A PRACTICAL GUIDE to Evaluating Teacher Effectiveness

April 2009

Olivia Little, ETS
Laura Goe, Ph.D., ETS
Courtney Bell, Ph.D., ETS

LEARNING POINT Associates®

ETS®

VANDERBILT

Contents

Introduction . . . 1
Defining Teacher Effectiveness . . . 1
Methods of Evaluating Teacher Effectiveness . . . 4
Valued-Added Models . . . 4
Classroom Observation . . . 6
Principal Evaluation . . . 8
Analysis of Classroom Artifacts . . . 9
Portfolios . . . 10
Self-Report of Practice . . . 11
Student Evaluation . . . 13
Creating a Comprehensive Teacher Evaluation System . . . 15
Recommendations . . . 17
References . . . 18
Appendix A. Validity and Measurement Terms . . . 23
Appendix B. Planning Guide . . . 25
Appendix C. Summary of Measures . . . 26
Appendix D. Sample of Existing Evaluation Systems . . . 30

Approaches to Evaluating Teacher Effectiveness: A Foundation for This Guide

This guide is based on *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis* (Goe, Bell, & Little, 2008). Articles for the research synthesis were identified through extensive Internet and library searches of keywords and phrases related to the topics of teacher effectiveness and measuring teacher performance from the last six to eight years. Additional articles, including older, seminal, nonempirical, and/or theoretical pieces were identified from broader Internet searches, reference lists of related articles, and recommendations of experts in the field.

Data Collection and Methods

The studies that were evaluated met the following criteria: (1) they were empirical, peer-reviewed journal articles; (2) they were published in English in the United States, Canada, Great Britain, Ireland, Australia, or New Zealand; (3) they addressed the K–12 student population and measured inservice teachers; (4) they included a measure of teacher effectiveness or classroom practice and included a student outcome measure *or* had implications for teacher effectiveness; and (5) they reported methods meeting accepted standards for quality research (e.g., reliable and validated instruments, appropriate study design, and necessary controls). The resulting synthesis includes approximately 120 studies that were thoroughly reviewed. The research synthesis focused primarily on studies measuring classroom processes and student outcomes, paying particular attention to studies that used value-added measures of teacher effectiveness.

The authors did not examine more indirect measures of teaching (e.g., teacher demonstrations of knowledge, teacher responses to theoretical teaching situations or structured vignettes, or parent satisfaction surveys). Instead, the synthesis focused on measures that more directly assess the processes and activities occurring during instruction and products that are created inside the classroom. The research synthesis excluded research on school effects, the effectiveness of curriculum or professional development implementations (unless research included measures specific to teachers), and other evaluations of educational interventions or programming. In addition, the research synthesis did not consider the research linking credentials, experience, or knowledge to teacher effectiveness, as this topic has been extensively reviewed (Goe, 2007). Though these are all important and related topics, they were beyond the scope of the research synthesis by Goe et al. (2008).

INTRODUCTION

There is increased consensus that highly qualified and effective teachers are necessary to improve student performance, and there is growing interest in identifying individual teachers' impact on student achievement. The No Child Left Behind (NCLB) Act mandates that all teachers should be highly qualified, and by the federal definition, most teachers now meet this requirement. However, it is increasingly clear that "highly qualified"—having the necessary qualifications and certifications—does not necessarily predict "highly effective" teaching—teaching that improves student learning. The question remains: What makes a teacher highly effective, and how can we measure it?

There are many different conceptions of teacher effectiveness, and defining it is complex and sometimes generates controversy. Teacher effectiveness is often defined as the ability to produce gains in student achievement scores. This prevailing concept of teacher effectiveness is far too narrow, and this guide presents an expanded view of what constitutes teacher effectiveness. The guide outlines the methods available to measure teacher effectiveness and discusses the utility of these methods for addressing specific aspects of teaching. Those charged with the task of identifying measures of teacher effectiveness are encouraged to carefully consider which aspects are most important to their context—whether national, state, or local. In addition, the guide offers recommendations for improving teacher evaluation systems. The conclusion indicates that a well-conceived system should combine approaches to gain the most complete understanding of teaching and that administrators and teachers should work together to create a system that supports teachers as well as evaluates them.

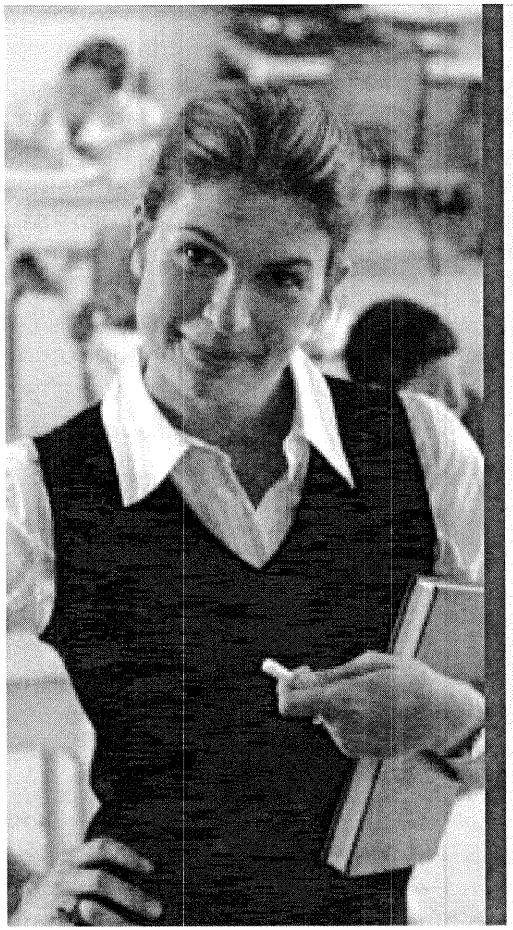
Defining Teacher Effectiveness

The way teacher effectiveness is defined impacts how it is conceived and measured and influences the development of education policy. *Teacher effectiveness*, in the narrowest sense, refers to a teacher's ability to improve student learning as measured by student gains on standardized achievement tests. Although this is one important aspect of teaching ability, it is not a comprehensive and robust view of teacher effectiveness.

There are several problems with defining teacher effectiveness solely in this way:

- **Teachers are not exclusively responsible for students' learning.** An individual teacher can make a huge impact; however, student learning cannot reasonably be attributed to the activities of just one teacher—it is influenced by a host of different factors. Other teachers, peers, family, home environment, school resources, community support, leadership, and school climate all play a role in how students learn.
- **Consensus should drive research, not measurement innovations.** Trends in measurement can be influenced by the development of new instruments and technologies. This is referred to as "the rule of the tool": if a person only has a hammer, suddenly every problem looks like a nail (Mintzberg, 1989). It is possible that the increase in data linking student achievement to individual teachers and new statistical techniques to analyze these data are contributing to an emphasis on measuring teacher effectiveness using student achievement gains (Drury & Doran, 2003; Hershberg, Simon, & Lea-Kruger, 2004; The Teaching Commission, 2004). This, in turn, may result in a narrowed definition of teacher effectiveness. Instead, important aspects and outcomes of teaching should be defined first; then, methods should be used or created to measure what has been identified. In other words, define the problem; then choose the tools.
- **Test scores are limited in the information they can provide.** Information is not available for some nontested subjects and certain student populations. Furthermore, basing teacher effectiveness on student achievement fails to account for other important student outcomes. Student achievement gains do not indicate how successful a teacher is at keeping at-risk students in school or providing a caring environment where diversity is valued. This method does not provide any additional information on student learning growth beyond the data gleaned through standardized testing. Standardized testing cannot provide information about those who teach early elementary school, special education, or untested subjects (e.g., art and music). It cannot evaluate the effectiveness of teachers who coteach and does not capture teachers' out-of-classroom contributions to making the school or district more effective as a whole.
- **Learning is more than average achievement gains.** Prominent researchers have promoted the idea that definitions of teacher effectiveness should encompass student social development in addition to formal academic goals (Brophy & Good, 1986; Campbell, Kyriakides, Muijs, & Robinson, 2004). Improving student attitudes, motivation, and confidence also contributes to learning. If the concept of effective teaching is limited to student achievement gains, differentiating between these factors becomes impossible. Was a teacher deemed effective because she focused class time narrowly on test-taking skills and test preparation activities? Or did the student achievement growth in her class result from inspired, competent teaching of a broad, rich curriculum that engaged students, motivated their learning, and prepared them for continued success? Teacher evaluations should be able to distinguish the former approach from the latter.

Given these critiques, a broader and more comprehensive definition of teacher effectiveness is necessary. The following five-point definition from Goe, Bell, & Little (2008, p. 8) is intended to focus measurement efforts on multiple components of teacher effectiveness. It is not proposed as a criticism of other useful definitions but as a means of clarifying priorities for measuring teaching effectiveness.



A Five-Point Definition of Teacher Effectiveness

Approaches to Evaluating Teacher Effectiveness: A Research Synthesis presents a five-point definition of teacher effectiveness developed through an analysis of research, policy, and standards that addressed teacher effectiveness. After the definition had been developed, the authors consulted a number of experts and strengthened the definition based on their feedback.

