Great Public Schools for Every Student

VISITING SCHOLARS SERIES

Spring 2010, volume 1

Research Brief

Teacher Evaluation Measures and Systems: How They Can Improve Teaching and Learning

Laura Goe
Educational Testing Service

Olivia Little
University of Wisconsin-Madison

It is not news to most that teachers matter to student learning. However, recent research has quantified the extent to which teachers matter, revealing that teachers may be the most important school-based factor in increasing student achievement.[1-3] It follows that improving teacher effectiveness may be the key to closing persistent achievement gaps that plague our education system. Understanding exactly what makes teachers effective is essential to this mission. Researchers have found scant evidence that teachers' particular qualifications, characteristics, or practices are strong predictors of student achievement.[4] While we know that some teachers are much more effective than others, we still do not know what combination of traits, qualifications, and practices are responsible for the huge variation in teacher effectiveness. Better understanding of what makes teachers effective will improve teacher selection and enable development of preparation programs and professional development offerings that contribute to more effective teaching practices and improved student learning.

The fundamental goal of teacher evaluation should be to improve teaching and learning. That is why the measures used and the evaluation systems themselves are so important. Evaluation systems that use multiple measures to examine both teaching and learning can be described as comprehensive teacher evaluation systems. Rather than the typical model of teacher evaluation that focuses almost exclusively on classroom-based teaching practice, comprehensive teacher evaluation systems focus on *what teachers do in classrooms, what teachers contribute in schools, and how well students learn.* Well-designed evaluation measures within a comprehensive teacher evaluation system can provide valuable feedback to teachers and administrators, identifying strengths and weaknesses and focusing attention on strategies for improving practice. What is the most efficient way to evaluate teacher effectiveness? While many measures, protocols, and proxies have been developed, no one measure has proven to be a fully accurate representation of the highly complex task of teaching. As Schacter concludes, "The bottom line is that simple, efficient, and cost-effective solutions to measure teacher quality have not been fruitful."[5]

This brief will discuss research findings and considerations for designing and implementing meaningful teacher evaluation. The first section will discuss key components of successful teacher evaluation systems. The second will look more specifically at particular categories of evaluation instruments that can be utilized within a system, and discuss what to consider when choosing instruments. The third will offer information about how districts and states are currently implementing comprehensive teacher evaluation systems and summarize the conceptual model, *i.e.,* how teacher evaluation can improve teaching and learning.

Research on Teacher Evaluation Systems

A recent review written for the NEA[6] examined five comprehensive teacher evaluation systems: the Teacher Advancement Program[7] (TAP); the Framework for Teaching[8] (FFT); the Professional Compensation System[9-10] in Denver (ProComp); Peer Assistance and Review[11-13] (PAR) in Toledo and other cities; and the Beginning Educator Support and Training Program[14] in Connecticut (BEST). This brief describes research on the effectiveness of each and highlights their common elements; each system shows promise as a model for designing a comprehensive evaluation approach. These systems have been consistently mentioned in research studies and policy reports as innovative approaches to reforming teacher evaluation, and they are established enough to have a research base. We focus primarily on the five programs listed above because they have been around long enough to accumulate considerable research evidence and recognition in the research literature. These five systems share a few features that research suggests are essential for establishing a credible and meaningful evaluation system. A successful evaluation system should: create linked and integrated systems which tie evaluation procedures to curricular standards, professional development activities, targeted support, and human capital decisions; use validated evaluation measures based on widely accepted standards of teaching, that attempt to capture a range of teaching behaviors and use multiple methods and evaluators; and establish credibility by involving multiple stakeholders in the design, development, and revision of the system and making procedures meaningful and transparent to all involved.

