

Teacher Evaluation Systems

The Window for Opportunity and Reform

Olivia Little

National Education Association
Research Department
Ronald D. Henderson, Director



*Great Public Schools
for Every Student*

Teacher Evaluation Systems

The Window for Opportunity and Reform

Olivia Little

National Education Association
Research Department
Ronald D. Henderson, Director



The National Education Association is the nation's largest professional employee organization, representing 3.2 million elementary and secondary teachers, higher education faculty, education support professionals, school administrators, retired educators, and students preparing to become teachers.

Copies of this publication may be purchased from the NEA Professional Library Distribution Center, P.O. Box 404846, Atlanta, GA, 30384-4846. Telephone 1-800-229-4200 for price information or go to the NEA Professional Library Web site at www.nea.org/books.

Reproduction: No part of this publication may be reproduced in any form without permission from NEA Research, except by NEA-affiliated associations and NEA members. Any reproduction of this material must contain the usual credit line and copyright notice. Address communications to Editor, NEA Research, 1201 16th Street, NW, Washington, DC 20036-3290.

Copyright © 2009 by the National Education Association
All Rights Reserved

Issues involving teacher evaluation systems are at the forefront of American education policy. At the National Education Association, we believe any discussion of teacher evaluation systems rightfully begins by asking:

- What do we know from prior research and practice about teacher evaluation systems, especially as they relate to student achievement and narrowing achievement gaps?
- What might an ideal system look like?
- Are there ways to examine what we have learned that will enable us to apply those lessons in a manner that supports student and teacher learning?

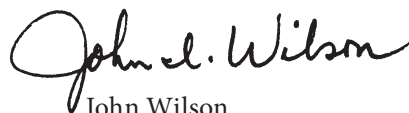
To that end, NEA commissioned a review of the research literature on teacher evaluation systems, particularly the way in which such systems serve to improve student achievement and narrow achievement gaps. This paper provides a basis for discussing how to design and implement teacher evaluation systems to meet those targets. It offers alternative ways of thinking about evaluation that might move us closer to a link between evaluation and student learning.

We hope this review is useful for revisiting ideas and generating new thoughts about the relationship between teacher evaluation and student learning. And we hope that our efforts in this regard will help us ensure a great public school for every student.

Sincerely,



Dennis Van Roekel
President



John Wilson
Executive Director

Contents

Executive Summary	vii
Introduction	xi
Structure of the Review.....	xii
Methods.....	xiii
Current Teacher Evaluation Systems	1
Teacher Advancement Program (TAP)	1
Framework for Teaching (FFT).....	4
Professional Compensation System (ProComp).....	6
Peer Assistance and Review (PAR)	7
Beginning Educator Support and Training Program (BEST).....	8
Implementing Comprehensive Evaluation Reform	11
Common Components of Successful Systems.....	11
Using Student Achievement Data	13
Linking Evaluation to Pay	16
Implementation Considerations	17
Conclusions	19
Implications for the Union.....	19
Opportunity for Change.....	20
Appendix	21
Summary of Evaluation Systems.....	22
References	25
About Olivia Little	30

Executive Summary

With research reinforcing the fact that teaching quality is key to improving education outcomes in this country, teacher evaluation has become a hot topic. Current evaluation practices do not adequately measure teaching effectiveness, and they remain disconnected from efforts to improve teaching. This has led to a system in which over 90 percent of teachers are classified as top performers and only a tiny percentage are deemed unsatisfactory, a system which allows underperforming teachers to remain in the workforce while failing to acknowledge and reward teachers who exhibit excellence. Recent political attention to these issues and subsequent calls for reform may present an unprecedented opportunity to introduce comprehensive change to our current system.

Based on a review of the research literature, this paper describes five current teacher evaluation systems that have been recognized as innovative and comprehensive approaches to evaluation reform:

- Teacher Advancement Program (TAP),
- Framework for Teaching (FFT),
- Professional Compensation System (ProComp),
- Peer Assistance and Review (PAR), and
- Beginning Educator Support and Training Program (BEST).

Research is reviewed on the effectiveness of each system, how it relates to student outcomes, and how it is received by teachers and administrators. Program features are summarized in the Appendix beginning on page 21.

Common Components of Successful Systems

These five programs are recognized as promising approaches to improving instruction, raising student achievement, gaining teacher support, increasing retention by taking a comprehensive rather than piecemeal approach to reform, and centering activities and procedures around instructional improvement and student learning. To be successful, approaches to evaluation reform should—

Establish a credible and meaningful evaluation system. Reform doesn't work if the people involved do not believe in it or worry it will be implemented unjustly.

- Involve multiple stakeholders in the development and revision of the system, and make sure the system is meaningful to everyone involved.

- Use validated, credible evaluation measures and ensure that they are faithfully implemented.
- Use multiple measures that evaluate multiple facets of what contributes to good teaching.

Create linked and integrated systems. Tie evaluation procedures to curricular standards, professional development activities, targeted support, and human capital decisions.

- Include embedded, ongoing professional development.
- Incorporate opportunities for career advancement within teaching.

Implementing Comprehensive Evaluation Reform

Calls for teacher evaluation reform include several elements that have remained controversial, most notably including student achievement data in teacher evaluation scores and linking evaluation to teacher pay. This review discusses how to incorporate these elements into a credible evaluation system and also addresses particular implementation considerations.

Using Student Achievement Data

- Involve teachers in deciding how to account for student learning and other relevant outcomes in evaluation, using a combination of measures so teachers feel they are being evaluated comprehensively and fairly.
- When using value-added measures, work to create the necessary data structures, collect complete data, use an appropriate standardized test, and employ an accepted model for the calculations. Consider the many limitations of value-added data to provide certain kinds of information, and supplement your evaluation with other methods.
- Consider other methods of assessing student achievement, such as the analysis of teacher assignments and student work or locally designed assessments aligned with curriculum.
- Think creatively about ways to assess multiple aspects and outcomes of teaching and learning, such as motivation, engagement, civic-mindedness, and social development.

Linking Evaluation to Pay

- Research indicates that standards-based evaluations can serve as a valid and reliable basis for a performance pay system, if they are supported by teachers and implemented well. An evaluation system should be established before the link to pay is made.
- Evaluation systems should be multifaceted and aligned with professional development and other school improvement efforts so that performance incentives do not result in a narrowing of teaching practices.

- Teaching performance should be just one element in a differentiated compensation system. Pay for other contributions—such as filling hard-to-staff positions, taking on additional responsibilities and leadership roles, and attaining relevant skills and certifications—should also be incorporated.

Implementation Considerations

- Ensure that evaluation is credible and useful by using accepted standards and a valid instrument, thoroughly training and calibrating raters, providing feedback and targeted support, and promoting transparency and communication between those doing the evaluating and those being evaluated.
- Leadership should demonstrate a firm commitment to reform through actions that allow teachers to meet evaluation requirements, and districts should consider implementing systems to hold leadership accountable.
- Allow sufficient time and resources to thoughtfully design and implement the system, expecting to encounter complications and refine the system along the way. Ensure there is adequate funding available.

Promising Models

It is this reviewer's conclusion that the time is right to rally for comprehensive evaluation reform, and that the systems presented here represent promising models on which to base evaluation systems. Education unions such as the NEA have an important voice in this process, with the responsibility to protect members from unfair evaluation practices while promoting improvements to the education system. In order to create and maintain successful evaluation reforms, unions and districts should approach the task in the spirit of sustained collaboration and compromise.

Introduction

There is widespread agreement among researchers and policymakers that teachers matter significantly in improving student learning. Because high-quality teaching may be the most important school-based factor in increasing student achievement (Darling-Hammond 2000, Rivkin, Hanushek, and Kain 2005, Wright, Horn, and Sanders 1997), measuring teaching quality has become a hot topic in the literature. Studies using value-added methodology—a statistical procedure for calculating teacher contributions to student gains on standardized achievement tests—have revealed that teachers vary widely in their effectiveness (Gordon, Kane, and Staiger 2006, Wright *et al.* 1997). However, these and follow-up studies have not yet uncovered exactly which combination of traits, qualifications, or practices are associated with these effective teachers. Even with value-added measures, then, it is necessary to measure multiple aspects of teaching in order to fuel further improvements in instruction (Bell *et al.* 2009). Researcher Mary Kennedy notes that there are many teacher qualities and practices that we care about and assess, but “what we lack is a strategy for organizing our assessments into a coherent system” (Kennedy 2008). The challenges lie in measurement—both what is most important to measure and how to measure it—as well as organization into a comprehensive, multifaceted system.

Examples abound demonstrating a need for change in current approaches to teacher evaluation. For instance, evaluation practices are typically locally determined and vary widely across districts. They most commonly consist of an observation by a principal or assistant principal, and most district evaluation policies provide little guidance on how often to observe, what criteria to follow, and how to use and share feedback from the process. Even less attention is typically paid to systematically training or calibrating administrators to ensure reliability and reduce bias in scoring (Brandt *et al.* 2007). The inadequacy of this approach is apparent, as evidenced by studies showing that over 90 percent of teachers are classified as top performers and only a tiny percentage are denied tenure or dismissed due to evaluation results. This trend prevails even in schools with dismal student achievement scores (Donaldson 2009, Weisberg *et al.* 2009).

Koppich writes that “Evaluation has two classic functions: improvement and accountability. Good evaluation is a continuation of good professional development” (Koppich 2008). However, current evaluation practices often lack alignment with curricular standards and professional development efforts and do not result in targeted instructional support (Heneman *et al.* 2006, Jerald 2009). Furthermore, the teacher salary

The challenges lie in measurement: both what is most important to measure and how to measure it.

schedule has remained the same for decades, with advancement and compensation based on experience and seniority—a system that essentially treats teachers as interchangeable parts. Such a system allows underperforming teachers to remain in the workforce while failing to acknowledge and reward teachers who exhibit excellence (Weisberg *et al.* 2009). In a system where almost everyone is rated as superior, constructive feedback is not provided, and opportunities for career advancement or performance-based promotion are extremely limited. There is little incentive for instructional improvement. Evaluation becomes a cursory procedure that has no significant effect on teaching practice. This often leads to a school culture in which neither teachers nor administrators take the evaluation process seriously, superior ratings are expected, and a less than superior rating is considered a personal affront rather than an opportunity for improvement (Donaldson 2009).

