Measuring teaching practice: A conceptual review

Courtney A. Bell, Educational Testing Service

Olivia M. Little, University of Wisconsin

Andrew J. Croft, Educational Testing Service

Drew H. Gitomer, Educational Testing Service

DRAFT: DO NOT CITE

Paper presented at the annual meeting of the
American Educational Research Association (AERA)

April 13-17, 2009, San Diego, CA.
Because of limitations in available methodologies for measuring classroom practice, the nature and quality of students’ learning environments have existed as a “black box” in many large-scale evaluation designs. (Clare & Aschbacher, 2001, p. 40)

Empirical evidence now supports what parents, students, and school administrators have always “known”: teachers matter. Multiple value-added studies have found considerable variation in teacher effects on students and these effects persist and predict students’ future outcomes (Gordon, Kane, & Staiger, 2006; Hanushek, Kain, & Rivkin, 1998; Lockwood et al., 2007; Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997). As such, the identification of effective teachers is crucial to the achievement of all students. Identifying effective teachers hinges on how one defines and measures teaching quality.

Value-added models have become increasingly popular measurement tools (Evans, 2008; Tong & O’Malley, 2006). Though value-added has great potential to identify high and low quality teaching, it is currently unable to help teachers or administrators figure out how to improve teaching. It can say who is increasing student achievement, but not what that person is doing to increase achievement. Thus, even with value-added, districts need measures that can help them target particular aspects of instruction to improve teaching.

In order to target particular aspects of instruction, these measures must be more proximal to teaching itself. This requires careful and systematic consideration of the acts that comprise teaching – e.g., interactions with students, teaching subject matter, and planning instruction. Only by gathering and assessing information about these acts of teaching, can districts hope to work with teachers to improve instruction.

Definitions of the Three Sources of Evidence Under Review

In this review we consider three sources of evidence on teaching performance – observation protocols, artifact protocols, and instructional collections – that might be used by districts, administrators, and researchers to judge and improve teaching practice. Certainly there are other evidentiary sources that can and have been used to judge teaching: surveys, teacher logs, and checklists; however, we limit this review to a small number of evidence sources that are most likely to be used to provide feedback to teachers regarding how to improve their instruction. This small set allows us to look carefully at the wide range of validity evidence available. We define these sources below.

Observation protocols are procedures based on categories and scoring scales used by someone to assess the quality of observed teaching performance. The purposes of those assessments range from collegial conversation to the improvement of instruction to the judgment of the quality of instruction. Further, the observation may be conducted by researchers, other teachers, or administrators and may be based on a live or video-taped lesson. The protocols we review here do not include observation checklists, which most often use a dichotomous scale – present or not present.
**Artifact protocols** are instruments which assess records of practice that are created as a part of regular classroom instruction. Examples include assignments, grading criteria, student work, lesson plans, etc. These protocols assess artifacts that are not created for an external audience. Information provided by the teacher to the researcher that goes beyond what is normally created for instruction is contextual in nature (e.g., grade level of the assignment, state standard being addressed by artifact, etc.). Another distinguishing feature of these protocols is that the researcher or district or school establishes the constraints which identify which artifacts will be included. There are narrow opportunities for the teacher to decide which artifacts are assessed by the protocol.

In contrast to artifact protocols, protocols that assess instructional collections assess evidence from both records that are created only for the external assessment audience and records that are created as part of regular classroom instruction. They often include diverse records ranging from assignments and student work to lesson plans, reflections, videos, and student interviews. The wide range of records intentionally provides the external audience with more than just contextual information. Instead, the records provide the teacher’s interpretation of the collection. Though protocols that assess instructional collections differ, they generally require a scorer to assess both the teacher’s interpretation of the records and the records themselves. Scoring for these protocols provides a holistic score across all the records in the collection. Portfolios are an example of a type of instructional collection.

This is not a review of instruments, though we do discuss instruments and provide an appendix that readers interested in specific instruments may find helpful. The goal of this conceptual review is to identify what researchers know about the validity and reliability of various evidentiary sources that are designed to provide actionable information. These evidentiary sources focus on the products, actions, and discourse that occur in classrooms by teachers and students. As such, they are closer to the acts of teaching and learning than many other evidentiary sources.

The review is organized as follows. First, we lay out a definition of teaching quality that articulates what is important to measure in order to improve instruction. Second, we describe our approach to validity, one which is consistent with Kane’s (2006) recent articulation of a unified construct. We then use that framework to summarize and evaluate existing scholarship on three evidentiary sources: observation protocols, artifact protocols, and instructional collections. Finally, we conclude with a discussion about future directions for the measurement of teaching performance.

**A Definition of Teaching Quality**

To be clear, this review is not focused on teacher quality, nor is it focused on teacher effectiveness. Both of these terms, which are in widespread use in the policy arena, emphasize the individual and we believe such a focus places undue and unproductive emphasis on the person, rather than the fact that teaching occurs within a complex system that is influenced by such factors as curriculum, school leadership, student and community characteristics, etc. Certainly the core constructs of teacher quality – credentials, course credits, test scores, beliefs, personality – are critically important to helping children learn. Further, the notion that we should
pay attention to whether teachers improve student learning as measured by standardized achievement scores, one aspect of teacher effectiveness, is also a very important idea. But student outcomes result from a set of factors that include teaching practices. Thus, this review focuses on understanding classroom and professional practices that can be improved upon through both individual and organizational development.