The first feature—*creating linked and integrated systems*—stresses the importance of aligning evaluation practices with other elements of the education system. Alignment is gaining attention as a crucial part of a successful system, something previous attempts to reform teacher evaluation and pay failed to incorporate.[15-16] An excellent example of alignment is TAP,[7, 16] which integrates four key elements: multiple career paths; ongoing applied professional growth; instructionally focused accountability; and performance-based compensation. In this system, accomplished teachers can assume leadership positions in which they deliver professional development, conduct teacher evaluations, and participate in other decision-making activities. Evaluations consist of well-established, evidence-based observation protocols that require intensive training and ongoing certification of evaluators. Job-embedded professional development is centered around evaluation standards, and teachers are compensated based on a combination of their evaluation outcomes, individual student achievement gains, and school-wide achievement gains.

The second feature—*using validated evaluation measures*—is crucial to ensuring that evaluations are credible, meaningful, and useful for providing information that will improve teaching. BEST, Connecticut's evaluation and support system for beginning teachers (now called TEAM), uses a portfolio assessment, one example of a well-developed protocol for examining a range of teaching behaviors, and multiple evaluators. BEST standards are based on years of research and development and are consistent with the well-established Interstate New Teacher Assessment and Support Consortium (INTASC) standards. Portfolios contain a variety of materials including lesson plans, videos of teaching, samples of student work, and out-of-class achievements. Portfolios are rated by three different evaluators who are experienced teachers in the same discipline as the teacher being evaluated. The evaluation is preceded by two years of mentoring and targeted support to prepare teachers for what is expected on the assessment.[14]

The third feature—*establishing credibility*—involves including administrators, teachers, teachers' unions, and other stakeholders in the design, development, and revision of a system. Gaining trust and buy-in from teachers in particular is essential for the system to effectively improve teaching. TAP will not be implemented in a school without agreement and buy-in from teachers. Denver's ProComp and the PAR program in Toledo are examples of innovative district-union partnerships that allowed for substantial reforms in teacher evaluation practices.

Each of the systems mentioned above has been researched to differing degrees, with FFT and TAP having the strongest evidence base. Both have shown positive relationships to student achievement, with some indication that the relationship is stronger when evaluation is more faithfully implemented.[15, 17] Rigorous evaluations of ProComp, PAR, and BEST have not yet been completed, but each shows evidence of success in either pilot and development work or preliminary analyses. Teachers and administrators tend to evaluate these systems favorably, and teachers report an increased sense of collegiality. Even those systems incorporating student achievement data and differentiated pay, such as TAP and ProComp, report high levels of collegiality among teachers, though it should be noted that surveys of teachers suggest that differentiated compensation is the least popular TAP component.[7]

Research on Teacher Evaluation Instruments

One essential element of a successful teacher evaluation system is the use of a well-established, validated evaluation protocol. Research can point to certain instruments that have been rigorously vetted, tested, and shown to measure teaching practice reliably; but research often reveals that validity evidence is sorely lacking for commonly used instruments.[18] It is important to note that instruments are not valid in and of themselves; validity depends on whether an instrument is implemented and interpreted as intended.[19] Establishing the validity of an evaluation protocol requires careful examination of several considerations—

Scoring and Training

- Is the instrument's scoring rubric based in research and consistent with widely-accepted standards of high quality teaching?
- Is the rubric specific and detailed enough to allow for reliable judgments of teaching practice?
- Are there procedures in place to ensure that raters can score reliably and consistently with "master" raters, other raters, and themselves over time?
- Are raters trained thoroughly and regularly to maintain reliability and reduce scoring bias?
- Are there guidelines for selecting raters who would provide the most accurate scores (*e.g.,* raters with content knowledge on a subject-specific protocol)?

Generalizability and Extrapolation

- Has research established an appropriate number of lessons or materials to be sampled in order to get a complete and stable picture of teaching practice?
- Do the instrument scores correlate with other accepted measures of teaching quality and with student outcomes of interest?
- Is the instrument practical and feasible for raters to implement?

Implication and Interpretation

- Is the instrument being used for the same purposes for which it was designed?
- Does the instrument capture what it is intended to, or is it biased by factors unrelated to teaching?
- Do the interpretations being drawn from the scores go beyond what the instrument is actually able to measure?

