NEPC

NATIONAL EDUCATION POLICY CENTER

GETTING TEACHER ASSESSMENT RIGHT

WHAT POLICYMAKERS CAN LEARN FROM RESEARCH

Patricia H. Hinchey

Penn State University

December 2010

National Education Policy Center

School of Education, University of Colorado at Boulder
Boulder, CO 80309-0249
Telephone: 303-735-5290
Fax: 303-492-7090

Email: NEPC@colorado.edu
http://nepc.colorado.edu

This is one of a series of briefs made possible in part by funding from The Great Lakes Center for Education Research and Practice.

GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE
http://www.greatlakescenter.org
GreatLakesCenter@greatlakescenter.org

Briefs published by the National Education Policy Center (NEPC) are blind peer-reviewed by members of the Editorial Review Board. Visit http://nepc.colorado.edu to find all of these briefs. For information on the editorial board and its members, visit: http://nepc.colorado.edu/editorial-board.

Publishing Director: Alex Molnar

Suggested Citation:
Hinchey, P. H. (2010). *Getting Teacher Assessment Right: What Policymakers Can Learn from Research*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/getting-teacher-assessment-right.

Getting Teacher Assessment Right: What Policymakers Can Learn from Research

Patricia H. Hinchey, Penn State University

Executive Summary

It is well established that teacher quality makes a difference in student learning. Since the implementation of No Child Left Behind in 2002, staffing every classroom with a high-quality teacher has been an official national priority. That goal entails an implicit requirement to assess teacher and teaching quality more rigorously than has been the case in the past. Despite decades of research on how best to assess teacher performance, however, no consensus has evolved on any single assessment strategy or collection of strategies—indicating that the problem of designing adequate and appropriate assessment is inherently complex and controversial. Such complexity has not, however, prevented the Obama administration from encouraging policymakers to define "good" teachers as those who produce gains in student achievement, measured by gains in standardized test scores.

Notwithstanding the federal enthusiasm for test scores, many researchers have warned against using a single measurement of any kind as the primary basis for such important personnel decisions as teacher retention, dismissal or pay. While there are important questions about what achievement scores can—and cannot—indicate about individual teachers, there is no question that placing excessive emphasis on test scores alone can have unintended and undesirable consequences that undermine the goal of developing an excellent teaching force.

Given the experience to date with an overwhelming focus on student achievement scores as a basis for high-stakes decisions, policymakers would do well to pause and carefully examine the issues that make teacher assessment so complex before implementing an assessment plan. To facilitate such examination, this brief reviews credible research exploring: the feasibility of combining formative assessment (a basis for professional growth) and summative assessment (a basis for high-stakes decisions like dismissal); the various tools that might be used to gather evidence of teacher effectiveness; and the various stakeholders who might play a role in a teacher assessment system. It also offers a brief overview of successful exemplars.

Based on the research reviewed, it is recommended that policymakers employ an assessment system that targets both continual improvement of the teaching staff and timely dismissal of teachers who cannot or will not improve. Steps toward that goal include that policymakers:

- **Be clear about the purposes of any assessment before selecting strategies. Where formative and summative assessment are to be combined, plan to address the challenges of dual-purpose systems.**
- **Involve all key stakeholders in system design.**

- **Rather than employing a single assessment tool, gather evidence from multiple sources. Combine strategies so that the weakness of any single tool is offset by the strengths of another.**
- **Be sure that the criteria for assessing performance, artifacts or other factors are credible and are well understood by teachers and assessors.**
- **Provide high-quality, ongoing training for assessors and routinely calibrate their efforts to ensure consistent application of criteria.**
- **Look to high-quality research on existing tools and programs to inform the design of assessment systems.**
- **Commit sufficient resources to produce high-quality, productive assessment.**

Getting Teacher Assessment Right: What Policymakers Can Learn from Research

Introduction

It is well established that teacher quality makes a difference in student learning.[1] Since the implementation of No Child Left Behind in 2002, staffing every classroom with a high-quality teacher has been an official national priority. That goal entails an implicit requirement to assess the quality of teachers and teaching more rigorously than has been the case in the past.[2] Despite decades of research on how best to assess teacher performance, however, no consensus has evolved on any single assessment strategy or collection of strategies—indicating that the problem of designing adequate and appropriate assessment is inherently complex and controversial. Such complexity has not prevented the Obama administration from encouraging policymakers to define "good" teachers as those who produce gains in student achievement, measured by gains in standardized test scores. His Race to the Top initiative, which offers competitive grant money rewarding states that link achievement data to individual teachers, has already prompted some states to pass laws mandating that teacher evaluation be tied to student achievement.[3]

Notwithstanding federal enthusiasm for test scores, many researchers have warned against using a single measurement of any kind as the primary basis for such important personnel decisions as teacher retention, dismissal or pay.[4] While there are important questions about

Policymakers who are considering employing test scores as the primary tool for teacher assessment would do well to pause and carefully examine research evidence.

what exactly achievement scores can—and cannot—indicate about individual teachers, there is no question that placing extreme emphasis on test scores alone can have unintended and undesirable consequences that undermine the goal of developing an excellent teaching force. NCLB's emphasis on high-stakes testing, for instance, has led not only to widespread cheating, but also to such counterproductive practices as school personnel encouraging academically struggling students to transfer or drop out.[5] While such practices might have led to higher achievement scores, no one would consider a teacher who promoted cheating or dropping out a "good" teacher.[6] The equation of higher test scores with high-quality teachers and teaching ignores such complications and their potential for harming students.

Given past drawbacks with basing high-stakes decisions exclusively on student achievement scores, policymakers who are considering employing test scores as the primary tool for teacher assessment would do well to pause and carefully examine research evidence. To facilitate such examination, this brief explores the research on several key questions:

- What exactly is to be assessed—and for what purposes?
- What measurement tools are available, and what are their strengths and weaknesses?
- Who is to do the assessing?
- What systemic models has research shown to be viable for credible and comprehensive assessment of teachers and teaching quality?

This research review provides a basis for concluding recommendations for policymakers.

Methods

Because of the current interest in using test scores as a basis for teacher dismissal, this brief focuses on the assessment of current classroom teachers. It therefore does not include an exploration of related research concerning assessment issues in teacher education, initial certification, and hiring.

The studies reviewed here are primarily research articles from peer-reviewed journals, although a few credible research reports from other sources are also included. Overall, 275 articles and reports were examined as potentially relevant to this review.[7]

What Exactly Is to Be Assessed—And For What Purposes?

High-stakes assessment can be problematic. That which is assessed is often distorted, and that which is not assessed is often neglected. When mandatory testing included only reading and math, for example, many schools narrowed the curriculum to those subjects at the expense of others, like science.[8] Because priority setting drives behavior, before asking *how* to assess teachers, it is essential to ask *what* is so important to teacher and teaching quality that it must be evaluated. Possibilities go well beyond test scores and range widely, from deep (untested) student learning and such teacher traits as honesty to specialized knowledge and skills, such as how to adapt learning activities for special-needs students.

What to assess is not, however, the only preliminary consideration. Just as it is often assumed that student achievement is a logical and sufficient way to assess teachers, it is also widely assumed that the point of such assessment is to make high-stakes personnel decisions. However, the question of how to use assessment data is more complex than it may first appear. Since the purpose of assessment also has implications for the choice of assessment tools, a discussion of two basic types of assessment—formative and summative—follows the discussion of assessment categories below.

