CECR Center for Educator Compensation Reform

Uses of Multiple Measures for Performance-Based Compensation

March 2012

Alyson Burnett
Ellen Cushing
Lauren Bivona
American Institutes for Research

Burnett, A., Cushing, E., & Bivona, L. (2012). *Uses of multiple measures for performance-based compensation*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Center for Educator Compensation Reform.

The Center for Educator Compensation Reform (CECR) would like to thank the following people for their time in reviewing drafts of this resource: Jan Gahala, American Institutes for Research; Carol Karton, American Institutes for Research; Robert Stonehill, Ph.D., American Institutes for Research; Gretchen Weber, American Institutes for Research; Matt Graham, Westat; Jenna Aurand Scott, Ph.D., Westat; Carolyn Lampila, U.S. Department of Education.

The work described in this paper was supported by the U.S. Department of Education through the Center for Educator Compensation Reform. The opinions expressed are those of the authors and do not necessarily reflect the view of the U.S. Department of Education, the Center for Educator Compensation Reform, or the institutional partners of the Center. Comments and suggestions are welcome.

The Center for Educator Compensation and Reform (CECR) was awarded to Westat — in partnership with Learning Point Associates, an affiliate of American Institutes for Research, Synergy Enterprises Inc., J. Koppich and Associates and the University of Wisconsin — by the U.S. Department of Education in October 2006.

The primary purpose of CECR is to support Teacher Incentive Fund (TIF) grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources. CECR also is charged with raising national awareness of alternative and effective strategies for educator compensation through a newsletter, a Web-based clearinghouse, and other outreach activities.

This work was originally produced in whole or in part by the CECR with funds from the U.S. Department of Education under contract number ED-06-CO-0110. The content does not necessarily reflect the position or policy of CECR or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by CECR or the federal government.

CECR

Center for
Educator Compensation
Reform

Allison Henderson, Director
Phone: 888-202-1513
E-mail: cecr@westat.com

34601.0312.83670507

Uses of Multiple Measures for Performance-Based Compensation

Introduction

Across the nation, states and local school districts are seeking to redesign their educator-evaluation systems. Newly funded initiatives from the U.S. Department of Education—such as Race to the Top, School Improvement Grants, and the Teacher Incentive Fund—call for innovative teacher and principal evaluation models. These new evaluation models can inform human capital decisions regarding recruitment, retention, professional development, evaluation, and compensation. Because no single measure can identify all strengths and weaknesses of teacher practice, performance-based compensation systems should include multiple measures of performance to accurately identify areas of needed support. When thoughtfully implemented, the use of multiple measures provides a more complete picture of teacher practice, increases the legitimacy of the performance-based compensation system, and promotes stakeholder buy-in.

Background on Teacher Evaluations

For decades, teacher evaluations relied almost exclusively on principal-conducted classroom observations. Toch (2008) describes the typical teacher evaluation in public education as "a single, fleeting classroom visit by a principal or other building administrator untrained in evaluation who wields a checklist of classroom conditions and teacher behaviors that often do not focus directly on the quality of instruction" (p. 32). The checklists, according to Toch, often include items about teacher dress or room safety while ignoring aspects of teaching that actually affect student learning.

Regrettably, traditional evaluation systems do not effectively distinguish superior, good, and weak teachers. Instead, they typically use binary rating scales that provide teachers with very little feedback. In 2009, Weisberg and colleagues released *The Widget Effect*, which concluded that 99 percent of teachers receive a satisfactory rating on binary rating scales and that even on scales with three or more rating levels, less than 1 percent receive an unsatisfactory rating. For example, a 2007 study of Chicago Public Schools (CPS) found that only 0.3 percent of CPS principals identified teachers' performance as unsatisfactory. In addition, 73 percent of principals who admitted to inflating teacher-evaluation scores thought the evaluation tool did not accurately or meaningfully assess performance (The New Teacher Project, 2007).

When evaluation systems identify only the very worst teachers, they do not serve the teachers: observers fail to recognize excellent teachers; poor performers receive insufficient support and feedback; and administrators may overlook key professional development needs (Weisberg, Sexton, Mulhern, & Keeling, 2009). Such findings have created public demand for evaluations that accurately measure teacher impact on student achievement.

Increasingly in recent years, school systems have implemented observation frameworks that consist of explicit, comprehensive expectations for teachers (Milanowski, Kimball, & White, 2004). These frameworks contain detailed descriptions of distinct levels of teaching practice and professionalism. The evaluations require teachers and administrators to collect evidence of teaching practice through observation notes, lesson plans, and student work samples (Milanowski et al., 2004). Recent studies suggest that observation frameworks are valid measures of teacher performance because they correlate with value-added scores (Kane & Staiger, 2012; Sartain, Stoelinga, & Brown, 2011; Tyler, Taylor, Kane, & Wooten, 2010). In addition, unlike previous observation processes, framework-based evaluation processes encourage and facilitate sharing specific feedback with teachers.

Despite the strengths of these complex frameworks, studies suggest that building-level administrators require extensive training if they are to make nuanced distinctions between more or less effective teachers. Jacob and Lefgren (2008) compared results of principal observations to student achievement scores and found that principals were fairly adept at identifying teachers whose students make the largest and the smallest standardized achievement gains in their schools; the principals were less able to distinguish among teachers in the middle of the distribution. A more recent study found that principals reliably rated low-performing and average teachers using Charlotte Danielson's *Framework for Teaching* but tended to rate more teachers as distinguished (the highest rating) than external observers (Sartain et al., 2011).[1] Successful implementation of observation frameworks requires clear standards and extensive, ongoing rater training.