These existing issues, coupled with the growing acknowledgement of the importance of teacher effectiveness in an increasingly knowledge- and technology-based society, have spurred calls for reform and accountability. Some of these calls for reform, including changes in the traditional teacher compensation system, are not new. Efforts in the 1970s and '80s to reform teaching through pay incentives and teacher career ladders were largely unsuccessful. However, advances in research since then have provided a solid evidence base for classroom and school practices that contribute to student learning. Testing requirements, such as those implemented by the No Child Left Behind (NCLB) Act, have led to the creation of data systems in each state that house standardized data on student achievement. More rigorous evaluation instruments have been developed, and the importance of aligning reforms rather than implementing them piecemeal has been gaining attention (Heneman *et al.* 2006, Podgursky and Springer 2007). The Obama administration's current focus on education reform as a top agenda item, along with the availability of funds from the American Recovery and Reinvestment Act (ARRA) for teacher quality improvement efforts, may present an unprecedented opportunity to introduce comprehensive change to our current system (Brodie 2009, Donaldson 2009).

Structure of this Review

In its first section, this paper discusses five current teacher evaluation systems that have been recognized in the research literature as innovative and comprehensive approaches to evaluation reform: the Teacher Advancement Program (TAP), the Framework for Teaching (FFT), the Professional Compensation System (ProComp), Peer Assistance and Review (PAR), and the Beginning Educator Support and Training Program (BEST). This paper will describe their elements and examine what the research literature has to say about their effectiveness, considering how the systems relate to student outcomes and how they are received by teachers and administrators. (Program features are summarized in the Appendix beginning on page 21.)

The paper's second section discusses issues related to the implementation of comprehensive evaluation reform. It presents the common components these five programs share that lend themselves to successfully designing a comprehensive evaluation system. It considers the issues of including student achievement data (specifically, value-added measures) in evaluation systems, of linking evaluations to performance pay, and of implementing a system with fidelity. The paper's conclusion discusses overall findings and implications for education unions as they engage in the teacher evaluation systems reform conversation.

Methodology

Literature for this review was obtained through database and Internet searches, expert recommendations, and article reference lists. Recent peer-reviewed journal articles were identified through searching the Educational Resources Information Center (ERIC) database using the subject terms “teacher evaluation,” “teacher effectiveness,” “instructional effectiveness,” and “instructional improvement,” as well as searching specifically for the names of the systems reviewed. Research and policy reports were identified through expert recommendation, reference lists found in other articles, and Web sites of prominent research and policy organizations involved in issues of teacher quality and educator compensation reform (*e.g.*, National Comprehensive Center for Teacher Quality, National Center for Educator Compensation Reform, Center for American Progress, RAND, New Teacher Project, National Governors Association, etc.).

Reviewed articles were those relating specifically to current United States education policy that address ongoing classroom evaluation for inservice teachers. Thus, this review does not consider measures based only on teacher qualifications, such as experience, certification, or knowledge. Nor does it examine evaluations of pre-service teachers. It is recognized that the knowledge and skills a teacher brings to teaching are crucial components of teacher quality, and that teacher education and recruitment policies are an important policy target. However, this review addresses how to identify and support those teachers who are already in the workforce. In order to sufficiently improve the quality of the nation’s teaching force, strategies for both recruiting higher quality teachers and improving the quality of existing teachers are necessary. This review focuses on the latter. Furthermore, this review considers evaluation systems and instruments currently being applied in states and districts, and does not include instruments that may exist in the research literature but which have only been used for research purposes.

Current Teacher Evaluation Systems

There are as many different evaluation systems as there are states and districts. Thus, it is difficult to get a handle on what is most popularly being used, how structured the systems are, and how they are being carried out. Through examining numerous research studies, reports, briefs, and news articles on teacher evaluation, I found that certain systems are consistently mentioned. These are the Teacher Advancement Program, The Framework for Teaching, the Professional Compensation System in Denver, Peer Assistance and Review in Toledo and other cities, and the Beginning Educator Support and Training Program in Connecticut. I will discuss each of these programs in turn, describing their elements and any evidence about their implementation and effectiveness.

Teacher Advancement Program

The Teacher Advancement Program (TAP), created by education reformer Lowell Milken of the Milken Family Foundation, is an innovative program that works to improve the recruitment and retention of talented teachers by restructuring the evaluation and rewards system within schools. It is a comprehensive approach based on four key elements: 1) multiple career paths, 2) ongoing applied professional growth, 3) instructionally focused accountability, and 4) performance-based compensation. Launched in 1999, TAP has been adopted by 130 schools across 14 states and the District of Columbia, and is now operated through the National Institute for Excellence in Teaching (Chait 2007, Solmon *et al.* 2007).

The four key elements work in concert to promote sustained instructional improvement within the school (Solmon *et al.* 2007).

- *Multiple career paths* give high-performing teachers the option of taking on additional instructional and leadership responsibilities tied to increased compensation. Career teachers are regular classroom teachers at the beginning level of the career ladder. Mentor teachers, representing the next step up, provide day-to-day instructional coaching to career teachers, conduct demonstration classes, team-teach, and help plan benchmark lessons. Master teachers, representing the highest teacher position, participate in curricular/assessment planning, oversee professional development efforts, and can conduct peer evaluations tied to pay (Agam and Wardell 2007). Thus, excellent classroom teachers can be recognized and promoted while remaining in the classroom and can contribute to the improvement of other teachers and the school as a whole.

- *Ongoing applied professional growth* requires job-embedded, evidence-based, targeted professional development with a focus on determining specific teacher needs based on analysis of student data. Because classroom teachers have frequent contact with mentor and master teachers, they receive continuous feedback about their teaching as well as sustained support for improvement. Teachers are active participants in their own professional growth, and the program instills in a school's culture the premise that all teachers, even great ones, can continue to learn and improve their practice.
- *Instructionally focused accountability* represents the program's evaluation element, which utilizes multiple methods of evaluating teacher performance based on rigorous, evidence-based standards. Teachers must be evaluated at least four times a year by multiple evaluators, including master and mentor teachers and the principal, using a scientifically validated rubric derived from several widely accepted sets of standards such as the Interstate New Teacher Assessment and Support Consortium (INTASC) standards, National Board for Professional Teacher Standards (NBPTS), and Charlotte Danielson's Framework for Teaching. Evaluators must undergo intensive training and pass a rigorous certification test to conduct evaluations, and they must be recertified every year. The system includes pre- and post-conferences between teachers and evaluators to provide feedback, identify areas for improvement, and target future professional support.
- *Performance-based compensation* is provided to teachers for taking on additional responsibilities and for their performance, which is based on their evaluation results and the academic growth of their students, measured both as growth in each individual teacher's classroom and in the entire school collectively. Schools can determine the relative weight given to each factor, with the program recommending an approximate breakdown of 50 percent for evaluations, 30 percent for individual student achievement growth, and 20 percent for school-wide achievement growth.

When a school adopts TAP, teachers and evaluators are given one year to become familiar with the rubric and participate in practice assessments before the official evaluation system with monetary consequences is put into place. The program is designed so that all these elements are aligned with one another, and all are focused on the ultimate goal of improving instruction and increasing student learning. In a report from the Center for American Progress, Jerald (2009) states that an analysis of TAP demonstrates how "it is possible to tightly align teacher compensation with other human resources reform policies, but that such alignment requires a highly intentional design and cannot be left to chance."

The National Institute for Excellence in Teaching conducted an evaluation to determine the effectiveness of TAP, examining whether schools implementing TAP had different student outcomes as compared to similar control schools in each state. The study included over 300 TAP and control schools across six states (Arkansas, Indiana, Minnesota, South Carolina, Florida, and Louisiana), although it is unclear from the report how control schools were selected. Student outcomes were measured as achievement gains on state standardized tests, both at the individual teacher level and the school-wide level, calculated using William Sanders' SAS® EVAAS® method. The

Adequate Yearly Progress of TAP schools vs. controls was also reported, as well as survey results regarding teacher opinions and acceptance of the TAP model (Solmon *et al.* 2007).

Overall findings indicated that teachers in TAP schools consistently outperformed teachers in the control schools across the six states, demonstrating both higher student achievement gains and proficiency levels. The study also found that TAP teachers were supportive of the four elements of TAP, and that support increased over time. Attitudinal measures revealed that teachers “experience[d] higher quality professional development as well as more opportunities for collaboration and collegiality, and ways to improve their effectiveness in the classroom” and that, “contrary to popular belief, performance pay has neither led to competition nor susceptibility to principal bias in TAP schools” (Solmon *et al.* 2007).

Teachers in TAP schools consistently outperformed teachers in control schools.

Schacter and Thum (2005) examined student achievement, teacher satisfaction, and implementation data in a study of four TAP schools in Arizona, matched to control schools based on achievement, school size, student demographics, and location. The study found that, on average, TAP teachers outperformed the controls by about 30 percent on student achievement gains in math, reading, and language arts over two years. Authors also note that achievement gains were related to the quality of implementation. Teachers in this study reported high levels of satisfaction for the elements of the program, and a heightened sense of teacher support and collegiality. However, in the second year of the program, satisfaction with the instructional accountability and performance-based pay elements declined.

