Approaches to Evaluating Teacher Effectiveness: A Research Synthesis

June 2008

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

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

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

3018_08/08

Acknowledgments

The authors wish to thank Arie Van der Ploeg, Amy Holcombe, Matthew Springer, Jeffrey Max, Jane Coggshall, Tricia Coulter, Waverely Van Winkle, Sarah Ohls, Carol Dwyer, Sabrina Laine, and Cynthia Tocci for helpful comments and support.

Contents

Abstract

This research synthesis examines how teacher effectiveness is currently measured. By evaluating the research on teacher effectiveness and the different instruments used to measure it, this research synthesis contributes to the discussion of appropriate rigor and relevance of measures for different purposes (i.e., formative vs. summative evaluation). The findings are presented along with related policy implications. In addition, the synthesis describes how various measures have been evaluated, explains why certain measures are most suitable for certain purposes (high-stakes evaluation vs. formative evaluation, for instance), and suggests how the results of the study might be used to inform the national conversation about teacher effectiveness. A comprehensive definition of the components and indicators that characterize effective teachers is provided, extending this definition beyond teachers' contribution to student achievement gains to include how teachers impact classrooms, schools, and their colleagues as well as how they contribute to other important outcomes for students. Through this synthesis, the National Comprehensive Center for Teacher Quality (TQ Center) hopes to provide some practical guidance in how best to evaluate teacher effectiveness.

Introduction

The process of evaluating the effectiveness of teachers has changed over time along with the definition of what effective teaching is, due in part to increasing state and federal attention to school-level and classroom-level accountability for student learning. Effective teaching has been defined in many ways throughout the years (Campbell, Kyriakides, Muijs, & Robinson, 2003; Cheng & Tsui, 1999; Cruickshank & Haefele, 1990; Good, 1996; Muijs, 2006), and methods for measuring teachers have changed as definitions and beliefs about what is important to measure have evolved. Although there is a general consensus that good teaching matters and that it may be the single most important school-based factor in improving student achievement (Darling-Hammond, 2000; Wright, Horn, & Sanders, 1997), measuring teacher effectiveness has remained elusive in part because of ongoing debate about what an effective teacher is and does. In a discussion of research-based indicators of effective teaching, Cruickshank and Haefele (1990) stated, "An enormous underlying problem with teacher evaluation relates to lack of agreement about what constitutes good or effective teaching" (p. 34).

Besides a lack of clear consensus on what an effective teacher is and does—or perhaps because of it—there is not a generally agreed-upon method for evaluating teacher effectiveness. Commonly used methods include classroom observations designed to measure teacher practices against some standard of effective teaching and value-added models that set out to measure the contribution of individual teachers to their students' achievement gains. This research synthesis, describes the various ways in which effective teaching can be conceptualized and measured and consists of the following sections:

- Rationale and Goals of This Study
- Important Definitions and Specifications
- Proposal of a Comprehensive Definition of Teacher Effectiveness
- Data Collection and Methods
- Validity and Considerations in Measuring Teacher Effectiveness
- Methods of Measuring Teacher Effectiveness
- Considering a Comprehensive Measure of Teacher Effectiveness
- Policy Recommendations and Implications

Rationale and Goals of This Study

The primary goal of this research synthesis is to help regional and state decision makers better understand what constitutes effective teaching and the advantages and disadvantages of the various measures commonly used to evaluate it. This study was commissioned by the National Comprehensive Center for Teacher Quality (TQ Center), which is charged with assisting regional comprehensive centers and the states they work with to understand and implement the highly qualified teacher requirements of the No Child Left Behind (NCLB) Act, with a particular emphasis on ensuring that students at risk for poor educational outcomes and students with special needs are served by highly qualified, effective teachers.

The TQ Center gathers data regularly to determine the most pressing needs of the states in regard to implementing the NCLB highly qualified teacher requirements. The TQ Center's needs-sensing data strongly suggest that states need more help identifying effective teachers in order to better respond to the NCLB equitable distribution requirement, which states that minority students and students living in poverty must have equal access to experienced, highly qualified teachers. This requirement focuses on ensuring equal access to highly qualified, experienced teachers; however, all students, particularly those who are at high risk for failure, also should have access to effective teachers. Mandating that teachers meet the minimum requirements to be considered highly qualified is a first step toward ensuring teacher effectiveness, but just meeting those requirements is no guarantee that teachers will be effective (Goe, 2007; Gordon, Kane, & Staiger, 2006).

The topic of this research synthesis is central to the mission of the TQ Center. A research-based discussion of teacher effectiveness—its definition and measurement as well as the advantages and drawbacks of different ways of measuring teacher effectiveness—can help inform states as they develop their own mechanisms for establishing teacher effectiveness more directly.

Important Definitions and Specifications

Evaluating teachers can be approached from three different but related angles: measurement of inputs, processes, and outputs. *Inputs* are what a teacher brings to his or her position, generally measured as teacher background, beliefs, expectations, experience, pedagogical and content knowledge, certification and licensure, and educational attainment. These measures are sometimes discussed in the literature as "teacher quality"; for instance, the NCLB requirement for highly qualified teachers refers specifically to teacher qualifications and credentials. *Processes,* on the other hand, refers to the interaction that occurs in a classroom between teachers and students. It also may include a teacher's professional activities within the larger school and community, but for the purposes of this research synthesis, classroom processes are the focus. *Outputs* represent the results of classroom processes, such as impact on student achievement, graduation rates, student behavior, engagement, attitudes, and social-emotional well-being. Other outcomes may involve contributions to the school or community in the form of taking on school leadership roles, educating other teachers, or strengthening relationships with parents, but again for the purposes of this research synthesis, student outcomes are the focus. Outputs can be referred to as "teacher effectiveness," although as discussed in the following section, *teacher effectiveness* as used in the research literature is often limited to mean impact on student achievement specifically.

The studies discussed in this research synthesis focus explicitly on teacher effectiveness in terms of gains in student achievement and on measures of classroom processes. The reasons for using this focus and selection criteria are described in the Data and Methods section. However, given the many terms discussed and their subtle distinctions, an argument can be made for a conceptualization of *teacher effectiveness* that is a broader and more encompassing term for the many facets that contribute to a teacher's success.

Defining Teacher Effectiveness

Clarifying the way *teacher effectiveness* is defined is important for two main reasons. First, what is measured is a reflection of what is valued, and as a corollary, what is measured is valued. Definitions nominate and shape what needs to be measured. If, for example, policy conversations revolve around scores from standardized tests, the significant outcomes can be narrowed to those that can be measured with standardized test scores. On the other hand, when policy conversations concern the interactions between teachers and students, the focus shifts to classrooms and documenting effective interactions among teachers and their students. In addition, different definitions lead to different policy solutions. When the conversation focuses on teacher quality, the discussion likely turns to improving teachers' scores on measures of knowledge or on signals of that knowledge, such as certification. When classroom processes are discussed, particular practices or approaches to teaching become the focus.

Given the importance of these distinctions, this research synthesis uses the term *teacher effectiveness* but does so with a much broader definition than is typically associated with that term in current policy conversations. In the remainder of this section, a more nuanced definition of *teacher effectiveness* is provided; this definition includes the varied roles teachers play as well as the varied student outcomes education stakeholders value.

Critiques of the Dominant Teacher Effectiveness Definition

Increasingly, policy conversations frame *teacher effectiveness* as a teacher's ability to produce higher than expected gains in students' standardized test scores. This focus on attributing gains on standardized tests to teachers and measuring the result of teaching by averaging test score gains has a number of strengths. It is parsimonious; it can be measured using data collected as part of NCLB requirements; and it has a certain amount of credibility—most would agree that an effective teacher *should* help students learn more than expected. This definition does, however, have serious limitations.

Teachers Are Not Solely Responsible for Students' Learning.

One critique concerns the problem of the assumptions of causality that underlie this approach. The approach requires the establishment of what part of an effectiveness score is attributable solely to the teacher. Making this determination is problematic not just for practical reasons but for logical reasons—assumptions are required that may be unreasonable. Fenstermacher and Richardson (2005) illustrate the problem with this scenario:

> If we presuppose a blank, receptive mind, encased within a compliant and passive learner, then we need travel only a very short logical distance to infer that teaching produces learning, and hence that what teachers do determines whether students learn. In the passive recipient view, it makes some sense to think of successful teaching arising solely from the actions of a teacher. That is, learning on the part of the student is indeed a direct result of actions by a teacher. Yet we all know that learners are not passive receptors of information directed at them. Learning does not arise solely on the basis of teacher activity. Assuming that the formulation offered above has merit, then it follows that success at learning requires a combination of circumstances well beyond the actions of a teacher. (pp. 190–191)

It can be argued that narrowing the definition of teacher effectiveness to reflect only student growth on standardized achievement measures takes this assumption too far. It is important to note that measures of teacher effectiveness can be calculated without regard to what takes place in classrooms and schools, if teacher effectiveness is narrowly defined as a given teacher's impact on the learning of his or her students as measured by standardized tests. With this narrow definition, other important ways that teachers contribute to successful students, communities, and schools are overlooked. Similarly, other influences on student outcomes, including other teachers, peers, school resources, community support, leadership, and school climate or culture, cannot be "parceled out" of the resulting score.

In the narrowest definition of teacher effectiveness, in which effectiveness is determined solely by student achievement gains, a teacher can be deemed effective compared to other teachers because his or her students performed better on the state test than the students' prior achievement would have predicted, without consideration of any other factors. In that case, it would be impossible to say whether the growth in achievement as reflected by test scores was the result of class time spent narrowly on test-taking skills and test preparation activities or whether

achievement growth was the result of inspired, competent teaching of a broad, rich curriculum that engaged students, motivated their learning, and prepared them for continued success.

Consensus Should Drive Research, Not Measurement Innovations.

Another critique of a teacher effectiveness model based on test scores concerns the degree to which innovations in measurement drive how teacher effectiveness is defined. Campbell et al. (2003) contend that trends in measurement of teacher effectiveness seem to follow the development of new instruments and technologies, focusing on the ability to measure something, rather than first defining effectiveness and *then* determining a technology for measuring it. They describe the sense of "…the horse and the cart being in the wrong places; the technology of measurement has been creating the concept of effectiveness rather than the concept requiring an appropriate technology. It follows that current concepts of teacher effectiveness may be open to question" (p. 350). These authors make an important point: just because it is possible to match teachers to their students' test scores and use this relationship as a measure of teacher effectiveness does not mean that this is the *only* way to evaluate teacher effectiveness.

The increased availability of data in which student achievement is linked to teachers along with statistical innovations in analyzing these data may be partly responsible for what appears to be a growing emphasis on measuring teachers' contributions to student achievement (Drury & Doran, 2003; Hershberg, Simon, & Lea-Kruger, 2004; The Teaching Commission, 2004) and a concomitant narrowing of the definition of teacher effectiveness. Students' knowledge is summarized in a test score, whereas teachers' effectiveness is reflected in their contribution to that test score.

Value-added models provide a classic example of a measure of teacher effectiveness driven by technological development. Using longitudinal linked teacher-student data, William Sanders was able to determine that students in some teachers' classrooms were scoring higher than their previous test scores would have predicted (Sanders & Rivers, 1996). Sanders' findings and his marketing of the technology to states for the purpose of evaluating schools and teachers have garnered considerable attention and contributed to the increased use of value-added methodologies.

In addition to the objection to innovations leading definitions, there are substantial issues with using student achievement test scores as measures of teaching effectiveness for all students. If, for example, students are dropping out of school at a higher rate because of testing-related graduation requirements, as some research suggests (Haney, 2000), then high school achievement scores are increasingly representing the scores of the "survivors" rather than *all* students. Such measurement issues raise questions about the validity of test scores as a measure of teacher effectiveness in secondary schools with high dropout rates.

Learning Is More Than Average Achievement Gains.

A final critique of this model suggests that 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. Though current policy conversations and some

research studies implicitly refer to teacher effectiveness as gains in student achievement, reviewing the literature on teacher evaluation revealed that definitions of teacher effectiveness provided by researchers have been more varied and broader in scope. For example, Campbell, Kyriakides, Muijs, and Robinson (2004) state, "Teacher effectiveness is the impact that classroom factors, such as teaching methods, teacher expectations, classroom organisation, and use of classroom resources, have on students' performance" (p. 3). This definition takes into consideration what occurs in the classroom, but the measure of effectiveness is still the students' performance. However, a number of researchers contend that there are other important outcomes besides students' performance on standardized tests that define effective teachers. More than 20 years ago, in their review of "process-outcome" research linking teacher behavior to student achievement, Brophy and Good (1986) made the following statement about their work:

> The research discussed is concerned with teachers' effects on students, but it is a misnomer to refer to it as "teacher effectiveness" research, because this equates "effectiveness" with success in producing achievement gain. What constitutes "teacher effectiveness" is a matter of definition, and most definitions include success in socializing students and promoting their affective and personal development in addition to success in fostering their mastery of formal curricula. (p. 328)

Brophy and Good's point becomes clear when the outcome measure of graduation is considered. In *A Highly Qualified Teacher in Every Classroom: The Secretary's Fourth Annual Report on Teacher Quality*, it is clear that improving graduation rates is an important goal that is tied to teaching: "While much of the work of NCLB has focused on elementary and middle schools, now, America must do more to prepare high school students for graduation, especially those most at risk of dropping out" (Office of Postsecondary Education, 2005, p. xii). Yet even though on-time promotion and high school graduation are important educational outcomes, they are ignored under an achievement-only definition of teacher effectiveness.

It could be that standards for judging effectiveness have become more focused on student achievement as the most important outcome due to increasing accountability pressures. Or it could be that the accessibility of linked student-teacher data, improvements in statistical methods, and increasingly powerful computers have made it possible to do analyses that were previously extremely difficult to perform. Most likely, it is a combination of those factors. Student achievement gains should be an important component in evaluating teacher effectiveness; however, the critiques of the achievement-focused view of teacher effectiveness are legitimate. The next section offers a broader view of teacher effectiveness and argues that other aspects of teaching must be a part of the conversation.

Considering a More Comprehensive Definition of Teacher Effectiveness

In light of these critiques, and given that teachers' roles involve much more than simply providing subject-matter instruction, it is appropriate to consider a broader and more comprehensive definition of effective teachers consisting of five points and formulated by evaluating discussions of teacher effectiveness in the research literature as well as in policy documents, standards, and reports (e.g., Berry, 2004; Brophy & Good, 1986; Campbell et al., 2003, 2004; Cheng & Tsui, 1999; Darling-Hammond & Youngs, 2002; Englert, Tarrant, & Mariage, 1992; Fenstermacher & Richardson, 2005; Gentilucci, 2004; Hamre & Pianta, 2005; Haycock, 2004; Interstate New Teacher Assessment and Support Consortium, 2001; Kyriakides, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; McColskey et al., 2005; Muijs, 2006; National Board for Professional Teaching Standards, 2002; Newmann, Bryk, & Nagaoka, 2001; Odden, Borman, & Fermanich, 2004; Office of Postsecondary Education & Office of Policy Planning and Innovation, 2003; Rivkin, Hanushek, & Kain, 2005; Schlusmans, 1978; Shavelson, Webb, & Burstein, 1986; Tucker & Stronge, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004; Watson & De Geest, 2005). In addition, after these five points were conceptualized, they were circulated among a number of experts on teacher quality and effectiveness for feedback and strengthened as a result the experts' input.

The five-point definition of effective teachers consists of the following:

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

This definition is intended to focus measurement efforts on multiple components of teacher effectiveness. It is proposed not as a criticism of other useful definitions, many of which were considered in the formation of these points, but as a means of *clarifying priorities* for measuring teaching effectiveness. The first point directly addresses student achievement gains on standardized tests, and the other points focus on teachers' contributions that may ultimately improve student learning, albeit indirectly. Clearly, student achievement gains on standardized tests are not the only—possibly not even the most important—outcome against which teacher

performance should be evaluated. A comprehensive evaluation of teacher effectiveness might be based on a composite that includes teachers' scores using a number of different measures.

Some may argue that teacher effectiveness should be limited to outcome measures, and thus process and behavior variables (e.g., having high expectations, using appropriate assessments, or collaborating with parents) should be excluded. However, because teachers impact student learning and growth through the processes and practices they employ, it is reasonable to state that an effective teacher can be observed to be doing things that research has suggested are likely to lead to improved student learning. It is necessary for these processes and practices to be measurable.

Although it is theoretically possible to identify indicators of all the components in the definition of effective teachers so that they can be measured and scored, there is a dearth of research in many of these areas. Most measures of teacher effectiveness focus on either student achievement gains attributed to the teacher *or* on classroom performance as measured with observation protocols. Actually *measuring* teachers' contribution to other outcomes—student attendance, promotion, and graduation—is less common. The fifth point in the definition is seldom measured or even considered as a component of teacher effectiveness, but it is particularly important given the increased emphasis on collaboration between general education teachers and those who focus on working with students with special needs (e.g., Abbott, Walton, Tapia, & Greenwood, 1999; Bauer, Johnson, & Sapona, 2004; Benner & Judge, 2000; Blanton, Blanton, & Cross, 1994; Blanton, Griffin, Winn, & Pugach, 1997; Fuchs & Fuchs, 1998; Gable, 1993; Hardman, McDonnell, & Welch, 1998; Interstate New Teacher Assessment and Support Consortium, 2001; Pugach, 2005). The next section describes the process through which the literature was selected and narrowed down in order to present information about various ways that teaching is measured and to make suggestions about how teacher effectiveness can be more comprehensively measured.

Data Collection and Methods

General Approach

The general approach to the identification and selection of articles for this synthesis was to start with broad categories and many search terms and then progressively narrow the group of studies down to only those that met certain criteria. While stricter criteria could have been applied, the authors of this synthesis are in agreement with Dynarksi (2008) who states, "Selective exclusion of research requires great caution, as selectivity can be interpreted as compromising scientific objectivity for purposes that educators cannot discern and may misinterpret" (p. 27). Rather than eliminate studies that might be informative for some purposes or audiences, the authors of this synthesis elected not to use narrow criteria. Dynarski also stated:

> Certainly it is possible that the findings from some studies are due to publication bias or arise from local conditions that are unusual or hard to replicate. But if syntheses review all the evidence and apply sound standards, educators can make up their own minds about whether the findings are credible or whether the implementation conditions are unrealistic and not useful to them. (p. 28)

Given that the purpose of this synthesis is to help policymakers, state leaders, and educational professionals sort out what the evidence says about teacher effectiveness, it seemed reasonable to let them weigh the evidence for themselves.

Stages of Development

Several stages were required to develop an appropriate set of articles to analyze for this synthesis. The authors served as reviewers of all articles and made decisions at each stage of the process based on their shared understanding of the identified criteria. In the case that one author was uncertain about whether an article met the criteria, she consulted with one of the other authors and discussed the uncertainty until a consensus was reached.

It is worth noting, however, that the literature on teacher effectiveness is large and disconnected. Scholars working in different fields theorize, conduct studies, and publish articles in very different journals. Sometimes these findings do not build on or connect with findings in other areas. This can mean that knowledge is less cumulative than one might like. As Kennedy (2007) notes, this means that reviews of research in such areas rely on the conceptual frameworks of the researchers. The authors of this research synthesis selected categories that they deemed to be reasonable; however, scholars in other disciplines might have used different categories.

Stage 1

The authors met on a number of occasions to discuss the purpose of the synthesis and develop a list of search terms that appeared to fit with that purpose.

Stage 2

Articles were identified through Internet and library searches of keywords and phrases related to the topics of teacher effectiveness and measuring teacher performance. ERIC and PsycInfo were the main databases used to identify relevant peer-reviewed articles within the last six to eight years, using the following search terms: *teacher effectiveness*, *teacher evaluation*, *value-added modeling*, *teaching methods*, *teacher improvement*, *teacher competencies*, *pedagogical content knowledge*, *instructional effectiveness*, *instructional improvement*, *research tools*, *videotape recordings*, *questionnaires*, *instructional material evaluation*, *teacher behavior*, *assignments*, *instructional development*, *beginning teacher induction*, *professional development*, *academic achievement prediction*, *educational measurement*, and *educational quality*. Additional articles, including older, seminal, nonempirical, and/or theoretical pieces, were identified from broader Internet searches, reference lists of related articles, and recommendations of experts in the field.

Stage 3

This search process yielded more than 1,600 studies. In order to narrow the results further, abstracts were reviewed to determine whether the studies met the following criteria:

- **Language and Location.** Studies were published in English, and research was conducted in the United States, Canada, Great Britain, Ireland, Australia, and New Zealand.
- **Population.** Research addressed the K–12 student population and measured inservice teachers.
- **Relevance.** Research addressed the topic of measuring effective teaching.

Approximately 300 articles meeting these criteria were then sent to the next stage.

Stage 4

The remaining 300 articles were reviewed more closely for relevance and methodological rigor. Studies chosen for this research synthesis met the following additional criteria:

- They were empirical.
- They included a measure of teacher effectiveness or classroom practice.
- They included a student outcome measure *or* had implications for teacher effectiveness.
- They reported methods meeting accepted standards for quality research (e.g., reliable and validated instruments, appropriate study design, and necessary controls).

Stage 5

The resulting collection of studies was then evaluated, and additional exclusions were made when deeper reading of studies revealed they did not meet the purposes or the quality standards of this synthesis. Studies that were of poor quality, off topic, out-of-scope, focused on higher education or prekindergarten education, or lacked descriptions of data and methods were

excluded. The resulting synthesis includes approximately 120 studies that were thoroughly reviewed.

As discussed, the search was narrowed by focusing on studies measuring classroom processes and outputs in the form of student outcomes, paying particular attention to studies measuring teacher effectiveness in terms of value-added student achievement measures. The search was limited in this way for two main reasons:

- A previous research synthesis commissioned by the TQ Center (see Goe, 2007) specifically addresses the links between measures of teacher quality and student outcomes, and this topic also has been addressed in a number of other research syntheses and reviews (e.g., Darling-Hammond & Youngs, 2002; Goe, 2007; Rice, 2003; Wayne & Youngs, 2003; Wilson & Floden, 2003). Though there is some overlap, this research synthesis is meant to be an extension of previous work, thus it focuses on processes and outputs rather than on inputs.
- The criteria was narrowed by only including processes occurring inside the classroom and outputs concerning student outcomes. This narrowing of scope was necessary to ensure that the amount of literature to be reviewed and synthesized was manageable enough to be transformed into a useable and informative document. The research synthesis mainly focuses on processes inside the classroom and student outcomes related to gains in student achievement because these are topics that are prevalent in the current education policy landscape and are areas in which states have indicated a need for more information and assistance.

Furthermore, this synthesis is limited to measuring teachers and does not address methods of measuring school effects, the effectiveness of curriculum or professional development implementations (unless they include measures specific to teachers), or other evaluations of educational interventions or programming. Though these are important and related topics, they are beyond the scope of this synthesis.