Assessment Categories

Despite many earlier efforts to develop one, there is no agreed-upon definition of teacher quality. Recently, however, several researchers have worked to clarify relevant factors.[9] Although terminology and specific categorizations vary in the literature, three common categories emerge: teacher quality, teacher performance, and teacher effectiveness (see Figure

1). *Teacher quality* refers to teacher characteristics such as education, experience, and beliefs. *Teacher performance* refers to what a teacher does, both inside and outside the classroom, and includes such elements as classroom interaction with students and collaborative activity with parents and others in the school community. *Teacher effectiveness* refers to teacher influence on student learning and includes such elements as student test scores and student motivation. Each

Teacher Quality	Teacher Performance	Teacher Effectiveness
Personal traits, skills, and understandings. • Education, experience, credentials, licensure • Content and pedagogical knowledge, including the ability to match pedagogy to context • Understanding of learners and their learning and development, including of specific populations like English Language Learners • Dispositions, beliefs, expectations, values	Teacher activities • Classroom activities and interaction between students and teachers • Learning activities provided or mentored outside the classroom • Teacher activities outside the classroom, in the school and the community	Teacher effects on students • Student achievement • Graduation rates • Student attitudes, behavior, motivation, social and emotional well-being

of these categories has potential for informing judgments about teachers and teaching; each appears routinely in research literature, although different researchers may define the same term a bit differently.

Figure 1. Categories of Teacher Assessment

Teacher Quality

Teacher quality can be thought of as those attributes the teacher brings to the classroom, including specialized knowledge. Some factors often included in this category (education, certification/licensure, and experience, for example) are frequently considered primarily during hiring, and so lie beyond the scope of this brief. However, there is widespread recognition that other personal qualities of teachers are important. Standards from both the National Council for Accreditation of Teacher Education (NCATE) and the Interstate New Teacher Assessment and Support Consortium (INTASC), for example, detail expected "dispositions."[10] Surveying existing literature, Thornton (2006) found that dispositions "often loosely equate to values, beliefs, attitudes, characteristics, professional behaviors and qualities, ethics and perceptions."[11] A common assumption is that teachers should be reflective, habitually monitoring their effectiveness and planning improvements.[12]

In the constellation of teacher characteristics receiving attention, teacher beliefs about students' capacity to learn are a particular concern because they shape a teacher's classroom choices.[13]

Recent studies have linked achievement gaps with negative teacher beliefs about students of color, students from low socioeconomic backgrounds, or both.[14] Changing teachers' negative preconceptions might even change classroom practice and help narrow achievement gaps.[15]

However, research has not as yet established the full complement of teacher characteristics that may affect student achievement. As Muñoz and Chang (2007) aptly summarize, "Teacher characteristics and student growth have an elusive relationship, but practice in the classrooms tells us that they are two intertwined concepts."[16] As these researchers note, policymakers will need "to make the best decision based on their particular context" about which teacher characteristics might be important to assess.

Teacher Performance

Teacher performance can be thought of as those things a teacher does, both inside and outside of the classroom. Because specialized knowledge does not automatically translate to effective classroom performance, it is necessary to assess not only what a teacher knows but also what a teacher can do. Teacher performance thus includes such instructional basics as how well a teacher plans learning activities, maintains a positive classroom environment, communicates with students, and provides productive feedback. It also includes activities outside the classroom, such as advising student groups, taking part in committees and other school-wide work, and communicating with parents.

To assess teacher performance requires having a set of performance criteria. For example, elements that Goe, Bell & Little (2008) consider essential include whether teachers "use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed . . . 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."[17] Kennedy (2008) includes as examples of relevant classroom practices "being organized, providing clear goals and standards, [and] keeping students on task"; as examples of typical practices outside the classroom, she includes "interacting with colleagues and parents, planning a curriculum that engages students, providing supervision to the chess club."[18]

Assessing a teacher's activities requires specifying clear criteria for desired behaviors. Often such criteria reflect the standards of professional organizations; many models are available.[19] To allow for variability in the teaching context, some models phrase expectations broadly enough to cover a wide range of activities in a wide variety of classroom contexts.[20] A broad goal, for example, might be that teachers "clearly state the goal of each class when it begins." Narrower guidelines are available in discipline-specific teaching standards formulated by several professional organizations, including those used for accrediting teacher education programs.[21] A discipline-specific criterion for language arts teachers, for example, might be that they "provide students practice in identifying and correcting common grammatical errors in their writing." Because assessment criteria can shape classroom behavior if performance assessment is well implemented, policymakers should choose them with great care.[22]

Teacher performance can be assessed. Heneman and colleagues (2006)[23] reviewed several studies of four sites implementing a well-known set of criteria (Danielson's 1996 *Framework for Teaching*)[24] and found that "the scores from standards-based performance evaluation systems can have a substantial positive relationship with student achievement and that the instructional practices measured by these systems contribute to student learning." There is also evidence, however, that the validity of evaluations varies significantly across evaluators, suggesting the importance of providing extensive training for evaluators and monitoring the credibility of their judgments.[25]

Teacher Effectiveness

Teacher effectiveness can be considered the result of a teacher's activities. It encompasses a wide range of outcomes, obviously including student learning. Academic achievement is critical, but as noted earlier, defining teacher effectiveness *only* in those terms ignores several other important ways that teachers affect students and the school community. The limitations of assessment based on student achievement are amplified when achievement is measured only by standardized test scores, with no consideration of such other classroom data as student projects, performances, papers, learning logs, and the like.

The current enthusiasm for using student test scores as the sole measure of teacher effectiveness stems from several sources—including convenience. Test score data are readily available because of NCLB requirements, and non-statisticians often perceive statistical analyses as objective, simple and reliable. Moreover, federal policy attaches high stakes to high scores, forcing school personnel to value them highly.

Also fueling the interest in test scores is the development of value-added modeling, which increases the capacity of researchers to isolate the effect of a single teacher from other influences on student achievement (such as prior teachers, home influences, school environment and student motivation). This modeling is sometimes known as Value-Added Assessment (VAA), and it uses complex formulas to estimate students' likely achievement gains in a given year. Actual gains are compared to this estimate, and classroom teachers are credited (or blamed) when students experience greater (or lesser) gains than expected.

However, while various VAA options exist, none is perfect:

> Trade-offs and risky assumptions are required in every case, so any given model is necessarily going to be imperfect. In the context of accountability, expectations for what any VAA-based tool can reasonably accomplish should be tempered, and the use of its estimates must be judicious.[26]

A steady stream of authoritative statements from the nation's foremost researchers has cautioned against the use of VAA to make high-stakes decisions,[27] both because of remaining methodological challenges and because "an overly narrow focus on standardized test scores as the most important—and in some cases, only—student outcome measure is not aligned with what the field agrees an effective teacher does."[28] Some researchers suggest including other important outcomes, such as whether students persist to graduation and whether they

demonstrate a positive attitude toward learning, toward themselves, and toward others. Another concern is whether students evidence a sense of engaged citizenship.[29]

A disincentive for including these latter types of attitudinal outcomes as a measure of teacher effectiveness comes from critics who have complained that the most important purpose of schools—to develop students' academic talents—has been elbowed aside by efforts to enhance students' self-esteem.[30] However, in 2009 the Educational Testing Service (ETS) sponsored a survey of existing research on the influence of noncognitive variables that found substantive empirical evidence indicating a correlation between achievement and student engagement (a category that includes such factors as student values and feelings).[31] That correlation was especially strong for reading and math.

While it is obvious that student learning should be factored into any assessment of teacher effectiveness, the overwhelming conclusion of top researchers is that value-added assessment alone is an invalid and unwise basis for making for high-stakes decisions. Just as teacher effectiveness should be combined with teacher quality measures and teacher performance measures, any measurement of teacher effectiveness that uses VAA should combine it with analyses of other evidence, such as classroom artifacts, student self-reports, parent surveys, and other key non-academic outcomes known to correlate with student learning.

