nea
NATIONAL
EDUCATION
ASSOCIATION
nea.org

*Great Public Schools
for Every Student*

Teacher Evaluation:

A Resource Guide for National Education Association Leaders and Staff

NEA Department Of

Teacher Quality

Table *of* Contents

Acknowledgments

The NEA Teacher Quality Department would like to thank the many contributors to this challenging initiative for their dedication and commitment. As debate at the national and state level continues to focus on effective teaching, it is critical for all education stakeholders to remain focused on the importance of developing and implementing comprehensive evaluation systems to support the NEA's vision of "great public schools for every student."
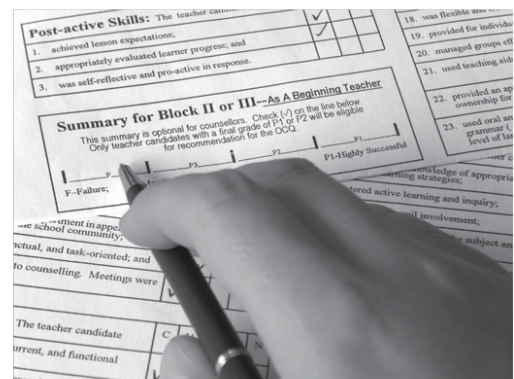
Today's Reality

As momentum to reform teacher evaluation systems continues to build at the district and state levels, America's educators need to play an active role in the policy discussion. The National Education Association (NEA) and its 3.2 million members have long advocated for valid and reliable assessments of teacher performance, and they must continue working to make certain that evaluation and assessment systems serve both teachers and students. Too often, the focus is solely on replacing underperforming teachers rather than on helping all teachers become more effective. The vast majority of teachers are serving students well, but a robust and meaningful evaluation system, by teachers and administrators, will benefit students and teachers alike.

Teacher evaluation has long been debated in policy circles, but only now are significant resources becoming available to revamp current evaluation systems. Federal funding and philanthropic contributions are now available to supplement state and local funds. In the federal Race to the Top (RTTT) state competition for $4.35 billion, for example, all 41 state applications mentioned teacher evaluation (Learning Point Associates, 2010). The Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET) project has provided an additional $45 million to strengthen approaches to teacher evaluation in several pilot districts.

State legislators have been proposing new laws and regulations pertaining to teacher evaluation, often in direct response to RTTT requirements. In a 2009 review of state legislation and regulation, NEA identified more than 25 new or proposed state laws and regulations regarding teacher evaluation. Virtually all of them focused on using evidence of student learning or achievement in the evaluation process. In some states, policymakers have consulted NEA affiliates and worked with them to develop evaluation systems that reflect a shared vision of teaching effectiveness. In other states, policymakers have developed evaluation systems with little input from the teaching profession or the national teacher unions, resulting in a lack of shared vision of teaching effectiveness.

Recent research, including the controversial and much debated study, *The Widget Effect* (The New Teacher Project, 2009), has prompted widespread activity on evaluation reform. The fact remains, however, that teacher evaluations have been sporadic, poorly designed, and unable to provide much useful information on teaching effectiveness (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2008; Duffet, Farkas, Rotherham, & Silva, 2008; Weisberg, Sexton, Mulhern, & Keeling, 2009). Contemporary thinking in the field holds that teacher evaluation should not be treated as a stand-alone process, but rather as part of a comprehensive approach to improve teaching and learning (Lasagna, Laine, & Behrstock-Sherratt, 2011). Pecheone and Chung Wei (2009) say that to do more than "tinker around the edges" will require a major commitment to innovation in teacher evaluation systems. Such a commitment, they believe, will require developing and testing evaluation systems that are feasible at a practical level and that stand up to public scrutiny. In working together to develop new teacher evaluation systems, Association leaders and state and local policymakers must recognize the comprehensive demands of their efforts.

Purpose and Principles

The purpose of the NEA's *Teacher Evaluation: A Resource Guide for Association Leaders and Staff* is to provide information and resources on the key issues that must be addressed in bargaining or advocating for a comprehensive teacher evaluation system in any district or state. Whether or not a state requires the use of student growth data as a significant component of a teacher's evaluation—and whether or not it requires the use of multiple measures of student learning—this guide will be helpful for understanding the key components of teacher evaluation as part of a comprehensive teacher growth and development system. Our goal is to help you become a better advocate for systems that are transparent, fair, comprehensive, and useful to practitioners.

This guide does not offer a single best solution for teacher assessment and evaluation, nor does it list all of NEA's recommendations on the subject. Links to NEA policies and position documents appear in the Additional Resources section on page 55. The Association supports the idea that evaluation systems must be developed at the state and local level and that they must be developed in partnership with teachers and their representatives, within the contexts of local schools and communities. While not always agreeing with every element of a locally developed system, NEA offers technical support and resources to affiliates in the spirit of helping to improve teacher effectiveness. While making no explicit policy recommendations, this guide aims to provide information to support the development of sound policies and comprehensive, robust evaluation systems.

Our aim is to help you work productively with your constituents to define effective teaching in your own local context, identify and incorporate multiple measures of effective teaching, select and adapt evaluation tools, and pilot new systems prior to launching a new system across the board.

A 2010 report from NEA's Professional Standards and Practice Committee summarizes the primary purpose for evaluating teachers:

> **"The *core purpose* of teacher assessment and evaluation should be to improve the knowledge, skills, dispositions, and classroom practices of professional educators."**

NEA supports using teacher assessment and evaluation to improve teacher practice in order to improve student learning. The association views teacher evaluation as only one component of a comprehensive teacher growth and development system.

NEA believes that teacher evaluation systems should be designed to enhance teaching practice, not to sort teachers into categories, reward those at the top, and fire those at the bottom. Unfortunately, an overly simplistic approach is often attractive to state legislators, governors, state and local board members, and other policymakers. But a simplistic approach of incentives and punishments sidesteps the more difficult and often costly work of improving teaching and learning. A very real danger of putting into place a system of scores, rewards, and punishments is that teacher morale could drop so precipitously that even highly effective teachers may decide to leave the profession at a time when teacher turnover is already a serious problem. Allowing evaluation systems to instead focus on facilitating *effective teaching* will make it possible not only to improve student learning, but also to strengthen teachers' commitment to the profession.

NEA has identified six key principles that must serve as the foundation for developing or reforming any teacher assessment and evaluation system (*Teacher Assessment and Evaluation*, NEA, 2010):

1. **Safe and open collaboration is necessary**. When assessment of teacher practices is transparent and openly collaborative, teachers can build professional communities and learn from one another. This process can occur only in nonthreatening environments of formative assessment and growth.

2. **Measures of teacher performance are most helpful and meaningful when they are based on multiple ratings and clear teaching standards**. Teachers need clear and actionable feedback based on comprehensive, transparent standards for teaching and student learning, and based on criterion-referenced assessments of teacher practice. Feedback is most useful as part of a comprehensive teacher development system. Summative evaluations of teachers should be based primarily on a single standard of effectiveness required for all teachers. Teachers who are consistently unable to meet that standard should be removed from the classroom.

3. **Evaluation systems must be integrated**. Integrated systems must link evaluation procedures with curricular standards, professional development activities, targeted support, and human capital decisions.

4. **Validated evaluation measures are essential**. Measures of teaching effectiveness need to be based on widely accepted standards that attempt to capture a range of teaching behaviors and use multiple evaluation methods.

5. **Teachers' input in determining performance and learning outcomes should be part of the evaluation process**. Although standards for teaching practice and student learning are essential, each teacher also should help to define a set of practices and student learning objectives to be assessed. Teacher input can provide vital learning goals for the unique, contextualized circumstances of each particular classroom.

6. **Assessment and evaluation systems need to be jointly created or designed, with local teacher association involvement**. The process must include teachers at the local level through collective bargaining or, where there is no collective bargaining, the organization representing teachers must agree to any assessment or evaluation system. This may be the most important principle of all. Ideals and visions need to be balanced with local context and political reality. There is no one-size-fits-all solution at a national level. Rather, NEA needs to work with its affiliates to craft local solutions based on the principles outlined in this report.

Getting Started: Defining Effective Teaching

For too long, policy discussions involving teacher evaluation have been mired in a reward-and-punishment framework characterized by the desire to: (1) measure the effectiveness of each teacher, (2) categorize and rank teachers, (3) reward those at the top, and (4) fire those at the bottom. Such a simplistic approach not only ignores the complexity of teaching but also overlooks the real purpose of teacher assessment and evaluation.

For many administrators and teachers, the notion of incorporating measures of effective teaching into teacher evaluation systems represents a huge change. Many current evaluation systems focus only on what teachers do—plan lessons, manage their classrooms, engage their students, use assessment in their instruction, and contribute to the professional community, for example. In many districts and states, however, evaluation systems are now being designed to assess not only what teachers do, but also the *outcomes* of what teachers do. In other words, evaluation systems are being developed to help teachers master content, refine their teaching skills, critically analyze their own performance as well as their students' performance, and make changes to improve teaching and learning in their classrooms. The best of these evaluation systems provide targeted support, assistance, and professional growth opportunities to match the needs of the teacher as well as the needs of students, schools, and districts.

But many questions about evaluation systems remain: What aspects of professional practice make teachers more (or less) effective? What and how much must students learn for a teacher to be deemed effective? How can real student learning be measured? If student learning is a function of teachers' collective efforts, then how do we **disaggregate** students' academic gains and attribute them to a specific teacher? In addition to academic gains, what other student outcomes should be considered? Social-emotional growth? Civic engagement? Graduation rates? The answers to these kinds of questions will inform a school system's definition of effective teaching and will drive the decisions about how to assess the desired outcomes.

Examples of Definitions of Effective Teachers and Effective Teaching

As you seek to define and describe effective teachers and effective teaching in your local context, you may find it useful to know how some researchers and scholars have done this. Some of the working definitions below focus on the teaching practice and its contribution to student learning; others identify the knowledge, skills, and dispositions that teachers need to be effective, and still others are combinations of the two approaches.

Effective teaching refers to instruction that enables all students to meet or exceed ambitious goals for student learning (adapted from Darling-Hammond, 2010). Effective teaching is in part a function of individual teacher talent, knowledge, and skills. But it is also largely influenced by the conditions in which teaching takes place: the school leadership, the quality of curriculum materials and resources, the opportunities teachers have for professional growth and learning, the size of teacher workloads, and the time teachers have to prepare, among other factors. A focus on *teaching* effectiveness rather than *teacher* effectiveness also allows for the fact that students learn from many teachers and makes use of that fact.

A report from the 2010 summit of North Carolina's National Board Certified Teachers, hosted by the Center for Teacher Quality (CTQ), says that effective teachers stimulate and nurture student

motivation, intellectual readiness, persistence, creativity, and the ability for students to apply knowledge and to work with others. Effective teachers use multiple measures of student growth that include teacher-developed assessments. Effective teachers also maintain high levels of student engagement, provide rigorous and relevant assignments, and sustain collegial professional learning communities (Byrd & Rasberry, 2011, p. 4).

The National Comprehensive Center for Teacher Quality (TQ Center) offers a comprehensive definition of effective teachers (Goe, Bell, & Little, 2008, p. 8):

1. 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.

2. Effective teachers contribute to positive academic, attitudinal, and social outcomes for students, as shown by regular attendance, self-efficacy, and cooperative behavior.

3. Effective teachers use diverse resources to plan engaging learning opportunities and structure them; to monitor **student progress** formatively and adapt instruction as needed; and to evaluate learning using multiple sources of evidence.

4. Effective teachers contribute to the development of classrooms and schools that value diversity and civic-mindedness.

5. 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.

In an August 5, 2010, Federal Register notice, U.S. Secretary of Education Arne Duncan offers a federal government definition of an effective teacher:

> E*ffective 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, multiple observation-based assessments of teacher performance.
>
> * *Student growth* means—
> the change in student achievement (as defined in this notice**) for an individual student between two or

Delaware State (NEA, Collective Bargaining Affiliate) Title 14 Education: Administrative Code

"Effective" shall mean that the teacher shall receive a Satisfactory Component Rating in at least three (3) Appraisal Components including the Student Improvement Appraisal Components:

1. Planning and Preparation
2. Classroom Environment
3. Instruction
4. Professional Responsibilities
5. Student Improvement

"Highly Effective" shall mean that the teacher has earned a Satisfactory Component rating in at least four (4) of the five (5) Appraisal Components

…and that the teacher's students achieve…more than one grade level improvement in an academic year.

more points in time. A state also may include other measures that are rigorous and comparable across classrooms.

** *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 nontested grades and subjects*: Alternative measures of student learning and performance, such as student scores on pretests 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 (U.S. Department of Education, 2010).

This proposed federal definition has raised controversy among policymakers and in the Association because it stipulates that effective teaching is based in *significant* part on measures of student learning growth, as measured by state or other standardized assessments. While the Department of Education does not define "significant," many states have set up their own definitions with legislation that requires that 50 percent or more of a teacher's evaluation be based on evidence of student growth.

NEA believes that using standardized test scores as the primary measure to determine the competency, quality, or effectiveness of any professional educator is inappropriate and cannot result in a valid assessment (see Appendix A).

NEA: Principles of Professional Practice

In 2007, the NEA drafted a set of ten principles to describe the essential knowledge, skills, and support that teachers must have to be effective. They are meant to guide teachers and local districts in working together to develop evaluation and assessment programs. The principles are as follows:

A quality teacher:

- Designs and facilitates instruction that incorporates the students' developmental levels, skills, and interests with content knowledge
- Develops collaborative relationships and partnerships with colleagues, families, and communities focused on meaningful and deep learning
- Provides leadership and advocates for students, for quality education, and for the education profession
- Demonstrates in-depth content and professional knowledge
- Participates in ongoing professional learning both individually and within the professional learning community

- Uses multiple and varied forms of assessment and student data to inform instruction, assess student learning, and drive school improvement efforts
- Establishes environments conducive to effective teaching and learning
- Integrates cultural competence and an understanding of the diversity of students and communities into teaching practice to enhance student learning
- Uses professional practices that recognize public education as vital to strengthening our society and building respect for the worth, dignity, and equality of every individual
- Strives to overcome the internal and external barriers that affect student learning

These principles are offered as guidance for developing teacher assessment policies rather than replacing or competing with professional teaching standards adopted by states, districts, or national organizations. They are not meant to distinguish between "effective" and "ineffective" teachers but rather to provide teachers with a quality continuum that can guide professional growth over a career.

Interstate Teacher Assessment and Support Consortium (InTASC): Model Core Teaching Standards: A Resource for State Dialogue

InTASC offers a set of model core teaching standards that outline what teachers should know and be able to do to ensure every K-12 student reaches the goal of being ready to enter college or the workforce in today's world.

The Model Core Teaching Standards are grounded in research. They articulate what effective teaching and learning would look like in a transformed public education system—a system that:

- Empowers every learner to take ownership of his or her own learning
- Emphasizes the learning of content and application of knowledge and skills to real world problems
- Values the differences each learner brings to the learning experience
- Leverages rapidly changing learning environments by recognizing the possibilities they bring to maximize learning and engage learners.