"The five-point definition of teacher effectiveness consists of the following:

- Effective teachers have high expectations for all students and help students learn, as measured by value-added or other test-based growth measures, or by alternative measures.
- Effective teachers contribute to positive academic, attitudinal, and social outcomes for students such as regular attendance, on-time promotion to the next grade, on-time graduation, self-efficacy, and cooperative behavior.
- Effective teachers use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed; and evaluate learning using multiple sources of evidence.
- Effective teachers contribute to the development of classrooms and schools that value diversity and civic-mindedness.
- Effective teachers collaborate with other teachers, administrators, parents, and education professionals to ensure student success, particularly the success of students with special needs and those at high risk for failure" (Goe et al., 2008, p. 8).

Methods of Evaluating Teacher Effectiveness

Given this broadened definition of teacher effectiveness, several methods to evaluate teaching and its many dimensions are presented in this section. Research findings on each method are discussed along with associated validity and measurement issues and the considerations to take into account when adopting a method for specific purposes. Two of the most widely used measures of teacher effectiveness—value-added models and classroom observations—are discussed. Then, other methods—principal evaluations, analyses of classroom artifacts, portfolios, self-reports of practice, and student evaluations—are examined. Appendix A offers a listing of validity and measurement terms used throughout this guide. Appendix B presents a planning guide for determining evaluation resources, design, and measures. Appendix C includes brief summaries of each measure presented. All possible teaching measures are not covered, but a more comprehensive list of instruments and studies can be found in *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis* (Goe et al., 2008).

Value-Added Models

Definition

Value-added models provide a summary score of the contribution of various factors toward growth in student achievement (Goldhaber & Anthony, 2004). The statistical models are complex, but the underlying assumptions are straightforward: students' prior achievement on standardized tests can be used to predict their achievement in a specific subject the next year. When most students in a particular classroom perform better than predicted on standardized achievement tests, the teacher is credited with being effective, but when most of his or her students perform worse than predicted, the teacher may be deemed less effective. Some models take into account only students' prior achievement scores; others include student characteristics (e.g., gender, race, and socioeconomic background); and still others include information about teachers' experience.

Value-added models are relatively new measures of teacher effectiveness, and supporters of their use (e.g., Hershberg et al., 2004; Sanders, 2000) argue that they provide an objective means of determining which teachers are successful at improving student learning. It is possible for teachers who are evaluated using classroom observations or other teaching measures to receive a high score but still have students with average or below-average achievement growth; however, value-added models directly assess how well teachers promote student achievement as measured by gains on standardized tests. Other researchers argue that these models are not yet fully understood and are theoretically and statistically problematic.

Research

Several studies compare teacher effectiveness as measured by value-added scores to effectiveness measured in other ways, such as observation of teaching practices, qualifications, or personal characteristics. The relationship between teaching practices and value-added scores depends on the observation instrument used and the evaluators' level of training (Heneman, Milanowski, Kimball, & Odden, 2006; Holtzapple, 2003; Kimball, White, Milanowski, & Borman, 2004). The studies that correlate value-added scores with teacher qualifications and characteristics produce mixed results, and most teacher quality variables do not show a strong ability to predict student achievement gains, with a few notable exceptions (see Goe, 2007, for a full review). Not enough research has been conducted to determine exactly which teacher behaviors or qualities value-added measures reflect.

Consider the following two examples:

- Rivkin, Hanushek, and Kain (2005) examined the relationship between value-added scores and observable teacher characteristics (e.g., education and experience) and concluded that the majority of teacher effectiveness could not be explained by observable characteristics. The study showed that teachers varied in their contribution to student achievement gains, but it did not reveal what caused the variation. This highlights a key problem with value-added measures: alone they do not provide an understanding of what effective teachers do that makes them effective.
- Schacter, Thum, and Zifkin (2006) considered whether teachers fostered student creativity in their classrooms and correlated observation scores with value-added achievement scores. They found that when teachers employed strategies to encourage student creativity, the result was improved student achievement. This study illustrates another key point: high-quality observational data combined with a high-quality value-added model can provide useful information about teaching that might lead to strategies for improving student outcomes. Value-added models may have great potential for improving instruction when combined with other measures, but additional research is needed to understand how to sort out which practices or constellations of practices lead to learning.

Considerations

Value-added models have several advantages. They directly examine how a teacher contributes to student learning and are considered highly objective by some because they do not involve raters making subjective judgments. They are generally cost-efficient and nonintrusive; they require no classroom visits, and test score data are already collected for NCLB purposes. They can reveal variation among teachers in their contributions to student learning and may be particularly useful in identifying teachers who fall at the top and bottom of that continuum. New or struggling teachers could benefit by observing teachers who are consistently deemed highly effective. Establishing these teachers' classrooms as "learning labs" for colleagues and researchers may provide valuable information about what practices and processes contribute to student achievement gains. Teachers consistently deemed less effective could be provided with help and support.

However, value-added scores must be interpreted with caution. Teachers vary greatly in their value-added scores, even within schools, but that variation has not been consistently and strongly linked to what teachers *do* in their classrooms. It may be that the classroom observation instruments typically used are not sensitive enough to capture the differences that influence student achievement, or it may be that value-added scores are measuring other elements of teaching that have not yet been conceptualized.

Several issues exist regarding the assumptions of value-added modeling. For example, there is much uncertainty in the statistical estimates for individual teachers (McCaffrey, Koretz, Lockwood, & Hamilton, 2004). Furthermore, value-added models focus only on data from standardized tests, which means they assume student test scores are valid, reliable indicators of student learning. Shavelson, Webb, and Burstein (1986) argue that linking teaching behaviors directly to student achievement outcomes can be problematic for several reasons: (1) it assumes that standardized tests are perfectly aligned with local curriculum when this is seldom the case; (2) it assumes that scores reflect improvements in students' cognition and capacity for understanding when summary scores from standardized tests do not adequately reflect this; (3) it assumes students' test performance is equated with their knowledge of the subject, even though their performance may be affected by other influences such as motivation, test-taking strategies, and attitudes toward testing; and (4) it averages test scores across all students in a classroom, ignoring differential learning, or a teacher's ability to target instruction to individual students' needs.

As stated in the National Research Council report on high-stakes testing, "Accountability for educational outcomes should be a shared responsibility of states, school districts, public officials, educators, parents, and students" (Heubert & Hauser, 1999, p. 3). Measuring teacher effectiveness through value-added models assumes that teachers are solely accountable for student achievement, rather than considering other influences (e.g., schools, families, or peers) that also contribute to student outcomes. Furthermore, certain methodological problems (e.g., incomplete student test score data and nonrandom assignment of students to teachers) threaten the validity of value-added models (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Teachers are not randomly assigned to schools, and students are not randomly assigned to teachers, meaning that the model cannot differentiate between how much student achievement growth is attributable solely to teachers and how much is attributable to other factors. Current models are not equipped to fully deal with problems such as missing data and nonrandom assignment (Rothstein, 2008a, 2008b). Given these many caveats, reliance on value-added measures as a primary means of evaluating teacher effectiveness may be premature.

Classroom Observation

Definition

Classroom observations are the most common form of teacher evaluation and vary widely in how they are conducted and what they evaluate. Observations can be created by the district or purchased as a product. They can be conducted by a school administrator or an outside evaluator. They can measure general teaching practices or subject-specific techniques. They can be formally scheduled or unannounced and can occur once or several times per year. The type of observation method adopted, its focus, and its frequency should depend on what the administration would like to learn from the process.

When measuring teacher effectiveness through classroom observations, valid and appropriate instruments are crucial. Equally important are well-trained and calibrated observers to utilize those instruments in standard ways so that results will be comparable across classrooms. Observations can provide significant, useful information about a teacher's practice if used thoughtfully, but districts must take great care to administer them in ways that minimize rater bias and other measurement concerns.

Research

Two observation protocols that are widely used and have been studied on a relatively large scale are Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (1996) and the University of Virginia's Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, & Hamre, 2006a). Both protocols are general across subject matter. The *Framework for Teaching* is general across grade levels, whereas CLASS has particular grade spans (i.e., early childhood, Grades K–5, and Grades 6–12). Both protocols also have formal procedures for training raters and establishing reliable scoring.

Framework for Teaching. Adaptations of the *Framework for Teaching* have been implemented and studied in several sites across the country and used for both formative and summative purposes. Most teachers found the framework credible and helpful for their teaching. *Framework for Teaching* scores were related to important outcomes such as student achievement, but the effects were modest and varied across the different sites (Gallagher, 2004; Heneman et al., 2006; Kimball et al., 2004; Milanowski, 2004). This variance may be caused by the modifications across sites, and it is still unclear whether adaptations of the *Framework for Teaching* work as well as the original version.