These are some of the essential questions that should be examined before employing any evaluation protocol. A more detailed discussion can be found in Goe *et al.*[20] and Bell *et al.*[18] We synthesized research on the reliability and validity of teacher evaluation protocols, focusing on direct measures of teaching practice such as observations, portfolio assessments, and analysis of instructional artifacts. Each category of instruments has its benefits and weaknesses, and some key findings are summarized below. For fuller descriptions of categories and examples of promising instruments, see the review papers.[18, 20, 21]

Most research has been conducted on observation protocols. Advantages of structured observations are that they tend to be viewed as credible and authentic by teachers and other stakeholders, and can provide rich information about teaching practices that can be used for both formative and summative purposes. Some have been shown to relate to other accepted measures of teaching and to gains in student achievement.[15, 22] However, the quality of this relationship will rely heavily on the training of raters and establishing scoring reliability. Portfolio assessments have the benefit of providing comprehensive and holistic information about teaching performance and can be useful for reflecting on teaching practices. However, establishing accurate and reliable scoring procedures can be cumbersome, and portfolios have not been conclusively linked to student achievement or to changes in teaching practice.[18] Analysis of instructional artifacts is an emerging approach to teacher evaluation, providing information about the quality, rigor, and relevance of teacher assignments and student work created within the classroom. Instruments being developed in this area show promise,[23-26] but are not yet well established in practice.

One caution when implementing instruments is that they all rely on the judgments of raters. Thus, it is extremely important to pay careful attention to the selection, training, and recalibration (retraining over time) of raters to ensure that evaluations are accurate and representative. The validity of a perfectly well-designed and rigorously tested instrument is threatened when placed in the hands of an untrained rater. Unfortunately, the research indicates a lack of attention to rater training and reliability of scoring, particularly when it comes to maintaining reliability over time and purposefully controlling for bias.

The Shifting Landscape of Teacher Evaluation

Teacher evaluation is undergoing substantial change due primarily to an increased emphasis on teacher effectiveness rather than teacher quality. Teacher quality is based primarily on teacher qualifications, which are poor predictors of who will be a good teacher. In contrast, teacher effectiveness focuses on student outcomes, such as student achievement growth, without consideration of teacher qualifications or characteristics. Federal funding tied to the use of student achievement growth as a measure of teacher performance has resulted in rapid changes in how teachers are evaluated, and focused attention on evaluating individual teachers' contributions to student achievement growth. In response to both incentives and pressures, innovative models are being developed around the country that incorporate student achievement growth and may also include stronger provisions for tying teacher evaluation results to professional growth opportunities for teachers. However, these models haven been so recently developed that little research exists on how such comprehensive models impact teaching and learning.

An example of one of the new comprehensive teacher evaluation systems is Washington, DC's recently implemented IMPACT system for teacher evaluation. IMPACT categorizes teachers for evaluation purposes, grouping teachers for whom value-added models (based on predicted student text scores) can be used and teachers for whom such models cannot be used because they teach in untested subjects, lower elementary, etc. For teachers for whom value-added cannot be used, evaluation scores are based primarily on teachers' performance in the classroom as measured by observations, with student learning growth as measured by classroom-based assessments making up only a small percentage of the score. They will be including differentiated compensation tied to teacher performance in the future.

Georgia has implemented a statewide plan teacher evaluation instrument called CLASS Keys which is notable for evaluating not only what teachers do in their classrooms but also their professional responsibilities such as promoting the active and sustained involvement of students, families, and the community in order to reinforce the continuous improvement of all students.[27] Georgia also evaluates teachers' contributions to student learning through both standardized test scores and other district-approved measures such as teacher-developed assessments, department or district common assessments, benchmark tests, student work samples, and portfolios. Like Washington, DC's IMPACT, Georgia's system is notable for its transparency and availability of documents describing what is being evaluated, the criteria used, and the evaluation process.