Recent policy has also focused increasingly on using student performance measures to assess teacher performance. Many states and districts have begun using standardized achievement tests in their teacher-evaluation systems. Some states and districts have developed measures, including value-added models (VAM) and student growth percentiles (SGP), that assess teacher contributions to student gains. Although these models measure an important aspect of teaching, they do not provide a complete picture of a teacher's performance.

With the limitations of student performance measures in mind, states and districts continue to improve instructional frameworks and classroom observation protocols. They are also developing other performance measures that provide a holistic understanding of teacher effectiveness. To increase observer agreement, states and districts should provide ongoing training to evaluators, often in the form of certification and calibration. Additionally, districts should combine observation scores with other measures of teacher performance.

Benefits of Using Multiple Measures

Despite the improvements in evaluation frameworks and the increased emphasis on student achievement in teacher evaluation systems, the use of multiple measures in performance-based compensation systems is preferable. First, using multiple measures can better capture the multifaceted nature of teaching. Decades of research confirm that teaching is complex and that effective teachers use myriad approaches to help students achieve. It stands to reason, then, that assessing teacher performance will require multiple measures (Goe, Bell, & Little, 2008). When appropriately combined, multiple measures provide a clearer picture of a teacher's performance. For example, a preliminary study from the Measures of Effective Teaching (MET) project found that combining teachers' observation scores, student feedback, and value-added data on state tests from a prior year created a statistically stronger indicator of effective teaching than observation scores alone (Kane & Staiger, 2012).

[1] Specifically, principals assigned at least one Distinguished rating to 52 percent of 257 teachers, but external observers assigned a Distinguished rating only to 24 percent of teachers (Sartain et al., 2011)

Another benefit of incorporating multiple measures is that it allows districts to make fair comparisons between teachers when determining compensation. (Box A provides an example of how Denver Public Schools uses multiple measures to determine teacher compensation) If evaluations consider only one performance measure, bias toward that measure will result in unfair compensation. For example, schools cannot generate value-added measures for all classroom teachers because of a lack of achievement data, but it would be unfair to make teachers of nontested subjects ineligible for performance-based compensation. Schools that use multiple measures can evaluate teachers based on a wide range of behaviors. In turn, these data can enable districts to better identify the best teachers for compensation decisions.

Box A: Denver ProComp

One method of using multiple measures for determining teacher compensation is to assign particular award amounts to each measure. Denver Public Schools uses this approach in its ProComp system. ProComp offers teacher incentives in four major categories: knowledge and skills, comprehensive professional evaluation, market incentives, and student growth. Each category corresponds to specific measures and dollar amounts. For example, in the student-growth category, teachers may receive rewards based on student growth objectives ($376), student growth percentiles ($2,403), school performance ($2,403), and/or schoolwide student growth ($2,403) (Denver Public Schools, 2011). This diversity of measures enables teachers of both tested and nontested subjects and grades to earn performance-based awards. As such, ProComp's use of multiple measures promotes fairness.

Using multiple measures can increase the quantity and quality of feedback that teachers receive. When the feedback is sufficiently detailed, teachers and administrators can use it to inform instructional and professional development decisions (Annenberg Institute for School Reform at Brown University, 2011; Goe et al., 2008). Some measures, such as classroom observations and interim student assessments, can help teachers identify their strengths and weaknesses of their practice in the short-term. Other measures—such as portfolios, end-of-course exams, and value-added scores—may not be as useful for formative purposes but may provide a longer term view of teachers' practice.

The combination of formative and summative measures can inform both short- and long-term professional growth plans. For example, classroom observations may reveal that the teacher mostly uses lower order questioning instead of engaging students in deeper thinking. Based on these results, teachers and administrators can help the teacher improve her questioning; supports might include coaching, targeted feedback on her questioning practices,

Box B: The TAP Model

The TAP model uses data from multiple measures to inform compensation, professional development, and professional growth decisions. In TAP, teachers are eligible for financial rewards based on teacher evaluations, individual classroom growth measured by VAM, and schoolwide growth measured by VAM (National Institute for Excellence in Teaching, 2012a). By using multiple measures, the TAP model underscores the importance of effective teaching and emphasizes student achievement while recognizing the collaborative nature of teaching. After observing instruction, mentor teachers provide individualized feedback to the instructor. In addition, TAP provides teachers with other multiple job-embedded professional development supports, including cluster groups, coaching, and classroom-based support (National Institute for Excellence in Teaching, 2012b). TAP also enables highly effective teachers to advance their careers and earn more money while remaining in the classroom by serving as mentor or master teachers (National Institute for Excellence in Teaching, 2012b). Thus, the TAP model combines multiple measures, performance-based compensation, and professional supports to improve teacher practice and student achievement.

lesson and unit planning that identifies key higher order questions, or collaboration with peers who use strong questioning practices.

