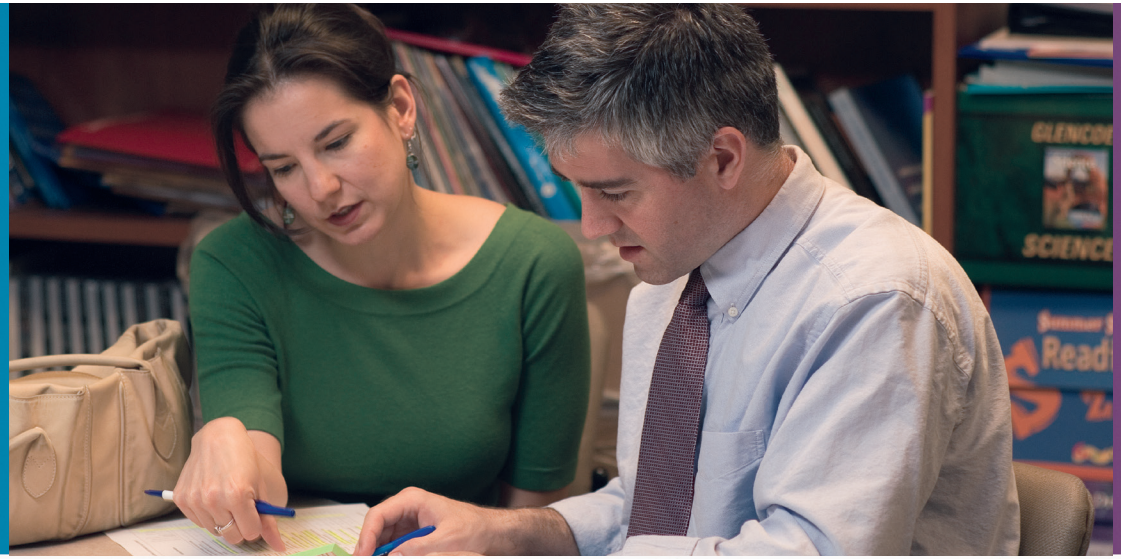
Center on
GREAT TEACHERS & LEADERS
at American Institutes for Research

Practical Guide to Designing Comprehensive Teacher Evaluation Systems



A Tool to Assist in the Development of Teacher Evaluation Systems

FEBRUARY 2014

AIR®
American Institutes for Research®

The content of this Practical Guide was originally developed in 2011 by the National Comprehensive Center for Teacher Quality and then updated in 2014 by the Center on Great Teachers and Leaders.

See **http://www.gtlcenter.org/tools-publications/online-tools/teacher-evaluation** to view the interactive, online version of the *Practical Guide to Designing Comprehensive Teacher Evaluation Systems*.

Practical Guide to Designing Comprehensive Teacher Evaluation Systems

Revised Edition

FEBRUARY 2014

Laura Goe, Ph.D.
ETS

Lynn Holdheide
American Institutes for Research

Tricia Miller, Ph.D.
Consultant

Contents

Rationale and Structure

Across the nation, states and districts are in the process of building better teacher evaluation systems that not only identify highly effective teachers but also systematically provide data and feedback that can be used to improve teacher practice. The *Practical Guide to Designing Comprehensive Teacher Evaluation Systems* is a tool designed to assist states and districts in constructing high-quality teacher evaluation systems in an effort to improve teaching and learning.

This tool is not a step-by-step guide to devising a teacher evaluation system. Rather, it is intended to facilitate discussion and promote coherence in the development process. The following assumptions have guided its construction:

- In response to federal initiatives and priorities as well as state legislation, states are motivated to improve their current evaluation systems to better identify successful teachers, assist less successful teachers, and help all teachers improve their practice.
- Most current definitions of *teacher effectiveness* (e.g., the Race to the Top definition) include teachers' contributions to student learning growth, and states need to consider measuring these contributions for all teachers.
- States are interested in systems that use multiple measures to assess various aspects of teachers' performance and instructional practice.
- States may be in various stages in terms of creating or revising teacher evaluation systems. This tool allows states to focus on the specific components of the system that are most relevant for them.
- In states where districts have substantial control over teacher evaluation systems, this tool may be used by districts or by consortia of districts for discussion and guidance.
- Teachers play a critical role in ensuring that the evaluation system is fair, valid, and successful; they should be active participants in designing, developing, implementing, and evaluating the system.

The guide begins with an overview of the factors influencing teacher evaluation reform today and continues with a discussion of approaches to balancing state accountability and district autonomy. The next section of the guide is structured around the following essential components of the design process as supported through research:

- Component 1a: Specifying Evaluation System Goals
- Component 1b: Establishing Standards
- Component 2: Securing and Sustaining Stakeholder Investment, and Cultivating a Strategic Communication Plan
- Component 3: Selecting Measures
- Component 4: Determining the Structure of the Evaluation System
- Component 5: Selecting and Training Evaluators
- Component 6: Ensuring Data Integrity and Transparency
- Component 7: Using Teacher Evaluation Results
- Component 8: Evaluating the System

Each subsection includes an overview of the component, resources and practical examples, and a series of guiding questions designed to help states organize their work and move strategically toward an evaluation system that functions to improve student learning and teacher performance.

Introduction

The research community has long recognized the importance of teachers to student achievement. Although research has shown that teachers are the most significant school-based factor in student achievement, traditional methods of evaluating teachers have not been able to capture or explain differences between effective and ineffective teachers.

Initial efforts to ensure quality education focused on teacher qualifications and degrees; however, research does not indicate that these factors significantly influence teacher effectiveness. For example, Rivkin, Hanushek, and Kain (2005) analyzed results from thousands of teachers and their students in Texas and determined that there were strong teacher effects on academic achievement, but variation in these effects could not be explained by education or experience.

Further, mounting evidence indicates that the United States is losing ground in comparison to other countries in terms of educational outcomes. One international study showed that U.S. students were outperformed in mathematics by students in 20 of the other 28 industrialized countries studied (Lemke et al., 2004). In addition, a recent Program for International Student Assessment study found that only five of the other 33 participating countries had lower scores in mathematics literacy than the United States (Fleischman, Hopstock, Pelczar, & Shelley, 2010). These types of findings resulted in increased concern about determining the best way to improve student learning through teacher performance and a shift in focus from analyzing teacher inputs (e.g., education, certification, and experience) to measuring teacher effects (e.g., student achievement and classroom practice).

Improving teacher quality and effectiveness is a complex issue, and the ability to identify high-performing and low-performing teachers is a necessary step toward pinpointing instructional strategies and pedagogy that result in improved student growth (e.g., evidence-based instructional strategies, strong student-teacher relationships). Unfortunately, traditional evaluation methods have not proven to be useful in meeting this challenge. In the past, teacher evaluation systems have varied widely in their rigor and utility. Most systems were based on classroom observations, usually conducted by principals but sometimes conducted by trained evaluators (see "Practical Example: Cincinnati Public Schools Evaluation System"). The steps taken after the observations differed considerably across states, districts, and even schools, with some schools linking results to professional growth plans for teachers and others filing the results away with little or no follow-up. The perfunctory, compliance-oriented approach to teacher evaluation in some districts likely did not contribute to tangible improvement in teaching and learning. Unfortunately, there has been little research on how these different approaches to classroom observation influenced teacher performance.

PRACTICAL EXAMPLE

Cincinnati Public Schools Evaluation System

Cincinnati teachers participate in a "comprehensive evaluation" during their first and fourth years of teaching; after the fourth year, they are evaluated every five years. Teachers are observed four times by teacher evaluators and once by a school administrator. Before they can become teacher evaluators, teachers must complete a three-step application process to become lead teachers. Lead teachers may then apply for positions such as teacher evaluators, consulting teachers, and program facilitators. Those selected for teacher evaluator positions are required to undergo extensive training in collecting and scoring evidence. Using videos, they are certified through a process of verifying the agreement of their scores with those of "master raters." Through this process, a high degree of reliability is ensured, meaning that a teacher's observation score is likely to be the same or nearly the same, regardless of which trained evaluator conducts the observation.

Source: Cincinnati Public Schools (n.d.)

In 2009, an investigation into the compliance-oriented approaches of evaluation systems conducted by The New Teacher Project sent shockwaves through the policy world. The study, titled *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*, examined large and small districts across several states where evaluation consisted primarily of classroom observations (Weisberg, Sexton, Mulhern, & Keeling, 2009). The following conclusions emerged from the study:

- Nearly all teachers received high ratings (*good* or *great*).
- Districts failed to recognize and reward excellence.
- Professional development was rarely tied to results and when it was, little support was offered to teachers.
- New teachers generally were rated above satisfactory, and tenure was seldom denied to teachers based on observation results.
- Poor performance rarely led to teacher dismissal.

The inability of evaluation systems to differentiate factors contributing to teacher effectiveness suggests that classroom observations, at least as they were used in most districts in the study, are of little use for improving and rewarding performance or identifying teachers who need support and training and those who should be dismissed.

Through funding opportunities including the American Recovery and Reinvestment Act (ARRA) and the Race to the Top competition, the federal government has encouraged states and districts to develop rigorous evaluation systems for use in high-stakes decisions including teacher advancement, compensation, distribution, and retention. These opportunities, coupled with the evidence of poorly functioning teacher evaluation systems, have resulted in a national urgency to create and implement comprehensive, strategic systems for evaluating teacher performance that identify, support, and develop teacher effectiveness and student growth.

In response to this urgency, many states have passed legislation mandating the development of rigorous, high-quality evaluation systems for use in high-stakes situations related to teacher employment and advancement. Advisory boards, committees, and multistate consortia are meeting to gather information on research and best practices related to the development, implementation, and use of these evaluation systems. This *Practical Guide* provides education policymakers and stakeholders with guidance on the key areas that should be addressed during the development and implementation of a new evaluation system.

State Accountability and District Responsibility in Teacher Evaluation Systems

Until recently, teacher evaluation has largely been considered the purview of districts or schools without much, if any, involvement from states. Starting with the highly qualified teacher requirements as codified in the Elementary and Secondary Education Act (ESEA), as reauthorized by the No Child Left Behind Act (NCLB), and continuing through the ARRA reform goals and assurances, states are expected to play an increasingly larger role in ensuring the quality and effectiveness of the nation's teaching force. This expectation creates a challenge for many states with a long history of local autonomy in most education matters.

Specific to teacher evaluation systems, states now must decide the extent to which the teacher evaluation model will make allowances for local flexibility and provide a balance between local and state control that encourages collective responsibility and accountability. This section includes an overview of several key roles states may play in assisting districts in the implementation of new evaluation requirements and descriptions of several models that balance state accountability with local autonomy.