With support from AFT Innovation Grants, a number of districts in New York and Rhode Island are convening teachers, administrators, district superintendents, and union leaders to work together to develop comprehensive teacher evaluation systems that include evidence of student learning growth. This effort requires a radically different way of thinking about teacher evaluation. Teachers are accustomed to being evaluated solely on inputs—what they are doing in the classroom when the evaluator (usually the principal) comes by. Comprehensive teacher evaluation systems require the inclusion of outputs—what students are actually learning in a given teacher's classroom. Teachers are willing to be held accountable for their own performance, but question how much of students' performance is the teacher's responsibility, and how much is the responsibility of the students themselves, along with parents, prior teachers, and communities. Unfortunately, the research provides little guidance in this area.

More examples of comprehensive teacher evaluation systems will crop up in the near future as states and districts expand or refine their current systems to include more than just a walk-through by the principal. It is too early to tell whether these changes to business-as-usual teacher evaluation will result in improved teaching and learning. It is certain that not all changes will be popular with teachers; in particular, using student achievement growth as a measure of teacher effectiveness is not generally embraced. However, information about student learning growth linked to individual teachers may provide valuable evidence than can help teachers and administrators identify weaknesses in curriculum, materials, and instructional practices. Triangulating student learning growth results with other evidence, such as results from classroom observations and an examination of classroom artifacts such as lesson plans, assignments, and student work, should yield powerful information that can be used to improve student learning.

Using Evaluation to Improve Teaching and Learning

The complex nature of teaching and learning suggest that evaluation tools and systems must also be complex. It is not enough for the principal to drop by a teacher's classroom, write some notes and check off some boxes, and expect teaching to improve. Nor is one-size-fits-all professional development offerings likely to improve teaching and learning. And using student test scores as a primary measure of teacher effectiveness is unlikely to result in better teaching and improved student learning. A comprehensive teacher evaluation system gathers evidence on teaching and learning using a varied set of measures and indicators. The logical next step is to use that evidence to set goals, develop strategies, and identify resources and opportunities that will enable teachers to grow professionally. Comprehensive teacher evaluation systems are not just about assessing teacher effectiveness; they are about providing specific guidance and support to ensure that all teachers reach their full potential. Better learning outcomes for students are sure to follow.

References

1. Darling-Hammond, L. 2000. Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives* 8(1).

2. Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417.
3. Wright, S. P., S. P. Horn, and W. L. Sanders. 1997. "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation." *Journal of Personnel Evaluation in Education* 11: 57.
4. Goe, L. 2007. *The Link between Teacher Quality and Student Outcomes: A Research Synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality:
5. Schacter, J. 2001. *Teacher Performance-based Accountability: Why, What and How.* Santa Monica, CA: Milken Family Foundation.
6. Little, O. M. 2009. *Teacher Evaluation Systems: The Window for Opportunity and Reform.* Washington, DC: National Education Association.
7. Solmon, L. C., *et al.* 2007. *The Effectiveness of the Teacher Advancement Program.* Santa Monica, CA: National Institute for Excellence in Teaching.
8. Danielson, C. 1996. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
9. Community Training and Assistance Center. 2004. *Catalyst for Change: Pay for Performance in Denver (final report).* Boston, MA: Author.
10. Azordegan, J., *et al.* 2005. *Diversifying Teacher Compensation.* Denver, CO: Education Commission of the States.
11. Escamilla, P., T. Clarke, and D. Linn. 2000. *Exploring Teacher Peer Review.* Washington, DC: National Governors Association Center for Best Practices.
12. Goldstein, J. 2007. "Easy to Dance to: Solving the Problems of Teacher Evaluation with Peer Assistance and Review." *American Journal of Education* 113(3): 479.
13. Toledo Federation of Teachers. 2009. "The Toledo Plan." Retrieved 10/2/09 from www.tft250.org/the_toledo_plan.htm.
14. Connecticut State Department of Education. 2009. A Guide to the BEST Program for Beginning Teachers 2008–2009. Hartford, CT: Author.
15. Heneman, H. G., *et al.* 2006. *Standards-based Teacher Evaluation as a Foundation for Knowledge- and Skill-based Pay*. Philadelphia, PA: Consortium for Policy Research in Education.
16. Jerald, C. 2009. *Aligned by Design: How Teacher Compensation Reform Can Support and Reinforce other Educational Reforms.* Washington, DC: Center for American Progress.
17. National Institute for Excellence in Teaching. 2010. *TAP Research Summary.* Washington, DC: Author.
18. Bell, C. A., *et al.* 2009. *Measuring Teaching Practice: A Conceptual Review.* San Diego, CA: American Educational Research Association.
19. Kane, M. T. 2006. "Validation," in R. L. Brennan (ed.), *Educational Measurement.*New York, NY: Praeger.
20. Goe, L., C. Bell, and O. M. Little. 2008. *Approaches to Evaluating Teacher Effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality.
21. Little, O. M., L. Goe, and C. Bell. 2009. *A Practical Guide to Evaluating Teacher Effectiveness.*

