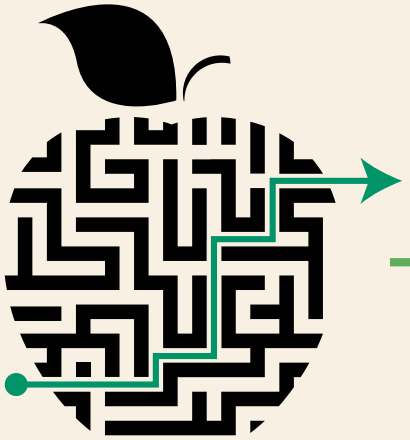


EVALUATING TEACHERS AND PRINCIPALS:

Developing Fair, Valid, and Reliable Systems

Kelly Burling, Ph.D.

Director, Center for Educator Effectiveness
919-627-8893
kelly.burling@pearson.com

EVALUATING TEACHERS AND PRINCIPALS: DEVELOPING FAIR, VALID, AND RELIABLE SYSTEMS

1. DEFINE THE CONSTRUCT

What is an effective educator?

2. DEPLOY MULTIPLE INDICATORS

What evidence characterizes good teaching and school leadership?

3. DEVELOP A CLEAR COMPOSITE RATING

What weights should each indicator have and who should be involved in the decision?

4. CLARIFY DIFFERENTIATED PERFORMANCE LEVELS

What distinguishes varying levels of educator effectiveness?

5. BUILD STRONG DATA ANALYSIS AND REPORTING TOOLS

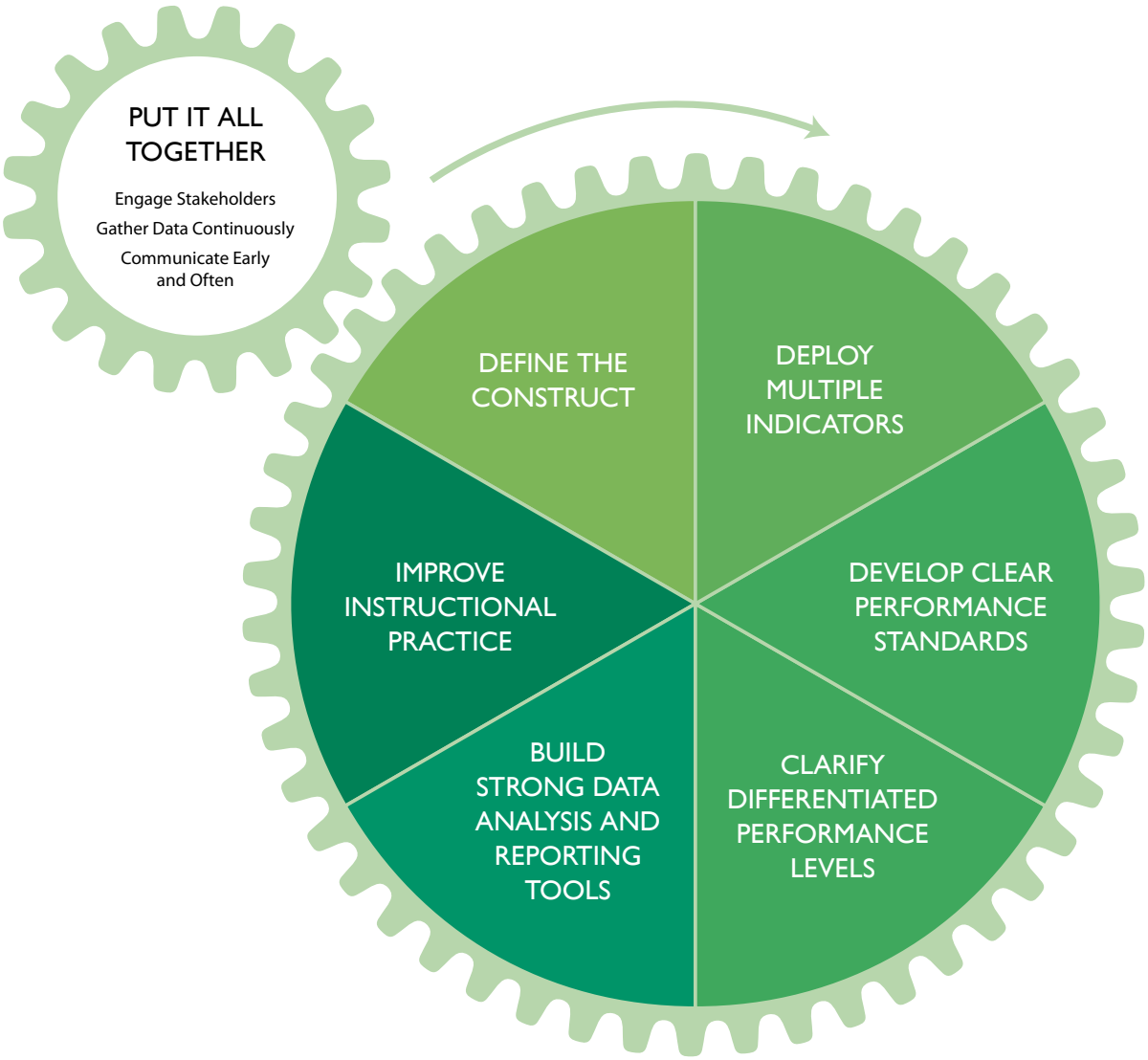
What does the information reveal about student, educator, and school performance?