There are ten model core teaching standards:

Standard #1: Learner Development. The teacher understands how learners grow and develop, recognizing that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas, and designs and implements developmentally appropriate and challenging learning experiences.

Standard #2: Learning Differences. The teacher uses understanding of individual differences and diverse cultures and communities to ensure inclusive learning environments that enable each learner to meet high standards.

Standard #3: Learning Environments. The teacher works with others to create environments that support individual and collaborative learning, and that encourage positive social interaction, active engagement in learning, and self motivation.

Standard #4: Content Knowledge. The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) he or she teaches and creates learning experiences that make the discipline accessible and meaningful for learners to assure mastery of the content.

Standard #5: Application of Content. The teacher understands how to connect concepts and use differing perspectives to engage learners in critical thinking, creativity, and collaborative problem solving related to authentic local and global issues.

Standard #6: Assessment. The teacher understands and uses multiple methods of assessment to engage learners in their own growth, to monitor learner progress, and to guide the teacher's and learner's decision making.

Standard #7: Planning for Instruction. The teacher plans instruction that supports every student in meeting rigorous learning goals by drawing upon knowledge of content areas, curriculum, cross-disciplinary skills, and pedagogy, as well as knowledge of learners and the community context.

Standard #8: Instructional Strategies. The teacher understands and uses a variety of instructional strategies to encourage learners to develop deep understanding of content areas and their connections, and to build skills to apply knowledge in meaningful ways.

Standard #9: Professional Learning and Ethical Practice. The teacher engages in ongoing professional learning and uses evidence to continually evaluate his/her practice, particularly the effects of his/her choices and actions on others (learners, families, other professionals, and the community), and adapts practice to meet the needs of each learner.

Standard #10: Leadership and Collaboration. The teacher seeks appropriate leadership roles and opportunities to take responsibility for student learning, to collaborate with learners, families, colleagues, other school professionals, and community members to ensure learner growth, and to advance the profession (www.ccsso.org/News_and_Events/Press_ Releases/CCSSO_Releases Model Core Teaching Standards. html).

National Board for Professional Teaching Standards: Core Propositions and Standards

The National Board for Professional Teaching Standards'(NBPTS) long-established standards of accomplished teaching identify the qualities of effective teaching. Much research shows that students of National Board Certified teachers (NBCTs) tend to learn more than students of their noncertified counterparts (Hakel, Konig, & Elliot, 2008). NBPTS standards cover more than 25 certification areas, all based on the NBPTS's Five Core Propositions:

Proposition 1: Teachers Are Committed to Students and Their Learning.

- NBCTs are dedicated to making knowledge accessible to all students. They believe all students can learn.

- They treat students equitably. They recognize the individual differences that distinguish their students from one another and they take account for these differences in their practice.

- NBCTs understand how students develop and learn.

- They respect the cultural and family differences students bring to their classroom.

- They are concerned with their students' self-concept, their motivation, and the effects of learning on peer relationships.

- NBCTs are also concerned with the development of character and civic responsibility.

Proposition 2: Teachers Know the Subjects They Teach and How to Teach Those Subjects to Students.

- NBCTs have mastery over the subject(s) they teach. They have a deep understanding of the history, structure, and real-world applications of the subject.

- They have skill and experience in teaching it, and they are very familiar with the skills gaps and preconceptions students may bring to the subject.

- They are able to use diverse instructional strategies to teach for understanding.

Proposition 3: Teachers Are Responsible for Managing and Monitoring Student Learning.

- NBCTs deliver effective instruction. They move fluently through a range of instructional techniques, keeping students motivated, engaged, and focused.

- They know how to engage students to ensure a disciplined learning environment, and how to organize instruction to meet instructional goals.

- NBCTs know how to assess the progress of individual students as well as the class as a whole.

- They use multiple methods for measuring student growth and understanding, and they can clearly explain student performance to parents.

Proposition 4: Teachers Think Systematically About Their Practice and Learn From Experience.

- NBCTs model what it means to be an educated person—they read, they question, they create, and they are willing to try new things.

- They are familiar with learning theories and instructional strategies and stay abreast of current issues in U.S. education.

- They critically examine their practice regularly to deepen knowledge, expand their repertoire of skills, and incorporate new findings into their practice.

Proposition 5: Teachers Are Members of Learning Communities.

- NBCTs collaborate with others to improve student learning.

- They are leaders and actively know how to seek and build partnerships with community groups and businesses.

- They work with other professionals on instructional policy, curriculum development, and staff development.

- They can evaluate school progress and the allocation of resources in order to meet state and local education objectives.
- They know how to work collaboratively with parents to engage them productively in the work of the school.

NBPTS uses the Five Core Propositions to develop certificate area standards. Those standards focus on the specific knowledge, skills, dispositions, and beliefs that support accomplished teaching. They emphasize the holistic nature of teaching and recognize how a teacher's professional judgment is reflected in action. The board standards also emphasize the importance of the particular contexts of teaching.

Conclusion

Finding common ground on the definition of teacher or teaching effectiveness is helpful for building a shared vocabulary. But more importantly, it establishes a common vision of what teachers should aim for and helps school leaders better understand how to encourage teachers to achieve their professional best.

Defining and measuring effectiveness in the classroom is both difficult and complex, which makes teachers' role in this process all the more important. **The Association recognizes that no system will be perfect and that no measurement will be exact in all circumstances and for all teachers. But NEA believes that a system that provides plenty of information about a teacher will be more useful than one that does not—both in supporting teachers in improving their practice and in supporting administrators who must be certain that summative decisions are based on sufficient evidence.** Thus, as observed by Little et al. (2009) from the TQ Center, it is important to "resist pressures to reduce the definition of effective teaching to a single score obtained on an observation instrument or through a value-added model" (p. 17). Comprehensive evaluation systems must contain many lenses for looking at teachers' abilities to promote student learning.

Activity: Moving to a Shared Understanding of Effective Teaching

In creating a fair and comprehensive evaluation system, it is helpful for all stakeholders to use common language and a shared understanding of what makes teaching effective. You can use the following activity at the outset of your collaboration—or as a "pause point" in the conversation—to help identify potential areas of miscommunication or false assumptions. The exercise also can help participants commit to common goals.

Instructions

Give each team member a copy of **Tool for Moving to a Shared Understanding of Effective Teaching**, which appears on the following page. Ask each person to fill out the first column on his or her own, then ask team members to work in small groups or in pairs, (according to their positions within the Association) to fill out the second column. Ask each small group to develop consensus on three priorities, taking note of where there are irreconcilable disagreements, if any. As a final step, ask the whole group to fill in the third column. You may want to use a projector so that everyone can see what is being added to the third column. The third column forms the basis for developing a shared understanding of effective teaching among the stakeholders.

You can begin with the broad definition adapted from Darling-Hammond, 2010: "*Effective teaching* refers to instruction that enables all students to meet or exceed ambitious goals for student learning." Then, you can encourage team members to focus on specifics that reflect the local situation to help build high-quality assessment tools.

Tool for Moving to a Shared Understanding of Effective Teaching

	Individual	Small Groups or Pairs	Whole Group
• 1. Three most important *instructional practices* that constitute effective teaching	• 1. • 2. • 3.	• 1. • 2. • 3.	• 1. • 2. • 3.
• 2. Three most important *dispositions* that influence effective teaching	• 1. • 2. • 3.	• 1. • 2. • 3.	• 1. • 2. • 3.
• 3. Three most important *outcomes* that result from effective teaching	• 1. • 2. • 3.	• 1. • 2. • 3.	• 1. • 2. • 3.
• 4. Three most important types or bases of knowledge that support effective teaching	• 1. • 2. • 3.	• 1. • 2. • 3.	• 1. • 2. • 3.
• 5. The basis of the above answers, write a consensus definition of effective teaching in your district.			

Defining and Incorporating Multiple Measures of Effective Teaching

A comprehensive evaluation system should include multiple measures of effective teaching. It is not enough to evaluate teachers using a single form of measurement, such as observations, student test scores, teaching portfolios, or classroom artifacts. Using multiple measures yields a more complete picture of a teacher's effectiveness in the classroom. It helps administrators, instructional coaches, and teachers themselves better understand areas of strength and possibilities for improvement. This way, everyone can more effectively identify opportunities for support and professional growth, which will in turn enhance student growth and learning.

There is widespread consensus in the research community that student achievement data by itself provides little insight into which elements of a teacher's practice may have contributed to an increase in student learning (Braun, 2005; Darling-Hammond, 2010; Goe, 2010). In fact, few student achievement tests accurately measure all the domains of learning that teachers address in the classroom (Baker, Bartson, Darling-Hammond, Haertel, Ladd, et al., 2010; Goe, 2009). How is it possible to assess a teacher's contribution to student growth and learning using achievement tests or other high-stakes standardized tests, given that most such tests have limited potential to effectively assess student growth and learning? Conversely, is it is possible to assess a teacher's contribution to student achievement when there is an absence of student growth and learning data? If observation data is to be used to measure teaching effectiveness, then it must assess teaching practices that increase student learning.

Linda Darling-Hammond (2010) states:

> To create systems that measure and encourage teacher effectiveness, it is important to use multiple measures of practice, performance, and outcomes so that a more complete picture of practice emerges, so that assessments are fair and produce the right incentives, and so that educators are encouraged to improve what they do instead of trying to game an unfair system.

In a system for assessing teacher effectiveness, three kinds of evidence should be considered in combination with one another:

- Contributions to *growth in student learning* and other student outcomes (based on Darling-Hammond's statement about data from classroom assessments and documentation, as well as standardized tests);

- Performance on teaching assessments measuring standards known to be associated with student learning (including *teacher performance assessments and standards-based teacher evaluations*);

- Evaluation of teaching practices that are associated with desired student outcomes and achievement of school goals (through *systematic collection of evidence* about teacher planning and instruction, interaction with parents and students, and contributions to the school).

The following are examples of the types of multiple measures that may be included in a comprehensive teacher evaluation system:

- Classroom observations
- Student growth data on standardized tests
- Other student growth data, such as district-based pre- and post-assessments tied to learning standards
- Student graduation data
- Teacher artifacts, such as lesson plans, curriculum plans, student data records, student work, student formative and summative course evaluation data, minutes from course team-planning meetings, curriculum maps, and teacher reflection notes
- Teacher interviews
- Teacher contributions to a school or district, such as serving on committees, developing curriculum, mentoring, or engaging community support
- Teacher self-assessments
- Student surveys of engagement, motivation, etc.[1]
- Parent surveys of engagement, motivation, etc.
- Records of student attendance and teacher attendance

Any comprehensive teacher evaluation system that includes multiple measures must be pilot-tested to ensure fidelity of implementation and to confirm that the measures being used are valid and reliable (see sidebar for definitions of those terms). Only if a measure of effective teaching is both valid and reliable can it be trusted to measure teachers' skills, knowledge, or dispositions. In addition, comparability of measures across classrooms is an important consideration. In its Race to the Top requirements, the U.S. Department of Education says that measures of student achievement growth used for teacher evaluations must be "rigorous and comparable across classrooms" (U.S. Department of Education, 2010). In other words, a measure must not only be comparable among students within a given class; the measure must be comparable across subject areas or grade levels, or at least be as rigorous as the measures used in other subject areas and grade levels.

Validity, Reliability, and Comparability

The following key terms should come up repeatedly in your conversations about teacher evaluation:

Validity is the degree to which an instrument measures what it was designed to measure. For example, instrument designed to assess teachers' content knowledge is not valid when used to assess student engagement.

Reliability is the degree to which an instrument can measure teacher performance consistently under similar conditions. For example, an instrument designed to assess a teacher's ability to engage students should yield consistent results regardless of the context.

Comparability is the extent to which teaching effectiveness has the same meaning both within and across grade levels and schools. For example, teacher evaluation findings would have the same when applied to different classrooms and schools.)

For further details, please see the glossary.

---

1 Since survey research is always based on a sample of the population, the success of the research depends on the representativeness of the population of concern. In addition, attention must be given to the response and nonresponse rates and the reliability and validity of the responses.

Implementing comprehensive teacher evaluation systems can be costly at the start, because of the time and money required to train both teachers and their evaluators, conduct multiple observations, and assess multiple pieces of data on teacher performance. But the time and money are usually well spent, because it is important for teachers to know that judgments about their effectiveness will not be based on a single data point, such as a standardized test score collected only once a year. All stakeholders involved in developing a new teacher -evaluation process—Association members as well as district- and state-level committee members—need to understand the importance of getting a complete picture of a teacher's effectiveness.

According to Goe (2010), multiple measures achieve the following ends:

1. **Strengthen teacher evaluation**. Multiple measures are needed to provide a complete picture of the teacher's contribution to student learning, including collaboration among teachers. The use of multiple measures allows everyone to have confidence in evaluation results.

2. **Contribute to teachers' professional growth**. Multiple measures give teachers insights about their practice, and such insights create learning opportunities.

3. **Set the stage for improvements in teaching and learning**. Multiple measures help to provide comprehensive information about student learning and identify students' areas of strength and weakness. This is particularly important for teachers of untested subjects and grades and for English Language Learners and students with disabilities.

North Carolina's Online Teacher Evaluation System: Ensuring Validity, Reliability and Transparency

In 2007, the North Carolina State Board of Education and the North Carolina Professional Teaching Standards Commission collaborated to revamp teaching standards in their state. Charged by the board to align the new standards to the board's mission and goals, the commission created a set of standards that incorporate 21st century skills and learning. The new standards called for a new evaluation system. In July 2007, the commission and board asked their regional education laboratory, Mid-Continent Research for Education and Learning (McREL), to assist with the design and implementation of an evaluation instrument that could rate the levels of the standards. With the support of the North Carolina Association of Educators, a field test was carried out in the fall of 2007, a pilot was conducted during the spring of the 2007–08 school year, and in August 2008, the system was officially launched in 13 districts across the state. By its third year of implementation, the instrument was being used in every district in the state.

The evaluation system is a multifaceted, comprehensive online system that collects data, using multiple measures, throughout the year. It requires recurring observations, teacher self-assessments, and submissions of professional artifacts. Each teacher is observed by two administrators in his or her school. Although each observer uses the same rubric to code, they are strictly prohibited from conferring with one another. The system allows teachers to offer self-assessments and submit their own artifacts, which adds a sense of transparency and objectivity to the process.

Every principal and assistant principal receives training before using the evaluation instrument. Initially, the training was done at the state level, but now administrators are trained at the district level. During the field test and pilot phases of implementation, the scoring results from the observers were sent to McREL for validity and reliability testing. Plans are currently under way for McREL to release an evaluation report on the first two full years of system implementation.

Multiple Measures of Teacher Performance: Delaware's System

Delaware's teacher evaluation system is more teacher-driven than most. The system assmes that teachers have greater expertise than anyone else for determining teaching effectiveness. The system requires "group alike" teachers meet with facilitators to decide on the multiple measures and evaluation processes to be used for their specific grade or subject area. In some grades and subjects, standardized test scores must be included.