Key State Roles

Interpreting Federal and State Regulations

Spurred on by the Race to the Top competition, many states have either recently passed new legislation or pointed to existing legislation concerning teacher evaluation, most of which is directly related to the four ARRA reform goals or assurances: ensuring the quality of standards and assessments, improving the collection and use of data, increasing teacher effectiveness and equitable distribution, and supporting struggling schools (Learning Point Associates, 2010). The language of this federal and state legislation permits varying degrees of flexibility in terms of how the evaluation system is to be implemented. As such, the responsibility for interpretation and implementation falls primarily to states. Implementing new teacher-evaluation laws and policies usually involves interpretation and overcoming challenges that may not have been anticipated by the policymakers. Recognition of these challenges and their potential variations according to local contexts should inform the training needs of personnel and contribute to the development of the evaluation model. Accordingly, states need to take proactive steps in helping districts interpret the new legislation and determining the best course of action toward full implementation.

For example, the Race to the Top application indicated that student achievement was to be a "significant" component of teacher evaluation. However, the federal government did not define *significant*, and currently there is not a research base to support differential weighting of the components in an evaluation system. As a result, many states looked to their own legislation to resolve any discrepancies in interpretation or implementation.

RESOURCES

Approaches to Evaluating Teacher Effectiveness: A Research Synthesis
http://www.gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf

This research synthesis examines how teacher effectiveness is measured and provides practical guidance for evaluating teacher effectiveness. It evaluates the research on teacher effectiveness and various measures. In addition, it defines components and indicators that characterize effective teachers, extending this definition beyond teachers' contributions to student achievement gains to include how teachers affect classrooms, schools, and colleagues as well as how teachers contribute to other important outcomes for students.

Methods of Evaluating Teacher Effectiveness **(Research-to-Practice Brief)**
http://www.gtlcenter.org/sites/default/files/docs/RestoPractice_EvaluatingTeacherEffectiveness.pdf

This brief is intended to help regional comprehensive centers and state policymakers as they consider evaluation methods to clarify policy, develop new strategies, identify effective teachers, or guide and support districts in selecting and using appropriate evaluation methods for various purposes. Included in this brief is a five-point definition of teacher effectiveness, which the authors developed by analyzing research, policy, and standards that address teacher effectiveness and by consulting experts in the field.

A Practical Guide to Evaluating Teacher Effectiveness
http://www.gtlcenter.org/sites/default/files/docs/practicalGuide.pdf

This guide offers a definition of teacher effectiveness that states and districts may adapt to meet local requirements, provides an overview of the many purposes for evaluating teacher effectiveness, and indicates which measures are most suitable to use under different circumstances. The guide also includes summaries of various measures, such as value-added models, classroom observations, analysis of classroom artifacts, and portfolios. The summaries include descriptions of the measures, along with a note about the research base and strengths and cautions to consider for each measure.

For example, several states codified the weight (percentage) of student achievement in the teacher evaluation system (e.g., Tennessee specified 50 percent, Rhode Island specified 51 percent, and Colorado specified 50 percent). Such state legislation was intended to drive changes in evaluation systems and provide better information about teachers' contributions to student learning growth. However, the legislation often did not address the other logistical and procedural aspects of teacher evaluation (e.g., how growth would be measured, the frequency of the evaluation, who would conduct the evaluations, and how evaluators would be trained). States should play a critical role in interpreting such legislation and be prepared to help districts address specific challenges, unintended consequences, and implementation considerations at the district level

Interpreting or Conducting Research

The dearth of available research-based methods and models of comprehensive teacher evaluation hinders states' abilities to offer assistance to districts. Although some research on the utility of specific measures of teacher performance exists (Goe, Bell, & Little, 2008), albeit limited, most has been conducted in low-stakes environments. Therefore, many states have chosen to assemble task forces and engage national experts in evaluation and measurement to secure recommendations and inform the conversation concerning teacher effectiveness and evaluation policy.

In many cases, states and districts may need to identify measures and conduct research during and after implementation. Given potential resource and human capacity limitations at the district level, states may need to play an active role in conducting research to ensure that the evaluation model is technically sound and therefore defensible, especially in situations in which teacher evaluation results will be used to make personnel and compensation decisions. These conversations and preliminary research could be instrumental in ensuring the validity of the results from comprehensive teacher evaluation systems and gaining educator and stakeholder support.

Models for State and District Evaluation Systems

Historically, models of teacher evaluation varied among schools, districts, and states and were largely dependent on the context in which the model evolved and was implemented. However, as federal guidance and policy lean toward more state responsibility for ensuring teacher quality and monitoring district teacher evaluation, states must determine their role and level of involvement—from providing limited guidance to taking a more directive approach.

For example, some states may elect to mandate a particular evaluation model, governing logistics (e.g., how often teachers are evaluated), format (e.g., selection of measures), and personnel decisions (e.g., what a rating means in terms of teacher tenure). Others may provide specific guidelines for the evaluation model while allowing the district flexibility in adapting those guidelines locally and in the implementation of the system. The level of flexibility will likely vary according to many factors (e.g., political context, local bargaining agreements, state size, and district capacity) and the state's goals. Some states, like Delaware, have mandated a statewide evaluation system. Other states allow every district to determine its own model for teacher evaluation as long as

stated requirements are met. States also may create or facilitate consortia among districts in the same region or those that share similar challenges (e.g., a rural consortium encompassing several contiguous districts or a statewide consortium of urban districts).

Various state options and accompanying stakeholder considerations are discussed in the following subsections. Note that this is not an exhaustive listing of options.

State-Level Evaluation System

Within a state-level evaluation system, the state provides a strict interpretation of legislation and prescribes the requirements for the teacher evaluation model. The state determines the components of the teacher evaluation model, which measures are to be used, how often evaluations are to be conducted, and by whom. Therefore, the state is instrumental in the design, implementation, and evaluation of the teacher evaluation model. Delaware is currently in the process of implementing this type of system (see "Practical Examples: Delaware's Evaluation System"). With significant contribution from local practitioners, the state has led the efforts related to the development of a comprehensive teacher evaluation model. After the model is finalized, all Delaware districts will be required to implement the model with little flexibility.

Elective State-Level Evaluation System

Within an elective state-level evaluation system, states may elect to provide a strict interpretation of state or federal legislation and dictate certain aspects of the evaluation model but allow flexibility in others. For example, the state may have legislation that mandates the use of student achievement as a significant factor, and district models would have to include measures of student achievement. Or the state may have specific language about which aspects of teacher evaluation are subject to local decision making and which aspects are state mandates that are not open to negotiation. The state may mandate the type of growth model and other measures the districts use to determine student growth, the attribution of growth to teachers, and the weight (percentage) of the components of the teacher evaluation system. The possible components of the evaluation model (e.g., observation protocols, portfolios, student/parent surveys) and processes for using them would be determined by the district. For instance, the state might offer the Framework for Teaching (Danielson Group, 2011) as an option that districts could elect to use but allow districts to choose different observation models as long as certain criteria are maintained. Thus, the state plays a major role in ensuring that certain components are part of the district models but allows for local flexibility in other aspects of the system. This option allows a continuance of established district models, provides flexibility for bargaining agreements, and continues the tradition of local control over teacher evaluation that exists in many states (see "Practical Examples: New York's Evaluation System").

District Evaluation System With Required Parameters

States that find it impractical to adopt a single statewide evaluation system may still deem it necessary to provide guidance to districts in implementing regulations and state priorities for teacher evaluation. Within a district evaluation system with required parameters, states play a much smaller role in the design and implementation of the teacher evaluation system at the district level. Guidance may be somewhat general, such as requiring states to implement an evaluation system that includes several components (e.g., observations, evidence of professional responsibilities, and evidence of teachers' contributions to student achievement).

The guidance also may be more restrictive, particularly if some aspects of the evaluation system are already in place. In this case, the state provides some level of guidance to districts and specifies the parameters for the district models. The state, therefore, does not play a major role in the evaluation process but provides some type of screening or approval to ensure district compliance in selecting models as well as an audit or follow-up mechanism to ensure that districts are working within the defined parameters. For instance, the state may indicate that districts can select their own evaluation model but require that new teachers be observed three times per year for at least 20 minutes per visit. The district has flexibility in the model selection, but the logistical parameters need to be met (see "Practical Examples: Ohio's Guidelines to Evaluation").

PRACTICAL EXAMPLES

Delaware's Evaluation System

Delaware already had a statewide evaluation system in existence prior to being awarded Race to the Top funds. This system included classroom observation and opportunities for professional growth. However, Delaware's existing system lacked a mechanism to measure student growth. Therefore, an external evaluation was conducted that included soliciting feedback from teachers and administrators through surveys, interviews, and focus groups. The state department collaborated with union representatives to ensure that the system would be accepted as comprehensive, valid, and fair. These results contributed to the design of a statewide model. However, given the timelines and implementation demands, it is not clear whether Delaware will ultimately use this model; alternatives are still being considered.

Adapted from page 12 of *Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects* by L. Goe and L. Holdheide. Copyright © 2011 National Comprehensive Center for Teacher Quality.

New York's Evaluation System

New York's system is an example of an elective state model, providing clear guidance about how new evaluation requirements will be phased in over several years. The system is based on a 100-point scale; 60 percent of the evaluation score will be based on locally negotiated processes (e.g., classroom observations by trained evaluators), and 40 percent will be based on a combination of state standardized tests and local assessments and measures, which will have to be developed by each school system.

Year 1: 20 percent of student growth is based on state assessments or comparable measures for teachers in the common branch subjects or ELA and mathematics in Grades 4–8 only, and 20 percent is based on other *locally selected* measures that are rigorous and comparable across classrooms.

Year 2: After two years, 25 percent will be based on standardized tests, and 15 percent will be based on locally selected measures.

Source: New York State Education Department (2011)

Ohio's Guidelines to Evaluation

Ohio has developed state teacher evaluation guidelines as follows.

The teacher evaluation system adopted by the district should:

- "Align to the *Standards for Ohio Educators.*"
- "Be systematic and ongoing in order to promote professional development and student learning."
- "Take into account experience, skill, longevity, and responsibility."
- "Use a variety of measures to collect evidence."
- "Include three or four clearly defined levels of performance to differentiate performance/effectiveness of teachers and encourage continuous professional growth. A performance appraisal rubric should also be developed."

Source: Ohio Department of Education (2010, pp. 9–10)

Factors for Stakeholder Consideration

Stakeholders might consider the following factors in selecting a particular model:

- Grant requirements as applicable (e.g., Race to the Top, School Improvement Grants, Teacher Incentive Fund)
- Existing or impending state legislation that affects the evaluation process
- Goals and priorities at the state and district levels
- State-level role in district practice
- The role of unions and bargaining agreements in local and state decisions
- The number and diversity of districts within a state as well as geographical distance between them
- The human and resource capacity at the state and local levels
- The training needed to implement the system with fidelity
- Stakeholder support for changes in teacher evaluation
- Technological capacity, including the ability to link teachers with students
- District models already in use and their level of acceptance at the local level
- Teachers' and administrators' preferences for certain types of measures

Note: Race to the Top and ARRA indicate that total district-level control with no state-level involvement or accountability is not supported at the federal level.

Table 1 lists some potential strengths and weaknesses within the various models.