Assessment Purposes: Summative and Formative

There are two very different purposes for assessment, each critical in its own right. Summative assessment is used to make a judgment, often a high-stakes decision—whether to award a teacher merit pay, for example, or whether to continue or terminate a teacher's employment. In contrast, formative assessment is used to gain information that can help teachers, even teachers who are already proficient, to improve or expand their abilities. Developing an excellent teaching force requires not only making good decisions about which teachers enter and remain in classrooms, but also finding ways to help teachers improve their skills.[32]

More than two decades ago, James Popham (1988) argued that each type of assessment is "splendid" in itself, but that they are "counter-productive when combined."[33] He summarized formative assessment as "fixing" the teacher, and summative assessment as "firing" the teacher, noting that "From the perspective of the teacher who is being fixed or fired, that distinction is profoundly important."[34] Assessment to improve practice requires that teachers be open to admitting weaknesses, which can happen only in a relatively non-threatening environment. In a formative situation, the evaluator functions as an ally, providing help to improve performance. When important career decisions are also to be based on the evaluation process, however, the environment may seem fraught with risk, especially for a teacher having significant difficulty.[35] In this case the evaluator functions as a potential enemy able to derail a career, and the assessment process may seem hostile. Teachers whose work can be improved but who are feeling at risk may understandably be inclined to hide, rather than confront, their problems—precluding valuable formative feedback.

Despite the inherent challenge of combining these assessment functions, a single system is frequently expected to serve both purposes, and often a single person—usually the principal—is

responsible for the assessment. Notwithstanding Popham's skepticism, one recent study suggests it may be possible. Milanowski (2005) divided new teachers in one district into two groups, one that received summative and formative feedback from a single source and one that received each type of feedback from a different source. He found "no major differences . . . in terms of openness to discussion of difficulties, reception and acceptance of performance feedback, stress, turnover intentions, actual turnover, or performance improvement."[36]

Moreover, some systems specifically designed to address both purposes have been successful. For example, locally developed systems of Professional Development Plans (PSPs) have shown promise for dual-purpose evaluation of experienced, competent teachers.[37] Also, peer assistance and review (PAR) strategies have shown promise for combined evaluation of both new and veteran teachers.[38] Other dual purpose systems have also been successful.[39]

While it appears that summative and formative assessment may be successfully combined, policymakers should remain aware of the challenges involved in doing so and address them as they plan.

What Measurement Tools Are Available?

Once *what* is to be assessed has been determined, policymakers can proceed to consider which measurement tools to use.[40] Several can be combined into comprehensive systems that assess multiple elements and provide multiple forms of data and judgments. Often, one tool can help offset weaknesses in another. For example, value-added assessments offer some information about student achievement but no information about what a teacher did to produce greater-than-expected (or less-than-expected) learning gains. Teacher observations, portfolios, and self-reports on classroom practice can help illuminate the important question of *how* gains were realized or losses were suffered. Moreover, these additional information sources may document high-quality teaching notwithstanding poor VAA results, and vice versa.

In a recent ambitious synthesis of research on teacher effectiveness,[41] Goe and co-authors organized assessment tools into seven categories and provided a useful table summarizing the purposes, benefits and drawbacks of each (reproduced in the Appendix below). Categories discussed here essentially parallel those of Goe and her colleagues, except that I have collapsed principal observation and classroom observation, yielding six (rather than seven) categories: *classroom observation, instructional artifacts, portfolios, teacher self-reports, student surveys,* and *value-added assessment*.[42]

Classroom observation

Classroom observation has long been a common method of teacher assessment, largely because it offers rich detail on a teacher's actual performance that can be used for both formative and summative purposes.[43] Classroom visits often take a class period or its equivalent, and procedures may be informal or highly structured to include the use of pre- and post-observation conferences.[44] Observers usually record their impressions of classroom events and characteristics, but calls for more objective evaluation have led to widespread use of observation

protocols. Many protocols are available, but while validity has been assessed for some of them,[45] Goe and colleagues caution that "[t]he degree to which observations can or should be used for specific purposes depends upon the instrument, how that instrument was developed, the level of training and monitoring raters receive, and the psychometric properties of the instrument."[46]

Based on their research, Kimball & Milanowski (2009) also caution that

> Providing evaluators with relatively detailed rubrics or rating scales describing generic teaching behaviors thought to promote student learning, coupled with initial training in applying them, is not enough to ensure that all evaluators' ratings will be positively related to student achievement.[47]

Confounding factors can include whether observers, especially principals, have enough time to do a thorough classroom assessment, whether they have sufficient familiarity with the wide variety of subjects and grades they must assess, and whether they are adequately trained in the use of the instruments.[48] Some worry that protocols may force observers to base judgments on overly narrow and prescriptive lists of teacher behaviors.[49] Ongoing training is necessary to ensure that observers apply criteria consistently over time.

Instructional artifacts

Instructional artifacts include a wide range of classroom-related materials, such as lesson plans, assignments, handouts, student work (including class work, homework, projects, and exams), scoring rubrics, and pictures of such classroom elements as writing on the board. Like observation, artifacts offer authentic evidence from the classroom and provide substantive detail on actual classroom activity, but analyzing them is less time consuming than is classroom observation. And, while teachers must spend time selecting artifacts, they need not generate any new materials.

As with observation, protocols are available to guide analysis. Different protocols target different criteria, such as how well materials reflect standards or the degree of intellectual challenge. Little peer-reviewed research has yet been conducted on artifact analysis as a credible means of teacher assessment, and criticisms of it have not yet been resolved.[50] Yet there is some evidence of promise for this measurement tool. For example, the National Center for Research on Evaluation, Standards, and Student Testing at UCLA has conducted several pilot studies on an instrument it has developed, the Instructional Quality Assessment, and found correlations with observation assessments, student work, and standardized achievement scores.[51] Researchers at the Consortium on Chicago School Research have similarly developed the Intellectual Demand Assignment Protocol; one study of this instrument found positive correlations between higher-scoring assignments and higher student achievement.[52] Pilot studies of a tool called the "Scoop Notebook" have also shown promise.[53]

More research is needed, but artifact analysis may be an informative part of a broader assessment system.

Portfolios

Portfolios include classroom artifacts, like those listed above, as well as a broader range of materials, such as samples from a teacher's journal or a statement of personal teaching philosophy—materials not in evidence in the classroom but nonetheless relevant to the teacher's activities.

The careful selection of evidence to build a coherent portrait of classroom performance requires extensive reflection by the teacher; formative evaluation and ongoing professional development are an inherent part of the portfolio approach. Ideally, teachers build portfolios gradually over time, so that their growth is evident. As for other tools, it is essential that teachers and assessors both have a clear understanding of the criteria by which the portfolio will be judged.

In a review of contextualized assessment tools (2000), Darling-Hammond and Snyder summarize the potential of portfolios:

> As assessment tools, portfolios that are structured around standards of practice are able to examine a teacher's practice both in context and in the light of a common set of expectations and benchmarks. By giving assessors access to teachers' thinking as well as to evidence of their behaviors and actions (e.g. through videotapes, lesson plans, assignments, and the like), portfolios permit the examination of teacher deliberation, along with the outcomes of that deliberation in teacher's actions and student learning.[54]

Because of the rich potential of portfolios to provide insight into multiple facets of the teacher's performance, they have become increasingly popular. Vermont, Connecticut, Washington state and Wisconsin have all adopted portfolio-based teacher assessment systems at some level.[55]

The very complexity of portfolios, however, can make them difficult to assess, and more research on their reliability and validity for assessment purposes is necessary before they should play a major role in accountability systems.[56] Links between portfolios and achievement have been found in the National Board for Professional Teaching Standards (NBPTS) system (discussed below), but other studies have not established a connection.[57]

Teacher self-reports

Teacher self-reports can be extremely valuable because teachers have unique, detailed information on such important elements as classroom context and teacher intentions. For example, observers can say *what* a teacher did but may have little understanding of *why*, an important consideration when assessing whether instruction has been effective or whether the teacher makes good instructional decisions. Moreover, self-reports can offer insight into the findings of other assessment measures, such as achievement scores, and so help identify appropriate professional development or other improvements.