Validity and Considerations in Measuring Teacher Effectiveness

Determining what type of teacher evaluation method is best for a given purpose includes taking account of the validity and reliability of the instrument or process being used. Validity is the "most fundamental consideration in assuring the quality of any assessment" (Millett, Stickler, Payne, & Dwyer, 2007, p. 4). Validity refers to the degree to which an interpretation of a test score, or in this case, a score from a measure of teacher effectiveness, is supported by evidence. For a measure of teacher effectiveness to be valid, evidence must support the argument that the measure actually assesses the dimension of teacher effectiveness it claims to measure and not something else. In addition, evidence that the measure is valid for the purpose for which it will be used is essential. Instruments cannot be valid in and of themselves; an instrument or assessment must be validated for particular purposes (Kane, 2006; Messick, 1989). For example, an observation-based score might be validated for professional development purposes but might not be validated for compensation purposes. Determining the validity of an instrument requires taking account of the evidence regarding what the instrument measures, what it does not measure, and how the scores are being used. This requires the user of the instrument to be well-informed about these issues and willing to make judgments about the degree to which there is sufficient evidence to use a particular instrument for the purpose under consideration.

In addition to concerns about validity, there are other measurement concerns. Blanton et al. (2003) identified six criteria that are particularly useful in informing this conversation [which are elaborated in Coggshall (2007)], and these criteria have been adapted and applied to the discussion of teacher effectiveness in the following pages.

Comprehensiveness refers to the degree to which a measure captures *all* of the various aspects of teacher effectiveness. For example, less comprehensive measures might only capture how well a teacher is able to represent mathematics in the classroom. More comprehensive measures would capture how teachers represent mathematics, how they scaffold student learning, and how well they work with colleagues.

Generality refers to how well an instrument captures the full range of contexts in which teachers work. If an instrument can be used to assess elementary and secondary teacher effectiveness in music and special education, the instrument can be said to have a high level of generality. Generality is particularly important if one intends to compare teachers across contexts.

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

Practicality refers to the logistical issues associated with a measure. These include the "costs, training requirements, and the developmental work required to adapt an existing model or measure" for one's own purpose (Blanton et al., 2003, p. 14). For example, creating valid and reliable instruments and processes for measuring teacher effectiveness is costly and time-

consuming. Adapting an existing instrument and process might be less of a drain on district or state resources.

Reliability refers to the degree to which an instrument measures something consistently. For example, it might be important to know whether scores on an instrument measuring teacher effectiveness vary by time of year, time of day, grade level, or subject matter. It is also important to note that instruments can be reliable without actually measuring what they were intended to measure. For example, an instrument might consistently measure teachers' use of flash cards. But if flash card use is not an important determinant of teacher effectiveness, then the instrument is reliable but not valid for the purpose of measuring teacher effectiveness.

Credibility is a specific type of validity—face validity—that is particularly important in measures of teacher effectiveness. If an instrument has strong credibility, many stakeholders from different groups (e.g., parents, teachers, administrators, and policymakers) view the measure as reasonable and appropriate.

In this research synthesis, these aspects of measurement—validity, comprehensiveness, generality, utility, practicality, reliability, and credibility—are used to describe and assess a range of approaches to measuring teacher effectiveness. Particular attention is given to issues of validity and reliability because the authors draw heavily from the research literature, which is very concerned with such issues.

In addition, careful attention is given to the purposes of instruments. The authors distinguish between high-stakes, low-stakes, formative, and summative assessments of teacher effectiveness. A formative evaluation is one that is intended to gather information that will be useful to improve a program, activity, or behavior. A summative evaluation is meant to make a final determination about a program, activity, or behavior at a specific point in time. For instance, a classroom observation may be an informal drop-in visit by a principal, or it may be a planned, formal observation conducted by highly trained professional evaluators with employment or tenure consequences. An informal evaluation that does not carry serious consequences and is meant to collect information for providing feedback to improve teaching is considered low-stakes and formative. In contrast, formal evaluations that carry substantial consequences and are conducted to gather information for a specific decision-making process are considered high-stakes and summative. Considering whether the intent of the evaluation is high-stakes or low-stakes and whether it is summative or formative in nature will have strong implications for choosing a measure that will provide valid results.

Methods of Measuring Teacher Effectiveness

The following sections present methods in teacher evaluation that are useful for measuring teacher effectiveness more broadly and providing information about what makes teachers effective. The discussion begins with the most widely used measure of teacher effectiveness, classroom observations. A review of other instruments that directly assess what teachers do in classrooms also is provided. These include principal evaluations; analysis of classroom artifacts (i.e., ratings of teacher assignments and student work); teaching portfolios; teacher self-reports of practice, including surveys, teaching logs, and interviews; and student ratings of teacher performance. Finally, teacher effectiveness as measured by value-added strategies is considered. For the scope of this discussion, more indirect measures of teaching, such as teacher demonstrations of knowledge, teacher responses to theoretical teaching situations (i.e., structured vignettes), or parent satisfaction surveys are not included. These measures can be extremely useful in assessing teaching competency; however, the authors chose to focus on measures that more directly assess the processes and activities occurring during instruction and products that are created inside the classroom. In addition, the research linking credentials, experience, or knowledge to teacher effectiveness is not considered. Though such work is terrifically important in discussions of initial teacher licensure, extensive reviews have already been conducted and widely publicized (e.g., Darling-Hammond & Youngs, 2002; Goe, 2007; Rice, 2003; Wayne & Youngs, 2003; Wilson & Floden, 2003).

Each of the sections that follow defines and describes the measure, provides examples and research findings on its use, and discusses its strengths and cautions, keeping in mind the previously described validity considerations and providing recommendations as appropriate. Coverage of instruments is not meant to be exhaustive but rather to accomplish the following: (1) to provide some researched examples of methods that are being employed by states or that are promising measures of teaching, and (2) to present knowledge of their uses and barriers. In addition, many commercially available products are not reviewed here but are examples of the broader class of instruments considered in this synthesis. Thus, in the interest of time, the synthesis considers the broader class of instruments and leaves it to the reader to consider the particular products. Table 1 presents a brief summary of the discussion on each method.

Table 1. Brief Summaries of Teacher Evaluation Methods

Measure	Description	Research	Strengths	Cautions
Classroom Observation	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, may be 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 teachers 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, expectations, 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 in 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 outstanding teachers whose classrooms can serve as "learning labs" as well as struggling teachers in need of support.	• Models are not able to sort out teacher effects from classroom effects. • Vertical test alignment is assumed (i.e., tests 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.

Classroom Observations

Description

Teacher observations take many forms, measure different aspects of teaching, and vary greatly in their implementation. They may be a district-developed set of categories that are used to give teachers' formative feedback. They may be a product purchased from an outside vendor that comes with rater training and scoring. Most often, observations occur somewhere between once and a few times during the school year, encompass roughly one lesson, and happen on a day agreed upon by the teacher and the rater. There is often a preobservation or postobservation conference between the rater and the teacher. The degree to which observations can or should be used for specific purposes depends on the instrument, how that instrument was developed, the level of training and monitoring raters receive, and the psychometric properties of the instrument. Review of the research suggests that observation scores have been related to important outcome measures such as student achievement (Gallagher, 2004; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004).

When measuring teacher effectiveness through classroom observations, valid and appropriate instruments are crucial as well as trained raters to utilize those instruments in standard ways so that results will be comparable across classrooms. The following example may help explain what is meant by a "trained rater":

Presume that there are four aspects of teaching effectiveness one wants to measure: teacher student interactions, classroom management, school community contributions, and subject matter knowledge. Each is measured on a three-point scale: *needs improvement*, *satisfactory*, and *excellent*. In rater training, raters would be taught the differences among *needs improvement*, *satisfactory*, and *excellent* classroom management. What, for example, causes a specific classroom management technique to go from satisfactory to excellent? Raters would need to practice applying those criteria to a number of lessons to make sure they understand when they actually are faced with diverse actions. Raters also would be taught where particular actions—say, scaffolding students' understanding of fractions—are to be scored. In that example, raters might want to have such scaffolding fall into the interactions domain, whereas, others might tend to score scaffolding as a part of the teacher's subject matter knowledge. As a part of the training, these issues would be discussed and practiced, and hopefully raters would learn to score observations accurately against the standards. If this were to happen, the raters would be calibrated to the standards. If raters could do this consistently for numerous lessons, they would be reliably trained.

Whoever is using the raters' scores also would want to be sure that throughout the school year, raters are consistently applying those criteria. It would be problematic if scores were more lenient in the beginning of the year (because, for example, the teachers are just getting started) and more stringent in the middle or end of the year (because raters had seen a lot of teaching). This would mean one's scores would partially depend on when they were observed. In addition to issues of what day during the year observations take place, users of observation protocols also should pay attention to whether or not there is information (and training) to help raters consistently apply the rating criteria across different times of the day and subject matter.

Depending on the protocol, trainers may or may not have investigated, thought about, or developed training materials to deal with these issues. These issues are critical for any protocol, but they are especially important if scores are going to be used for high-stakes purposes such as tenure and compensation.

Given those technical considerations, observations can provide important, useful information about a teacher's practice if used thoughtfully. Districts must be careful, however, because observations are susceptible to rater biases in ways that some of the other measures of teacher effectiveness are not.

Examples

Examples of observation protocols that are widely used and have been studied on a relatively large scale include Charlotte Danielson's (1996) *Enhancing Professional Practice: Framework for Teaching* and the University of Virginia's Classroom Assessment Scoring System (CLASS) for prekindergarten and K–5 (Pianta, La Paro, & Hamre, 2006). The *Framework for Teaching* is meant to be used across subject matter and grade levels. CLASS also can be used across subject matter but has particular grade spans (early childhood, K–5, and 6–12).

In addition to these instruments, there are countless numbers of additional observation protocols that are less widely used, some of which have no published validity information and others of which have been used in very limited contexts—most often in research projects in which scores are not reported to teachers or used for any purpose outside the research project. A subset of these more narrowly used instruments is comprised of several promising subject-specific protocols. These protocols are particularly noteworthy, given the increasing focus on the role of subject-specific knowledge for teaching and the increasing call for teachers to have more and more relevant subject matter knowledge. Examples of these include the Reformed Teaching Observation Protocol (RTOP) for mathematics and science (Piburn & Sawada, 2000), the Quality of Mathematics in Instruction (QMI) in mathematics (Blunk, 2007), and the TEX-IN3 for literacy (Hoffman, Sailors, Duffy, & Beretvas, 2004). Though these three specific instruments are regarded as promising, they have not been widely used by anyone beyond the developers, and there is little published data on how these instruments function. RTOP has the most information (e.g., MacIsaac, Sawada, & Falconer, 2001; Piburn & Sawada, 2000; Sawada et al., 2002), whereas QMI is the newest and is still in the beginning stages of documentation (e.g., Blunk, 2007). For practitioners interested in modifying generic protocols to include more subject matter, these would be excellent resources. They also might be useful for districts interested in using subject-specific protocols for formative feedback.

Danielson's Framework. Danielson's (1996) *Framework for Teaching* is one of the most commonly used observation protocols in districts (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007). Danielson based the framework on research she and colleagues conducted in developing Praxis III, an observational protocol designed by ETS for assessing the classroom performance of beginning teachers. ETS researchers worked with many teachers and other educators to do the following:

- Define a holistic view of teaching.
- Describe the complex relationships of teachers and students.

- Examine the importance of tailoring teaching to the individual, developmental, and cultural differences of students.
- Consider the influence of the subject being taught on teaching.
- Spell out the implications of all this for teacher assessment.

The *Framework for Teaching* is described on the Danielson Group website as "a research-based set of components of instruction, aligned to the INTASC standards, and grounded in a constructivist view of learning and teaching." It consists of four domains, broken down into 22 components and 76 smaller elements. Teachers are evaluated against a detailed rubric, which can be used to rate each of the 76 elements as *unsatisfactory*, *basic*, *proficient*, or *distinguished*. The framework can be used for several purposes, such as reflection and self-assessment, mentoring and induction, peer coaching, and supervision. Although it can be used for summative evaluation, providing feedback for formative use is key. According to the Danielson Group website:

> The *Framework* may be used for many purposes, but its full value is realized as the foundation for professional conversations among practitioners as they seek to enhance their skill in the complex task of teaching. The *Framework* may be used as the foundation of a school or district's mentoring, coaching, professional development, and teacher evaluation processes, thus linking all those activities together and helping teachers become more thoughtful practitioners.

The *Framework for Teaching* has been implemented and studied in districts including Cincinnati, Ohio; Reno/Sparks, Nevada; Coventry, Rhode Island; and Los Angeles, California; with several studies finding that teachers who scored higher on the *Framework for Teaching* were associated with greater gains in student achievement (Gallagher, 2004; Kimball et al., 2004; Milanowski, 2004; Milanowski, Kimball, & Odden, 2005). These findings vary by subject matter (reading and mathematics) and grade level and are small to modest sized correlations. It is important to note that there was wide variation in rater training, rater's relationship with the teacher (peer, supervisor, etc.), the degree of adherence to Danielson's recommendations for use, the use of the scores, and the number of observations conducted for each teacher. This variation may be partially responsible for the range of findings.

For example, the school in Los Angeles that used the *Framework for Teaching* adopted a subject specific version, used it for simultaneous formative and summative feedback, and linked scores with skills-based merit pay. In Cincinnati, a nonsubject-specific version was used for both formative and summative purposes and was linked to skills-based merit pay. Research reports on these sites reported observations taking place between three and six times per year by a hired teacher evaluator (who was released from teaching duties for three years) and/or an administrator. In Nevada, principal and assistant principals used multiple sources of evidence to assign scores on a nonsubject-specific version of the modified framework. This information was used formatively and summatively, but it was not used for compensation. In Coventry, Rhode Island, principals and department heads conducted observations based on a modified version of the *Framework for Teaching*. Frequency of observations was dependent on tenure status, and scores were not intended to be linked to pay. In the two cases in which *Framework for Teaching*

scores were used for compensation decisions, they were used with other information (e.g., credentials, experience).

This variation and the research documentation of the instrument suggest a number of important points. First, a good proportion of teachers in each site find the framework credible and helpful for their teaching (Heneman, Milanowski, Kimball, & Odden, 2006). Scores have been used in four districts and 179 schools across the country for both formative and summative purposes, which suggests it is possible to use the *Framework for Teaching* in various contexts and purposes. The framework is general with respect to grade level and subject matter area. It does not capture subject specific aspects of teaching, though at least one district was able to develop subject-specific versions. The research does not indicate whether modified versions of the instrument perform as well as versions that adhere to Danielson's recommendations. In addition, it is not evident whether the instrument functions differently (or is implemented differently) at different grade levels. Finally, the *Framework for Teaching* values a constructivist approach to teaching.

Classroom Assessment Scoring System (CLASS). This observation instrument was developed at the University of Virginia as a measure of classroom quality in preschool and in the early elementary grades. A number of studies have been conducted to examine the relationship between scores on CLASS and students' academic and social growth, as described in this section. CLASS was conceptually based on theories of child development, and the dimensions characterize interactions between students and teachers (Pianta, La Paro, & Hamre, 2007). In CLASS, "the focus is on what teachers *do* with the materials they have and in the interactions they have with students" (Pianta, La Paro, et al., 2007, p. 1). Although the instrument started out as a measure of classrooms in early elementary settings, protocols have now been developed for prekindergarten, Grades K–5, and Grades 6–12.

The CLASS framework is a theoretically driven and empirically supported conceptualization of classroom interactions organized into three major domains: emotional support, classroom organization, and instructional support. Each domain has a set of more specific dimensions of classroom interactions that are deemed to be important to students' academic and social development. The *emotional support* construct refers to the teacher's ability to establish a classroom climate and set of relationships that enhance students' social and emotional functioning. The *classroom organization* construct refers to classroom processes related to the organization and management of students' behavior, time, and attention in the classroom. The *instructional support* construct refers to teaching that is consistent with both theories of how students learn best and domain-specific models of content.

CLASS uses time-sampling in the form of observation cycles. A cycle is defined as a 30-minute period in which the first 20 minutes is used for observations and note-taking and the next 10 minutes are used for scoring. CLASS has been used both in real-time observations and videotaped lessons. The authors of CLASS found that "four cycles provides a representative sampling of classrooms" (Pianta, La Paro, et al., 2007, p. 10). Based on two large studies using CLASS, researchers also found that scores are relatively stable across the school year. There are, however, small differences in mean scores around the holidays and toward the end of the year.

The developers of CLASS offer training to groups interested in using the protocol. Training consists of a two-day training and scoring session in which potential raters watch numerous 20-minute training videos that have been consensus-scored by at least three master raters. At the end of the training, potential raters take a reliability test on five 20-minute segments of videotaped teaching. A rater is considered to have achieved sufficient reliability if he or she produces a score within one point of the master raters' consensus score for that video clip. The training materials thus far have been successful, achieving an average inter-rater reliability of 87 percent (Pianta, La Paro, et al., 2007).

Currently, there is little information on the Grades 6–12 version of CLASS works, but there is extensive validity and reliability data on the elementary and prekindergarten versions, and those data are promising (Pianta, La Paro, et al., 2007). The data on the prekindergarten and K–5 versions come from six studies in more than 1,700 PK–5 classrooms in urban, rural, and suburban settings across the country. Scores on CLASS or its precursor have been related to academic gains, other developmental markers, and student behavior (Hamre & Pianta, 2005; Howes et al., 2008; Rimm-Kaufman, La Paro, Downer, & Pianta, 2005).

Although the information provided suggests that the prekindergarten and K–5 instruments are of high-quality, there are a number of considerations to keep in mind:

- There is little information about the secondary instrument, and thus it should be used with caution.
- The protocol can be used across subject matters, but it is targeted at grade levels. The protocol does have an instructional support domain but is limited in terms of the kind of subject-specific information it can generate for formative purposes.
- There are increasing numbers of districts and schools using the protocol; however, the research does not reveal whether or how districts adapt or use the instrument.

In addition, it is not known whether districts find it affordable or doable to keep raters trained at reliable and calibrated levels. Many researchers find the scores from CLASS to be meaningful, but again, there is not much information about how teachers view CLASS scores.

Strengths and Cautions

As a class of instruments, observation protocols have a number of strengths. Teacher observations often *seem* valid. To the degree that observational ratings reflect who teachers and administrators believe is a good teacher, stakeholders can support their use. This makes it particularly important for a given protocol to be developed to reflect stakeholders' ideas about best practice and to be implemented in robust, defensible ways. When observation protocols clash with stakeholders' beliefs and/or are implemented in biased ways, the validity of results is weakened. Thus, including stakeholders' views about the content and implementation of observation protocols may be beneficial.

Another strength is that observation protocols have been and could be used as a part of teacher compensation. They have been modestly to moderately linked to student achievement, depending on the instrument. They also have been used both formatively and summatively, suggesting that the same instrument can serve multiple purposes for districts. For formative use, observations can provide rich feedback about teachers' areas of strengths and weaknesses. This type of rich feedback could be used productively for formative evaluations of teachers. The rater/evaluator can share with the teacher the results of the evaluation and then use those results to help develop (cooperatively) a plan of professional development and personal growth that will lead to a closer alignment to the effective teaching practices that are valued.

There are a number of cautions that are worth bearing in mind, however, considering the use of observations for evaluation of any form. The most popular and well-researched instruments are generic and may not take account of subject-specificity in ways that could support teachers as they endeavor to teach more students increasingly ambitious content. Many protocols have been used in research projects only by the researchers themselves (or by one other researcher who was not involved in the protocol's development). This lack of field testing introduces two significant concerns. First, for many instruments, it is not evident whether it is possible for districts to use the protocols effectively for nonresearch purposes. This issue might be resolved by a review of the instruments themselves and a conversation with the developers, but nonetheless, it is important to note there is little research to guide practitioners on this issue. In addition, because many protocols have not been used to improve practice, it is not known whether the district can expect to see a change in teachers' practice when a particular protocol is used. This is a serious gap in the understanding of how these protocols might improve practice.

In addition, the link between observations and student achievement and other outcome measures (e.g., graduation and citizenship) is another concern. Though there have been some studies that link teachers' scores on observation protocols to gains in student achievement (Gallagher, 2004; Kimball et al., 2004; A. Milanowski, 2004), there is much work to be done. For example, there is little research that links scores on well-validated observation protocols with other student outcomes of interest. Observations teachers may tell a great deal about how well a given teacher's practice aligns with what is believed to be good practice, but without linking this information to student outcomes, determining effectiveness is difficult.

A final set of concerns about observation protocols involves the issue of raters. Proper training is essential because raters are making moment-by-moment judgments about what they see. McGreal (1990) contends, "The high inference nature of rating scales places the burden of selecting a rating directly upon the evaluator" (p. 50). Considerable progress has been made in developing methods for ensuring more consistent ratings through evaluator training and calibration sessions. However, there is no assurance that a given state or district actually employs these methods, meaning that different evaluators might give very different scores to the same teacher, depending on their views of good teaching. Measuring teacher effectiveness through observations can be very uneven, which threatens the utility and credibility of the protocols themselves.

Principal Evaluations

Description

Classroom observation conducted by principals or vice-principals is one of the most common forms of teacher evaluation (Brandt et al., 2007). The format varies by district; for instance, a principal evaluation can consist of a formal observation using a validated instrument, conducted at a predetermined time, coupled with pre-interviews and post-interviews with teachers, and used for both formative and summative purposes (Heneman, Milanowski, et al., 2006). It also can be an informal drop-in visit by the principal, used to develop a quick impression of how and what a teacher is doing in the classroom.

Principal evaluations differ from evaluations performed by district personnel, researchers, or other outside evaluators who are hired and trained to conduct evaluations. Principals are most knowledgeable about the context of their schools and their student and teacher populations, and thus may be likely to compare the school's teachers to each other rather than to the larger population of teachers in the district or state. They may employ evaluation techniques that serve multiple purposes:

- To provide summative evaluation scores for school, district, state, or federal accountability purposes.
- To inform decisions about tenure or dismissal.
- To identify teachers in need of remediation.
- To provide formative feedback to improve teachers' practice.

Although these factors can make principals valuable sources of information about their schools and teachers, they also have the potential to introduce bias in either direction to principals' interpretation of teaching behaviors.

Examples

Although principal evaluation is the most common component of teacher evaluation systems, there is not a lot of solid evidence on the validity of these evaluations. One recent study by Brandt and colleagues (2007) examined district policies on teacher evaluation in several Midwestern districts. They found that principals and administrators typically conducted the evaluations, which were primarily focused on making decisions about which beginning teachers should be retained and released. District policies were more likely to offer guidance on the process of conducting evaluations than to instruct administrators on the potential uses of the evaluation results. Two particularly relevant findings from the study are that most evaluations were summative—for high-stakes employment decisions, rather than formative—for helping teachers grow in the profession. Furthermore, only 8 percent of districts mentioned evaluator training as a component of their teacher evaluation systems. Thus, although the use of high-stakes, summative assessment was prevalent, the evidence that assessments were used in a reliable and valid manner was not. These findings may be regional rather than national; however, they raise the concern that career consequences are being based on the assessments of evaluators who may have little understanding of how to use the instrument in ways that ensure valid scores.