Implicit in any review of teaching is a normative conception of good teaching. There have been countless efforts to define good teaching (Campbell, Kyriakides, Muijs, & Robinson, 2004; Fenstermacher & Richardson, 2005; McColskey et al., 2005; Shavelson, Webb, & Burstein, 1986; Tucker & Stronge, 2005). Proposed definitions typically have been driven by available measurement capabilities targeted toward easily quantifiable characteristics of teachers and students, such as student achievement or teacher knowledge. Others have placed an emphasis on non-cognitive attributes of teachers (e.g., resourcefulness) or non-cognitive student outcomes (e.g., self-efficacy). Researchers have called for increased attention to the teacher’s activity outside the classroom in the larger context of the school as an organization and also the surrounding community (Cheng & Tsui, 1999).

This review defines high quality teaching as teaching that enables all students to learn and develop the subject matter knowledge, personal characteristics, and interpersonal skills that facilitate high academic achievement, job acquisition, and participation in a democracy. This definition focuses our attention on four areas: 1) classroom practices, 2) professional activities outside of the classroom, and how these first two areas affect 3) student learning, and 4) student attributes that ultimately bear on academic achievement, acquiring a job, and participating in a democracy (e.g., motivation, self-efficacy, etc.). We use these areas to help us assess the presence of critical pieces of validity evidence in observation protocols, assignment protocols, and instructional collections.

The above definition of teaching quality acknowledges that it is important to pay attention to what teachers do (areas one and two) and the outcomes teachers’ actions contribute to (areas three and four). This view of teaching bridges narrow definitions that focus exclusively on process and other inputs (Wayne & Youngs, 2003) or on outcomes (Hanushek, 1997; Muijs, 2006). It also suggests that as we consider what the field knows and does not know about teaching quality, we must be very careful to identify the specific aspects of teaching quality a given evidence source can speak to. For example, classroom observations are well suited to support claims about classroom interactions between teachers and students; while instructional collections, which draw on evidence from a range of sources – lesson plans, reflections, assignments, observations – may be more suited to support claims about an individual’s teaching practice broadly conceived.

Our Approach to Validity

Our approach to investigating these instruments is consistent with the prevailing unitary view of validity (American Psychological Association, National Council on Measurement in Education, & American Educational Research Association, 1999). In particular, we are interested in an argument based approach to the validation and validity of instruments as Kane (2006) defines them. He explains, “[Validation is] the process of evaluating the plausibility of proposed
interpretations and uses, and [validity is] the extent to which the evidence supports or refutes the proposed interpretations and uses” (p.17).

Kane (2006) suggests there are two types of arguments that should be specified in a validation effort, the interpretive argument and the validity argument. The interpretive argument lays out the “network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances” (p. 23). The validity argument evaluates the interpretive argument in light of empirical data and analytic reasoning. Specific interpretive and validity arguments must be viewed in light of the proposed uses of the scores. In other words, instruments should not be thought of as valid or invalid. Instead, specific interpretations of scores from a given instrument may be valid for a particular use. The interpretive argument lays bare the connections among assumptions, interpretations, and conclusions and the validity argument assesses their strength. The specific methods necessary for validation are determined by the nature of the interpretive argument. Thus, we specify both the uses of the instruments and a general interpretive argument that is consistent with those uses. This serves to bound the review so that claims about the validity of the instruments is possible.

This review has two goals. The first goal is to understand where the evidentiary sources of teaching quality are in a broader validation process. The second goal of the review is to assess the degree to which evidence supports or refutes particular interpretations and uses of these evidentiary sources. Below, we describe the interpretations and uses on which we focus our review.