Self-reports can take several forms, including surveys, teaching journals or logs, and interviews. These reports may be relatively unstructured or highly structured, and they may explore fairly generic topics (such as assessment practices) or very specific ones (such as how a particular math concept is taught). Studies on the validity of self-reports have yielded mixed results, and some are

concerned about the normal human tendency to make favorable self-reports.[58] In addition, there is some evidence that data collected only annually, which is highly dependent on long-term memory, is less reliable than data collected more frequently, as when teachers report on a specific day at its end.[59] Another concern is that teachers and others may have different understandings of the same terms (*challenging* or *successful*, for example), confounding results.[60]

Questions of validity concerning self-reports preclude using them as a primary basis for high-stakes decisions. However, they are relatively inexpensive, can yield detailed information useful in both formative and summative assessment, and can promote reflection and professional development. Moreover, incorporating teacher self-reports conveys the important message that the contextual knowledge of practitioners is respected and valued, and so helps to promote stakeholder buy-in.

Student surveys

Although some adults may have reservations about the ability of students to assess their teachers, there is some persuasive evidence that student surveys, even at the elementary school level, can be valid sources of information. A 1995 study based on a review of research on elementary students' teacher ratings found evidence that elementary students are "no more vulnerable than others to rating leniency and halo" (extending one positive characteristic into a positive global rating).[61] Like teachers, students have unique knowledge of the classroom.

One study involving nearly 1,000 teachers, 35 teachers and four principals found student ratings of teachers to be good predictors of student achievement as measured by the district's criterion-referenced examinations.[62] The quality of a survey instrument used will, of course, affect results. In this case, researchers worked with instruments that had demonstrated validity and reliability in prior research. Another study involving over 400 teachers in 27 K-12 schools similarly found that student ratings were both reliable and valid. The researchers concluded that student assessments "are not popularity contests" and that students can and do "distinguish between merely liking a teacher and recognizing one who enables their learning."[63]

As with other instruments, even proponents of student ratings don't recommend using them as the sole source of information, but rather as part of a comprehensive assessment system. "[H]igh student ratings do not necessarily mean the same thing as good teaching. Perhaps the best interpretation is that high student ratings in conjunction with at least several other indicators are a good indicator of quality teaching."[64]

Value-added assessment

As noted above, value-added assessment (VAA) is a means of measuring how much academic growth can be linked to a particular teacher. Complex formulas predict the amount of growth for students in a given year, and particular teachers are assumed to be responsible for students meeting, exceeding, or missing the expected gains.

There are many influences on student learning beyond the teacher, however, and early versions of VAA were criticized as unable to control adequately for such influences as the socioeconomic

status of students and schools and for the validity of the tests used to measure achievement.[65] Researchers have been working for years on improving VAA formulas, and enough progress has been made to allow for distinguishing between particularly strong and particularly weak teachers—as long as one accepts the importance of test scores as an outcome.[66] However, no perfect formula has yet been devised:

> The myriad factors that influence cognitive growth over extended periods of time, the purposeful sorting of families and teachers into schools and classrooms, compensatory behavior on the part of families, and the imperfections of tests as measures of knowledge complicate efforts to estimate measures of teacher effectiveness, including the overall variance of teacher value added and the ranking of teachers by quality of instruction. Even within-school rankings are subject to biases and the vagaries of sampling variability. Along with possible distortions of classroom time allocation and teaching methods in an effort to increase scores, these problems raise concerns about the use of tests for high stakes purposes (p. 534).[67]

An additional limitation of VAA, as noted earlier, is that it provides no information on what a teacher may or may not be doing to produce specific scores, and therefore it therefore offers no information helpful for improving practice. This weakness may, however, be offset by complementary observational data.[68]

Like the other tools catalogued here, VAA may provide useful information as part of a broader assessment system using multiple sources of data, but is not in itself a reliable method for assessing teacher performance.[69]

Who Should Assess?

For many years, teacher assessment was routinely the responsibility of the principal (or assistant principal). However, as suggested earlier, relying on a single administrator for teacher assessment has proven problematic and has been criticized as well for failing to identify weak teachers.[70] While some recent research suggests that principals can be effective assessors, there has been growing interest in newer alternatives—fueled not only by weaknesses in the traditional approach but by increasing calls for teachers to monitor peer performance, as is common in other professions.[71] Several systems have evolved that distribute responsibility for high-stakes decisions among multiple stakeholders and that give teachers themselves key roles in both formative and summative assessment.

Thus, policymakers designing assessment systems must make choices about who will be involved in teacher evaluation. This section provides a brief review of research relevant to that issue.

Principal Ratings

Summarizing criticism of traditional evaluation by principals, Calabrese and colleagues (2004)[72] identified several common themes: that principals' ratings do not adequately identify poor and marginal teachers; that the process is a time-wasting ritual with little or no effect on personnel

decisions and staff development; and that experienced teachers are often dissatisfied with evaluators' skill and feedback as well as their failure to connect assessment to professional development. Just as it is important to take contextual influences into account when considering student achievement, however, it is also important to consider how context may influence principal ratings.

In their study involving 80 classroom teachers and eight principals, Calabrese and colleagues found that both groups had negative feelings about their district's traditional "top down" assessment process. However, in analyzing subjects' comments, the researchers identified the mandated use of a particular rating instrument and the lack of opportunity to provide detailed and useful feedback as the real problem. While teachers often blamed principals for stressful and ineffective evaluations, principals saw themselves as victims of a system imposed upon them that they had no voice in designing. Thus, this study suggests that the weaknesses often ascribed to principals may be linked instead to a poor, externally imposed process.

> Like many of their colleagues throughout the US, [teachers in this study] endure evaluation systems based on reward or punishment. Teachers endure this process and continue to develop a deepening resentment toward principals who are systematically forced to participate....Principals...found themselves caught in an enigma. On the one hand, they desired the less adversarial formative role; on the other hand, they had no choice, but to operate as the summative evaluator.[73]

This study suggests that if the assessment process were more collaborative, principal ratings might be more useful and better received. The teachers and principals surveyed professed the same goals for assessment—accountability and an effective aid to professional growth—but the

Although context is critical, and although some principals may be uncomfortable giving strong negative feedback, principals can provide valuable assessments.

imposition of a rigid structure subverted them.[74]This would suggest that a preliminary concern with relying on principals for evaluation is the need to ensure good conditions. Principals themselves appear to believe that major barriers to effective evaluation include insufficient time, tenure, and restrictive rules.[75] The strength of an observational protocol or other data collection tool can also affect principals' ratings, as can the amount and quality of training a principal receives (if any) in collecting and interpreting data, the consequences of the evaluation, and whether the principal is held accountable for the quality of the evaluations. Another concern is whether a principal has the necessary subject-area knowledge for all disciplines.

Although context is critical, and although some principals may be uncomfortable giving strong negative feedback,[76] principals can provide valuable assessments. In a 2008 study, Jacob and Lefgren found that principals can reliably identify both the weakest and strongest teachers, even though they are less able to make fine distinctions in the middle range.[77] In these researchers' view, the "findings provide compelling evidence that good teaching is, at least to some extent, observable by those close to the education process, even though it may not be easily captured in those variables commonly available to the econometrician."[78] They also note that principal

observation can mitigate concerns about teachers pursuing improved test scores at the expense of meaningful learning.

Research on the correlation between principal ratings and student achievement is mixed, however, and it is possible that using value-added measures and principal observation together might better predict student achievement than using either alone.[79]

Peer Review

Historically perceived as women's work, teaching has suffered from both low status and low pay. Yet there has been growing support for recognizing teachers as skilled professionals uniquely qualified to assess their colleagues, as do professionals in other fields. Calls for a more collaborative assessment process with more emphasis on professional development have fueled interest in—and often union support for—involving teachers in assessment.[80] There is, however, scant empirical research on peer review, which can take many different forms. To offer a sense of the possibilities for peer assessment and of the existing research, two of the best-known programs are briefly described here.

