how to establish and measure learning objectives.* The SLO's appropriateness, rigor, and acceptability are determined through the use of a rubric that considers the following questions:

- What are the needs?
- What and who is targeted?
- What will students' learn?
- How will you know whether they learned it?
- What is your goal for student achievement?
- How rigorous is your SLO?

Information regarding this system and the rubric can be accessed at **http://www.austin.isd.tenet.edu/inside/initiatives/compensation/releases.phtml**.

*SLOs are used to determine incentives and are not an integral part of the evaluation of teachers at this time.

Measuring Student Learning Growth for Teachers in the Arts and Other Nontested Subjects

Not all standards can be adequately assessed with a multiple-choice test. Many subjects require students to perform or create a product to demonstrate mastery of the standards. For these subjects, one or several of the four Ps (i.e., portfolios, performances, products, and projects) will likely be required to assess music students' ability to play scales on their chosen instruments; art students' ability to create works of art in various mediums; foreign language students' ability to speak the language they are studying; and family and consumer science students' ability to budget, plan, and prepare a wholesome family meal.

For these subjects, the focus is on designing appropriate tasks (e.g., performance, activities) that demonstrate students' mastery of standards and then developing appropriate pretests that allow districts/schools to determine students' knowledge and skills at the *beginning* of the course. In some cases, students can perform the *same* task: music students' can play the same piece of music at different points in time to show progress; art students can draw a still life; drama students can perform a monologue; and so on. In other cases, it may not be feasible for students to perform the same task. In these instances, it may be useful to identify the specific knowledge and skills that students need to know to successfully demonstrate mastery of a particular standard and then identify or develop tasks to serve as pretests from which progress on those standards can be determined.

Measuring Student Outcomes for "Caseload" Educators

Not every educator has a classroom. And some educators are responsible for services delivered to the entire school, not just a class. These related personnel (e.g., counselors, school psychologists, librarians, school

nurses, and speech therapists) may work with individuals but also with small or large groups of students. Although many states do not require the evaluation of such personnel in parallel with teachers, these "caseload" educators are included in the educator evaluation system in a number of states and districts. To measure their contributions to student learning growth, it may be helpful to think of them as having "caseloads." For example, a school counselor may have a caseload that includes:

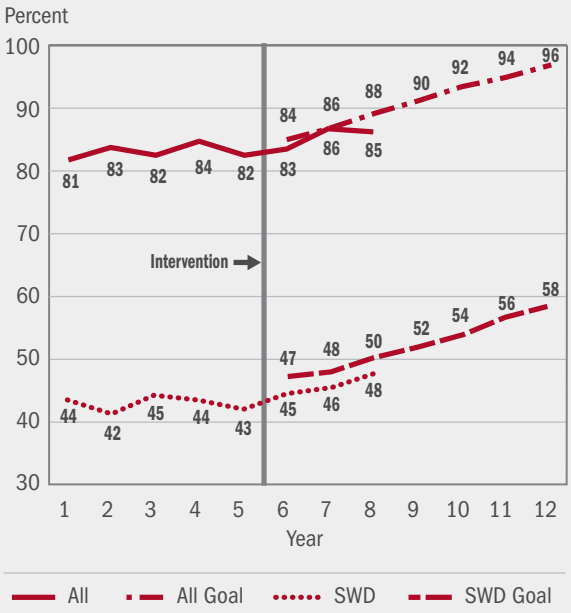
- All the students in the school (i.e., providing services such as career counseling at the high school level).
- Students experiencing emotional or behavioral problems.
- Students in crisis because of family events or relationship issues.
- Students with frequent unexcused absences.
- Teachers (e.g., providing professional development on recognizing the signs of physical or sexual abuse and what the law requires them to do).

Caseload educators may not be directly involved with academic content, making determining their contribution to academic achievement more difficult. These personnel may want to document their contributions to growth in terms of both educational successes and other types of outcomes. For example, a high school guidance counselor may want to track the proportion of students enrolling in AP classes, the proportion of students engaging in extracurricular activities, or the proportion of students for whom attendance rates have increased.