Consistent with other findings, teacher and principal surveys conducted by the TAP Foundation have found high levels of support and generally positive evaluations of the program. The teacher survey included approximately 1,700 current TAP teachers from Arizona, Indiana, Arkansas, Colorado, South Carolina, Florida, Louisiana, and Minnesota (Agam *et al.* 2006). Findings showed that teachers were supportive of TAP’s four elements and that support increased over time. Support did not vary by a teacher’s role (*e.g.*, career, master, or mentor teacher), but beginning teachers (with less than five years’ teaching experience) were more supportive than veteran teachers. Performance-based pay was the least popular among the four TAP elements, with ratings that were closer to average but still not generally negative. For instance, the authors report that “despite a lower level of support for this factor, nearly 70 percent of respondents agree or agree very much that more effective teachers should be paid more” and “only 14 percent of teachers reported that most of the teachers in their schools preferred the current step and column salary schedule to performance-based compensation” (Agam *et al.* 2006). They also note that collegiality among TAP teachers remained high.

Results from the principal survey were also largely positive, with 88 percent of principals expressing satisfaction with the quality of master teachers at their school and 76 percent expressing satisfaction with the quality of their mentor teachers. Overall, principals were happy with how the TAP elements were being implemented in their school, with 91 percent reporting that instructionally focused accountability was well-implemented, 80 percent reporting that professional growth activities were well-implemented, and 77 percent reporting that multiple career paths were well-implemented. Forty-two percent of principals felt that performance-based pay was well implemented, with another 15.3 percent feeling neutral about its implementation, and 33 percent reporting

they had not yet implemented that element. The authors note that this lower rating is not surprising given that performance pay continues to be controversial on a national scale (Agam and Wardell 2007).

Minnesota has introduced the Quality Compensation (Q Comp) program, an adaptation of TAP, which is now being used by 39 school districts and 21 charter schools throughout the state. District participation is voluntary, and all teachers in participating schools are eligible for pay incentives (Chait 2007). Hezel Associates (2009) conducted a formative and summative evaluation of the program, which included focus group interviews and online surveys with teachers, administrators, and district officials in Q Comp sites across the state. They also utilized surveys of community stakeholders, student performance data, and comparison case studies between seven Q Comp schools matched to similar non-Q Comp schools. The authors conclude that at the state and district levels, Q Comp is popular and has led to improved instructional practices, increased collaboration, meaningful professional development, more constructive evaluations, and broader teacher decision-making opportunities. Respondents felt the program provided a unifying framework for viewing and evaluating instruction that led to more consistent student learning expectations and teaching strategies. Furthermore, the authors found “a significant and positive relationship between the number of years a school has been implementing Q Comp and student achievement” (Hezel Associates 2009).

The TAP Web site highlights its successes in several other locations including Chicago, IL; Algiers Charter School Association and Forest Hill Elementary School in Louisiana; across 43 schools in South Carolina; and Richardson Independent School District and other schools in Texas. Several reports and news articles describe significant gains in student achievement in TAP schools, and glowing support from teachers and administrators on TAP’s contribution to cohesion, collaboration, accountability, professionalism, and instructional improvement in their schools.

These research results are encouraging, and they portray TAP as a promising reform effort that, when implemented faithfully, can produce substantive results. TAP is evidence-based and includes elements that have been consistently identified in the research literature as essential components of comprehensive reform and school improvement (Schacter and Thum 2005). However, because the program is still new and so complex, it will benefit from continued research exploring its effects over time, how its components can be refined and better implemented, and whether research conducted in different contexts and by other outside researchers continues to yield positive results.

TAP is a promising reform effort that can produce substantive results.

The Framework for Teaching

The Framework for Teaching (FFT) was created by Charlotte Danielson (1996) as a system for evaluating and improving instruction. It is derived from the same research base as other well-known standards, including INTASC, NBPTS, and Praxis III, it is grounded in a constructivist view of teaching, and it has served as a basis for evaluation systems in several districts and states. FFT consists of four domains: 1) planning and preparation, 2) classroom environment, 3) instruction, and 4) professional responsibilities. These domains are broken down into 22 components and 76 smaller elements on a detailed evaluation rubric, which can be used to rate each of the elements as Unsatisfactory, Basic, Proficient, or Distinguished (Danielson 2009).

FFT was created as a tool for both formative and summative assessment, and can be used for several purposes, such as teacher reflection and self-assessment, mentoring and induction, peer coaching, and evaluation and supervision. Depending on the intended use of the framework, various sources of data can be collected for evaluation, including observation, pre- and post-observation conferences, lesson videos, teaching portfolios, instructional artifacts, and teacher interviews. Danielson states that, although FFT can be used in many ways “its full value is realized as the foundation for professional conversations among practitioners as they seek to enhance their skill in the complex task of teaching” (Danielson Group 2009). In other words, FFT serves as a useful “framework” with which to link together improvement, evaluation, and other human capital development activities. Thus, FFT can be viewed as both an instrument and a system.

Of the systems reviewed here, FFT comes up most often in the peer-reviewed research literature. This is likely due to the fact that it has been around longer than the other systems described, has been utilized widely across districts and states, and can be freely accessed. Cincinnati’s Teacher Evaluation System (TES) is one prominent example of a system based on FFT. The main body of research on FFT has been conducted by researchers from the Consortium for Policy Research in Education (CPRE) over several years. Heneman *et al.* (2006) present a review of much of this work, consolidating research findings from a series of studies examining FFT as the basis of a standards-based evaluation system. These studies were conducted across four sites: Cincinnati, OH; Vaughn Charter School in Los Angeles, CA; Washoe County (Reno/Sparks), NV; and Coventry, RI.

Overall, these studies found that scores on the FFT positively correlated with student achievement, as measured by value-added gains on standardized tests. The magnitude of the gains was small to moderate and varied across the four sites. The authors speculate that this variation is due to differences in implementation and training procedures. The highest correlations were found for Vaughn Charter School and, in Cincinnati, sites that required evaluations to be conducted by multiple evaluators. In addition, Cincinnati evaluators were required to participate in intensive, high-quality training before conducting evaluations. Thus, more consistent and rigorous implementation of the evaluation standards likely led to the stronger correlations with student achievement (Heneman *et al.* 2006).

Scores on the FFT positively correlated with student achievement.

Teachers and administrators responded very positively to the components of the framework, feeling that the standards were understandable and credible, that they reflected good teaching, and that they helped improve professional conversations about practice. Teachers reported that their instruction benefited from the system, improving their lesson planning, classroom management, and reflection skills. Teachers were less likely to report changing their practices in deeper ways, those identified by the “distinguished” levels of FFT such as focusing more on student-initiated activities and empowerment. However, the researchers note that the level of feedback and focus of professional development efforts in these schools did not enable instructional change at these higher levels, pointing to the importance of aligning professional development efforts and feedback procedures with evaluation standards (Heneman *et al.* 2006).

The authors conclude that research on FFT indicates it is possible to utilize validated standards-based evaluation systems as a basis for knowledge- and skill-based

pay plans, but careful attention must be paid to implementation issues. For instance, to ensure accuracy and reliability, raters should be thoroughly trained and calibrated, multiple evaluators should be used, strong leadership and buy-in from teachers and other stakeholders should be established, and the evaluation system and other parts of the human resource management system should be tightly aligned (Heneman *et al.* 2006).

Professional Compensation System

The Professional Compensation System (ProComp) in Denver, CO, has received national attention as a bold performance-pay initiative designed through a successful collaboration between district and union leaders. The system was developed based on a four-year pilot program from 1999 to 2003 for which teachers developed their own annual objectives based on student achievement data and received financial incentives for meeting those objectives. An extensive evaluation of the pilot program conducted by the Community Training and Assistance Center (CTAC) states that, “As teachers were learning about developing and meeting measurable annual objectives, the schools and the district were learning about the necessary alignment of the curriculum, assessment, student data, human resources and other parts of the larger system with Pay for Performance” (CTAC 2004). Thus the program was conceived and developed as a comprehensive system for evaluation and compensation reform.

ProComp was conceived and developed as a comprehensive system for evaluation and compensation reform.

The pilot evaluation found that high-quality objectives were positively related to higher average student achievement across all grade levels, teachers who met their objectives had students with higher average achievement, and achievement improved as the length of teacher involvement in the pilot program increased. Over the course of the pilot, teachers learned to create higher quality objectives and teachers and administrators felt that the program served to focus efforts around achievement and effectively utilize student data to improve achievement. Pilot teachers felt that collaboration improved, and they were less opposed to performance pay than were control teachers. Pilot teachers raised issues of fairness and consistency in regard to administrator evaluations of their objectives, but they felt that a fair evaluation system could be achieved (CTAC 2004).

Denver’s current compensation system, ProComp, was developed from the findings and recommendations of this pilot. The system was voted in by teachers in 2004, and veteran teachers were allowed the choice of opting in to the system or remaining on the traditional salary schedule. Teachers hired after 2006 were enrolled in ProComp automatically. Teachers are compensated based on a combination of factors, which include teacher-determined student achievement objectives, achievement growth on state exams, performance evaluation results, professional development participation, advanced degree or certification attainment, and taking hard-to-staff positions. Incentives are additional to the base salary and are available to all teachers (Azordegan *et al.* 2005, Koppich 2008).

It is often noted that ProComp’s success in implementation is based largely on securing buy-in from teachers, unions, and the community at the outset, involving multiple stakeholders in the development and decision-making process, and allowing flexibility, choice, and multiple options within the system, especially when it came to the controversial issue of performance pay. A 2008 *Education Week* article revealed how delicate these collaborative relationships can be, describing difficult negotiations between

It is often noted that ProComp’s success in implementation is based largely on securing buy-in from teachers, unions, and the community at the outset, involving multiple stakeholders in the development and decision-making process, and allowing flexibility, choice, and multiple options within the system, especially when it came to the controversial issue of performance pay. A 2008 *Education Week* article revealed how delicate these collaborative relationships can be, describing difficult negotiations between

district and union leaders on how to adjust the program's pay scheme. The article noted that one preliminary study showed less than expected gains in student achievement related to ProComp, but cautioned that these were early results that might change as the system became more established. Currently, an independent evaluation of ProComp's effectiveness is being conducted, and the results will reveal more about how successful the program has been at improving instruction and raising student achievement (Honawar 2008). So, while the verdict is still out on ProComp's success at achieving its goals, it remains an informative model for promoting collaboration and gaining stakeholder support in the development of a new and innovative system.