6. IMPROVE INSTRUCTIONAL AND LEADERSHIP PRACTICE

How can the information target professional development to boost educator practice, student learning outcomes, and school efficacy?

PUT IT ALL TOGETHER

How do those responsible for creating educator effectiveness systems build capacity for continuous improvement?


PUT IT ALL TOGETHER
Engage Stakeholders
Gather Data Continuously
Communicate Early and Often
DEFINE THE CONSTRUCT
DEPLOY MULTIPLE INDICATORS
DEVELOP CLEAR PERFORMANCE STANDARDS
CLARIFY DIFFERENTIATED PERFORMANCE LEVELS
BUILD STRONG DATA ANALYSIS AND REPORTING TOOLS
IMPROVE INSTRUCTIONAL PRACTICE

EVALUATING TEACHERS AND PRINCIPALS: DEVELOPING FAIR, VALID, AND RELIABLE SYSTEMS

"Meaningful teacher [and principal] evaluation involve[s] an accurate appraisal of the effectiveness of teaching [and leading], its strengths and areas for development, followed by feedback, coaching, support and opportunities for professional development. It is also essential to celebrate, recognize and reward the work of teachers [and principals]

(OECD, 2009)

Teachers and school leaders are the most important school-based factors in student achievement. In order to support effective instruction and leadership, states and districts are designing and implementing aligned teacher and principal evaluation systems. These systems need to be valid, reliable, and legally defensible, particularly when the stakes are high and ratings may determine training and continuing development investments; impact compensation and promotion choices; and influence tenure and employment decisions. As such, the same standards and principles applied in high stakes student assessment systems must be used, including psychometric rigor and technical documentation.[1] In order to create a robust and comprehensive educator effectiveness system, we advise engaging stakeholders upfront in a thoughtful process to consider the best way to define effectiveness, measure it in a fair and reasonable way, and distinguish between varying levels of educator quality. One of the primary goals of educator effectiveness systems should be to generate meaningful feedback that can be used to improve teacher and leader quality.

A wide range of stakeholders needs to be engaged upfront, including education policymakers, practitioners (teachers, principals, curriculum specialists, union representatives), researchers, technology experts, human resources experts, community leaders, parents, and students. These various parties need to come together to design and implement a thoughtful and fair system that offers meaningful information on educators' strengths, weaknesses, and professional development needs.