Adding a third measurement may strengthen understanding of teacher practice. For example, student learning objectives, also sometimes referred to as student growth objectives, provide information about teacher attainment of specific goals. Unlike some other measures of student growth, student learning objectives can target subgroups or specific skills and content. Like observations, these results can draw teachers' attention to potential growth areas. For example, if all of the teacher's students met their goals for mathematics except for the English Language Learners (ELLs), the teacher could seek advice as to how she or he might better support these learners. Professional development sessions, school ELL specialists, or reading materials may help the teacher improve his or her practice. Thus, the combination of multiple measures can provide teachers with specific feedback on their instruction and can help inform future professional growth plans. For another example, see Box B about the System for Teacher Advancement (TAP) model, which combines multiple measures, performance-based compensation, and job-embedded professional development to encourage student achievement growth and improved teacher practice.

In addition to providing teachers with better feedback on their practice, the use of multiple measures can also help states and districts gain buy-in for their performance-based compensation plans. Parents, educators, teacher unions, community members, and policymakers would likely have different ideas about what aspects of teaching are most important and the best ways to evaluate educators. Many of these stakeholders would be uneasy with the idea of a system that evaluated teachers solely based on student test scores, but would likely support a system that considered several aspects of teaching that the stakeholders deemed important. Stakeholder buy-in is critical for states and districts when implementing performance-based compensation, and using multiple credible measures may encourage stakeholders to support the new systems.

Measures for Teacher Evaluation

States and districts can select from a number of performance measures to assess teaching practice, including student assessments and artifacts, classroom observations, portfolios or evidence binders, teacher essays, self-report measures, measures of student growth, and surveys. When selecting measures, states and districts should be mindful of the implications for using those measures to make high-stakes decisions, including compensation. The selection of measures would require careful consideration of the relative strengths and weaknesses of each measure and how to balance those strengths and weaknesses. For each measure, states and districts should consider the following questions:

- Is this measurement valid? (Does it measure what it is designed to measure?)
- Is this measurement reliable? (Does it produce accurate and consistent results?)
- How does this measure support the professional judgment of the evaluator?
- What training or supports will be needed to reasonably implement these measures?
- What are the costs associated with this measure? How feasible is it?
- How might assigning importance to the measure affect the measure itself (e.g., "teaching to the test" or manipulating student surveys to punish unpopular teachers)?

The following table describes numerous measurements and analytical techniques. The table has four columns: the first lists the measurement; the second column briefly describes it; the third column identifies the strengths of the measure; and the fourth column discusses some of its limitations.

Measure	Description	Strengths	Limitations
Student Assessment Measures			
State Achievement Tests	*NCLB* requires states to assess students in Grades 3 through 8 and in high school (U.S. Department of Education, 2004). One measure of teacher performance is student performance on state assessments (Miller & Scott, 2012).	State achievement tests provide a common measure of student performance across a state and allow for comparisons across teachers and schools within and across districts. These assessments often tie to state and district standards, and the test content reflects state priorities. Further, many state achievement tests have high levels of validity (they measure what they are designed to measure) and reliability (they produce consistent and accurate results). Finally, because states already use these tests for *NCLB* reporting, using state achievement tests to assess teacher performance incurs few additional costs (Goe et al., 2008).	Although student achievement tests reflect state standards, they do not capture teachers' efforts to expand student knowledge beyond what is required for grade-level proficiency. Subsequently, test scores may not capture true performance of low- or high-performing students (Little, Goe, & Bell, 2009; May et al., 2009). State achievement tests usually consist of 40 to 50 multiple choice questions; consequently, they often contain only one or two test items per assessed objective and do not assess all objectives (May et al., 2009). State achievement tests also may not adequately measure higher level thinking skills, such as analysis and evaluation (Toch, 2008). Because states do not typically test all subjects, the use of student achievement tests is limited to select subjects and grades. In addition, studies comparing the rigor of state proficiency standards against nationally normed[2] standards, such as the National Assessment of Educational Progress (NAEP), suggest that standards and assessments vary widely across states (Bandeira de Mello, 2011; Peterson & Hess, 2008). Thus, student achievement tests are not appropriate for interstate comparisons. Student performance on state assessment tests can help teachers identify the strengths and weaknesses of their performance, but do not directly evaluate instructional quality (Goe, 2010).

[2] A nationally normed test uses a common standard for the entire nation.

Measure	Description	Strengths	Limitations
Student Assessment Measures (continued)			
Standardized Achievement Tests	Standardized achievement tests are nationally normed and measure student knowledge and skills in relation to other students (Beaupré, 1995-2011; Herndon, 1980). Examples include the Terra Nova tests, Stanford Achievement Tests, Iowa Tests of Basic Skills, the NAEP exams, and Advanced Placement (AP) exams (Beaupré, 1995-2011).	Test makers typically design these tests to have high validity and reliability, and many have been evaluated in research studies (National Center on RTI, 2011). Because these tests are nationally normed, evaluators can compare test results across classrooms, districts, and states (Herndon, 1980). Unlike state achievement tests, which tend to be high-stakes assessments, many standardized achievement tests are considered low-stakes exams.	Standardized achievement tests are expensive to purchase, and administering these exams for every grade and subject can be an expense too large for most states and districts (Buckley & Marion, 2011). Further, because tests are not typically created for particular states or districts, they need to be aligned to the specific curriculum standards of the district, state, or school (Buckley & Marion, 2011).
End-of-Course Assessments	End-of-course assessments are summative assessments of student learning. States or districts often mandate them, and they correspond to the content of the course (Buckley & Marion, 2011).	End-of-course assessments serve as a way to measure student achievement in otherwise untested subjects and grades (Prince et al., 2009). Because states and districts frequently develop end-of-course assessments, states and districts can often use the results to compare educators across schools within a state or district (Buckley & Marion, 2011). Locally created end-of-course assessments can be customized to fit local standards and can align with local curricula more easily than standardized or state tests (Prince et al., 2009).	Few states and districts have the financial resources to create tests in every subject and validate them sufficiently for high-stakes accountability (Buckley & Marion, 2011). Further, when used alone, end-of-course assessments cannot measure growth (Prince et al., 2009).