Although teachers lead the process, the state must approve teachers' recommended assessments top ensure falidity and reliability, and the state monitors how well the assessments are working over time. A key benefit of Delaware's approach is that it provides apportunities for professional growth and reflection for the teachers involved, while also creating a system that is supported by those who use it. (for more information, visit the Mid-Atlantic Comprehensive Cente website: www.macc.ceee.gwu.edu.)

Once multiple measures of effective teaching have been defined and tested in practice, the next critical design step is to determine how each measure will be weighted, based on how much it contributes to a teacher's overall performance rating. Those who are developing evaluation systems need to know how to advocate for the fair use of various measures in determining a teacher's overall evaluation. Decisions about weighting the different assessments of teacher practice are considered fair if the most relevant and reliable measures of a teacher's performance are weighted more heavily than less relevant and reliable measures (Goe, 2010). Another question to be considered is to what extent, if any, the weighting should be determined by a teacher's placement on a career ladder (e.g., career teacher, mentor). As the weightings are collectively negotiated in bargaining states or agreed upon by teachers' representatives and the district in nonbargaining states, they should adhere to federal policies and state regulations as well as local priorities.

In some instances, the weighting of an individual measure can be informed by the teacher's experience and number of years in the classroom. For example, for evaluations of new teachers, classroom observations might play a relatively large role and student growth might play a smaller role, since limited data is available. For evaluations of experienced teachers, the teacher might be given the option to demonstrate effectiveness based on leadership roles, mentoring, and collaboration in ways that new teachers cannot. Finally, decisions about how performance-based ratings are determined vary for teachers in different subject areas, because value-added scores can only be calculated for teachers of tested subjects and grades. The following boxes provide examples of weighting multiple measures, weighting based on place on a career ladder, and weighting based on availability of value-added data. These examples are intended only to provide information and should not be interpreted as recommendations or endorsements by the Association.

The TAP System: Teacher Performance Evaluation Domains

TAP™: *The System for Teacher and Student Advancement* was created in 1999 and currently serves 7,500 teachers in school districts across 13 states and the District of Columbia. Each teacher receives a performance assessment score based on multiple measures: a skills, knowledge, and responsibilities rubric; classroom achievement gains; and student achievement gains. Each teacher is given an averaged performance rating for each evaluation based on indicators in each of four domains:

1. Designing and Planning of Instruction
2. The Learning Environment
3. Instruction
4. Responsibilities

In each domain, performance is rated on a five-point scale. The ratings are then averaged and assigned a single score. Further, each domain is assigned a weight based on whether a teacher is at the stage of career, mentor, or master teacher:

Weighting Based on Place on Career Ladder

Domain Weights	Career	Mentor	Master
Designing and Planning Instruction	15%	15%	15%
The Learning Environment	5%	5%	5%
Instruction	75%	60%	40%
Responsibilities	5%	20%	40%

Although TAP demonstrates differential weightings across the educator career ladder, TAP teachers in untested grades and subject areas are *not* evaluated based on multiple robust measures. Rather, their evaluations are based *solely* on the school-wide value-added scores of teachers in mathematics and language arts in their schools. (**Note: NEA does not recommend teacher evaluation models in which test scores are the predominant measure of teaching effectiveness. See Appendix B for limitations.**)

Weighting by Teacher Type: A Hypothetical Example

One option in creating a teacher evaluation system is to weight measures based on the type of classroom teacher being evaluated. Unlike TAP, which bases weights only on whether a teacher is a career, mentor, or master teacher, some evaluation systems use different weights for classroom teachers in different subject areas, as illustrated in the following hypothetical example. In an actual situation, teachers from each of these different circumstances should be involved, alongside their Association representatives, in selecting the measures and identifying the percentage weightings of the various measures of teaching effectiveness for their situation.

Weighting Based on Teacher Type

Domain Weights	General Education Teachers with Value-Added Data	General Education Teachers without Value-Added Data	Special Education Teachers
Individual value-added student learning data	10%	0%	0%
Classroom observations	30%	40%	30%
Commitment to school community	15%	15%	15%
School value-added student achievement data	10%	10%	10%
Teacher-assessed student achievement data	15%	15%	15%
Individual professional growth plan	10%	10%	10%
Develop and monitor two student learning outcomes /year	10%	10%	20%

An example such as the above could be adapted for evaluating preschool teachers, counselors, librarians, or other school staff. **(Note: NEA does not recommend teacher evaluation models in which test scores are the predominant measure of teaching effectiveness.)**

Understanding Value-Added Assessment

(*Adapted from Value-Added Measures Fact Sheet, NEA Collective Bargaining and Member Advocacy*)

In a value-added model (VAM), states or districts use student achievement data, and sometimes other student background data, as statistical controls in order to isolate the specific effects of a teacher, school, or program on student academic progress. VAMs aim to answer the question of whether, on average, a student's change in performance met a growth expectation [based on what can be gleaned from his or her past performance] (Goldschmidt, 2005).

The adequate yearly progress (AYP) goals and associated sanctions established under the *No Child Left Behind Act* of 2001 and recent Race to the Top grant criteria have spawned tremendous interest in statistical models that evaluate teacher effectiveness on the basis of student test scores. Proponents of VAMs argue that "objective" data about whether students have learned must be taken into consideration and that, despite its flaws, a VAM is the best model available to show whether or not the test scores of a teacher's students are improving over time. However, growth models are complex, have rigorous data requirements, and require more human resources and psychometric expertise than most states and districts can provide. While some VAMs might provide insight into student achievement, questions about the reliability and validity of these models and the estimates they produce make it premature to use teacher effectiveness estimates for high-stakes decisions related to teacher performance or pay.

Comments from research experts (excerpted from the NEA Fact Sheet):

There is a lack of consensus in the research community about whether VAMs can accurately isolate the effects of a single teacher, especially over an extended period.

- "That some value-added models will be reliable but not others, and that value-added modeling may be only reliable in some settings, are important limitations. They suggest that in contexts such as statewide teacher-accountability systems, large-scale value-added modeling may not be a viable solution." (Koedel & Betts, 2009)

- "Both our … analyses find significant fade-out of teacher effects from one year to the next, raising important concerns about whether unbiased estimates of short-term teacher impact are misleading in terms of long-term impacts of a teacher… It is not clear what should be made out of such 'fade out' effects. Obviously, it would be troubling if students are simply forgetting what they have learned, or if value-added measured something transitory (like teaching to the test) rather than true learning. This would imply that value-added overstates long-term teacher effectiveness." (Kane & Staiger, 2008)

Teacher input may be an important influence on achievement, but it is not the only influence.

- "[T]reating the output of a value-added analysis as an accurate indicator of a teacher's relative contribution to student learning is equivalent to making a causal interpretation of a statistical estimate…. In the absence of randomization, causal interpretations can be misleading. In reality, the classroom placement of students and teachers is far from random. In most districts, parents

often influence where their children go to school and even to which class and teacher they are assigned. Similarly, teachers may select the school and classroom where they are placed. Thus, the students assigned to a particular teacher may not be representative of the general student population with respect to their level and rate of growth in achievement, parental support, motivation, study habits, interpersonal dynamics and other relevant characteristics. It is very difficult for the statistical machinery to disentangle these intrinsic student differences from true differences in teacher effectiveness." (Braun, 2005)

Students are not necessarily assigned to classrooms and teachers on a random basis. This has major ramifications. Since value-added models are built on the assumption of random assignment, this compromises the results.

- "It is worth considering how violations of the static selection assumption might arise in practice.… School principals 'track' students and do not randomly assign teachers to tracks. Monk (1987) finds that most school principals randomly or evenly distribute students in elementary grades, apparently because principals want to even out the workload among teachers. But he also finds that some principals try to match students to teachers who have skills particularly well suited to student needs." (Harris, 2009)

The role of administration is often not considered in VAMs.

- "[I]t becomes quite difficult to account for the impact of school administration. If we were making within-school comparisons, we might reasonably assume that the impact of school administration affects all teachers relatively equally, so there is no need to account for it. But when estimating VAMs for accountability, this approach fails because very few teachers are observed in multiple schools, which would aid in isolating the effect of the teacher from the effect of administrators." (Harris, 2009)

VAMs may be expensive to adopt.

- "Data system requirements need to be addressed. To measure growth, there must be a capacity to track individual student scores from year to year (and sometimes from one district to another within a state). This capacity often requires a statewide student identification system." (Goldschmidt, 2005)

- "Training is required to build capacity among the teachers, administrators, media, legislators, and general public to understand the additional complexities that occur when using data from more than one point in time. Even changing to the simplest of growth models will require a significant retooling of training materials." (Goldschmidt, 2005)

VAM numbers don't say much about what teachers do to influence student achievement.

- "These analyses treat the classroom as a black box…and do not tell us why some classrooms are more effective than others, nor do they give us a very good picture of the potential improvements in student achievement that might be produced if we combined particularly effective instructional conditions into powerful instructional programs." (Rowan, 2002)

"The statistical methods behind value-added indicators of teachers' effectiveness—inscrutable to virtually all teachers—produce estimates, and even the most sophisticated estimates are subject to error, bias, and misinterpretation. This is [of particular concern] where measures of academic achievement are involved." (Miller, 2009)

Researchers have warned against using value-added estimates for high-stakes purposes.

- "VAM results should not serve as the sole or principal basis for making consequential decisions about teachers. There are many pitfalls to making causal attributions of teacher effectiveness on the basis of the kinds of data available from typical school districts. We still lack sufficient understanding of how seriously the different technical problems threaten the validity of such interpretations." (Braun, 2005)

- "[T]he research community has cautioned against the heavy reliance on test scores—even when sophisticated VAM methods are used—for high stakes decisions such as pay, evaluation, or tenure. For instance, the Board on Testing and Assessment of the National Research Council of the National Academy of Sciences stated, '…VAM estimates of teacher effectiveness should not be used to make operational decisions, because such estimates are far too unstable to be considered fair or reliable.'" (Baker, 2010)

Degrees of Controversy Over the Use of Value-Added Measures	
VAM to be used in a Formative Evaluation designed to inform and improve practice	Less controversial
VAM to be used as a "trigger" to examine a teacher's performance more closely, but not to be used as part of a final evaluation	Less controversial
VAM to be used as one of several measures of student learning, so that student learning then becomes one of several measures of teaching effectiveness—a measure within multiple measures within multiple measures	Acceptable as a measure within multiple measures
VAM to be used as a significant percentage of a summative evaluation	Controversial (e.g., defining significant)
VAM to be used as a sole measure	Highly controversial

Understanding Multiple Measures of Student Learning

A common source of confusion is the difference between multiple measures of effective teaching and multiple measures of student learning. **Multiple measures of teaching effectiveness** refers to different approaches that assess teachers' practices, skills, and dispositions, as well as contributions to student growth and professional learning communities within schools and districts. **Multiple measures of student learning**, by contrast, refer to the myriad of approaches used to assess student growth across subjects, grades, and educational settings. Student achievement is only one measure that attempts to capture student growth by using test scores. Student achievement cannot capture student growth and learning that occurs in non-testing settings. Consequently, multiple measures or multiple approaches are needed to collect a plethora of evidence that can be used to provide a more holistic account of student growth and learning. These measures or approaches may include standardized tests, curriculum-based tests, performances (e.g., musical performances), products created (e.g., artwork), projects, or portfolios of student work.

Evidence of good teaching and evidence of student learning are both critical components of a comprehensive teacher evaluation system. Although student learning is often measured by standardized tests, it is important to recognize that standardized tests cannot account for the work of teachers in untested grades, subjects, and student populations. Most standardized tests are limited in their ability to accurately assess student learning. One persistent challenge to designing and implementing comprehensive evaluation systems is the narrow selection of student growth measures for the majority of teachers for whom there is no student data from standardized tests (Prince, Schuermann, Guthrie, Withman, Milanowski, et al., 2010). Student learning, like teacher evaluation, must include multiple measures that illustrate evidence of growth in knowledge and skills.

Examples of student learning:

- Pre- and posttests
- Written work scored by a common rubric
- Group work or presentations scored by a common rubric
- Student learning objectives
- End-of-course papers or portfolios
- Students' oral and written presentations
- Project-based inquiry activities
- Teacher-generated information about student growth and goals
- Evaluations of effective engagement, critical thinking, or self-efficacy

Examples of specific assessments that constitute multiple measures of student learning:

- Benchmark assessments
- Dynamic Indicators of Basic Early Literacy Skills (DIBELS)
- Developmental Reading Assessment (DRA)
- Curriculum-based assessments (e.g., Everyday Mathematics)
- Unit tests from district- or state-approved textbooks
- Off-the-shelf standardized assessments that are aligned to the district or state's content standards
- Rigorous teacher-created assessments that are aligned to the district or state's content standards
- Rigorous portfolios of student work that are aligned to the district or state's content standards (Goe, 2010)

For measuring student growth in untested grades and subject areas, three emerging approaches include student learning objectives, other assessments, and measures of collective performance.

Student learning objectives are a participatory method of setting measurable goals or objectives, based on the students taught, the subject matter taught, the baseline performance of the students, and the measurable gain in student performance during the course of instruction. For example, a student learning objective might specify that 75 percent of students in a given class will improve their scores on a writing rubric by two levels. Teachers can develop student learning objectives based on the mathematics and English Language Arts content standards that would satisfy the testing requirements required by the *Elementary and Secondary Education Act* (ESEA) for grades 3–8 and one year of high school.

Other assessments that can be used to measure student growth include standardized assessments in additional subject areas. They may be based on teacher-developed or other classroom assessments, but they must be aligned with established learning outcomes and content standards, and they must be considered rigorous and comparable across classrooms. Additional assessments can include the development or adaptation of other measures of student growth for subjects or grades not required by ESEA. These measures may take a variety of forms, including: early reading measures; end-of-course assessments; common **formative assessments**; benchmark, interim, or unit assessments; and measures of English language proficiency. In any of these forms, however, the measures must be considered rigorous and comparable across classrooms. The assessments may be developed at either the SEA or LEA level with input from teachers. Teacher-developed assessments of student learning or growth also may fall into this category.

Finally, **measures of collective performance** refer to measures required by ESEA and other assessments used to measure the performance of groups of teachers. Measures of collective performance may assess performance of teachers throughout a school, grade level, instructional department, or other group of teachers. These measures can take a variety of forms, such as assigning school-wide, department-wide, or even district-wide student growth measures or assessing attainment of learning goals for individual teachers (Reform Support Network, 2010). However, these collective performance measures should *not* be used for high-stakes decisions such as pay or employment continuation. Rather, they should be used as formative tools designed to inform and strengthen the collective practice of effective teaching.

Tables 1 and 2 list various instruments for measuring effective teaching and describe the strengths and limitations of each. Note that each instrument has limitations, which is another reason to consider multiple measures. Table 1 presents tools for measuring teachers' practice, and Table 2 presents tools for measuring teachers' effect on student learning. Questions to consider include the following:

- What tradeoffs must a district make between comprehensiveness and feasibility?
- What will be the consequences of these tradeoffs?
- How will the instruments be validated? In other words, how can they be tested to make sure they measure what they intend to measure?
- What instruments are in place already, and how challenging would a change in evaluation instruments be for evaluators and for teachers?
- How will teachers be involved in the selecting and piloting of instruments?
- What resources are needed to monitor and sustain the effectiveness of the instrument?