This document is designed to guide stakeholders in developing educator effectiveness systems. We offer six considerations for establishing a fair and valid system:

1. Define the construct: what is an effective educator?
2. Deploy multiple indicators: what evidence characterizes good teaching and school leadership?
3. Develop a clear composite rating: what weights should each indicator have and who should be involved in the decision?
4. Clarify differentiated performance levels: what distinguishes varying levels of educator effectiveness?
5. Build strong data analysis and reporting tools: what does the information reveal about student, educator, and school performance?
6. Improve instructional and leadership practice: how can the information target professional development to boost educator practice, student learning outcomes, and school efficacy?

1.

DEFINE THE CONSTRUCT

What is an effective educator?

Crafting a valid and reliable educator evaluation system begins with clearly defining teacher and school leader effectiveness. The way in which educator effectiveness is defined will impact how it is measured.

Educating our students is a complex activity that involves specific knowledge, skills, and abilities as well as non-routine thinking and extensive professional judgment. Intricate, meaningful exchanges between students and educators all depend on educator expertise. Therefore, the definition of educator effectiveness should be robust and comprehensive. While there is no one widely agreed upon characterization of educator effectiveness, there is a growing research base, which helps to define what it means to be an effective teacher or leader.

For teachers, effectiveness includes demonstrating a positive impact on student outcomes over time and demonstrating teacher practices and behaviors linked to those outcomes, such as high expectations for students; planning and preparation; instructional expertise; classroom management; assessing student learning; reflection on teaching practices; and demonstration of leadership in schools.[2]

For school leaders, effectiveness includes demonstrating a positive impact on staff development, school planning, student outcomes over time, and demonstrating leadership practices and behaviors linked to those outcomes. In particular, leaders establish a strong school culture; support and develop highly effective teachers; and manage systems and align resources to support learning.[3]

This emphasis on performance is a departure from current practice, which often relies on "inputs" such as educator degrees and years of experience to determine quality. These inputs are an insufficient standard for quality. An accurate definition of effectiveness needs to focus on the impact of the work educators do in practice, not just the training and other factors they bring into the classroom. Clearly articulated educator practice standards[4] and job-task analyses of educators with demonstrated and sustained positive impacts on student growth and achievement are good building blocks from which to construct an underlying conceptual framework.

Strategies for defining educator effectiveness

- Review educator practice standards, including Interstate Teacher Assessment and Support Consortium (InTASC), Interstate SchoolLeaders Licensure Consortium (ISSLC) Standards for School Leaders, and state teaching standards)
- Conduct job-task analyses on educators with demonstrated and sustained positive impact on student growth and achievement
- Convene stakeholder meetings
- Review literature on best practices

2. DEPLOY MULTIPLE INDICATORS

What evidence characterizes good teaching and school leadership?

No single measure can capture a teacher or principal's contribution to student learning and growth. Therefore, collecting feedback and evidence from multiple indicators over time is critical.

An educator's ability to improve student learning is an important indicator of effectiveness. Student test data can be utilized to determine student growth, achievement, achievement gap-closing, and educator value-add over a given period of time. In addition to standardized tests and large-scale state testing programs, evaluations may include results from formative, classroom-based assessments or progress against student learning objectives. There are many issues to consider when measuring student achievement, including the need to:

1. Create unique identifiers for students and teachers.
2. Ensure data systems can link students and teachers.
3. Ensure data systems are longitudinal.
4. Identify additional student characteristics that factor into student outcomes and need to be measured (e.g., free and reduced-price lunch, race/ethnicity, gender, special education, and English Language Learner status).
5. Collect objective student data for students in core subjects and grades. This can include standardized test scores, periodic diagnostic assessments, and benchmark assessments that show student growth.
6. Determine how student outcomes will be measured in non-tested subjects and grades. This requires identifying, collecting, and evaluating alternative sources of evidence of student learning, such as new assessments, annual classroom achievement goals, and group metrics (by school or grade). Indicators will likely need to be different for educators in K-2, high school, special education, and non-core subject areas.
7. Statistically define how student growth will be measured (e.g., using an annual score difference or a projection measure).
8. Statistically define how teacher value-add will be measured. Value-added modeling rewards educators according to the amount of academic growth that students make over the course of a school year, regardless of students' beginning levels of academic achievement.[5]
9. Plan on collecting and monitoring at least three years of student data for stable value-added estimates. Research has shown that fewer than three years of data is not sufficient for making determinations about educator effectiveness.
10. Recognize that well-designed systems take time to roll out. The timeline should allow for one year of development and one year for validation before full implementation.