Measure	Description	Strengths	Limitations
Student Assessment Measures (continued)			
School- or Teacher-Developed Assessments	A school, teacher, or group of teachers develop these tests to assess student knowledge or mastery. Examples could include an end-of-chapter test, a midterm, student portfolios, or performance tasks (Buckley & Marion, 2011).	These tests are relatively inexpensive to create and administer. They serve as a way to measure student achievement of typically untested subjects and grades. Because teachers typically develop tests for their own students, they have the flexibility to tailor the tests to the specific goals and may be more likely to support the use of student assessment for evaluation purposes (Buckley & Marion, 2011). If assessments are developed by a group of teachers teaching the same subject and content, teachers can compare results across classes. Group discussions of student performance may encourage teachers to share best practices or develop collective plans for improving student achievement. Finally, the test development process itself can improve teachers' practices through the exercise of designing and improving their assessments (CTAC, 2008).	Because individual schools or teachers develop the tests, they do not permit comparison beyond school walls. The rigor and quality of the tests are also questionable because the testing environment is unlikely to be as controlled as with standardized tests (CTAC, 2008). A teacher may use the same questions for all classes year after year, which could enable students to decrease the test's validity by circulating questions (CTAC, 2008). Also, individual schools and teachers are unlikely to have sufficient resources to develop assessments with the validity needed for high stakes accountability (Buckley & Marion, 2011). Finally, unless teachers administer well-aligned pretests, teacher-developed tests cannot measure growth.
Teacher Observation Measures			
Classroom Observation by Principal or Outside Evaluator	A principal or outside evaluator evaluates a lesson, often using a protocol or rubric, either during an informal walkthrough or a formal session (Hinchey, 2010).	A variety of stakeholders, including teachers, principals, and community members, generally consider classroom observations to be an effective measure of teacher quality (Little et al., 2009). When the observation process includes use of valid observation forms and rigorous training, observers can provide detailed information about teacher practices that can be useful for both formative and summative purposes (Goe et al., 2008). Observations can also capture information that is not included in student achievement scores, including student engagement, classroom environment, and teacher questioning techniques—all of which are important for student learning.	Valid observation forms, rigorous training and recalibration, and sufficient observation time are necessary in order for observations to be valid and reliable for high-stakes accountability (Graham, Milanowski, & Miller, 2012). Properly training evaluators and ensuring agreement between observers can be costly in time and financial resources (Graham et al., 2012). Furthermore, the time required to observe teachers, document evidence, and conduct pre- or post-observation conferences may strain the capacity of already busy school administrators (Malen et al., 2011; Sartain et al., 2011). Observations may not be able to provide teachers with formative feedback if observers are not experts in the subjects of the teachers they observe.

Measure	Description	Strengths	Limitations
Teacher Observation Measures (continued)			
Peer Review	Peer and/or master teachers conduct classroom observations. These peer teachers can be from the observed teacher's school or another school. Peer observers usually specialize in the same content area as the teachers they observe (Goldstein, 2004).	Administrators, teachers, and peer observers can benefit from the peer review process. By having expert teachers observe their peers, administrators are able to reduce their administrative burden (Goldstein, 2004). Teacher observers with subject area expertise may be able to provide peers with more detailed feedback on the lesson content than administrators (Goldstein, 2004). In return, peer observers may benefit through exposure to a variety of instructional strategies, some of which may be new to the reviewer (Weems & Rodgers, 2010). Compared to the cost of hiring external evaluators, it is generally less expensive to use existing staff. Additionally, research suggests that peer review models can be effective at pushing ineffective teachers out of the classroom (Goldstein, 2004).	Peer review, like other classroom observation methods, presents similar challenges. Successful peer review processes require valid observation forms, rigorous training, and sufficient time to observe and meet with teachers (Goe et al., 2008). Providing time to master teachers or expert teachers to observe and meet with teachers can be a costly logistical challenge. Although the peer review process can be a useful exercise for everyone involved, it can also create tensions between the peer observer and the observed teacher, especially if the observer's feedback is negative.
Other Approaches to Teacher Evaluation			
Instructional Artifacts	Using standardized protocols, raters select and evaluate specific artifacts of teachers' work, such as letters to parents, open-house handouts, student assessments, grading guidelines, lesson plans, or student work (Goe, Holdheide, & Miller, 2011).	Evaluators can use artifacts to assess teachers of all subjects and grade levels. Artifacts supplement other measures well because they assess areas of teachers' practice that may not be evident in achievement data or classroom observations. For example, artifacts can assess teachers' professionalism, instructional preparation, communication with parents, and engagement with the community. Another advantage of artifacts is that they require little additional effort on behalf of teachers because artifacts are preexisting student and teacher work products. When rubrics and protocols are valid and raters receive sufficient training, artifacts can serve as a useful indicator of instructional quality (Goe, 2011).	Although artifact collection requires teachers to collect only evidence of work they already do, evaluators need sufficient training to ensure that they score artifacts consistently and provide teachers with appropriate guidance (Goe et al., 2008). The review process is time intensive, requiring review of multiple materials and provision of feedback so that teachers can their improve practice. Studies suggest that the subjectivity of such assessments renders instructional artifact analysis insufficiently reliable for high-stakes accountability (Steele, Hamilton, & Strecher, 2010). Additionally, raters need sufficient expertise in all subject areas (Little et al., 2009). With the exception of student work, most artifacts do not demonstrate student performance or growth.