Peer Assistance and Review (PAR)

Peer Assistance and Review, widely known as PAR, first appeared in the early 1980s in Toledo, Ohio. The design calls for a joint union-administration panel to administer a program in which experienced, highly skilled teachers serve as mentors and primary assessors for new teachers, for veteran teachers having difficulty, or for both. Because the reviewing teachers, often called consulting teachers or CTs, are released from their own classrooms, there is substantial cost involved for classroom replacements.

The PAR system depends on a clear system of expectations and all stakeholders having a shared understanding of them. CTs both help mentees meet standards and assess their progress; their recommendations to rehire or terminate a teacher carry great weight with the oversight panel making the ultimate decision.

Credible research on PAR is growing, with preliminary findings showing that, while the cost per teacher in a PAR program was $4,000 to $7,000, it also created savings including efficiencies from higher retention of new teachers and lower arbitration and dismissal costs.[81] In addition, stakeholders "felt strongly that PAR not only was a worthwhile investment but that it also saved the district money."[82]

Additional research has focused on implementation of 1999 PAR legislation in California. In one district studied for several years, PAR increased dismissals, and "[t]he community of educators created by PAR and the PAR panel appears to have proved a more rigorous, evidence-based check on classroom teaching performance."[83] The accountability provided by the oversight panel appears crucial in providing support for the CTs who provide summative evaluations of colleagues.

Other Exemplars Employing Peer Review

While PAR operates completely within a district, peer review may also be structured through external agencies. In Connecticut, for example, the state's Beginning Educator Support and Training (BEST) program requires beginning teachers to submit second-year portfolios, scored by a cohort of experienced teachers that the state trains and employs as assessors. An in-depth study of BEST has found substantive gains in student achievement in mathematics and literacy, which appear to be linked to Connecticut's teacher assessment policies.[84, 85]

While these programs tend to focus on less experienced teachers, the National Board for Professional Teaching Standards (NBPTS) offers national certification to highly skilled teachers through another type of peer-review process. Applicants must take subject-matter tests and submit a detailed portfolio that includes such materials as classroom videotapes, written analyses of teacher objectives, and artifacts demonstrating student learning. The portfolios are reviewed by teachers accomplished in the same subject and at the same high experience and skill level as the candidates.

Research on the outcomes of NBPTS certification is mixed,[86] but when the National Research Council reviewed all available studies and issued a report, it concluded that "national board certification distinguishes more effective teachers from less effective teachers with respect to student achievement. The differences are small (and not entirely consistent) in absolute terms, but when considered in terms of teacher value-added contributions to achievement, they are substantively meaningful."[87]

As is evident from such programs, teachers themselves may play an important role in the assessment of their peers. Those designing comprehensive teacher evaluation systems should seriously consider including this element.

Systemic Models

Because every tool for assessing teacher performance has both strengths and weaknesses, and because assessment can have multiple goals, it is better to develop a comprehensive assessment system than to adopt a single measure of performance. Several such systems have already been developed, and experience suggests they have promise for helping to nurture and promote a highly skilled teaching staff.

Given the recent interest in merit pay, the National Education Association (NEA) recently commissioned a review of research literature on linking assessment systems to teacher compensation. The review paid particular attention to the impact of specific assessment systems on both student achievement and achievement gaps.[88] It identified five programs as "promising approaches to improving instruction, raising student achievement, gaining teacher support, increasing retention by taking a comprehensive rather than piecemeal approach to reform, and centering activities and procedures around instructional improvement and student learning."[89] Following is a brief summary of that study's findings on each program.[90]

Three of the models identified as promising have already been discussed here. Danielson's Framework for Teaching (FFT) has the longest history and appears most often in the research

It is important to note that outcomes are determined by implementation as well as by design.

literature, and its scores have been found to be positively correlated with value-added measures of student achievement. PAR research has found major advantages to distributed responsibility for personnel decisions and extensive contact between consulting teachers and mentees.[91] Connecticut's BEST has also been found to have positive results.[92]

The remaining two promising programs identified in the NEA report are the Teacher Advancement Program (TAP) and Denver's Professional Compensation System (ProComp). TAP was designed by Lowell Milken of the Milken Family Foundation. It integrates assessment within a system linking accountability to compensation.[93] While some results are promising, the program is complex, and more outside research is needed on its effects over time and in more varied contexts.[94] The ProComp system, developed collaboratively by the Denver district and union leaders, is also tightly linked to compensation.[95] It stresses teachers developing and then striving to meet high-quality objectives for student learning, and it financially rewards teachers for realizing them. Olivia Little, the author of the NEA report, cites as particular advantages of ProComp its flexibility, choice and varied options, and she finds the model informative in terms of fostering collaboration and stakeholder support for a new assessment system. An independent assessment of ProComp's effects on achievement is underway; a preliminary report has identified some positive trends in outcomes.[96]

An earlier guide for policymakers by Linda Darling-Hammond and Cynthia Price (2007)[97] also noted several promising systems. In addition to BEST, TAP, PAR and ProComp, these researchers cite the NBPTS' national certification as an effective assessment.

While research indicates promise for these programs, it is important to note that outcomes are determined by implementation as well as by design. Different schools or districts may implement a program with greater or lesser fidelity to the design and with more or less commitment to key components.

Discussion

A teacher evaluation system focused solely on high-stakes decisions like tenure or compensation will not meet contemporary needs. If each classroom is to be staffed with a highly skilled teacher, an assessment system must do more than weed out weak teachers. As explained by Darling-Hammond and Prince:

> Clearly, meeting the expectation that all students will learn to high standards will require a transformation in the ways in which our education system attracts, prepares, supports, and develops expert teachers who can teach in more powerful ways.

> An aspect of this transformation is developing means to evaluate and recognize teacher effectiveness throughout the career, for the purposes of licensing, hiring, and granting tenure; for providing needed professional development; and for identifying expert teachers who can be recognized and rewarded. A goal of such recognition is to keep talented teachers in the profession and to identify those who can take on roles as mentors, coaches, and teacher leaders who develop curriculum and professional learning opportunities, who redesign schools, and who, in some cases, become principals.[98]

It is important, then, for policymakers to think clearly about the assessment needs and goals of their particular context, to make careful decisions among options, and to commit sufficient resources for successful implementation.

Since any teacher assessment system must address multiple goals, it should rely on multiple sources of information. At the moment, value-added assessment is being strongly promoted as a primary indicator of teacher effectiveness. However, policymakers should remember that good

Since any teacher assessment system must address multiple goals, it should rely on multiple sources of information.

policy requires a sturdier base than momentary popularity. No value-added model provides a sufficient and reliable indicator of teacher effectiveness. Adding to an already overwhelming consensus, the Economic Policy Institute recently convened a panel of the nation's top experts, who reached the following conclusion:

> A review of the technical evidence leads us to conclude that, although standardized test scores of students are one piece of information for school leaders to use to make judgments about teacher effectiveness, such scores should be only part of an overall comprehensive evaluation. Some states are now considering plans that would give as much as 50% of the weight in teacher evaluation and compensation decisions to scores on existing tests of basic skills in math and reading. Based on the evidence, we consider this unwise. Any sound evaluation will necessarily involve a balancing of many factors that provide a more accurate view of what teachers in fact do in the classroom and how that contributes to student learning. . . . [T]here is broad agreement among statisticians, psychometricians, and economists that student test scores alone are not sufficiently reliable and valid indicators of teacher effectiveness to be used in high stakes personnel decisions, even when the most sophisticated statistical applications such as value-added modeling are employed.[99]

Policymakers interested in reliable teacher assessment must look beyond value-added scores, no matter how enticing some claims appear.