In addition to student achievement, other indicators should be included to reflect the diverse types of evidence that illustrate good teaching and/or leadership, including:

Observations

Observations are performance-based evaluations that can provide useful information about an educator's practice. However, observations must be conducted carefully, with trained evaluators using valid rubrics and formal observation protocols in order to minimize rater bias and other measurement concerns. Ongoing calibration of evaluators is also recommended.

Surveys of Students, Parents, and Staff

Surveys help educators understand the perspectives of various members of the school community and the conditions critical to student and school success. Surveys have the benefit of being cost-efficient and time-efficient, however survey results are subject to bias and should be considered as part of a larger collection of evaluation measures.

Peer-to-Peer Reviews

This strategy allows educators to review, evaluate, and comment on the work of their colleagues using common standards and frameworks. As with observations, it is important to provide training and ongoing calibration, rubrics, and protocols when conducting peer-to-peer reviews in order to minimize bias.

Portfolios

Portfolios are collections of materials compiled by teachers or principals to exhibit evidence of practice, school activities, and student progress. The portfolio process requires educators to reflect on the materials and explain why certain artifacts were included. However, it can be difficult to verify consistency in scoring portfolios and to obtain reliability between scorers.

Local Indicators

To reflect local context, evaluations may also factor in indicators such as student and teacher attendance, graduation/dropout rates, and others. When considering local indicators, it is important that educators are held accountable for only those factors for which there is evidence that they can impact change.

Staff Planning and Development

Recognizing that principals play a major role in teacher quality, school leaders can be evaluated on their ability to attract teachers, develop and grow them, and retain those who are high-performing. Teacher selection instruments, induction and professional development programs, and teacher evaluations that include instructional feedback can be collected to assess school leader competencies. Teacher turnover analyses might also provide insight into which teachers are leaving and why.

The appropriateness of each of these types of indicators for use in an effectiveness system is impacted by many variables that must be considered in the local context, such as the timing and availability of the data; the cost and feasibility of collecting and analyzing the data; and comparability at the school, district, and/or state level. Additionally, it is important to ensure that the information gathered can be used to provide educators with valuable feedback and support in improving their practice. The ultimate goal of an effectiveness system is to strengthen the quality of teachers and school leaders.

In addition to thinking about which indicators to use, it is just as important to consider how those measures will be implemented. For example, in order for observations to be reliable, evaluators need to be trained in how to apply scoring criteria accurately and consistently, and the data collection process needs to be monitored. Keep in mind that ensuring that data are complete and accurate, and that raters are trained and calibrated can be costly. Striking a balance between valid measurement and realistic implementation is critical.

3. DEVELOP A CLEAR COMPOSITE RATING

What weights should each indicator have and who should be involved in the decision?

Once the various indicators to measure educator effectiveness have been identified, a determination needs to be made on how to combine them into a composite measure that represents overall performance. The intent is to create the most valid (accurate) and reliable (consistent) rating of educator effectiveness using all of the available information. Each piece of data is weighted—or given different amounts of emphasis—and then they are combined to form a single composite score.