Caseload educators, and their associated goals, will likely vary according to the discipline and needs at the school, building, classroom, group, or individual student level. For example, a school with attendance issues may concentrate on attendance, whereas others may turn their attention toward AP course enrollment, reduction in incidences of bullying, or increased interactions between educators and parents.

Documented progress toward goals can be charted and monitored on an Excel spreadsheet, as illustrated in Figure 1. Likewise, intervention implementation can be tracked and monitored to determine effectiveness.

Figure 1. Sample of Documented Progress for Student Attendance



Source: Reschly and Holdheide (2010)

Alignment With Federal Priorities

Some measures are more likely than others to comply with federal priorities and state legislative mandates; however, these various approaches generally lack supporting research, leaving states and districts to their own devices to determine which options are most feasible. State and district priorities, financial resources, human capacity strengths and limitations, professional development needs, and system capacity issues should be contemplated prior to making decisions. General guidelines for selecting measures include the following:

- Avoid "reinventing the wheel." If tests already exist that can be used for measuring teachers' contributions to student learning, consider them first and determine whether they are useful in differentiating among levels of teacher effectiveness.

- Evaluate the available evidence for using the assessment as a measure of student growth for teacher evaluation.
  - Continue to evaluate the evidence by collecting and analyzing data resulting from the use of particular measures, including correlating measures with each other.
- Focus on measures that meet federal and state requirements and priorities by putting them to the following test:
  - Measures *must* show students' growth "between two or more points in time."
  - Measures *must* be "comparable across classrooms."
    - Consistency of measures across all teachers in a grade/subject ensures comparability of results.
    - For the four Ps—portfolios, products, performance, and projects—common rubrics should be used and agreement should be established as to how they will be used and who will score them.
  - Measures *must* be "rigorous."
    - Measures must be based on appropriate grade-level and subject standards.
    - Measures must demonstrate high expectations for student learning (i.e., on track to produce college- and career-ready graduates).
- Involve teachers and administrators in decision-making processes. They will benefit from their involvement, and their participation in considering appropriate measures will ensure greater "buy-in" for the results of the process.
- Choose measures that have the potential to help teachers improve their performance by:
  - Motivating teachers to examine their own practice against specific standards.
  - Allowing teachers to participate in or co-construct the evaluation (e.g., "evidence binders").
  - Giving teachers opportunities to discuss the results with evaluators, administrators, colleagues, teacher learning communities, mentors, and coaches.
- Choose measures that are directly and explicitly aligned with:
  - Teaching standards.
  - Professional development offerings.
- Include protocols and processes that teachers can examine and comprehend.

Application to All Grades and Student Populations

Assessing the effectiveness of teachers of students with disabilities and English learners presents challenges to determining teacher effectiveness due to the unique and varied roles these teachers assume (Holdheide et al., 2010). Likewise, measuring growth using standard measures for students with disabilities can be problematic, as standards-based models to determine growth are not based on individualized student goals.

The general tendency is to identify a different system or set of measures for special education teachers or English language specialists. Students with special needs and English learners have varying levels of ability and are taught in many different settings (e.g., general education classroom, resource room, separate classroom). Therefore, the types of assessment used to determine student growth may vary depending on the curriculum taught in the specified setting. Many students with special needs receive services in the general education classroom in which the assessments for determining student growth could (or should) be the same (possibly with accommodations) as that of students without disabilities, especially if these measures are vertically equated. For example, states may use the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment (Good & Kaminski, 2002, 2011) to determine student progress in reading

and the effectiveness of teachers in teaching reading, particularly if the state does not have a standardized measure of reading in early grades. The DIBELS assessment would be appropriate for general education students, including students with disabilities who are participating in the general education curriculum.

The appropriateness of each content-specific or grade-specific assessment should be considered, and appropriate accommodations should be provided as needed. Similarly, some students with disabilities are working toward alternative standards, such as a life skills curriculum, which is not reflected in the standardized tests. In this scenario, different assessments need to be identified in order to measure student growth toward those alternative standards. Therefore, participation by teachers of students with disabilities is essential as states assemble teams to design and develop appropriate measures in all achievement areas included in the standard curriculum. Special education teachers who serve in inclusion models and engage in coteaching are able to bring a perspective to this work that addresses the needs of general and special education students, thereby contributing to the design of appropriate assessments in the areas not currently tested with standardized measures. Separate teams of special educators who instruct toward alternative standards also may be developed, as their measures would vary considerably due to content and ability level.