ProComp secured buy-in at the outset from teachers, unions, and the community.

Peer Assistance and Review

Peer Assistance and Review (PAR) is a system in which experienced and accomplished teachers take on the role of "consulting" teachers, serving as evaluators and mentors to their peers. PAR systems are generally geared toward new teachers and struggling veteran teachers, but they could be adapted to include all teachers. PAR consists of two main elements—assistance and review. Assistance involves observing and working with new and struggling teachers to help them improve their practice, by providing instructional support, suggesting strategies, and modeling teaching. The review element allows consulting teachers to conduct formal evaluations and make employment recommendations regarding renewal and dismissal (Escamilla, Clarke, and Linn 2000, AFT/NEA 1998). While many states and districts have used peer assistance as a way to help improve teaching, this section considers programs that tie the assistance and review components together in their evaluation systems.

The most well-known example of a PAR system is the Toledo Plan, first implemented in 1981. New teachers and experienced teachers recommended for remediation are assigned to a consulting teacher who oversees their professional development and conducts ongoing evaluation. The evaluation process is described as "one of continuous mutual goal-setting using classroom observations and follow-up conferences where the [classroom teacher] and consulting teacher can analyze and set practical goals for improvement based on detailed evaluation criteria" (Toledo Federation of Teachers 2009). Consulting teachers are released from regular classroom teaching and are given a pay bonus in order to fulfill their duties. A districtwide review board made up of teachers and administrators, with chairmanship rotating between the assistant superintendent and the union president, administers the program and selects consulting teachers. Consulting teachers periodically report to the board on the progress of each teacher in their caseload and, at the end of the year, make formal recommendations for renewal or dismissal. Consulting teachers must justify their recommendations to the board using evaluation evidence collected throughout the year, and the board then chooses whether or not to accept the recommendations (Toledo Federation of Teachers 2009).

PAR has been implemented in several other districts, including Columbus, OH, Rochester, NY, Chicago, IL, and throughout California. Research on PAR indicates that it does increase the number of teachers who are dismissed or not renewed because they are unable to improve their practice with targeted support (Goldstein 2007, AFT/NEA 1998). This is considered a major improvement from the prevailing systems, which almost never dismiss teachers or deny them tenure (Weisberg *et al.* 2009). Evidence from PAR

as implemented in Columbus indicates that the program has improved teacher retention rates, particularly for teachers of color, and that first-year teachers rate the program and its professional development and support components very positively (*National Conference on Teacher Quality 2009*).

PAR was shown to be advantageous compared to traditional evaluation.

The most comprehensive study of PAR so far is an evaluation of a program implemented in California, one based closely on the Toledo model. Goldstein (2007) conducted a thorough four-year longitudinal case study of an urban district in California, describing six areas in which PAR was shown to be advantageous compared to traditional evaluation (namely principal observation). These were: 1) more time spent on evaluation due to the release of consulting teachers from other responsibilities; 2) increased linkage between professional development and evaluation, including the matching of evaluators and teachers by grade and subject and the use of performance standards; 3) improved transparency of the system due to the ongoing consulting and regular reports to the review board; 4) improved labor relations, with the teachers' union and administration working together; 5) higher levels of confidence in decisions made about tenure and dismissal due to the ongoing collection of evaluation evidence for the review board; and 6) increased accountability, with a higher percentage of teachers being dismissed as compared to almost none. Goldstein notes that these results did not come without significant challenges, and cautions districts to pay careful attention to implementation issues such as selecting consulting teachers in an accepted and unbiased way, utilizing agreed upon standards of practice and performance, and making sure all high-stakes decisions are fully documented and justifiable.

PAR appears to be another very promising model that employs strategies of distributed leadership, rigorous standards-based evaluation, and ongoing professional support in conjunction with one another. Goldstein (2007) notes the value of distributed leadership in particular, which serves to remove the burden of evaluation from the plate of busy principals, places it in the hands of those with subject-specific skills and knowledge, and provides them with the release time to meet with teachers on a continual basis. For instance, in the Columbus PAR model consulting teachers were expected to conduct at least 20 observations and 10 conferences with new teachers throughout the year, and double that amount for struggling veteran teachers. In the California model described by Goldstein consulting teachers were to visit their evaluatees once a week on average, sometimes unannounced, and conduct at least three formal observations throughout the year. The frequency and consistency of teacher contact and support in this model present a major contrast to once-a-year professional development workshops and quick, cursory principal observations.

Beginning Educator Support and Training Program

Connecticut's Beginning Educator Support and Training Program (BEST) has gained attention for its use of portfolios in evaluating beginning teachers, which is one element of a larger support and improvement system meant to recruit and retain talented teachers. First-year teachers receive structured instructional training and mentoring, which gives them an opportunity to develop their practice. They then submit a portfolio during their second year, which includes daily lesson plans, video segments of their teaching, and samples of student work. Portfolios are evaluated according to evidence-based standards

which, aligned with the INTASC standards, are based on four elements: 1) instructional design, 2) instructional implementation, 3) assessment of learning, and 4) ability to analyze teaching and learning. Each portfolio is scored by three trained raters who are experienced teachers in the same discipline as the teacher being evaluated. Teachers are provided detailed feedback on their portfolio, and they must receive a satisfactory score in order to gain full certification in the state of Connecticut. If teachers do not pass the assessment during their second year they undergo further professional development and submit another portfolio the following year. Those who do not pass in their third year are denied certification and cannot teach in Connecticut public schools (Connecticut State Department of Education 2009).

Although this system specifically targets beginning teachers in Connecticut, it represents a linked professional development and evaluation model that can be adapted to more experienced teachers. The opportunity for different professional roles, such as mentors, professional development leaders, and scorers, could be utilized in the creation of career-ladder options (Miller, Morley, and Westwater 2002). Portfolios are considered comprehensive evaluation instruments with the ability to assess multiple facets of teaching both inside and outside the classroom and that can be applied to any grade level or subject matter. On the other hand, portfolios can be considered burdensome by teachers, and it is difficult to establish reliable scoring of portfolios (Goe, Bell, and Little 2008). When including portfolio assessment as part of an evaluation system, it is recommended that teachers be given adequate release time and support to fulfill portfolio requirements, and that careful attention be paid to establishing and maintaining scoring accuracy. BEST provides a good model of this, with its embedded professional support for teachers and its thorough selection and training process for scorers.

Wilson and colleagues describe how BEST has been a key element in a purposeful, 15-year-long process to reform education policy and practice in Connecticut, resulting in “large, steady gains in student achievement and a plentiful supply of well-qualified teachers” in the state (Wilson, Darling-Hammond, and Berry 2001). While it is difficult to sort out the effects of the evaluation system itself, the authors do note that improvements in teaching seem to be the most significant driver behind these student achievement gains. One unpublished study found that teachers who received high scores on the BEST portfolio had students who significantly outperformed students of teachers with lower portfolio scores. Both beginning teachers and experienced teachers serving in mentor or evaluator roles expressed positive views about the program and felt that it improved their teaching (Toch and Rothman 2008). Thus, BEST appears to be another promising, comprehensive, aligned approach that would benefit from continued research and experimentation.

BEST appears to be another promising approach.

Implementing Comprehensive Evaluation Reform

With the exception of FFT, comprehensive evaluation reform is not prevalent in the peer-reviewed research literature. This is actually somewhat expected, given that these are fairly recent reforms involving complex educational policy issues and thus are more likely to be discussed and evaluated in research reports and news articles. It is also the case that, with complex and multi-faceted systems operating in the real world it is extremely difficult to conduct “pure” experiments that can be used to determine causal relationships. For example, it is generally both unfeasible and unethical to design experiments in which teachers are randomly assigned to schools and students are randomly assigned to teachers. Thus, we must rely more heavily on controlled quasi-experiments and thorough program evaluations, which can shed a lot of light on how a system functions and whether it is accomplishing its intended goals.

TAP, FFT, ProComp, PAR, and BEST, though studied to differing degrees, all show potential to improve instruction, raise student achievement, gain teacher support, and improve retention. They all do this by taking a comprehensive rather than piecemeal approach to reform and by centering activities and procedures around the one outcome that truly matters—student learning. The evaluation components of these programs are all widely accepted, evidence-based principles about what constitutes quality teaching and what leads to student success. They share many common criteria that tend to be aligned with well-documented and accepted standards such as INTASC, NBPTS, and Praxis III, which come from a shared research base. This is a positive sign indicating that consensus exists on what constitutes effective teaching.

Common Components of Successful Systems

Successful systems share several common components that come up continually in reports about effectively measuring teaching and reforming evaluation and compensation systems (e.g., Chait 2007, Donaldson 2009, Heneman *et al.* 2006, Jerald 2009, Koppich 2008, Little, Goe, and Bell 2009, Toch and Rothman 2008, Weisberg *et al.* 2009). The following main points appear to be essential to successfully implementing comprehensive evaluation reform.

Establish a credible and meaningful evaluation system. Reform doesn’t work if the people involved do not believe in it or worry it will be implemented unjustly. There are several ways to ensure credibility and transparency of an evaluation system.