Recommendations for Developing a Teacher Assessment System

Based on the research reviewed, it is recommended that policymakers employ an assessment system that targets both continual improvement of the teaching staff and timely dismissal of teachers who cannot or will not improve. Steps toward that goal include that policymakers:

- **Be clear about the purposes of any assessment before selecting strategies. Where formative and summative assessment are to be combined, plan to address the challenges of dual-purpose systems.**
- **Involve all key stakeholders in system design.**
- **Rather than employing a single assessment tool, gather evidence from multiple sources. Combine strategies so that the weakness of any single tool is offset by the strengths of another.**
- **Be sure that the criteria for assessing performance, artifacts or other factors are credible and are well understood by teachers and assessors.**
- **Provide high-quality, ongoing training for assessors and routinely calibrate their efforts to ensure consistent application of criteria.**
- **Look to high-quality research on existing tools and programs to inform the design of assessment systems.**
- **Commit sufficient resources to produce high-quality, productive assessment.**

Notes and References

[1] See, for example, Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, 8(1). Retrieved August 10, 2010 from http://epaa.asu.edu/ojs/article/viewFile/392/515.

[2] See, for example:

Yariv, E. (2006). "Mum effect": Principals' reluctance to submit negative feedback. *Journal of Managerial Psychology*, 21(6): 533-546.

Painter, S. Barriers to evaluation: Beliefs of elementary and middle school principals. *Planning and Changing*, 32(1): 58-70.

[3] Banchero, S. (2010, June 1). Race to top leaves some school reformers weary. *Wall Street Journal* (Online). Retrieved from AB/INFORM Global database June 22, 2010.

[4] See, for example:

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved October 13, 2010, from http://epaa.asu.edu/ojs/article/view/810.

National Academies. (2009, October 7). Education innovations funded by 'race to the top' should be rigorously evaluated. [Press release]. Retrieved October 13, 2010, from http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=12780.

Darling-Hammond, L., & Prince, C. D. (2007). *Strengthening teacher quality in high-need schools: Policy and practice*. Council of Chief State School Officers.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59-63.

[5] Nichols, S.L. & Berliner, D. (2007). *Collateral damage: How high stakes testing corrupts America's schools*. Cambridge, MA: Harvard University Press.

[6] For a detailed discussion of the role of values in teacher assessment, see Campbell, R.J., Kyriakides, L., Muijs, D. and Robinson, W. (2004). Effective teaching and values: Some implications for research and teacher appraisal. *Oxford Review of Education*, (30)4: 451-465. Retrieved June 26, 2010 from JSTOR.

[7] Databases accessed during electronic searches included the Educational Resources Information Center (ERIC) via CSA Illumina, AB/INFORM Global (ProQuest), and Teachers College Record. Websites consulted as a source of useful references included those of the National Comprehensive Center for Teacher Quality, the Center for Teaching Quality, the Consortium for Policy Research in Education, the Consortium for Research on Educational Accountability and Teacher Evaluation, and the National Network for the Study of Education Dispositions.

Keyword searches involved various combinations of several terms, including: teacher, assessment, performance evaluation, performance appraisal, teacher attitudes, teacher dispositions, teacher quality, teacher characteristics, teacher performance, walk through, observation, observation protocols, classroom observation, portfolios, artifacts, value added, self-assessment, self-study, principal, videotapes, student surveys, and self reports.

[8]See, for example, Dillon, S. (2006, March 26). Schools cut back subjects to push reading and math. *New York Times*. Retrieved June 22, 2010 from http://www.nytimes.com/2006/03/26/education/26child.html.

[9] See, for example, Darling-Hammond, L., & Prince, C. D. (2007). *Strengthening teacher quality in high-need schools: Policy and practice*. Council of Chief State School Officers.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59-63.

[10] For NCATE standards, see http://www.ncate.org/public/standards.asp; for INTASC standards, see http://www.ccsso.org/content/pdfs/corestrd.pdf.

[11] Thornton, H. (2006). Dispositions in action: Do dispositions make a difference in practice? *Teacher Education Quarterly*, 33(2), 53-68, (¶4). Retrieved June 23, 2010, from ProQuest Education Journals.

[12] This assumption flows from a long line of work on reflective practice sparked by Donald Schön's seminal 1984 text, *The reflective practitioner: How professionals think in action*. New York: Basic Books.

[13] Gill, M.G. & Hoffman, B. (2009). Shared planning time: A novel context for studying teachers' discourse and beliefs about learning and instruction. *Teachers College Record*, 111(5): 1242-1273. Retrieved July 19, 2010, from http://www.tcredord.org/library [ID Number 15241].

[14] Lynn, M., Bacon, J.N., Totten, T.L., Bridges, T.L.III, & Jennings, M.E. (2010). Examining teachers' beliefs about African American male students in a low-performing high school in an African American school district. *Teachers College Record*, 112(1): 289-330.

Love, A. & Kruger, A.C. (2005). Teacher beliefs and student achievement in urban schools serving African American students. *Journal of Educational Research*, 99(2): 87-99.

[15] Johnson, C. (2009). An examination of effective practice: Moving toward elimination of achievement gaps in science. *Journal of Science Teacher Education*, 20(3): 287-306.

On the influence of beliefs on practice, see also Airasian, P.W. & Horn, S. (2000). CREATE NEWS: Consortium for research on educational accountability and teacher Education. *Journal of Personnel Evaluation in Education*, 14(4), 319+. +? Retrieved July 9, 2010, from SpringerLink database.

[16] Muñoz, M. A. & Chang, F. C. (2007). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal of Personnel Evaluation in Education*, 20(3-4): 147-164. Retrieved June 26, 2010, from SpringerLink database.

[17] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality, 8.

[18] Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 61.

[19] Such standards and models often specify relevant teacher knowledge as well, but they are discussed as part of the performance category because they generally stress what a teacher does more than what a teacher knows.

[20] See, for example, Charlotte Danielson's Framework for Teaching, http://www.danielsongroup.org/theframeteach.htm; the Interstate New Teacher Assessment and Support Consortium (INTASC) standards, http://www.wresa.org/Pbl/The%20INTASC%20Standards%20overheads.htm.

[21] See, for example, Science Teaching Standards, produced by the National Research Council, http://www.nap.edu/openbook.php?record_id=4962, or those from the National Council of Social Studies, http://www.socialstudies.org/standards/teacherstandards.

[22] For a detailed exploration of weaknesses in the criteria-based approach to assessing performance, see Stickney, J.A. (2009). Wittgenstein's contextualist approach to judging "sound" teaching: Escaping enthrallment in criteria-based assessments. *Educational Theory*, 59(2), 197-215.

[23] Heneman, H.G.III, Milanowski, A. Kimball,S.M., and Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay. [CPRE Policy Brief]. Retrieved July 12, 2010 from http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf.

[24] See http://www.danielsongroup.org/theframeteach.htm.

[25] Kimball, S.M. & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making with a standards-based evaluation system. *Education Administration Quarterly*, 45(1): 34-70.

[26] Wiley, E. W. (2006). *A practitioner's guide to value added assessment*. Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved July 7, 2010 from http://www.epicpolicy.org/files/Wiley_APractitionersGuide.pdf.

[27] See, for example,

Baker, E.L., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J. & Shepard, L. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers.* Briefing Paper # 278. Washington, DC: Economic Policy Institute. Retrieved December 7, 2010 from http://www.epi.org/publications/entry/bp278.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved October 13, 2010, from http://epaa.asu.edu/ojs/article/view/810.

National Academies. (2009, October 7). Education innovations funded by 'race to the top' should be rigorously evaluated. [Press release]. Retrieved October 13, 2010, from http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=12780.

Darling-Hammond, L. (2007). *Recognizing and enhancing teacher effectiveness: A policy maker's guide.* Washington, DC: Council for Chief State School Officers.

Martineau, J.A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, 31(1): 35-62.

Braun, H. (2005). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS.

McCaffrey, D., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). Evaluating the value-added models for teacher accountability (Report MG-158). Santa Monica, CA: Rand Corporation.

[28] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality, 6.