Student progress on the individualized education program (IEP) has emerged as a potential source for measuring teacher effectiveness for students with disabilities. In one sense, it is not surprising because most IEPs contain individualized goals that are aligned with state standards, including measureable objectives that are monitored regularly for student progress. However, IEPs were never intended to be used as a tool to measure teacher effectiveness, and using them this way likely will raise legal and other potentially contentious issues. Though the individualized nature of the IEP and the detailed description of present levels and objectives for growth are positive features, standardized measures based on the general curriculum are still needed to assess teacher effectiveness.

STANDARDIZED EVIDENCE COLLECTION

Many states and districts are attempting to build comprehensive teacher evaluation systems that are responsive to federal priorities but are finding that there is little research to support the use of particular systems, weights, or measures. Because few states and districts currently have evaluation systems that incorporate multiple measures, there has been little opportunity to conduct research on how these measures perform. The question remains: Do the various measures in some weighted combination accurately identify teachers at different levels of effectiveness? Until systems with multiple measures and various weighting schemes are employed over time and evaluated by researchers, states and districts must be guided by general knowledge about how to use measures in a way that yields results that are rigorous and comparable.

One general method to ensure greater rigor in how multiple measures of all types are used is to implement *standardized* evidence collection. Everyone is familiar with the term *standardized test*. A standardized test is a test that is given according to specific rules that ensure that the test results will be comparable across students, schools, and districts. Specific rules also can be created and followed for all types of measures. By standardizing evidence collection, greater comparability across teachers is possible. Table 2 offers some suggestions for standardizing evidence collection for different types of measures of student learning growth.

Table 2. Standardizing Evidence Collection for Different Types of Measures

Type of Measure	How to Standardize Evidence Collection	Challenges
Curriculum-based pretests and posttests	Ensure that all teachers give the tests on the same day at the same time and allow students the same amount of time for completion. Teachers should agree to limitations on test preparation for posttests.	Accurately determining growth may be difficult in schools where students are particularly advanced versus schools where students begin the year below grade level. Adjustments may need to be made to account for these differences. Some students may do very well on the initial pretest, making it impossible to show growth. Providing those students with additional challenging curriculum and enrichment activities may allow them to show growth.
Student portfolios	Engage all teachers who plan to use student portfolios in the process of determining what constitutes acceptable evidence for various levels of performance (i.e., characteristics of a "beginning" versus "advanced" still life drawing). Develop or adopt appropriate rubrics and forms for teachers to use in establishing students' beginning performance levels on the knowledge and skills needed to meet the grade/content standards reflected in the portfolio. The same rubrics and forms can be used to evaluate the portfolio at the end of the course.	Portfolios should include not only the students' work but also the teachers' scoring rubric and comments and the students' reflections (i.e., how the student plans to improve upon the work). They should not be a catch-all for multiple iterations of an essay or other unrelated work. Teachers need to work together to create or adopt a rubric and scoring approach to ensure that they all agree on the characteristics of a "beginning" versus "advanced" effort. Schools/districts need to provide time to allow teachers to meet repeatedly during the year.
Classroom-based tests (e.g., DIBELS and the Diagnostic Reading Assessment)	Provide training for elementary teachers in the appropriate use of these instruments, how often they should be used, and how to record results so that student growth across time points can be determined.	Classroom-based tests were designed primarily to help teachers track progress and adjust instruction accordingly. Because students differ in reading ability in early elementary grades and have a range of growth trajectories, it will be challenging to compare relative teachers' contributions.
Student performance	Provide all art teachers in the district with the opportunity to meet and agree upon levels of performance (i.e., characteristics of a "beginning" performance and an "advanced" performance and how to document the performances to serve as evidence). The same applies to other classes for which a product or performance is the basis for the grade (e.g., music, drama, industrial arts classes).	If teachers do not have standards and a curriculum for the grade/subject, then they must first agree on what students should know and be able to do in a particular grade and subject before they can determine what different levels of performance should look like.
Other classroom-based evidence	Create opportunities for teachers in particular grades and subjects to meet together and agree upon ways to assess student learning. For example, timed multiplication drills might be used to document students' growth in skills over time, but teachers must agree to a set of materials and a timeframe for conducting the drills.	Teacher-created assessments, worksheets, student journals, records of experiments, and other types of evidence can be excellent sources of documentation of student growth between two points in time, but there must be some consistency across classrooms and teachers to make such evidence comparable.