Other studies have examined the accuracy and predictive value of principal evaluations by comparing subjective principal ratings of teachers to value-added scores of student achievement (Harris & Sass, 2007b; Jacob & Lefgren, 2005, 2008; Medley & Coker, 1987; Wilkerson, Manatt, Rogers, & Maughan, 2000). These studies required principals to rate teachers in their school using a scale created by the researcher. Because these ratings were not based on a specific observation and were not tied to any official decision making, these studies are distinct from the context of principal evaluation as it generally occurs in schools, but they do raise noteworthy issues about the accuracy of principals' judgments. The main finding from these studies is that principal ratings are significantly correlated with teacher value-added scores, but the correlation is usually low. Principals were found to be fairly accurate at identifying teachers in the top or bottom group of effectiveness but were less successful at distinguishing between teachers in the middle (Jacob & Lefgren, 2008). Note, however, that the same result has been found for value-added measures (e.g., Archibald, 2007; McCaffrey et al., 2003). Principals were better able to predict value-added scores at the elementary level than they were at the secondary level (Jacob & Lefgren, 2008) and were better at making reasonable judgments about which teachers would improve achievement in mathematics than they were in making judgments about which teachers would improve achievement in reading (Harris & Sass, 2007b; Wilkerson et al., 2000).

Findings do indicate that principal ratings are better predictors of teacher value-added scores than several standard measures of teacher quality (e.g., experience, certification, and education) (Harris & Sass, 2007b); however, some of the specific findings present a mixed picture. Harris and Sass (2007b) found that principal ratings were as accurate at predicting future student achievement gains as value-added measures of teacher effectiveness, whereas Jacob and Lefgren (2008) found principal ratings to be less accurate predictors than value-added measures. Wilkerson and colleagues (2000) found that student ratings of teachers were better predictors of achievement than principal ratings. Jacob and Lefgren (2008) also explored some of the speculations behind why the correlation between principal ratings and value-added scores was lower than expected and found that principals may tend to pay more attention to the mean level of achievement in a teacher's class and not the relative improvement that students made (i.e., they do not account for differences in classroom composition). In addition, they found that principals may tend to focus on their most recent observations of a teacher rather than considering the teacher's long-term performance. Their data support the notion that a combination of principal ratings and value-added measures is a stronger predictor of student achievement than either alone.

Strengths and Cautions

Given the many areas a principal must attend to simultaneously and in the interest of reducing the subjectivity and potential bias inherent in observation, it is advisable for administrators to employ a specific and validated observation protocol when conducting teacher evaluations (see the Classroom Observations section on page 20 for examples), especially if the information is to be used in any high-stakes decision making. When choosing an instrument, careful attention should be paid to its intended and validated use. As discussed in the observation section, administrators should be fully trained on the instrument, rater reliability should be established, and periodic recalibration should occur.

Observations should be conducted several times per year to ensure reliability, and a combination of announced and unannounced visits may be preferable to ensure that observations capture a more complete picture of the teacher's practices. Another consideration is the focus of the evaluation. For instance, an observation assessing deep or specific content knowledge may be better conducted by a peer teacher or content expert, as a principal or administrator may not be equipped with the specialized knowledge to make the best judgments necessary for this type of evaluation (Stodolsky, 1990; Weber, 1987; Yon, Burnap, & Kohut, 2002). Using a combination of principal and peer raters is another consideration that may increase the credibility of the evaluation.

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

Analysis of Classroom Artifacts

Description

Another method that has been introduced to the area of teacher evaluation is the analysis of classroom artifacts, such as lesson plans, teacher assignments, assessments, scoring rubrics, and student work. The classroom artifacts that a teacher selects and creates and the student work that is generated can provide insight into the types of opportunities to learn that students are presented with on a day-to-day basis. Depending on the goals and priorities of the evaluation, artifacts may be judged on a wide variety of criteria including rigor, authenticity, intellectual demand, alignment to standards, clarity, and comprehensiveness. Though the examination of teacher lesson plans or student work is often mentioned as a part of teacher evaluation procedures, few systems employ a structured and validated protocol for analyzing artifacts to evaluate the quality of instruction. Use of a valid protocol for analyzing teacher assignments and student work introduces a potentially useful compromise in terms of providing a window into actual classroom practice, as evidenced by classroom artifacts, while employing a method that is less labor-intensive and costly than full classroom observation.

Examples

Instructional Quality Assessment (IQA). The most work on this has been done by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) located at the University of California–Los Angeles. CRESST researchers have worked extensively to develop the Instructional Quality Assessment (IQA), a protocol that can be used both for evaluating the instructional quality of a classroom and for providing feedback to teachers for purposes of professional development. IQA consists of protocols for rating the quality of teachers' assignments and student work in reading comprehension and mathematics. Rubrics focus on quality of discussion, rigor of lesson activities and assignments, and quality of expectations communicated to students (Matsumura, Slater, Junker et al., 2006). CRESST has conducted several pilot studies on IQA, finding that the rubrics are generally correlated with quality of observed instruction, quality of student work, and standardized student test scores (Clare & Aschbacher, 2001; Junker et al., 2006; Matsumura, Garnier, Pascal, & Valdés, 2002; Matsumura & Pascal, 2003; Matsumura, Slater, Junker et al., 2006). These studies also indicate reasonable reliability for the instrument, though more work may be needed to confirm its dependability and stability. For instance, work has been conducted to determine the ideal number of assignments that should be collected to maximize accuracy of scores while minimizing teacher time and effort.

Intellectual Demand Assignment Protocol (IDAP). Newmann and colleagues of the Consortium on Chicago School Research have conducted another branch of work on analyzing instructional artifacts (Newmann et al., 2001; Newmann, Lopez, & Bryk, 1998). These researchers were interested in determining the authenticity and intellectual demand of classroom assignments and created rubrics for scoring teacher assignments and student work in mathematics and reading. The rubric assesses the degree to which the assignment involves construction of knowledge, promotes disciplined inquiry, and exhibits value beyond school. The authors collected "typical" and "challenging" assignments from Chicago elementary school teachers, which were rated by trained scorers according to the rubric (see Newmann et al., 1998). Scorers were able to achieve high levels of interrater reliability using the rubrics, with greater than 90 percent agreement within one point for the different subjects and grades scored. IDAP scores were matched to student achievement gains in each teacher's classroom. Findings showed that in classrooms with higher-scoring assignments, student learning gains on the Iowa Test of Basic Skills were 20 percent higher than the national average; in classrooms with lower-scoring assignments, learning gains were 22 percent to 25 percent lower than the national average. Use of high-demand assignments appeared unrelated to student demographics and prior achievement and benefited students with high and low prior achievement alike.

Scoop Notebook. Another example is the Scoop Notebook—developed and piloted by Borko, Stecher, Alonzo, Moncure, and McClam (2005) and further analyzed by Borko, Stecher, and Kufner (2007)—used to evaluate classroom practices through the examination of artifacts reflecting the teaching and learning process. Materials in the notebook included handouts, scoring rubrics, writing on the board, student class work, student homework, and projects. In a pilot study of 13 middle-school mathematics and science teachers, teachers provided two examples of "high" and "average" quality work for each set of class work or homework collected over a five- to seven-day period. Teachers also took pictures of artifacts in the classroom (e.g.,

writing on the board) and answered reflective questions about lessons. Multidimensional scoring guides were developed by the researchers using mathematics and science education standards and were rated by two or more trained raters. Although rater agreement was higher than would be predicted by chance, there were clear areas in which raters were inconsistent, and they appeared to be better at judging a lack of evidence rather than the presence of evidence. Some teachers found the process to be beneficial to their instruction, particularly reflecting on the lessons. Ratings also were found to be reasonably consistent with observational measures, but no links were made to student achievement in this small pilot study.

Strengths and Cautions

Analysis of classroom artifacts is a promising method to provide a comprehensive view of a teacher's quality of instruction and gain a deeper understanding of his or her intentions and expectations. It may prove to be a practical and feasible method, as the artifacts have already been created by the teacher and the procedures do not appear to place unreasonable burdens on teachers (Borko et al., 2005). This method has the potential to provide summative information about instruction as well as rich formative information and opportunity for reflection to teachers.

However, several cautions should be taken into consideration. As with the other methods discussed so far, accurate scoring is essential to preserving the validity of the instruments. This requires adequate training and calibration of scorers and also may require scorers to possess some knowledge of the subject matter being evaluated. Some studies also have noted that a lack of variation in quality of assignments (i.e., teachers at a school consistently assign very low-quality assignments) can make it difficult to validate the scoring rubrics (e.g., Matsumura, Patthey-Chavez, Valdés, & Garnier, 2002). More research needs to be done to investigate the reliability and stability of ratings and explore links to student achievement. There remains a lack of research documenting the use of these instruments in practice, and they have yet to be validated by independent research efforts. Thus, much more work is needed to validate the use of this method in actual evaluation settings before it should be considered as a primary means for teacher evaluation.

Portfolios

Description

Portfolios are a collection of materials compiled by teachers to exhibit evidence of their teaching practices, school activities, and student progress. They are distinct from analyses of instructional artifacts in that portfolio materials are collected and created by the teacher for the purpose of evaluation and are meant to exhibit exemplary work, as opposed to a sampling of artifacts that are already being used in a teacher's classroom. The materials gathered are intended to demonstrate fulfillment of certain predetermined standards, and often portfolios are designed to promote teacher reflection and improvement in addition to being used for evaluation. Examples of portfolio materials include teacher lesson plans, schedules, assignments, assessments, student work samples, videos of classroom instruction and interaction, reflective writings, notes from parents, and special awards or recognitions. Part of the exercise for teachers is choosing a feasible number of artifacts that will represent the full range of their teaching practices and larger

school contributions while demonstrating how their performance meets the given standards. The portfolio process often requires a defense of why artifacts were included and how they relate to the standards (Painter, 2001).

Portfolios are commonly used in teacher preparation programs as a requirement for licensure, but states have increasingly adopted portfolio assessments for use in evaluating both beginning and experienced teachers. Vermont reformed their performance assessment program beginning in 1988, implementing a unique system that used performance assessments, namely portfolios, as a main source of evaluation instead of an addition to a more traditional program (Koretz, Stecher, Klein, & McCaffrey, 1994). Connecticut also has a well-known program, the Beginning Educator Support and Training (BEST) program, which requires teachers to complete portfolios as part of their continuing licensure requirements. Washington State's Professional Certificate Program offers an advanced certification that requires the completion of a classroom-based portfolio (see Office of Superintendent of Public Instruction, n.d.), and the state of Wisconsin has a voluntary Master Educator License that requires a teacher to demonstrate advanced proficiency on a portfolio assessment aligned with the Wisconsin Educator Development and Licensure Standards (see Wisconsin Department of Public Instruction, 2008). To illustrate the uses of portfolios in evaluation, Connecticut's BEST program and the well-known advanced certification program of the National Board for Professional Teaching Standard (NBPTS) are described in the following section.

Examples

Connecticut's Beginning Educator Support and Training (BEST) Program. The BEST program is a two-year induction, support, and assessment program for new teachers in the state of Connecticut. The first year consists of seminars, workshops, and meetings with an assigned mentor teacher, giving new teachers an opportunity to develop their practice. During the second year, teachers submit a portfolio for assessment of their practice, and a satisfactory evaluation is required for teachers to obtain full certification and remain teaching in the state. Teachers who do not pass the assessment must undergo further professional development and resubmit the portfolio during the third year; if they do not pass in the third year, they are no longer permitted to teach in Connecticut public schools. As a part of the program, teachers are entitled to school-based support in the form of mentorship, release time, and content-specific instructional support and to state-based support in the form of professional development seminars, conferences, and Internet-based resources. In turn, beginning teachers are expected to fulfill the requirements of the BEST program and keep their certification up to date using the resources provided to them (Connecticut State Department of Education, 2007; Pecheone & Stansbury, 1996).

The evaluation standards for BEST portfolios are culled from Connecticut's Common Core of Teaching standards and are based on demonstrating foundational skills that are believed to be common across teachers in all grade levels and subjects as well as establishing knowledge and competency in discipline-specific areas. BEST portfolios include "daily lesson plans for a five- to eight-hour unit of instruction with one class; two to four videotaped segments of teaching equaling in total approximately 30–40 minutes; examples of the work of two students; and reflective commentaries on teaching and learning that took place during the unit" (Connecticut State Department of Education, 2007, p.. 22). Portfolios are scored by experienced teachers in

the same discipline as the teacher being evaluated. These assessors are hired by the Connecticut State Department of Education, work for two years at the department as teachers in residence, and must participate in at least 50 hours of comprehensive training in scoring and pass reliability assessments. After portfolios are scored, teachers are provided with an individualized performance summary, which discusses their performance according to the categories of designing and implementing instruction, assessment of learning, and analysis of teaching. Portfolios are scored based on a series of discipline-specific guiding questions and performance indicators, which are included in portfolio handbooks so that teachers are fully aware of the evaluation criteria as they create their portfolios (Connecticut State Department of Education, 2007).

National Board for Professional Teaching Standards (NBPTS) Certification. NBPTS offers a certification system to recognize accomplished teachers who meet high and rigorous standards, and a main component of their evaluation is a portfolio assessment (the other component is an assessment of subject matter knowledge). NBPTS offers 25 certificates that cover a variety of subject areas and student developmental levels. Standards for certification in each area are created by committees of expert teachers and specialists in education, child development, and other relevant areas. The portfolio requirement consists of four different entries, three of which are classroom based and one which exhibits work with families, the community, colleagues, and the larger profession. Contents of the portfolios include video of instructional practice, video of teacher-student interactions, and student work samples; all entries must be accompanied by detailed reflection and analysis of the instruction represented. Portfolios are evaluated by assessors who have completed intensive training through NBPTS and met qualification requirements by demonstrating an understanding of the NBPTS standards, directions, scoring guides, and rubrics. Teachers and school counselors, especially those who have achieved National Board Certification, are eligible to apply to become assessors (National Board for Professional Teaching Standards, 2008).

Much research has been conducted on NBPTS certification. There are several studies linking NBPTS certification to gains in student achievement (e.g., Cavalluzzo, 2004; Clotfelter, Ladd, & Vigdor, 2006; Goldhaber & Anthony, 2004; Vandevoort et al., 2004), though there are also studies that do not find a relationship (e.g., Cunningham & Stone, 2005; McColskey et al., 2005; Sanders, Ashton, & Wright, 2005). In a recent evaluation commissioned by the U.S. Department of Education on the effects of NBPTS certification, the Committee on Evaluation of Teacher Certification determined that NBPTS certification is successful in identifying high-performing teachers, but not enough evidence exists to determine whether the process itself leads to improvements in practice or whether teachers who are already effective complete the process (Hakel, Koenig, & Elliott, 2008). Because NBPTS participation is strictly voluntary, findings from studies examining the impact of the NBPTS process on teachers can be hard to interpret. Teachers who pursue the certification are a self-selected group and may differ in significant ways from the teaching population as a whole (Pecheone, Pigg, Chung, & Souviney, 2005). Though the NBPTS process tends to be viewed by teachers as contributing to their learning and professional growth, these findings are based mainly on teacher or administrator perceptions (Pecheone et al., 2005) and have not yet been verified by studies using more direct measures of learning (Hakel et al., 2008).

Validity and Reliability Research. Portfolios can offer a very comprehensive and in-depth portrait of teaching activities; however, their complexity can raise concerns about the ability of scorers to evaluate them reliably. In a study on the implementation of the Vermont teacher assessment program, Koretz et al. (1994) discuss problems with the portfolio rating system in establishing rater reliability and distinguishing real differences in the quality of student work contained in the portfolios. They also describe related difficulties with establishing validity of the measure and using it for school accountability purposes.

Johnson, McDaniel, and Willeke (2000) point out that studies that have examined the interrater reliability of large-scale portfolio assessments have found that the percentage of agreement is usually between 45 percent to 75 percent, and correlations between raters rarely reach .80, which is considered by some as a necessary threshold of reliability. [The study cites Nunnally's (1978) argument "that test reliability of .80 was necessary for review of group means and at least .90 was necessary for reporting individual scores" (Johnson et al., 2000, p. 74)]. Thus, although some of these correlations are moderately high, they are lower than desirable for use in any high-stakes decision making. Johnson et al. also demonstrate that reliability is affected by the type and number of items being scored. In their investigation of interrater reliability for a smaller-scale family portfolio assessment, they examine separately the interrater reliability of ratings on six individual criteria, the composite of those six ratings, and an overall holistic rating. They found that in general, the reliability of rating individual criteria was consistently lower than the composite score and somewhat lower than the holistic score. They also conducted a decision study to determine the number of raters necessary to achieve a reasonable level of reliability for each of these categories, finding that three raters were desirable for the individual rating or the holistic rating but that two raters were sufficient for the composite rating.

Tucker, Stronge, Gareis, and Beers (2003) examined the validity and usefulness of teaching portfolios in assessing teacher performance for both accountability and professional development purposes. In teams of two, researchers rated a random stratified sample of 24 portfolios from elementary, middle, and secondary teachers, based on 18 teacher responsibilities specified by the district covering four major domains (instruction, assessment, management, and professionalism). Perceptions of the usefulness of portfolios were measured via survey and follow-up focus groups with teachers and administrators. Authors found that portfolios were able to document the fulfillment of the 18 teaching responsibilities and included representation of each of the four major domains, and 90 percent of the artifacts submitted had content validity (i.e., were relevant to the domains). Professionalism was the most highly represented domain, illustrating the role of portfolios in documenting aspects of teacher performance that cannot be measured through classroom observation. Administrators found that portfolios gave them a broader view of teacher activities and allowed them to make "finer distinctions about the quality of teacher performance" (Tucker et al., 2003, p. 572). Both teachers and administrators viewed portfolios as fair and accurate, but teachers expressed concerns about feasibility. There were mixed results regarding the usefulness of portfolios for professional growth, with some teachers reporting them helpful for reflecting on practice but with little evidence of impact on teaching practices. Tucker et al. suggest that teachers may need further training in order to make the connection between teaching reflections and changes in instructional practice.

Overall, these studies illustrate that although portfolios are an effective method for tapping into broader concepts of teacher development and providing valuable information to teachers about their practice, several issues in scoring portfolios exist, and more research is needed to fully assess their reliability and validity. Due to these concerns, some studies advise against the use of portfolios as a stand-alone assessment in high-stakes decision making (e.g., Johnson et al., 2000). In addition, there is a lack of studies that investigate the relationship between scores on portfolio assessments and student outcomes, and this area deserves much more research.

Strengths and Cautions

Portfolios do offer several advantages over some of the other measures of evaluation discussed. They are generally considered useful for providing a broad and varied view of a teacher's many capabilities and providing formative information and opportunities for teacher reflection that can enhance performance. They can be used with teachers in any subject or grade level and thus are useful in multiple contexts. They are a very comprehensive measure, with the ability to assess aspects of teaching that are not readily observable in the classroom and extend beyond classroom instruction. They also have high face validity, generally being viewed by teachers and administrators as "authentic" assessments that are relevant and useful to their teaching practice. Portfolio assessments provide the opportunity to actively involve teachers in the evaluation process and give them personal ownership of their improvement and professional growth, helping to reform the conception of evaluation as something done *to* teachers *by* administrators (Tucker et al., 2003).

As this discussion indicates, more research on the reliability and validity of portfolios as a performance assessment is needed before they should play a substantial role in evaluation for accountability purposes. They present a useful opportunity for providing formative assessment to teachers, though teachers may need training in order to learn how to choose relevant artifacts (Painter, 2001) and reflect on their practice in a way that fosters improvement and leads to actual changes in practice (Tucker et al., 2003). They also can become quite cumbersome for teachers, requiring a significant time commitment if they are to gain the most benefit from the portfolio process, thus it is recommended that teachers are provided with support and time to complete portfolio requirements. In a study of beginning teacher performance assessments in California, Mitchell, Scott, Hendrick, and Boyns (1998) found that the amount of priority placed on the program by the school and district was related to teachers' perceptions of fairness and helpfulness of the assessments (cited in Pecheone et al., 2005). This demonstrates how buy-in and support from the administration can be crucial to the success of a performance assessment program.

Tucker et al. (2003) make some useful observations and suggestions based on their validity study. They recommend that to maximize the efficacy of portfolio assessments, it is useful to include complete units of study (e.g., lesson plans, teaching strategies, sample assessments, and scoring rubrics; student work with teacher comments that pertain to the specific unit; and reflections on the artifacts the teacher chose to include with explanations of their relevance and importance). They also recommend the use of portfolios inclusively but not exclusively in the evaluation of teachers, as a complement to data collected through classroom observation, conferences, and client surveys.

Self-Reports of Teacher Practice

Description

This section examines different categories of self-report measures of teacher performance. These measures prompt teachers to report on what they are doing in the classroom and may take the form of surveys, instructional logs, and interviews. These measures vary widely depending on the focus, the level of detail they attempt to gather, and the intended use of the scores. Mullens (1995) describes several considerations in reference to designing large-scale survey measures of teaching, such as whether or not the aspects measured bear a relationship to student achievement or other outcomes of interest, whether the measures can inform policy and decision making aimed at educational improvement, and whether the measures can be used appropriately with the population of interest. For instance, as discussed in the observation section, survey measures may focus on broad and overarching aspects of teaching that are thought to be important in all contexts, or they may focus on specific subject matter, content areas, grade levels, or techniques. Survey measures may consist of straightforward checklists of easily observable behaviors and practices; they may contain rating scales that attempt to assess the extent to which certain practices are used or aligned with certain standards; or they may set out to measure the precise frequency of use of practices or standards. Thus, this class of measures is quite broad in scope, and considerations in choosing or designing a self-report measure will depend largely on its intended purpose and use.

Examples

Surveys. Several large-scale and well-known teaching surveys focus on measuring reform-oriented practices or enactment of curriculum. Examples of large-scale surveys include those developed by the National Center for Education Statistics (NCES); the Trends in International Mathematics and Science Study (TIMSS); Reform-Up-Close and the *Surveys of Enacted Curriculum* (SEC); and studies by the RAND Corporation, including the School Reform Assessment Project, Validating National Curriculum Indicators, and the California Learning Assessment System (CLAS). Some of these are broad and meant to be used with all teachers (e.g., NCES survey), whereas others are subject-specific and focused on content (e.g., TIMSS and CLAS surveys). Mullens (1995) identifies four broad dimensions of classroom instruction that are critical for large-scale surveys to address: pedagogy, professional development, instructional materials and technology, and topical coverage within courses. According to Mullens (1995), "All four dimensions under consideration have an established or expected relationship to student achievement and could provide interesting information about variation in achievement. Of the four, pedagogy and topical coverage within courses are more related to the teacher/student interaction and may therefore have a stronger relationship with student achievement" (p. 16).