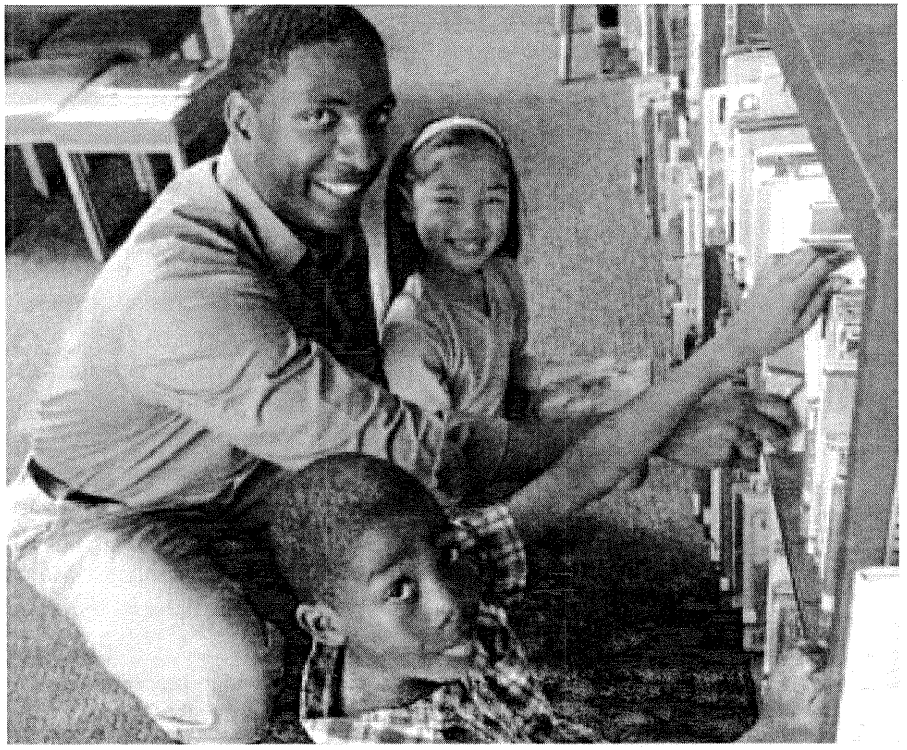
CLASS. CLASS was first developed to assess classroom quality in preschool and early elementary school. It is based on theories of child development and focuses on interactions between students and teachers (Pianta, La Paro, & Hamre, 2006b). In urban, rural, and suburban classrooms across the country, studies have found promising validity and reliability results for the prekindergarten and Grades K–5 versions of CLASS. CLASS ratings were relatively stable across the school year and correlated with academic gains, improved student behavior, and other developmental markers (Hamre & Pianta, 2005; Howes et al., 2008; Rimm-Kaufman, La Paro, Downer, & Pianta, 2005). However, there is little information on the Grades 6–12 version of CLASS, and more research on its validity and reliability is needed. Additional research also is needed to verify how CLASS functions in practice (e.g., whether districts find it affordable or feasible to keep raters trained at reliable and calibrated levels).

Other Observation Protocols. There are also numerous observation protocols that are less widely used and studied, some of which were created for a limited context. Among these more narrowly used instruments are several promising subject-specific protocols. Examples include the Reformed Teaching Observation Protocol for mathematics and science (Piburn & Sawada, 2000), the Quality of Mathematics in Instruction for mathematics (Blunk, 2007), and the TEX-IN3 for literacy (Hoffman, Sailors, Duffy, & Beretvas, 2004). Though these instruments are regarded as promising, they have not yet been widely used or studied by anyone but the developers. For practitioners interested in modifying generic protocols to include more subject matter, these observation protocols would be excellent resources. They also might be useful for districts interested in using subject-specific protocols for formative feedback.

Considerations

A main strength of formal observation protocols is that they are often perceived as credible by multiple stakeholders. Observations are considered the most direct way to measure teaching practice because the evaluator can see the full dynamic of the classroom. They have been modestly to moderately linked to student achievement, depending on the instrument. Observations have been used both formatively and summatively, suggesting that the same instrument can serve multiple purposes for districts.

However, many protocols have not been used or studied by anyone but the developers and need to undergo more independent study. More work is needed to link scores on well-validated observation protocols with student achievement and other student outcomes of interest, such as graduation and citizenship. Rater reliability is also a key concern, although progress has been made in developing methods to train and calibrate evaluators to ensure more consistent ratings. There is no assurance that a given state or district actually employs these methods, however, meaning that different evaluators might give very different scores to the same teacher depending on their views of effective teaching. In this case, measures of teacher effectiveness through observations can fluctuate, threatening the utility and credibility of the protocols themselves. Thus, when using observations, care should be taken to select validated instruments and properly train and calibrate raters in order to obtain the most accurate results.



Principal Evaluation

Definition

One of the most common forms of teacher evaluation is principal or vice-principal classroom observations (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007). Principal evaluation can vary widely by district—from a formal process using validated observation instruments for both formative and summative purposes (Heneman et al., 2006) to an informal, unannounced, or infrequent classroom visit to develop a quick impression of what a teacher is doing in the classroom. Whenever an evaluation involves classroom observation, the concerns raised in the previous subsection apply. In this subsection, principal evaluation is considered a special case of classroom observation, and some of its distinct issues are detailed.

Principal evaluations differ from those performed by district personnel, researchers, or other outside evaluators who are hired and trained to conduct evaluations. Principals are most knowledgeable about the context of their schools and their student and teacher populations, but they may not be well trained in methods of evaluation. They may employ evaluation techniques that serve multiple purposes: to provide summative scores for accountability purposes, inform decisions about tenure or dismissal, identify teachers in need of remediation, or provide formative feedback to improve teachers' practice. Although these factors can make principals a valuable source of information about their schools and teachers, they also have the potential to introduce bias in either direction to principals' interpretation of teaching behaviors.

Research

Because principal evaluation procedures vary so much by district, little research exists on their overall validity. A recent study (Brandt et al., 2007) considers evaluation policies in several Midwestern districts, finding that principals and administrators typically conduct evaluations. Most of the evaluations considered in the study were summative (for high-stakes employment decisions) rather than formative (for helping teachers grow in the profession). Districts were more likely to offer guidance on the process of conducting evaluations rather than on the appropriate application of the evaluation results. Of greatest concern, only 8 percent of districts mentioned evaluator training as a component of their teacher evaluation systems (Brandt et al., 2007, p. 6). So although most evaluations were being used for high-stakes, summative purposes, there was little evidence that they were being used in a reliable and valid manner.

Other studies have examined subjective principal ratings of teachers compared to value-added scores of student achievement (Harris & Sass, 2007; Jacob & Lefgren, 2005, 2008; Medley & Coker, 1987; Wilkerson, Manatt, Rogers, & Maughan, 2000). In these studies, principals rated teachers in their school using a researcher-created instrument. These ratings were not based on a specific observation and were not tied to any official decision making, so they are distinct from the context of principal evaluation as it generally occurs in schools. However, the studies raise noteworthy issues about the accuracy of principals' judgments. Results are mixed, showing on the one hand that principal evaluations may be as accurate as value-added models in identifying teachers' ability to improve student achievement but on the other, that principal ratings may be biased by various factors and are more accurate in some contexts than others.

Considerations

Because principals must attend to several areas simultaneously and any evaluation used for decision-making purposes should minimize subjectivity and potential bias, administrators should employ a specific and validated observation protocol when conducting teacher evaluations (see Classroom Observation subsection on p. 6). When choosing an instrument, pay careful attention to its intended and validated use. Administrators should be fully trained on the instrument, rater reliability should be established, and periodic recalibration should occur. Observations should be conducted several times per year to ensure reliability, and a combination of announced and unannounced visits may be preferable to ensure that observations capture a more complete picture of teacher practices. If the focus of the evaluation is to assess deep or specific content knowledge, it may be better to ask a peer teacher or content expert to conduct the evaluation, as a principal or administrator may not have the specialized knowledge to make informed judgments (Stodolsky, 1990; Weber, 1987; Yon, Burnap, & Kohut, 2002). Using a combination of principal and peer raters may increase the credibility of the evaluation.

To incorporate all these ideas, principals should consider a system of evaluation that serves both formative and summative purposes and involves teachers in the process. If principals are viewed as uninformed or unjust evaluators, teachers may not take evaluation procedures seriously. Making teachers aware of the evaluation criteria ahead of time, providing feedback afterward, giving them the opportunity to discuss their evaluation, and offering them support to target the areas in which they need improvement are components that will strengthen the credibility of the evaluation. Evaluation systems are more likely to be productive and respected by teachers if the processes are well explained and understood by teachers, well aligned with school goals and standards, used formatively for teaching development, and viewed as a support system for promoting schoolwide improvement.

Analysis of Classroom Artifacts

Definition

Another method that has been introduced to the area of teacher evaluation is the analysis of classroom artifacts. This method considers lesson plans, teacher assignments, assessments, scoring rubrics, student work, and other artifacts to determine the quality of instruction in a classroom. The idea is that by analyzing classroom artifacts, evaluators can glean a better understanding of how a teacher creates learning opportunities for students on a day-to-day basis. Depending on the goals and priorities of the evaluation, artifacts may be judged on a wide variety of criteria including rigor, authenticity, intellectual demand, alignment to standards, clarity, and comprehensiveness. Although the examination of teacher lesson plans or student work is often included in teacher evaluation procedures, this subsection specifically addresses structured and validated protocols for analyzing artifacts to evaluate the quality of instruction.