[29] See, for example: Rubin, B.C., Hayes, B. & Benson, K. (2009). "It's the worst place to live": Urban youth and the challenge of school-based civic learning. *Theory Into* Practice, 48(3): 213-221; Callahan, R.M., Muller, C., & Schiller, K.S. (2008). Preparing for citizenship: Immigrant high school students' curriculum and socialization. *Theory and Research in Social Education*, 36(2): 6-31.

[30]Typical of such criticism is the work of E.D. Hirsch, Jr., including The Schools We Need and Why We Don't Have Them (1999) and The Knowledge Deficit: Closing the Shocking Education Gap for American Children (2007). Less scholarly, but potentially influential, criticism is also widely available in such works as: Stout, M. (2001). The feel-good curriculum: The dumbing down of America's kids in the name of self-esteem. New York: De Capo Press.

[31] Lee, J. & Shute, V.J. (2009). The influence of noncognitive domains on academic achievement in K-12 [ETS RR-09-34]. Princeton, NJ: ETS.

[32] See, for example, Darling-Hammond, L., & Prince, C. D. (2007). *Strengthening teacher quality in high-need schools: Policy and practice*. Council of Chief State School Officers.

[33] Popham, W.J. (1998). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education*, 1: 269-273. Retrieved July 8, 2010, from the SpringerLink database.

[34] Popham, W.J. (1998). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education*, 1: 269-273. Retrieved July 8, 2010, from the SpringerLink database.

[35] Conley, S. & Glasman, N.S. (2008). Fear, the school organization, and teacher evaluation. *Educational Policy*, 22(1): 63-85.

[36] Milanowski, A.T. (2005). Split roles in performance evaluation—A field study involving new teachers. *Journal of Personnel Evaluation in Education*, 18(3), 153-169.

[37] Holland, P.E. & Adams, P. (2002). Through the horns of a dilemma between instructional supervision and the summative evaluation of teaching. *International Journal of Leadership in Education*, 5(3), 227-247. Retrieved July 8, 2010 from Ebscohost database.

[38] See, for example, Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education*, 113(2): 479-508.

[39] Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform*. Washington, DC: National Education Association. Retrieved July 8, 2010, from http://forum.mdischools.net/sites/default/files/forum.mdischools.net/2009_NEA_teacherevaluationsystems.pdf.

Heneman, H.G.III, Milanowski, A. Kimball,S.M., and Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay*. [CPRE Policy Brief]. Retrieved July 12, 2010, from http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf.

[40] The following overview intends primarily to catalogue tools validated by credible research; for a more detailed and technical discussion, readers are advised to consult the work cited above.This overview intends primarily to catalogue tools validated by credible research; for a more detailed and technical discussion, readers are advised to consult the

cited work: Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[41] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[42] For a detailed discussion of choices and issues involved with classroom observation, see Weade, G. & Evertson, C.M. (1991). On what can be learned by observing teaching. *Theory into Practice*, 30(1): 37-45.

[43] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[44] Although brief visits known as "walk-throughs" have been discussed as a means of formative evaluation (perhaps in part because traditional classroom observation is known to be particularly time-consuming), no studies of their effects appear in research literature. For a brief description, see http://www3.learningpt.org/tqsource/GEP/GEPEvalType.aspx?tid=1.

[45] National Comprehensive Center for Teacher Quality. (n.d.) Guide to teacher evaluation products. Retrieved July 15, 2010, from http://www3.learningpt.org/tqsource/GEP/GEPEvalType.aspx?tid=1.

[46] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[47] Kimball, S.M. & Milanowski, A.(2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, 45(1), 34-70. Retrieved July 11, 2010 from the Sage database.

[48] Kimball, S.M. & Milanowski, A.(2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, 45(1), 34-70. Retrieved July 11, 2010 from the Sage database.

[49] See, for example, Darling-Hammond, L. & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher* Education, 16: 523-524; Peterson,K.D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice* (2nd ed). Thousand Oaks, CA: Corwin Press; Prybylo, D. (1998). Beyond a positivist approach to teacher evaluation. *Journal of School Leadership*, 8(6): 558-583.

[50] Denner, P.R., Salzman, S.A., & Bangert, A. (2001). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education*, 15(4): 287-307.

[51] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[52] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[53] Borko, H., Stecher, B. M., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The scoop notebook and rating guide* (CSE Tech. Rep. No. 707). Los Angeles: Center for Evaluation, Standards and Student Testing (CRESST), University of California at Los Angeles. Accessed October 13, 2010 from http://www.cse.ucla.edu/products/reports/r707.pdf ; 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.

[54] Darling-Hammond, L. & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher* Education, 16: 523-524.

[55] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[56] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[57] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[58] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality, 39.

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

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

[61] Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal*, 1(57), 57-78.

[62] 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, p. 179. *Journal of Personnel Evaluation in Education*, 14(2), 179-192. Retrieved July 30, 2010, from ABI/INFORM Global.

[63] Peterson, K.D, Wahlquist, C. & Bone, K. (2000). Student Surveys for School Teacher Evaluation, p. 148. Journal of Personnel Evaluation in Education, 14(2), 135-153. Retrieved July 30, 2010, from ABI/INFORM Global.

Similar results were also found in 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), 237-247 ; and in Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal*, 1(57), 57-78.

[64] Peterson, K.D, Wahlquist, C. & Bone, K. (2000). Student Surveys for School Teacher Evaluation, p. 148. Journal of Personnel Evaluation in Education, 14(2), 135-153. Retrieved July 30, 2010, from ABI/INFORM Global.

[65] See, for example, Berk, R.A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education*, 1(4), 345-363.

[66] Harris, D.N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, 4(1): 319-350.

[67] Ishii, J. & Rivkin, S.G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy*, 4(1): 520-536.

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

[69] See, for example:

Baker, E.L., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J. & Shepard, L. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers*. Briefing Paper # 278. Washington, DC: Economic Policy Institute. Retrieved December 7, 2010 from http://www.epi.org/publications/entry/bp278.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved October 13, 2010, from http://epaa.asu.edu/ojs/article/view/810.

National Academies. (2009, October 7). Education innovations funded by 'race to the top' should be rigorously evaluated. [Press release]. Retrieved October 13, 2010, from http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=12780.

[70] Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. New York: The New Teacher Project. Retrieved October 13, 2010, from http://widgeteffect.org/

[71] See, for example, Darling-Hammond, L. (1986). A proposal for evaluation in the teaching profession. *The Elementary School Journal*, 86(4): 530-551.

[72] Calabrese, R.L., Sherwood, K., Fast, J. & Womack, C. (2004). Teachers' and principals' perceptions of the summative evaluation conference: An examination of Model I theories-in-use. *The International Journal of Educational Management*, 18(2/3): 110.

[73] Calabrese, R.L., Sherwood, K., Fast, J. & Womack, C. (2004). Teachers' and principals' perceptions of the summative evaluation conference: An examination of Model I theories-in-use. *The International Journal of Educational Management*, 18(2/3): 116-17.

[74] Other studies have also demonstrated that context can impede the performance of principals as primary evaluators. See, for example:

Cooper, B.S., Ehrensal, P.A.L., & Bromme, M. (2005). School-level politics and professional development: Traps in evaluating the quality of practicing teachers. *Educational Policy*, 19(1): 112-125.

Painter, S. (2001). Barriers to evaluation: Beliefs of elementary and middle school principals. *Planning and Changing*, 32(1-2): 58-70.

[75] Painter, S. (2001). Barriers to evaluation: Beliefs of elementary and middle school principals. *Planning and Changing*, 32(1-2): 58-70.

[76] Yariv, E. (2006). "Mum effect": principals' reluctance to submit negative feedback. *Journal of Managerial Psychology*, 21(6): 533-546.

[77] 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.

Jacob, B.A. & Lefgren, L. (2006). When principals rate teachers: The best—and the worst—stand out. *Education Next*, 6(2): 58-64.

[78] Jacob, B.A. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 130.

[79] Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality, 27.