Table 1. Evaluation Model Strengths and Weaknesses

Model	Strengths	Weaknesses
State-Level Evaluation System	■ Measures and dimensions are the same statewide. ■ Data collection can be standardized. ■ Districts can be directly compared. ■ Evaluating the system and results will be easier. ■ System is perceived as fair because all districts are held to the same standards. ■ There is increased system reliability because changes from year to year affect all districts.	■ Local flexibility and ownership are diminished. ■ The system fails to consider local context. ■ It is difficult to obtain statewide support. ■ There is variance in district resources. ■ The system may be subject to local bargaining agreements. ■ The system may be seen as unfair by low-capacity districts forced to implement the same model as districts with greater capacity. ■ Local variations in school year and testing times may result.
Elective State-Level Evaluation System	■ The system allows for some local flexibility. ■ Data collection can still be standardized for certain components. ■ Districts can be directly compared in certain areas. ■ Reliability is strong in required components. ■ The system allows for continuance of locally developed models.	■ Local flexibility in certain areas is diminished. ■ The system presents more challenges for state oversight. ■ Data aggregation of teacher results may be more difficult.
District Evaluation System With Required Parameters	■ Local ownership and buy-in are increased. ■ Districts have the ability to address local priorities within the model. ■ The system allows for continuance of locally developed models.	■ It is difficult to compare progress or results. ■ Data aggregation may present considerable challenges. ■ Reliability is vulnerable across districts. ■ Training to ensure fidelity would likely be conducted at the district level, meaning more district resources are required. ■ Resources may be limited.

Development and Implementation of Comprehensive Teacher Evaluation Systems

This section is divided into eight subsections describing the critical components of designing a comprehensive teacher evaluation system:

- Component 1a: Specifying Evaluation System Goals
- Component 1b: Establishing Standards
- Component 2: Securing and Sustaining Stakeholder Investment, and Cultivating a Strategic Communication Plan
- Component 3: Selecting Measures
- Component 4: Determining the Structure of the Evaluation System
- Component 5: Selecting and Training Evaluators
- Component 6: Ensuring Data Integrity and Transparency
- Component 7: Using Teacher Evaluation Results
- Component 8: Evaluating the System

Each subsection discusses the relevance of the component in the design process and concludes with a series of questions to guide the development process.

COMPONENT 1a

Specifying Evaluation System Goals

Goal setting is an imperative, and often challenging, first step in designing a teacher evaluation system. The establishment of explicit, well-defined goals lays the foundation for a comprehensive, sustainable evaluation system. Some states have defined teacher evaluation goals and purposes in recent legislation and/or policy. In most scenarios, however, stakeholders are left to define effective teaching and achieve consensus on the evaluation system's purpose. There is a general tendency to oversimplify this crucial step; however, agreement about goal selection focuses and guides all decisions throughout the design process. The methods and weighting used for various components of the resulting system and any actions informed by evaluation results (e.g., professional development targeted to a challenge area) should reflect the evaluation system goals.

Stakeholders should exercise caution when selecting goals, keeping in mind that the ultimate objective of teacher evaluation is to improve teaching and learning. Systems designed exclusively for accountability are less likely to have an impact on teacher practice than those tied to professional learning opportunities and growth.

At the same time, if a goal of the teacher evaluation system is to make personnel and compensation decisions, there is an increased need to ensure that measures are technically defensible. The higher the stakes, the greater the need to establish reliable and valid measures that can accurately differentiate among more and less effective teachers. Likewise, if the goal of the evaluation system is to improve teacher practice, ensuring a link to professional learning within the evaluation cycle is crucial.

Reviewing current state and district initiatives is another important step of the goal selection process. Gaining clarity in state and district reform initiatives enables consistency among programs and prevents fragmentation in which human resource capacity and professional development decisions are made in isolation. Integrating and embedding the evaluation system goals into large state and district reform initiatives will facilitate coherence and strengthen the system's credibility and implementation.

Stakeholders might consider the guiding questions for Component 1a as they work to develop the overall vision and goal of the evaluation system.

Specifying Evaluation System Goals

NOTES

SYSTEM GOALS AND PURPOSES

1. Have the goals and purposes of the evaluation system been determined?

GUIDING QUESTIONS

- What type of impact do stakeholders hope to achieve (e.g., better teacher retention, improved student test scores, increased teacher capacity)?
- Will teacher evaluation results be used for personnel and compensation decisions?
- Will teacher evaluation results be used to improve teacher practice?
- Will teachers be held accountable for student academic growth?
- What type of reform efforts are most important to the teachers union? (if applicable)
- Will incentives be offered to teachers according to performance?
- Will support be available for teachers identified in need?
- What financial and human capital resources are available?
- Are state teacher performance standards established?

GOAL DEFINITION

2. Are the goals explicit, well-defined, and clearly articulated for stakeholders?

GUIDING QUESTIONS

- Are the goals stated in measurable terms?
- Can a model of teacher evaluation conceivably meet these goals?
- Do all the training and explanatory materials portray a consistent message?

GOAL ALIGNMENT

3. Have the evaluation system goals been aligned to the state strategic plan or other teacher reform initiatives?

GUIDING QUESTIONS

- Are there other teacher quality initiatives occurring within the state?
- How will the efforts in teacher evaluation affect other quality initiatives (e.g., curriculum, professional learning, certification)?
- How can reform efforts be aligned to create a coherent system?
- Is there flexibility for district input/alignment with district initiatives?

Establishing Standards

After the goals and purposes of an evaluation system are determined, the state or district needs to ensure alignment between these goals and teacher standards. This task often begins with defining the term *effective teacher,* then breaking that definition into teacher standards, competencies, and achievement-related outcomes (see "Defining Teacher Effectiveness" on page 12). Most states already have teacher standards in place, for use in hiring and traditional evaluation processes. However, as outlined previously, Race to the Top requirements and potential forthcoming mandates in the reauthorization of ESEA demand evaluation systems with the capacity to determine teacher effectiveness through measures of teacher performance and student growth. It is important, therefore, that teacher standards not only define what is valued in a teacher but indicate knowledge, skills, or practices that can be measured reliably, correlated to student growth, and aligned with professional learning opportunities.

Finally, the standards or criteria should include concise descriptions so that the broad statements are clearly articulated in a meaningful way at the practitioner level for shared understanding. These standards will form the basis from which definitions of desired behaviors can be created—the rating scale for the evaluation system.

Many states have referred to the National Board for Professional Teaching Standards (2014) for reference in standard development. In addition, the Council of Chief State School Officers Interstate Teacher Assessment and Support Consortium (InTASC) released its *Model Core Teaching Standards: A Resource for State Dialogue* in April 2011.

These standards are an update of the 1992 InTASC Standards that were primarily designed for licensing new teachers (see Council of Chief State School Officers, 2011). The professional practice standards can be used for all stages of a teacher's career. Both sets of standards have either been adopted or used as the basis for standards development.

State stakeholders might consider the guiding questions for Component 1b (see page 13) as they develop or revise teacher standards.

DEFINING TEACHER EFFECTIVENESS

GTL Center Definition

- "Effective teachers have high expectations for all students and help students learn, as measured by value-added or other test-based growth measures, or by alternative measures.
- "Effective teachers contribute to positive academic, attitudinal, and social outcomes for students such as regular attendance, on-time promotion to the next grade, on-time graduation, self-efficacy, and cooperative behavior.
- "Effective teachers use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed; and evaluate learning using multiple sources of evidence.
- "Effective teachers contribute to the development of classrooms and schools that value diversity and civic-mindedness.
- "Effective teachers collaborate with other teachers, administrators, parents, and education professionals to ensure student success, particularly the success of students with special needs and those at high risk for failure" (Goe et al., 2008, p. 8).

Federal Definition

"*Effective teacher* means a teacher whose students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth (as defined in this notice). A method for determining if a teacher is effective must include multiple measures, and effectiveness must be evaluated, in significant part, on the basis of student growth (as defined in this notice). Supplemental measures may include, for example, high school graduation rates (as defined in this notice) and college enrollment rates, as well as evidence of providing supportive teaching and learning conditions, strong instructional leadership, and positive family and community engagement" (Secretary's Priorities for Discretionary Grant Programs, 2010, p. 47288).

"*Student growth* means the change in student achievement (as defined in this notice) for an individual student between two or more points in time. A State may also include other measures that are rigorous and comparable across classrooms" (Secretary's Priorities for Discretionary Grant Programs, 2010, p. 47290).

"*Student achievement* means—

"(a) *For tested grades and subjects:* (1) A student's score on the State's assessments under the ESEA; and, as appropriate, (2) other measures of student learning, such as those described in paragraph (b) of this definition, provided they are rigorous and comparable across schools.

"(b) *For non-tested grades and subjects:* Alternative measures of student learning and performance, such as student scores on pre-tests and end-of-course tests; student performance on English language proficiency assessments; and other measures of student achievement that are rigorous and comparable across schools" (Secretary's Priorities for Discretionary Grant Programs, 2010, p. 47290).

Establishing Standards

NOTES

DEFINITION OF EFFECTIVE TEACHER

1. Has the state defined what constitutes an effective teacher?

GUIDING QUESTIONS

- Will the state or district go beyond a teacher's ability to improve student learning in its definition of an effective teacher?
- Will the use of evidence-based teaching practices be a factor in identifying an effective teacher?
- Will behavioral and social outcomes be a factor in identifying an effective teacher?
- Will effective collaboration be a contributing factor in identifying an effective teacher?
- Will a teacher's professionalism be a factor in identifying an effective teacher?
- What characteristics, behaviors, and values should a highly effective teacher demonstrate?
- What type of classroom environment should a teacher create in his or her classroom?
- Should a highly effective teacher demonstrate leadership? If so, what might that look like?
- What content knowledge do the teachers need to translate to their students?

TEACHING STANDARDS

2. Has the state established teaching standards?

GUIDING QUESTIONS

- Are there existing state teaching standards that can be used to guide system development?
- Are the standards written in a manner that reflects measures of teacher performance and student growth?
- Do the standards explicitly define desired teaching competencies?
- Have levels of teaching performance been established within the standards?
- Have the standards been written in a manner in which evaluation system results will yield reliable information on teacher performance according to the identified standards?
- Have sample performance indicators been developed to provide examples of observable behavior?
- Was public comment a step in developing teaching standards?

COMPONENT 2

Securing and Sustaining Stakeholder Investment, and Cultivating a Strategic Communication Plan

Stakeholder Investment

Evaluation systems are much more likely to be accepted, successfully implemented, and sustained if stakeholders are included in the design process. Stakeholder involvement throughout the design, implementation, assessment, and revision of teacher evaluation systems increases the likelihood that the system is perceived as responsive, useful, and fair. Potential stakeholder representation could include the following:

- Teachers (including various levels, content areas, and specialists)
- Teacher union representatives
- Related services personnel
- School board members
- Superintendents
- School principals
- Teacher preparation programs, parents
- Students
- Business and community leaders

Involving teachers in the initial stages of development and throughout the implementation process will likely increase teachers' collaboration, support, and promotion of state and district efforts. Teachers are in the best position to inform this process because they can discern what will work in their classrooms.