- Involve multiple stakeholders in the development and revision of the system, and make sure that the system is meaningful to everyone involved. It is especially important to gain trust and secure buy-in from administrators, teachers, and teachers unions as they are the ones who will be implementing the system. Policymakers, parents, students, and community members can also be involved. Make decisions collaboratively and continue to assess the system and refine it based on evaluation outcomes.
- Use validated, credible evaluation measures and ensure that they are faithfully implemented. Teachers should be made knowledgeable about the standards against which they are to be evaluated, and they should feel that the standards are valid components of high-quality teaching. Evaluators should be thoroughly trained on the evaluation instrument, the reliability of their scoring should be established, and they should be periodically reassessed to ensure they are still scoring reliably, with recalibration training provided as needed. Evaluations should occur several times a year to gain a broad assessment of a teacher's practice, and multiple evaluators should be used if possible so scores can be calibrated against one another.
- Use multiple measures that evaluate multiple facets of what contributes to good teaching. Teaching is a complex behavior that involves many different skills and competencies. While some consensus exists on what constitutes good teaching, many questions remain about what matters most and how to measure it. Incorporating several measures—measures that are considered valid and meaningful by local stakeholders—will help to establish a system that everyone can trust.

Evaluations should be tied to curricular standards, professional development, and targeted support.

Create linked and integrated systems. Evaluation should be tied to curricular standards, professional development activities, and targeted support. Once an evaluation system is fully developed, it should also be aligned with human capital decisions such as recruitment, hiring, retention, compensation, career advancement, and remediation. A credible system should be established before being used to make high-stakes decisions.

- Include embedded, ongoing professional development. Professional development should be aligned with evaluation standards and evaluation results should be used to provide targeted, individualized feedback and support to teachers. Evaluators and teachers should work collaboratively to identify and address areas of weakness in a system focused on instructional improvement rather than solely on accountability.
- Incorporate opportunities for career advancement within teaching. Providing differentiated roles and upward mobility for teachers without requiring them to move into a purely administrative role helps to retain the most capable teachers and professionalize the field. Teachers should be given career pathways that allow them to develop their leadership and instructional skills, to be recognized for their excellence, and to be rewarded for taking on additional responsibilities.

Using Student Achievement Data

Calls for the inclusion of student achievement data in teacher evaluation systems are prevalent and strong, and they have particular political weight right now with support from the Obama administration, the National Governors Association, and several educational research institutions (e.g., Brodie 2009, Dillon 2009, Goldrick 2002, Weisberg *et al.* 2009). Along with this push come concerns and cautions that the concept of teacher effectiveness should not become too narrowly focused on standardized student achievement and should consider the many ways that teachers and school systems contribute to student learning and development (Goe *et al.* 2008). Propositions to hold teachers increasingly accountable for their students' achievement are often met with skepticism and alarm by teachers who recognize that many factors contribute to student achievement that are beyond their control, and standardized tests are far from perfect indicators of student learning. These are valid concerns and, assuming that the pressure to consider student data in evaluation is here to stay, there are strategies for addressing these concerns. Utilizing student data appropriately may present opportunities for furthering instructional improvement and recognizing deserving teachers.

Value-added measurement. One reason for the recent push to incorporate student data into evaluation has been the introduction of value-added measurement. This technique allows gains in student achievement to be calculated from several years of standardized achievement test data. Based on a student's previous test achievement, the method is used to predict what the student's achievement is expected to be the next year. If that student's achievement increases (or decreases) by more than expected, those gains are attributed to the teaching the student received. If a teacher consistently has students who score higher than expected over several years, that teacher is considered "effective." This method is an improvement over previous student achievement measures, which considered the average achievement of a teacher's classroom as compared to other classrooms. Measuring average achievement using the latter method disadvantages those teachers who work with lower-achieving students and does not control for the many outside factors that contribute to achievement. Value-added methodology attempts to isolate those outside factors and take them out of the equation.

Value-added methodology has its benefits and drawbacks. Looking at student achievement gains is often considered more objective and relevant than other evaluation measures, since it is not susceptible to personal biases (as introduced in observation or portfolio scoring) and it focuses directly on the outcomes of teaching. Value-added is particularly useful for identifying very high- and low-performing teachers, and thus can be used to determine which teachers deserve recognition and which are struggling and need targeted support. Teachers identified as highly effective can be tapped as mentors or model teachers to help other teachers improve their practice (Goe 2008). Evaluation systems in both Dallas and Houston provide examples of how value-added data is used to distinguish high-performing teachers and learn from their practices, identify teachers' areas of strength and weakness to provide targeted support, inform pay and career advancement decisions, and evaluate programs and practices implemented by the district. For example, Goe (2008) describes how Dallas has utilized value-added data to examine

Many factors beyond teachers' control contribute to student achievement, and standardized tests are imperfect indicators of student learning.

the alignment of professional development and other school improvement efforts by identifying high- and low-performing schools using value-added scores and comparing their “professional development...instructional practices, staff cohesiveness, administrative leadership and support, and a number of other school and student factors.”

However, there are numerous limitations and concerns around using value-added measurement, especially regarding higher-stakes uses. Value-added measures are not completely clean, unbiased measures of teacher effects, but inevitably include the influence of additional classroom-level factors such as curricular quality, peer effects, school climate, and availability of resources (Braun 2005). It is important to recognize that value-added scores represent a relative ranking, usually calculated based on the performance of all teachers in a school or district. Thus, a teacher who is effective in one district could be considered average in another or vice versa, making cross-district and statewide comparisons problematic (Goe 2008). There remains disagreement over which statistical specifications are most accurate, for instance whether or not student background characteristics should be included in the calculation. It is still unclear whether value-added scores remain stable over time, in different contexts or across different standardized tests. The method requires linked student and teacher data over at least three years, and the procedures are very sensitive to missing data, which is prevalent in school data systems where students often switch locations. Furthermore, value-added scores are directly related to the quality of the tests used to calculate them, so using standardized tests that are not aligned with the curriculum being taught will misrepresent a teacher’s effectiveness score (McCaffrey *et al.* 2003).

Even if these technical considerations could be addressed, the concern remains that value-added measures can only provide a very limited amount of information about teaching. Prince *et al.* (2008) from the Center on Educator Compensation Reform point out that a significant majority of teachers cannot be evaluated using value-added mea-

A significant majority of teachers can’t be evaluated using value-added measures.

sures. These include teachers in non-tested subjects, such as music, art, and physical education; teachers of non-tested grades, including preK through second and the high school grades; and teachers of English Language Learners and students with disabilities. They stress that if an evaluation system is to include incentives based on student achievement data, the performance criteria must be made applicable and available to all teachers, not just to those for whom value-added scores can be calculated. Goe (2008) emphasizes that value-added scores give us no information about what teachers are doing that makes them effective. Thus, evaluation methods such as observation must be used in concert with value-added measures to obtain information about how to improve teaching practice. In addition, it remains unclear how to handle situations in which teachers co-teach or otherwise share responsibility for students and how to account for potential “spillover” effects between teachers. Recent research shows that achievement gains in certain subjects may affect gains in other subjects but that spillover effects are complex and do not apply equally to all subjects. For instance, the study found that math teachers contributed to student achievement gains in reading, but the spillover did not occur the other way around (Koedel 2007).

Distrust of standardized testing, a reluctance to promote even more test-taking activities in school, and differential treatment for teachers in non-tested areas are legitimate teacher concerns about being judged based on student achievement growth.

A discussion of the merits and pitfalls of standardized testing is beyond the scope of this paper, but it remains the case that the quality of a value-added score will reflect the quality of the test used. Given the fact that standardized testing is unlikely to disappear anytime soon, it may be most constructive to focus on using test scores responsibly and advocating for testing improvements. Ideally, the act of “teaching to the test” is less of a problem if the test is high-quality, well-conceived, and aligned with accepted curricular standards. Value-added scores will be most valid if they are based on a trusted and relevant test that is used consistently from year to year and if they are calculated carefully with the most complete and accurate data possible.

Research on the use of value-added measures in instructional accountability systems concludes that value-added should be utilized with several caveats and that much more research is needed to improve our understanding of what these measures are capturing. Value-added may be most appropriate for identifying those teachers at either extreme of the performance distribution, but it is less reliable at ranking teachers in the middle of the distribution and can provide only a limited amount of performance information. Interestingly, however, Goldhaber and Hansen (2008) point out that the reliability of value-added measures, although possibly low, is comparable to that of performance measures used in other sectors of the economy. Likewise, given the extremely limited usefulness of the current evaluation system, value-added measures may have a role in a re-conceptualized system. They may be very useful as one component of an evaluation system, but until more research is conducted it is not recommended that they be used as the sole criteria for accountability.

Additional measures of teaching and learning. Given the many subject areas and grade levels for which value-added measurement is not possible, it is necessary to consider alternate ways of evaluating teaching practice and student achievement. One important point is that student learning can be conceptualized and thus measured in many different ways. As mentioned above, the best evaluation systems will utilize several indicators of performance, which may include standards-based evaluations, value-added student achievement scores, leadership activities and other contributions to the school, and specific skills and qualifications. This principle holds for considering student achievement data as well. While student growth on standardized tests may be one element to consider, student learning can also be measured by locally designed assessments that are aligned with curricular standards. In taking this approach, however, states and districts must keep in mind that a substantial investment of time and resources is necessary to create high-quality assessments from scratch (see Prince *et al.* 2008).

Another method to consider is the analysis of classroom assignments. Two promising instruments have been developed that analyze teacher assignments and student work using standards-based rubrics, with encouraging results. These are the Intellectual Demand Assignment Protocol (IDAP) developed by Newmann *et al.* from the Consortium on Chicago School Research (Newmann, Bryk, and Nagaoka 2001, Newmann, Lopez, and Bryk 1998) and the Instructional Quality Assessment (IQA) developed by Matsumura *et al.* through the National Center for Research on Evaluation, Standards, and Student Testing (Junker *et al.* 2006, Matsumura *et al.* 2006).

Value-added may be most appropriate for identifying those teachers at either end of the performance distribution. Not recommended is that value-added be the sole criteria for accountability.

Districts may want to think creatively about ways to assess and include outcomes beyond student achievement.