Research

Most of the research in this area has focused on the Instructional Quality Assessment (IQA) developed by UCLA's National Center for Research on Evaluation, Standards, and Student Testing (CRESST). IQA rubrics use classroom assignments and student work to assess the quality of classroom discussion, rigor of lesson activities and assignments, and quality of expectations communicated to students. Pilot studies found that scores generally correlate with quality of observed instruction, quality of student work, and standardized student test scores (Clare & Aschbacher, 2001; Junker et al., 2006; Matsumura, Garnier, Pascal, & Valdés, 2002; Matsumura & Pascal, 2003; Matsumura et al., 2006). These studies also found reasonable reliability for the instrument, though more work may be needed to confirm its dependability and stability (e.g., to determine the ideal number of assignments that should be collected to maximize accuracy of scores while minimizing teacher time and effort).

Another branch of work on analyzing instructional artifacts has been conducted through the Consortium on Chicago School Research (Newmann, Bryk, & Nagaoka, 2001; Newmann, Lopez, & Bryk, 1998) to develop the Intellectual Demand Assignment Protocol (IDAP). This protocol assesses the authenticity and intellectual demand of classroom assignments by analyzing teacher assignments and student work in mathematics and reading. Pilot studies found that scorers can achieve high levels of interrater reliability using the rubrics and that IDAP scores correlate with standardized test score gains. Teachers' use of high-demand assignments was unrelated to student demographics and prior achievement and benefited students with high and low prior achievement alike.

Considerations

Analyzing classroom artifacts is practical and feasible because the artifacts have already been created by teachers, and the procedures do not appear to place unreasonable burdens on teachers (Borko, Stecher, Alonzo, Moncure, & McClam, 2005). This technique may be a useful compromise in terms of providing a window into actual classroom practice, as evidenced by classroom artifacts, while employing a method that is less labor-intensive and costly than full classroom observation. It has the potential to be used both summatively and formatively. However, accurate scoring is essential to the validity of this method. Scorers must be well trained and calibrated and, in some cases, should possess knowledge of the subject matter being evaluated. More research is needed to verify the reliability and stability of ratings, explore links to student achievement, and validate the instruments in different contexts before analysis of classroom artifacts should be considered a primary means for teacher evaluation.

Portfolios

Definition

Portfolios are a collection of materials compiled by teachers to exhibit evidence of their teaching practices, school activities, and student progress. Portfolios are distinct from analyses of instructional artifacts in that materials are collected and created by the teacher for the purpose of evaluation. The portfolio process often requires teachers to reflect on the materials and explain why artifacts were included and how they relate to particular standards. They may contain exemplary work as well as evidence that the teacher is able to reflect on a lesson, identify problems in the lesson, make appropriate modifications, and use that information to plan future lessons. Examples of portfolio materials include teacher lesson plans, schedules, assignments, assessments, student work samples, videos of classroom instruction and interaction, reflective writings, notes from parents, and special awards or recognitions.

Research

Two major examples of programs that use portfolio assessments to evaluate teaching include the National Board for Professional Teaching Standards (NBPTS) certification and Connecticut's Beginning Educator Support and Training (BEST) program. These programs include carefully developed scoring rubrics and requirements for training scorers, who are generally experienced teachers with knowledge of the subject matter being evaluated.

Much research has been conducted on NBPTS certification in particular, but studies linking NBPTS certification to student achievement gains have produced mixed results (Cavalluzzo, 2004; Clotfelter, Ladd, & Vigdor, 2006; Cunningham & Stone, 2005; Goldhaber & Anthony, 2004; McColskey et al., 2006; Sanders, Ashton, & Wright, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004). A recent review of research determined that NBPTS certification can successfully identify high-performing teachers, but it is unclear whether the process itself leads to improvements in practice or whether effective teachers opt to complete the process (Hakel, Koenig, & Elliott, 2008). Results are also difficult to interpret because NBPTS participation is strictly voluntary, and those who pursue the certification are a self-selected group that may differ in significant ways from the teaching population as a whole (Pecheone, Pigg, Chung, & Souviney, 2005). Similarly, other teaching portfolio studies have not produced conclusive results about their reliability or validity in measuring teacher effectiveness.

Considerations

One of the most beneficial aspects of teaching portfolios is their comprehensiveness—they can capture effective teaching that occurs both inside and outside of the classroom and can be specific to any grade level, subject matter, or student population. Research shows that portfolios are useful tools in self-reflection and formative assessment, and they are often seen as beneficial by teachers and administrators. However, their use for summative or high-stakes assessment has not been validated. Most studies deal with teacher and administrator perceptions and do not measure actual improvements in teaching or student learning as a result of the portfolio process. Issues have been found in scoring portfolios. It is difficult to verify consistency in scoring and obtain reliability between scorers, and it is unclear whether materials included in portfolios are accurate representations of a teacher's practice. In addition, portfolios and corresponding reflections are considered a time burden by some teachers, so built-in time to develop portfolios should be provided to teachers if portfolios are required as part of a school evaluation or improvement system.

Self-Report of Practice

Definition

Teacher self-report measures ask teachers to report on what they are doing in the classroom and may take the form of surveys, instructional logs, or interviews. Like observations, self-report measures may focus on broad and overarching aspects of teaching that are thought to be important in all contexts, or they may focus on specific subject matter, content areas, grade levels, or techniques. They may consist of straightforward checklists of easily observable behaviors and practices; they may contain rating scales that assess the extent to which certain practices are used or are aligned with certain standards; or they may require teachers to indicate the precise frequency of use of practices or standards. Thus, this class of measures is quite broad in scope, and considerations in choosing or designing a self-report measure will depend largely on its intended purpose and use.

Research

Examples of teacher self-report methods include large-scale surveys, instructional logs, and teacher interviews.

Large-Scale Surveys. Large-scale surveys often focus on measuring reform-oriented practices or enactment of curriculum. Some examples include surveys from the National Center for Education Statistics, Trends in International Mathematics and Science Study (TIMSS), Reform-Up-Close study, *Surveys of Enacted Curriculum*, School Reform Assessment Project, Validating National Curriculum Indicators, and California Learning Assessment System. Large-scale surveys generally address four main dimensions of classroom instruction: (1) pedagogy, (2) professional development, (3) instructional materials and technology, and (4) topical coverage within courses (Mullens, 1995). Some researchers have found survey responses to be consistent with related measures (e.g., Porter, Kirst, Osthoff, Smithson, & Schneider, 1993), whereas others have found serious problems with the reliability and validity of self-reported practices (e.g., Burstein et al., 1995). One study suggests that surveys may be able to indicate which practices are used most relative to others but are less reliable in indicating the precise amount of time spent on those practices (Mayer, 1999).

Instructional Logs. In contrast to large-scale surveys, instructional logs require teachers to keep a frequent and detailed record of teaching. Logs are highly structured and require very specific information about content coverage and instructional practices. One notable study reveals issues with the validity of logs, finding that teacher and researcher reports did not always correspond (Camburn & Barnes, 2004). Rater agreement was sensitive to several factors, from the frequency of the instructional activity to the content being covered. This study raises an important issue: individual raters inherently bring different values, knowledge, and interpretations into their evaluations. Although logs may have potential for providing a detailed account of teaching practices, further investigation is needed to address these validity issues.

Teacher Interviews. Interviews are most often used as supplements to other measures of effective teaching and can play a unique role in gathering information on perceptions and opinions that describe the "whys" and "hows" of teacher performance and its impact. Studies such as the Study of Instructional Improvement (Ball & Rowan, 2004) and the RAND Mosaic II (Le et al., 2006) use interviews to help explain and verify the information they obtain from other measures of teaching. Interview protocols can be highly structured or largely open-ended and can produce more detailed, in-depth information than survey measures. Few studies examine the reliability or validity of interview protocols as a whole. In one example, researchers developed an interview protocol to assess professional standards and student learning (Flowers & Hancock, 2003). The protocol was highly structured, including specific questions on instructional activities, intentions, actions, and a detailed scoring rubric completed by trained evaluators. The study reported high rater reliability and content validity for the protocol, demonstrating that interviews can meet these criteria given their design.

Considerations

Teacher self-reports have certain advantages, and this method may be one useful element in a teacher evaluation system. Self-report data can tap into a teacher's intentions, thought processes, knowledge, and beliefs, and they can be useful for teacher self-reflection and formative purposes. In addition, consideration of teacher perspective and teacher involvement in their evaluations are important factors. Teachers are the only ones with full knowledge of their abilities, classroom context, and curricular content, and they can provide insights that an outside observer may not recognize. Surveys tend to be cost-efficient, generally unobtrusive to collect, and capable of gathering a large array of data.