Clarifying expectations in terms of stakeholder purpose, level and duration of commitment, and authority in decisions will assist in sustaining stakeholder investment throughout the process. Individual members bring to bear unique sets of skills, experiences, and interests, and the level of involvement of each stakeholder may shift during the process of designing and implementing the evaluation system. Defining stakeholders' roles and responsibilities, while capitalizing on their expertise, may cultivate a high level of active participation. Stakeholders could play an integral role in the following tasks:

- Determining the standards and criteria for the system
- Mobilizing administrator, teacher, and community support
- Facilitating practitioner groups to obtain input and feedback
- Marketing the system and publicizing the findings
- Interpreting policy implications
- Investigating and/or securing federal, state, or private sector funding

RESOURCE

Communication Framework for Measuring Teacher Quality and Effectiveness: Bringing Coherence to the Conversation
http://www.gtlcenter.org/sites/default/files/docs/NCCTQCommFramework.pdf

This framework can be used by regional comprehensive center staff, state education agency personnel, and local education agency personnel to promote effective dialogue about the measurement of teacher quality and effectiveness. The framework consists of the following four components: communication planning, goals clarification, teacher quality terms, and measurement tools and resources.

Communication Plan

Early on in the process, stakeholders should consider communication needs. A strategic communication plan detailing steps to inform the broader school community about implementation efforts, results, and future plans may increase the potential for statewide adoption. Misperceptions and opposition can be minimized if the state and districts communicate a clear and consistent message. A strategic communication plan first identifies the essential messages and audiences. Potential key audiences could include pilot participants, school personnel, families, and the external community.

Stakeholders would then determine the most effective channel of communication for its purpose and target audience. Written, spoken, and/or electronic communication strategies may include the following:

- Online communications
- Community information nights
- Quarterly memos
- Weekly e-mail updates
- Media relations materials
- Word of mouth
- Events
- Workshops
- Videos
- CDs
- Press releases
- Newsletters

Communication plans should take into account the duration of the process of improving the evaluation system including its initiation and all implementation phases. For example, communication needs during the design of the system will be different from those during implementation and the process of gathering feedback. Plans should include updates on efforts to build the evaluation system, celebrations of successes as the work moves forward, and recognition of stakeholder contributions. Communicating success in terms of implementation efforts, changes in teacher practice, and student outcomes can be a powerful way to ensure buy-in and secure stakeholder investment. Highlighting successes also reinforces, inspires, and energizes teachers.

Stakeholders might consider the guiding questions for Component 2 as they develop a strategic communication plan.

Securing and Sustaining Stakeholder Investment, and Cultivating a Strategic Communication Plan

STAKEHOLDER GROUP

1. Has the stakeholder group been identified for involvement in the design of the evaluation model?

GUIDING QUESTIONS

- Who are the crucial stakeholders (e.g., teachers, union representatives, teacher preparation faculty, community members, administrators, state and local personnel and leaders)?
- What state departments will be affected by changes in teacher evaluation? Should a representative be included in the stakeholder group?
- Are other stakeholder groups already established? If yes, could one of these groups be expanded to include teacher evaluation work?
- Do representatives from other stakeholder groups need to be appointed to this stakeholder group to ensure that reform efforts are aligned?

NOTES

GROUP ROLES AND EXPECTATIONS

2. Have the group expectations and individual roles been established?

Group Expectations

GUIDING QUESTIONS

- Will the group have authority in making decisions, or will it serve in an advisory capacity?
- What is the group's purpose? Will it help design the system, provide recommendations, and/or provide approval?
- What level of commitment will stakeholders be required to make (e.g., how frequently the team will meet, for how many months)?
- Does legislation dictate the work of the stakeholder group?

Stakeholder Roles

GUIDING QUESTIONS

- What skills, experience, and knowledge does each stakeholder bring to the table?
- What roles need to be filled (e.g., marketing, mobilizing support, interpreting legislation)?
- Will some stakeholders be involved in designing the system? Communicating plans and progress? Designing research?

COMMUNICATION PLAN

3. Does the group have a strategic communication plan to keep the broader school community informed?

Content

GUIDING QUESTIONS

- What information needs to be communicated to stakeholders?
- Will pilot results be communicated?
- Will progress on the design, implementation, and success of the evaluation system be shared?
- Will teacher evaluation results be reported?

Target Audience

GUIDING QUESTIONS

- Which stakeholders should be kept informed about the development, implementation, and results of efforts related to teacher evaluation?
- Who will be the target audience (e.g., pilot participants, teachers, administrators, students, parents, community)?
- Will communication efforts be varied according to audience (e.g., board members require more detailed updates than community members)?
- How will personnel outside of the stakeholder group be kept informed?

Mode

GUIDING QUESTIONS

- Do channels of communication with stakeholders already exist?
- Does the state have a public communications department that could assist in marketing?
- What forms of communication will be utilized (e.g., website, e-mails, newsletters, public announcements)?

Timing

GUIDING QUESTIONS

- Does the plan include communication strategies throughout the development process (e.g., in the beginning, during, and after each phase)?
- Has the plan considered optimal timing for communicating evaluation efforts and results?

NOTES

FEEDBACK

4. Has the stakeholder group determined a process to ensure that constituent feedback is integrated into the systems' redesign efforts?

Who

GUIDING QUESTIONS

- From whom does the group wish to solicit feedback (e.g., pilot participants, teachers, legislators, administrators, parents)?

Methods

GUIDING QUESTIONS

- What methods will the state use to obtain feedback from affected school personnel during the design process (e.g., surveys, focus groups)?
- Are there teacher groups or electronic mailing lists that could be accessed to obtain stakeholder feedback?
- Are there teachers of certain student populations and content areas in which focus groups should be considered?
- Has the group considered an internal or external evaluation to determine the effectiveness of the system (from a teacher/principal perspective) during implementation?

Response

GUIDING QUESTIONS

- Who will consolidate the stakeholder feedback? How will it be incorporated into the redesign process?
- How will the group respond to stakeholder feedback (e.g., Q&A document, FAQ newsletter?)
- What weight will constituent feedback hold?
- Will student outcomes be considered before changes are considered?

Selecting Measures

The evaluation system's purpose and teacher standards should inform the types of outcomes and practices that will be assessed through the evaluation system, which in turn, will inform the methods and measures to be used. Selecting appropriate measures is a critical component of the design process. Measures should yield reliable information on whether teaching standards have been demonstrated and evaluation system goals have been realized.

Current federal definitions of teacher effectiveness have focused strongly on student growth. This focus was made clear in the Race to the Top competition, which required states to develop evaluation systems that "differentiate effectiveness using multiple rating categories that take into account data on student growth . . . as a significant factor" (Race to the Top application, D[2][ii], p. 34). Race to the Top guidance also indicates that multiple measures of evaluating teacher performance should be used, a belief that is echoed by the research and policy communities. Multiple measures of teacher outcomes allow for a more comprehensive view of a teacher's effectiveness based on a variety of evidence. Although summative student achievement data are relevant, data on teacher performance are most useful for targeting professional development and specifically addressing areas in which growth is needed.

According to Goe and Holdheide (2011), multiple measures:

- Strengthen teacher evaluation.
  - Provide a more complete picture of teachers' contributions to student learning.
  - Contribute to greater confidence in the results of teacher evaluations.
  - Provide more information about collaboration for student success.
- Contribute to teachers' professional growth.
  - Create opportunities for teachers to learn from their colleagues.
  - Provide teachers with greater insights into how their instruction affects student learning.
- Set the stage for improved teaching and learning.
  - Offer more complete evidence about students' learning growth, particularly in nontested subjects and grades.
  - Provide more complete evidence of learning growth for English language learners (ELLs) and students with disabilities.
  - Contribute to a more comprehensive view of students' strengths and areas in which they need improvement.

RESOURCE

Guide to Evaluation Products
http://resource.tqsource.org/GEP/

This guide can be used by states and districts to explore various evaluation methods and tools that represent the "puzzle pieces" of an evaluation system.

The guide includes detailed descriptions of more than 75 teacher evaluation tools that are currently implemented and tested in districts and states throughout the country.

The following information is provided for each tool:

- Research and resources
- Information on the teacher and student populations assessed
- Costs, contact information, and technical support offered

There are many potential measures of teacher performance that a state or district could use as part of the evaluation process. Measures of student growth provide specific feedback as to whether a student has progressed as expected in the course of a year. Potential measures include the following:

- Value-added models
- Other growth models
- Other measures (e.g., curriculum-based measures)
- Student learning objectives
- Subject specific tests

Although evidence of teacher effectiveness can be demonstrated, in part, through student growth measures, such measures are limited in distinguishing evidence of instructional quality. These measures are better able to capture teacher practice, identify learning needs, and guide professional growth. Potential measures include the following:

- Observation instruments
- Performance rubrics
- Portfolios/evidence binders
- Teacher self-assessments
- Parent/student surveys

Each measure has its inherent strengths and weaknesses (see Little, Goe, & Bell, 2009). Likewise, each measure could fulfill a particular evaluation system purpose. Therefore, measure selection is dependent on the overall purpose of the evaluation system. For instance, if the purpose of the system is to improve teacher capacity and collaboration, the selected measures might include an assortment of measures that provide evidence of instructional quality.

Table 2 reviews potential teacher evaluation goals and identifies the measurement types that are most appropriate to meet those goals. Research and policy have not suggested a particular number of measures that should comprise an evaluation "system"; however, policy does indicate that evidence of student learning should be a "significant" component within teacher evaluation. Hence, a measure of student growth is necessary to provide the "hard" data that effective instructional practices (as demonstrated through evidence of instructional quality measures) lead to student growth (as demonstrated through student growth model measures).

Table 2. Matching Measures to Specific Purposes

Purpose of Evaluation of Teacher Effectiveness	Value-Added	Classroom Observation	Analysis of Artifacts	Portfolios	Teacher Self-Reports	Student Ratings	Other Reports
Find out whether grade-level or instructional teams are meeting specific achievement goals.	X						
Determine whether a teacher's students are meeting achievement growth expectations.	X		X				
Gather information in order to provide new teachers with guidance related to identified strengths and shortcomings.		X	X	X			X
Examine the effectiveness of teachers in lower elementary grades for which no test scores from previous years are available to predict student achievement (required for value-added models).		X	X	X			X
Examine the effectiveness of teachers in nonacademic subjects (e.g., art, music, and physical education).		X		X		X	X
Determine whether a new teacher is meeting performance expectations in the classroom.		X	X	X		X	X
Determine the types of assistance and support a struggling teacher may need.		X	X		X	X	
Gather information to determine what professional development opportunities are needed for individual teachers, instructional teams, grade-level teams, etc.	X	X			X		X
Gather evidence for making contract renewal and tenure decisions.	X	X					X
Determine whether a teacher's performance qualifies him or her for additional compensation or incentive pay (rewards).	X	X					
Gather information on a teacher's ability to work collaboratively with colleagues to evaluate needs of and determine appropriate instruction for at-risk or struggling students.				X	X		X
Establish whether a teacher is effectively communicating with parents/guardians.				X			X
Determine how students and parents perceive a teacher's instructional efforts.						X	
Determine who would qualify to become a mentor, coach, or teacher leader.	X	X	X	X			X

Adoption of particular measures can be guided by the following factors:

- Evaluation system purpose
- Strength of measures
- Application to all student populations and teaching contexts
- Human and resource capacity strengths and limitations

Evaluation System Purpose

As mentioned previously, goal selection guides all decisions in the design process, particularly in measure selection. Systems designed with higher stakes (e.g., personnel dismissal and renewal decisions) point to measures that are technically defensible (e.g., valid and reliable), whereas systems designed to improve teacher capacity point to measures of instructional quality. Frequent reflection on the evaluation system's purpose will help direct measure selection.