The key to designing a fair, reliable, and defensible system is minimizing measurement error. While any educator evaluation system is subject to some inaccuracy, the goal is to demonstrate by "professionally acceptable methods" that the evaluation is "predictive or significantly correlated to" educator effectiveness.[6]

The reliability and validity of the composite score depends, in large part, on the reliability and validity of the indicators comprising it. To maximize the reliability and validity of the composite measure, it makes sense to assign heavier weights to indicators that have high reliabilities and that are highly valid. Indicators with low reliabilities and low validity will lower the reliability and validity of the composite score and should, therefore, not be weighted heavily. Complexity arises when there are measures with high validity and relatively low reliability and the reverse. When such cases arise, weighting decisions need to be a balance between measurement and policy considerations.

One suggested approach to finding the right balance in creating a composite score is to engage stakeholders and experts in a modified standard-setting process. In this exercise stakeholders are provided information about each of the indicators, including their validity and reliability, the opportunity to vote for their preferred weighting, and, through rounds of discussion and voting, the opportunity to share their reasoning and learn about the views of other stakeholders.

Strategies for setting up a modified standard-setting system

- Select stakeholder panelists to serve on the committee (e.g., policymakers, educators, researchers, technology experts, human resources experts, community leaders, parents, and students)
- Explain the indicators to the panelists
- Make recommendations of weightings in multiple rounds
- Build consensus through discussion and collaboration
- Provide feedback between each round to provide additional context to the next round of recommendations
- Document the process and the outcomes when making final recommendations to policymakers

4.

CLARIFY DIFFERENTIATED PERFORMANCE LEVELS

What distinguishes varying levels of educator quality?

Current evaluation systems, based largely on classroom observation, fail to identify true variation in educator quality, with the vast majority of educators identified as satisfactory. The addition of multiple indicators and a focus on student outcomes make it much more likely that ratings will clearly distinguish levels of performance.

In order to account for measures of professional practice, a rubric is required that distinguishes performance levels. This rubric should be research-based and define the practices and behaviors of excellent educators, while including differentiated expectations for novice and veteran educators. These standards, known as performance level descriptors (PLDs), must be specific enough to provide useful performance information to educators, and general enough to describe educator practice across a broad spectrum of grade levels and subjects. Moreover, the rubric should show a progression—a roadmap for improvement—from one level to the next.

It is recommended that the rubric include at least four levels to describe differences in educator effectiveness—for example, inadequate, sufficient, good, excellent—in order to be meaningful, provide expectations for good practice, and differentiate educator quality.

A psychometrically sound standard setting method should be implemented to determine cut scores between performance levels, much like when setting performance standards on statewide assessment systems. The standard setting process relies on high quality performance level descriptors to define each level.

Strategies for establishing differentiated performance levels

- Select the standard-setting methodology
- Use PLDs to ensure common understanding of the characteristics of performance at each level
- Select panelists to serve on the committee
- Train panelists to form a key conceptualization of those individuals on the borderline of a performance level
- Use evidence of practice to help participants understand performance expectations at each level
- Make recommendations of cut scores in multiple rounds
- Provide feedback between each round to provide additional context to the next round of recommendations
- Document the results of the process to provide final recommendations to key stakeholders

Since the language of rubrics is open to interpretation and the language defining complex performance levels is necessarily broad, it may be useful to include a set of benchmark cases (actual responses, entries, or evidence of practice) to provide concrete examples or reference points. Benchmarks have many uses, including providing guidance to scorers and helping educators, stakeholders, and the public to understand the distinction between performance levels.

5.

BUILD STRONG DATA ANALYSIS AND REPORTING TOOLS

What does the information reveal about student, educator, and school performance?