One example of a thoughtfully developed and tested large-scale survey is the SEC, which were created as practical and reliable tools for data collection and reporting on instructional practices and content being taught in K–12 mathematics, science, and English language arts (ELA) classes. Blank, Porter, and Smithson (2001) describe how SEC data can be used in schools. The survey is conducted online, so results are tabulated and made accessible to schools in a variety of

formats. Data from the SEC allow administrators to examine differences between schools and teachers, compare instruction to standards, and evaluate the alignment between practices and standards. Like any effective evaluation instrument, it also provides a framework for communicating about practices and instruction, which can guide teacher reflection and lead to increased discussion and collaboration among colleagues. Blank et al. (2001) address concerns in the study about potential inconsistencies or inaccuracies in teacher responses due to factors such as differing interpretations of the terminology used and the time lag in reporting (teachers reported on their practices for the entirety of the semester or year). They also address concerns about low response rates; however, they express confidence in the accuracy of the teacher reports, citing findings from an earlier related study of Reform-Up-Close (Porter, Kirst, Osthoff, & Smithson, 1993), which compared teacher practices as measured by daily logs, independent observation, and teacher survey reports and found survey data to be highly correlated with the more detailed and frequently collected log measures.

Other studies also have investigated the validity of self-report survey data by comparing multiple measures. A study conducted by RAND Corporation researchers examined teachers' instructional practices using both self-report survey data and analysis of artifacts from teachers' classroom activities (Burstein et al., 1995). Researchers collected homework, quizzes, classroom exercises, projects, and exams from 70 mathematics teachers in California and Washington. They also analyzed daily logs kept for five weeks by the participating teachers, which described their instructional practice. The researchers found problems with the validity of the survey responses, stating, "instructional goals cannot be validly measured through national surveys of teachers. The data are inconsistent not only with artifact data but also with teachers' own self-reports on other survey items such as those describing their exam formats" (Burstein et al., 1995, p. 54). This finding raises concerns about the use of self-report survey data to represent teacher practices. It also might suggest that evaluating classroom artifacts, while considerably more expensive, may provide better evidence of actual teacher practices than self-report data. However, more research is needed to examine the validity of these measures.

Mayer (1999) conducted a study to examine the validity of teacher self-report data on instructional practices by surveying Algebra I teachers on their use of practices that reflected teaching standards set forth by the National Council of Teachers of Mathematics (NCTM). The author calculated the time teachers reported spending on certain practices aligned with the standards, comparing this with observational measures of the time they spent engaging in those practices. The study found that observational and survey measures were highly correlated but that survey measures were systematically inflated. It also determined that measures of individual practices were not reliable; however, composite measures of teaching practices were valid, and relative rankings of practices used were generally consistent. In other words, the survey could indicate the extent to which a teacher utilized a group of instructional practices as compared to other teachers but could not accurately measure the amount of time spent on individual practices. In addition, when a teacher reported using certain practices, the survey did not reveal anything about the level or quality of their implementation. Though sample sizes were small, these findings reveal important distinctions about the quality of information that can be gleaned from self-report survey data.

Logs. In contrast to broad surveys, instructional logs require teachers to keep a frequent and detailed record of teaching. The logs are highly structured and ask for specific information regarding content coverage and use by both the teacher and students. Much of the development and research work in the area of instructional logs has been conducted by researchers from the Consortium for Policy Research in Education (CPRE), as part of their larger Study of Instructional Improvement. The study is a comprehensive examination of measures of teaching, using multiple methods to gather data on instruction, including questionnaires, instructional logs, classroom observations, and teacher interviews. Ball and Rowan (2004) describe how the logs came to be developed: "Because gathering annual data on daily instruction likely often misrepresented actual practice, more frequently administered logs emerged as an approach to gathering information about content covered" (p. 4).

Camburn and Barnes (2004) examined the validity of these instructional logs, focusing on language arts lessons, by comparing teacher log responses with responses given by third-party observers. The log consisted of 150 items, including detailed information on content and emphasis on curricular areas. Thirty-one teachers who were pilot-testing the logs in eight public elementary schools were observed for one day, and both the teachers and observers completed a log for each lesson. One of the main findings revealed that teacher and researcher reports did not always agree, and scores between researchers were nearly always more highly correlated than scores between researchers and teachers, indicating that "researchers and teachers may have brought different perspectives to bear when completing the language arts log, perhaps drawing on different knowledge and experiences" (p. 59). Authors speculated that because observers have a more limited experience with the classroom than teachers, they may lack certain contextual information or interpret information differently when making judgments that reflect how a teacher perceives his or her intentions and practices. The importance of establishing a common understanding of terminology between teachers and raters also was raised, as differing interpretations of glossary terms may have contributed to inconsistencies in ratings. The study also found that rater agreement was affected by the degree of detail in the category being scored, the frequency of the instructional activity, and the content being covered.

In addition, Camburn and Barnes (2004) suggest that the ability to create a clear shared understanding with teachers through a log remains a challenge and is a significant threat to construct validity. They argue that researchers may face a trade-off between measuring subtle differences in content use that may affect student learning and the use of categories that measure broader aspects of instruction. They explain that "the former approach, which parses instruction more finely, makes interrater agreement more difficult to obtain and poses a threat to the validity of the measures. The latter approach may miss nuances in instruction that are theoretically and empirically important but may yield more valid measurement" (pp. 65–66). This study raises an important issue, which relates to the aforementioned studies: discrepancies between teacher self-reports of practice and third-party observer reports may not simply reflect inaccuracy on the part of the teacher but may uncover a larger issue concerning the differing values, knowledge, and interpretations that these two parties inherently bring into their evaluations. This is certainly an area worthy of further investigation.

Interviews. Another method for investigating teachers' self-reported practices is to utilize an interview protocol. Interviews are most often used as supplements to other measures of teaching

and are particularly useful in providing qualitative information that supports or explains results obtained from more quantitative measures. Studies that attempt to triangulate several measures of teaching in order to ensure accuracy of the results may employ an interview protocol, such as the aforementioned Study of Instructional Improvement (see Ball & Rowan, 2004) and the RAND Mosaic Study (see Le et al., 2006). The Mosaic Study examined the use of reform-oriented teaching practices, employing several measures including teacher surveys, instructional logs, structured vignettes, and observers' ratings of classrooms. An interview protocol was developed to investigate whether teachers felt that local systemic reforms and other policies were influencing their practices. This illustrates the very unique role interviews can play in gathering information on perceptions and opinions that may inform the "whys" and "hows" of measuring teacher performance and its impact.

Interview protocols can be highly structured or largely open-ended and can be a means for gathering data on practice that is more detailed or in-depth than survey measures. They are generally locally designed and intended for use in the context for which they were created. Few studies examine the reliability or validity of interview protocols intended to be used on a larger scale. One example is a study by Flowers and Hancock (2003), which describes the development of an interview protocol focused on professional standards and student learning. They describe the advantage of their interview protocol as a "method of collecting data from multiple sources while avoiding the shortcomings of singularly focused evaluation systems" (p. 163). The interview questions require teachers to provide specific examples of their instructional activities, intentions behind the activities, and specific actions they have taken to monitor and improve student learning. The protocol includes a structured scoring rubric with detailed criteria included for each rating. Evaluators must be trained on the interview protocol and scoring rubric, and teachers should be provided with the interview procedure and standards prior to the interview so that they can prepare materials in advance and formulate any clarifying questions they may have. This study reports high interrater reliability and rater consistency for the protocol, and extensive feedback from experts in the field helped to establish its content validity.

Strengths and Cautions

Teacher self-report methods may be one useful element in a teacher evaluation system, as they do have certain advantages. Self-report data can tap into a teacher's intentions, thought processes, knowledge, and beliefs better than the other methods discussed, and they can be useful for teacher self-reflection and formative purposes. In addition, it is important to consider the perspectives of teachers and involve them in their own evaluation because they are the only ones with full knowledge of their abilities, classroom context, and curricular content, and thus can provide insight that an outside observer may not recognize. Surveys are a cost-efficient, generally unobtrusive way to gather a large array of data at once. Using one instrument, data can be collected on instructional practices as well as administrative support, professional development opportunities, relationships with students, school climate, working conditions, demographic or background information, and perceptions or opinions that may have bearing on the effectiveness of a teacher.

Teacher self-report measures may be an efficient means of obtaining information about instructional practices without incurring the high costs of observation or other measures and can

be particularly useful as a first step toward investigating some question of interest (e.g., establishing some basic level of standard use and understanding among teachers) (Cohen & Hill, 2000; Spector, 1994). However, extreme caution should be taken not to base potentially consequential decisions on results of self-report measures. Research findings on the reliability and validity of these methods have produced mixed results. Concerns have been raised in the literature about self-report responses being susceptible to social desirability, defined by Moorman and Podsakoff (1992) in the organizational psychology literature as "the tendency on the part of individuals to present themselves in a favourable light" (p. 132). This phenomenon would include both the conscious misrepresentation of teaching practices to "look good" as well as unintentional misreporting due to a teacher's perception that he or she is correctly implementing a practice when in fact it is not being implemented with fidelity. Potential biases may lead to both overreporting and underreporting of practices, making the data difficult to interpret. Although this phenomenon has been widely researched in the psychology literature, more research is needed to determine the extent of its effect in the context of education and teaching. Some of the inconsistency caused by socially desirable responding may be controlled by ensuring the confidentiality and anonymity of teacher responses, gathering data longitudinally rather than just at one point in time, and gathering data from more than one source. However, these measures are not likely to eliminate all bias (Spector, 1994).

Several additional concerns warrant attention when selecting, designing, or administering self-report measures. An issue raised by several studies is the importance of ensuring consistent interpretations of terminology and a shared understanding of what the measures entail (Ball & Rowan, 2004; Blank et al., 2001; Mullens, 1995). This may require training of both teachers and outside raters (if applicable) on the survey or log measure in order to elicit the intended information. In addition, consideration should be taken to determine how broad or how detailed a survey needs to be to inform its desired purpose. Mullens (1995) notes, "Because the number of questions and the respondent burden by necessity must be limited, …indepth questions often preempt items representing a broader range of inquiry and may result in specific and often detailed information about a relatively narrow range of interest" (p. 18). Conversely, gathering information on a wider range of topics or practices may result in an insufficient amount of detail. Blank et al. (2001) also make the point that selecting a random and/or representative sample and ensuring high response rates are important considerations for obtaining valid self-report measures. Their study indicates that response rates were highest when teachers were given in-house time and support to complete the measures. In addition, teachers were more likely to complete measures when they received something of personal value from the process. Blank et al. (2004), therefore, recommend providing teachers with results that may inform their practice and assuring teachers that responses are confidential and will not be used in any way for accountability purposes.

Student Ratings

Description

It can be argued that student opinions of a teacher are an important consideration in any teacher evaluation system because students have the most contact with teachers and are the direct consumers of a teacher's services. Given their extensive experience with teachers, it seems that

valuable information can be obtained through student evaluations of teachers in the form of surveys or rating scales. However, student ratings of teachers are sometimes not considered a valid source of information because of potential biases that may affect their ratings and lack of knowledge about the full context of teaching. For example, studies have investigated whether student ratings are influenced by student age or academic level, expected or actual grades, and level of course challenge (e.g., Worrell & Kuterbach, 2001). As with teacher self-report measures, the reliability and validity of student ratings depend to some extent on the instrument used, how it is developed, how it is administered, and the level of detail it attempts to measure. The following example studies investigate the validity of student ratings for evaluating teachers.

Examples

Peterson, Wahlquist, and Bone (2000) examined whether student ratings could provide reliable and valid information to teacher evaluation. An item analysis of 9,765 student surveys, which varied by grade level (primary, elementary, and secondary), showed that students responded reliably and validly when rating their classroom teachers, though scores tended to be skewed toward high satisfaction. The study also revealed that students of different age groups may focus on different aspects of teaching. Findings showed that younger students were more concerned with teacher-student relationship (e.g., "teacher shows caring and respect"), whereas older students placed more weight on student learning. The study also reported that teachers were favorable toward having student ratings as one part of their larger evaluation system, attesting to the face validity of student ratings.

There is also evidence that student ratings can be valid predictors of student achievement. A study of schools in Cyprus by Kyriakides (2005) included a student survey of teacher practices in which the rating scales relating to teacher-student relationship and the degree of cooperation between teacher and students were highly correlated with achievement gains for mathematics and Greek language as well as with affective outcomes of schooling. In a study that compared principal ratings, student ratings, and teacher self-ratings to measures of student achievement on criterion-referenced tests in mathematics and reading, Wilkerson et al. (2000) found that student ratings were more highly correlated with student achievement than the other ratings and were the best predictor of student achievement across all subjects. These studies provide convincing evidence that student ratings of teaching are worth considering for inclusion in teacher evaluation systems.

Strengths and Cautions

There are several persuasive arguments for considering student ratings of teachers as part of the teacher evaluation process. In an empirical literature review on using public secondary school students' ratings to evaluate teachers, Follman (1992) notes that students are the most direct clients of teachers and, thus, have a broader and deeper experience with teachers than other potential evaluators, including principals, administrators, peers, or parents. A teacher's first responsibility is to his or her students, and students are in turn the most frequent source of feedback on a teacher's performance. Follman (1992, 1995) goes on to conclude that although validity concerns, such as rating leniency and halo effects (i.e., when an opinion on one trait or aspect of teaching influences all other ratings in the same direction) may affect student

evaluations of teaching, they do not seem to affect students more so than adult raters. Secondary students were shown to be capable of providing reliable ratings, validly reporting classroom events and teacher interactions, and judging whether or not a teacher is "meritorious."

In a study showing that high-achieving secondary school students could rate teaching behaviors as reliably and validly as college students, Worrell and Kuterbach (2001) note that student ratings are cost-efficient and time-efficient, can be collected anonymously, and can be used to track changes over time. They also require minimal training, though employing a well-designed rating instrument that includes detailed items measuring meaningful teacher behaviors would be important in maintaining the validity of the results.

However, researchers caution that student ratings should not be stand-alone evaluation measures because students are not usually qualified to rate teachers on curriculum, classroom management, content knowledge, collegiality, or other areas associated with effective teaching (Follman, 1992; Worrell & Kuterbach, 2001). Overall, the reviewed studies recommend that student ratings be included as part of the teacher evaluation process but not as the primary or sole evaluation criterion.

Value-Added Models

Description

Value-added measures provide a summary score of the "contribution of various factors toward growth in student achievement" (Goldhaber & Anthony, 2003, p. 38). Value-added models can be defined of "a collection of complex statistical techniques that use multiple years of students' test score data to estimate the effects of individual schools or teachers" (McCaffrey et al, 2003, p. xi). Although value-added models also may be used to evaluate schools for accountability purposes, this research synthesis concerns their use for evaluating teachers in terms of their effectiveness relative to other teachers.

Measuring effectiveness at the classroom level, rather than at the school level, is increasingly the focus of effectiveness research (Creemers & Reezigt, 1996). Researchers have focused on trying to determine teacher effectiveness by examining teachers' contribution to student achievement gains for many years, but a lack of valid measures and instrumentation has hampered the process. Only in the last 10–15 years have researchers had the necessary combination of sufficient computing power, extensive data on student achievement linked to individual teachers, and appropriate statistical models with which to determine effectiveness in terms of teachers' contributions to student learning. The result is a set of sophisticated statistical models that are used with linked student-teacher data to measure teachers' contributions to the student achievement growth of the students they taught in a given year.

Value-added models are promising, controversial, and increasingly common as a method of determining teacher effectiveness (when effectiveness is construed as teachers' contributions to achievement). However, it is also the method that is the least understood by most education professionals and teachers. Unlike classroom observations in which the teachers actually meet their evaluator, value-added model evaluators conduct their analyses from afar.

The models are complex; however, the underlying assumptions are straightforward: students' prior achievement on standardized tests can be used to predict their achievement in a specific subject the next year. Whether the student met, exceeded, or failed to reach the predicted score forms the basis for the teachers' effectiveness score. When most students in a particular classroom perform better than predicted on standardized achievement tests, the teacher is credited with being an effective teacher, but when most students' in a particular classroom fail to meet predicted gain scores, the teacher may be deemed less effective. In some models, students' prior achievement scores are the basis for calculations of effectiveness, whereas other models include students' gender, race, and socioeconomic background, and still others include information about teachers' experience.

Examples

Heneman, Milanowski, et al. (2006) conducted a multiyear mixed-methods study investigating the validity of teacher evaluation systems in four sites throughout the country. The instruments they examined were modifications of Danielson's (1996) *Framework for Teaching* and included planning and preparation, the classroom environment, instruction, and professional responsibilities. They used a value-added model in which achievement was estimated based on prior achievement and other student characteristics and found positive relationships between teacher evaluation scores and student achievement gains, although there was substantial variability across sites (and within sites). Although the study focused on the evaluation instruments, there was a fairly high correlation in two sites between what the teachers were observed to be doing in their classrooms and the achievement gains of their students. The authors theorized that the higher correlation was likely due to using multiple evaluators and, in Cincinnati, highly trained evaluators. At the sites with lower correlation, there was a single evaluator with less training conducting the evaluations.

Heneman, Milanowski, et al. (2006) focus attention on one type of research that may prove to be useful in establishing the validity of various measures of teacher effectiveness. This type of research correlates scores on various measures to draw conclusions about the information the measures can actually provide. For example, they speculated that finding links between what teachers did and student test scores was in part dependent on the performance of the classroom evaluators, not just the performance of the teachers. Although they did find some connection between teachers' performance and student test scores, the findings were not consistent across sites, suggesting that using value-added strategies instead of classroom observations as a measure of teacher effectiveness does not necessarily result in more valid assessments.

Holtzapple (2003) used Danielson's (1996) *Framework for Teaching* to compare student achievement with teachers' evaluation scores using a value-added model of predicted achievement versus actual achievement in Cincinnati. The author found a correlation between the observation scores and the value-added scores for teachers: teachers who received low ratings on the instructional domain of the teacher evaluation system had students with lower achievement, teachers with *advanced* or *distinguished* rankings on this instrument generally had students with higher-than-expected scores, and teachers rated *proficient* had students with average gains.

Interestingly, one of the sites investigated by Heneman, Milanowski, et al. (2006) was Cincinnati, and it was one of the sites that had higher correlations between the observations and the value-added scores. Cincinnati had highly trained raters conducting evaluations, which may explain the correlation. If, in fact, observable teacher practices lead to improved student learning, then there certainly should be a correlation between these two measures.

A similar study by Kimball et al. (2004) examined the relationship between teacher evaluation scores and student achievement in nine grade-test combinations in one county. Using an adaptation of Danielson's (1996) *Framework for Teaching,* this study estimated teacher effects on student achievement and determined that teacher practices contributed slightly to student achievement. However, only two of the correlations were statistically significant. This finding suggests that there is still much to learn about what value-added models are actually measuring because the research is not providing strong, consistent correlations between what teachers do in their classrooms and value-added scores.

Other researchers have calculated value-added scores for teachers and then tried to correlate them with other explanatory information. For example, Aaronson, Barrow, and Sander (2007) conducted a study using Chicago public high school data, focusing on mathematics. They calculated value-added scores for teachers and then attempted to correlate these scores with teacher characteristics including age, experience, degree level, certification, and undergraduate major. They found that almost none of the variance in teacher effectiveness—except having an undergraduate major in mathematics or science—was accounted for by these characteristics. The authors concluded that the differences in teachers were not to be found among the teacher characteristics for which they had data. This study demonstrates an unfortunate fact about value-added scores—they reveal nothing about *why* teachers vary in their effectiveness as measured by student achievement score gains. Thus, it is impossible to either predict which teachers will be most effective or help less effective teachers improve.

Rivkin et al. (2005) attempted to correlate observable teacher characteristics, such as education and experience and unobservable components to student achievement gains in Texas. They determined that observable teacher characteristics have small but significant effects on student achievement gains but found that the majority of teacher effectiveness cannot be explained by these observable characteristics. In other words, they demonstrated that teachers vary in their contribution to students' achievement score gains, but they could not explain what caused the variation. Again, this study points out a key problem with value-added measures—they do not enhance understanding of what effective teachers do that makes them effective.

Another study focused on whether teachers fostered student creativity in their classrooms and used observation scores as predictors of student achievement gains (Schacter, Thum, & Zifkin, 2006). After multiple classroom observations, the researchers found that most teachers did not employ teaching strategies that encouraged students' creativity, but when they did, the result was improved student achievement. This study illustrates an important point about using value-added models: High-quality observational data, when combined with a sound value-added model, may provide useful information about differences in teaching that could lead to strategies for improving student outcomes. In this instance, if the teacher behaviors that promoted student

creativity could be taught to other teachers, better student achievement might result. On the other hand, without the observational data, the authors would know only that students of some teachers had better achievement gains—but they would not know what practices were responsible for those differences. Clearly, value-added models have great potential for improving instruction when combined with observational data, though there are still questions to be answered. Chief among them is how to sort out the impact of one particular variable—teaching for creativity, for instance—from all of the other interactions between teachers and students that lead to learning.

Value-added models also are being used for research projects examining teacher preparation programs, such as the Carnegie-funded Teachers for a New Era (Sanders & Rivers, 2006) and Louisiana State University's value-added assessment of teacher preparation (Noell, Porter, & Patt, 2007). The goal of these studies is to better understand the relationship between what teachers learn in preparation programs and their students' achievement gains. Unfortunately, the inability to get appropriate longitudinally linked student-teacher data has hampered such efforts.

Brief summaries of other studies appear in the appendixes. There is little validity evidence linking value-added scores to teacher characteristics or practices—or even to school characteristics or practices. Teachers vary greatly—even within schools—in their effectiveness as measured by standardized test scores, but that variation has not been consistently and strongly linked to what teachers do in their classrooms. This suggests that either classroom observation instruments are not sensitive enough to capture the differences that matter in terms of student achievement or that other things are being measured that have not yet been conceptualized. So, although it is possible to say that students in one classroom learned more than students in another, it is not possible to say with any certainty why that occurred. Thus, value-added models are limited in their usefulness because the information gleaned from them is essentially a "black box"—the classroom context and teacher characteristics, qualifications, and practices that produced the value-added scores are unknown. This speaks to the importance of having additional components of a useful system of evaluating teacher effectiveness.