Strength of Measures

All measures have their own inherent strengths and weaknesses. Not all measures are equally useful nor equally valid and reliable. Measures should be selected based on the following:

- Ability to accurately measure student progress
- Demonstrated impact on student achievement
- Demonstrated impact on teacher practice

Federal priorities (Secretary's Priorities for Discretionary Grant Programs, 2010) provide guidance on student growth measures stipulating that such measures must:

- Be rigorous.
- Measure progress between two points in time.
- Be comparable across classrooms.

At the same time, these measures must be valid and reliable for their intended purposes. In other words, the measure or assessment must accurately and fairly measure what the student is supposed to learn, whether the student learned the material, and how results can be attributed to individual teachers (Herman, Heritage, & Goldschmidt, 2011). Existing potential measures of student growth are not yet likely to meet all these criteria; therefore, stakeholders should factor the measure's strength in terms of the technical adequacy of the instrument as measurement selection is being considered. Likewise, measuring teacher practice through observations or a review of classroom artifacts requires trained raters so that the scores teachers receive are not dependent on who observes them or analyzes artifacts. Demonstrated validity and reliability within such measures also should guide the selection process. (*Note*: the Appendix provides an overview of measures including descriptions, research base, strengths, and cautions.)

Application of Measures to All Student Populations and Teaching Contexts

Applicability to all teaching contexts and student populations also should be considered in the measure selection process. A measure's validity and reliability with all teachers, student populations, and local contexts play an important role in maintaining implementation fidelity and

RESOURCE

Alternative Measures of Teacher Performance **(Policy-to-Practice Brief)**
http://www.gtlcenter.org/sites/default/files/docs/TQ_Policy-to-PracticeBriefAlternativeMeasures.pdf

This Policy-to-Practice Brief introduces five current examples of measures of teacher performance. The goal is to assist regional comprehensive centers and state education agencies in building local capacity to incorporate the use of alternative measures of teacher performance into the overhaul of state evaluation systems—especially in states with looming legislative deadlines.

yielding valid and useful results. For example, implementing teacher evaluation systems in rural districts may be more challenging. These districts may lack the financial and human resources to implement a system with fidelity, which will likely result in less management support and fewer resources for professional learning. Likewise, the increasing diversity of our nation's classrooms is another factor to consider in measure selection. For example, certain measures may not be appropriate or yield useful information for teachers of students with disabilities, ELLs, or gifted students. Holdheide, Goe, Croft, and Reschly (2010) address the following specific challenges in evaluating teachers of at-risk populations and measuring student growth in these populations:

- Statewide assessment results may be unavailable (e.g., students working toward alternative standards) or not viable.
- Learning trajectories may be different for students with disabilities and ELLs.
- The "ceiling effect" for gifted students may prevent adequate measurement of student growth.
- Attribution of student growth when multiple teachers are responsible for instruction and observation of teacher practice with multiple teachers in the classroom can be complicated.

Investigation into how measures apply to all teachers and contexts may increase the overall validity and reliability of measures. States need to consider these specific challenges and, if chosen, help districts develop feasible solutions to ensure successful implementation.

Human and Resource Capacity Strengths and Limitations

Each potential measure has associated expenses that need to be factored into the decision-making process. Likewise, some measures require more human capacity than others. Both human and resource capacity strengths and limitations need to be considered in the selection of measures. Implementing measures without regard to the demands they place on teachers, administrators, and others will likely yield results that lack validity or are not implemented with fidelity and thus fail to affect teacher performance and student learning.

Stakeholders may consider the following guiding questions for Component 3 during the measurement selection process.

RESOURCE

Challenges in Evaluating Special Educators and English Language Learner Specialists (Research & Policy Brief)
http://www.gtlcenter.org/sites/default/files/docs/July2010Brief.pdf

This Research & Policy Brief offers the following recommendations for states and districts::

- Include special education and ELL administrators and teachers in the process of revamping/designing evaluation models.
- Identify a common framework that defines effective teaching for all teachers. Where appropriate, include differentiated criteria for special education teachers and ELL specialists.
- In addition to—or in the absence of—appropriate standardized assessment data, incorporate other concrete evidence of teachers' contributions to student learning.

Selecting Measures

GUIDING FACTORS IN MEASURE SELECTION

1. Did stakeholders consider all the recommended factors in selecting measures?

Evaluation System's Purpose

GUIDING QUESTIONS

- Does the selected measure provide data to inform progress on the evaluation system's goals?
- Does the measure match the purpose of the evaluation?
- If necessary, does the measure provide valid and reliable data to make high-stakes decisions (e.g., dismissal)?
- Does the measure provide data on effective teaching practices and professional development needs?

Strength of Measures

GUIDING QUESTIONS

- Does the measure have research on its:
  - Ability to measure student progress?
  - Demonstrated impact on student achievement?
  - Demonstrated impact on teacher practice?
- What processes are in place (or need to be) to ensure the fidelity of the measure?
- Is the measure an accurate and fair indicator of what a student is supposed to learn?
- Is the measure an accurate and fair indicator of teacher practice?

Application to All Teaching Contexts and Student Populations

GUIDING QUESTIONS

- Do teaching context and student populations need to be differentiated to provide reliable and valid data?
- Are there specific training needs that should be considered for various teaching contexts and student populations?
- Can the measure be implemented with limited human and resource capacity?
- Can the measure of student growth be attributed accurately to multiple teachers?

NOTES

Human and Resource Capacity

GUIDING QUESTIONS

- What human and resource capacity is necessary to implement the measure reliably and with validity?
- Can resources be pulled between and within districts to implement the measure?

Specific Questions for Measuring Growth in Tested Subjects

TEACHERS' CONTRIBUTIONS TO STUDENT LEARNING GROWTH

1. Does the state intend to use teachers' contributions to student learning growth (determined using standardized test results) as a factor in teacher evaluation (e.g., value-added models and other growth models)?

Plan to Use Other Measures →

GUIDING QUESTIONS

- Will the other measures be rigorous and comparable across classrooms?
- Is there evidence that the other measures can differentiate among teachers who are helping students learn at high levels and those who are not?
- Will excluding student achievement as a factor be acceptable to the state legislature and the community?

Satisfied With Current System →

GUIDING QUESTIONS

- Are legislative changes required to implement an evaluation system that includes student growth as a component?
- Who would support or oppose linking teacher and student data? Why? How will these concerns be addressed?
- Will the other measures be rigorous and comparable across classrooms?

Plan to Use Student Achievement Growth →

NOTES

TEACHERS OF TESTED SUBJECTS

2. Has a growth model for teachers of tested subjects been selected?

GUIDING QUESTIONS

- What statistical model of longitudinal student growth will promote the most coherence and alignment with the state's accountability system? Examples: Colorado Growth Model, value-added models
- How will the state or district choose a model? Will the task force meet with experts? Will the state assessment office investigate options?
- Do these measures meet the federal requirements of rigor: *between two points in time* and *comparability?*

PERCENTAGE OF RESULTS BASED ON GROWTH MODEL

3. Has the percentage of teacher evaluation results that will be based on the growth model been determined?

GUIDING QUESTIONS

- What percentage will be supported by the education community?
- What will the state define as significant?
- Is legislation necessary to determine the percentage?
- Are the assessments reliable and valid to support a significant portion of the evaluation to be based on student progress?

IDENTIFICATION OF TEACHERS

4. Have teachers for whom the growth model will be factored into evaluation results been identified?

Teacher Inclusion/ Exclusion Criteria

GUIDING QUESTIONS

- Will all teachers of tested subjects be included?
- What is the minimum number of students required for a teacher to be evaluated with student growth (e.g., five students per grade/ content area)?
- Are there certain student populations in which inclusion in value-added or other growth models may raise validity questions (e.g., students with disabilities, ELLs)?
- Can students working toward alternative assessments be included in the growth model?
- How will the state or district choose a model? Will the task force meet with experts? Will the state assessment office investigate options?

NOTES

DATA LINKAGE

5. Can student achievement data be accurately linked to teachers (data integrity)?

Data Integrity

GUIDING QUESTIONS

- What validation process can be established to ensure clean data (e.g., teachers reviewing student lists, administrators monitoring input)?
- Can automatic data validation programs be developed?
- Are there certain student populations in which inclusion in value-added or other growth models is not appropriate (e.g., students with disabilities, ELLs)?

Teaching Context/ Extenuating Circumstances

GUIDING QUESTIONS

- Has the teacher attribution process been established for coteaching situations?
- How will teachers with high student absenteeism rates or highly mobile students be evaluated?
- Has a focus group been held with teachers to determine fair attribution?

DETERMINATION OF ADEQUATE GROWTH

6. Has a process been established to determine adequate student growth?

GUIDING QUESTIONS

- What does the research suggest regarding the number of years teacher data should be collected in order to use it as part of teacher evaluation?
- Will the learning trajectory be different for at-risk, special needs, or gifted students?
- Has the "ceiling effect" been addressed?
- Will the use of accommodations affect the measure of student growth?
- Does this measure meet the federal requirements of rigor: *between two points in time* and *comparability?*

Specific Questions for Alternative Growth Measures in Tested and Nontested Subjects

MEASURES OTHER THAN STANDARDIZED TESTS

1. Does the state intend to use measures other than standardized tests to determine student growth (e.g., classroom-based assessments; interim or benchmark assessments; curriculum-based assessments; the Four Ps: projects, portfolios, performances, products)?

Plan to Use Measures Other Than Standardized Tests but Not Student Achievement Growth

Satisfied With Our Current Evaluation System

GUIDING QUESTIONS

- Will the measures be rigorous and comparable across classrooms?
- Is there evidence that the other measures can differentiate among teachers who are helping students learn at high levels and those who are not?
- Will excluding student achievement as a factor be acceptable to the legislature and the community?

Plan to Include Student Achievement Growth

GUIDING QUESTIONS

- Are legislative changes required to implement an evaluation system that includes student growth using other measures as a component?
- What would be the challenges of using other measures of growth besides standardized assessment data?
- Will the measures other than standardized tests be rigorous and comparable across classrooms?

NOTES

NOTES

IDENTIFICATION OF TEACHERS

2. Have the teachers who meet the criteria for use of measures other than standardized tests been identified?

GUIDING QUESTIONS

- Will all teachers (in both tested and nontested subjects) be evaluated with alternative growth measures? Only teachers of nontested subjects?
- Which teachers fall under the category of nontested subjects?
- Are there teachers of certain student populations or situations in which standardized test scores are not available or appropriate to utilize?
- Will contributions to student learning growth be measured for related services personnel?