A second crucial point is that one method or formula for evaluation may not be appropriate for assessing all teachers. Prince *et al.* (2008) present various options being used by districts to evaluate those teachers who cannot be evaluated using value-added, focusing specifically on teachers of non-tested subjects and grades and on teachers of students with disabilities and English Language Learners. For some of these groups, standards-based performance assessments may be a more appropriate option for evaluating teachers than student achievement. As one example, they describe how standardized testing is not considered developmentally appropriate for assessing young children in grades preK–2. Thus, it would make more sense to utilize a teaching assessment targeted toward this student population, such as the Classroom Assessment Scoring System (La Paro, Pianta, and Stuhlman 2004). Finally, it is important to remember that other aspects of teacher practice and student development are important outcomes to schools and communities, such as increased student motivation, engagement, civic-mindedness, and social/emotional well-being, as well as teacher contributions to school functioning and the larger educational context (Goe *et al.* 2008). One new study shows that effective teachers not only contribute to student learning, but can significantly improve the practices of their peers as well (Jackson and Bruegmann 2009). Thus, districts may want to think creatively about ways to assess these additional outcomes and include them in their evaluation systems. (For further details about methods and instruments for evaluating teaching performance depending on the priorities of your evaluation system, refer to Goe *et al.* 2008, Little *et al.* 2009, and Coggshall *et al.* 2008.)

The bottom line is that current calls to incorporate student achievement data into evaluations are strong and may not go away, but they may present a true opportunity to inform the evaluation process if used appropriately. In order to successfully do so, involve teachers in deciding how to account for student learning and other relevant outcomes in evaluation using a combination of measures so teachers feel that they are being evaluated comprehensively and fairly. If opposition to standardized test data is very strong, try to utilize other assessment measures that are more closely aligned with local evaluation standards and professional development efforts. If there is interest in using value-added measures, work to create the necessary data structures, collect complete data, use an appropriate standardized test that is aligned with curriculum, and employ a well-researched model for the calculations. There should be some combination of measures that teachers, administrators, and other stakeholders can agree on. Work toward reaching a consensus through compromise, and be open-minded to the many options out there. Do not lose sight of the fact that the evaluation system must be designed to benefit individual teachers, the school as a whole, and most importantly the students.

Linking Evaluation to Pay

Compensation reform is typically the most controversial component of a fully integrated system. There seems to be agreement that teachers should be rewarded for their contributions and recognized for excellence, but skepticism remains that this can be accomplished in a fair and consistent way. This review suggests that a credible, validated, standards-based evaluation system linked to professional support should be established first before trying to tie pay directly to teaching performance, and that this can be done

given careful planning and genuine investment (Heneman *et al.* 2006). However, it is also the reality that, like calls for the use of student data, calls for compensation reform are strong, unlikely to disappear, and already being implemented (Azordegan *et al.* 2005, Dillon 2009, Hassel and Hassel 2007). Again, this presents an opportunity for improving education if implemented strategically.

The research on standards-based evaluations such as TAP and FFT show that they can be valid and reliable enough to use as the basis for a performance pay system (Milanowski 2004, Odden 2004, Schacter and Thum 2004), but it is crucial to use a system that is supported by teachers and to implement it well. This means basing performance incentives on multiple agreed-upon criteria, making incentives available to teachers of all grades and subjects, not placing an artificial cap on the number of teachers who can receive rewards, rewarding both individual and group performance, and building flexibility for different needs and comfort levels into the program (Prince *et al.* 2008, TeacherSolutions 2007).

A review of research on performance pay shows that teachers do respond positively to financial incentives; however, it cautions that a system must be well-designed so that the incentives cause changes in teacher behavior that actually result in the desired educational outcomes. For example, rewards based solely on standardized achievement gains may lead to higher test scores, but they may not lead to increased student engagement or thorough coverage of a broad and rich curriculum. Evaluation must be multifaceted and aligned with professional development and other school improvement efforts in order to minimize the potential for “gaming the system” by narrowing teaching practices or even cheating (Podgursky and Springer 2007).

It is also important to remember that performance can be considered just one element in a differentiated compensation system. A brief from the National Governors Association outlines several “pay for contribution” options in addition to performance pay, including pay for filling hard-to-staff positions and skill shortages, taking on additional responsibilities and leadership roles, and attaining relevant skills and certifications shown to be related to teacher effectiveness (Hassel and Hassel 2007). TAP and ProComp provide convincing models for how to structure pay incentives so that they include multiple elements and gain the support of teachers. In addition, the Center for Educator Compensation reform presents several options for structuring pay incentives when dealing with teachers who cannot be evaluated using standardized achievement measures, including important considerations about when it is most appropriate to use school-based, team-based, or individual-based performance incentives (Prince *et al.* 2008).

Although standards-based evaluations can be used as a basis for performance pay, it is crucial to use a system supported by teachers.

Implementation Considerations

A perfectly designed system can fall apart if not implemented with fidelity to the procedures and with attention paid to the agreed-upon goals. Several of the recommendations discussed are crucial here, such as securing buy-in from all stakeholders, committing to comprehensive change, and ensuring that evaluation is credible and useful by 1) establishing accepted, evidence-based teaching standards, 2) using a valid instrument, 3) thoroughly training and recalibrating raters, 4) employing multiple evaluators, and 5) establishing a process for providing feedback and targeting support. A system should

promote transparency such that teachers can easily understand what is expected of them, and it should serve to facilitate increased communication between evaluators and evaluatees. Teachers should feel that they are benefiting from the system, rather than simply being judged by it (Heneman *et al.* 2006). Below are some additional considerations related to implementation.

Demonstrated commitment from the leadership. District and school leaders must convey their commitment and support for reform throughout the process, not just in the form of rhetoric, but through actions that allow teachers to meet the goals of the new system (Prince *et al.* 2008). Otherwise it is likely that teachers will view reforms as another passing fad in education. For instance, if a new evaluation system includes procedures requiring significant teacher time, such as completing a portfolio assessment, teachers should be provided with release time and other resources so that the new assessment does not become overly burdensome for them. Likewise, if principals are being asked to undergo thorough training and conduct evaluations more frequently, they must be relieved of other duties in order to accomplish this. Increased leadership accountability will also help to instill confidence in the system. For example, evaluators could be held accountable for conducting evaluations accurately and providing detailed feedback, district-level principal evaluations could be tied to the same student outcomes that are considered in teacher evaluations, and districts could consider rewarding or sanctioning schools based on student test results, comparisons to similar schools, and parent and student survey data (Donaldson 2009, Toch and Rothman 2008).

Allowing sufficient time and resources. Heneman *et al.* (2006) caution, “This commitment is not for the faint of will, time, or budget; it is for those who want to invest in creating a high-quality teaching force with the competencies needed to help kids learn in a standards-based world.” Allow adequate time to thoughtfully design the system, and conduct pilot testing before full implementation if possible. Even after the system is in place, it will take time for the new procedures to function smoothly and become fully accepted. District and school leaders should expect to see some unintended consequences and complications in the first few years, making it important to continually gather feedback from teachers and other stakeholders, and use it to reevaluate and refine the system. A comprehensive approach will also be costly at the outset, particularly if it includes a performance pay component. It is important to secure adequate and reliable funding to fully implement the reform (Hannaway and Rotherham 2008).

Conclusion

In light of the growing recognition that high-quality teaching is central to the future of education in this country, it is difficult to ignore calls for reform to our current teacher evaluation process. Research calls for a long-term, comprehensive approach to reform that links teacher evaluation with other important school improvement measures such as professional development, hiring, retention, compensation, and dismissal. The good news is that there isn't a need to reinvent the wheel. Several research-supported models for implementing comprehensive teacher evaluation systems exist and are currently being used in districts across the country. Some of these systems are still new and awaiting further evidence of their effectiveness, but they are worth exploring and may serve as valuable models for adapting a system to local circumstances. It is clear, however, that change will require serious investment on the part of education leaders, and collaboration in this process is crucial.

Implications for the Union

Teachers unions represent an important voice in the debate on accountability. The unions are responsible for protecting their members from unfair evaluation and compensation practices while promoting the improvement of the education system as a whole. This leads to a situation where traditional labor-management relationships are less appropriate than more collaborative arrangements (Johnson *et al.* 2007). The reforms presented in this review were often made possible through innovative partnerships between administrations and unions. District-union collaborations were key to the successful implementation of systems such as ProComp in Denver, PAR in Toledo, and Q Comp in Minnesota (Chait 2007, Jupp 2005, Toledo Federation of Teachers 2009).

Representing such a large and diverse profession as teaching can pose challenges. In one of the few studies focusing on the views of local union presidents, these leaders expressed having to meet the needs of two very different populations: beginning teachers and veteran teachers. Beginning teachers tended to have greater expectations for professional support and be more supportive of evaluation and compensation reforms. Veteran teachers tended to remember the original conditions that led to the single salary schedule and were more inclined to maintain the traditional compensation system (Johnson *et al.* 2007). Catering to the different needs and standpoints of these two groups might prove daunting, and the union should look for ways to accommodate both groups. An informative model can be found in ProComp, which was implemented with an option for

Reforms were often made possible through innovative partnerships between administrations and unions.

veteran teachers to opt out of the performance pay system if they chose. Hannaway and Rotherham (2008) note that successful systems such as ProComp, TAP, Q Comp, and the Toledo Plan all require a certain proportion of teachers in a school to elect to participate before the system can be implemented, reinforcing the idea that teacher buy-in and flexible participation requirements are key to successful reform.

Collaboration and compromise will be vital to fueling education reforms and improving student learning.

Finally, collaboration and compromise will be vital to fueling education reforms and improving student learning in the long-run. Striving for a hybrid approach to evaluation using multiple measures of teaching performance and student outcomes (Toch and Rothman 2008) has the potential to gain consensus among stakeholders while broadening the definitions of teaching and learning. In an article on district-union collaboration, Brad Jupp, the union leader in Denver when ProComp was instituted, stressed the importance of “avoiding false choices,” stating that, “It’s a false choice to say, let’s look at state tests or let’s look at individual student results. In Denver, we said, ‘Let’s do both’” (Varlas 2009). He also noted that, sometimes experimentation and bold action are necessary, and in the case of ProComp results did not confirm the fears of critics but led to increased motivation and focus on improving school performance and student growth.