Measure	Description	Strengths	Limitations
Other Approaches to Teacher Evaluation (continued)			
Portfolios	A portfolio is a collection of artifacts that can include teacher videos, lesson plans, rationales for teaching, and teaching philosophies. Often, the portfolio process requires teachers to reflect on their practice (Tucker, Stronger, Gareis, & Beers, 2003).	Portfolio creation is a reflective process that requires teachers to think about their effectiveness and provide evidence of their practice (Danielson, 2007; Weems & Rodgers, 2010). A well-compiled portfolio provides a comprehensive overview of a teacher's practice, including aspects that achievement data or classroom observations may not demonstrate. Portfolios enable teachers to demonstrate excellence in a variety of ways, rather than using measures that focus on one aspect of teaching (Weems & Rodgers, 2010), and give teachers control over the basis of their evaluation (Goe et al., 2008; Tucker et al., 2003). Additionally, districts can use portfolios to evaluate teachers of all subjects and grade levels.	Portfolios tend to focus on teachers rather than students. Unless student work or test scores are included, portfolios cannot demonstrate student achievement or growth. The preparation and review of portfolios can be very time-consuming. Portfolios require teachers to devote a significant amount of nonteaching time to compile materials, organize them appropriately, and write rationales or reflections on their practice. Portfolios can also be unwieldy documents that are difficult and time-consuming to review; often, rater training is needed to increase reliability (Tucker et al., 2003). Teachers tend to select their most exemplary work for portfolios, thereby giving raters an unrepresentative impression of teacher practice (Goe et al., 2011). The process may also unfairly favor teachers who are more articulate or are skilled writers. As with observations and artifacts, judging portfolio quality and teacher practice is challenging (Weems & Rodgers, 2010). Developing specific rubrics and training raters is an expensive but necessary step to ensure consistent scoring. It may also be necessary to find raters with expertise in each teacher's subject area (Little et al., 2009).

Measure	Description	Strengths	Limitations
Other Approaches to Teacher Evaluation (continued)			
Student and Parent Surveys	Students and/or parents complete surveys about teachers' behaviors. Topics often include teachers' efforts to engage and challenge students (Goe et al., 2011).	Surveys provide the perspective of those most affected by teachers: their students (Goe et al., 2011). Surveys can gauge intangible aspects of teacher performance, such as perceived teacher expectations and rapport with students. These measures can provide teachers with specific feedback on how they can improve their interactions with students and parents. Because they can be administered across an entire school or district at once, surveys are often cost- and time-efficient (Goe et al., 2011). System-wide distribution enables evaluators to compare survey results across classrooms within a school or district. A recent study from the Measures of Effective Teaching (MET) project found that the relationship between observation scores and student growth grew stronger when researchers combined observations results with student surveys (Kane & Staiger, 2012). Other studies have found that high-quality survey instruments can be valid measures of teacher performance and predictors of student achievement (Peterson et al., 2000; Wilkerson, Manatt, Rogers, & Maughan, 2000).	As with most surveys, achieving high response rates can be challenging. Schools may struggle to get parents to return the surveys. Additionally, the survey process may prevent illiterate or non-English-speaking parents from completing the survey. Although surveys can provide valuable insight into aspects of teacher practice that other measures do not capture, some argue that students and parents may not be knowledgeable about the complexity of teaching (Goe et al., 2008). Peterson, Wahlquist, and Bone (2000) caution that having high survey ratings does not necessarily equate with being a good teacher. Furthermore, collecting feedback from young students may be difficult. Districts should also consider that students will likely find out when surveys affect teacher compensation or employment status. Some students may see such a survey as an opportunity to punish a teacher they dislike. Teachers may also attempt to manipulate the survey results, even if they are not present during survey administration.
Self-Report Measures	These measures can take a variety of forms. They require teachers to reflect upon and document their practice, using surveys, instructional logs, or interviews (Goe et al., 2011).	Self-report measures provide information about teacher beliefs, intentions, and expectations that other measures often do not capture. The self-report process promotes teachers' reflection on their practice because it requires teachers to document and describe their areas of strength and professional growth needs. Self-report surveys are relatively inexpensive and easy to implement because large amounts of data can be collected at once (Goe et al., 2011).	Reliability assessments of these measures produce mixed results (Goe et al., 2008). Teachers may over-report or underreport practices either intentionally or unintentionally; teachers may intentionally misreport their practices in order to receive higher ratings, or may unintentionally misreport their practices because they misperceive the correctness of their implementation (Goe et al., 2008).