IDENTIFICATION OF MEASURES

3. Have measures to determine student learning growth been identified?

Content Standards

GUIDING QUESTIONS

- Do content standards exist for all grades and subjects?
- Is there a consensus on the key competencies that students should achieve in the content areas?
- Can these content standards be used to guide selection and development of measures?

Measure Selection

GUIDING QUESTIONS

- Which stakeholders need to be involved in determining or identifying measures?
- What type of meetings or facilitation will stakeholder groups require to select or develop student measures?
- How will growth in performance subjects (e.g., music, art, physical education) be determined to demonstrate student growth?
- Will the state use classroom-based assessments, interim or benchmark assessments, curriculum-based assessments, and/or the Four Ps (i.e., projects, portfolios, performances, products) as measures?
- Are there existing measures that could be considered (e.g., end-of-course assessments, DIBELS, DRA)?
- Could assessments be developed or purchased?

FEDERAL REQUIREMENTS

4. Do these measures meet the federal requirements of rigor: *between two points in time* and *comparability?*

Validity and Reliability

GUIDING QUESTIONS

- Does the measure accurately and fairly measure what the student is supposed to learn?
- Does the measure assess what it is intended to assess?
- Can the measure accurately indicate levels of student growth in the course of a year?
- Can student growth be accurately linked to teachers' efforts?
- Are there appropriate assessments for all grades and all teachers, including special educators and ELL specialists?

RESEARCH

5. Are there plans to conduct research during implementation to increase confidence in the measures?

GUIDING QUESTIONS

- Are federal, state, or private funds available to conduct research?
- How will the content validity be tested?
- Can national experts in measurement and assessment be appointed to assist in conducting this research?

Specific Questions for Observation Measures

MEASURE OF INSTRUCTIONAL QUALITY

1. Does the state intend to use measures other than observations as indicators of instructional quality?

GUIDING QUESTIONS

- If observations will not be used, how will the results from other measures be used to guide and strengthen teacher practice?
- Will the other measures be able to detect teacher strengths and weaknesses?
- Will the other measures be able to identify effective teaching practices?
- Will the other measures be able to identify professional development needs?

NOTES

RESEARCH BASE

2. Is there a research base for this observation tool?

GUIDING QUESTIONS

- Has the tool/instrument been piloted?
- Can results from the tool/instrument be correlated with improved student achievement?
- Have any research studies been conducted on this tool/instrument?

APPLICABILITY

3. Is the observation instrument applicable to all teachers and teaching contexts?

GUIDING QUESTIONS

- Is there any teacher population that requires differentiation in the observation process? For example, do teachers of special populations (e.g., special education students, ELLs) require different instruments and/or different observers?
- Will teachers serving in a coteaching capacity need to be observed with a different or modified tool, or will specialized training be required for evaluators to appropriately use the tool in these settings?
- Will teachers of specific content areas benefit from a more specialized tool that focuses on evidence-based practices in the content area?

PROCESS

4. Has the observation process been thoroughly specified?

Evaluators

GUIDING QUESTIONS

- Who will conduct the teacher observations (e.g., administrators, master teachers, peers)?
- Could expert teachers be appointed to conduct the observations?
- Will building administrators have the time and expertise to conduct the observations?
- Will more than one evaluator observe each teacher?

Frequency

GUIDING QUESTIONS

- How often will observations be required? Will the frequency vary depending on teachers' levels of experience?

Training and Interrater Reliability

GUIDING QUESTIONS

- What training and/or certification will be required to qualify as an evaluator?
- How will the district or state ensure that evaluators can use the observation instrument with fidelity?
- How will the district or state ensure interrater reliability? During training? Over time?

Teacher Reflection

GUIDING QUESTIONS

- Will teachers have access to all observation forms and materials in advance?
- Will teachers' self-assessments on the instruments (to be compared to the evaluator's assessment) be part of the process?
- Will preobservation and/or postobservation conferences be conducted?
- How will the observation instruments support teachers in reflecting on their practice?

COMPONENT 4

Determining the Structure of the Evaluation System

When determining the structure of the system, stakeholders must consider the designated levels of performance; the frequency of evaluations, as applicable; and a number of other factors related to implementation. In designating the number and description of levels, states must ensure that the level designations (e.g., *developing, proficient, exemplary*) work for teachers at different experience levels. Likewise, the instruments must be sensitive enough to identify the appropriate level of reliability.

In addition, it is important that the frequency of evaluation is considered separately for each measure used. Classroom observations, for example, are often conducted several times throughout the year, whereas analyses of teacher artifacts may be performed at a different frequency. The teacher's level of performance or experience also may be a factor in determining the appropriate frequency of evaluation. Beginning teachers, or teachers with identified areas of weakness, may be evaluated more frequently than teachers who have reached exemplary or master status.

States may elect to mandate specific format requirements or allow for local flexibility. When making these determinations, states should consider implementation fidelity and reliability, local bargaining restraints, and resource limitations.

As mentioned previously, all measures are not equally reliable and useful. States also may want to consider the measure's strength in comparison with the other measures used within the evaluation system. Measures that have higher validity and reliability may be used with more confidence. The measure's weight within a system may be dependent on its validity, its impact on student achievement, the information it provides to help teachers improve their practice, or other considerations. In some scenarios, states may gradually increase the weight of a measure as confidence in the measure increases and technical rigor is enhanced. For instance, states may determine that current assessments have not been validated for the purposes of teacher evaluation. In this case, data need to be collected and analyzed and compared with other types of evidence to determine whether the results are valid. As the system is evaluated and results, which increase or decrease confidence in the measures, are obtained, the weights may need to be revisited. The measure's weight also may be reflective of the evaluation system's goals. If collaboration between teachers is a priority, a rubric measuring teacher capacity to collaborate may be weighted more heavily. Or if the ultimate goal of the system is to increase teacher capacity to implement evidence-based practices, the observation instrument may carry more weight.

RESOURCE

Teacher Evaluation Models in Practice
http://resource.tqsource.org/evalmodel/

This interactive online resource responds to the need for detailed information about the design, implementation, and delivery of teacher evaluation models in practice in districts and states. It includes an overview of district evaluation models with links to their documentation, tools, training materials, and resources. It also contains lessons learned from an in-depth examination of district efforts by national experts in measurement and assessment.

Stakeholders might consider the guiding questions for Component 4 as they determine the structure of the evaluation system.

Determining the Structure of the Evaluation System

MULTIPLE MEASURES

1. Will the state promote or use multiple measures?

GUIDING QUESTIONS

- Will a single measure be sufficient in making defensible decisions regarding teacher effectiveness?
- Will a single measure accurately capture teacher capacity in terms of ability to elicit improved student achievement and implement evidence-based instructional strategies?

NOTES

WEIGHT OF MEASURES

2. Has the state determined the percentage (weight) of each measure in the overall teacher rating?

GUIDING QUESTIONS

- Will each measure be weighted differently depending on:
  - Its relation to student achievement?
  - Its reliability and validity?
  - Its face validity?
- Will the weight of each measure fluctuate depending on the level of reliability and validity that is proven over time?
- Will the weight of each measure vary depending on teaching discipline and context?

LEVELS OF PROFICIENCY

3. Have the levels of teaching proficiency been determined?

GUIDING QUESTIONS

- How many levels of proficiency can be explicitly defined?
- Can rubrics be developed to ensure fidelity?
- How often can data be generated?
- What implementation limitations should be considered (e.g., how frequently assessments can be conducted)?
- Will baseline data be analyzed prior to making decisions regarding teacher proficiency levels?

FAILURE TO MEET PERFORMANCE LEVELS

4. Have consequences been determined for failure to meet acceptable performance levels?

GUIDING QUESTIONS

- Are opportunities for teachers to improve going to be embedded in the evaluation cycle?
- Are the measures technically defensible to make personnel and compensation decisions?
- Will teacher supports be provided to assist teachers with unacceptable performance?
- How much time and assistance will be provided for a teacher to demonstrate improvement before termination is considered?
- Will teacher performance affect tenure?

NOTES

Selecting and Training Evaluators

Most evaluation measures require some level of training. The amount of training required to implement the evaluation system is highly dependent on the type of measure being considered. For example, value-added measures of student growth would require training related to the technical aspects of the system and how the data can be interpreted. Observations would require a substantial investment in training for evaluators to ensure interrater reliability as well as training for teachers and administrators in using the results to inform practice. States need to consider their own human capacity strengths and limitations in making decisions about measurement types to ensure that implementation fidelity is maintained. Moreover, local capacity limitations should be considered. For example, it may be unrealistic to mandate an evaluation system that requires a large investment in training raters if state and district budgets are tight. Districts may need flexibility in funding and implementing evaluation models with the resources they have.

Implementation fidelity is most important when the selected measures are dependent on human scoring with observation instruments or rubrics. Effective evaluator selection and training is essential if the integrity of the system is to be maintained, ensuring that the resulting scores are fair and defensible. Including targeted evaluator training with explicit decision rules and examples of evidence that would justify one performance rating over another may help with interrater reliability. Training, coupled with feedback and support, will likely lead to a high level of integrity.

Likewise, with measures dependent on personnel, evaluators may have difficulty when observing someone outside of their area of expertise. Most observation instruments (e.g., Charlotte Danielson's Framework for Teaching, Classroom Assessment Scoring System, and others) are designed to evaluate all teachers without regard to content area. However, trained evaluators with knowledge of specialist roles and subject-matter competence may be seen as more credible and pick up on nuances in instruction that other raters would miss. States could use mentors or teacher leaders with expertise in content areas as evaluators to ensure appropriate frequency, duration, and feedback related to content/discipline.

Stakeholders might consider the guiding questions for Component 5 during the evaluator selection and training process.

Selecting and Training Evaluators

PERSONNEL

1. Do the selected measures require trained personnel to use rubrics or other sources of documentation to determine the level of teacher effectiveness?

GUIDING QUESTIONS

- If personnel are not utilized to determine teacher proficiency, are there other personnel training needs (e.g., interpreting value-added scores, tracking progress-monitoring data)?

NOTES

TRAINING AND GUIDELINES

2. Will the state provide training or guidelines on evaluator/reviewer selection and training?

Selection

GUIDING QUESTIONS

- What criteria will be used to select evaluators or reviewers?
- Who will be eligible to conduct the evaluations?
- Which personnel will conduct evaluations or approve student learning targets?
- Will the state require evaluators or reviewers to have content knowledge and/or experience in the subject area/level being evaluated?
- Could teacher-to-teacher evaluations or reviews be considered?

Training

GUIDING QUESTIONS

- How will the state ensure implementation fidelity?
- Will the state offer specialized training for the evaluation of or review of specific content or specialty area teachers?
- To what extent will the training provide opportunities for guided practice paired with specific feedback to improve reliability?
- Will the state provide examples and explicit guidance in determining levels of proficiency and approval?

RETRAINING

3. Does the state have a system in place to retrain evaluators or reviewers if the system is not implemented with fidelity?