Self-report measures can be particularly useful as a first step toward investigating some questions of interest—for instance, in establishing a basic level of standard use and understanding among teachers (Cohen & Hill, 2000; Spector, 1994). However, summative or high-stakes decisions should not be based on the results of self-report measures. Research on the reliability and validity of these methods is mixed, and self-report responses may be susceptible to biases such as social desirability (Moorman & Podsakoff, 1992). For example, teachers may misrepresent their actual teaching practices to "look good," or they may unintentionally misreport their practices, believing that they are correctly implementing a practice when, in fact, they are not. Potential biases may lead to both overreporting and underreporting of practices, making the data difficult to interpret.

To minimize potential reporting bias, it is best to gather data from more than one source, gather data longitudinally rather than just at one point in time, and ensure teachers that their responses will be strictly confidential and anonymous. Another crucial issue is making sure that the terminology used in the measures is clear and understandable and that teachers and raters will be able to consistently interpret what information the measures request (Ball & Rowan, 2004; Blank, Porter, & Smithson, 2001; Mullens, 1995). This may require training of teachers and raters on the survey or log measure in order to elicit the intended information. In addition, administrators should consider how broad or detailed the instrument needs to be to inform the desired purpose of the evaluation. If considering a preexisting instrument, administrators should select one that has been widely used and validated by research for their intended purpose.

Student Evaluation

Definition

Student evaluations most often come in the form of a questionnaire that asks students to rate teachers on a Likert-type scale (usually a four-point or five-point scale). Students may assess various aspects of teaching, from course content to specific teaching practices and behaviors. Given that students have the most contact with their teachers and are the most direct consumers of teachers' services, it seems that valuable information could be obtained from evaluations of their experience. However, student ratings are rarely taken seriously as part of teacher evaluation systems.

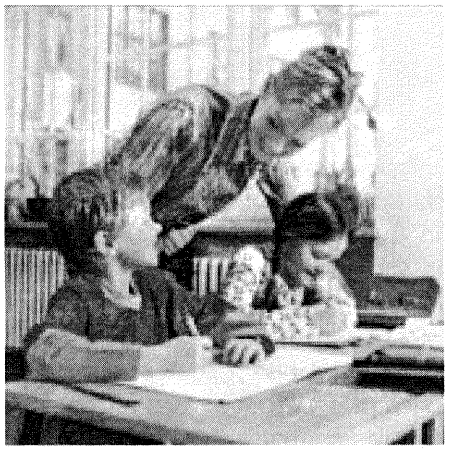
Student ratings of teachers are sometimes not considered a valid source of information because students lack knowledge about the full context of teaching, and their ratings may be susceptible to bias. There is concern that students may rate teachers on personality characteristics or how they are graded rather than instructional quality. Students are considered particularly susceptible to rating leniency and "halo" effects. For example, if they rate a teacher highly on one trait or aspect of teaching, they might be influenced to rate that teacher highly on other, unrelated items.

Research

Research suggests that these worries might be exaggerated and that student feedback can be a valuable component of a teacher evaluation system. Several studies conclude that students can respond reliably and validly when rating their classroom teachers and do not seem to be more susceptible to bias than college students or other adult groups (Follman, 1992, 1995; Worrell & Kuterbach, 2001). Worrell and Kuterbach (2001) found that student ratings tended to be skewed toward high satisfaction but were reliable overall. The study also showed that students of different age groups focused on different aspects of teaching. For example, younger students were more concerned with teacher-student relationships, whereas older students focused more on student learning. Furthermore, student ratings have been shown to correlate with measures of student achievement (Kyriakides, 2005; Wilkerson et al., 2000). For example, Wilkerson et al. (2000) found that student ratings were more highly correlated with student achievement than teacher effectiveness ratings given by principals and teachers themselves. In this study, student ratings were the best predictor of student achievement across all subjects.

Considerations

Student ratings are cost-efficient and time-efficient, can be collected unobtrusively, and can be used to track changes over time (Worrell & Kuterbach, 2001). They also require minimal training, although it is necessary to employ a well-designed questionnaire that measures meaningful teacher behaviors to maintain the validity of the results. However, researchers caution that student ratings should not be stand-alone evaluation measures, as students are not usually qualified to rate teachers on curriculum, classroom management, content knowledge, collegiality, or other areas associated with effective teaching (Follman, 1992; Worrell & Kuterbach, 2001). Overall, studies recommend that student ratings be included as part of the teacher evaluation process but never as the primary or sole evaluation criterion.



Creating a Comprehensive Teacher Evaluation System

In many states, teacher effectiveness is determined based on results from a single measure, typically classroom observations and sometimes value-added models. However, using one or even both of these measures cannot account for the many significant ways teachers contribute to the success and well-being of their students, classrooms, and schools. Creating a comprehensive score for teachers that includes multiple measures is necessary to capture important information that is not included in most classroom observation protocols or value-added scores. Of course, it is not practical or feasible to employ *all* the measures presented in this guide, but by considering the priorities of the school and the intended purpose of evaluation, administrators can strategically choose evaluation measures to create a *system* that accomplishes its various goals. Appendix D contains a sample of existing evaluation systems for consultation in the development of new evaluation systems.

In devising such systems, it is crucial to consider the following main points:

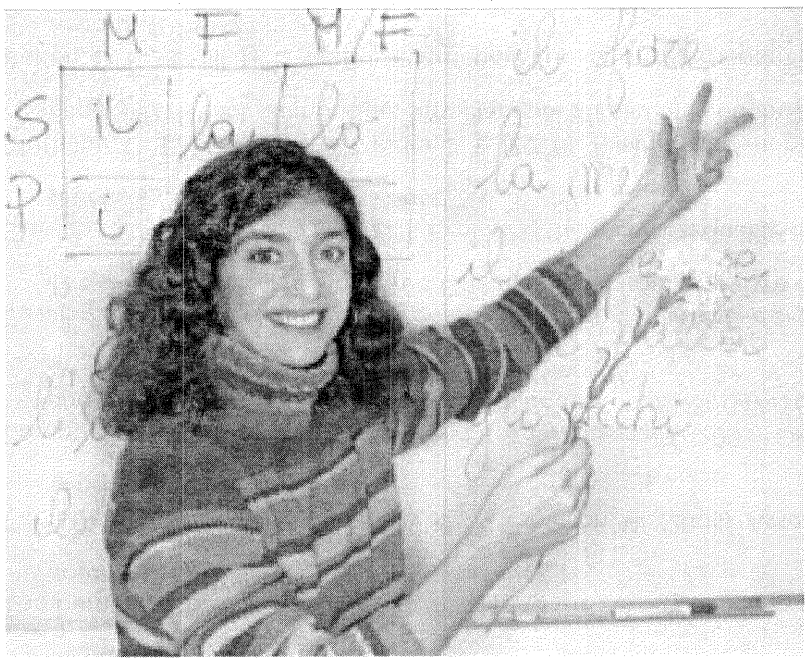
- Teaching contexts differ greatly across subjects, grades, intentional groupings of students in schools, and subgroups of students and between schools with different student populations and local circumstances. Consider teacher effectiveness in light of these different contexts, and incorporate measures that take into account differences in subject matter, teacher activities, student background, personal characteristics, and school culture and organization (Campbell, Kyriakides, Muijs, & Robinson, 2003).
- Use teacher effectiveness results to improve instruction. There are many ways to conceptualize teacher effectiveness and many different uses for teacher evaluation results, but the ultimate goal of evaluation is the same: to improve instruction and student learning. Evaluations should provide information that can be used to identify weaknesses in instruction and to design appropriate strategies for improving instruction. Effective evaluation systems will integrate summative and formative processes so that summative results are not isolated from professional development efforts but are used in conjunction with formative data to support teachers and help them improve.
- Measures of teacher effectiveness (e.g., classroom observation protocols or value-added models) are not valid in and of themselves for determining teacher effectiveness. Instruments are validated for a particular purpose, and their validity is dependent on whether they are used as intended. A crucial step in obtaining valid information is deciding what is important and then finding (or perhaps creating) a measure that will yield tangible evidence about teachers' performance in that area. Using a broadened definition of teacher effectiveness, there is no single measure that will provide valid information on all the ways teachers contribute to student learning. Multiple measures capturing different aspects of teacher effectiveness should be employed. See Table 1 to determine appropriate measures for specific purposes.