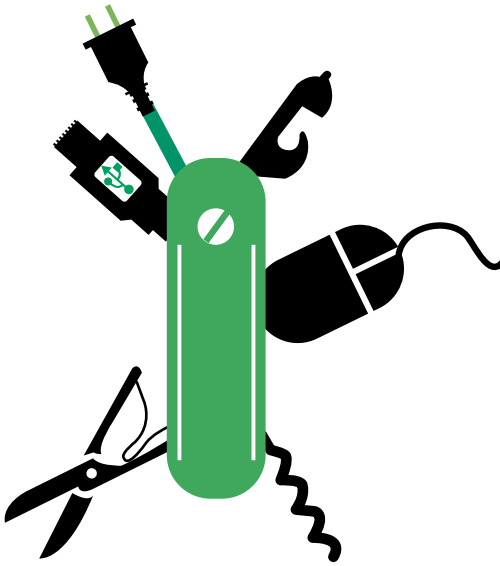
After creating a system of multiple indicators and performance levels that result in a single, composite educator rating, it is important to think about how the data will be used to support and enhance educator development. Data analysis and reporting tools are needed to generate ongoing effectiveness ratings and provide clear feedback to educators on their performance and areas for development. The information also allows state, district, and school leaders to identify trends, pinpoint educators' strengths and weaknesses, and efficiently allocate instructional and professional development resources. These tools serve to empower school leaders and teachers; strengthen practice; drive institutional efficiencies; and ultimately improve student achievement.

Strategies for building data analysis and reporting tools

- Intuitive, role-based platform – To gain teacher and administrator buy-in, the information captured needs to be high-quality and user-friendly. All data must be timely and accurate. Establishing good data systems requires significant investments of time and money in order to ensure that a variety of users (school board, central office staff, principals, teachers, and others) can access relevant data to help assess and improve educator performance.
- Comprehensive but easy-to-understand reports – The platform should provide a variety of focused reports organized by data source or performance measure in the multiple indicators system. Having drill-down functionality makes it easy to disaggregate data by district, building, teacher or principal, and indicator.
- Connection to Professional Development – These tools can support educator development with convenient access to aligned resources for further professional development. Educator evaluation reports, including multi-measure reporting tools and professional growth plans, should be directly aligned to a comprehensive library of professional development resources that can be assigned automatically based on performance or recommended by an evaluator.

6.

IMPROVE INSTRUCTIONAL PRACTICE



How can the information target professional development to boost educator practice, student learning outcomes, and school efficacy?

Educator effectiveness systems provide an individualized assessment of an educator's strengths and weaknesses, and therefore can help educators improve their practice. With a comprehensive view and solid information, districts can align individualized educator professional development plans with school and district goals. The most effective systems enable school and district leaders to gauge return on investment, demonstrating whether professional development activities are building educators' skills and affecting real change in student learning.

Strategies for improving instructional practice

- Ensure educators have the tools to understand and draw conclusions based on the data
- Align professional development activities to rubrics to help teachers and school leaders move from one performance level to the next
- Provide opportunities for collaboration amongst colleagues by facilitating data interpretation and discussion sessions in which educators can see their practice and outcomes relative to others and then have discussions about practices being used in their classrooms
- Identify educators to become mentors and leaders based on high performance
- Evaluate the efficacy of professional development programs, funding practices, etc.

PUT IT ALL TOGETHER

How do those responsible for creating educator effectiveness systems build capacity for continuous improvement?

Developing an educator effectiveness system that integrates accountability and evaluation to promote continuous improvement requires an inclusive, data-driven approach that embraces ongoing communication. As such, there are three critical activities that states and districts should consider as they plan, implement, and continually evaluate their educator effectiveness systems.

Engage Stakeholders

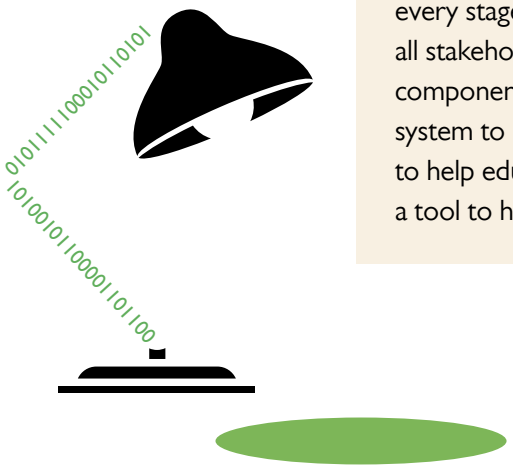
If a central purpose of the new evaluation system is to provide valid and reliable feedback to educators with a goal of improving their performance and student learning, then the evaluation process itself necessitates engagement of and input from all affected stakeholders. Continuous input from teachers, students, parents, administrators, school board members, policymakers, and others will help develop a system that is fair and valid. These stakeholders need to remain engaged after the design phase and throughout implementation to ensure a cycle of continuous improvement.