GUIDING QUESTIONS

- How will the state address personnel time limitations for conducting evaluations or reviews?
- If evaluators or reviewers are not implementing the system with fidelity, what mechanisms will be in place to retrain them?
- Will evaluators or reviewers be monitored regularly for checks in reliability?

COMPONENT 6

Ensuring Data Integrity and Transparency

Data infrastructure that can be used to collect, validate, interpret, track, and communicate teacher performance data will be necessary to inform stakeholders, guide professional learning, and assess the measures and the teacher evaluation system as a whole. The evaluation system goals can guide this development and influence the required data elements.

An integral step in this process is ensuring that the data are sound. Data integrity is crucial in all types of data-based decision making—whether making high-stakes personnel decisions or targeting professional learning activities. Verifying and cleaning existing data and establishing the means to collect the needed data elements require a thorough understanding of available and potential data sources. Therefore, collaboration between teachers (who know their students and their classrooms) and information technology personnel (who know the data) to structure the data collection will lead to greater accuracy.

Transparency of measures and resulting data is also a key factor in measure selection. Measures that provide real-time feedback, are accessible and easily understood, and have direct application to teacher practice are more likely to have an immediate impact on teaching and learning. If teachers and administrators are expected to enter information into data portals, ensuring that these portals are user-friendly will be critical as states scale up evaluation efforts.

Stakeholders might consider the guiding questions for Component 6 to ensure data integrity and transparency.

Ensuring Data Integrity and Transparency

DATA INFRASTRUCTURE

1. Is the data infrastructure to collect teacher evaluation data established?

GUIDING QUESTIONS

- Does the state or district have the data infrastructure to link teachers to individual student data including unique identifiers for both teachers and students?
- Have the critical questions that stakeholders want the evaluation system to answer been identified? Will the data system collect sufficient information to answer them?
- Have information technology personnel been brought into the discussion?
- Do districts have the technology and human capacity to collect data accurately?

NOTES

DATA VALIDATION

2. Is there a data validation process to ensure the integrity of the data?

Validation

GUIDING QUESTIONS

- What validation process can be established to ensure clean data (e.g., teachers reviewing student lists, administrators monitoring input)?
- Have criteria been established to ensure teacher/student confidentiality?
- Can computerized programs be used/developed for automatic data validation?

Training

GUIDING QUESTIONS

- What training will personnel need to ensure accurate data collection?
- Which personnel at the state and district levels will require training to ensure accuracy in data entry and reporting?

NOTES

REPORTING

3. Can teacher evaluation data be reported (aggregated/ disaggregated) to depict results at the state, district, building, or classroom levels?

Teacher Data

Student Data

GUIDING QUESTIONS

- Do administrators or teachers have access to the teacher evaluation data?
- Is there a system whereby teachers or administrators can make changes when errors are found?
- Is the data collection methodology or database easily understood and user-friendly?
- Have teachers been trained to extrapolate and use the data to inform teacher practice?
- Are administrators, teachers, and parents (as appropriate) trained in how to use the database and interpret teacher evaluation results?

USE OF DATA

4. Is there a plan for how the teacher evaluation data will be used?

Data Sharing

GUIDING QUESTIONS

- How frequently should teacher evaluation data be shared with the education community?
- What teacher evaluation data would be relevant, easily understood, and appropriate to share with the education community?
- Will administrators and teachers have access to the teacher evaluation data?
- How will evaluation results be shared with the community (e.g., website, press releases, town meetings)?

Data Use

GUIDING QUESTIONS

- Will teacher evaluation data be used to inform changes in the teacher evaluation design?
- Will administrators, teachers, and parents (as appropriate) be trained in how to use the database and interpret teacher evaluation results?
- Will data be used to identify teachers in need of support and target professional learning?
- Will data be used to identify highly effective teachers and potential mentors?

Using Teacher Evaluation Results

Selecting Trigger Points for Action

If a state plans to use its evaluation system for personnel decisions, designations of when action will be triggered need to be determined and communicated to the teacher workforce. For example, if evaluation results are tied to teacher advancement, will the teacher need to achieve exemplary ratings for three consecutive evaluation cycles prior to promotion? Will achieving exemplary ratings during two of four cycles trigger advancement? If ameliorative action is indicated, in how many evaluation cycles will improvement be expected?

Targeting Professional Development

Using evaluation results to support professional learning is likely the most significant phase of the evaluation cycle. An evaluation system's capacity to reliably identify highly effective and ineffective teachers is important. However, ensuring that teacher ratings can reliably detect teacher strengths and weaknesses is essential for accurately targeting professional development. Evaluation results can then be used to identify individual, school, and districtwide needs; target professional learning; gauge teacher growth; and identify potential mentors. Providing job-embedded, ongoing, individualized professional learning and support is necessary for teacher evaluation to have positive impacts on teacher practice.

As professional development is incorporated into the evaluation cycle, stakeholders need to evaluate outcomes to determine whether the efforts have improved teaching practice. This process goes beyond a simple evaluation of the professional learning activity, moving toward a continual, longitudinal reflection and analysis of teacher participation, support, and outcomes related to student achievement. Investing in the technical infrastructure to collect, link, and analyze professional development and teacher evaluation results over time may improve the overall effectiveness of professional learning efforts.

Stakeholders might consider the guiding questions for Component 7 as they contemplate professional development needs.

RESOURCE

Job-Embedded Professional Development: What It Is, Who Is Responsible, and How to Get It Done Well
http://www.gtlcenter.org/sites/default/files/docs/JEPD%20Issue%20Brief.pdf

This Issue Brief provides specific recommendations for states to support high-quality job-embedded professional development (p. 10):

- "Help build a shared vocabulary."
- "Provide technical assistance."
- "Monitor implementation."
- "Identify successful job-embedded professional development practices within the state."
- "Align teacher licensure and relicensure requirements with high-quality job-embedded professional development."
- "Build comprehensive data systems."

Using Teacher Evaluation Results

NOTES

TRIGGER POINTS FOR ACTION

1. Have trigger points for action using evaluation results been established?

GUIDING QUESTIONS

- Does the state intend to align evaluation results to human resource decisions?
- At what point will evaluation results warrant a promotion or dismissal?
- How many evaluation cycles will be used to ensure that opportunity for professional growth is provided?
- How will evaluation results be shared with teachers? When will teachers be notified of next steps toward professional growth or termination?

EVALUATION CYCLE

2. Is professional development an integral component of the evaluation cycle?

GUIDING QUESTIONS

- Is a goal of the evaluation system to improve teacher capacity? If so, how will the evaluation system affect teacher practice?
- Will teachers identified as ineffective have sufficient opportunities and support to improve before termination is considered?
- Will personnel decisions be defensible if teachers were not provided an opportunity and the resources to improve?
- What resources, including time and personnel, are dedicated to teacher improvement?

EVALUATION RESULTS

3. Will teacher evaluation results be used to target professional development activities?

GUIDING QUESTIONS

- How will professional development opportunities be determined for teachers, schools, and the district?
- How will data obtained through the various teacher evaluation measures inform professional development offerings?
- How can the evaluation system be retooled to reliably detect teacher strengths and weaknesses?
- Can teacher evaluation results be used to identify teachers for roles such as mentor teachers, master teachers, and consulting teachers?

RESEARCH

4. Are professional learning activities provided in a manner that is supported in research?

GUIDING QUESTIONS

- What human and fiscal resources can be used to provide job-embedded professional development?
- Can teacher application and reflection be built into the professional learning activity?
- Are professional learning activities "job-embedded" or a one-time-only session?
- Do teachers have common planning times to reflect upon new practices?
- Can opportunities for teachers to observe effective teachers be provided?
- Will professional learning communities be established?

NOTES

EVALUATION SYSTEMS

5. Are systems established to evaluate professional learning efforts?

Evaluating the Training

GUIDING QUESTIONS

- What mechanism will be established to ensure that participant feedback is obtained (e.g., training evaluation, follow-up survey)?
- What procedures will be established to ensure that active participation and application are integral parts of the professional development activity?

Reviewing the Outcomes

GUIDING QUESTIONS

- Can the evaluation measure(s) detect teacher growth as a result of professional development efforts?
- Can demonstrated teacher growth be correlated to improved student achievement?
- What mechanism will be established to follow up on teachers to ascertain whether teacher practice has been improved as a result of the professional learning efforts (e.g., follow-up survey or observation)?

Modifying the Process

GUIDING QUESTIONS

- Can the system identify which professional learning opportunities are or are not effective?
- Are changes in the evaluation system necessary to correlate teacher and student growth with participation in professional learning activities?
- How will results (e.g., evaluations and outcomes) be used to improve professional development offerings and strategies?

Evaluating the System

Systematically evaluating the performance of the evaluation model in terms of its goals and results and modifying its structure, processes, or format accordingly ensures system efficacy and sustainability. States need to identify the factors that will determine whether the system is effective.

For example, the state and districts will want to know whether:

- Stakeholders value and understand the system.
- Student performance is improved.
- Teacher practice is affected.
- Teacher retention is improved.
- The system is implemented with fidelity.

States have used external and internal review processes to collect and analyze data. Surveys of teachers, administrators, and stakeholders may be valuable for this process. Ultimately, researchers should work closely with stakeholders to ensure that the design allows important questions to be answered.

Stakeholders might consider the guiding questions for Component 8 when determining the evaluation process for the system.

Evaluating the System

EVALUATION PROCESS

1. Has a process been developed to systematically evaluate the effectiveness of the teacher evaluation model?

GUIDING QUESTIONS

- How will the stakeholders know whether the new teacher-evaluation model is effective?
- Has the model been piloted, or are there plans to pilot the model prior to statewide or districtwide implementation?
- Is there a plan for securing stakeholder and participant feedback?
- Will research be conducted in conjunction with implementation to provide validation?
- Are the goals of the evaluation system a good measure of effectiveness?
- Will research be conducted to determine whether there is a correlation between growth model scores and observation ratings?

NOTES

EFFECTIVENESS OUTCOMES

2. Have outcomes been established to determine the overall effectiveness of the evaluation system?

GUIDING QUESTIONS

- Have the stakeholders identified factors that should be considered in determining whether the evaluation system is effective (e.g., participant satisfaction, improved teacher practice, other improved student outcomes)?
- Are resources available to conduct an internal or external assessment of the evaluation model?
- Has the data infrastructure been established to track data over a period of time to determine teacher and student growth?
- Have explicit benchmarks or targets been established to determine effectiveness?
- In review of baseline data, what would be acceptable performance targets?

OTHER ASPECTS OF TEACHER QUALITY

3. Will other aspects of teacher quality that affect teacher performance be reviewed to determine whether they have been influenced by the evaluation system?

GUIDING QUESTIONS

- If the teacher evaluation plan includes modifications in tenure, promotion, or compensation, how will the state conduct research to determine the level of effectiveness on teacher retention and improved teacher capacity?
- Will the teacher evaluation plan include working in collaboration with teacher preparation programs to ensure that candidates are prepared with the competencies for which they will be held accountable when they begin teaching?
- Will data be collected on teacher effectiveness to determine whether effective teachers are equally distributed throughout the state—including both high-performing and low-performing schools?
- Will research be conducted to determine whether professional development efforts have resulted in improved teacher practice and student outcomes?