Measure	Description	Strengths	Limitations
Approaches to Measuring Student Growth			
Simple Growth or Gain	Measures of simple growth or gain compute the difference between student performance on a pre-test and post-test (Miller & Scott, 2012).	Simple growth or gain scores use longitudinal measures so they are able to capture student performance over time (Miller & Scott, 2012).	These measures are only effective when at least some of the same content appears on the pre-test and post-test (Miller & Scott, 2012). In addition, simple growth or gain approaches do not account for contextual factors that may mitigate student achievement.
Value-Added Models (VAMs)	VAMs use previous student test data to predict students' level of achievement in the next school year. Models use these data to determine a particular teacher's effect on student growth. A variety of VAMs exist. Most are regression or analysis of variance (ANOVA) based, and many consider student and school characteristics (Miller & Scott, 2012).	Because these measures often control for some school and non-school factors that may affect student achievement, VAMs provide more valid comparisons of student outcomes than student achievement measures (Miller & Scott, 2012). VAMs focus on growth rather than proficiency and thus do not penalize teachers for working with students below proficiency levels (Holdheide et al., 2012). For these reasons, many supporters of VAMs perceive them to be more objective than other measures (Little et al., 2009). Rivken (2007) argues that principals who are knowledgeable about their schools can contextualize VAM results and use them to make informed decisions about the teachers and instruction of their schools. VAMs can also be useful when looking at larger patterns, such as the distribution of "effective" teachers across schools (Goe, 2008).	Given the complex formulas used in VAMs, teachers and other stakeholders may have difficulty understanding how VAMs assess teacher performance. VAMs are unreliable when teachers have small class sizes or a high percentage of students with missing records (Milanowski, 2011). They also tend to be highly unstable from year to year due to measurement error (Goldhaber & Hansen, 2010; Meyer & Dokumaci, 2010; Newton, Darling-Hammong, Haertel, & Thomas, 2010; Steele et al., 2010). VAMs also present design issues. First, the quality of VAMs is dependent upon the quality of the measures incorporated into the model, since VAMs use tests as proxies of student achievement (Rivkin, 2007); measurement error, test coverage of topics, and a lack of construct validity can all affect VAMs. (Braun, 2010; Meyer & Dokumaci, 2010; Rivken, 2007). Though VAMs control for some variables, they may not account for all factors that may affect student achievement (i.e., the effects of a recent parent divorce). In addition, the collaborative nature of teaching may make it difficult to disentangle one teacher's contributions to student learning from another's (Valli, Croninger, & Walters, 2007). In addition, tutors, support staff, mentors, and other teachers can all contribute to students' learning (Steele et al., 2010). When using VAM, multiple decision rules decide how student growth will be attributed to teachers. Additionally, VAMs can only assess teachers of tested grades and subjects. As with student achievement measures, VAMs do not document teachers' use of effective or ineffective practices.

Measure	Description	Strengths	Limitations
Approaches to Measuring Student Growth (continued)			
Schoolwide VAM	Schoolwide VAMs use student achievement data to estimate the contributions of teachers to student academic growth. Unlike individual value-added measures, schoolwide VAMs measure student growth at the school level (Holdheide et al., 2012; Miller & Scott, 2012).	This model allows teachers of nontested subjects and grades to be accountable for student performance and can encourage elective teachers to incorporate tested subjects (reading and math) into instruction. Schoolwide growth models also have the potential to increase teacher collaboration for the good of all students (Holdheide et al., 2012).	Schoolwide models do not directly measure the effectiveness of individual teachers (Buckley & Marion, 2011). Teachers of nontested subjects and grades may believe that they have less ability to contribute to outcomes, which may decrease their buy-in and satisfaction with the evaluation system. Conversely, schoolwide VAM models may present a free-rider problem, wherein teachers of nontested subjects receive evaluations based on the efforts of their peers rather than their own efforts (Lavy, 2007). In addition, schoolwide VAM can devalue nontested subjects and grades that are not included in the model (Holdheide et al., 2012).
Student Growth Percentiles (SGPs)	Like VAMs, SGPs use previous student test data to determine teacher contributions to student learning. With SGPs, each student's percentile rank is calculated from one year to the next to determine student growth associated with a particular teacher (Marion & Buckley, 2011).	As with VAMs, SGPs evaluate an educator's contribution to student learning. They allow districts and states to compare test score growth across groups of students with similar test histories in the same grade and subject (Miller & Scott, 2012).	VAMs and SGPs share many limitations. SGPs do not describe teacher use of effective or ineffective practices and are only useful in assessing teachers of tested grades and subjects. These measures cannot attribute the cause of student gain to teachers because schools do not randomly assign students to teachers (Steele et al., 2010). Like VAMs, SGPs attribute growth to one teacher when multiple teachers could be involved in students' learning—support staff, tutors, mentors, student teachers, and so forth (Steele et al., 2010).
Student Learning Objectives (SLOs)	SLOs are goals set by a teacher or group of teachers that specify specific learning targets for students. These targets include what students will know or be able to perform after completing a quarter, semester, or school year (Miller & Scott, 2012).	SLOs are highly flexible and can be set for students of any grade level or subject. Unlike some other measures, all teachers can create SLOs because they do not depend on the availability of standardized achievement tests. Though teachers are not required to standardize test scores when creating SLOs, the SLO process requires teachers to analyze trend and baseline data to set rigorous yet achievable targets. The SLO process encourages teachers to reflect on their practice and their students' outcomes (CTAC, 2008). Additionally, the process provides teachers the opportunity to give input on how schools measure student learning, which may increase teacher support for a performance-based evaluation system (Buckley & Marion, 2011).	SLOs may not be comparable across classrooms if the process is not standardized or objective (Holdheide et al., 2012). In addition, attaching SLOs to high-stakes decisions could incentivize teachers to set easily obtainable goals (Holdheide et al., 2012). Implementing SLOs requires significant time and resources at both the district and school levels. Teachers need training on how to set appropriate growth targets, interpret student data, identify trends, and adjust instruction. They also need time to complete the process. Principals need professional development so that they know how to ensure that SLOs are comparable, rigorous, and realistic. They also need sufficient training to ensure that they evaluate teacher performance in a systematic and fair way. In addition, principals must make time to review each teacher's SLOs, and district or state personnel must monitor SLO quality across schools (Holdheide et al., 2012).