Strengths and Cautions

Value-added models are a relatively new way to measure teacher effectiveness, and there are researchers who support their use (e.g., Hershberg et al., 2004; Sanders, 2000). These researchers argue that value-added models provide an objective means of determining which teachers are successful at improving student learning as measured by gains on standardized tests. It is possible for teachers evaluated with a classroom observation instrument to receive a high score but still have students with average or below-average achievement growth. In addition, observation instruments can be used to evaluate teachers on their use of teaching practices that reflect experts' beliefs about good teaching, but there is a dearth of empirical evidence that specific teaching practices improve student learning (see Goe, 2007, for a synthesis of this research). This mismatch between what teachers do in their classrooms and student achievement gains may be due in part to the difficulty of measuring differences in teaching practices with standardized achievement outcomes (see Valli et al., 2004, for a discussion of these difficulties). Because value-added measures focus only on actual student gains on standardized tests, the extent to which teachers' practices reflect an instructional ideal is not relevant. Under this model,

teacher effectiveness is based on confidence that student test scores are valid and reliable indicators of student learning.

Value-added results may be able to help identify exemplary teachers. Across schools and even within schools, there are considerable differences among teachers in terms of their contributions to student learning (Rivkin et al., 2005; Rockoff, 2004). New or struggling teachers may benefit by observing highly effective teachers, but these outstanding teachers are often identified through their reputation. Value-added scores provide a means to identify highly effective teachers whose practices contribute the most to student learning gains. Establishing these teachers' classrooms as "learning labs" for colleagues and researchers may provide valuable information about what practices and processes contribute to student achievement gains. It would be especially useful to identify—and learn from—teachers who are successfully teaching students who are at-risk for poor educational outcomes.

Despite these potentially positive uses for value-added models, some researchers express reservations and describe serious concerns about their use for assessing teacher effectiveness (e.g., Bracey, 2004; Braun, 2005b; Kupermintz, 2003; McCaffrey et al., 2003; Thum, 2003). In his critique of value-added models, Bracey (2004) makes an interesting point: "V[alue added assessment] is not a theory of what makes a good teacher in all the complexity that that might require. It was developed as an atheoretical method, a technology" (p. 333). Bracey highlights a key issue of using value-added methods as a means of evaluating teacher effectiveness—that good teaching is complex, and the "technology" of value-added models examines what appears to be the results of that complex process, without regard to the causes.

Heubert and Hauser (1999), in their National Research Council report on high-stakes testing, recommended that, "accountability for educational outcomes should be a shared responsibility of states, school districts, public officials, educators, parents, and students" (p. 3). Using value-added models as the primary means of evaluating teacher effectiveness is not recommended because it holds teachers solely accountable for achievement, rather than including others who also contribute to student outcomes. Using a single score for a teacher as a measure of his or her effectiveness suggests that all, or nearly all, of the student learning in a particular subject or classroom in a given year was the product of a single teacher's efforts.

It is not just the use of value-added models that is subject to cautions from researchers. Berliner (1976) discussed the "obstructions to the study of teacher effectiveness," identifying the lack of "replicable findings relating teaching behavior to student achievement in natural classroom settings" as a key issue and noted that "instrumentation problems connected with the independent and dependent variables commonly used in research on teacher effectiveness" (p. 5) made data collection and analysis problematic. More than 30 years later, the same "obstructions" hamper the work of evaluating teacher effectiveness, particularly using student achievement to measure teacher effectiveness.

In fact, criticisms of using test scores to measure teacher effectiveness are not new. Shavelson et al. (1986) critique the process-product research that was popular in the 1970s in which researchers studied the link between teacher behaviors and student outcomes. Their appraisal of the process-product research focuses on the following four factors:

- Perfect alignment of local curriculum with the standardized test is assumed, when such alignment seldom exists, resulting in teachers being judged by their adherence to the *test's* curriculum.
- Standardized tests are strictly summative, but summary scores are not adequate reflections of improvements in students' cognition; thus, important information about students' capacity for understanding is not tested.
- Students' performance on the test is equated with their knowledge of the subject, even though the tests may be inaccurate measures of that knowledge, due to motivation, test-taking strategies, and attitudes toward testing—all of which are "extra-knowledge" influences that may affect test scores.
- Aggregating test scores across all students in a classroom may mask teachers' contributions to student learning by ignoring differential learning among students that actually reflects teachers' abilities to target appropriate instruction based on individual needs.

Shavelson et al. (1986) argue for measuring teacher effectiveness in ways that "attend to the organization of instruction in classrooms and differences in students' reactions to it" (p. 57).

The concerns about what value-added models can and cannot measure in terms of teacher effectiveness have not prevented the growth of value-added models as a seemingly objective measure of teacher effectiveness. Many states—including North Carolina, Pennsylvania, Ohio, Tennessee, Louisiana, and Florida—now use some type of value-added modeling, though they do not all use the results as a means of ranking teachers. However, an increasing number of states and school districts are exploring the use of value-added models either instead of or as a component of their previous systems of evaluating teacher effectiveness. Given this increased use of value-added models in this way, it is important to consider whether they are valid measures of teacher effectiveness.

McCaffrey et al. (2003) have argued that incomplete data and confounding influences that impact student scores that may not be included in the models (e.g., school effects) present major challenges to using value-added models for determining teacher effectiveness. In fact, Braun (2005a) has stated that what are typically called "teacher effects" are more accurately termed "classroom effects." This distinction is made because student learning is impacted by many variables in classrooms besides the teacher. It is not possible to sort out what part of a students' growth (or lack of growth) is solely attributable to the teacher's efforts. Thus, it is possible to see that students in one classroom had greater gains in achievement; however, the statistical models reveal nothing about why this is so, nor how much of the difference in student gains was due to effective teaching rather than other variables.

Another issue that has been raised by researchers is the impact of value-added measures of nonrandom assignment of students to teachers. Students are assigned to teachers by a number of methods. Different schools use different strategies, but the result is that the students in a given classroom were likely assigned to that classroom for a reason. If all students were randomly assigned to classrooms, there would be much more confidence in the resulting scores from the

use of value-added models. Several researchers have conducted studies that examine the impact of nonrandom assignment on value-added scores and concluded that there are no currently used models that adequately deal with the problem of nonrandom assignment (e.g., Rivkin & Ishii, 2008; Rothstein, 2008a, 2008b).

Finally, the validity of using value-added models for measuring teacher effectiveness is dependent in part on whether the statistical models are correctly specified and whether the inferences drawn are appropriate and defensible. The causative elements are not usually included in the modeling. Teachers teach, but *what* and *how* they teach are not part of the statistical model. So even though it has been determined that teachers differ in effectiveness in terms of producing student learning gains, ways to replicate those differences are not apparent. Even if teachers could be cloned, the teaching context (students, curriculum, resources, parental support, school leadership, etc.) would vary. Teachers may be differentially effective (i.e., a teacher who is successful in one context may be less successful in another).

Toward a Comprehensive View of Teacher Effectiveness

In many states, teacher effectiveness is assessed by focusing on results from a single measure, typically classroom observations and less commonly, teachers' contributions to student achievement growth (value-added models are one mechanism for examining this growth). Revisiting the five-point definition of teacher effectiveness, it is clear that using one or even both of these methods of measuring teacher effectiveness fails to indicate the many important ways in which teachers contribute to the success and well-being of their students, classrooms, and schools. Thus, creating a comprehensive score for teachers that includes multiple measures is one possible way to capture information that is not included in most classroom observation protocols or in scores developed using value-added models.

What types of measures might be included in this comprehensive measure? Here are some options for collecting data (from New Mexico's teacher performance evaluation guidelines): "review of videotape (of lesson); written documentation of activities; locally developed survey of staff, students, and/or parents; review of student work and performance; review of the teacher's contribution to the school's vision, mission, and outcomes; portfolios; information gained through peer observation and/or peer coaching; anecdotal records; reflective journals; self-evaluations; instructional artifacts; other formats" (New Mexico 3-Tier Licensure Implementation Teacher Training Work Group, 2005, p. 9). Unfortunately, there is little empirical evidence of the validity of these various methods for measuring teacher effectiveness, and in many cases, there are no standardized instruments for data collection. Instead, the collection of data—and decisions about what is important to collect—is left up to local decision makers.

Considering Teaching Contexts

Deciding how teacher effectiveness should be measured is not necessarily the sole purview of policymakers, researchers, and bureaucrats. Given that teaching contexts vary widely, it is essential that local input is considered when decisions are made about what to prioritize in a composite measure of teacher effectiveness. For example, a district with a high percentage of English language learners may want to consider teachers' ability to communicate effectively with these students and their parents as part of their composite measure of teacher effectiveness. Similarly, an urban school that has a high proportion of student dropouts may want to include a measure of teachers' documented efforts to assist at-risk students as part of their composite measure of teacher effectiveness. And a school in which teacher collegiality has been lacking might want to consider evidence of ways in which teachers initiate, lead, or support efforts to work together in professional learning communities.

Given that instruments and protocols for measuring teachers' leadership activities or contributions to improvement in school climate have yet to be developed in some cases, and standardized in most cases, it is not possible to make recommendations about what a state or local education agency should include in the creation of a valid composite measure of teacher effectiveness. Rather, it is recommended that the definition of teacher effectiveness be broadened, that it be inclusive of state and local priorities, and that it consider teaching contexts. Obviously, some schools have little or no problem with student attendance or dropouts, whereas other schools may lose days of students' learning time or lose students altogether. In some

schools, then, a measure of ways in which teachers have worked toward improving attendance or preventing students from dropping out would be a low priority, whereas such a measure would be a high priority in other schools.

Another consideration is that teaching contexts differ greatly across subjects and grades, and some types of measures may be more suitable for certain types of contexts. Campbell et al. (2003) critique teacher effectiveness models that are applied equally to all school levels and contexts, without regard to what may distinguish effectiveness in a particular subject, grade, or context. They argue for incorporating five dimensions of differential teacher effectiveness: "differences in activity, differences in subjects and/or components of subjects, differences in pupils' background factors, differences in pupils' personal characteristics, differences in cultural and organisational context" (p. 354).

Most classroom observation protocols, including Praxis III and Charlotte Danielson's (1996) *Framework*, are intended for use in all classrooms without regard to context. The CLASS instrument, however, has a Grades PK–3 version that has been extensively tested (La Paro, Pianta, & Stuhlman, 2004) as well as a more recently developed middle- and secondary version that is currently being piloted (Pianta, Hamre, Haynes, Mintz, & La Paro, 2007). These different versions of CLASS take into account the differences in teaching contexts at those levels. However, it may be possible to use a single instrument to evaluate teachers in different subjects, grade levels, and school contexts. The differences would then have to be accounted for in the scoring rather than in observation.

In their choice of teaching preparation programs, teachers select a grade level and subject in which they feel they have the most to offer their students. In their choice of schools, teachers select a context in which they feel they are likely to be successful. Yet many evaluation instruments do not acknowledge that teachers may be differentially successful depending on the context. What does this mean in terms of teacher effectiveness? First, teachers are not interchangeable—a teacher that performs well in one classroom may feel challenged in another classroom. Thus, an evaluation of teacher effectiveness should be specific to a context, subject, and grade level, and teachers should be compared with or ranked against teachers who are in similar contexts, subjects, and grade levels. In addition, evaluating a secondary science teacher's effectiveness on the same scale as that of a kindergarten teacher's effectiveness may be problematic, particularly if there is a need to identify exceptional teachers in specific contexts, grades, or subjects. This need might arise from a number of situations, including identifying a suitable mentor for a novice middle-school ELA teacher, rewarding exceptional teaching at the elementary level, recruiting teachers who have proven to be especially able to work with at-risk students for a special program within a school, or even offering an incentive for transferring to a hard-to-staff school. Lastly, taking into account teachers' evaluations when making hiring or transfer decisions might ensure a better match to open positions. A teacher's record of effectiveness in a specific setting may be a factor worth considering.

Using Teacher Effectiveness Results to Improve Instruction

There are many different purposes for evaluating teacher effectiveness; a key reason is to identify weaknesses in instruction and develop ways to address them. For this reason, one goal of evaluating teaching effectiveness should be to collect information that will be useful in designing appropriate strategies to improve instruction. Approaches to improving instruction may involve professional development, individualized work with a curriculum specialist, college coursework, and study teams within or across schools. Smylie and Wenzel (2006), citing a number of successes among school districts around the country, recommend a "human resources management" approach to improving instruction, wherein vertical and horizontal alignment of practices enable school leaders to carry out instructional objectives. They reported on three Chicago elementary schools that coordinated and aligned human resources to improve practices, including "teacher recruitment and induction, professional development activities, communication of expectations for teacher performance, specification of classroom teaching strategies, provision of encouragement and incentives, principal supervision and evaluation, and removal of poorly performing teachers" (p. 24).

Other sites may choose a more individualized approach to improving instruction, allowing teachers to plan their own professional growth. Denver's Professional Compensation System (ProComp) is an example of a district that has created a sophisticated system that permits considerable flexibility for teachers to decide how they will improve instruction (for additional information, see the ProComp website at www.denverprocomp.org). In collaboration with principals and supervisors, teachers can create a plan for their professional development, including taking courses (with tuition reimbursement) that will address gaps in their knowledge. Teachers and their supervisors can use evaluation results (from classroom observations and student achievement gains) to help them determine areas that need to be addressed.

Although there are many possible approaches besides those mentioned, the point is that evaluating teacher effectiveness should ultimately lead to improved instruction. In addition, under the broad definition of teacher effectiveness presented in this synthesis, evaluations also can be used to identify other areas in which teachers are performing well or they may need additional support. For example, if a district's priority is decreasing referrals to special education by identifying and providing assistance to at-risk students, it may be necessary to create opportunities for teachers to collaborate with colleagues and other education professionals during the school day.

A Final Note About Validity

When designing systems for evaluating teacher effectiveness and using the results of such evaluation, it is important to keep in mind that ways of measuring teacher effectiveness—such as classroom observation protocols or value-added models—are not valid in and of themselves for determining teacher effectiveness. Rather, their validity lies in their ability—when used correctly—to accurately and reliably measure what they were intended to measure. For classroom observation instruments, validity lies in the instrument's ability to measure how well a teacher exemplifies standards of practice that have been deemed important for that grade level, subject, and teaching context by some group of experts. For value-added measures, validity lies

in how well the model accurately captures an individual teacher's contribution to student achievement growth in a particular subject area.

At this juncture, researchers still have a long way to go toward clearly establishing the validity of various instruments for the purpose of measuring teacher effectiveness. There have been many research studies published to establish the validity of various measures of teacher effectiveness (e.g., examining how a score from an observation instrument correlates with a value-added score); however, validity cannot be determined by correlating results from measures based on two different constructs. Rather, validity must be determined by how well a given teacher's performance matches the construct—whether that means keeping at-risk students in school, contributing to a positive classroom environment, or having a high value-added score. Thus, the crucial step in getting valid information is deciding what is important and then finding (perhaps creating) a measure that will yield concrete evidence about teachers' performance on what is important. In a broad definition of teacher effectiveness, such as the one suggested, there is no single measure that will provide valid information on all the ways teachers contribute to student learning and growth and to their schools. Multiple measures—each designed to measure different aspects of teacher effectiveness—must be employed.

Policy Recommendations and Implications

The following set of recommendations is designed to provide guidance to entities that are considering how best to measure teacher effectiveness:

- Resist pressures to reduce the definition of teacher effectiveness to a single score obtained with an observation instrument or through using a value-added model. Although it may be convenient to adopt a single measure of teacher effectiveness, there is no *single* measure that captures everything important that a teacher contributes to educational, social, and behavioral growth of students, not to mention ways teachers impact classrooms, colleagues, schools, and communities.
- Consider the purpose for the evaluation of teacher effectiveness before deciding on the appropriate measure to employ. Scores from a value-added model may provide information about a teacher's contribution to student learning, but it would be less helpful in providing teachers with guidance on how to improve their performance.
- In considering the validity of various ways of measuring teacher effectiveness, keep in mind that the validity does not lie solely with the quality of the instrument or model but also with how well the instrument measures the construct and how the instrument is used in practice. Even a good classroom observation instrument in the hands of untrained evaluators may result in vastly different scores for similar teacher practices. And using a value-added model when large amounts of student data are missing may yield scores that fail to reflect the teacher's actual contribution to student learning.
- Seek other measures, or create appropriate measures, to capture important information about teachers' contributions that go beyond student achievement score gains. This may mean developing a measure that captures evidence of an individual teacher's leadership activities within the school, his or her collaboration with other teachers to strategize ways to help students who are at risk for failure, or participation in a study group to align the curriculum with state standards.
- Include education stakeholders in decisions about what is important to measure. Although a state legislature or task force may ultimately decide upon how teacher effectiveness will be measured, listening to the voices of teachers, principals, curriculum specialists, union representatives, parents, and students will help assure greater acceptance of the measurement system. Ultimately, this also will contribute to greater validity; the validity of a measure can be threatened by noncompliance or active resistance to the measure.
- Keep in mind that valid measurement may be costly. Ensuring that data is complete and accurate and that raters are trained and calibrated is essential in order to ensure the validity of the scores of the most commonly used measures of teacher effectiveness. Developing and validating new measures based on local priorities also will require adequate funding.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95–135.

Abbott, M., Walton, C., Tapia, Y., & Greenwood, C. R. (1999). Research to practice: A "blueprint" for closing the gap in local schools. *Exceptional Children*, *65,* 339–352.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Archibald, S. J. (2007). How well do standards-based teacher evaluation scores identify high-quality teachers? A multilevel, longitudinal analysis of one district. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 68*, 1235–1235.

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *Elementary School Journal, 107*(2), 199–219.

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal, 5*(1), 3–10.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Bauer, A. M., Johnson, L. J., & Sapona, R. H. (2004). Reflections on 20 years of preparing special education teachers. *Exceptionality, 12*(4), 239–246.

Benner, S. M., & Judge, S. L. (2000). Teacher preparation for inclusive settings: A talent development model. *Teacher Education Quarterly, 27*(3), 23–38.

Berliner, D. C. (1976). Impediments to the study of teacher effectiveness. *Journal of Teacher Education, 27*(1), 5–13.

Berry, B. (2004). *Making good on what matters most: A review of teaching at risk: A call to action (the report of The Teaching Commission).* Chapel Hill, NC: Southeast Center for Teaching Quality.

Betebenner, D. (2004). *An analysis of school district data using value-added methodology* (CSE Tech. Rep. No. 622). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles. Retrieved August 22, 2008, from http://www.cse.ucla.edu/products/reports/r622.pdf

Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology, 35*(1), 61–79.

Blank, R. K., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science. Results from the survey of enacted curriculum project.* Washington, DC: Council of Chief State School Officers.

Blanton, L. P., Blanton, W. E., & Cross, L. S. (1994). An exploratory study of how general and special education teachers think and make instructional decisions about students with special needs. *Teacher Education and Special Education, 17*(1), 62–74.

Blanton, L. P., Griffin, C. C., Winn, J. A., & Pugach, M. C. (Eds.). (1997). *Teacher education in transition*. Denver, CO: Love Publishing Company.

Blanton, L. P., Sindelar, P. T., Correa, V., Hardman, M., McDonnell, J., & Kuhel, K. (2003). *Conceptions of beginning teacher quality: Models for conducting research* (COPSSE Doc. No. RS-6). Gainesville: Center on Personnel Studies in Special Education (COPSSE), University of Florida. Retrieved August 22, 2008, from http://www.coe.ufl.edu/copsse/docs/RS-6/1/RS-6.pdf

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

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.

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. Retrieved August 22, 2008, from http://www.cse.ucla.edu/products/reports/r707.pdf

Bracey, G. W. (2004). Value-added assessment findings: Poor kids get poor teachers. *Phi Delta Kappan*, *86,* 331–333.

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest region* (Issues & Answers Report, REL 2007–No. 030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, REL Midwest. Retrieved August 22, 2008, from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2007030.pdf

Braun, H. I. (2005a). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved August 22, 2008, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Braun, H. I. (2005b). Value-added modeling: What does due diligence require? In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 19–39). Maple Grove, MN: JAM Press.

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

Burry, J. A., Chissom, B. S., & Shaw, D. G. (1990). *Validity and reliability of classroom observations: A paradox*. Paper presented at the National Council on Measurement in Education, Boston, MA.

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

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

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education, 29*(3), 347–362.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2004). Differentiated teacher effectiveness: Framing the concept. In *Assessing teacher effectiveness: Developing a differentiated model* (pp. 3–11). New York: Routledge.

Cavalluzzo, L. C. (2004). *Is National Board Certification an effective signal of teacher quality?* (Report No. (IPR) 11204). Alexandria, VA: The CNA Corporation. Retrieved August 22, 2008, from http://www.cna.org/documents/cavaluzzostudy.pdf

Cheng, Y. C., & Tsui, K. T. (1999). Multimodels of teacher effectiveness: Implications for research. *The Journal of Educational Research, 92*(3), 141–150.

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

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

Coggshall, J. (2007). *Communication framework for measuring teacher quality and effectiveness: Bringing coherence to the conversation*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved August 22, 2008, http://www.tqsource.org/publications/NCCTQCommFramework.pdf

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

Connecticut State Department of Education. (2007). *A guide to the BEST program for beginning teachers: 2007–2008*. Hartford, CT: Author.

Creemers, B. P. M., & Reezigt, G. J. (1996). School-level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement, 7*(3), 197–228.

Cruickshank, D. R., & Haefele, D. L. (1990). Research-based indicators: Is the glass half-full or half-empty? *Journal of Personnel Evaluation in Education, 4*(1), 33–39.

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

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1–22.

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

The Danielson Group. (n.d.). *Description> The framework for teaching* [Website]. Retrieved August 22, 2008, from www.danielsongroup.org/theframeteach.htm

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved August 22, 2008, from http://epaa.asu.edu/epaa/v8n1/

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" tell us? *Educational Researcher, 31*(9), 13–25.

Doherty, R. W., Hilberg, R. S., Epaloose, G., & Tharp, R. G. (2002). Standards Performance Continuum: Development and validation of a measure of effective pedagogy. *Journal of Educational Research, 96*(2), 78–89.

Dolezal, S. E., Welsh, L. M., Pressley, M., & Vincent, M. M. (2003). How nine third-grade teachers motivate student academic engagement. *Elementary School Journal, 103*(3), 239–267.

Dossett, D., & Munoz, M. (2003). *Classroom accountability: A value-added methodology*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Drury, D., & Doran, H. (2003). The value of value-added analysis. *NSBA Policy Research Brief, 3*(1),1–4.

Dynarski, M. (2008). Bringing answers to educators: Guiding principles for research syntheses. *Educational Researcher, 37*(1), 27–29.

Englert, C. S., Tarrant, K. L., & Mariage, T. V. (1992). Defining and redefining instructional practice in special education: Perspectives on good teaching. *Teacher Education and Special Education, 15*(2), 62–86.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record, 107*(1), 186–213.

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

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

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

Fuchs, D., & Fuchs, L. S. (1998). Researchers and teachers working together to adapt instruction for diverse learners. *Learning Disabilities Research of Practice, 13*, 126–137.

Gable, R. A. (1993). Unifying general and special education teacher preparation: Some cautions along the road to educational reform. *Preventing School Failure, 37*(2), 5–10.