Whether utilizing existing measures, designing new ones, or using a combination of both, states and districts need to ensure that the measure or method utilized does not take time away from teaching. Instead, these assessments need to be an integral part of the teaching cycle that can quickly gauge student growth and inform teacher practice. Adding complicated, labor-intensive measures and processes will likely result in an upheaval from the education community and threaten the validity of the results.

Measures That May Improve Teacher Performance

All measures are not created equally in terms of how much they can inform a teacher about his or her practice and success in teaching specific content. Measures that are distant from the classroom, such as standardized tests administered once per year, are less likely to influence teaching practice and student learning in a timely manner, whereas measures that are aligned with an integral part of the curriculum and instructional sequence may provide useful information to the teacher about which skills and knowledge students have already mastered. This type of feedback, such as results from a pretest administered early in the year, can be used to guide instructional decisions.

In addition, ongoing assessments and examination of student work, especially in cooperation with colleagues, may not be included as part of teacher evaluation but may be useful for teachers in determining next steps for their students. When teachers know areas in which the students are experiencing difficulty, they can use that information to make the necessary instructional adjustments (e.g., reteaching), allowing extra opportunities for practice, instruction in small groups, peer tutoring, computer-assisted instruction, individual tutoring, or other changes in the method or type of instruction. In addition, teachers find value in working together to examine and score student work (e.g., essays, portfolios, or projects). Discussions with other teachers about the differences between an outstanding piece of work and a good one can be valuable to teachers in thinking about how to target specific criteria in their own instruction.

Little attention has been paid to how the instruments and processes of teacher evaluation can inform professional growth opportunities. A feedback loop should be established that allows teachers and those who support them to identify areas of student weakness and strategize ways to improve instructional practices, resulting in improved student performance. Evaluation results should feed directly into that loop, providing specific, timely information in a format that is useful to teachers, administrators, and support personnel.

STATE GUIDANCE TO DISTRICTS

Districts will look to states for specific guidance about how to evaluate teachers' contributions to student learning growth, particularly in the nontested subjects and grades. There are several areas in which they need guidance.

Comparability: Across or Within Districts?

In order to better understand the differences among teacher effectiveness across schools and districts and identify teachers who are performing at high levels or those who are struggling, all teachers ideally would be evaluated in exactly the same way, using exactly the same measures. The state must first decide whether to insist on comparability *within* or *across* districts. A statewide system would be based on across-district comparability, whereas a district model would be based on within-district comparability. The following questions may be useful in making this decision:

- Is there a single set of subject-specific and grade-specific state standards for students that all districts use? If not, comparability across districts will be problematic.

- Do all districts throughout the state use the same curriculum and textbooks for all subjects? If not, it may be difficult to identify a common set of assessments that are appropriate for all districts.
- Do all districts have the same school calendar (e.g., start and end dates for the students, standardized testing dates, breaks, and holidays)? If not, it may be difficult to standardize the assessment process so that students are assessed at the same time across the state. The more standardized the assessment process is, the more comparable results will be.
- Do various types of educators in all districts across the state have the same job descriptions? The job description for some educators, particularly counselors, special educators, school nurses, librarians, and itinerant teachers, may vary widely from district to district.

If state staff answer "no" to any or all of these questions, they may want to consider comparability within rather than *across* districts. However, states could still provide guidelines to districts to ensure as much comparability as possible, given the district-to-district differences. For more information about appropriate guidance, see Goe, Holdheide, and Miller (in press).