Conclusion

As states and districts reform their evaluation systems, and, in many instances, tie performance directly to compensation, they need to evaluate and select from a menu of performance measures. There are tradeoffs associated with each measure. Because no single measure adequately captures the complexity of teaching, evaluation systems should include multiple measures of teacher effectiveness. Additionally, the mix of measures should align to the evaluation's purpose. A tight fit between measures and purposes can result in a more comprehensive and fair performance-based evaluation system that leads to greater buy-in among teachers, principals, and other stakeholders. When carefully selected and properly implemented, use of multiple measures of teacher performance can enhance performance-based development systems.

References

Annenberg Institute for School Reform at Brown University. (2011). *Straight talk on teaching quality: Six game-changing ideas and what to do about them.* Providence, RI: Author. Retrieved March 5, 2012, from *http://annenberginstitute.org/VUE/wp-content/pdf/StraightTalk.pdf*

Bandeira de Mello, V. (2011). *Mapping state proficiency standards onto the NAEP scales: Variation and change in state standards for reading and mathematics, 2005–2009* (NCES 2011-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC: Government Printing Office.

Beaupré, D. (1995-2011). *Testing our schools: A guide for parents*. PBS Frontline. [Website]. Retrieved March 8, 2012 from *http://www.pbs.org/wgbh/pages/frontline/shows/schools/etc/guide.html*

Braun, H. (2010). *Issues in measuring student growth and conducting productivity analyses.* Austin, TX: Educational Testing Service, Center for K–12 Assessment and Performance Management. Retrieved March 5, 2012, from *http://www.k12center.org/rsc/pdf/BraunPresenterSession2.pdf*

Buckley, K., & Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February 21, 2012, from *http://colegacy.org/news/wp-content/uploads/2011/10/Summary-of-Approaches-for-non-tested-grades_7-26-11.pdf*

Community Training and Assistance Center (CTAC). (2008). *Tying earning to learning: The link between teacher compensation and student learning objectives*. Boston, MA: Author. Retrieved February 21, 2012, from *http://www.ctacusa.com/PDFs/Rpt-TyingEarningtoLearning-2008.pdf*

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: ASCD.

Denver Public Schools. (2011). *About ProComp* [Website]. Retrieved February 21, 2012, from *http://denverprocomp.dpsk12.org/about/*

Goe, L. (2008). *Using value-added models to identify and support highly effective teachers* (Key Issue). Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 21, 2012, from *http://www2.tqsource.org/strategies/het/UsingValueAddedModels.pdf*

Goe, L. (2010). *Evaluating teaching with multiple measures.* Washington, DC: American Federation of Teachers. Retrieved February 21, 2012, from *http://www.lauragoe.com/LauraGoe/EvalTchgWithMultipleMeasures.pdf*

Goe, L., Bell, C., & Little, O. (June 2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 21, 2012, from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Goe, L., Holdheide, L., & Miller, T. (2011). *A practical guide to designing comprehensive teacher evaluation systems.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 21, 2012, from http://www.tqsource.org/publications/practicalGuideEvalSystems.pdf

Goldhaber, D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance* (CEDR working Paper 2010–3). Seattle: University of Washington, Center for Education Data and Research. Retrieved February 21, 2012, from http://cedr.us/papers/working/CEDR%20WP%202010-3_Bad%20Class%20Stability%20%288-23-10%29.pdf

Goldstein, J. (2004). Making sense of distributed leadership: The case of peer assistance and review. *Educational Evaluation and Policy Analysis, 26*(2), 173–197.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*. Washington, DC: Center for Educator Compensation Reform. Retrieved March 23, 2012, from http://cecr.ed.gov/pdfs/Inter_Rater.pdf.

Herndon, E. B. (1980). *Your child and testing.* Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED195579). Retrieved March 5, 2012, from http://www.eric.ed.gov/PDFS/ED195579.pdf

Hinchey, P. H. (2010). *Getting teacher assessment right: What policymakers can learn from research.* Boulder: University of Colorado, National Education Policy Center. Retrieved March 5, 2012, from http://nepc.colorado.edu/files/PB-TEval-Hinchey_0.pdf

Holdheide, L., Browder, D., Warren, S., Buzick, H., & Jones, N. (January 2012). *Summary of using student growth to evaluate educators of students with disabilities: Issues, challenges, and next steps.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 21, 2012, from http://isbe.net/peac/pdf/using_student_growth_summary0112.pdf

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective evaluation in education. *Journal of Labor Economics, 26*(1), 101–136.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved February 21, 2012, from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Lavy, V. (2007). Using performance-based pay to improve the quality of teachers. *The Future of Children, 17*(1), 87–109.