Table 1. Tools to Assess Teachers' Attainment of Standards Associated With Effective Teaching (Measures of Practice)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Classroom Observation Instruments	Classroom observations are the most common form of teacher evaluation and vary widely in how they are conducted and what they assess. High-quality classroom observation instruments are standards based and contain well-specified rubrics that delineate consistent assessment criteria for each standard of practice. Trained evaluators use a classroom observation instrument to make consistent judgments of teachers' practice in the classroom. The observation instrument should incorporate evidence from pre- and post-observation conferences with the teacher. Observation instruments also may be used to assess teacher practice via videos of their lessons.	• Professionally credible as they often are developed by teachers and experts in teacher education • Adaptable for different types of teachers • Can measure many aspects of practice that are associated with effective teaching • Usually based on vetted standards of professional practice • Some observation instruments have been demonstrated to be related to student achievement • Can provide useful formative and summative information	• Reliability across observers requires extensive training, which can be expensive and time-consuming • Many observation instruments not validated in terms of being demonstrated to be related to student learning • Assesses teacher practice but not effects on student learning • Teachers are often left out of any training related to the evaluation instrument, which includes a shared vision of effective teaching and shared language describing the vision. • Some live video-streaming of classrooms do not incorporate pre- and post-observation conferences. In these situations, observers do not have the benefit of assessing the teaching in context.
Examples: Charlotte Danielson's Framework for Teaching, Cincinnati's Teacher Evaluation System, Teacher Advancement Program protocol, Classroom Assessment Scoring System (CLASS), Protocol for Language Arts Teaching Observations (PLATO)			

Table 1. Tools to Assess Teachers' Attainment of Standards Associated With Effective Teaching (Measures of Practice)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Teacher Portfolios	Teacher portfolios are exhibits of evidence of teaching practice, school activities, and **student progress**. They usually are compiled by the teacher him- or herself. Portfolios may include teacher-created lesson or unit plans, descriptions of the classroom context, assignments, student work samples, videos of classroom instruction, notes from parents, and teachers' analyses of their students learning in relation to their instruction.	• Professionally credible • Comprehensive; can measure aspects of teaching not readily observable • Useful for teacher self-reflection • Can be done electronically • May include some assessment of student learning • Can provide important formative and summative information on teacher practice	• Time-consuming for teachers and to a lesser extent scorers • May not represent day-to-day practice well • Portfolio scores rarely have been shown to be consistently related to student achievement • May not include robust measures of student learning
Examples: National Board for Professional Teaching Standards portfolio assessment, Teacher Performance Assessment (TPA), Performance Assessment of California Teachers (PACT)			
Evaluation of Teacher-Developed Student Artifacts	These instruments may rate lesson plans, teacher assignments, teacher-created assessments, and scoring rubrics on particular criteria, such as rigor, authenticity, intellectual demand, and alignment to standards, clarity, and comprehensiveness.	• Professionally credible • Adaptable for different types of teachers • Noninvasive, does not need to be done in real time • Captures many aspects of teacher practice • Can provide important formative and summative information on teacher practice	• Few validated systems exist • Comparability across different types of teachers has not been established • Reliability across observers requires extensive training • Does not include measures of student learning
Examples: Instructional Quality Assessments (IQA), Teacher Work Sample Methodology			

Table 1. Tools to Assess Teachers' Attainment of Standards Associated With Effective Teaching (Measures of Practice)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Student Surveys	These questionnaires generally ask students to rate teachers on an extent scale (e.g., from 1 to 5, where 1 = very effective, and 5 = not at all effective) regarding various aspects of teachers' practice (e.g., course content, usefulness of feedback, starting classes on time) as well as how much students say they learned or the extent to which they were engaged. They very often are not used for teacher evaluation at the precollegiate level.	• Provides perspective of schools' primary clients • Can provide information on how to improve relationships with students • Have been validated in certain contexts • Can provide important formative information on teacher practice	• Students are not able to observe much of what goes into a teachers' practice and therefore might not capture important information • Have not been validated for summative decisions in PK–12 classrooms • How do very young children participate? Reading levels?

Examples: Tripod Surveys developed by Ron Ferguson at Harvard University, Gallup Student Engagement Surveys, Georgia plans to use teacher-focused student surveys as part of their teacher evaluation system starting at Grade 4 (according to the GA Race to the Top [RTTT] grant application)

Table 1. Tools to Assess Teachers' Attainment of Standards Associated With Effective Teaching (Measures of Practice)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Parent Surveys	These questionnaires generally ask parents to rate teachers on an extent scale (e.g., from 1 to 5, where 1 = very effective, and 5 = not at all effective) regarding various aspects of teachers' practice (e.g., course content, usefulness of feedback, quality of homework) as well as the extent to which they are satisfied with the teachers' instruction. They very often are not used for teacher evaluation.	• Can provide perspective of schools' primary clients • Can provide information on how to improve relationships with students • May provide information that can be used for to inform teaching practice.	• Parents not able to observe much of what goes into a teachers' practice and therefore might not capture important information about a teachers' classroom practice • Have not been validated for summative decisions • Do not include reliable information on student learning growth • Participation is often low, therefore resulting information that may not be valid or reliable • Language barriers, literacy rates, or other factors may prevent parents from participating in surveys.
Examples: Tripod Surveys, Utah's Pay for Performance pilot program uses parent satisfaction surveys, Georgia plans to use parent surveys for teachers of students in Grades K–3 (according to the GA Race to the Top grant application)			

Table 1. Tools to Assess Teachers' Attainment of Standards Associated With Effective Teaching (Measures of Practice)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Teacher Self-Reports	Teacher self-report measures may take the form of surveys, instructional logs, or interviews. They ask teachers to report on what they are doing in the classroom, the extent to which they are meeting standards, and in some cases analyze the impact of their practice. They may consist of checklists, rating scales, rubrics, and may require teachers to indicate the frequency of particular practices.	• Can measure unobservable aspects of teacher quality • Can be easily administered • Can promote teacher self-reflection and analysis • Can promote a sense of self-efficacy • Can provide useful formative information	• Reliability and validity for summative decisions not well-established • Instructional logs have been used only for research purposes, so their validity for evaluation is questionable • Do not include independent measures of impact on student learning

Examples: Study of Instructional Improvement instructional logs, self-assessments, National Board for Professional Teaching Standards

Table 1. Tools to Assess Teachers' Attainment of Standards Associated With Effective Teaching (Measures of Practice)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Assessments of Professionalism/Commitment to School Community/ Advocacy for Students	Some evaluation systems include measures of the extent to which teachers demonstrate a commitment to professionalism These may or may not be captured in a classroom observation instrument, but because some of these activities are not obvious in the classroom—for example, the extent to which teachers are leaders of their professional learning communities or the extent to which they seek our opportunities to engage in professional development, or to the extent which they advocate for students particularly those who are often underserved —they may require a separate tool or rubric that evaluators use based on varied sources of evidence and conversations with teachers.	• Can be adapted for different types of teachers • Can incentivize collaboration and commitment • Can provide useful formative information	• Can be subjective; reliability across observers difficult to establish • Little research available that establishes a direct relationship between teachers' professionalism and commitment to student growth • Because advocacy for students is influenced by context and constituents, it may be subjective and difficult to establish reliability across observers
Examples: Charlotte Danielson Framework			

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Student Growth Objectives or Goal-Driven Professional Development Plans	These evaluation tools offer teachers the opportunity to set their own high but feasible objectives for their students' growth in collaboration with their principal and/or other colleagues. The assessments teachers use may be common or standardized exams or teacher-developed assessments. Some tools require teachers to specify the professional development they will participate in to ensure their students achieve their growth objectives.	• Can be used to gauge teachers' contributions to outcomes in untested subjects (e.g., social studies, biology, music) • Can focus teachers' practice on achieving particular outcomes based on their analyses of student learning needs • Can promote collaboration among teachers and instructional leaders • Can incentivize teachers to engage in professional learning opportunities that will help them achieve goals	• Safeguards must be in place to ensure that goals set are feasible yet also high and rigorous • Comparability among teachers may be problematic
Example: Student Learning Objectives–Austin Independent School District; Student Growth Objectives–Denver Public Schools; New Mexico Professional Development Plans			

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Measures of Student Engagement or Educational Attainment	These can include classroom observations, and/or or self-reports of student engagement, as well as measuring Advanced Placement course participation rates, graduation rates, dropout rates, and student absenteeism.	• Can assess and incentivize other important teaching outcomes • Can provide some formative and summative information on the effects of teachers' practice	• Difficult to attribute individual teacher contribution to such outcomes • Have not been widely tested in the field to determine validity, reliability, etc.

Example: These currently are not used for individual teacher evaluation but are being considered for example in Georgia, where according to the Georgia Race to the Top application, "plans to invest in the development, testing, and evaluation of alternative quantitative measures to assess student engagement and student achievement" (p. 106). The Maryland RTT application suggests that local school systems can "propose alternative priorities for annually measuring student growth and learning, such as—at the high-school level—gains in Advanced Placement participation and exam performance or decreases in the dropout rate" (p. 146). The Measures of Effective Teaching Study funded by the Gates Foundation is currently examining a measure of student engagement as a valid predictor of student growth. The Tripod Survey was used in this study to measure student engagement.

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Measures of Student Work That Show Evidence of Growth	These may take the form of alternative assessments of student learning, for example, writing samples, portfolios of student work, student oral presentations, capstone projects, and the like.	• Can be more authentic assessments of student learning than standardized tests and therefore more valid assessments of teachers' contributions to student learning • Can provide important formative and summative information about a teacher's practice	• Validity and reliability not well-established • Very difficult to standardize and therefore difficult to establish comparability and reliability • Psychometric properties of such assessments are not well-understood
Example: According to the Massachusetts' Race to the Top application, MA will work to develop student performance assessments, and teachers and teacher teams will require training on how to gather student work that demonstrates individual student learning, as well as typical student learning in a class. The state also will develop district-based assessments that are comparable across subjects and grades. Finally, the state will include student work samples as evidence of student learning.			

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Gain Score Models or Pre-Test/Posttest Methods**	Gain Score Models measure the difference between an earlier and a later test score, so unlike Value-Added Models (VAMs), they only require two test scores. This approach also relies on vertically equated assessments.	• Are preferable to "status" models because they indicate change in student learning over time • Can provide information on some teachers' contributions to student outcomes • Allows for comparisons between teachers • Can provide some formative and summative information on the effects of teachers' practice	• Cannot adequately control for students' background characteristics • Can be prone to error if tests are too easy or too difficult • Can lead to false attributions of value—that is, in team- or co-teaching situations, cannot separate out contributions of one teacher or another • Growth models can encourage noneducative test prep, cheating, etc. • Validity and reliability can only be achieved if the administration of the tests and interpretation of the scores are consistent across all classrooms • Growth models can disincentivize collaboration depending on how they are used

Example: Hillsborough County, FL; Eagle County, CO; and MA, MD, and NY among other Race to the Top winners will develop pre- and posttest measures of student learning for teachers in subjects not tested by the state assessment system

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Value-Added Models**	Value-Added Models (VAMs) measure the gains that students make and adjust those gains for student, teacher, or school characteristics. The gains are interpreted as the "value" that a teacher adds because the gains are presumed to be net of all other influences. This presumption is much debated. Stronger VAMS use well-designed vertically equated standardized achievement tests that measure relevant concepts and test students at least three times. Some tests are vertically equated, which means that a given score on the fourth-grade version of a test represents the same level of performance as that same score on the fifth-grade version of the test. Not all states have vertically equated assessments.	• As all growth models, VAMs are preferable to "status" or attainment models that measure student proficiency at one point in time because they indicate change in student learning over time • Can provide information on some teachers' contributions to student outcomes (only those in tested subjects and grade levels) • Can allow for comparisons between teachers • More likely to measure impact of teachers versus other student and school background factors than other kinds of growth models	• Estimates of teaching effectiveness can be unstable from year to year, which is why most VAMs require three or more years of data • Difficult to verify the accuracy of the measures due to lack of transparency. • Value added measures are not available for the majority of teachers • Results do not provide sufficient information on how teachers can improve their effectiveness • Can lead to false attributions of value—that is, in team- or co-teaching situations, cannot separate out contributions of one teacher or another • Requires vertically scaled exams • Growth models can encourage noneducative test prep, cheating, etc. • Growth models can disincentivize collaboration depending on how they are used
	Example: Tennessee Value-Added Assessment System (TVAAS); Ohio's Education Value-Added Assessment System; Dallas Value-Added Accountability System (DVAAS)		

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Normative Growth Models**	Normative Growth Models compare growth in student achievement to the amount of growth made by a representative population of students on the same test. A vertical scale is not necessary.	• Are preferable to "status" models because they indicate change in student learning over time • Can provide information on some teachers' contributions to student outcomes • Can allow for comparisons between teachers • Unlike other growth models, these do not rely on vertically-scaled exams	• Cannot be computed for teachers in untested grades and subjects • Can lead to false attributions of value—that is, in team- or co-teaching situations, cannot separate out contributions of one teacher or another • Growth models can encourage noneducative test prep, cheating, etc. • Cannot adequately control for student background characteristics • Growth models can disincentivize collaboration depending on how they are used
Example: Massachusetts Student Growth Percentile (SGP) (not currently used to measure teaching effectiveness)			

Table 2. Tools to Assess Teachers' Impact on Student Outcomes (Measures of Effects)

Evaluation Measurement Tool-Practice	Description	Strengths	Limitations
Categorical Growth Models**	Categorical Growth Models calculate student growth based on changes in performance category placement (e.g., from "beginning" to "proficient") from year to year. Changes in all possible category placements are judged subjectively and each is assigned a value that indexes its importance.	• Are preferable to "status" models because they indicate change in student learning over time • Can provide information on some teachers' contributions to student outcomes • Can allow for comparisons between teachers • Are more easily understandable by various stakeholders than VAMs	• Cannot adequately control for student's background characteristics • Growth models may encourage noneducative test prep, cheating, etc. • Growth models can disincentivize collaboration depending on how they are used
Example: Florida Value Tables, Minnesota Growth Model			

* This table is adapted substantially from Little, Goe, & Bell, 2009.

** Note that one option for using measures of student growth in teacher evaluation is to calculate *school-wide* student growth using any of the statistical models described in this table and use that as one of multiple measures of teachers' contributions to school and student outcomes. Depending on how *school-wide* student growth data is used (e.g. inform teaching practice), it could have the potential benefit of motivating all teachers in the school to work with their students on particular knowledge and skills (for example, all teachers can promote student writing in their classes). However if used in high-stake situations (e.g., continuation of employment), it may put undue pressure on teachers whose primary responsibility is the tested subject area.

It is important to note that the majority of the measures of teacher practice and teacher effects presented in Tables 1 and 2 have not been tested rigorously in the field for their validity, reliability, or comparability, nor have they been tested for the purposes of making summative decisions that involve hiring, promotion, compensation, due process, tenure, or dismissal.