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

Gentilucci, J. L. (2004). Improving school learning: The student perspective. *The Educational Forum, 68*(2), 133–143.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved August 22, 2008, from http://www.tqsource.org/publications/LinkBetweenTQandStudentOutcomes.pdf

Goldhaber, D., & Anthony, E. (2003). *Teacher quality and student achievement* (Urban Diversity Series No. 115). New York: ERIC Clearinghouse on Urban Education.

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* (Working Paper). Seattle, WA: Center on Reinventing Public Education. Retrieved August 22, 2008, from http://www.crpe.org/cs/crpe/download/csr_files/wp_crpe6_nbptsoutcomes_apr04.pdf

Good, T. L. (1996). Teaching effects and teacher evaluation. In J. P. Sikula, T. J. Buttery, & E. Guyton (Eds.), *Handbook of research on teacher education* (pp. 617–665). New York: Macmillan.

Good, T. L., Grouws, D. A., & Ebmeier, H. (1983). *Active mathematics teaching*. New York: Longman.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job: The Hamilton Project* (Discussion Paper 2006-01). Washington, DC: The Brookings Institution. Retrieved August 22, 2008, from http://www.brookings.edu/~/media/Files/rc/papers/2006/04education_gordon/200604hamilton_1.pdf

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

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*(2), 625–638.

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

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved August 22, 2008, from http://epaa.asu.edu/epaa/v8n41/

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: Using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology, 27*(5), 711–731.

Hardman, M. L., McDonnell, J., & Welch, M. (1998). *Special education in an era of school reform: Preparing special education teachers*. Washington, DC: Federal Resource Center for Special Education.

Harris, D. N., & Sass, T. (2007a). *The effects of NBPTS-certified teachers on student achievement* (CALDER Working Paper No. 4). Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research. Retrieved August 22, 2008, from http://www.caldercenter.org/PDF/1001060_NBPTS_Certified.pdf

Harris, D. N., & Sass, T. R. (2007b). *What makes for a good teacher and who can tell?* Paper presented at the 2007 summer workshop of the National Bureau of Economic Research. Cambridge, MA.

Haycock, K. (2004). The real value of teachers: If good teachers matter, why don't we act like it? *Thinking K–16, 8*(1), 1–2. Retrieved August 22, 2008, from http://www2.edtrust.org/NR/rdonlyres/5704CBA6-CE12-46D0-A852-D2E2B4638885/0/Spring04.pdf

Heistad, D. (1999). *Teachers who beat the odds: Value-added reading instruction in Minneapolis 2nd grade classrooms*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Heneman, H. G., Kimball, S., & Milanowski, A. (2006). *The teacher sense of efficacy scale: Validation evidence and behavioral prediction* (WCER Working Paper No. 2006-7). Madison, WI: Wisconsin Center for Education Research. Retrieved August 22, 2008, from http://www.wceruw.org/publications/workingPapers/Working_Paper_No_2006_07.pdf

Heneman, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay* (CPRE Policy Briefs No. RB-45). Philadelphia: Consortium for Policy Research in Education. Retrieved August 22, 2008, from http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf

Hershberg, T., Simon, V. A., & Lea-Kruger, B. (2004). Measuring what matters. *American School Board Journal, 191*(2), 27–31.

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

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

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

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

Interstate New Teacher Assessment and Support Consortium. (2001). *Model standards for licensing general and special education teachers of students with disabilities: A resource for state dialogue*. Washington, DC: Council of Chief State School Officers. Retrieved August 22, 2008, from http://www.ccsso.org/content/pdfs/SPEDStds.pdf

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (Faculty Research Working Papers Series No. RWP05-040). Cambridge, MA: Harvard University John F. Kennedy School of Government. Retrieved August 22, 2008, from http://ksgnotes1.harvard.edu/Research/wpaper.nsf/rwp/RWP05-040/$File/rwp_05_040_jacob.pdf

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.

Johnson, R. L., McDaniel, F., II, & Willeke, M. J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *American Journal of Evaluation, 21*(1), 65–80.

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., et al. (2006). *Overview of the instructional quality assessment* (CSE Tech. Rep. No. 671). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles. Retrieved August 22, 2008, from http://www.cse.ucla.edu/products/reports/r671.pdf

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

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21*(4), 345–363. Retrieved August 22, 2008, from http://ed-web3.educ.msu.edu/digitaladvisor/Research/Articles/approx.pdf

Kennedy, M. M. (2007). *Monitoring and assessing teacher quality*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Kupermintz, H. (2002). *Teacher effects as a measure of teacher effectiveness: Construct validity considerations in TVAAS (Tennessee Value-Added Assessment System)* (CSE Tech. Rep. No. 563). Boulder: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of Colorado.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.

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

Kyriakides, L., Demetriou, D., & Charalmbous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research, 48*(1), 1–20.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal, 104*(5), 409–426.

Le, V.-N., Stecher, B. M., Lockwood, J. R., Hamilton, L. S., Robyn, A., Williams, V. L., et al. (2006). *Improving mathematics and science education: A longitudinal investigation of*

the relationship between reform-oriented instruction and student achievement (Monograph No. MG-480-NSF). Arlington, VA: RAND Corporation. Retrieved August 22, 2008, from http://www.rand.org/pubs/monographs/2006/RAND_MG480.pdf

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255–270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125–150.

Lutz, S. L., Guthrie, J. T., & Davis, M. H. (2006). Scaffolding for engagement in elementary school reading instruction. *Journal of Educational Research, 100*(1), 3–20.

MacIsaac, D., Sawada, D., & Falconer, K. (2001). *Using the Reformed Teaching Observation Protocol (RTOP) as a catalyst for self-reflective change in secondary science teaching*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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.

Matsumura, L. C., & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice* (CSE Tech. Rep. No. 602). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles. Retrieved August 22, 2008, from http://www.cse.ucla.edu/products/reports/r602.pdf

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

Matsumura, L. C., Patthey-Chavez, G. G., Valdés, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103*(1), 3–25.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., et al. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment* (CSE Tech Rep. No. 681). Los Angeles: Center for the Study of Evaluation National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles. Retrieved August 22, 2008, from http://www.cse.ucla.edu/products/reports/r681.pdf

Matsumura, L. C., Slater, S. C., Wolf, M. K., Crosson, A., Levison, A., Peterson, M., et al. (2006). *Using the instructional quality assessment toolkit to investigate the quality of reading comprehension assignments and student work* (No. CSE Report 669). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles. Retrieved August 22, 2008, from http://www.cse.ucla.edu/products/reports/r669.pdf

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

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project.* Arlington, VA: RAND Corporation. Retrieved August 22, 2008, from http://www.rand.org/pubs/technical_reports/2007/RAND_TR506.sum.pdf

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2004). *The promise and peril of using value-added modeling to measure teacher effectiveness* (Research Brief No. RB-9050-EDU). Santa Monica, CA: RAND Corporation. Retrieved August 22, 2008, from http://www.rand.org/pubs/research_briefs/2005/RAND_RB9050.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation. Retrieved August 22, 2008, from http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101. Retrieved August 22, 2008, from http://www.rand.org/pubs/reprints/2005/RAND_RP1165.pdf

McColskey, W., Stronge, J. H., Ward, T. J., Tucker, P. D., Howard, B., Lewis, K., et al. (2005). *Teacher effectiveness, student achievement, and National Board Certified Teachers*. Arlington, VA: National Board for Professional Teaching Standards. Retrieved August 22, 2008, from http://www.nbpts.org/UserFiles/File/Teacher_Effectiveness_Student_Achievement_and_National_Board_Certified_Teachers_D_-_McColskey.pdf

McGreal, T. L. (1990). The use of rating scales in teacher evaluation: Concerns and recommendations. *Journal of Personnel Evaluation in Education, 4*(1), 41–58.

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

Mendro, R. L., Jordan, H. R., Gomez, E., Anderson, M. C., Bembry, K. L., & Schools, D. P. (1998). *An application of multiple linear regression in determining longitudinal teacher effectiveness.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

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

Milanowski, A. T., Kimball, S. M., & Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A. E. Schwartz, R. Rubenstein & J. Zabel (Eds.), *Measuring school performance and efficiency: Implications for practice and research* (pp. 137–162). Larchmont, NY: Eye on Education.

Millett, C. M., Stickler, L. M., Payne, D. G., & Dwyer, C. A. (2007). *A culture of evidence: Critical features of assessments for postsecondary student learning*. Princeton, NJ: Educational Testing Service. Retrieved August 22, 2008, from http://www.ets.org/Media/Resources_For/Higher_Education/pdf/4418_COEII.pdf

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

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research & Evaluation, 12*(1), 53–74.

Muijs, D., & Reynolds, D. (2003). Student background and teacher effects on achievement and attainment in mathematics: A longitudinal study. *Educational Research and Evaluation, 9*(3), 289–314.

Mullens, J. E. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for the teacher follow-up survey* (NCES Working Paper No. 95-15). Washington, DC: National Center for Education Statistics. Retrieved August 22, 2008, from http://nces.ed.gov/pubs95/9515.pdf

National Board for Professional Teaching Standards. (2002). *What teachers should know and be able to do*. Arlington, VA: Author. Retrieved August 22, 2008, from http://www.nbpts.org/UserFiles/File/what_teachers.pdf

National Board for Professional Teaching Standards. (2008). 2008 Guide to National Board Certification. Arlington, VA: Author. Retrieved August 22, 2008, from www.nbpts.org/userfiles/File/2008_Guide_Web_PDF_final.pdf

National Institute of Child Health and Human Development Early Child Care Research Network. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*(3), 305–323.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago: Consortium on Chicago School Research. Retrieved August 22, 2008, from http://ccsr.uchicago.edu/publications/p0a02.pdf

Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report*. Chicago: Consortium on Chicago School Research. Retrieved August 22, 2008, from http://ccsr.uchicago.edu/publications/p0f04.pdf

New Mexico 3-Tier Licensure Implementation Teacher Training Work Group. (2005). *A handbook on annual evaluation for New Mexico teachers holding a level 1 license*. Retrieved August 22, 2008, from http://www.teachnm.org/docs/EvalHandbookI_9.27.05.pdf

Mitchell, D. E., Scott, L. E., Hendrick, I. G., & Boyns, D. E. (1998). *The California beginning teacher support and assessment program: 1998 statewide evaluation study*. Riverside, CA: California Educational Research Cooperative.

Noell, G. H. (2005). *Assessing teacher preparation program effectiveness: A pilot examination of value added approaches* (Tech. Rep.). Baton Rouge, LA: Louisiana Board of Regents. Retrieved August 22, 2008, from http://asa.regents.state.la.us/TE/technical_report_200405.pdf

Noell, G. H. (2006). *Value added assessment of teacher preparation* (Annual Report): Baton Rouge, LA: Louisiana Board of Regents. Retrieved August 22, 2008, from http://asa.regents.state.la.us/TE/2005-technical-report.pdf

Noell, G. H., Porter, B. A., & Patt, R. M. (2007). *Value added assessment of teacher preparation in Louisiana 2004-2006*. Baton Rouge, LA: Louisiana Board of Regents. Retrieved August 22, 2008, from http://www.regents.state.la.us/Academic/TE/2007/VAA%20TPP%20Technical%20Report%2010-24-2007.pdf

Nunnally, J. C. (Ed.). (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4–32.

Office of Postsecondary Education & Office of Policy Planning and Innovation. (2003). *Meeting the highly qualified teachers challenge: The Secretary's second annual report on teacher quality*. Washington, DC: U.S. Department of Education. Retrieved August 22, 2008, from http://www.ed.gov/about/reports/annual/teachprep/2003title-ii-report.pdf

Office of Postsecondary Education. (2005). *A highly qualified teacher in every classroom: The Secretary's fourth annual report on teacher quality*. Washington, DC: U.S. Department of Education. Retrieved August 22, 2008, http://www.ed.gov/about/reports/annual/teachprep/2005Title2-Report.pdf

Office of Superintendent of Public Instruction. (n.d.). *Certification* [Website]. Retrieved August 22, 2008, from http://www.k12.wa.us/certification

Painter, B. (2001). Using teaching portfolios. *Educational Leadership, 58*(5), 31–34.

Patthey-Chavez, G. G., Matsumura, L. C., & Valdés, R. (2004). Investigating the process approach to writing instruction in urban middle schools. *Journal of Adolescent & Adult Literacy, 47*(6), 462–477.

Pecheone, R. L., & Stansbury, K. (1996). Connecting teacher assessment and school reform. *Elementary School Journal, 97*(2), 163–177.

Pecheone, R. L., Pigg, M. J., Chung, R. R., & Souviney, R. J. (2005). Performance assessment and electronic portfolios: Their effect on teacher learning and education. *The Clearing House 78*(4), 164–176.

Perry, K. E., Donohue, K. M., & Weinstein, R. S. (2007). Teaching practices and the promotion of achievement and adjustment in first grade. *Journal of School Psychology, 45*(3), 269–292.

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

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). *Classroom assessment scoring system manual, middle/secondary version*. Charlottesville, VA University of Virginia.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2006). *Classroom assessment scoring system: Preschool (pre-k) version* (Vol. Manual). Charlottesville, VA: Center for Advanced Study of Teaching and Learning.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). *Classroom assessment scoring system* (Observation Protocol Manual). Baltimore: Paul H. Brookes.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal, 102*(3), 225–238.

Piburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP) reference manual* (ACET Tech. Rep. No. IN00-3). Tempe: Arizona Collaborative for Excellence in the Preparation of Teachers, Arizona State University. Retrieved August 22, 2008, from http://cresmet.asu.edu/prods/rtop_files/RTOP_Reference_Manual.pdf

Porter, A. C., Kirst, M. W., Osthoff, E. J., & Smithson, J. L. (1993). *Reform up close: An analysis of high school mathematics and science classrooms. Final report to the National Science Foundation*. Madison, WI: Wisconsin Center for Education Research.

Pugach, M. C. (2005). Research on preparing teachers to work with students with disabilities. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 549–590). Mahwah, NJ: Lawrence Erlbaum Associates.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington, DC: The Economic Policy Institute.

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

Rivers-Sanders, J. C. (1999). *The impact of teacher effect on student math competency achievement.* Unpublished doctoral dissertation, University of Tennessee, Knoxville, TN.

Rivkin, S. G., & Ishii, J. (2008). *Impediments to the estimation of teacher value-added*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Ross, S., Stringfield, S., Sanders, W. L., & Wright, S. P. (2003). Inside systemic elementary school reform: Teacher effects and teacher mobility. *School Effectiveness and School Improvement, 14*(1), 73–110.

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

Rothstein, J. (2008b). *Student sorting and bias in value added estimation: Selection on observables and unobservables*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI. Retrieved August 22, 2008, from http://www.wceruw.org/news/events/VAM%20Conference%20Final%20Papers/Student Sorting&Bias_JRothstein.pdf

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study elementary school mathematics: A close look at curriculum and teaching in the early grades. *Elementary School Journal, 105*(1), 103–127.

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329–339.

Sanders, W. L., Ashton, J. J., & Wright, S. P. (2005). *Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress*. Arlington, VA: National Board for Professional Teaching Standards. Retrieved August 22, 2008, from http://www.nbpts.org/UserFiles/File/SAS_final_NBPTS_report_D_-_Sanders.pdf

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (No. R11-0435-02-001-97). Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., & Rivers, J. C. (2006). The value of evidence in reforming teacher education. *Teachers for a New Era Quarterly Newsletter,* 2(3). Retrieved August 22, 2008, from http://www.teachersforanewera.org/newsletters/newsletters/TNE%20Newsletter%20V2-N3.doc

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., et al. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 1–20.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review, 23*, 411–430.

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

Schalock, H. D., Schalock, M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon State College. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 15–45). Thousand Oaks, CA: Corwin Press.

Schlusmans, K. (1978). *What is an effective teacher?* Paper presented at the Conference of the International Association for Educational Assessment, Baden, Austria.

Schweinle, A., Meyer, D. K., & Turner, J. C. (2006). Striking the right balance: Students' motivation and affect in elementary mathematics. *Journal of Educational Research, 99*(5), 271–293.

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

Shymansky, J. A., Yore, L. D., & Anderson, J. O. (2004). Impact of a school district's science reform effort on the achievement and attitudes of third- and fourth-grade students. *Journal of Research in Science Teaching, 41*(8), 771–790.

Smylie, M. A., & Wenzel, S. A. (2006). *Promoting instructional improvement: A strategic human resource management perspective*. Chicago: Consortium on Chicago School Research. Retrieved August 22, 2008, from http://ccsr.uchicago.edu/publications/p84.pdf

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

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

Stone, J. E. (2002). *The value-added achievement gains of NBPTS-certified teachers in Tennessee: A brief report*. Arlington, VA: National Board for Professional Teaching Standards.

Taylor, D. M. (2006). Refining learned repertoire for percussion instruments in an elementary setting. *Journal of Research in Music Education, 54*(3), 231–243.

The Teaching Commission. (2004). *Teaching at risk: A call to action*. New York: Author. Retrieved August 22, 2008, from http://www.ecs.org/html/offsite.asp?document=http%3A%2F%2Fftp%2Eets%2Eorg%2Fpub%2Fcorp%2Fttcreport%2Epdf

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36.

Thum, Y. M. (2003). Measuring progress toward a goal estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research, 32*(2), 153–207.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Tytler, R., Waldrip, B., & Griffiths, M. (2004). Windows into practice: Constructing effective science teaching and learning in a school change initiative. Research report. *International Journal of Science Education, 26*(2), 171–194.

Valli, L., Croninger, R., Alexander, P., Chambliss, M., Graeber, A., & Price, J. (2004). *A study of high-quality teaching: Mathematics and reading.* Paper presented at the American Educational Research Association, San Diego, CA.

Valli, L., Croninger, R. G., & Walters, K. (2007). Who (else) is the teacher? Cautionary notes on teacher accountability systems. *American Journal of Education, 113*(4), 635–662.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board Certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved August 22, 2008, from http://epaa.asu.edu/epaa/v12n46/v12n46.pdf

Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching, 36*(10), 1110–1126.

Watson, A., & De Geest, E. (2005). Principled teaching for deep progress: Improving mathematical learning beyond methods and materials. *Educational Studies in Mathematics, 58*(2), 209–234.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89–122.

Weber, J. R. (1987). *Teacher evaluation as a strategy for improving instruction. Synthesis of literature*. Elmhurst, IL: North Central Regional Educational Lab. Retrieved August 22, 2008, from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/fd/7d.pdf

Webster, W. J. (2005). The Dallas school-level accountability model: The marriage of status and value-added approaches. In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 233–271). Maple Grove, MN: JAM Press.

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

Wilson, S. M., & Floden, R. (2003). *Creating effective teachers: Concise answers for hard questions. An addendum to the report "Teacher preparation research: current knowledge, gaps, and recommendations."* Washington, DC: AACTE Publications.

Wisconsin Department of Public Instruction. (2008). Summary of the Wisconsin Master Educator Assessment Process (WMEAP) and the Master Educator License [Website]. Retrieved August 22, 2008, from http://dpi.state.wi.us/tepdl/wmeapsumm.html

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

Wright, S. P. (2004). *Advantages of a multivariate longitudinal approach to educational value-added assessment without imputation*. Paper presented at the National Evaluation Institute, Colorado Springs, CO. Retrieved August 22, 2008, from http://www.wmich.edu/evalctr/create/2004/Wright-NEI04.pdf

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.

Wright, S. P., & Sanders, W. L. (2008). *Decomposition of estimates in a layered value-added assessment model*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI. Retrieved August 22, 2008, from http://www.wcer.wisc.edu/news/events/WrightSanders_Decomposition.pdf

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

Appendix A. Tools for Measuring Teacher Effectiveness

This table provides an overview of the different methods of measuring teacher effectiveness discussed in this paper. Selected articles are meant to be representative examples of the various methodologies and instrumentations currently being utilized in each category, to provide readers with an idea of how teachers are measured under different circumstances and in multiple contexts.

Author (Year)	Measure	Description
Observational Protocols		
Blunk (2007)	Quality of Mathematics in Instruction (QMI)	Discussed the development and technical properties of the QMI protocol, examining reliability of scores, interrater reliability, and validity in terms of relationships between the instrument and other measures of mathematical knowledge for teaching.
Hamre and Pianta (2005)	Classroom Observation System (COS)	Examined the quality of instructional and emotional support in Grade 1 classrooms using the COS-1, and analyzed the ways in which teacher support moderates children's risk of school failure.
Hoffman, Sailors, Duffy, and Beretvas (2004)	TEX-IN3 Observation System	Described development, reliability, and validity of the TEX-IN3 protocol for assessing the overall effectiveness of the classroom literacy environment, including its relationship to student literacy attainment.
Kimball, White, Milanowski, and Borman (2004)	*Framework for Teaching*	Assessed the relationship between a teacher evaluation system based on Danielson's (1996) *Framework for Teaching* and student achievement gains on standardized tests in reading and mathematics in Washoe County school district, Nevada.
La Paro, Pianta, and Stuhlman (2004)	Classroom Assessment Scoring System (CLASS)	Described the development of the CLASS for observing and assessing emotional and instructional quality in Grades PK–3 classrooms, and investigated its validity and reliability in a prekindergarten setting.
Milanowski (2004)	*Framework for Teaching*	Investigated the relationship between teacher evaluation scores based on Danielson's (1996) *Framework for Teaching* and student achievement gains on district and state tests in reading, mathematics, and science in Cincinnati public schools.

Author (Year)	Measure	Description
Observational Protocols		
Piburn and Sawada (2000)	Reformed Teaching Observation Protocol (RTOP)	Presented development, reliability, and validity information on the RTOP, designed to measure reformed teaching in mathematics and science. Reported associations with student learning at middle, high school, and college levels.
Principal Evaluations		
Brandt, Mathers, Oliva, Brown-Sims, and Hess (2007)	District teacher evaluation systems	Investigated existing teacher evaluation policies in a diverse sample of Midwestern school districts and reported on the requirements, specifications, and guidance provided to schools by these policies.
Harris and Sass (2007)	Researcher-created teacher rating scale	Compared subjective principal ratings of teachers in Grades K–12 to value-added student achievement in mathematics and reading to assess principals' ability to evaluate teacher effectiveness.
Jacob and Lefgren (2008)	Researcher-created teacher rating scale	Compared subjective principal evaluation scores of elementary teachers to value-added mathematics and reading student achievement scores, teacher experience, and teacher education.
Medley and Coker (1987)	Researcher-created teacher ranking form	Examined accuracy of principals' judgments of teacher performance and effectiveness compared to value-added student achievement measures in mathematics and reading.
Instructional Artifacts		
Borko, Stecher, Alonzo, Moncure, and McClam (2005)	Scoop Notebook	Examined reliability, validity, and feasibility of the Scoop Notebook, a protocol for gathering and rating quality of classroom artifacts, and reported pilot study results from middle school mathematics and science teachers.
Matsumura, Patthey-Chavez, Valdés, and Garnier (2002)	Quality of writing assignments and teachers' feedback	Examined the relationships between quality of teachers' writing assignments, written feedback, and student revisions in Grade 3 classrooms, relating these factors to students' improvements in writing.
Matsumura, Slater, Junker, Peterson, Boston, Steele, et al. (2006)	Instructional Quality Assessment (IQA)	Described development and reliability of IQA for rating teacher assignments and student work, and examined its relationship to student achievement in middle school language arts and mathematics classes.