Measures

States need to provide guidance to districts in selecting appropriate standards-based measures for documenting student growth. The following questions may help in determining the type of guidance to provide:

- Does the state want to approve all measures used by districts? If not, the state can provide the districts with guidelines and criteria for acceptable measures and leave approval of measures up to the districts.
- Does the state or district have a valid test that measures students' progress toward mastery of grade-level and subject standards? If not, other measures will have to be identified, purchased, or created to provide valid indicators of student growth. Districts can pool resources to share the costs of assessments and measures as a more cost-effective approach than each district attempting to pay these costs individually.
- Do districts have the capacity to implement processes for assessing student growth? If not, districts may need to join with other districts in regional or other purposeful consortiums to take advantage of economies of scale. For example, a number of rural districts might share information and resources, whereas an urban district might join forces with other urban districts in the state to form a consortium to share resources.

Exceptions

After a state or district adopts specific measures and processes for determining student learning growth, decision makers need to consider how to manage "exceptions" to the established processes for using these measures. For example, should a teacher be held accountable if the student was only assigned to his or her class for a portion of the school year? Or what happens if the student rarely attends school? Should the same level of accountability or attribution be assigned? Should working conditions be considered as a factor in determining teachers' contributions to student learning growth? States and districts, working closely with teachers, administrators, and stakeholder groups, need to determine which exceptions to include and how to include them in ways that will ensure fairness and comparability.

Approaches to handling these exceptions may be left up to districts, but states may provide guidance or limit options to ensure greater comparability across districts.

Table 3. Priorities, Challenges, and Potential Solutions

Priority	Challenges	Potential Solutions
Measuring student growth between "two points in time"	Students complete only the pretest but not the posttest or vice versa. Students fail to turn in required work (e.g., a portfolio or project being used as the postmeasure).	With large numbers of students (e.g., at the secondary level), eliminate the student from the pool of students used to calculate the average student growth for the teacher. With smaller class sizes, it is important to include as many students as possible to reduce the margin of error. Allowing a review of other student work (homework or classwork), comparing current work or scores to those from previous years, or devising standards-based projects for students to complete are possible options, though imperfect at best.
Ensuring "rigor" of assessments	The measures used are complex, and it is difficult to determine rigor. There is little agreement about what rigor is and how it is reflected in the measures.	For a portfolio, project, or other multi-part measure, break down the components by the standard(s) being addressed. Will success on these components provide a clear indication of students' mastery of standards-based knowledge or skills? Subject and grade-level standards should provide the focus for all measures. If the measure is not adequate to show progress toward mastery of standards-based skills and knowledge, it is not rigorous. In addition, demonstration of mastery of the knowledge and skills should be possible with the measure.
Making certain that measurement is "comparable across classrooms"	Raters are not adequately trained in scoring students' work for portfolios, projects, performances, and products (the four Ps) that are being used as measures of students' growth.	Essays and the four Ps (i.e., portfolios, projects, performances, and products) all require training with scoring rubrics to ensure that all raters agree on what each level of the rubric looks like. Raters may be teachers, administrators, district personnel, or people hired specifically for scoring, but they must be trained to a high level of agreement. In addition, retraining and calibration should be conducted periodically to ensure that raters are still in agreement on interpreting the evidence. Training involves examining and discussing student work and rating it, then discussing rating decisions until agreement is reached.
	Teachers acting as raters do not have time in their schedules to work with "like" teachers on scoring writing samples, portfolios, projects, performances, products (the four Ps), and so on.	When teachers are trained as raters, it is important that they are given time to work together on scoring student work. Greater reliability and thus greater comparability will be achieved with multiple raters working together. Using some scheduled professional development time, grade-level or subject-level meeting time, or team time may be necessary.
	Pretests and posttests are not given in a standardized way.	Results will not be comparable across classrooms unless specific practices are followed in giving pretests and posttests. These practices require a commitment and coordination across schools within a district to (1) choose a date/time that all schools agree to for pretesting of a subject/grade; (2) ensure that teachers are properly instructed on how to prepare students for the pretests and posttests; (3) give the tests at the same time of day; and (4) give tests for a predetermined length of time.

Ongoing Research on Systems, Models, and Measures

Changes in teacher evaluation policies have occurred at a dizzying pace, outstripping researchers' ability to study the validity and fairness of the systems themselves and the individual components of the systems. Although research has been conducted on some of the measures, studies generally focus on low-stakes evaluation systems. (For a review of research on measures, see Goe, Bell, and Little, 2008.) There is little research on using student achievement growth as a measure of teacher effectiveness in a high-stakes system in which the results could mean commendation or probation, rewards or even dismissal. Planning for and consistently evaluating the relative quality of results from the use of various measures is important to increasing ability to accurately determine teacher effectiveness.