The assessment of effective teaching for both summative and formative purposes is a new field, and careful experimentation is required. It is critical that the Association help lead the efforts to support effective teaching and improve student learning. Implementing robust teacher evaluation systems has the potential to provide the information needed to improve instructional practice. Teachers themselves are in a better position than anyone else to accurately determine what types of assessment of their practice will be most meaningful for continued professional growth. Improved assessment of teaching effectiveness will help school leaders and policymakers determine how to change the system and shape policies to enhance teaching and learning in our nation's schools.

For multiple measures of student learning to be used effectively, evaluation systems must appropriately account for them. All measures must be regularly tested for rigor and comparability across grades and schools. For example, what counts as an excellent paper in one classroom must be considered equally in a classroom down the hall or across the city. Students and teachers alike should understand the criteria and be tested or graded with comparable measures, at similar times during the year, with the same time limits, with the same preparation, and under otherwise similar circumstances. Developing rigorous and comparable measures of student learning is no easy feat. The questions below provide guidance as to how districts and states can develop multiple measures of student learning to assess teacher performance throughout a district.

Questions to consider when developing or selecting measures of student learning:

- Does the approach allow for the assessment of student learning over time?
- How can we establish the following?
  * The proposed measure assesses the expected knowledge and skills appropriately, in terms of the content of questions or tasks included and coverage of the subject area.
  * No students are disadvantaged by the specific questions or tasks included.
  * The measure appropriately distinguishes among students.
  * Scores based on the measure accurately reflect meaningful changes in student learning in the subject area, either in strictly comparative terms (e.g., some students learned more than other students) or in terms of growth toward a standard (e.g., some students made more progress than others toward a goal that will help them be successful).
- How can we ensure that student learning is being measured consistently across classrooms?
- What steps in development and administration are needed to ensure that scores will have the same meaning within a given subject area and that student growth will have a similar interpretation across subjects?
- Is the approach transparent and understandable to teachers and other stakeholders?

- What capacity is needed to develop and implement the measures now and over time?
- How easily can data from the model be used along with other data to assess teaching effectiveness?
- What provisions are in place to ensure ongoing review, calibration, and adjustments when needed?

Developing a Teacher Evaluation System That Includes Multiple Measures of Student Learning: The Case of Hillsborough County, Florida

Hillsborough County Public Schools and the local union have recently developed promising new policies on multiple measures for teacher evaluation. Specifically, the district is moving from a system in which evaluation was based solely on principals' ratings to one in which both teacher performance and student learning are used to assess teaching effectiveness.

Hillsborough is using pre- and posttests in each grade and subject. This includes Florida Comprehensive Assessment Test (FCAT) for some grades and subjects, and various national norm-referenced tests for other grades and subjects. Content supervisors in the district have developed additional tests with input from teachers, and those tests have been assessed for validity.

When the pre- and posttests were initially designed for every class, there was some concern that certain specialty tests were not very rigorous and that tests in different subject areas could not be compared. But over the years, the tests improved and concerns subsided, largely because the district included input from teachers in the full range of subject areas. In the case of noncore academic subjects, a combination of district-developed pre- and posttests and FCAT reading and mathematics scores are used. **This student growth measure accounts for 60 percent of the performance rating for a teacher, and the evaluation accounts for 40 percent**. In each evaluation, the principal assesses instructional effectiveness, planning and preparation, professional behaviors, techniques of instruction, and classroom management.

Creating this system has taken several years and significant financial investment by the district. Districts or states that lack the resources to implement similar reforms can still look to Hillsborough County for ideas. **(Note: NEA does not recommend teacher evaluation models in which standardized achievement test scores are the primary measure of teaching effectiveness.)**

Selecting and Adapting Teacher Evaluation Tools

After developing a teacher evaluation system, the next step is to develop teacher evaluation tools. This is where you specify how your definition of teaching effectiveness will be measured in actual classrooms, using tools such as rubrics, surveys, observation instruments, portfolios, artifacts, student learning measures, or a combination of approaches. Evaluation tools will determine what can be learned about a teacher based on his or her evaluation scores.

It is important to select or develop an evaluation tool or set of tools that:

- Are standards-based, so that everyone understands what is expected of teachers in a district or in districts across the state

- Provide multiple measures or sources of evidence of teaching practice and effectiveness

- Have been pilot-tested in the field and shown potential to be valid and reliable; in other words, they measure what they intend to measure time and time again

- Have professional credibility, reflected by the fact that they were developed with teacher input and measure important aspects of teaching practice and effectiveness

- Provide feedback that teachers can use to improve their practice

- Differentiate among several levels of teaching practice and teaching effectiveness

In developing an evaluation system, consider the following questions:

What standards for professional practice will provide a basis for the evaluation tools?

Defining effective teaching and selecting the standards for professional teaching practice are critical to the process of developing an evaluation system, because they provide the foundation for the evaluation instrument that will be used to assess the teaching practice. The standards also establish a vision for effective teaching, so that everyone can work toward a shared objective. Without standards, there is no common language and understanding. Teachers deserve clear standards and clear guidance about how they will be assessed by reviewers.

You might begin by asking which, if any, of the following standards currently inform the teacher evaluation system and whether another set of standards would be more appropriate:

- State professional teaching standards
- Locally developed teaching standards
- InTASC Model Core Teaching Standards www.ccsso.org/Resources/Programs/Interstate_Teacher_Assessment_Consortium_(InTASC).html
- National Board for Professional Teaching Standards www.nbpts.org/the_standards
- Other _____

Who will be evaluated using the proposed system? Will different measures be used for different staff?

The same evaluation tool will not work for every type of educator. Effective teaching practices vary by grade level and by subject. Effective teaching in a music class requires different professional practice and outcome measures than effective teaching in an algebra class. The practices and outcomes of an effective instructional leader will be different from that of an effective guidance counselor. The list below offers examples of the different roles that staff play, which you need to take into account in determining how many variations of your evaluation tool will be needed:

- Teachers in core content areas
- Teachers in noncore content areas, such as physical education and the arts
- Teachers of English Language Learners
- Special education teachers
- Student support staff, such as school psychologists, occupational therapists, and guidance counselors
- Mentors, coaches, and instructional leaders such as reading specialists
- Nonteaching staff, such as administrative assistants and custodians
- Principals and other administrators
- Other _____

Given time and resource constraints, are the tools practical? What tradeoffs between practicality and comprehensiveness must be made?

As the Tables 1 and 2 depict, some evaluation tools require more time, training, materials, and expertise to implement effectively than do others. Sustaining a robust evaluation system that meaningfully differentiates performance and supports teacher development requires significant investments of time, training, materials, and expertise. The goal is to design a system that provides sufficient evidence for appropriate decision making but does not drown the evaluator in evidence or data. An example of a well-designed evaluation system that collects a significant amount of data is the National Board for Professional Teaching Standards system. While using this system for all teachers in a district would require significant training, some practitioners believe it may be a cost-effective solution.

How will the ratings from each of these instruments be weighted to tally a final evaluation rating?

If multiple instruments are employed to come up with one evaluation rating, scores from some instruments may have greater weight than scores from others, depending on how highly valued, or respected, the results are. For example, an overall teacher evaluation rating may consist of a value-added score, principal observation score, peer observation score, and school-wide value-added score. Some states require that a certain percentage be based on student learning growth. Some districts

weight the measures in different ways depending on the type of teacher and the available measures for each category or grouping. Decisions about weighting different instruments should be done at the local level through collective bargaining or, where there is no bargaining, agreed to by the organization representing teachers.

How many levels of proficiency will the evaluation system be able to detect?

Principal evaluation ratings traditionally have classified teachers as either satisfactory or unsatisfactory. It is important to consider the purpose of the evaluation system, because having two levels **might** be sufficient to provide important information about whether a teacher is reaching a minimum standard of performance for high-stakes decisions such as job retention. However, having only two levels in an evaluation system designed to improve instructional practice does not adequately capture variation in teacher performance. Such a limited model also fails to recognize real differences in performance or provide adequate information to inform teachers' efforts to improve. Some states now require by law or state code that teacher evaluation systems include four or more rating levels. Including several levels allows state and district officials to more accurately evaluate how programs and strategies impact teaching effectiveness and detect incremental improvements or shortcomings. Teachers tend to prefer having multiple rating levels as a professional acknowledgment that what they do is complex and important. However, the more levels within an evaluation system, the smaller the differences between levels, which means evaluators must be highly skilled in their ability to understand and accurately distinguish the differences (Perlman, 2002). There is currently no agreement in the education community on the optimal number of levels for an evaluation system, but many well-regarded systems use four levels. For example, Cincinnati's Teacher Evaluation System, which is based on Charlotte Danielson's framework, uses four levels.

How do the evaluation tools differentiate among different levels of experience?

Research has demonstrated that teachers become increasingly effective during their first few years of practice. This is the case for most other professions as well. So, when it comes to evaluating performance, being able to differentiate between novice and experienced teachers is important. Measurement instruments must be able to capture different levels of effectiveness and detect growth over time. For example, some rubrics distinguish between emerging and more accomplished practice in ways that can make it acceptable for first-year teachers to perform at the emerging level but not teachers in their second or third year of practice—unless the experienced teacher is teaching a new subject area or grade level.

It might not be possible to measure the performance of first- or second-year teachers using instruments based on student assessments, because there is no track record of student growth measurements upon which to make valid judgments. Evaluation systems might need to use different instruments for different levels of experience or increase the frequency or length of observations for new teachers. Ohio is one state that is working to develop evaluation system guidelines that promote differentiation in teacher evaluation tools based on teachers' levels of experience.

Will classroom observation tools employ checklists, rubrics, or narratives?

All three approaches can provide teachers with useful information about their performance. Narratives and rubrics can provide more information than checklists can, but they are often more time-consuming. The use of narratives raises issues of comparability among different classrooms. Comparability is important, because it helps to ensure fairness and sets a high bar for all teachers in the system.

Using rubrics or narratives in observations is very time-consuming. To make those approaches more feasible, some districts and states have changed their job description for principals, so that principals can focus more on instructional leadership and spend less time on budget, operations, and school discipline. In some cases, other personnel may also receive training that qualifies them to conduct teacher evaluations.

It requires significant thought to determine whether to use checklists, rubrics, or narrative forms and to develop the precise design of the evaluation instrument. As solutions are proposed, refined, and negotiated by teacher representatives, the voices of educators themselves need to be heard.

How often and for how long will teachers be observed?

Formal observations by principals or other administrators factor into teachers' summative evaluation ratings. The frequency and duration of those observations are often established in teachers' collective bargaining agreements or codified in state law.

Teacher Observations: The following key questions pertain to implementation and policy. They require thoughtful consideration:

- Selecting a Rubric
  * Will the state select a predesigned rubric or develop its own?
  * Will the state mandate a rubric to districts or allow local flexibility?
  * Will the same rubric be used for all teachers?
- Selecting and Training Evaluators
  * How will evaluators be selected?
  * What training will evaluators receive?
  * Will evaluators be required to demonstrate competency before administering evaluations?
- Conducting Observation
  * How many observations will be required (for novice and accomplished teachers)?
  * When will the observations take place?
  * Will the length of the observation be mandated?
- Collecting Information
  * What information will be collected to support observation findings?
  * How will the results be shared with the teachers?
  * How will the district or state ensure that the results are valid and reliable?
  * Will there be a grievance process, and if so, how will it operate?
- Refining the Process
  * How will the evaluator training be monitored?
  * How will inter-rater reliability be monitored and by whom?
  * How will this information be used to further refine the evaluation system?

To meet the needs of members and the district, it is important to determine whether to change the regulations or work within them. For example, some systems require that new teachers be observed twice per year and tenured teachers once every three years, for at least 30 minutes per observation. Such infrequent observation is not likely to be enough to meaningfully support teacher development, comprehensively assess the quality of a teacher's instruction, or ensure that effective teaching practices are being implemented in every classroom.

Increasing the frequency or duration of observations may enable evaluators to collect evidence that is substantial enough for making high-stakes decisions, such as nonrenewal or continued employment. To promote teacher development, formal observations can also be supplemented with informal observations that have a formative evaluation purpose. To ensure fairness and equity, contracts should include language pertaining to observations for formative evaluation.

To what extent does the evaluation system assess teachers on aspects of their practice that they can control?

Teachers may have control over some aspects of their working conditions and effectiveness. To a large extent, they can control their reactions to those working conditions. However, there is much that teachers cannot control. For example, teachers cannot control the following:

- The planning of instructional time allocated by content areas, and the reallocation of instructional time to prepare students for high-stakes tests
- Assignment to managerial and organizational tasks, such as lunchroom duties, hall duties, paperwork activities, and reports
- Physical arrangements of a classroom that could preclude various types of instructional grouping arrangements
- The mix of students assigned to them, and those students' needs, interests, and levels of readiness
- The degree to which the required curriculum is aligned with the assessments they are required to administer—even when their performance evaluations may be based on those assessments
- Their workload in terms of class size, student load, number of lessons they must prepare and extracurricular responsibilities
- The quality of the professional learning opportunities available to them
- The quality of their school leadership

The degree to which teacher evaluation tools are sensitive to these parameters is important. For example, most educators are wary of having their performance judged solely on students' test scores. To some degree, this concern is justified, since much of what test scores are known to reflect is beyond teachers' direct control. Some experts argue that student socioeconomic backgrounds, access to healthcare, neighborhood crime rates, housing stability, and other out-of-school factors significantly influence student achievement. They believe that even with high-quality instruction, children with cumulative disadvantages will perform less well than other students (see A B*roader*, B*older*, A*pproach to* E*ducation*, at www.boldapproach.org/statement.html). Certain statistical models attempt to account for these factors, but no one knows whether they do so adequately. Some models may better address these factors than others.

In-school factors that can affect the quality of learning include teaching arrangements (co-teaching and team teaching), school leadership, availability of resources, and the school's professional climate. Teaching in a community of professionals who share a vision, for example, can enhance student learning. Measuring these kinds of in-school factors must be given serious consideration if we are to provide a fair and accurate picture of the context in which the teaching and learning occurs. Beyond holding teachers accountable, we need to ensure that the school as a whole supports ongoing professional development, focuses on teaching and learning for all students, promotes collaboration among teachers, and supports effective teaching for all students.

How will the measures assist in the development of specific performance goals and targeted professional development?

The primary goal of teacher evaluation is to help teachers become more effective. That means the evaluation tools must provide administrators, instructional coaches, peer reviewers, and teachers themselves with useful information that can guide improvement efforts. For example, student growth measures may indicate that a teacher's instruction is weak in an area—say, reading comprehension. This is important to know, but that fact provides no guidance for how that teacher can improve in a given area. A well-designed observation instrument may provide such guidance. Teachers need both pieces of information to develop specific and measurable performance goals that are likely to result in improved practice.

Does the district have the ability to link individual teachers' data with their students' data, including unique identifiers for both students and teachers?