Author (Year)	Measure	Description
Instructional Artifacts		
Newmann, Bryk, and Nagaoka (2001)	Intellectual Demand Assignment Protocol (IDAP)	Used IDAP to rate the intellectual demand and authenticity of teacher assignments and student work, and related scores to achievement gains in mathematics and writing on standardized tests in Grades 3, 6, and 8.
Schalock, Schalock, and Girod (1997)	Teacher Work Sample Methodology (TWSM)	Described the TWSM as used at Western Oregon State College.
Portfolios		
Hakel, Koenig, and Elliot (2008)	NBPTS Certification	Reviewed research and presented an evaluation of the impacts of National Board Certification, investigating its effects on teachers, student achievement, and teacher quality in the education field as a whole.
Koretz, Stecher, Klein, and McCaffrey (1994)	Vermont teacher portfolio assessment	Reported on the implementation of the Vermont teacher assessment program, focusing on the value of the program to teachers and administrators and discussing difficulties in establishing reliability and validity of the portfolio system.
Teacher Self-Report Measures		
Blank, Porter, and Smithson (2001)	Surveys of Enacted Curriculum (SEC)	Reported on the two-year project to develop the SEC in mathematics and science, discussing central research findings, advances in the survey measures, and important applications of the surveys and data tools.
Camburn and Barnes (2004)	Teaching log	Examined validity of an English language arts (ELA) teaching log for measuring teacher practice.
Flowers and Hancock (2003)	Interview protocol	Described and reported validity and reliability data for an interview protocol developed to evaluate teachers' ability to accurately assess and modify instruction for improving student learning.
Mayer (1999)	Researcher-created teacher survey mirroring questions from the NCES TFS (1992)	Examined reliability and validity of self-reported teacher survey data on instructional practices in mathematics compared to observational measures.
Student Ratings		
Kyriakides (2005)	Researcher-created student survey	Examined validity of student-rated teacher behaviors using student achievement gains.
Peterson, Wahlquist, and Bone (2000)	District-created student surveys	Examined validity and reliability of student ratings of teacher performance.

Author (Year)	Measure	Description
Student Ratings		
Worrell and Kuterbach (2001)	Researcher-created student survey using items from the Teacher Behavior Inventory (TBI)	Examined reliability and validity of student ratings of low-inference teaching behaviors in a sample of "academically talented" students.
Value-Added Models		
Dossett and Munoz (2003)	Longitudinal value-added model	Described and compared value-added models and proposed a longitudinal value-added model to examine how student and teacher characteristics affect student achievement gains.
Heistad (1999)	Minneapolis value-added system	Examined teacher effects in reading using the Minneapolis value-added system.
Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007)	Multivariate Bayesian formulation of a longitudinal model	Introduced a multivariate Bayesian formulation of a longitudinal model developed to parameterize the long-term effects of past teachers on future years.
Mendro, Jordan, Gomez, Anderson, Bembry, and Schools (1998)	Hierarchical linear modeling (HLM)	Described the use of an HLM model to analyze longitudinal teacher effectiveness data.
Sanders and Horn (1998)	Tennessee Value-Added Assessment System (TVAAS)	Presented findings on teacher effects as measured by TVAAS.

Appendix B. Technical Considerations in Assessing Teacher Effectiveness

This table provides representative examples of articles that discuss the technical elements of different measures, including reliability and validity analyses, methodological issues, and considerations involved in using each.

Author (Year)	Description
Comparative Methods	
Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, and Guiton (1995)	Investigated the validity of large-scale teacher surveys for gathering information about school curriculum, comparing survey responses with analysis of benchmark data including course textbooks, assignments, exams, and teaching logs.
Holtzapple (2003)	Examined the validity of teacher evaluation ratings (adapted from the *Framework for Teaching*) compared to comprehensive principal evaluations based on observations and portfolios and to value-added student achievement gains.
Kennedy (1999)	Compared different measures used to collect information about student learning, including standardized tests, classroom observations, teacher logs, responses to vignettes, questionnaires, and interviews, through review of the empirical literature.
Le, Stecher, Lockwood, Hamilton, Robyn, Williams, et al. (2006)	Reported on the Mosaic II study, which examines relationships between reform-oriented instructional practices and student outcomes in mathematics and science, using multiple measures such as teacher surveys, daily logs, structured vignettes, classroom observations, and interviews.
McColskey, Stronge, Ward, Tucker, Howard, Lewis, et al. (2005)	Examined relationship between NBPTS certification and teachers' value-added effectiveness, then compared the teaching practices of NBCTs and other teachers identified as effective, using observations, surveys, instructional artifacts, and interviews.
Shavelson, Webb, and Burstein (1986)	Reviewed literature on the measurement of teaching, focusing on measurement of teacher effectiveness, classroom processes, and teachers' cognitive processes.
Wilkerson, Manatt, Rogers, and Maughan (2000)	Compared ratings of teacher performance from principal evaluations, student ratings, and teacher self-evaluations, and examined which are most strongly related to student achievement gains.

Author (Year)	Description
Observation-Based Evaluation	
Burry, Chissom, and Shaw (1990)	Presented features of valid classroom observation procedures, a five-step measurement schemata, and a systematic classroom observation procedure designed to reduce measurement error.
McGreal (1990)	Discussed the different types of observational rating scales that are used in teacher evaluation, highlighting the concerns and recommendations for the use of each type.
Pianta, La Paro, et al. (2007)	Provided information on the psychometric properties of the Classroom Assessment Scoring System (CLASS), which is derived from the Classroom Observation System (COS), including reliability and validity findings in Grades PK–6.
Schacter and Thum (2004)	Described the design, development, and validation of a classroom observation system to judge teacher performance based on predetermined standards, and examined the relationship of ratings to student achievement gains in mathematics and English language arts (ELA).
Instructional Artifacts	
Junker, Weisburg, Matsumura, Crosson, Wolf, Levison, et al. (2006)	Presented an overview of the process for building and piloting the Instructional Quality Assessment (IQA), a formal toolkit for rating instructional quality based on classroom observation and student assignments in reading and mathematics.
Matsumura, Garnier, Pascal, and Valdés (2002)	Examined the technical quality of a measure rating the quality of teacher assignments in language arts (a precursor to IQA), focusing on reliability and stability of scores and their relationship to student achievement.
Newmann, Lopez, and Bryk (1998)	Described a protocol to evaluate the authenticity and intellectual demand of teacher assignments in writing and mathematics, examining how often students encountered challenging assignments and the connection between level of demand and quality of student work.
Portfolios	
Johnson, McDaniel, and Willeke (2000)	Investigated the interrater reliability of a small-scale family portfolio assessment, examining reliability differences between individual analytic ratings, the composite analytic rating, and an overall holistic rating.
Tucker, Stronge, Gareis, and Beers (2003)	Examined the use of portfolios in teacher evaluation for both accountability and professional development purposes, discussing their validity, their contribution to the evaluation process, and teacher and administrator perceptions of their use.
Student Ratings	
Follman (1992)	Presented an empirical literature review on using public secondary school students' ratings to evaluate teachers, exploring reliability and validity findings and presenting conclusions and recommendations.

Author (Year)	Description
Student Ratings	
Follman (1995)	Presented an empirical literature review on using public elementary school students' ratings to evaluate teachers, exploring reliability and validity findings and presenting conclusions and recommendations.
Value-Added Models	
Amrein-Beardsley (2008)	Discussed methodological issues with education valued-added assessment system, noting in particular that the models have not undergone external review and validity studies have not been done.
Kupermintz (2003)	Examined validity of teacher evaluation measures produced by Tennessee Value-Added Assessment System (TVAAS).
Martineau (2006)	Demonstrated that the even vertically scaled assessments used for value-added assessment of teachers result in "remarkable distortions" in teacher estimates.
McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004)	Used simulated data to illustrate problems with the general multivariate, longitudinal mixed-model of value-added assessment.
Raudenbush (2004)	Examined what kinds of effects can and cannot reasonably be estimated using value-added analyses.
Rivkin and Ishii (2008)	Discussed difficulty of using value-added models given nonrandom assignment of student to teachers, but suggested that some methods may compensate for some problems. However, none appear able to resolve all issues.
Rothstein (2008b)	Illustrated how biases can impact results from using value-added measures in the presence of nonrandom assignment of students to teachers. The author found that "even well-controlled models may be substantially biased" (p. 1).
Rothstein (2008a)	Argued that widely used value-added models result in inaccurate estimates of teacher quality due to nonrandom assignment of students to teachers. Stated that value-added models "need further development and validation before they can support causal interpretations or policy applications."
Tekwe, Carter, Ma, Algina, Lucas, Roth, et al. (2004)	Examined four examples of assessment systems that use student achievement as a measure of teacher effectiveness.
Wright (2004)	Compared several different statistical approaches to value-added modeling to demonstrate the benefits of using a more complex, multivariate longitudinal approach to calculating value-added measures.
Wright and Sanders (2008)	Addressed criticisms of "complexity and lack of transparency" by comparing the Sanders value-added model with three other models to illustrate how the model "constructs teacher effects from student data" (p. 1).

Appendix C. Outcomes of Interest in Teacher Evaluation

This table provides representative examples of the many different ways to measure and conceptualize student outcomes, including the specific instruments reported in each study.

Author (Year)	Measure	Description
Student Achievement: Standardized Tests		
Hamre and Pianta (2005)	Woodcock-Johnson Psycho-educational Battery-Revised (WJ-R)	Examined how teachers' instructional and emotional support toward students moderates their risk of school failure.
Matsumura, Slater, Junker, Peterson, Boston, Steele, et al., (2006)	Stanford Achievement Test, 10th ed. (SAT-10)	Described development of IQA ratings of teacher assignments and student work, and reported pilot data showing relationship of IQA to student achievement gains in mathematics and reading.
Newmann, Bryk, Nagaoka (2001)	Iowa Test of Basic Skills (ITBS); Illinois State standardized tests	Examined relationship between intellectual demand of classroom assignments and value-added measures of student achievement in mathematics and reading.
Noell (2006)	ITBS; Louisiana State LEAP-21 test	Examined effect of teacher preparation on value-added student achievement measures in English language arts (ELA), mathematics, science, and social studies.
Rivkin, Hanushek, and Kain (2005)	Texas Assessment of Academic Skills (TAAS)	Examined how teacher education and experience relate to value-added student achievement scores in reading and mathematics.
Rockoff (2004)	Comprehensive Test of Basic Skills (CTBS); TerraNova CTBS; Metropolitan Achievement Test (MAT)	Examined how teacher experience relates to effectiveness, through value-added measures of student achievement in mathematics and reading.
Shymansky, Yore, and Anderson (2004)	Third International Mathematics and Science Study (TIMSS)	Investigated teacher implementation of a science professional development program and its effect on student science achievement and attitudes about science.
Thum (2003)	Stanford Achievement Test, 9th ed. (SAT-9)	Examined teacher effects on student achievement in language arts, mathematics, and reading as measured by value-added models, partialling out student and classroom covariates.
Student Achievement: District-Created Tests		
Jacob and Lefgren (2008)	District-created "core" exams	Examined principals' ability to distinguish between more and less effective teachers, comparing principal evaluation scores with value-added measures of student achievement in mathematics and reading, teacher experience, and education.

Author (Year)	Measure	Description
Student Achievement: District-Created Tests		
Patthey-Chavez, Matsumura, and Valdés (2004)	Mechanics, Usage, Grammar, and Spelling (MUGS) ratings created by district and teachers' union	Explored the nature of teacher feedback on reading and writing assignments to urban middle school students, and examined its relationship to student writing improvement.
Perry, Donohue, and Weinstein (2007)	District-created tests geared to California State academic standards	Investigated effect of socially and cognitively supportive teaching practices on students' reading and mathematics achievement and on behavioral and socioemotional adjustment.
Wilkerson, Manatt, Rogers, and Maughan (2000)	Criterion-referenced tests created collaboratively by district and researchers	Examined relationship between teacher performance ratings given by principals, students, and teacher self-ratings to student achievement in reading, ELA, and mathematics.
Student Engagement		
Dolezal, Welsh, Pressley, and Vincent, (2003)	Percentage of time students were on task, cognitive demand of teachers' assignments/activities	Examined instructional practices in classrooms with different levels of student engagement and identified practices used by highly engaging teachers.
Lutz, Guthrie, and Davis (2006)	Researcher-developed rubric for rating student engagement	Examined how teachers scaffold student engagement and how engagement relates to science-literacy achievement.
National Institute of Child Health and Human Development Early Child Care Research Network (2005)	Percentage of time students were on task, quality of classroom activities	Examined nature and quality of classroom climate, including classroom structure, social climate, and quality of activities.
Student Behavior		
Hamre and Pianta (2001)	Teacher reports of student work habits, number of disciplinary infractions, student suspension	Investigated whether kindergarten teachers' perceptions of their relationships with students predict later student academic and behavioral outcomes.
Hamre and Pianta, (2005)	Student-Teacher Relationship Scale (STRS)	Examined how teachers' instructional and emotional support toward students moderates their risk of school failure.
Perry, Donohue, and Weinstein (2007)	Pupil Behavior Rating Scale (PBRS)	Investigated effect of socially and cognitively supportive teaching practices on students' reading and mathematics achievement and behavioral and socioemotional adjustment.

Author (Year)	Measure	Description
Social/Emotional Outcomes		
Birch and Ladd (1997)	Teacher Rating Scale of School Adjustment (TRSSA), Loneliness and Social Dissatisfaction Questionnaire (LSDQ), School Liking and Avoidance Scale (SLAS)	Examined relationship between three dimensions of the teacher-child relationship (closeness, dependency, and conflict) and aspects of children's early adjustment to school.
Schweinle, Meyer, and Turner (2006)	Experience Sampling Form (ESF)	Investigated the relationship between student-reported levels of motivation and affect, and examines associated classroom practices.
Student Attitudes		
Shymansky, Yore, and Anderson (2004)	Research-created Likert-type questionnaire assessing attitudes toward science and science careers	Investigated teacher implementation of a science professional development program and its effect on student science achievement and attitudes about science.

Appendix D. Comprehensive List of Studies With Summaries

This table includes the articles that were examined in this research synthesis.

Author (Year)	Category	Summary
Borko, Stecher, Alonzo, Moncure, and McClam (2005)	artifacts	Examined the reliability, validity, and feasibility of the Scoop Notebook, which measures the use of reform-oriented teaching practices by analyzing classroom artifacts. Pilot results found ratings to be reasonably consistent with observational measures.
Borko, Stecher, and Kuffner (2007)	artifacts	Included final data collection and scoring tools for the Scoop Notebook to rate reform-oriented teaching practices using classroom artifacts. Provided descriptions of procedures, rating guides, administration details, and potential uses.
Clare and Aschbacher (2001)	artifacts	Examined reliability and validity of teacher assignment ratings on the literacy Instructional Quality Assessment (IQA). Found quality of teachers' assignments was associated with quality of student writing.
Junker, Weisburg, Matsumura, Crosson, Wolf, Levison, et al., (2006)	artifacts	Presented an overview of the development and validation of IQA for rating instructional quality based on classroom observation and student assignments in reading and mathematics. Reported on a large pilot study of IQA and discusses future work and directions.
Matsumura, Garnier, Pascal, and Valdés (2002)	artifacts	Examined the technical quality of a measure for rating the quality of teacher assignments. Found that use of high-quality assignments was related to student achievement on SAT-9 language arts sections.
Matsumura, Patthey-Chavez, Valdés, and Garnier (2002)	artifacts	Investigated the quality of teacher writing assignments and written feedback in Grade 3. Feedback helped improve students' writing mechanics, but overall quality of assignments and feedback was low and students showed little improvement in writing content or organization.
Matsumura, Slater, Junker, Peterson, Boston, Steele, et al. (2006)	artifacts	Described the development of IQA ratings of teacher assignments and student work. Pilot data showed IQA ratings predicted student achievement gains on the SAT-10 in mathematics and reading.

Author (Year)	Category	Summary
Matsumura, Slater, Wolf, Crosson, Levison, Peterson, et al. (2006)	artifacts	Described the theoretical framework behind IQA and presented pilot results in reading comprehension. Discussed problems with interrater reliability and stability of scores in this small sample, and made suggestions for future data collection.
Newmann, Lopez, and Bryk (1998)	artifacts	Examined assignments and work samples from Grades 3, 6, and 8 writing and mathematics classes and rated them for authenticity and intellectual demand. Found that quality of assigned work was generally low, but high-quality work related to improved student performance.
Newmann, Bryk, and Nagaoka (2001)	artifacts	Examined the relationship between teachers' use of intellectually demanding assignments as measured by IDAP rubrics and student achievement. Assignment quality was predictive of ITBS and state proficiency test scores in language arts and mathematics.
Patthey-Chavez, Matsumura, and Valdés (2004)	artifacts	Explored the nature of teacher feedback on reading and writing assignments to urban middle school students. Students responded to surface-level teacher feedback, but overall received little feedback and showed little writing improvement.
Archibald (2007)	observation	Examined standards-based teacher evaluation scores based on Danielson's (1996) *Framework for Teaching* at a site in Nevada, while controlling for student-, teacher- and school-level characteristics. Found that scores were a positive predictor of student achievement, but school-level characteristics were also a significant factor.
Baker, Gersten, Haager, and Dingle (2006)	observation	Examined the validity of the English Language Learner Classroom Observation Instrument (ELLOI). Findings showed that each subscale correlated moderately with student achievement gains in reading.
Birch and Ladd (1997)	observation	Examined ratings of teacher-student relationship (closeness, conflict, and dependency). Found these dimensions were related to student visual and language achievement, social and affective outcomes, and school engagement.

Author (Year)	Category	Summary
Blunk (2007)	observation	Discussed the development of the QMI protocol and investigated its reliability and validity. Found that four to five lessons are necessary for adequate reliability, and QMI scores significantly correlated with other measures of mathematical knowledge for teaching.
Burry, Chissom, and Shaw (1990)	observation	Described features of valid classroom observation and included a systematic observation procedure purported to reduce measurement error in conducting classroom observations. Discussed psychometric issues and evaluated the conditions under which classroom observations are valid.
Doherty, Hilberg, Epaloose, and Tharp (2002)	observation	Investigated the reliability and validity of the Standards Performance Continuum (SPC) observation instrument. Results from small pilot studies showed high interrater reliability, reasonable correlations with other observational measures, and modest predictive relationships with student achievement in ELA.
Dolezal, Welsh, Pressley, and Vincent (2003)	observation	Examined instructional practices in Grade 3 classrooms with different levels of student engagement. Discussed several mechanisms used in concert by highly engaging teachers, which seemed to increase student motivation.
Gallagher (2004)	observation	Examined the relationship between scores on a teacher evaluation based on Danielson's (1996) *Framework for Teaching* and student achievement in literacy, mathematics, and English language arts (ELA) at a site in Los Angeles. Found that composite and literacy scores were positively related to gains on the SAT-9.
Good, Grouws, and Ebmeier (1983)	observation	Described four studies investigating classroom processes in mathematics and the impact of an intervention on instructional practices and student achievement. Used observational data to identify and describe practices associated with higher and lower achievement, and examined how the intervention produced student achievement gains.

Author (Year)	Category	Summary
Hamre and Pianta (2005)	observation	Examined teachers' level of instructional and emotional support in Grade 1 classrooms using the COS-1. High support was associated with improved achievement for students at high risk of school failure.
Harachi, Abbott, Catalano, Haggerty, and Fleming (1999)	observation	Explored teaching practices related to a professional development intervention using the RHC Classroom Observation System. Revealed positive student involvement and proactive classroom management strategies related to students' social competency and school commitment and overall positive instructional practices related to decreases in student antisocial behavior.
Heneman, Kimball, and Milanowski (2006)	observation	Examined the relationship between teacher self-efficacy, teacher performance scores based on Danielson's (1996) *Framework for Teaching*, and student achievement, at a site in Nevada. Self-efficacy correlated with performance, but performance did not correlate with student achievement.
Heneman, Milanowski, Kimball, and Odden (2006)	observation	Examined the validity, acceptability, and usability of teacher evaluation measures adapted from Danielson's (1996) *Framework for Teaching* in four different sites. Evaluation scores related to student achievement gains, especially when schools used trained and multiple observers.
Hoffman, Sailors, Duffy, and Beretvas (2004)	observation	Described the development, validity, and reliability of the TEX-IN3 Observation System for evaluating the classroom literacy environment. Found that each aspect of TEX-IN3 had predicted relationships with student achievement.
Holtzapple (2003)	observation	Examined the validity of a teacher evaluation based on Danielson's (1996) *Framework for Teaching* at a site in Cincinnati. Evaluation scores positively related to student achievement on Ohio state proficiency tests for reading, mathematics, science, and social studies in Grades 3–8.

Author (Year)	Category	Summary
Howes, Burchinal, Pianta, Bryant, Early, Clifford, et al. (2008)	observation	Explored the dimensions of classroom quality in state-funded prekindergarten programs using the COS. Found that quality of instruction and closeness of teacher-child relationships related to academic outcomes.
Kimball, White, Milanoski, and Borman (2004)	observation	Analyzed the relationship between teachers' evaluation scores adapted from Danielson's (1996) *Framework for Teaching* and student achievement at a site in Nevada. Evaluation scores correlated slightly with student gains on CTBS/Terra Nova and Nevada proficiency tests.
La Paro, Pianta, and Stuhlman (2004)	observation	Explored the development, reliability, and validity of CLASS for observing and assessing emotional and instructional quality in Grades PK–3 classrooms. Ratings revealed generally positive classroom environments, and scales were related to a similar instrument.
Lutz, Guthrie, and Davis (2006)	observation	Examined how teachers scaffold student engagement. High literacy achievement was associated with moderate to high engagement in learning and high complexity of literacy tasks. Scaffolding appeared to foster student engagement with complex tasks.
MacIsaac, Sawada, and Falconer (2001)	observation	Investigated the feasibility of using the RTOP to promote teacher reflection and understanding of reform teaching. Found that the RTOP had high validity and credibility for this purpose and that teachers were open to using the RTOP in this way.
McGreal (1990)	observation	Evaluated the different types of observational rating scales that are used in teacher evaluation. Discussed validity concerns, measurement issues, and recommendations for using each type of scale.
Milanowski (2004)	observation	Analyzed the relationship between teacher evaluation scores based on Danielson's (1996) *Framework for Teaching* and student achievement at a site in Cincinnati. Evaluation scores related to student CTBS and state proficiency test scores in mathematics and reading.