As states and districts implement evaluation systems that include multiple measures of student learning, it will be possible to evaluate the usefulness of various measures in differentiating among educators' levels of performance. This type of research should result in enhanced ability to conduct teacher evaluations that provide a nuanced, comprehensive, and accurate picture of teachers' contributions to student learning growth.

Considerations for States: Moving Forward

Without a research base to guide states' efforts, the TQ Center encourages caution and careful deliberation in designing and implementing high-stakes evaluation systems that measure teachers' contributions to student learning growth. States may consider the following as they move forward:

- Partner with national and regional comprehensive centers in conducting needs assessments and outlining steps to take in determining appropriate measures and processes.
- Bring stakeholders (e.g., teachers, administrators, parents, school board members, union representatives, business leaders) to the table early in the discussions about measures and seek their help in communicating results.
- If the state does not currently have grade-level and subject standards for all courses, adopting such standards is important to ensure appropriate rigor in measuring student learning growth.
- The following steps can be used for selecting measures:
  - Categorize teachers by whether they are in tested or nontested subjects and grades.
  - Develop indicators within data systems to link teachers to appropriate student growth data.
  - Determine whether there are existing measures that might be useful in measuring student growth, and establish an approval process and/or listing of acceptable measures.
  - Secure content expertise to help evaluate coverage (i.e., whether measures exist to show learning growth for all teachers).
  - When gaps are found in existing measures, purchase or develop appropriate measures.
  - Consider alternative assessments as well as how measures need to be modified or differentiated through accommodations for students with special needs.

- Conserve resources by encouraging districts to join forces with other districts or regional groups to determine appropriate measures for nontested subjects and grades. This approach also contributes to greater comparability because teachers will be using the same measures across schools, districts, and regions.
- Consider whether human resources and capacity are sufficient to ensure fidelity of implementation.
- Develop a communication strategy to increase awareness and buy-in. Consider "frequently asked questions" pages on state and district websites and other means of sharing information about how and why measures were chosen and how they will be used.
- Establish a plan to evaluate measures to determine whether they can effectively differentiate among teacher performance.
- Evaluate processes and data each year and make needed adjustments.

CONCLUSION

There is little doubt that teacher evaluation has been permanently and irrevocably changed. No longer is a score on a principal's observation checklist acceptable as evidence that a teacher is effective in the classroom. Linking teachers with student outcomes—including evidence of their growth in standards-based knowledge and skills—will become increasingly common. Moving forward in a responsible, deliberate, and cautious manner will ensure that the results are valid and defensible.

REFERENCES

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Battelle for Kids. (2009). *The importance of accurately linking instruction to students to determine teacher effectiveness.* Columbus, OH: Author. Retrieved February 18, 2011, from http://static.battelleforkids.org/images/BFK/Link_whitepagesApril2010web.pdf

Betebenner, D. W. (2008). *A primer on student growth percentiles.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February 18, 2011, from http://www.cde.state.co.us/cdedocs/Research/PDF/Aprimeronstudentgrowthpercentiles.pdf

Braun, H., Chudowsky, N., & Koenig, J. A. (2010). *Getting value out of value-added: Report of a workshop.* Washington, DC: National Academies Press. Retrieved February 18, 2011, from http://www.nap.edu/openbook.php?record_id=12820&page=1

Feng, L., & Sass, T. R. (2009). *Special education teacher quality and student achievement.* Unpublished working paper.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 18, 2011, from http://www.tqsource.org/publications/LinkBetweenTQandStudentOutcomes.pdf

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 18, 2011, from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Goe, L., Holdheide, L., & Miller, T. C. (in press). *A practical guide to designing state teacher evaluation systems.* Washington, DC: National Comprehensive Center for Teacher Quality.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of early basic literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., & Kaminski, R. A. (2011). *DIBELS next assessment manual.* Longmont, CO: Sopris.