Washington, DC: National Comprehensive Center for Teacher Quality.

22. Pianta, R. C., K. M. La Paro, and B.K. Hamre. 2007. *Classroom Assessment Scoring System*. Baltimore, MD: Brookes Publishing.
23. Junker, B., *et al.* 2006. *Overview of the Instructional Quality Assessment.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
24. Matsumura, L. C., *et al.* 2006. *Measuring Reading Comprehension and Mathematics Instruction in Urban Middle Schools: A Pilot Study of the Instructional Quality Assessment.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
25. Newmann, F. M., A. S. Bryk, and J. K. Nagaoka. 2001. *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Chicago, IL: Consortium on Chicago School Research.
26. Newmann, F. M., G. Lopez, and A. S. Bryk. 1998. *The Quality of Intellectual Work in Chicago Schools: A baseline Report.* Chicago, IL: Consortium on Chicago School Research.
27. Georgia Department of Education. 2009. "CLASS Keys: Classroom Analysis of State Standards." Retrieved 3/16/2010 from http://www.doe.k12.ga.us/tss_teacher.aspx.

About Laura Goe

Laura Goe, Ph.D., is a Research Scientist at Educational Testing Service in Princeton, New Jersey. She received her doctorate from UC–Berkeley where she was the Research Director for the Bay Area Consortium for Urban Education, working to improve teacher quality and supply in urban schools. Prior to Berkeley, Laura taught at-risk middle school students in Mississippi and Tennessee. She is the Principal Investigator for Research and Dissemination for the federally-funded National Comprehensive Center for Teacher Quality, and she recently finished a term as co-editor of the AERA journal, *Educational Evaluation and Policy Analysis*. Laura's research interests include teacher qualifications, measuring teacher quality, and teacher effectiveness.

About Olivia Little

Olivia Little is a former employee of the Educational Testing Service, where she worked in the Learning and Teaching Research Center on issues of measuring teacher effectiveness, equitable distribution of teachers, and teacher professional development. She also served as a staff member for the National Comprehensive Center for Teacher Quality, a national resource that assists states in strengthening quality teaching. At the Center she co-authored *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis* and the accompanying policy brief, *A Practical Guide to Evaluating Teacher Effectiveness.* She is currently a graduate student at the University of Wisconsin–Madison pursuing a Ph.D. in Human Development and Family Studies, with primary interests in anti-poverty programming and family policy. She continues to collaborate with ETS colleagues in writing about the evaluation of teaching.

About the Visiting Scholars Series

NEA Research hosts the Visiting Scholars Series as a forum intended to help link policy initiatives with educational scholarship. Prominent scholars are asked to link their research to recommendations for closing achievement gaps. All Research Briefs are available on *InsideNEA* at http://insidenea.nea.org/NEABiz/ResearchInfo/Pages/VisitingScholars.aspx. For more information about the Visiting Scholars Series, contact:

Judith McQuaide
NEA Research, ext. 7375
jmcquaide@nea.org

Gwen Williams
NEA Research, ext. 7368
gwilliams@nea.org