Opportunity for Change

The recent research and policy attention surrounding the issue of teacher evaluation and educational improvement presents an ideal window of opportunity to work toward comprehensive reform. New funds are being made available through the American Recovery and Reinvestment Act and interest from philanthropic organizations such as the Gates Foundation. The Gates Foundation has announced plans to fund teacher evaluation improvement initiatives (Wolfe 2009) in addition to providing research grants to learn more about defining and measuring effective teaching and retaining and rewarding excellent teachers (Robelen 2008). Other exciting research initiatives are emerging—such as the Strategic Management of Human Capital project led by Allan Odden and James Kelley of the Consortium for Policy Research in Education—to better understand how to align human capital management systems in public education (see, for example, Strategic Management of Human Capital 2009). The promising models of comprehensive teacher evaluation in this review show that it is possible to move toward true reform, and the timing is right to act now.

Appendix

Summary of Evaluation Systems

Evaluation System	Where the System Operates	Features of the System	Research Findings
Teacher Advancement Program (TAP)	In schools across: Arizona Arkansas Colorado Illinois Indiana Louisiana Minnesota North Carolina Ohio Pennsylvania South Carolina Tennessee Texas Washington, DC	<ul style="list-style-type: none"> • Includes four integrated system elements: <ol style="list-style-type: none"> (1) Multiple career paths (2) Ongoing applied professional growth (3) Instructionally focused accountability (4) Performance-based compensation • Evaluations are based on well-researched and accepted standards such as INTASC and NBPTS • Performance rewards are determined using both standards-based evaluation and student achievement data • Requires extensive evaluator training and recalibration 	<ul style="list-style-type: none"> • Evaluations of TAP schools found that TAP teachers consistently outperformed teachers in similar non-TAP schools in both student achievement gains and proficiency • Teacher and principal surveys have found high levels of support for the system • Teachers felt the program promoted collaboration and professional growth • Performance-based compensation was the least popular element, but was not evaluated negatively and was preferred over the traditional compensation system
Framework for Teaching (FFT)	Implemented and adapted in several states and districts; has been studied in: Cincinnati, OH Los Angeles, CA Washoe County, NV Coventry, RI	<ul style="list-style-type: none"> • Bases evaluation and improvement around four domains: <ol style="list-style-type: none"> (1) planning and preparation (2) classroom environment (3) instruction (4) professional responsibilities • Standards are aligned with INTASC and NBPTS • Evaluation rubrics can be used formatively or summatively, and can be applied to multiple sources of data (e.g., observations, portfolios, video of teaching, student work, etc.) 	<ul style="list-style-type: none"> • Observation scores using the FFT correlated positively with student achievement • Stronger relationships may be due to better evaluator training and implementation • Teachers and administrators responded favorably to the components of the framework, saying it helped improve professional conversations about practice and benefitted their teaching
Professional Compensation System (ProComp)	Denver, CO	<ul style="list-style-type: none"> • Teachers work with principals to establish individual annual performance objectives based on student achievement • Teachers are compensated for a combination of factors including annual objectives, student achievement growth on state exams, performance evaluation results, professional development participation, advanced degree or certification attainment, and taking hard-to-staff positions 	<ul style="list-style-type: none"> • Pilot evaluation found high quality objectives were positively related to higher average student achievement, and achievement improved as the length of teacher involvement in the program increased • Pilot teachers learned to create higher quality objectives over the course of the program • Teachers and administrators felt the program effectively focused efforts around achievement, and promoted collaboration

Summary of Evaluation Systems

Evaluation System	Where the System Operates	Features of the System	Research Findings
Peer Assistance and Review (PAR)	Implemented in several locations, including: Toledo, OH Columbus, OH Rochester, NY Chicago, IL Districts throughout California	<ul style="list-style-type: none"> Accomplished “consulting” teachers are released from regular teaching duties to serve as evaluators and mentors to their peers, and are compensated accordingly Consulting teachers provide both frequent, targeted professional development and conduct formal evaluations of beginning and remediated teachers Consulting teachers make annual recommendations regarding tenure and dismissal to a district-wide review board made up of teachers and administrators and headed by district and union leaders 	<ul style="list-style-type: none"> Results found an increased number of underperforming teachers were dismissed under PAR compared to traditional evaluation systems The distributed leadership model of PAR led to more time spent on evaluation, increased linkage with professional development, improved transparency of the system, improved labor relations, and increased accountability Teachers and administrators evaluated PAR positively, and some evidence suggested it helped improve teacher retention
Beginning Educator Support and Training Program (BEST)	All districts in Connecticut	<ul style="list-style-type: none"> First year teachers receive structured instructional training and mentoring Teachers must submit a second-year portfolio, including daily lesson plans, video segments of their teaching, and samples of student work Portfolios are evaluated using INTASC-based standards on four elements: <ol style="list-style-type: none"> instructional design, instructional implementation, assessment of learning, and ability to analyze teaching and learning Each portfolio is scored by three trained raters who are experienced teachers in the same discipline as the teacher being evaluated 	<ul style="list-style-type: none"> Portfolios are considered comprehensive and multi-faceted evaluation instruments, but it is difficult to establish reliable scoring of portfolios and they can be burdensome for teachers Throughout Connecticut’s 15-year implementation of education reform, student achievement and recruitment of high-quality teachers have steadily increased in the state Some findings indicate the BEST portfolio may contribute to improved teaching and relate to student achievement gains

References

- AFT/NEA. 1998. *Peer Assistance and Peer Review: An AFT/NEA Handbook*. Prepared for Shaping the Profession that Shapes the Future: An AFT/NEA Conference on Teacher Quality. Retrieved 10/2/09 from www.aft.org/pubs-reports/downloads/teachers/par-hndbk.pdf.
- Agam, K., D. Reifsneider, and D. Wardell. 2006. *The Teacher Advancement Program (TAP): National Teacher Attitudes*. Teacher Advancement Program Foundation. Retrieved 10/2/09 from www.tapsystem.org/publications/publications.taf?page=reports.
- Agam, K., and D. Wardell. 2007. *2007 Annual TAP Principal Survey: Select Findings*. Santa Monica, CA, and Washington, DC: National Institute for Excellence in Teaching. Retrieved 10/2/09 from www.tapsystem.org/publications/publications.taf?page=reports.
- Azordegan, J., P. Byrnett, K. Campbell, J. Greenman, and T. Coulter. 2005. *Diversifying Teacher Compensation*. Denver, CO: Education Commission of the States. Retrieved 10/2/09 from www.eric.ed.gov:80/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/b9/39.pdf.
- Bell, C. A., O. M. Little, A. J. Croft, and D. H. Gitomer. 2009. *Measuring Teaching Practice: A Conceptual Review*. Paper presented at the American Educational Research Association.
- Brandt, C., C. Mathers, M. Oliva, M., Brown-Sims, and J. Hess. 2007. *Examining District Guidance to Schools on Teacher Evaluation Policies in the Midwest Region* (Issues & Answers Report, REL 2007-No. 030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved 10/2/09 from ies.ed.gov/ncee/edlabs/regions/midwest/pdf/techbrief/tr_00408.pdf.
- Braun, H. I. 2005. "Value-added Modeling: What Does Due Diligence Require?" In R. W. Lissitz, ed., *Value-added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.
- Brodie, J. M. 2009. "Duncan Outlines Priorities at Principals Conference: Teacher Evaluations Cited as Education Reform Priority." *Education Daily* 42.
- Community Training and Assistance Center. 2004. *Catalyst for Change: Pay for Performance in Denver Final Report*. Boston, MA: Author. Retrieved 10/2/09 from www.ctacusa.com/PDFs/Rpt-CatalystChangeFull-2004.pdf.

- Chait, R. 2007. *Current State Policies that Reform Teacher Pay: An Examination of Pay-for-performance Programs in Eight States*. Washington, DC: Center for American Progress. Retrieved 10/2/09 from www.americanprogress.org/issues/2007/11/pdf/teacher_pay.pdf.
- Coggshall, J., J. Max, and K. Bassett. 2008. *Key Issue: Using Performance-based Assessment to Identify and Support High-quality Teachers*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved 10/2/09 from www.tqsource.org/publications/keyIssue-June2008.pdf.
- Connecticut State Department of Education. 2009. *A Guide to the BEST Program for Beginning Teachers 2008–2009*. Hartford, CT: Author.
- Danielson, C. 1996. *Enhancing Professional Practice: A Framework for Teaching* (1st edition). Alexandria, VA: Association for Supervision and Curriculum Development.
- Danielson, C. 2009. "A Framework for Learning to Teach." *Educational Leadership* 66(8). Retrieved 10/2/09 from www.ascd.org/publications/educational_leadership/summer09/vol66/num09/A_Framework_for_Learning_to_Teach.aspx.
- Danielson Group. 2009. "The Framework for Teaching." Retrieved 10/2/09 from charlottedanielson.com/theframeteach.htm.
- Darling-Hammond, L. 2000. "Teacher Quality and Student Achievement: A Review of State Policy Evidence." *Education Policy Analysis Archives* 8:(1). Retrieved 10/2/09 from epaa.asu.edu/epaa/v8n1/.
- Dillon, S. 2009. "Dangling Money, Obama Pushes Education Shift." *The New York Times* August 16. Retrieved 10/2/09 from www.nytimes.com/2009/08/17/education/17educ.html?emc=eta1.
- Donaldson, M. L. 2009. *So Long, Lake Wobegon? Using Teacher Evaluation to Raise Teacher Quality*. Washington, DC: Center for American Progress. Retrieved 10/2/09 from www.americanprogress.org/issues/2009/06/teacher_evaluation.html.
- Escamilla, P., T. Clarke, and D. Linn. 2000. *Exploring Teacher Peer Review*. Washington, DC: National Governors Association Center for Best Practices. Retrieved 10/2/09 from www.nga.org/Files/pdf/000125PEERREVIEW.pdf.
- Goe, L. 2008. *Key Issue: Using Value-added Models to Identify and Support Highly Effective Teachers*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved 10/2/09 from www2.tqsource.org/strategies/het/UsingValueAddedModels.pdf.
- Goe, L., C. Bell, and O. M. Little. 2008. *Approaches to Evaluating Teacher Effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved 10/2/09 from www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf.
- Goldhaber, D., and M. Hansen. 2008. *Assessing the Potential of Using Value-added Estimates of Teacher Job Performance for Tenure Decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, Urban Institute. Retrieved 10/2/09 from www.urban.org/UploadedPDF/1001265_Teacher_Job_Performance.pdf.
- Goldrick, L. 2002. *Improving Teacher Evaluation to Improve Teaching Quality*. Washington, DC: National Governors Association Center for Best Practices. Retrieved 10/2/09 from www.nga.org/Files/pdf/1202IMPROVINGTEACHEVAL.pdf.