Matching teachers' data with their students' data allows districts and states to use student outcome measures in teacher evaluations through, for example, value-added or other growth model measures. Some districts and a small majority of states can link student test scores with the students' teachers, but teachers themselves should have the ability to verify that they are teaching the students whose scores they are linked to. This is not a simple process, especially in places where student mobility is high. However, it is important to be able to verify student-teacher data links. Knowing whether a teacher's data are accurately linked to students' data will determine the types of student learning outcomes that can be used in teacher evaluation.

Even so, it is important that high-stakes decisions, such as due process (tenure), dismissal, or pay, not be based solely on student test scores and **never** be based on a single test score. On the other hand, teachers' value-added scores can be used to guide efforts to enhance teaching effectiveness if the scores are supplemented with other measures of teaching effectiveness. Value-added measures are not diagnostic. Because they are only a single measure, they cannot provide information for making teaching more effective.

Does the district have a plan to ensure data accuracy?

Having accurate data is critical. For example, if student test scores are improperly entered or ascribed to the wrong teachers, teaching effectiveness ratings will be incorrect. Moreover, teachers' overall evaluation ratings need to be accurately entered and properly stored to ensure that misinformation

does not lead to poor decisions. One way to ensure accuracy is through transparency. The need for accuracy and transparency must be addressed in the planning stages. It must be determined how information access will be handled. For example, a system might or might not allow teachers themselves to check to make sure that they get credit only for the students they teach.

How will the district, the Association, or both evaluate the validity and reliability of the measures being used?

To make responsible high- or medium-stakes decisions on the basis of teacher evaluation tools, it is especially critical to be certain that they measure what they intend to measure (validity), time and time again (reliability). It is important to be able to demonstrate that the chosen measures are fair, valid, reliable, and transparent in your setting. An overall review of the evaluation system should answer the following questions:

- Does the evaluation system achieve the purposes for which it was designed?
- How well does the evaluation system support effective teaching for all students?
- Based on the agreed-upon definition of effective teaching, are the measures valid?
- Do the measures meet high standards of reliability in every school and for every teacher?

Working with the District: Piloting the Model

The next step in developing an evaluation system is to pilot the new teacher evaluation model in one or several schools in the district or in several districts within a state. A pilot test should last for at least one full year, so that each phase of the evaluation process—from who does the evaluating to how the results are used—can be implemented and assessed. This is your opportunity to make sure that the system produces the intended outcomes within the allocated resources, and that the system is sustainable. If problems are discovered, the pilot gives you the chance to refine the system or, if necessary, make changes before it is implemented across a school, district, or state. **It is imperative for all parties to agree that while the model is being refined, any resulting data will not be used for high-stakes decisions**.

As an Association leader, you should be involved in all phases of the pilot. You might notice problems with the initial design of the evaluation system that others might overlook. You might become aware, for example, that distinguishing between formative and summative assessments is critical to teachers, even though this fact might not occur to others involved in the reform process. As a critical partner in field-testing the model, the union can address teachers' questions about the program and how it will work when fully implemented. Just as importantly, the union should be advocating on behalf of members and giving voice to their concerns.

State and local leaders and staff should ask critical questions at the outset and throughout the pilot. The pilot will need to address the following questions:

- What criteria must schools meet to qualify as a pilot site for the evaluation model?

- How long will the pilot last?

- Are there any conflicts between the proposed evaluation tools and the collective bargaining agreement? If so, what challenges are likely to arise in the negotiation process when labor and management attempt to resolve the conflicts?

- Who will train the teachers, administrators, and evaluators to utilize the new evaluation tools? How will the union make sure that the training addresses reliability issues and equips evaluators to assess specific content areas and specialists effectively? Will evaluators themselves be subject to frequent review to ensure reliability?

- Who will serve as evaluators? Will the criteria and selection process be transparent? And how will their work be supported and funded?

- When will the training occur? Will teachers receive compensation to complete training after the school day, or will substitute teacher coverage be provided to allow teachers to complete their training during the school day?

- How will the evaluation data collected during the pilot be used? Will the pilot results be free of consequences, or will the data collected be used to make decisions about employment and promotions?

- Which measures will be used for formative purposes, and which will be used for summative purposes? Who will make the determination, and how?

- Will feedback be collected systematically at each stage of the pilot so that changes to the policy can be made accordingly? How will that happen? Will changes be made to the system part-way through the pilot stage, or will all changes be made after the pilot, but before the reforms are launched district-wide?

Keep professional growth at the forefront of discussions about improving and implementing the system, so that it does not get overlooked in the face of immediate implementation issues. If teachers participating in the pilot see their performance improving as a result of the new system, they will be more supportive of the program and, as a result, implementation will go more smoothly.

Feedback from teachers, principals, and other staff members should be collected throughout the pilot to ensure that the new system targets and promotes professional growth. The following are some questions to ask of staff during the pilot:

- Do teachers believe that the new system leads to more targeted professional development that can improve their performance?
- Are individual teachers allowed to set goals in the new system?
- Do teachers feel they have adequate information about the new evaluation system? Do they know where to go with questions?
- Do teachers believe the new system is fair?
- Is there anything about the new tools that is confusing to evaluators? Do evaluators feel adequately trained to use the new tools?
- How will communication be handled between the Association and members during the development and piloting of the system? Such communication will be essential for maintaining transparency in implementing the system district-wide.

Once the pilot is complete, all stakeholders must analyze the feedback carefully, including state and local leaders. This could be an onerous task, given the variety in stakeholders, the many measures of teachers' contributions to student learning, and the complexity and nuances of effective teaching. Many aspects of the evaluation system will likely require further clarity and attention. You can expect to receive questions and feedback on a range of issues. For example, there may be questions about paperwork, to questions about what kinds of meetings will be needed and whether the new system will affect teachers' job security or salary.

Throughout the development process, the Association must be an equal partner, because of its knowledge and understanding of what it means to be an effective teacher. The Association is committed to supporting state and local leaders as they implement and monitor new teacher evaluation systems.

Review the feedback from the pilot and ask the following questions to yourself and to those with whom you are bargaining or jointly developing the system:

- Which aspects of the evaluation system are not working as intended? Can they be modified, or do they need to be replaced?

- Is there any indication from the pilot that the evaluation results are influencing district-wide professional development planning?
- In practice, does the evaluation system align with district or state professional teaching standards, InTASC standards, or other teaching standards?
- Are instructional coaches or other resources available to support professional development needs identified in the evaluation results?
- Will there be adequate funding and resources to support widespread improvements to professional development based on the evaluations?

Even as states and local districts continue to develop, implement, and monitor teacher evaluation systems, new initiatives are under way that focus on teaching effectiveness and its relationship to student learning. We need to help shape initiatives that will enhance student learning and contribute to greater teaching effectiveness. Such initiatives must be coupled with sound policies if we are to create a more effective teaching profession.

Conclusion

NEA is committed to helping teachers lead in the development of new evaluation systems. This resource guide highlights some of the key areas that Association leaders need to address as they reform their teacher evaluation systems. State and local affiliates have a critical role to play in developing new evaluation systems—from defining teaching effectiveness at the local level to making sure that pilot tests result in changes to the proposed teacher evaluation system. Together, we can build comprehensive systems that improve the knowledge, skills, dispositions, and instructional practices of professional educators while taking into account the context of the local situation.

References

Baker, E. L., Bartson, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). "Problems with the use of student test scores to evaluate teachers." Washington, DC: Economic Policy Institute. Retrieved April 15, 2011, from http://epi.3cdn.net/ b9667271ee6c154195_t9m6iij8k.pdf

Blank, R. (2010). "State growth models for school accountability: Progress on development and reporting measures of student growth." Washington, DC: The Council of Chief State School Officers.

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2008). "Examining district guidance to schools on teacher evaluation policies in the Midwest region." Naperville, IL: REL Midwest.

Braun, Henry. (2005). "Using student progress to evaluate teachers: A primer on value-added model." Educational Testing Service. Retrieved April 13, 2011, from www.ets.org/Media/Research/pdf/ PICVAM.pdf

Braun, H., Chudowsky, N., Koenig, J. (2010). "Getting value out of value-added: Report of a workshop." Washington, DC: The National Academies Press. Retrieved April 15, 2011, from www.nap.edu/ catalog.php?record_id=12820

Buddin, R., McCaffrey, D. F., Kirby, S. N., & Xia, N. (2007). "Merit pay for Florida teachers: Design and implementation issues" (Working Paper WR-508-FEA). Santa Monica, CA: RAND Corporation.

Byrd, A., & Rasberry, M. (2011). *Teacher and Teaching Effectiveness*. Center for Teaching Quality. Retrieved April 15, 2011, from www.teachingquality.org/node/1118

Corcoran, S. P. (2010). "Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice." Providence, RI: Annenberg Institute for School Reform at Brown University.

Darling-Hammond, L. (2000). "Teacher quality and student achievement: A review of state policy evidence." *Education Policy Analysis Archives*, 8 (1-48).

Darling-Hammond, L., & Ducommun, C. E. (2010). "Recognizing and developing effective teaching in all schools: What policymakers should know and do." Washington, DC: National Education Association and the American Association of Colleges of Teacher Education. Retrieved April 13, 2011, from http://bit.ly/nyVFiZ

Duffett, A., Farkas, S., Rotherham, A. J., & Silva, E. (2008). "Waiting to be won over: Teachers speak on the profession, unions, and reform." Washington, DC: Education Sector.

Goe, L. (2010). "Evaluating teachers with multiple measures." Washington, DC: American Federation of Teachers. Retrieved April 15, 2011, from www.atfunion.org/files/dmfile/ EvaluatingTeachingwithMultipleMeasures.pdf

Goe, L., & Little, O. (2010). "Teacher evaluation measures and systems: How they can improve teaching and learning." Washington, DC: National Education Association. Retrieved March 15, 2010, from http://bit.ly/oxNHRU

Goe, L., Bell, C., & Little, O. (2008). "Approaches to evaluating teacher effectiveness: A research synthesis." Washington, DC: National Comprehensive Center for Teacher Quality.

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., et al. (2005). "Policymakers' guide to growth models for school accountability: How do accountability models differ?" Washington, DC: Council of Chief State School Officers. Retrieved April 15, 2011, from www. ccsso.org/Documents/2005/Policymakers_Guide_To_Growth_2005.pdf

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). "Identifying effective teachers using performance on the job." Washington, DC: Brookings Institution.

Hakel, M. D., Koenig, J. A., & Elliott, S. W. (Eds.). (2008). *Assessing Accomplished Teaching: Advanced-Level Certification Programs*. Washington, DC: National Academies Press.

Hamilton, L., Stecher, B., Marsh, J., McCombs, J., Robyn, A., Russell, J., et al. (2007). "Standards-based accountability under No Child Left Behind experiences of teachers and administrators in three states." Santa Monica, CA: RAND Corporation.

Harris, D. N. (2009). "Would accountability based on teacher value-added be smart policy? An examination of the statistical properties and policy alternatives." *Education Finance and Policy*, 4(4), 319–350.

Kane, T. J., Taylor, E.S., Tyler, J. H., & Wooten, A. L. (2011). *Evaluating Teacher Effectiveness*. Retrieved May 18, 2011, from http://educationnext.org/evaluating-teacher-effectiveness

Kane, T. J., & Staiger, D. O. (2008). "Estimating teacher impacts on student achievement: An experimental evaluation" (Working Paper No. 14607). Washington, DC: National Bureau of Economic Research.

Koedel, C., & Betts, J. R. (2009). "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique" (Working Paper No. 2009-01). Washington, DC: National Bureau of Economic Research.

Kupermintz, H. (2003). "Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value-added assessment system." *Educational Evaluation and Policy Analysis*, 25(3), 287–298.

Lasagna, M., Laine, S., & Behrstock-Sherratt, E. (2011). *Improving Teacher Quality: A Guide for Education Leaders*. San Francisco: Jossey-Bass.

Learning Point Associates. (2010). "State legislations: Emerging trends reflected in the state phase I Race to the Top applications." Naperville, IL: Author. Retrieved April 15, 2011, from www.learningpt.org/pdfs/RttT_State_Legislation.pdf

Little, O. (2009). "Teacher evaluation systems: The window for opportunity and reform." Washington, DC: National Education Association. Retrieved March 15, 2011, from http://bit.ly/o5o8Kd

Little, O., Goe, L., & Bell, C. (2009). "A practical guide to evaluating teacher effectiveness." Washington, DC: National Comprehensive Center for Teacher Quality.

Martineau, J. A. (2006). "Distorting value-added: The use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability." *Journal of Educational and Behavioral Statistics*, 31(1), 35–62.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). "Evaluating value-added models for teacher accountability." Santa Monica, CA: RAND Corporation. www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Miller, R. (2009). "Adding value to discussions about value-added: A new framework for talking about teacher effectiveness." Washington, DC: Center for American Progress.

National Education Association. (2010). "Teacher assessment and evaluation: The National Education Association's framework for transforming education systems to support effective teaching and improve student learning." Washington, DC: Author. Retrieved April 15, 2011, from www.nea.org/assets/docs/TeachrAssmntWhtPaperTransform10_2.pdf

Pecheone, R. L., & Chung Wei, R. (2009). Review of the report "The widget effect: Our national failure to acknowledge and act on teacher differences." East Lansing, MI: Great Lakes Center for Education Research and Practice. Retrieved April 15, 2011, from http://greatlakescenter.org/docs/Think_Twice/TT_Pecheone_Widget.pdf

Perlman, C. (2002). "An introduction to performance assessment rubrics." In C. Boston (Ed.), *Understanding Scoring Rubrics: A Guide for Teachers* (pp. 25–33). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Withman, P. J., Milanowski, A. T., & Thorn, C. A. (2010). "The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades." Washington, DC: Center for Educator Compensation Reform. Retrieved April 15, 2011, from www.cecr.ed.gov/pdfs/guide/other69Percent.pdf

Reform Support Network. (2010). "Measuring student growth for teachers in nontested grades and subjects: A primer." Washington, DC.

Rothstein, J. (2010). "Teacher quality in educational production: Tracking, decay, and student achievement." *The Quarterly Journal of Economics*. Retrieved April 15, 2011, from http://hvrd.me/oDNheS

Rowan, B., Correnti, R., & Miller, R. J. (2002). "What large-scale, survey research tells us about teacher effects on student achievement: Insights from the 'Prospects' study of elementary schools." Philadelphia: CPRE Publications.

Secretary's Priorities for Discretionary Grant Programs, 75 Fed. Reg. 47,288 (proposed Aug. 5, 2010). Retrieved April 15, 2011, from www2.ed.gov/legislation/FedRegister/other/2010-3/080510d.pdf

US Department of Education. (2010). "Race to the Top application for phase 2 funding." CFDA Number 84.395A. Retrieved April 15, 2011, from www2.ed.gov/programs/racetothetop/applicant.html

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: New Teacher Project.