Table 1. Matching Measures to Specific Purposes

Purpose of Evaluation of Teacher Effectiveness	Value-Added	Classroom Observation	Analysis of Artifacts	Portfolios	Teacher Self-Reports	Student Ratings	Other Reports
Find out whether grade-level or instructional teams are meeting specific achievement goals.	x						
Determine whether a teacher's students are meeting achievement growth expectations.	x		x				
Gather information in order to provide new teachers with guidance related to identified strengths and shortcomings.		x	x	x			x
Examine the effectiveness of teachers in lower elementary grades for which no test scores from previous years are available to predict student achievement (required for value-added models).		x	x	x			x
Examine the effectiveness of teachers in nonacademic subjects (e.g., art, music, and physical education).		x		x		x	x
Determine whether a new teacher is meeting performance expectations in the classroom.		x	x	x		x	x
Determine the types of assistance and support a struggling teacher may need.		x	x		x	x	
Gather information to determine what professional development opportunities are needed for individual teachers, instructional teams, grade-level teams, etc.	x	x			x		x
Gather evidence for making contract renewal and tenure decisions.	x	x					x
Determine whether a teacher's performance qualifies him or her for additional compensation or incentive pay (rewards).	x	x					
Gather information on a teacher's ability to work collaboratively with colleagues to evaluate needs of and determine appropriate instruction for at-risk or struggling students.				x	x		x
Establish whether a teacher is effectively communicating with parents/guardians.				x			x
Determine how students and parents perceive a teacher's instructional efforts.						x	
Determine who would qualify to become a mentor, coach, or teacher leader.	x	x	x	x			x

Note. "X" indicates appropriate measures for the specified purpose.

Recommendations



The following guidelines sum up the suggestions presented in this guide for conceptualizing and creating a comprehensive system to best measure teacher effectiveness:

- **Resist pressures to reduce the definition of teacher effectiveness to a single score obtained on an observation instrument or through a value-added model.** It may be convenient to adopt a single measure of teacher effectiveness; however, there is no single measure that captures every significant teacher contribution.
- **Consider the purpose of the teacher evaluation before deciding on the appropriate measure to employ.** Value-added scores may provide information about a teacher's contribution to student learning, but they would be less helpful in providing teachers with guidance on how to improve their performance.
- **Remember that validity depends on how well the instrument measures what you have deemed important and how the instrument is used in practice.** Even a high-quality instrument is not valid unless it is being used appropriately and for its intended purpose. A reliable classroom observation protocol may be wildly inaccurate or inconsistent in the hands of an untrained evaluator, and a value-added score calculated with large amounts of missing student data may grossly misrepresent a teacher's contribution to student learning.
- **Seek out or create appropriate measures to capture important information about teachers' contributions that go beyond student achievement score gains.** This may include measures that capture teachers' leadership activities within the school, their collaboration with other teachers to strategize ways to help students at risk for failure, or their participation in a study group to align the curriculum with state standards.
- **Include education stakeholders in decisions about what is important to measure.** Although a state legislature or task force may ultimately decide how teacher effectiveness will be measured, listening to the voices of teachers, principals, curriculum specialists, union representatives, parents, and students will help assure greater acceptance of the measurement system. In addition, this will help ensure that the validity of the system is not threatened by noncompliance or active resistance.
- **Keep in mind that valid measurement may be costly.** Ensuring that data are complete and accurate and that raters are trained and calibrated is essential to guarantee the validity of scores from teacher effectiveness measures. In addition, developing and validating new measures based on local priorities will require adequate funding.

References

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal, 105*(1), 3–10.

Blank, R. K., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science. Results from Survey of Enacted Curriculum* project. Washington, DC: Council of Chief State School Officers. Retrieved March 3, 2009, from http://www.ccsso.org/content/pdfs/newtools.pdf

Blunk, M. L. (2007). *The QMI: Results from validation and scale-building.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment, 10*(2), 73–104.

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest region* (Issues & Answers, 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 March 3, 2009, from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2007030.pdf

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Burstein, L., McDonnell, L. M., VanWinkle, J., Ormseth, T. H., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal, 105*(1), 49–73.

Campbell, R. J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated mode.* London: Routledge Falmer.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education, 29*(3), 347–362. Retrieved March 3, 2009, from http://www.jstor.org/stable/3595446?seq=16

Cavalluzzo, L. C. (2004). *Is national board certification an effective signal of teacher quality?* [Report No. (IPR) 11204]. Alexandria, VA: The CNA Corporation. Retrieved March 3, 2009, from http://www.cna.org/documents/cavaluzzostudy.pdf

Clare, L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7*(1), 39–59.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). *Teacher-student matching and the assessment of teacher effectiveness* (NBER Working Paper No. 11936). Cambridge, MA: National Bureau of Economic Research.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294–343.

Cunningham, G. K., & Stone, J. E. (2005). Value-added assessment of teacher quality as an alternative to the National Board for Professional Teaching Standards: What recent studies say. In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 209–232). Maple Grove, MN: JAM Press. Retrieved March 3, 2009, from http://www.education-consumers.com/Cunningham-Stone.pdf

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Drury, D., & Doran, H. (2003). The value of value-added analysis [Policy Research Brief, 3(1)]. Alexandria, VA: National School Boards Association.

Flowers, C. P., & Hancock, D. R. (2003). An interview protocol and scoring rubric for evaluating teacher performance. *Assessment in Education, 10*(2), 161–168.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal, 75*(3), 168–178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal, 25*(1), 57–78.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 79–107.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved March 3, 2009, from http://www.tqsource.org/publications/LinkBetweenTQandStudentOutcomes.pdf

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved March 3, 2009, from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* (Working Paper). Seattle, WA: Center on Reinventing Public Education. Retrieved March 3, 2009, from http://www.crpe.org/cs/crpe/download/csr_files/wp_crpe6_nbptsoutcomes_apr04.pdf

Hakel, M. D., Koenig, J. A., & Elliott, S. W. (2008). *Assessing accomplished teaching: Advanced-level certification programs.* Washington, DC: National Academies Press.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967.

Harris, D. N., & Sass, T. R. (2007). *What makes for a good teacher and who can tell?* Unpublished manuscript.

Heneman, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay* (CPRE Policy Briefs No. RB-45). Philadelphia, PA: Consortium for Policy Research in Education. Retrieved March 3, 2009, from http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf

Hershberg, T., Simon, V. A., & Lea-Kruger, B. (2004). Measuring what matters. *American School Board Journal, 191*(2), 27–31. Retrieved March 3, 2009, from http://www.cgp.upenn.edu/pdf/measuring_what_matters.pdf

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Hoffman, J. V., Sailors, M., Duffy, G. R., & Beretvas, S. N. (2004). The effective elementary classroom literacy environment: Examining the validity of the TEX-IN3 Observation System. *Journal of Literacy Research, 36*(3), 303–334.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207–219.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27–50.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (Faculty research working paper series No. RWP05-040). Cambridge, MA: Harvard University John F. Kennedy School of Government.

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

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., et al. (2006). *Overview of the instructional quality assessment* (CSE Technical Report 671). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved March 3, 2009, from http://www.cse.ucla.edu/products/reports/r671.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York: Praeger Publishers.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54–78.

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction, 40*(2), 44–66.

Le, V., Stecher, B. M., Lockwood, J. R., Hamilton, L. S., Robyn, A., Williams, V. L., et al. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement.* Santa Monica, CA: RAND. Retrieved March 3, 2009, from http://www.rand.org/pubs/monographs/2006/RAND_MG480.pdf

Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8*(3), 207–229.

Matsumura, L. C., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice* (Technical Report No. 602). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., et al. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment* (CSE Technical Report No. 681). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved March 3, 2009, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/e0/f6.pdf

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29–45.

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2004). *The promise and peril of using value-added modeling to measure teacher effectiveness* (Research Brief No. RB-9050-EDU). Santa Monica, CA: RAND. Retrieved March 3, 2009, from http://www.rand.org/pubs/research_briefs/RB9050/RAND_RB9050.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND. Retrieved March 3, 2009, from http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

McColskey, W., Stronge, J. H., Ward, T. J., Tucker, P. D., Howard, B., Lewis, K., et al. (2006). *Teacher effectiveness, student achievement, and National Board Certified Teachers.* Arlington, VA: National Board for Professional Teaching Standards. Retrieved March 3, 2009, from http://www.education-consumers.com/articles/W-M%20NBPTS%20certified%20report.pdf

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*(4), 242–247.

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

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Mintzberg, H. (1989). *Mintzberg on management, inside our strange world of organisations.* New York: Free Press.

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65*(2), 131–149.

Mullens, J. E. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for the Teacher Followup Survey* (NCES-WP-95-15). Washington, DC: National Center for Education Statistics.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research. Retrieved March 3, 2009, from http://ccsr.uchicago.edu/publications/p0a02.pdf

Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report.* Chicago, IL: Consortium on Chicago School Research. Retrieved March 3, 2009, from http://ccsr.uchicago.edu/publications/p0f04.pdf

Pecheone, R. L., Pigg, M. J., Chung, R. R., & Souviney, R. J. (2005). Performance assessment and electronic portfolios: Their effect on teacher learning and education. *The Clearing House, 78*(4), 164–176. Retrieved March 3, 2009, from http://eds.ucsd.edu/people/PerformanceAssessmentTCH.pdf

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2006a). *Classroom assessment scoring system: Preschool (pre-k) version.* Baltimore, MD: Brookes.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2006b). *Classroom assessment scoring system: Observation protocol manual.* Baltimore, MD: Brookes.

Piburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP): Reference manual* (Technical Report No. IN00-3). Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers. Retrieved March 3, 2009, from http://cresmet.asu.edu/instruments/RTOP/RTOP_Reference_Manual.pdf

Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J. L., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms.* Madison: Wisconsin Center for Education Research.

Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *Elementary School Journal, 105*(4), 377–394.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458. Retrieved March 3, 2009, from http://edpro.stanford.edu/Hanushek/admin/pages/files/uploads/teachers.econometrica.pdf

Rothstein, J. (2008a). *Do value-added models add value? Tracking, fixed effects, and causal inference.* Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison.

Rothstein, J. (2008b, April 22–24). *Student sorting and bias in value added estimation: Selection on observables and unobservables.* Paper presented at the National Conference on Value-Added Modeling.

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329–339. Retrieved March 3, 2009, from http://www.sas.com/govedu/edu/opp_hurdles.pdf

Sanders, W. L., Ashton, J. J., & Wright, S. P. (2005). *Final report: Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress.* Arlington, VA: National Board for Professional Teaching Standards. Retrieved March 3, 2009, from http://www.nbpts.org/UserFiles/File/SAS_final_NBPTS_report_D_-_Sanders.pdf

Schacter, J., Thum, Y. M., & Zifkin, D. (2006). How much does creative teaching enhance elementary school students' achievement? *Journal of Creative Behavior, 40*(1), 47–72.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50–91). New York: Macmillan.

Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior, 15*(5), 385–392.

Stodolsky, S. S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175–190). Thousand Oaks, CA: Sage.

The Teaching Commission. (2004). *Teaching at risk: A call to action.* New York: The Teaching Commission. Retrieved March 3, 2009, from http://www.ecs.org/html/offsite.asp?document=http%3A%2F%2Fftp%2Eets%2Eorg%2Fpub%2Fcorp%2Fttcreport%2Epdf

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved March 3, 2009, from http://www.nbpts.org/UserFiles/File/National_Board_Certified_Teachers_and_Their_Students_Achievement_-_Vandevoort.pdf

Weber, J. R. (1987). *Teacher evaluation as a strategy for improving instruction. Synthesis of literature* (ERIC ED287213). Elmhurst, IL: North Central Regional Educational Laboratory. Retrieved March 3, 2009, from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/fd/7d.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 Evaluation in Education, 14*(2), 179–192.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education, 14*(4), 236–247.

Yon, M., Burnap, C., & Kohut, G. (2002). Evidence of effective teaching perceptions of peer reviewers. *College Teaching, 50*(3), 104–110.

Appendix A. Validity and Measurement Terms

The following terms are used in discussions of validity and measurement throughout this guide. The definitions helped frame the recommendations and evaluations of each method for measuring teacher effectiveness presented in this guide. Thus, these terms double as a list of considerations to take into account when assessing an evaluation instrument or designing an evaluation program.

Calibration refers to a periodic assessment of whether raters are continuing to score reliably. Raters trained to use a certain instrument may "drift" from their original training. For example, there is a tendency for raters to score teachers differently at the beginning of the year compared to later in the year—after observing more teachers, they may become more lenient or more stringent in their scoring. This potentially results in teachers receiving different scores for the same performance. Valid evaluation systems will protect against this rater "drift" by establishing rater reliability not just at the beginning of the process but periodically throughout the year and will provide continued training to recalibrate raters to reliable levels. Ensuring that raters are still scoring the instrument as its developers intended can be accomplished through "double-scoring"—having specially trained "master coders" observe the same lessons observed by raters in the field and verifying the raters' interpretations.

Comprehensiveness refers to the extent to which a measure can capture all the various aspects of teacher effectiveness (e.g., how well a teacher represents mathematics in the classroom, scaffolds student learning, and works collaboratively with colleagues).

Credibility is a specific type of validity—also called *face validity*—that refers to how many stakeholders from different groups (e.g., parents, teachers, administrators, and policymakers) view the measure as reasonable and appropriate and support its use.

Generality refers to how well an instrument captures the full range of contexts in which teachers work. An instrument that can assess teacher effectiveness across multiple subjects and grade levels is more general, and this is particularly useful if the intent is to compare teachers across contexts.

Formative evaluation gathers information with the intention of providing feedback to improve a program, activity, or behavior. Formative feedback is meant to promote reflection and growth rather than to make definitive judgments.

High-stakes evaluations are those that use summative information to make decisions that carry significant consequences to teachers, such as tenure, dismissal, and pay decisions.

Low-stakes evaluations are often informal or unstructured evaluations that do not carry substantial consequences to teachers. Low-stakes evaluations are conducted for purely formative purposes.

Practicality refers to the logistical issues associated with a measure, such as costs, feasibility, adaptability, training, and other resources required to implement the measure.

Reliability refers to the degree to which an instrument measures something consistently. A validated instrument must be evaluated for how reliable the results are across different raters and contexts. When the various methods for measuring teacher effectiveness are discussed, there is often reference to rater reliability—whether or not raters have been trained to score reliably. This involves being able to do the following: rate consistently with standards, rate consistently with other raters (referred to as *interrater reliability*), and rate consistently across observations and contexts—ratings should not be influenced by factors such as the time of day, time of year, or subject matter being taught, and they should be consistent across different observations of the same teacher.

Summative evaluation gathers information with the intention of making a final determination about a program, activity, or behavior at a specific point in time. Summative evaluation is meant to make definitive judgments that inform decisions involving tenure, dismissal, performance pay, and teaching assignments.

Utility refers to how useful scores from an instrument are for a specific purpose. For example, scores from an instrument that ignores teaching context may not be useful in identifying contexts that appear to support more effective teaching. The experience of other researchers or practitioners with an instrument makes it possible to better anticipate its potential uses and limitations.

Validity refers to the degree to which an interpretation of an evaluation score is supported by evidence. For a measure of teacher effectiveness to be valid, evidence must support the argument that the measure actually assesses the dimension of teacher effectiveness it claims to measure and not something else. It is also essential to have evidence that the measure is valid for the purpose for which it will be used. Instruments cannot be valid in and of themselves; an instrument or assessment must be validated for particular purposes (Kane, 2006; Messick, 1989).

Appendix B. Planning Guide

	Recommendation	Planning Questions
Resources	Consider what resources you will need to carry out an evaluation of teacher effectiveness. Resources include not just dollars but also people, time, data, and cooperation from schools and districts.	What are the federal, state, and local resources that we could use to evaluate teacher effectiveness? Are they sufficient for the task?
Design	Decide on the purpose of the evaluation—formative assessment to help improve teaching practice, summative assessment as part of credentialing or tenure, and/or summative assessment to reward effective teaching.	What purpose(s) will measuring teacher effectiveness serve?
	Measure what is most important to you, your administrators, your teachers, and other education stakeholders. Administrators and teachers will focus on these measures as they strive to improve.	What is most important for our state? Student achievement? Classroom practice? Other school outcomes (e.g., graduation rates, narrowing achievement gaps, college attendance)? How do we know these are most important to measure?
	Involve teachers and stakeholders in developing a system for measuring teacher effectiveness. Having the participation of teachers, administrators, and other stakeholders may increase the validity of the instruments and processes involved in measuring effectiveness.	Which stakeholders should be involved in designing a system for measuring teacher effectiveness?
	Incorporate a way for teachers to understand how they can improve and provide learning opportunities for them to develop their teaching skills.	How will the results of effectiveness assessments be communicated to teachers so that they can improve their teaching? How will we build into the evaluation system ways for teachers to grow professionally?
Measures	Choose a set of measures and processes for which adequate resources are available.	Does the data system currently in place support value-added measures of teacher effectiveness (student achievement data for individual students linked to specific teachers)?
	Differentiate among teachers. Not all measures of teacher effectiveness are appropriate for every grade level and subject.	What measures can we use for subjects and grade levels for which no standardized test is available? How can we ensure that all teachers receive a fair evaluation of their effectiveness, even when different measures are used?
	Consider multiple measures. More measures will provide more information which can be used by teachers, schools, and teacher preparation programs to address teacher performance.	How can we combine measures to develop a more comprehensive picture of teacher practice and how it affects student outcomes?