Conclusion and Recommendations

Designing a comprehensive teacher evaluation system in an effective and sustainable manner is a difficult process, especially with few research-based models to consider. States are charged with overseeing this process—which for many is unfamiliar territory because, historically, evaluation in most states has been left up to districts. Using teacher evaluation to improve teacher practice in schools should be the ultimate goal of state and district efforts. Identifying areas in which teacher practice can be improved and providing targeted professional learning opportunities to teachers should go a long way toward addressing the persistent achievement gaps in our nation's schools.

Too often, teacher evaluation is seen as a mechanism for enforcing personnel decisions rather than cultivating effective teaching. Adding to the challenges of creating comprehensive teacher evaluation systems is the relationship between state and district leaders and teachers. Building trust and ensuring collaboration toward common goals requires substantial resources, including time, patience, and resilience. To further the development of direct links between teacher evaluation and instructional improvement, states and districts need to nurture an educational climate in which evaluation is considered a fair and transparent appraisal (not punitive), and teachers are highly invested in the process. The core of evaluation reform efforts should be human capacity building at all levels so that states, districts, and schools can identify and learn from top-performing teachers, support discouraged and less successful teachers, and continue to develop all teachers toward their full potential.

References

American Recovery and Reinvestment Act of 2009, Pub. L. No.111-5, 123 Stat. 115 (2009). Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_bills&docid=f:h1enr.txt.pdf

Cincinnati Public Schools. (n.d.). *Teacher evaluation* [Website]. Retrieved from http://www.cps-k12.org/about-cps/tes

Colorado Department of Education. (2013). *Colorado growth model* [Website]. Retrieved from http://www.cde.state.co.us/accountability/coloradogrowthmodel

Council of Chief State School Officers. (2011). *InTASC model core teaching standards: A resource for state dialogue.* Washington, DC: Author. Retrieved from http://www.ccsso.org/Documents/2011/InTASC_Model_Core_Teaching_Standards_2011.pdf

Danielson Group. (2011). *The framework for teaching* [Website]. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Elementary and Secondary Education Act (No Child Left Behind Act of 2001), Pub. L. No. 107-110, 115 Stat. 1425 (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

Fleischman, H. L., Hopstock, P.J., Pelczar, M. P., & Shelley, B.E. (2010). *Highlights from PISA 2009: Performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international context* (NCES 2011-004). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2011/2011004.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 from http://www.gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects* (Research & Policy Brief). Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/MeasuringTeachersContributions.pdf

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Guidance for developing and selecting student growth measures for use in teacher evaluation* (Extended Version). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/policy/TestScoresTeacherEval.pdf

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 from http://www.gtlcenter.org/sites/default/files/docs/July2010Brief.pdf

Learning Point Associates. (2010). *Evaluating teacher effectiveness: Emerging trends reflected in the state Phase 1 Race to the Top applications.* Naperville, IL: Learning Point Associates. Retrieved from http://www.learningpt.org/pdfs/RttT_Teacher_Evaluation.pdf

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., et al. (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. perspective* (NCES 2005-003). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2005/2005003.pdf

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/practicalGuide.pdf

National Board for Professional Teaching Standards. (2011). *Certificate areas* [Website]. Retrieved from http://www.nbpts.org/certificate-areas

New York State Education Department. (2011). *Draft teacher and principal evaluation regulations.* Albany, NY: Author. Retrieved from http://usny.nysed.gov/rttt/docs/100.2subpart30-2terms-draft-for-comment.pdf

North Carolina Professional Teaching Standards Commission. (2013). *North Carolina professional teaching standards.* Raleigh, NC: Author. Retrieved from http://www.ncpublicschools.org/docs/humanresources/district-personnel/evaluation/standardsteacher.pdf

Ohio Department of Education. (2010). *Ohio teacher evaluation system: A guide to support high quality teacher evaluation.* Columbus, OH: Author.

Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

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

U.S. Department of Education. (2010). *Phase I Race to the Top application for initial funding.* Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/application.doc

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 from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Appendix. Summary of Teacher Evaluation Measures

Measure	Description	Research	Strengths	Cautions
Classroom Observation	Used to measure observable classroom processes, including specific teacher practices, holistic aspects of instruction, and interactions between teachers and students. Can measure broad, overarching aspects of teaching or subject-specific or context-specific aspects of practice.	Some highly researched protocols have been found to link to student achievement, though associations are sometimes modest. Research and validity findings are highly dependent on the instrument used, sampling procedures, and training of raters. There is a lack of research on observation protocols as used in context for teacher evaluation.	■ Provides rich information about classroom behaviors and activities. ■ Is generally considered a fair and direct measure by stakeholders. ■ Depending on the protocol, can be used in various subjects, grades, and contexts. ■ Can provide information useful for both formative and summative purposes.	■ Careful attention must be paid to choosing or creating a valid and reliable protocol and training and calibrating raters. ■ Classroom observation is expensive due to cost of observers' time; intensive training and calibrating of observers adds to expense but is necessary for validity. ■ This method assesses observable classroom behaviors but is not as useful for assessing beliefs, feelings, intentions, or out-of-classroom activities.
Principal Evaluation	Is generally based on classroom observation, may be structured or unstructured; uses and procedures vary widely by district. Is generally used for summative purposes, most commonly for tenure or dismissal decisions for beginning teachers.	Studies comparing subjective principal ratings to student achievement find mixed results. Little evidence exists on validity of evaluations as they occur in schools, but evidence exists that training for principals is limited and rare, which would impair validity of their evaluations.	■ Can represent a useful perspective based on principals' knowledge of school and context. ■ Is generally feasible and can be one useful component in a system used to make summative judgments and provide formative feedback.	■ Evaluation instruments used without proper training or regard for their intended purpose will impair validity. ■ Principals may not be qualified to evaluate teachers on measures highly specialized for certain subjects or contexts.
Instructional Artifact	Structured protocols used to analyze classroom artifacts in order to determine the quality of instruction in a classroom. May include lesson plans, teacher assignments, assessments, scoring rubrics, and student work.	Pilot research has linked artifact ratings to observed measures of practice, quality of student work, and student achievement gains. More work is needed to establish scoring reliability and determine the ideal amount of work to sample. Lack of research exists on use of structured artifact analysis in practice.	■ Can be a useful measure of instructional quality if a validated protocol is used, if raters are well-trained for reliability, and if assignments show sufficient variation in quality. ■ Is practical and feasible because artifacts have already been created for the classroom.	■ More validity and reliability research is needed. ■ Training knowledgeable scorers can be costly but is necessary to ensure validity. ■ This method may be a promising middle ground in terms of feasibility and validity between full observation and less direct measures such as self-report.

Measure	Description	Research	Strengths	Cautions
Portfolio	Used to document a large range of teaching behaviors and responsibilities. Has been used widely in teacher education programs and in states for assessing the performance of teacher candidates and beginning teachers.	Research on validity and reliability is ongoing, and concerns have been raised about consistency/stability in scoring. There is a lack of research linking portfolios to student achievement. Some studies have linked National Board for Professional Teaching Standards certification (which includes a portfolio) to student achievement, but other studies have found no relationship.	■ Is comprehensive and can measure aspects of teaching that are not readily observable in the classroom. ■ Can be used with teachers of all fields. ■ Provides a high level of credibility among stakeholders. ■ Is a good tool for teacher reflection and improvement.	■ This method is time-consuming on the part of teachers and scorers; scorers should have content knowledge of the portfolios. ■ The stability of scores may not be high enough to use for high-stakes assessment. ■ Portfolios are difficult to standardize (compare across teachers or schools). ■ Portfolios represent teachers' exemplary work but may not reflect everyday classroom activities.
Teacher Self-Report Measure	Teacher reports of what they are doing in classrooms. May be assessed through surveys, instructional logs, and interviews. Can vary widely in focus and level of detail.	Studies on the validity of teacher self-report measures present mixed results. Highly detailed measures of practice may be better able to capture actual teaching practices but may be harder to establish reliability or may result in very narrowly focused measures.	■ Can measure unobservable factors that may affect teaching, such as knowledge, intentions, expectations, and beliefs. ■ Provides the unique perspective of the teacher. ■ Is very feasible and cost-efficient; can collect large amounts of information at once.	■ Reliability and validity of self-report is not fully established and depends on instrument used. ■ Using or creating a well-developed and validated instrument will decrease cost-efficiency but will increase accuracy of findings. ■ This method should not be used as a sole or primary measure in teacher evaluation.
Student Survey	Used to gather student opinions or judgments about teaching practice as part of teacher evaluation and to provide information about teaching as it is perceived by students.	Several studies have shown that student ratings of teachers can be useful in providing information about teaching; may be as valid as judgments made by college students and other groups; and, in some cases, may correlate with measures of student achievement. Validity is dependent on the instrument used and its administration and is generally recommended for formative use only.	■ Provides perspective of students who have the most experience with teachers. ■ Can provide formative information to help teachers improve practice in a way that will connect with students. ■ Makes use of students, who may be as capable as adult raters at providing accurate ratings.	■ Student ratings have not been validated for use in summative assessment and should not be used as a sole or primary measure of teacher evaluation. ■ Students cannot provide information on aspects of teaching such as a teacher's content knowledge, curriculum fulfillment, and professional activities.

Value-Added Model	Used to determine teachers' contributions to students' test score gains. May also be used as a research tool (e.g., determining the distribution of "effective" teachers by student or school characteristics).	Little is known about the validity of value-added scores for identifying effective teaching, though research using value-added models does suggest that teachers differ markedly in their contributions to students' test score gains. However, correlating value-added scores with teacher qualifications, characteristics, or practices has yielded mixed results and few significant findings. Thus, it is obvious that teachers vary in effectiveness, but the reasons for this are not known.	■ Provides a way to evaluate teachers' contribution to student learning, which most measures do not. ■ Requires no classroom visits because linked student/teacher data can be analyzed at a distance. ■ Entails little burden at the classroom or school level because most data are already collected for NCLB purposes. ■ May be useful for identifying outstanding teachers whose classrooms can serve as "learning labs" as well as struggling teachers in need of support.	■ Models are not able to sort out teacher effects from classroom effects. ■ Vertical test alignment is assumed (i.e., tests essentially measure the same thing from grade to grade). ■ Value-added scores are not useful for formative purposes because teachers learn nothing about how their practices contributed to (or impeded) student learning. ■ Value-added measures are controversial because they measure only teachers' contributions to student achievement gains on standardized tests.

Center on
GREAT TEACHERS & LEADERS
at American Institutes for Research

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

www.gtlcenter.org

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

www.air.org

This work was originally produced in whole or in part by the Center on Great Teachers and Leaders with funds from the U.S. Department of Education under cooperative agreement numbers S283B120021 and 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 Center on Great Teachers and Leaders is administered by American Institutes for Research and its partners: the Council of Chief State School Officers and Public Impact.

1123_02/14