Additional Resources

Overview of Teaching Effectiveness

Berry, B. (2009). "Children of poverty deserve great teachers: One union's commitment to changing the status quo." Carrboro, NC: Center for Teaching Quality; Washington, DC:

National Education Association. Retrieved April 15, 2011, from www.nea.org/assets/docs/HE/children_of_poverty_09.pdf

Berry, B. (2010). "Teacher effectiveness: The conditions that matter most and a look to the future." Carrboro, NC: Center for Teaching Quality. Retrieved April 15, 2011, from www.teachingquality.org/sites/default/files/TG_effectiveness_mar30_1.pdf

Bill & Melinda Gates Foundation. (2010). "Empowering effective teachers: Strategies for implementing reforms." Seattle: Author. Retrieved April 15, 2011, from www.gatesfoundation.org/united-states/Documents/empowering-effective-teachers-empowering-strategy.pdf

Chait, R. (2009). "Ensuring effective teachers for all students: Six state strategies for attracting and retaining effective teachers in high-poverty and high-minority schools." (2009, May 18). Washington, DC: Center for American Progress. www.americanprogress.org/issues/2009/05/effective_teachers.html

Chait, R. (2009). "From qualifications to results: Promoting teacher effectiveness through federal policy." Washington, DC: Center for American Progress. www.americanprogress.org/issues/2009/01/qualifications_to_results.html

Chait, R., & Miller, R. (2010). "Treating different teachers differently: How state policy should act on differences in teacher performance to improve teacher effectiveness and equity." Washington, DC: Center for American Progress. www.americanprogress.org/issues/2010/03/different_teachers.html

Goe, L., Bell, C., & Little, O. (2008). "Approaches to evaluating teacher effectiveness: A research synthesis." Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved April 15, 2011, from www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Green, E. (March 7, 2010). "Building a better teacher." *The New York Times Magazine*.

Harris, D., & Goe, L. (Hosts). (2009). "Evaluating teacher effectiveness: The what, how, and why of educator evaluation." (Webcast). Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved April 15, 2011, from www.tqsource.org/webcasts/evaluateEffectiveness/

Hope Street Group. (2009). "Policy 2.0: Using open innovation to reform teacher evaluation systems." Los Angeles: Author. Retrieved April 15, 2011, from www.hopestreetgroup.org/content/images/stories/documents/policy2.0policypaper.pdf

Mathers, C., & Oliva, M. (with Laine, S. W. M.). (2008). "Improving instruction through effective teacher evaluation: Options for states and districts." Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved April 15, 2011, from www.tqsource.org/publications/February2008Brief.pdf

Raue, K., MacAllum, K., & Ristow, L. (2008). "Building systems to recognize teachers of excellence: Lessons from the Ohio Teacher Incentive Fund." Columbus: Ohio Department of Education. Retrieved April 15, 2011, from http://education.ohio.gov/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=62741

Sawyer, L. (2001). "Revamping a teacher evaluation system." *Educational Leadership*, 58(5), 44–47.

Steiner, L. (2009). "Determining processes that build sustainable teacher accountability systems." Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved April 15, 2011, from www.tqsource.org/publications/October2009Brief.pdf

Toch, T., & Rothman, R. (2008). "Rush to judgment: Teacher evaluation in public education." Washington, DC: Education Sector.

Zinth, J. D. (2010). "Teacher evaluation: New approaches for a new decade." Denver: Education Commission of the States. Retrieved April 15, 2011, from www.ecs.org/clearinghouse/86/21/8621.pdf

Communication

Coggshall, J. (2007). "Communication framework for measuring teacher quality and effectiveness: Bringing coherence to the conversation." Retrieved April 15, 2011, from www.tqsource.org/publications/NCCTQCommFramework.pdf

Center for American Progress. (2009). "Stimulating excellence: Unleashing the power of innovation in education." Washington, DC: Author. Retrieved April 15, 2011, from www.americanprogress.org/issues/2009/05/pdf/education_entrepreneurs.pdf

Hansen, M. (2009). "State intervention and contract choice in the public teacher labor market." Seattle: University of Washington and Center on Reinventing Public Education.

Hess, F. (2009). "Cages of their own design." *Educational Leadership*, 67(2), 28–33.

Koppich, J., Prince, C. D., Guthrie, J. W., Schuermann, P. J. (2010). "Stakeholder engagement and communication: Guide to implementation: Resources for applied practice." Washington, DC: Center for Educator Compensation Reform. Retrieved April 15, 2011, from www.cecr.ed.gov/pdfs/guide/CECRStakeholderEngagement.pdf

Meyers, J., Meyers, B., Gelzheiser, L., & Borelli, C. (1997). "Making connections in school reform: An examination of communication strategies." Paper presented at the annual meeting of the American Educational Research Association, Chicago, March 24–28.

Taylor, W. L., & Rosario, C. (2007). "Fresh ideas in collective bargaining: How new agreements help kids." Washington, DC: Citizens' Commission on Civil Rights.

Toch, T., & Rothman, R. (2008). "Rush to judgment: Teacher evaluation in public education." Washington, DC: Education Sector.

West, K. L., & Mykerezi, E. (2009). "Effects of teachers' unions on qualification-specific and incentive-based teacher compensation." Washington, DC: National Council on TeacherQuality. Retrieved April 15, 2011, from www.nctq.org/tr3/conference/docs/tr3_conf_west_and_mykerezi.pdf

Wright, R. E. (2003). "Difficulties in marketing the concept of merit pay for primary and secondary teachers." *Research for Educational Reform*, 8(3), 38–45.

Developing Multiple Measures

Antonioni, D. (1996). "Designing an effective 360-degree appraisal feedback process." *Organizational Dynamics*, 25(2), 24–38.

Cavalluzzo, L. C. (2004). "Is National Board Certification an effective signal of teacher quality?" Alexandria, VA: CNA Corporation.

Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching* (2nd ed.). Alexandria, VA: ASCD.

Denner, P. R., Salzman, S. A., & Bangert, A. W. (2001). "Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples." *Journal of Personnel Evaluation in Education*, 15(4), 287–307.

Ellett, C. D., & Teddlie, C. (2003). "Teacher evaluation, teacher effectiveness, and school effectiveness: Perspectives from the USA." *Journal of Personnel Evaluation in Education*, 17(1), 101–128. Retrieved April 15, 2011, from www.ctu.edu.vn/centers/cfl/learningresource/ebooks/10.pdf

Escamilla, P., Clarke, T., & Linn, D. (2000). "Exploring teacher peer review." Washington, DC: National Governors Association. Retrieved April 15, 2011, from www.nga.org/Files/pdf/000125PEERREVIEW.pdf

Goe, L., & Croft, A. (2009). "Methods of evaluating teacher effectiveness." Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved April 15, 2011, from www.tqsource.org/publications/RestoPractice_EvaluatingTeacherEffectiveness.pdf

Goldhaber, D. (2005). "Teacher licensure tests and student achievement: Is teacher testing an effective policy?" Seattle: Center on Reinventing Public Education.

Goldhaber, D. (2007). "Everyone's doing it, but what does teacher testing tell us about teacher effectiveness?" *Journal of Human Resources*, 52(4), 765–794.

Goldhaber, D., & Anthony, E. (2004). "Can teacher quality be effectively assessed?" Washington, DC: Urban Institute.

Goldschmidt, P., Roschewski, P., Choi, K. Auty, W., Hebbler, S., Blank, R., et al. (2005). "Policymaker's guide to growth models for school accountability: How do accountability models differ?" Washington, DC: Council of Chief State School Officers. Retrieved April 15, 2011, from www.ccsso.org/Documents/2005/Policymakers_Guide_To_Growth_2005.pdf

Halverson, R., Kelley, C., & Kimball, S. (2004). "Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instruction practice." In W. K. Hoy & C. G. Miskel (Eds.), *Educational Administration, Policy, and Reform: Research and Measurement Research and Theory in Educational Administration*. Greenwich, CT: Information Age Press.

Harris, D. N., & Rutledge, S. A. (2007). "Models and predictors of teacher effectiveness: A review of the literature with lessons from (and for) other occupations." New York: Columbia University Teachers College.

Hibpshman, T. (2004). "A review of value-added models." Frankfort: Kentucky Education Professional Standards Board.

Johnson, S. M., Papay, J. P., Fiarman, S. E., Munger, M. S., & Qazilbash, E. K. (2010). "Teacher to teacher: Realizing the potential of peer assistance and review." Washington, DC: Center for American Progress. Retrieved April 15, 2011, from www.americanprogress.org/issues/2010/05/pdf/par.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). "Evaluating value-added models for teacher accountability." Santa Monica, CA: RAND Corporation. Retrieved April 15, 2011, from www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

Moore, Z., & Bond, N. (2002). "The use of portfolios for in-service teacher assessment: A case study of foreign-language middle-school teachers in Texas." *Foreign Language Annals*, 35(1), 85–92.

Muncey, D. E., & Conley, S. (1999). "Teacher compensation and teacher teaming: Sketching the terrain." *Journal of Personnel Evaluation in Education*, 12(4), 363–385.

Oakley, K. (1998). "The performance assessment system: A portfolio assessment model for evaluating beginning teachers." *Journal of Personnel Evaluation in Education*, 11(4), 323–341.

Painter, B. (2001). "Using teaching portfolios." *Educational Leadership*, 58(5), 31–34.

Peterson, K. D. (2000). *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Peterson, K. D., Wahlquist, C., Bone, K., Thompson, J., & Chatterton, K. (2001). "Using more data sources to evaluate teachers." *Educational Leadership*, 58(5), 40–43.

Pool, J. E., Ellett, C. D., Schiavone, S., & Carey-Lewis, C. (2001). "How valid are the National Board for Professional Teaching Standards assessments for predicting the quality of actual classroom teaching and learning? Results of six mini case studies." *Journal of Personnel Evaluation in Education*, 15(1), 31–48.

RAND Corporation. (2005). "The promise and peril of using value-added modeling to measure teacher effectiveness" (Education Research Brief). Santa Monica, CA: Author. Retrieved April 15, 2011, from www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf

Rivers, J. C., & Sanders, W. L. "Teacher quality and equity in educational opportunity: Findings and policy implications." In W. M. Evers & L. T. Izumi (Eds.), *Teacher Quality* (pp. 13–23). Stanford, CA: Hoover Institution Press. Retrieved April 15, 2011, from http://media.hoover.org/sites/default/files/documents/0817929320_13.pdf

Stronge, J. H., & Tucker, P. D. (2003). "Handbook on teacher evaluation: Assessing and improving performance." Larchmont, NY: Eye on Education.

Tucker, B. (2009). "Beyond the bubble: Technology and the future of student assessment." Washington, DC: Education Sector.

Tung, R. (2010). "Including performance assessments in accountability systems: A review of scale-up efforts." Quincy, MA: Nellie Mae Education Fund.

TQ Center Resources

Goe, L. (2008). "Key issue: Using value-added models to identify and support highly effective teachers." Washington, DC: National Comprehensive Center for Teacher Quality. Tools, tips, and strategies for using longitudinal statistical information to explore teacher effectiveness. Retrieved April 15, 2011, from www2.tqsource.org/strategies/het/UsingValueAddedModels.pdf

Goe, L., & Stickler, L. M. (2008). "Teacher quality and student achievement: Making the most of recent research." Washington, DC: National Comprehensive Center for Teacher Quality. This brief identifies several teacher quality variables—including specific teacher qualifications, characteristics, and classroom practices—that are strong and consistent predictors of student achievement. Retrieved April 15, 2011, from www.tqsource.org/publications/March2008Brief.pdf

Holdheide, L. R., Goe, L., Croft, A., & Reschly, D. J. (2010). "Challenges in evaluating special education teachers and English Language Learner specialists." Washington, DC: National Comprehensive Center for Teacher Quality. This TQ Research & Policy Brief offers policy and practice recommendations for regions, states, and districts to help in their efforts to create valid, reliable, and comprehensive evaluation systems for all teachers as they work to improve the achievement of all students. Retrieved April 15, 2011, from www.tqsource.org/publications/July2010Brief.pdf

National Comprehensive Center for Teacher Quality. (2010). "Enhancing teacher evaluation: effective practices for evaluating teachers of all students." (Webcast). Washington, DC: Author. This webcast includes an overview of the major issues related to evaluating teachers of at-risk populations, a discussion of the existing research, and a review of evaluation strategies and recommendations. Retrieved April 15, 2011, from www.tqsource.org/webcasts/evaluatingTeachers/index.php

NEA Resources

National Education Association. (2010). "Teacher assessment and evaluation: The National Education Association's framework for transforming education systems to support effective teaching and improve student learning." Washington, DC: Author. Retrieved April 15, 2011, from www.nea.org/assets/docs/TeachrAssmntWhtPaperTransform10_2.pdf

National Education Association. (2010). "Bargaining/advocacy guidance on value-added models (VAMs): Ask the right questions and get the right answers." Washington, DC: Author. Retrieved May 23, 2011, from http://bit.ly/rm06fb

National Education Association. (2009). "Teacher evaluation systems: The window for opportunity and reform." Washington, DC: Author. Retrieved April 15, 2011, from http://bit.ly/o5o8Kd

Glossary: Building a Common Language

Attainment model is a method for measuring how students perform at one point in time. For example, the percentage of fourth graders who were scoring at proficient or above in 2010.

Classroom observations are used by evaluators to make judgments of teachers' practice in the classroom. Classroom observations are the most common form of teacher evaluation and vary widely in how they are conducted and what they assess. High-quality classroom observation instruments are based on standards and contain well-specified rubrics that delineate consistent assessment criteria for each standard of practice. Evaluators should be trained to ensure accuracy and consistency in scoring. A transparent system ensures that all educators who will be observed know how the process will be conducted and how the findings will be used.

Comparability is the extent to which student learning is being measured consistently across classrooms, schools, and districts so that scores have the same meaning both within and across subject areas and grade levels.

Disaggregate is to separate into different categories or to show individual results.

Effective teacher is an individual teacher who produces substantial student outcomes. Effective teacher is often narrowly defined as a teacher who contributes more to student test scores than other teachers. This focus implies evaluating teacher performance by treating teachers and students as single units rather than interconnected members of a learning community. Not to be confused with effective teaching (which addresses the complexity of teaching by recognizing contextual factors).

Effective teaching consists of instruction that enables all students to meet or exceed ambitious goals for student learning (adapted from Darling-Hammond, 2010). Effective teaching is in part a function of individual teacher talent, knowledge, and skills, but it is also highly influenced by the conditions in which teaching takes place—the school leadership, the quality of curriculum materials and resources, the opportunities teachers have for professional growth and learning, the size of their workloads, and the time teachers have to prepare, among other factors.

Evaluation of student artifacts and work judged according to rubrics means that evaluators rate lesson plans, teacher assignments, teacher-created assessments, scoring rubrics, and student work on particular criteria, such as rigor, authenticity, intellectual demand, alignment to standards, clarity, and comprehensiveness. Evaluators typically use an evaluation tool or rubric to make judgments about the quality of student artifacts.

Evaluation tools are models, rubrics, instruments, and protocols that are used by evaluators to assess teachers' performances.

Formative teacher evaluation is the assessment of teachers' practices for the purposes of supporting or improving teachers' practices.

Formative assessments are assessments used while the learning process is under way or recently completed, and designed to inform current or future instruction.

Goal-driven professional development plans are evaluation tools that offer teachers the opportunity to set their own ambitious but feasible objectives for their professional growth in collaboration with their principal or other colleagues. Some tools require teachers to specify the professional development in which they will participate to ensure their students achieve their growth objectives.