[80] For a review of recent expansion, see: Sawchuk, S. (2009, Nov. 18). Judging their peers: An old concept that calls for teachers to assess their own is gaining traction as evaluation comes under the spotlight. *Education Week*, 29(12): 20-23.

[81] Munger, M.S., Johnson, S.M., Fiarman, S.E., & Papay, J.P. (2009, April). Beyond dollars and cents: The costs and benefits of teacher peer assistance and review. Paper presented at the annual meeting of the American Educational Research Association, San Diego. Accessed August 11, 2010 from http://www.gse.harvard.edu/~ngt/new_papers/JPP_AERA_2009.pdf.

[82] Munger, M.S., Johnson, S.M., Fiarman, S.E., & Papay, J.P. (2009, April). Beyond dollars and cents: The costs and benefits of teacher peer assistance and review. Page 9. Paper presented at the annual meeting of the American Educational Research Association, San Diego. Accessed August 11, 2010 from http://www.gse.harvard.edu/~ngt/new_papers/JPP_AERA_2009.pdf.

[83] Goldstein, J. (2009). Designing transparent teacher evaluation: The role of oversight panels for professional accountability. *Teachers College Record*, 111(4), 893.

See also Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education*, 113(3): 479-508.

[84] Wilson, S.M., Darling-Hammond, L. & Berry, B. (2001). *A case of successful teaching policy: Connecticut's long-term efforts to improve teaching and learning*. Retrieved August 11, 2010 from http://depts.washington.edu/ctpmail/PDFs/Connecticut-WDHB-02-2001.pdf.

[85] BEST has been replaced by a new system, the Teacher Education and Mentoring (TEAM) program, currently in its first year of implementation. The new effort involves "guided" mentoring and the completion of specific "instructional modules." Effects remain to be seen. An overview of the new program is available online at http://www.sde.ct.gov/sde/lib/sde/pdf/team/team_program_guidelines_adoption_board_report_06022010.pdf.

[86] See the NBPTS website for an overview of studies and the mixed results produced: http://www.nbpts.org/resources/research.

[87] Hakel, M.D., Koening, J.A., & Elliott, S. W. (Eds.) (2008). *Assessing accomplished teaching: Advanced-level certification programs*. Washington, DC: The National Academies Press. Accessed online August 11, 2010 from http://www.nap.edu/openbook.php?record_id=12224&page=1.

[88] Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform*. Washington, DC: National Education Association.

[89] Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform*. Washington, DC: National Education Association, p. vii.

[90] Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform*. Washington, DC: National Education Association.

[91] Goldstein, J. (2009). Designing transparent teacher evaluation: The role of oversight panels for professional accountability. *Teachers College Record*, 111(4), 893; Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education*, 113(3): 479-508.

[92] Heneman, H. G., A. Milanowski, S. M. Kimball, and A. Odden. 2006. Standards-based Teacher Evaluation as a Foundation for Knowledge- and Skill-based Pay. Philadelphia, PA: Consortium for Policy Research in Education. Retrieved October 15, 2010 from http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf.

[93] For details, see TAP's website at http://www.tapsystem.org/.

[94] Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform*. Washington, DC: National Education Association.

[95] For details, see ProComp's website at http://denverprocomp.dpsk12.org/.

[96] Wiley, E. W., Spindler, E. R, & Subert, A.N. *Denver ProComp: An outcomes evaluation of Denver's Alternative Teacher Compensation System 2010 report*. Boulder, CO: University of Colorado, Department of Education. Retrieved October 15, 2010 from http://static.dpsk12.org/gems/newprocomp/ProCompOutcomesEvaluationApril2010final.pdf.

[97] Darling-Hammond, L. & Prince, C.D. (2007). *Strengthening teacher quality in high need schools—Policy and practice*. Washington, DC: Council of Chief State School Officers: Washington, DC: Council of Chief State School Officers.

[98] Darling-Hammond, L. & Prince, C.D. (2007). *Strengthening teacher quality in high need schools—Policy and practice*. Washington, DC: Council of Chief State School Officers: 3.

[99] Baker, E.L., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J. & Shepard, L. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers*. Briefing Paper # 278. Washington, DC: Economic Policy Institute: 1-2. Retrieved December 7, 2010 from http://www.epi.org/publications/entry/bp278.

Appendix: Brief Summaries of Teacher Evaluation Methods

Measure	Description	Research	Strengths	Cautions
Classroom Observations	Used to measure observable classroom processes, including specific teacher practices, holistic aspects of instruction, and interactions between teachers and students. Can measure broad, overarching aspects of teaching or subject-specific or context-specific aspects of practice.	Some highly researched protocols have been found to link to student achievement, though associations are sometimes modest. Research and validity findings are highly dependent on the instrument used, sampling procedures, and 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 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.	• Careful attention must be paid to choosing or creating a valid and reliable protocol and training and calibrating raters • Classroom observation is expensive due to cost of observers' time; intensive training and calibrating of observers adds to expense but is necessary for validity. • This method assesses observable classroom behaviors but is not as useful for assessing beliefs, feelings, intentions, or out-of-classroom activities.
Principal Evaluation	Is generally based on classroom observation, maybe by structured or unstructured; uses and procedures vary widely by district. Is 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 exists that training for principals, is limited and rare, which would impair validity of their evaluations.	• Can represent a useful perspective based on principals' knowledge of 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.

Measure	Description	Research	Strengths	Cautions
Instructional Artifact	Structured protocols used to analyze classroom artifacts in order to determine the quality of instruction in a classroom. May include 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 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 method may be a promising middle ground in terms of feasibility and validity between full observation and less direct measures such as self-report.
Portfolio	Used 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/stability in scoring. There is a lack of research linking portfolios to student achievement. Some studies have linked NBPTS certification (which includes a portfolio) to student achievement, but other studies have found no relationship.	• Is comprehensive and can measure aspects of teaching that are not readily observable in the classroom. • Can be used with teachers of all fields. • Provides a high level of credibility among stakeholders. • Is a good tool for teacher reflection and improvement.	• This method is time-consuming on the part of teachers and scorers; scorers should have content knowledge of the portfolios. • The stability of scores may not be high enough to use for high-stakes assessment. • Portfolios are difficult to standardize (compare across teaches or schools). • Portfolios represent teachers' exemplary work but may not reflect everyday classroom activities.

Measure	Description	Research	Strengths	Cautions
Teacher Self-Report Measure	Teacher reports of what they are doing in classrooms. May be assessed through surveys, instructional logs, and interviews. 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 harder to establish reliability or may result in very narrowly focused measures.	• Can measure unobservable factors that may affect teaching, such as knowledge, intentions, expectation, and beliefs. • Provides the unique perspective of the teacher. • Is very feasible and cost-efficient; can collect large amounts of information at once.	• Reliability and validity of self-report is not 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 method should not be used as a sole or primary measure in teacher evaluation.
Student Survey	Used to gather student opinions or judgments about teaching practice as part of teacher evaluation and to provide information about teaching as it is perceived by students.	Several studies have shown that student ratings of teachers can be useful in providing information about teaching; may be as valid as judgments made by college students and other groups; and, in some cases, may correlate with measures of student achievement. 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 n a way that will connect with students. • Makes use of students, who may be as capable as adult raters at providing accurate ratings.	• Student ratings have not been validated for use in summative assessment and should not be used as a sole or primary measure of teacher evaluation. • Students cannot provide information on aspects of teaching such as a teacher's content knowledge, curriculum fulfillment, and professional activities.

Measure	Description	Research	Strengths	Cautions
Value-Added Model	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 does suggest 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. Thus, it is obvious that teachers vary in effectiveness, but the reasons for this are not known.	• Provides a way to evaluate teachers' 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 is already collected for NCLB purposes. • May be useful for identifying upstanding 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 essentially measure 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.

Source: Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. Washington, D.C.: National Comprehensive Center for Teacher Quality, p. 16-19. Reproduced by permission of Laura Goe.