Gather Data Continuously for the Purposes of Continuous Improvement

Data and information need to guide all development stages of the new educator effectiveness system. Through the collection and interpretation of data on each aspect of the system as well as on the system as a whole, the state or district can make informed, evidence-based decisions. Furthermore, data collected on the implementation of the system and ways in which educators interpret results from the system allows states and districts the opportunity to continuously improve the system so it best meets the needs of the educators, students, districts, and states.

Communicate Early and Often

Changes of this scale that so directly touch the lives of educators, students, and parents are challenging under even the best of circumstances. Communication at every stage of the process is critical to success. It cannot be overemphasized that all stakeholders must be given advance notice of the evaluation system's process, components, criteria, and how the evaluation results will be used. For this new system to be effective, it will take a massive cultural shift in beliefs and behaviors to help educators understand how the evaluation process can and will be used as a tool to help them improve their practice on behalf of students.



[1] See *The Standards for Educational and Psychological Testing* for technical principles and requirements for developing high stakes assessment systems. Codeveloped by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999.

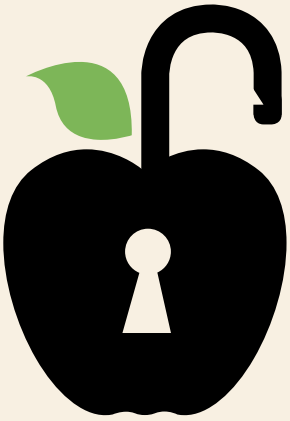
[2] Laura Goe, Courtney Bell, and Olivia Little, *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. National Comprehensive Center for Teacher Quality, 2008; *State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies*, National Council on Teacher Quality, October 2011; and Linda Darling-Hammond, *Evaluating Teacher Effectiveness*, Center for American Progress, October 2010.

[3] Stephen Davis, Linda Darling-Hammond, Michelle LaPointe, and Debra Meyerson, *School Leadership Study: Developing Successful Principals*, Stanford Educational Leadership Institute, 2005; Robert J. Marzano, Tim Waters, and Brian McNulty, *School Leadership That Works: From Research to Results*, Association for Supervision and Curriculum, 2005; *Evaluating Principals: Balancing Accountability with Professional Growth*, New Leaders for New Schools, 2010.

[4] Standards to consider include the Interstate Teacher Assessment and Support Consortium (InTASC), National Board for Professional Teaching Standards (NBPTS), the Interstate School Leaders Licensure Consortium (ISSLC) Standards for School Leaders, and state teaching and leadership standards.

[5] Cynthia D. Prince, Vanderbilt University; Julia Koppich, Ph.D., J. Koppich and Associates; Tamara Morse Azar, Westat; Monica Bhatt, Learning Point Associates; and Peter J. Witham, Vanderbilt University. *Research Synthesis: What Does Research Suggest About Ways to Measure Teacher Effectiveness so that Determination of Performance-Based Rewards is Accurate, Reliable, and Defensible?* Center for Educator Compensation Reform, 2010.

[6] Albemarle Paper Co. v. Moody, 422 U.S. 405, 431 (1975). See also Association of Mexican-American Educators v. California, 937 F. Supp. 1397, 1419-20 (N.D. Cali. 1996) ("Defendants' validity evidence is not perfect. But it is not required to be, under either legal or scientific standards.")



To learn more about the issues discussed here or to get information about Pearson's educator effectiveness initiatives, please contact:

Kelly Burling, Ph.D.

Director, Center for Educator Effectiveness
919-627-8893
kelly.burling@pearson.com

http://educatoreffectiveness.pearsonassessments.com