Herman, J. L., Heritage, M., & Goldschmidt, P. (in press). *Guidance for developing and selecting student growth measures for use in teacher evaluation.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Holdheide, L., Goe, L., Croft, A., & Reschly, D. (2010). *Challenges in evaluating special education teachers and English language learner specialists* (Research & Policy Brief). Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 18, 2011, from http://www.tqsource.org/publications/July2010Brief.pdf

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446, 118 Stat. 2647 (2004). Retrieved February 18, 2011, from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=108_cong_public_laws&docid=f:publ446.108.pdf

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique.* Cambridge, MA: National Bureau of Economic Research. Retrieved February 18, 2011, from http://economics.missouri.edu/working-papers/2009/WP0902_koedel.pdf

McCaffrey, D., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606. Retrieved February 18, 2011, from http://www.mitpressjournals.org/doi/pdf/10.1162/edfp.2009.4.4.572

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002). Retrieved February 18, 2011, from http://www.ed.gov/policy/elsec/leg/esea02/index.html

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades.* Washington, DC: Center for Educator Compensation Reform. Retrieved February 18, 2011, from http://www.cecr.ed.gov/guides/other69Percent.pdf

Reschly, D. R., & Holdheide, L. R. (2010, September 21). Figure presented during discussion group convened at "Evaluating and Rewarding Effectiveness and Navigating the Evolving Landscape," a symposium cohosted by the National Center on Performance Incentives, Battelle for Kids, and Vanderbilt Peabody College, Nashville, TN.

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes.* Washington, DC: Economic Policy Institute.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458. Retrieved February 18, 2011, from http://www.econ.ucsb.edu/~jon/Econ230C/HanushekRivkin.pdf

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256. Retrieved February 18, 2011, http://www.sas.com/govedu/edu/ed_eval.pdf

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (No. R11-0435-02-001-97). Knoxville: University of Tennessee Value-Added Research and Assessment Center. Retrieved February 18, 2011, http://www.mccsc.edu/~curriculum/cumulative%20and%20residual%20effects%20of%20teachers.pdf

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains.* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved February 18, 2011, from http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Secretary's Priorities for Discretionary Grant Programs, 75 Fed. Reg. 47,288 (proposed Aug. 5, 2010). Retrieved February 18, 2011, from http://www2.ed.gov/legislation/FedRegister/other/2010-3/080510d.pdf

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89–122.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Brooklyn, NY: The New Teacher Project. Retrieved February 18, 2011, from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

ABOUT THE NATIONAL COMPREHENSIVE CENTER FOR TEACHER QUALITY

The National Comprehensive Center for Teacher Quality (TQ Center) was created to serve as the national resource to which the regional comprehensive centers, states, and other education stakeholders turn for strengthening the quality of teaching—especially in high-poverty, low-performing, and hard-to-staff schools—and for finding guidance in addressing specific needs, thereby ensuring that highly qualified teachers are serving students with special needs.

The TQ Center is funded by the U.S. Department of Education and is a collaborative effort of ETS, Learning Point Associates, and Vanderbilt University. Integral to the TQ Center's charge is the provision of timely and relevant resources to build the capacity of regional comprehensive centers and states to effectively implement state policy and practice by ensuring that all teachers meet the federal teacher requirements of the current provisions of the Elementary and Secondary Education Act (ESEA), as reauthorized by the No Child Left Behind Act.

The TQ Center is part of the U.S. Department of Education's Comprehensive Centers program, which includes 16 regional comprehensive centers that provide technical assistance to states within a specified boundary and five content centers that provide expert assistance to benefit states and districts nationwide on key issues related to current provisions of ESEA.

NATIONAL COMPREHENSIVE CENTER FOR TEACHER QUALITY

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
877.322.8700 | 202.223.6690

www.tqsource.org

Copyright © 2011 National Comprehensive Center for Teacher Quality, sponsored under government cooperative agreement number S283B050051. All rights reserved.

This work was originally produced in whole or in part by the National Comprehensive Center for Teacher Quality with funds from the U.S. Department of Education under cooperative agreement number S283B050051. The content does not necessarily reflect the position or policy of the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by the federal government.

The National Comprehensive Center for Teacher Quality is a collaborative effort of ETS, Learning Point Associates, and Vanderbilt University.

0039_03/11

LEARNING POINT Associates®
An Affiliate of American Institutes for Research®

ETS®

VANDERBILT®