- Goldstein, J. 2007. "Easy to Dance to: Solving the Problems of Teacher Evaluation with Peer Assistance and Review." *American Journal of Education* 113(3): 479–508.
- Gordon, R., T. J. Kane, and D. O. Staiger. 2006. *Identifying Effective Teachers Using Performance on the Job*. Washington, DC: Brookings Institution. Retrieved 10/2/09 from www.brookings.edu/~media/Files/rc/papers/2006/04education_gordon/200604hamilton_1.pdf.
- Hannaway, J., and A. J. Rotherham. 2008. *Collective Bargaining in Education and Pay for Performance*. Nashville, TN: National Center on Performance Incentives. Retrieved 10/2/09 from www.performanceincentives.org/data/files/directory/ConferencePapersNews/Hannaway_et_al_2008.pdf.
- Hassel, E. A., and B. C. Hassel. 2007. *Improving Teaching Through Pay for Contribution*. Washington, DC: National Governors Association Center for Best Practices. Retrieved 10/2/09 from www.nga.org/Files/pdf/0711IMPROVINGTEACHING.PDF.
- Heneman, H. G., A. Milanowski, S. M. Kimball, and A. Odden. 2006. *Standards-based Teacher Evaluation as a Foundation for Knowledge- and Skill-based Pay*. Philadelphia, PA: Consortium for Policy Research in Education. Retrieved 10/2/09 from www.cpre.org/images/stories/cpre_pdfs/RB45.pdf.
- Hezel Associates. 2009. *Quality Compensation for Teachers: Summative Evaluation*. Syracuse, NY: Author. Retrieved 10/2/09 from archive.leg.state.mn.us/docs/2009/other/090321.pdf.
- Honawar, V. 2008. "Model Plan of Merit Pay in Ferment." *Education Week* 27.
- Jackson, C. K., and E. Bruegmann. 2009. *Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers*. Cambridge, MA: National Bureau of Economic Research.
- Jerald, C. 2009. *Aligned By Design: How Teacher Compensation Reform Can Support and Reinforce Other Educational Reforms*. Washington, DC: Center for American Progress. Retrieved 10/2/09 from www.americanprogress.org/issues/2009/07/pdf/teacher_alignment.pdf.
- Johnson, S. M., M. L. Donaldson, M. S. Munger, J. P. Papay, and E. K. Qazilbash. 2007. *Leading the Local: Teachers Union Presidents Speak on Change, Challenges*. Washington, DC: Education Sector. Retrieved 10/2/09 from www.educationsector.org/usr_doc/UnionLeaders.pdf.
- Junker, B., Y. Weisberg, L. C. Matsumura, A. Crosson, M. K. Wolf, A. Levison, *et al.* 2006. *Overview of the Instructional Quality Assessment*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved 10/2/09 from www.cse.ucla.edu/products/reports/r671.pdf.
- Jupp, B. 2005. "The Uniform Salary Schedule: A Progressive Leader Proposes Differential Pay." *Education Next* 5(3).
- Kennedy, M. M. 2008. "Sorting Out Teacher Quality." *Phi Delta Kappan* 90(5).
- Koedel, C. 2007. *Teacher Quality and Educational Production in Secondary School*. Nashville, TN: National Center on Performance Incentives. Retrieved 10/2/09 from www.performanceincentives.org/data/files/news/PapersNews/Koedel_2007a_Revised.pdf.

- Koppich, J. E. 2008. *Reshaping Teacher Policies to Improve Student Achievement*. Berkeley, CA: Policy Analysis for California Education. Retrieved 10/2/09 from gse.berkeley.edu/research/pace/reports/PB.08-3.pdf.
- La Paro, K. M., R. C. Pianta, and M. Stuhlman. 2004. "The Classroom Assessment Scoring System: Findings from the Prekindergarten Year." *The Elementary School Journal* 104(5): 409–426.
- Little, O. M., L. Goe, and C. Bell. 2009. *A Practical Guide to Evaluating Teacher Effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved 10/2/09 from www.tqsource.org/publications/practicalGuide.pdf.
- Matsumura, L. C., S. C. Slater, B. Junker, M. Peterson, M. Boston, M., Steele, *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. Retrieved 10/2/09 from www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/e0/f6.pdf.
- McCaffrey, D. F., J. R. Lockwood, D. M. Koretz, and L. S. Hamilton. 2003. *Evaluating Value-added Models for Teacher Accountability*. Santa Monica, CA: RAND. Retrieved 10/2/09 from www.rand.org/pubs/monographs/2004/RAND_MG158.pdf.
- Milanowski, A. 2004. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79(4): 33–53.
- Miller, J. B., V. S. Morley, and B. Westwater. 2002. "The Beginning Educator Support and Training Program in Connecticut." *Journal of Physical Education, Recreation, and Dance* 73(4): 24–27.
- National Conference on Teacher Quality. 2009. "Toward a Seamless Transition: Columbus Peer Assistance and Review Program." Retrieved 10/2/09 from www.ed.gov/inits/teachers/exemplarypractices/d-2.html.
- 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. Retrieved 10/2/09 from csr.uchicago.edu/publications/p0a02.pdf.
- 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. Retrieved 10/2/09 from csr.uchicago.edu/content/publications.php?pub_id=50.
- Odden, A. 2004. Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education* 79(4): 126–137.
- Podgursky, M., and M. G. Springer. 2007. "Credentials Versus Performance: Review of the Teacher Performance Pay Research." *Peabody Journal of Education* 82(4): 551–573.
- Prince, C. D., P. J. Schuermann, J. W. Guthrie, P. J. Witham, A. T. Milanowski, and C. A. Thorn. 2008. *The Other 69 Percent: Fairly Rewarding the Performance of Teachers of Non-tested Subjects and Grades*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education. Retrieved 10/2/09 from cecr.ed.gov/guides/other69Percent.pdf.
- Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417–458.

- Robelen, E. W. 2008. "Gates Revamps Its Strategy for Giving to Education." *Education Week* 28. Retrieved 10/2/09 from www.edweek.org/ew/articles/2008/11/11/13gates.h28.html.
- Schacter, J., and Y. M. Thum. 2004. "Paying for High- and Low-quality Teaching." *Economics of Education Review* 23(4): 411–430.
- Schacter, J., and Y. M. Thum. 2005. "TAPping into High-quality Teachers: Preliminary Results from the Teacher Advancement Program Comprehensive School Reform." *School Effectiveness and School Improvement* 16(3): 327–353.
- Solmon, L. C., J. T. White, D. Cohen, and D. Woo. 2007. *The Effectiveness of the Teacher Advancement Program*. Santa Monica, CA, and Washington, DC: National Institute for Excellence in Teaching. Retrieved 10/2/09 from www.tapsystem.org/publications/publications.taf?page=reports.
- Strategic Management of Human Capital. 2009. "Strategic Management of Human Capital in Public Education." Retrieved 10/2/09 from www.smhc-cpre.org/about/.
- TeacherSolutions. 2007. *Performance-pay for Teachers: Designing a System that Students Deserve*. Hillsborough, NC: Center for Teaching Quality. Retrieved 10/2/09 from www.teacherleaders.org/sites/default/files/TS2008_0.pdf.
- Toch, T., and R. Rothman. 2008. *Rush to Judgment: Teacher Evaluation in Public Education*. Washington, DC: Education Sector. Retrieved 10/2/09 from www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf.
- Toledo Federation of Teachers. 2009. "The Toledo Plan." Retrieved 10/2/09 from www.tft250.org/the_toledo_plan.htm.
- Varlas, L. 2009. "Bold Opportunities for District-union Collaboration on Teacher Quality." *Education Update* 51.
- Weisberg, D., S. Sexton, J. Mulhern, and D. Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved 10/2/09 from widgeteffect.org/downloads/TheWidgetEffect.pdf.
- Wilson, S. M., L. Darling-Hammond, and B. Berry. 2001. *A Case of Successful Teaching Policy: Connecticut's Long-term Efforts to Improve Teaching and Learning*. Seattle, WA: University of Washington, Center for the Study of Teaching and Policy. Retrieved 10/2/09 from [/depts.washington.edu/ctpmail/PDFs/Connecticut-WDHB-02-2001.pdf](http://depts.washington.edu/ctpmail/PDFs/Connecticut-WDHB-02-2001.pdf).
- Wolfe, F. 2009. "Gates Foundation to Award Grants for Evaluation Systems." *Education Daily* August 7.
- 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: 57–67.

Olivia Little is a former employee of the Educational Testing Service (ETS), 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.



Great Public Schools for Every Student

NEA Research
1201 16th Street, N.W.
Washington, D.C. 20036-3290
www.nea.org