Appendix C. Summary of Measures

Measure	Description	Research	Strengths	Cautions
Value-Added Models	Statistical models used to determine teachers' contributions to students' test score gains. May also be used as a research tool (e.g., determining the distribution of "effective" teachers by student or school characteristics).	Little is known about the validity of value-added scores for identifying effective teaching, though research using value-added models suggests that teachers differ markedly in their contributions to students' test score gains. However, correlating value-added scores with teacher qualifications, characteristics, or practices has yielded mixed results and few significant findings. Teachers vary in effectiveness, but research has not determined why.	• Provides a way to evaluate teachers on their contribution to student learning, which most measures do not. • Requires no classroom visits because linked student/teacher data can be analyzed at a distance. • Entails little burden at the classroom or school level because most data are already collected for NCLB purposes. • May be useful for identifying outstanding teachers whose classrooms can serve as "learning labs" as well as struggling teachers in need of support.	• Models are not able to sort out *teacher* effects from *classroom* effects. • Vertical test alignment is assumed (i.e., tests are measuring essentially the same thing from grade to grade). • Value-added scores are not useful for formative purposes because teachers learn nothing about how their practices contributed to (or impeded) student learning. • Value-added measures are controversial because they measure *only* teachers' contributions to student achievement gains on standardized tests.
Classroom Observation	Classroom observations are used to measure observable classroom processes, including specific teacher practices, holistic aspects of instruction, and interactions between teachers and students. They can measure broad, overarching aspects of teaching, or subject-specific or context-specific aspects of practice.	Some highly researched protocols have been linked to student achievement, though associations are sometimes modest. Research and validity findings are highly dependent on the instrument used, sampling procedures, and the training of raters. There is a lack of research on observation protocols as used in context for teacher evaluation.	• Provides rich information about classroom behaviors and activities. • Is credible—generally considered a fair and direct measure by stakeholders. • Depending on the protocol, can be used in various subjects, grades, and contexts. • Can provide information useful for both formative and summative purposes.	• Choosing or creating a valid and reliable protocol and training and calibrating raters are essential to obtaining valid results. • Expensive due to cost of observers' time; intensive training and calibrating of observers adds to expense but is necessary for validity. • Assesses observable classroom behaviors, but not as useful for assessing beliefs, feelings, intentions, or out-of-classroom activities.

Measure	Description	Research	Strengths	Cautions
Portfolios	A collection of teaching materials and artifacts assembled by the teacher to document a large range of teaching behaviors and responsibilities. Has been used widely in teacher education programs and in states for assessing the performance of teacher candidates and beginning teachers.	Research on validity and reliability is ongoing, and concerns have been raised about consistency of scoring. There is a lack of research linking portfolios to observed changes in teaching practice or student achievement. Some studies have linked NBPTS certification (which includes a portfolio) to student achievement, but other studies have found no relationship.	• Is comprehensive; can measure aspects of teaching that are not readily observable in the classroom. • Can be used with teachers of all fields. • Has a high level of credibility among stakeholders. • Is a useful tool for teacher reflection and improvement.	• This measure is time-consuming for teachers and scorers; scorers should have content knowledge of the portfolios they score. • Stability of scores may not be high enough to use for high-stakes assessment. • Portfolios are difficult to standardize (compare across teachers or schools). • Portfolios represent teachers' exemplary work but may not reflect everyday classroom activities.
Self-Report of Practice	Teacher reports of their practices, techniques, intentions, beliefs, and other teaching elements assessed through surveys, instructional logs, or interviews. Measures cover a broad spectrum and can vary widely in focus and level of detail.	Studies on the validity of teacher self-report measures present mixed results. Highly detailed measures of practice may be better able to capture actual teaching practices but may be more difficult to establish reliability or may result in very narrowly focused measures.	• Can measure unobservable factors that may affect teaching, such as knowledge, intentions, expectations, and beliefs. • Provides the unique perspective of the teacher. • Is feasible and cost-efficient; can collect large amounts of information at once.	• Reliability and validity of self-report has not been fully established and depends on instrument used. • Using or creating a well-developed and validated instrument will decrease cost-efficiency but will increase accuracy of findings. • This measure should not be used as the sole or primary measure in teacher evaluation.

Measure	Description	Research	Strengths	Cautions
Principal Evaluation	Are generally based on classroom observation, may be structured or unstructured; procedures and uses vary widely by district. Generally used for summative purposes, most commonly for tenure or dismissal decisions for beginning teachers.	Studies comparing subjective principal ratings to student achievement find mixed results. Little evidence exists on validity of evaluations as they occur in schools, but evidence indicates that training for principals is limited and rare, which would impair validity of their evaluations.	• Represents a useful perspective based on principals' knowledge of their school and context. • Is generally feasible and can be one useful component in a system used to make summative judgments and provide formative feedback.	• Evaluation instruments used without proper training or regard for their intended purpose will impair validity. • Principals may not be qualified to evaluate teachers on measures highly specialized for certain subjects or contexts.
Analysis of Classroom Artifacts	Structured protocols used to analyze classroom artifacts in order to determine the quality of instruction in a classroom. Artifact examples: lesson plans, teacher assignments, assessments, scoring rubrics, and student work.	Pilot research has linked artifact ratings to observed measures of practice, quality of student work, and student achievement gains. More work is needed to establish scoring reliability and determine the ideal amount of work to sample. Lack of research exists on the use of structured artifact analysis in practice.	• Can be a useful measure of instructional quality if a validated protocol is used, if raters are well-trained for reliability, and if assignments show sufficient variation in quality. • Is practical and feasible because artifacts have already been created for the classroom.	• More validity and reliability research is needed. • Training knowledgeable scorers can be costly but is necessary to ensure validity. • This measure may be a compromise in terms of feasibility and validity between full observation and less direct measures such as self-report.

Appendix D. Sample of Existing Evaluation Systems

- The Beginning Educator Support and Training Program (BEST) (http://www.ctbest.org) [This Connecticut program is currently being revamped due to new legislation (see http://24.248.88.133/Resources/2008_BEST_C1.htm)].
- Delaware Performance Appraisal System (http://www.doe.k12.de.us/performance/dpasii/default.shtml)
- Florida District Performance Appraisal System Checklist (http://www.fldoe.org/profdev/pa.asp)
- Minnesota Q-Comp – Quality Teacher Compensation, Part of the National Institute for Excellence in Teaching (http://cfl.state.mn.us/MDE/Teacher_Support/QComp/index.html)
- New Mexico Evaluation Guidelines (http://www.teachnm.org/annual_assessment.html)
- North Carolina Public School Employee Evaluation Standards and Instruments (http://www.ncpublicschools.org/fbs/personnel/evaluation/)
- Ohio Value-Added Support (http://portal.battelleforkids.org/Ohio/home.html?sflang=en)
- South Carolina Performance Appraisal System (ADEPT) (http://www.scteachers.org/ADEPT/index.cfm)
- Ten Indicators of a Quality Teacher Evaluation Plan (http://www.sde.ct.gov/sde/cwp/view.asp?a=2641&q=320432)
- Tennessee Framework for Evaluation and Professional Growth Guidelines and Manuals (http://www.state.tn.us/education/frameval/)
- Wisconsin Master Educator Assessment Process and the Master Educator License (http://dpi.wi.gov/tepdl/wmeapsumm.html)

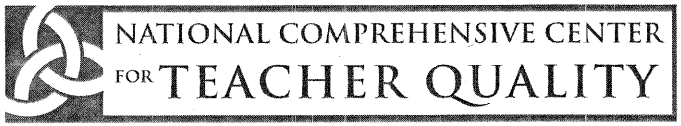
Measure	Description	Research	Strengths	Cautions
Student Evaluation	Surveys or rating scales used to gather student opinions or judgments about teaching practice and to provide information about teaching as perceived by students. Measures can vary widely in focus and level of detail.	Several studies show that student ratings of teachers may be as valid as judgments made by college students and other groups and, in some cases, may correlate with measures of student achievement; thus students can provide useful information about teaching. Validity is dependent on the instrument used and its administration and is generally recommended for formative use only.	• Provides perspective of students, who have the most experience with teachers. • Can provide formative information to help teachers improve practice in a way that will connect with students. • Can potentially provide ratings as accurate as those provided by adult raters.	• Student ratings have not been validated for use in summative assessment and should not be used as the sole or primary measure of teacher evaluation. • Students cannot provide information on aspects of teaching such as a teacher's content knowledge, curriculum fulfillment, or professional activities.

About the National Comprehensive Center for Teacher Quality

The National Comprehensive Center for Teacher Quality (TQ Center) was created to serve as the national resource to which the regional comprehensive centers, states, and other education stakeholders turn for strengthening the quality of teaching—especially in high-poverty, low-performing, and hard-to-staff schools—and for finding guidance in addressing specific needs, thereby ensuring that highly qualified teachers are serving students with special needs.

The TQ Center is funded by the U.S. Department of Education and is a collaborative effort of ETS, Learning Point Associates, and Vanderbilt University. Integral to the TQ Center's charge is the provision of timely and relevant resources to build the capacity of regional comprehensive centers and states to effectively implement state policy and practice by ensuring that all teachers meet the federal teacher requirements of the No Child Left Behind (NCLB) Act.

The TQ Center is part of the U.S. Department of Education's Comprehensive Centers program, which includes 16 regional comprehensive centers that provide technical assistance to states within a specified boundary and five content centers that provide expert assistance to benefit states and districts nationwide on key issues related to the NCLB Act.



1100 17th Street NW, Suite 500
Washington, DC 20036-4632
877-322-8700 • 202-223-6690
www.tqsource.org

Copyright © 2009 National Comprehensive Center for Teacher Quality, sponsored under government cooperative agreement number S283B050051. All rights reserved.

This work was originally produced in whole or in part by the National Comprehensive Center for Teacher Quality with funds from the U.S. Department of Education under cooperative agreement number S283B050051. The content does not necessarily reflect the position or policy of the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsements by the federal government.

The National Comprehensive Center for Teacher Quality is a collaborative effort of ETS, Learning Point Associates, and Vanderbilt University.

3335_04/09