Growth measures are assessments of students' improvements in learning from one point in time to another point in time.

Growth models or individual student growth models generally refer to models that measure progress by tracking the achievement scores of the same students from one year to the next, with the intent of determining whether the students in a given group are making progress (Goldschmidt, Roschewski, Choi, Auty, Hebbler, et al., 2005). For example, the model can compare the performance of this year's eighth graders with the performance of the same students the previous year in the seventh grade. Growth models control for mobility of students between schools from year to year as well as students' prior achievement and the effects of their family background (Blank, 2010). Types of growth models include growth-to-proficiency, linear growth, projection, transition table, and value-added growth models.

Growth-to-proficiency models measure whether students are on track to meet standards for proficient and above.

Multiple measures of student learning are the various types of assessment of students' learning, for example, value-added or growth measures, curriculum-based tests, pre- and posttests, capstone projects, oral presentations, performances, and artistic or other projects.

Multiple measures of teaching effectiveness are the various types of assessments of teachers' performance, for example, classroom observations, student test-score data, self-assessments, and student or parent surveys.

Parent surveys are questionnaires that typically ask parents to rate teachers on a scale from 1 to 5 (where 1 = very effective, and 5 = not at all effective) for various aspects of teachers' practice (e.g., course content, usefulness of feedback, quality of homework, quality of communication) as well as the extent to which they are satisfied with the teacher's instruction.

Pre- and posttests of student growth are student-achievement tests that measure the content of the curriculum of a particular course. They are taken at the beginning of some time period (usually a semester or year) and then toward the end of that time period to obtain a measure of student growth. Many pre- and posttest models also include midyear assessments and formative assessments for teachers to adjust instruction throughout the course or year.

Reliability is a measure of the degree to which an instrument measures something consistently. A validated instrument must be evaluated for how reliable the results are across raters and contexts. Discussion of methods for measuring teaching effectiveness often make reference to rater reliability—whether or not raters have been trained to score reliably. Scoring reliably means being able to do the following: rate consistently with standards, rate consistently with other raters (referred to as interrater reliability), and rate consistently across observations and contexts. Ratings should not be influenced by factors such as the time of day, time of year, or subject matter being taught, and they should be consistent across observations of the same teacher (from Little, Goe, & Bell, 2009).

Self-assessments are surveys, instructional logs, or interviews in which teachers report on their work in the classroom, the extent to which they are meeting standards, and in some cases the impact of their practice. Self-assessments may consist of checklists, rating scales, or rubrics, and they may require teachers to indicate the frequency of particular practices.

Standards for professional practice are a set of ideals for what behaviors, skills, knowledge, and dispositions teachers should exhibit.

Student growth data are information about change in students' performance on some measure such as a test between two or more points in time.

Student progress is the extent to which individual students are moving through the learning process.

Student surveys are questionnaires that typically ask students to rate teachers on a scale from 1 to 5, where 1 = very effective, and 5 = not at all effective, for various aspects of teachers' practice (e.g., relevance of course content, usefulness of feedback) as well as how much students say they learned or the extent to which they were engaged.

Summative teacher evaluation is the assessment of teachers' practice for the purpose of making high-stakes personnel decisions.

Teacher growth and development system is a comprehensive performance assessment system that incorporates multiple measures of both teacher evaluation and student learning and has the intent of improving the knowledge, skills, dispositions, and classroom practices of professional educators. Beyond a simple evaluation system, a teacher growth and development system is connected closely to other key aspects of the teaching continuum (induction, professional development, promotion, etc.) and recognizes the importance of formative assessment for the improvement of teaching.

Teacher portfolios and evidence binders are a collection of materials that exhibit evidence of teaching practice, school activities, and student progress. The teacher usually compiles them himself or herself. They may include teacher-created lesson or unit plans, descriptions of the classroom context, assignments, student work samples, videos of classroom instruction, notes from parents, and the teacher's analyses of their students' learning in relation to their instruction. Evidence binders often provide specific requirements for inclusion and require a final teacher-led presentation of the work to an evaluation team.

Teacher quality refers to teachers who are certified, have a bachelor's degree, and have a major in their subject area. Qualified teachers hold credentials certifying that they have successfully completed a state-approved (often nationally accredited) teacher preparation program, have demonstrated their good character (usually through a criminal background check), and hold a bachelor's degree. Moreover, states almost always require an examination of content and pedagogy for state certification.

Unique identifiers are numbers that are assigned to each student and teacher in a school; they can be matched to data about that student's or teacher's performance.

Validity is a measure of the degree to which an interpretation of an evaluation score is supported by evidence. For a measure of teaching effectiveness to be valid, evidence must support the argument that the measure actually assesses the dimensions of teaching effectiveness it claims to measure and not something else. It also is essential to have evidence that the measure is valid for the purposes for which it will be used. Instruments cannot be valid in and of themselves; an instrument or assessment must be validated for particular purposes (Gordon, Kane, & Staiger, 2006; Messick, 1989, as cited in Little, Goe, & Bell, 2009).

Value-added student test scores are a measure of the contribution that a teacher makes on the growth of his or her students' test scores, taking into account those students' test scores in prior years and also often taking into account socioeconomic and other factors that might affect growth in scores.

Appendices

Appendix A
Resolution D-20. Education Employee Evaluation

The National Education Association believes that formal performance-based evaluations must be augmented by formative evaluation components in order to ensure the continuing competency of all education employees in their respective fields.

Effective evaluation procedures supported by professional development programs will enable all education employees to keep abreast of developments in their areas of specialization. Such procedures, with sufficient resources, can help ensure job competency, identify deficiencies in performance, and provide options such as counseling, training programs, a remediation plan, and opportunities to observe peers.

If after such an evaluation and after being given sufficient time, training, and opportunity for improvement, a person is then formally reevaluated and incompetence can be documented, dismissal proceedings with guaranteed due process may be instituted. Such proceedings must be implemented by administrators or evaluators who are properly trained and held accountable for appropriate and fair evaluation systems.

The Association also believes that the use of student achievement measures such as standardized test scores or grades to determine the competency, quality, or effectiveness of any professional educator is inappropriate and is not a valid measure.

The Association further believes that classroom teachers, without fear of discipline or negative evaluation, must be given the discretion to modify the pace of predetermined progress rates, dictated pacing guides, and mandated scripted lesson pacing charts.

The evaluation procedure should be cooperatively developed and maintained in conjunction with representatives selected by the local affiliate and should include the following:

- Clear performance expectations that are specific to the job description
- Regular observation of job performance with advance notice and discussion of evaluation visits and a timely consultation after each visit
- A written evaluation report that is provided to the person being evaluated
- Opportunity for a written response prior to the placement of the evaluation in the personnel file
- An employee improvement plan that will not interfere with any earned pay increase or longevity credit
- A provision for an alternative evaluator or an opportunity for an alternative evaluation report to ensure a fair and unbiased evaluation of the education employee
- An unbiased appeals process with an evidentiary hearing under oath

Appendix A
Resolution D-20. Education Employee Evaluation

The Association further believes that procedures for evaluation of administrators should include evaluations by education employees who are directly supervised by them.

By participating in an evaluation process, an education employee shall not waive his or her right to due process in any subsequent contractual or legal proceeding (National Education Association Handbook, 2011).

Appendix B

Evaluation System Models
(The presentation of the models do not represent an endorsement of NEA as best practice.)

Evaluation System Model		
Models	**Strengths**	**Limitations**
Austin, TX Student learning objectives with pay for performance, group and individual student learning objectives assessed with comprehensive rubric Teachers determine two student learning objectives per semester or year to track using pre- and post-assessments	Teachers take an active role in determining student learning outcomes Good professional opportunity for teachers If objectives are of high quality and teachers plan instruction to meet them, then students should benefit	Heavily dependent on administrators understanding and time commitment to supervision Not "comparable across classrooms" because teachers set the objectives and they can vary widely "Rigor" dependent upon the evaluator's interpretation and/or having an appropriate rubric
http://archive.austinisd.org/inside/initiatives/compensation/slos.phtml		
Chicago, IL The Excellence in Teaching Project uses Charlotte Danielson's Framework for Teaching to define effective practice, encourage conversations about instruction, and identify areas for professional growth.	Focuses on true collaboration within schools and across the district Project intends to establish ways to provide continuous professional growth for teachers without requiring them to leave the classroom Evaluators must be trained for interrater reliability in their use of rubrics	Student growth as a "significant" factor will be determined at the district level, which makes comparability across districts difficult to attain Principals note that implementing the framework with fidelity requires a lot of time, which must be added to their ever-expanding list of responsibilities
http://www.isbe.net/PEAC/default.htm		

Appendix B

Evaluation System Models
(The presentation of the models do not represent an endorsement of NEA as best practice.)

Evaluation System Model		
Models	**Strengths**	**Limitations**
Delaware Teacher participation in identifying grade and subject measures that then must be approved by the state.	Teachers take an active role in determining student learning outcomes Good professional opportunity for teachers If measures are of high quality and teachers plan instruction to meet them, then students should benefit	Teachers are not trained psychometricians and should not be expected to develop tests that are valid and reliable for high-stakes decisions Not "comparable across classrooms" because measures can vary widely between subjects Time-consuming
www.doe.k12.de.us/csa/dpasii/student_growth/default.shtml		
Georgia Comprehensive rubric, includes student achievement.	Addresses multiple aspects of teaching that are often overlooked by other frameworks Model is aligned to standards	Time-consuming for evaluators in terms of both training and implementation
System: www.gadoe.org/tss_teacher.aspx Rubric: bit.ly/na88Lx		
Hillsborough, FL District is creating assessments and tests for all subjects. Hillsborough has more than 600 assessments that examine student growth for nearly all grades and subjects.	Tests are used in subjects outside the NCLB-tested grades and subjects to give a measure of student growth that will be associated with teacher effectiveness	Time-consuming Resource-intensive— Hillsborough has funding from the Gates Foundation for this work
http://communication.sdhc.k12.fl.us/empoweringteachers/		

Appendix B

Evaluation System Models
(The presentation of the models do not represent an endorsement of NEA as best practice.)

Evaluation System Model		
Models	**Strengths**	**Limitations**
Massachusetts All judgments of practice are guided by the four Principles of Effective Teaching or Effective Administrative Leadership. The entry point is the educator's self-assessment, which is followed by multiple observations and discussions with evaluators and/or peers. Multiple measures of student learning and outcomes are considered in the overall performance rating; these must include the MCAS Student Growth Percentile, a state-specific metric, if available. Other evidence of educator practice must be considered. The processes, procedures and details–such as the emphasis placed on student learning outcomes–are determined through collective bargaining.	Reflective practice and improving practice are at the heart of the reform recommendations, which apply to all educators: teachers, principals, and superintendents Uses three different elements –observations/evidence of classroom practice, multiple measures of student learning and outcomes, evidence of contributions to professional culture – to create a whole picture of the educator's performance Multiple measures of student learning and outcomes are significant in that they must validate the judgments about practice on the four standards; they are not weighted or given a set percentage	Complete change from current practice, which is evaluator-driven, to one that is educator-driven. Will require professional development for all licensed personnel to understand their role in the new system The MCAS Student Growth Percentile applies to only 17 percent of teaching work force; applies to almost 100 percent of the administrative work force. Districts are charged with developing standards-based pre/post assessments to address the RTTT requirement. As written, the proposed regulations lack specificity about the content, time and resources devoted to the professional growth or improvement plans that are informed
www.doe.mass.edu/boe/docs/0311/item1_breakthroughframework.pdf		
Minnesota Q Comp Program Local districts design and collectively bargain a plan that contains five components—career ladder/advancement options, job-embedded professional development, teacher evaluation, performance pay, and an alternative salary schedule.	The district develops all the programs, which must be done collaboratively with teachers The program addresses components that are important in the systemic development of teacher effectiveness	Ensuring rigor across the Q Comp sites can be difficult The state evaluation of the program does not ensure that the program is effective at meeting its goals There is not enough funding to support all the districts that have expressed interest
http://education.state.mn.us/MDE/Teacher_Support/QComp/index.html		

Appendix B

Evaluation System Models
(The presentation of the models do not represent an endorsement of NEA as best practice.)

Evaluation System Model		
Models	**Strengths**	**Limitations**
New Haven, CT A collaboratively designed proposed teacher evaluation and development system with a set of guiding design principles that allow the new system to (1) enable professional evaluation and coaching for all teachers; (2) support deep individualized development for teachers aligned to student learning goals; and (3) allow for the consequential recognition of both outstanding and poor performance.	Teachers take an active role in determining student learning outcomes Good professional opportunity for teachers If objectives are of high quality and teachers plan instruction to meet them, then students should benefit	Heavily depends on evaluator's understanding and time commitment to supervision Not "comparable across classrooms" because teachers set the objectives and they can vary widely "Rigor" depends upon the evaluator's interpretation and having an appropriate rubric
www.nhps.net/scc/index		
Rhode Island Student learning objectives combined with teacher observations, professionalism, and student growth.	Combination of student learning objectives and other measures illustrates a multiple-method system in development Professionalism is addressed by this system Teachers take an active role in determining student learning outcomes	Heavily depends on administrators implementing the system Uses student growth as a large percentage of measuring teacher effectiveness
www.ride.ri.gov/assessment/DOCS/Asst.Sups_CurriculumDir.Network/Assnt_Sup_August_24_rev.ppt		

Appendix B

Evaluation System Models
(The presentation of the models do not represent an endorsement of NEA as best practice.)

Evaluation System Model		
Models	**Strengths**	**Limitations**
Teacher Advancement Program (TAP) TAP requires that teachers in tested subjects be evaluated with value-added models All teachers are observed in their classrooms (using a Charlotte Danielson–type instrument) at least three times per year by different observers (usually one administrator and two teachers who have been appointed to the role) Teacher effectiveness (for performance awards) determined by combination of value-added and observations Teachers in untested subjects are given the school-wide average for their value-added component, which is combined with their observation scores.	Value-added becomes everyone's responsibility, which should encourage efforts from teachers in untested subjects to support teachers in tested subjects Multiple yearly observations should be more informative and produce more reliable information about practice Evaluators must attend frequent recalibration trainings Professional development must align with results	Concerns about "fairness" when student achievement and progress toward learning goals "counts" for only a few teachers Teachers in nontested subjects are given the school-wide average for their value-added component, which is combined with their observation scores Tells you nothing about how teachers in other subjects are performing in terms of student learning growth (grades are not always good indicators) The model is costly to implement and sustain
www.tapsystem.org/		
Washington, DC, IMPACT Variation in how groups of teachers are measured—50% standardized tests for some groups, 10% other assessments for untested subjects and grades.	An example of a multiple-measure teacher evaluation system Has differentiated assessments for different types of teachers (e.g., teachers in tested grades, teachers in untested grades, counselors)	Uses student growth as a large percentage of measuring teacher effectiveness Resource-intensive—requires additional staff to serve as evaluators System implementation caused much contention in the district
http://1.usa.gov/pvroc3		