Author (Year)	Category	Summary
Muijs and Reynolds (2003)	observation	Examined the relationship between student social background, classroom social context, classroom organization, and teacher behavior. Teacher behavior accounted for a large portion of between-classroom and between-school variance in student achievement, while student background characteristics accounted for little.
National Institute of Child Health and Human Development Early Child Care Research Network (2005)	observation	Examined the nature and quality of classroom climates using the COS-3. Found that most Grade 3 classrooms had positive social climates but low instructional quality, with one-third of activities rated as unproductive.
Perry, Donohue, and Weinstein (2007)	observation	Investigated socially and cognitively supportive teaching practices in Grade 1 classrooms. Supportive practices were associated with improvements in mathematics and reading achievement, inter- and intrapersonal social behaviors, and self-perceived academic competence.
Pianta, La Paro, et al. (2007)	observation	Provided information on the psychometric properties of CLASS (and its precursor COS), including reliability and validity findings in Grades PK–6. Studies have shown that CLASS and COS are related to student academic and social progress and that high levels of reliability can be obtained using standard rater training procedures.
Pianta, La Paro, Payne, Cox, and Bradley (2002)	observation	Examined teacher-student interactions in kindergarten classrooms using the COS-K. Positive interactions were associated with students' observed social and on-task behavior and teachers' reports of social and academic competence.
Piburn and Sawada (2000)	observation	Presented the development, reliability, and validity of the RTOP for measuring reformed teaching in mathematics and science, and provided a guide for its use. Found that high interrater reliability could be achieved with appropriate training and that scores linked to student achievement at all levels.

Author (Year)	Category	Summary
Rimm-Kaufman, La Paro, Downer, and Pianta (2005)	observation	Examined the quality of classroom environments in kindergarten using the COS-K. Found students' on- and off-task behavior and aggression toward peers were related to quality of classroom setting, but compliance with teachers' requests was not related.
Sawada, Piburn, Judson, Turley, Falconer, Benford et al. (2002)	observation	Presented data collected for more than two years on the RTOP for measuring reform teaching in mathematics and science from public school, college, and university settings. Established high levels of interrater reliability and internal consistency and links with student achievement gains.
Schacter and Thum (2004)	observation	Examined the development and use of a standards-based performance rubric for evaluating teachers. Found that performance scores were highly predictive of elementary school student achievement on the SAT-9 in mathematics, reading, and ELA.
Schacter, Thum, and Zifkin (2006)	observation	Explored instructional creativity and quality in upper elementary school classes using the observational Creative Teaching Framework protocol. Found that few teachers elicited student creativity, especially in high-minority and low-achieving classes, but fostering creativity was associated with student achievement gains.
Schweinle, Meyer, and Turner (2006)	observation	Examined teaching processes associated with different levels of student-reported motivation and effect in elementary mathematics classes. Constructive teacher feedback, humor, and cooperative learning arrangements were positively associated with student attitudinal and motivational outcomes.
Shymansky, Yore, and Anderson (2004)	observation	Investigated teacher utilization of reformed teaching practices in elementary science. Implementation of practices was associated with improved student attitudes toward science, but no effects were found for achievement.

Author (Year)	Category	Summary
Taylor (2006)	observation	Examined instructional practices used in an instrumental music class. Certain practices, such as constructive feedback and targeted instruction, were associated with improved student performance.
Watson and De Geest (2005)	observation	Examined the practices and beliefs of teachers implementing a mathematics reform program for low-achieving students. Students showed improved mathematics achievement and attitudes. Changes were not tied to specific instructional practices but to teachers' freedom to innovate and underlying principles and beliefs about student learning.
Shavelson, Webb, and Burstein (1986)	other	Reviewed literature on the measurement of teaching, focusing on measurement of teacher effectiveness, classroom processes, and teachers' cognitive processes. Addressed reliability and validity issues and provided implications for measurement.
Smylie and Wenzel (2006)	other	Discussed how the strategic use of human resource management (HRM) at the school, district, and state levels can support and promote instructional improvement. Described the use of HRM practices and their impact in three Chicago elementary schools.
Kennedy (1999)	other, multiple methods	Compared different methods used to measures of student learning, including standardized tests, classroom observations, teacher logs, vignettes, questionnaires, and interviews, through review of the empirical literature. Discussed validity and practicality issues and made recommendations for measurement.
Le, Stecher, Lockwood, Hamilton, Robyn, Williams, et al. (2006)	other, multiple methods	Explored the use of reform-oriented teaching practices in mathematics and science, comparing information from surveys, logs, vignettes, observations, and interviews. Found weak relationships between reform practices and student achievement, though relationships may be affected by achievement measure used.

Author (Year)	Category	Summary
McColskey, Stronge, Ward, Tucker, Howard, Lewis, et al. (2005)	other, multiple methods	Examined the relationship between National Board Certification and student achievement in Grade 5, comparing practices of National Board Certified teachers (NBCTs) and non-NBCTs using multiple methods. Found no differences in average student achievement gains or most other teaching measures, with the exception of planning practices and cognitive challenge of assignments.
Porter, Kirst, Osthoff, and Smithson (1993)	other, multiple methods	Investigated school policy, instruction, and enacted curriculum in secondary mathematics and science using surveys, daily logs, observations, and interviews. Indicated a high correlation between survey and log measures of instruction and presented other findings related to instruction and policy.
Wilkerson, Manatt, Rogers, and Maughan (2000)	other, multiple methods	Compared K–12 principal evaluations to student ratings, teacher self-evaluations, and value-added measures of teacher effectiveness in mathematics, reading, and ELA. Student ratings were best able to predict student achievement on criterion-referenced tests.
Johnson, McDaniel, and Willeke (2000)	portfolios	Investigated the interrater reliability of portfolios using a small-scale family portfolio assessment. Found that reliability was lowest for individual analytic scores, higher for holistic scores, and highest for the composite analytic score. Fewer raters were needed to establish reliability for the composite score than for the others.
Koretz, Stecher, Klein, and McCaffrey (1994)	portfolios	Described the Vermont performance assessment program, which uses portfolios to evaluate teachers in writing and mathematics. Discussed difficulties in establishing interrater reliability and problems with validity and feasibility of the portfolio assessments.
Pecheone and Stansbury (1996)	portfolios	Described the development of a portfolio assessment for beginning elementary teachers in Connecticut. Discussed strategies and challenges associated with creating the assessment, focusing on implementing standards, establishing validity, and building statewide capacity.

Author (Year)	Category	Summary
Pecheone, Pigg, Chung, and Souviney (2005)	portfolios	Explored how a standards-based portfolio assessment promotes learning opportunities in teacher education, comparing traditionally created portfolios with those created electronically. Discussed opportunities and challenges in using performance assessments to measure changes in teacher learning.
Tucker, Stronge, Gareis, and Beers (2003)	portfolios	Examined the validity of portfolios as an assessment of teacher performance. Found they provided adequate documentation of teacher responsibilities and useful information to administrators, but perceptions of their feasibility and contributions to professional growth were mixed.
Cavalluzzo (2004)	portfolios (NBPTS)	Examined whether teacher experience, certification, subject matter, education, and National Board Certification relate to student achievement gains in Grades 9–10 mathematics. Students with NBCTs made larger gains than those with teachers who failed or withdrew from the certification process.
Clotfelter, Ladd, and Vigdor (2006)	portfolios (NBPTS)	Investigated whether teacher experience, licensure test scores, undergraduate institution, and National Board Certification relate to Grade 5 reading and mathematics achievement. National Board Certification had a modest but significant effect on reading achievement.
Cunningham and Stone (2005)	portfolios (NBPTS)	Critiqued National Board Certification as a valid means of recognizing teacher effectiveness and evaluated four studies that examine value-added results for NBCTs compared with non-NBCTs. Concluded that National Board Certification does not identify highly effective teaching as measured by value-added scores.
Goldhaber and Anthony (2004)	portfolios (NBPTS)	Examined the relationship between National Board Certification and value-added gains in elementary student achievement in reading and mathematics. NBCTs were more effective than non-NBCTs in improving student achievement, and National Board Certification could successfully identify effective teachers among NBPTS applicants.

Author (Year)	Category	Summary
Hakel, Koenig, and Elliott (2008)	portfolios (NBPTS)	Reviewed the research on National Board Certification to determine the impact of the National Board Certification process on teachers and the education field. Concluded that National Board Certification is able to identify high-performing teachers, but more direct evidence is needed to establish whether the process itself contributes to improvements in teacher knowledge and instruction.
Harris and Sass (2007a)	portfolios (NBPTS)	Investigated the impact of National Board Certification on mathematics and reading achievement in Grades 3–10. National Board Certification showed a small association with teacher productivity in some cases, but its ability to identify high-quality teachers varied by subject, grade level, and the achievement test given.
Stone (2002)	portfolios (NBPTS)	Examined the relationship between National Board Certification and teacher value-added effectiveness scores in Grades 3–8 in Tennessee. NBCTs were not found to be exceptionally effective in bringing about student achievement gains.
Sanders, Ashton, and Wright (2005)	portfolios (NBPTS)	Examined the relationship between National Board Certification and teacher value-added effectiveness scores in Grades 4–8. NBCTs were not reliably more effective teachers than non-NBCTs studied due to large within-group variability.
Vandevoort, Amrein-Beardsley, and Berliner (2004)	portfolios (NBPTS)	Examined the relationship between National Board Certification and student achievement in Grades 3–6. NBCTs were judged as superior teachers and leaders by supervisors and contributed to robust student achievement gains on SAT-9 tests.
Brandt, Mathers, Oliva, Brown-Sims, and Hess (2007)	principal evaluation	Described district policies on teacher evaluation in a diverse sample of Midwestern districts. Analyzed policy documents to determine specifications on evaluation processes, content, standards, and use of evaluation results and presented several findings.

Author (Year)	Category	Summary
Harris and Sass (2007b)	principal evaluation	Examined the relationship between principals' subjective ratings and teacher effectiveness through value-added measures in a Florida district. Found a positive and significant correlation.
Jacob and Lefgren (2005)	principal evaluation	Compared teachers' subjective assessments by principals to value-added measures of student achievement in reading and mathematics for Grades 2–6. Principal ratings were related to student gains on "Core" exams, especially in mathematics.
Jacob and Lefgren (2008)	principal evaluation	Examined principal ratings of teacher effectiveness and student achievement gains in mathematics and reading for Grades 2–6. Ratings showed small correlations with achievement but also evidence of bias. Principals were less accurate at identifying teachers in the middle range of effectiveness than those in the extremes.
Medley and Coker (1987)	principal evaluation	Examined accuracy of principals' judgments on teacher performance in elementary schools. Correlations between principal ratings and teacher effectiveness as measured by student gains in mathematics and reading were very low.
Follman (1992)	student ratings	Presented an empirical literature review on using public secondary school students' ratings to evaluate teachers, exploring reliability and validity findings and presenting conclusions and recommendations.
Follman (1995)	student ratings	Presented an empirical literature review on using public elementary school students' ratings to evaluate teachers, exploring reliability and validity findings and presenting conclusions and recommendations.
Kyriakides (2005)	student ratings	Examined whether student ratings can provide reliable and valid information to teacher evaluation. Student rating scales measuring teacher-student relationship and cooperation were highly correlated with mathematics and language achievement gains and with affective schooling outcomes in Grade 6 classrooms in Cyprus.

Author (Year)	Category	Summary
Peterson, Wahlquist, and Bone (2000)	student ratings	Examined the use of K–12 student ratings in teacher evaluation. Found that student rating scales showed reasonable reliability and validity but were somewhat upwardly skewed, teachers were generally receptive to the ratings, and students at different grade levels weighted certain classroom aspects differently.
Worrell and Kuterbach (2001)	student ratings	Examined the validity of ratings of low-inference teacher behaviors provided by academically talented high-schoolers. Ratings were slightly upwardly skewed but were associated with teaching behaviors and not overwhelmingly affected by course challenge, expected grade, or student ability level.
Flowers and Hancock (2003)	teacher self-report measures (interviews)	Described development, administration procedures, and scoring rubric for an interview protocol to evaluate teacher performance. Evidence suggested that the protocol is aligned with standards and can be reliably and consistently scored.
Blank, Porter, and Smithson (2001)	teacher self-report measures (surveys)	Reported on the development of the SEC in mathematics and science, discussing central research findings, advances in the survey measures, and important applications of the surveys and data tools. Found that the SEC provided reliable, efficient, and comparable data on curriculum.
Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, and Guiton (1995)	teacher self-report measures (surveys)	Examined survey responses on curriculum enactment of secondary mathematics teachers, comparing them to information from textbooks, assignments, daily logs, and exams. Discussed what the survey can and cannot assess, and described validity concerns.
D'Agostino, Welsh, and Corson (2007)	teacher self-report measures (surveys)	Measured how teachers align practices with state standards and how those standards are tested in Grade 5 mathematics. Found that match between how standards were taught and tested and the interaction between the match and emphasis on standards were best predictors of student achievement.

Author (Year)	Category	Summary
Hamre and Pianta (2001)	teacher self-report measures (surveys)	Examined how teacher-student relationship in kindergarten, defined as teacher-perceived closeness, conflict, and dependency, related to student achievement and behavior in later grades. Found all three dimensions had differing effects on student achievement, work habits, and disciplinary problems.
Kyriakides, Demetriou, and Charalmbous (2006)	teacher self-report measures (surveys)	Attempted to generate measurable criteria of teacher evaluation by eliciting teacher opinions about criteria that should be included in teacher evaluations and comparing responses to principles of teacher effectiveness research. Found general teacher agreement on criteria.
Mayer (1999)	teacher self-report measures (surveys)	Examined the validity and reliability of self-reported teacher survey data on instructional practices. Found that self-reports can determine the relative but not exact amounts of time spent on certain practices.
Mullens (1995)	teacher self-report measures (surveys)	Reviewed research on several large-scale survey measures of instruction and evaluated their applicability for inclusion in the teacher follow-up survey. Discussed the measures in terms of their relationship to student achievement, relevance to policy, appropriateness for a large-scale sample, and level of specificity.
Tytler, Waldrip, and Griffiths (2004)	teacher self-report measures (surveys)	Described the validation of the Science in Schools (SiS) Component Map, by comparing reported practices of science teachers deemed as effective to observational measures. Discussed effective practices and how SiS components relate to a more holistic view of teaching.
Von Secker and Lissitz (1999)	teacher self-report measures (surveys)	Examined science instructional practices in schools undergoing science reform. Found that more laboratory inquiry and less teacher-centered instruction were associated with higher student achievement, while emphasis on critical thinking was not related.

Author (Year)	Category	Summary
Camburn and Barnes (2004)	teacher self-report measures (teaching logs)	Examined the validity of teachings log for measuring instruction, focusing on Grades 1–5 ELA. Found discrepancies between reports of instruction provided by teachers and third-party observers (researchers). Discussed problems with establishing validity of the log.
Rowan, Harrison, and Hayes (2004)	teacher self-report measures (teaching logs)	Examined the use of teaching logs to measure enacted mathematics curriculum at the elementary level. Found high variation in the content and difficulty of lessons from day to day. Discussed implications for reliably and feasibly measuring curriculum using this method.
Aaronson, Barrow, and Sander (2007)	value-added	Attempted to estimate the importance of teachers in Chicago public high schools. Found that teacher effects were positively related to student mathematics achievement, particularly for lower-ability students.
Ballou, Sanders, and Wright (2004)	value-added	Used a modification of the Tennessee Value-Added Assessment System (TVAAS), which includes controls for student SES and demographics to examine teacher effects. Teacher effects were related to gains on CTBS/Terra Nova tests of reading, ELA, and mathematics. Lagged year test score was an appropriate proxy for student background variables.
Betebenner (2004)	value-added	Examined residuals from value-added models of effectiveness and their relation to school and teacher demographic variables. Showed that the most variation in effectiveness was at the teacher level.
Bracey (2004)	value-added	Presented research findings surrounding the use of value-added. Discussed the importance of understanding how teachers improve student learning and discussed several concerns about using value-added measures.
Braun (2005)	value-added	Reported on the use of value-added measures and concerns raised by the literature. Meant as a "layperson's guide" to the practical, technical, and philosophical issues associated with value-added.

Author (Year)	Category	Summary
Dossett and Munoz (2003)	value-added	Described and compared several value-added approaches. Examined how student and teacher characteristics affect student achievement gains, using a proposed longitudinal value-added model.
(Heistad (1999)	value-added	Examined the stability of teacher effectiveness ratings in Grade 2 reading. Reported demographic, attitudinal, and instructional correlates that were associated with highly effective teachers.
Hershberg, Simon, and Lea-Kruger (2004)	value-added	Discussed the importance of value-added measurement in promoting educational improvement. Described relevance to school effectiveness, NCLB goals, and accountability systems.
Kupermintz (2002)	value-added	Examined the validity of measures of teacher effectiveness from TVAAS. Highlighted weaknesses in the model in terms of capturing teachers' unique contributions to student achievement, and questions the usefulness of the scores for comparing teachers or capturing desirable outcomes of teaching.
Kupermintz (2003)	value-added	Examined the mechanism used in TVAAS for calculating estimates of teacher effectiveness and considered relationships between these estimates and factors such as student ability and socioeconomic background. Described perceived weaknesses in the system and calls for additional research to validate TVAAS.
Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007)	value-added	Measured teacher effects using value-added models with different specifications and controls. Suggested that results are sensitive to ways in which student achievement is measured.
Lockwood, McCaffrey, Mariano, and Setodji (2007)	value-added	Illustrated the following with urban school district data: it is difficult to disentangle student background characteristics from teacher effects; teacher effects "dampen" over time so that Value-added models, which include teacher effects, may be misspecified; when missing data is from low-performing students, the teachers' scores may be biased upwards. Discussed possible revisions to the models to adjust for these problems.

Author (Year)	Category	Summary
Lockwood, Louis, and McCaffrey (2002)	value-added	Investigated the performance of rank or percentile estimators used to rank teachers based on student achievement. Showed that use of value-added modeling to determine teacher rankings is inherently flawed.
Martineau (2006)	value-added	Examined concerns with calculating longitudinal value-added student achievement measures using scales that span wide grade, developmental, and content ranges. Demonstrated mathematically that when scales span different content areas, distortions in value-added estimates can result.
McCaffrey and Hamilton (2007)	value-added	Examined Pennsylvania's Value-Added Assessment System, focusing on attitudes toward and use of value-added data for decision making. Found that educators did not make significant use of the information the system provides to improve teaching and learning.
McCaffrey, Lockwood, Koretz, and Hamilton(2004)	value-added	Presented research on value-added models as part of a systematic review and evaluation of leading value-added approaches. Discussed the use of value-added models for measuring teacher effects, reviewed recent applications, and presented important statistical and measurement issues that might affect the validity of value-added model inferences.
McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004)	value-added	Used simulated data to illustrate problems with the general multivariate, longitudinal mixed-model of value-added assessment. One finding revealed that student correlations are robust over time only in schools that serve similar student populations.
Mendro, Jordan, Gomez, Anderson, Bembry, and Schools (1998)	value-added	Investigated the application of multiple linear regression techniques (particularly HLM) in determining longitudinal teacher effectiveness. Found teacher effects were related to students' gains on the ITBS.
Noell (2005)	value-added	Described value-added approaches being used in Louisiana and examined how student and teacher demographics relate to teacher effectiveness. Found some effects for teacher experience and certification.

Author (Year)	Category	Summary
Noell (2006)	value-added	Examined the effect of teacher preparation programs on value-added measures of student achievement in ELA, mathematics, science, and social studies in Grades 4–9. Prior student achievement was the strongest predictor of value-added outcomes, and within-program variation was too high to detect any effects of preparation.
Noell, Porter, and Patt (2007)	value-added	Examined the feasibility of using student achievement, teacher, and curriculum databases to assess the efficacy of teacher preparation programs in Louisiana. Described implementation of the data system and issues and implications of findings.
Raudenbush (2004)	value-added	Examined what kinds of effects can and cannot reasonably be estimated using value-added analyses. Discussed several considerations at the school, teacher, and student levels and provided implications for using value-added models in accountability systems.
Rivers-Sanders (1999)	value-added	Examined residual and cumulative teacher effects on student learning in mathematics for Grades 4–8. Found that all students benefited from highly effective teachers, with the lower-achieving 50 percent benefiting most.
Rivkin, Hanushek, and Kain (2005)	value-added	Investigated the influence of schools and teachers on student achievement gains in mathematics and reading in Grades 3–7. Unobserved differences in teacher quality accounted for most of the difference in achievement; observable teacher characteristics showed some small effects.
Rockoff (2004)	value-added	Examined how teacher fixed effects and experience relate to student achievement in Grades K–6. Both teacher effects and experience had a small effect on student CTBS and Metropolitan Achievement Test scores in mathematics and reading.
Ross, Stringfield, Sanders, and Wright (2003)	value-added	Examined differences in teacher effects between those in restructured vs. nonrestructured elementary schools. Teachers in restructured schools showed higher value-added student achievement gains.

Author (Year)	Category	Summary
Sanders and Horn (1998)	value-added	Reviewed studies examining factors related to student achievement gains as measured by TVAAS. Found that teacher effects significantly explained differences in student achievement while student and classroom characteristics did not.
Sanders and Rivers (1996)	value-added	Described the use of TVAAS to determine teacher effectiveness in a sample of elementary school teachers. Demonstrated how teacher effectiveness made both additive and cumulative contributions to students' gains on TCAP achievement tests.
Sanders, Saxton, and Horn (1997)	value-added	The developer of the Tennessee Value-Added Assessment System (TVAAS) discussed how it works to identify good teachers.
Tekwe, Carter, Ma, Algina, Lucas, Roth, et al. (2004)	value-added	Investigated the impact of differences between three statistical models for assessing school performance using value-added models. Found correlations between the measures, but also some discrepancies.
Thum (2003)	value-added	Presented a three-level education production function model with empirical Bayes residuals to measure student achievement gains. Found teacher effectiveness was difficult to measure with a high degree of certainty.
Valli, Croninger, and Walters (2007)	value-added	Examined the validity of using value-added models to measure teachers, showing that some forms of instructional design rely on multiple teachers, and these designs were pervasive, particularly in higher-poverty schools. Raised questions about holding individual teachers responsible for student learning.
Webster (2005)	value-added	Described the Dallas value-added system, focusing not on the technical aspects but the practical aspects of implementing and using the system as part of teacher evaluation.

Author (Year)	Category	Summary
Wright, Horn, and Sanders (1997)	value-added	Examined how intraclassroom heterogeneity, student achievement level, and class size influence teacher effects as measured by TVAAS. Teacher effects were the dominant factors affecting student academic gain, whereas classroom context variables had little influence on academic gain.
Wright (2004)	value-added	Compared several different statistical approaches to value-added modeling to demonstrate the benefits of using a more complex, multivariate longitudinal approach to calculating value-added measures. Presented results from using each of the models.