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 21, 2012, from http://tqsource.org/publications/practicalGuide.pdf

Malen, B., Rice, J. K., Jackson, C., Hoyer, K. H., Hyde, L., Bivona, L., et al. (2011). *Implementation, payouts, and perceived effects: A formative analysis of Financial Incentive Rewards for Supervisors and Teachers (FIRST).* Prince George’s County, MD: Prince George’s County Public School System.

Marion, S., & Buckley, K. (2011). *Approaches and considerations for incorporating student performance results from ’non-tested‘ grades and subjects into educator effectiveness determinations.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February 21, 2012, from http://colegacy.org/news/wp-content/uploads/2011/10/Marion-Buckley_Considerations-for-non-tested-grades_2011.pdf

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009–013). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance Retrieved February 21, 2012, from http://ies.ed.gov/ncee/pdf/2009013.pdf

Meyer, R. H., & Dokumaci, E. (2010). *Value-added models and the next generation of assessments*. Austin, TX: Educational Testing Service, Center for K–12 Assessment and Performance Management. Retrieved March 5, 2012, from http://www.k12center.org/rsc/pdf/MeyerDokumaciPresenterSession4.pdf

Milanowski, A. (2011). *Resolving some issues in using value-added measures of productivity for school and teacher incentives: Ideas from technical assistance and TIF grantee experience*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Center for Educator Compensation Reform. Retrieved March 5, 2012, from http://cecr.ed.gov/ pdfs/CECR_HP_ValueAdded.pdf

Milanowski, A., Kimball, S., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites*. Madison, WI: Consortium for Policy Research in Education.

Miller, J. & Scott, J. (2012). *Understanding the Basics of Measuring Student Achievement*. Washington, DC: Center for Educator Compensation Reform. Retrieved March 12, 2012, from http://cecr.ed.gov/pdfs/Understanding_Basics.pdf

National Center on RTI (2011). *Progress monitoring tools* [Website]. Retrieved February 21, 2012, from http://www.rti4success.org/ progressMonitoringTools

National Institute for Excellence in Teaching. (2012a). *TAP elements of success: Ongoing applied professional growth* [Website]. Retrieved February 21, 2012, from http://www.tapsystem.org/action/action.taf?page=oapg

National Institute for Excellence in Teaching. (2012b). *TAP elements of success: Performance-based compensation* [Website]. Retrieved February 21, 2012, from http://www.tapsystem.org/action/action.taf?page=oapg

New Teacher Project, The. (2007). *Hiring, assignment, and transfer in Chicago Public Schools* [Presentation]. Brooklyn, NY: Author. Retrieved February 21, 2012, from http://tntp.org/assets/documents/TNTPAnalysis-Chicago.pdf?files/ TNTPAnalysis-Chicago.pdf

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23). Retrieved February 21, 2012, from http://epaa.asu.edu/ojs/article/view/810/858

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 135–153.

Peterson, P. E., & Hess, F. M. (2008). *Few states set world-class standards*. Education Next. Retrieved February 21, 2012, from http://educationnext.org/files/ednext_20083_70.pdf

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Center for Educator Compensation Reform. Retrieved March 5, 2012, from http://cecr.ed.gov/guides/other69Percent.pdf

Rivkin, S. G. (2007). *Value-added analysis and education policy* (Brief 1). Washington, DC: Urban Institute, National Center for Analysis of Longitudinal Data in Education Research (CALDER). Retrieved March 5, 2012, from *http://urbaninstitute.org/UploadedPDF/411577_value-added_analysis.pdf*

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation.* Chicago: Consortium on Chicago School Research: University of Chicago Urban Education Institute. Retrieved February 21, 2012, from *http://ccsr.uchicago.edu/publications/Teacher%20Eval%20Report%20FINAL.pdf*

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems.* Santa Monica, CA: RAND. Retrieved February 21, 2012, from *http://www.rand.org/pubs/technical_reports/TR917.html*

Toch, T. (2008). Fixing teacher evaluation: Evaluations pay large dividends when they improve teaching practices. *Education Leadership, 66*(2), 32–37.

Tucker, P. D., Stronger, J. H., Gareis, C. R., & Beers, C. S. (2003). The efficacy of portfolios for teacher evaluation and professional development: Do they make a difference? *Educational Administration Quarterly, 39*(5), 572–602.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review: Papers and Proceedings, 100*, 256–260.

U.S. Department of Education. (2004). *Testing: Frequently asked questions* [Website]. Retrieved March 5, 2012, from *http://www2.ed.gov/nclb/accountability/ayp/testing-faq.html*

Valli, L., Croninger, R. G., & Walters, K. (2007). Who (else) is the teacher? Cautionary notes on teacher accountability systems. *American Journal of Education, 113*, 635–662.

Weems, D. M. & Rodgers, C. B. H. (2010). Are U.S. teachers making the grade? A proposed framework for teacher evaluation and professional growth. *Management in Education, 24*(1), 19–24.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Brooklyn, NY: The New Teacher Project. Retrieved February 21, 2012, from *http://widgeteffect.org/downloads/TheWidgetEffect.pdf*

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360° feedback™ for teacher evaluation. *Journal of Personnel and Evaluation, 14*(2), 179–192.