All the evidentiary sources under investigation provide data on various aspects of teaching quality. However, the specific instruments within each evidentiary source are used toward different ends. Some instruments were designed to assess a particular reform or intervention (McCaslin et al., 2006; Sterbinsky, Ross, & Redfield, 2006) while others provide a score that is incorporated into a larger licensure decision (McColskey et al., 2005). We review this literature with two uses in mind: the improvement of teaching practice and making high stakes personnel decisions (e.g., dismissal and compensation). While there are other important uses for these instruments (e.g., the evaluation of professional development programs), we focus on these two uses because the current political climate presses school districts to improve teaching quality in order to meet the accountability demands of the 2001 reauthorization of the Elementary and Secondary Education Act (also known as the No Child Left Behind Act). This press has generated a great deal of interest in methods of teacher evaluation in the service of helping students meet adequate yearly progress (AYP). For example, the Teacher Incentive Fund monies are being used by many districts to implement both value-added assessments and observation based assessments to improve teaching practice ("Prince George's County Public Schools Financial Incentive Rewards for Supervisors and Teachers," 2007). Regarding education policy, the Obama administration is focused on making improvements in teacher effectiveness and establishing equitable distribution of qualified teachers ("The American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education," 2009). This federal focus is likely to result in even more local interest in how best to measure teaching quality.
The two uses of these instruments do not get equal attention in the review. We focus more on the support for using these instruments to support high stakes personnel decisions than using them for the improvement of teaching practice. We do this for two reasons. First, although some researchers have used these instruments for the improvement of teaching practice (Crosson et al., 2006), research on this use is very scarce. Second, we are persuaded by the literature on professional development that teacher learning is complex and requires specific supports in order for it to work (Garet, Porter, Desimone, Birman, & Yoon, 2001; Wilson & Berne, 1999). Thus, in order for us to know whether and how a specific instrument may improve practice, there will need to be detailed accounts of the supports that accompanied the use of the instrument (e.g., how and when the instrument was used, the quality of the training, the follow-up with teachers on the training, evidence of teaching learning). This level of detail is generally missing from reports that do use instruments to improve teaching.

Adapting Kane’s (2006) examples of interpretive arguments, we lay out an interpretive argument for the three sources of evidence we consider. We then assess the validity arguments for each of the three sources. We make some important assumptions that are critical to understanding the review. First, we presume that to varying degrees, all of the instruments measure some aspects of teaching quality. Teaching quality is a large construct and any one instrument focuses on a very narrow part of the larger construct via a specific source of evidence (e.g., the cognitive demand in assignments or the emotional support provided to students during a classroom lesson). Second, we presume that other traits of teachers (e.g., extroversion, warmth, etc.) and teaching context (location of the schools, curriculum materials used in the district, etc.) are important to teaching quality but they are not what these instruments attempt to capture. Finally, though all the instruments under review measure various aspects of teaching quality, we refer to the types of claims these instruments intend to make as judgments of teaching performance (See Figure 1).
The interpretive argument outlines four related sets of assumptions and interpretations that should be evaluated in the validity argument. I) Scoring: from the observed performance to the observed score on an observation, artifact, or portfolio protocol; II) Generalization: from the observed score to the score one could expect on all artifacts, all lessons, or all collections of reflections and artifacts; III) Extrapolation: from the score one could expect on all artifacts, lessons, or collections of reflections and artifacts to the level of teaching quality; and IV) Interpretation: from the level of teaching quality to the description of teaching performance.

I. Scoring
1.1 The scoring rule is appropriate.
1.2 The scoring rule is applied accurately and consistently.
1.3 The scoring is bias free.
1.4 The data fit the scoring model.

II. Generalization
2.1 The sample adequately represents the quality of all lessons, or artifacts, or collections of artifacts.
2.2 The sample is sufficiently large that random error is controlled for.

III. Extrapolation
3.1 The score on all lessons, artifacts, or collections of artifacts is related to the teaching quality an individual is able to enact.
3.2 There are not systematic errors that undermine the extrapolation to teaching quality.

Adapted from Kane (2006) Figure 2.2 (p. 33)
IV. Implication
4.1 The implications associated with teaching performance are appropriate.
4.2 The properties of the observed scores on the lessons, artifacts, and collections of artifacts support the implications associated with the judgments of teaching performance.

For each of the three sources of evidence, we use this interpretive argument as the basis for a validity argument specific to the evidentiary source.

Methods of Locating, Reviewing and Synthesizing Articles

This review was conducted by first identifying empirical studies on how teaching quality is measured, and then examining how different evidentiary sources of teaching quality were used in those studies. Articles were selected using the process described below, the instruments used to measure aspects of teaching practice were identified in each article, and available reliability and validity evidence on these instruments was extracted. The authors reviewed all articles, engaged in continued conversation to create a shared understanding of the study purpose and validation framework, made decisions at each stage of the review process, and discussed any decision discrepancies until they were resolved.

The process of locating studies occurred through three main avenues. First, based on a previously completed paper examining measures of teacher effectiveness (Goe, Bell, & Little, 2008), an initial database was created of studies that had been identified for this project and were deemed relevant to the current paper. This initial database included 113 articles. Second, a comprehensive literature search was conducted, and articles were narrowed down according to the procedure described below. This yielded an additional 277 articles which were added to the database. Third, throughout the search and review process other articles were identified, both empirical and theoretical, through reference lists, conference presentations, and expert suggestions. Forty articles were added to our database through this route. As this is an ongoing project, we are continuing to add articles to our database. This paper reports on the results we have obtained from our current total of 430 articles.

Literature Search

To obtain the majority of our data, a comprehensive literature search was conducted in ERIC and PsycInfo. The authors met on a number of occasions to discuss the purpose of this review and to determine a broad list of search descriptors that were likely to encompass studies of instructional quality. In ERIC, the following descriptors were entered:

- Assignments
- Beginning Teacher Induction
- Instructional Development
- Instructional Effectiveness
- Instructional Improvement
- Instructional Material Evaluation
- Pedagogical Content Knowledge
- Professional Development
- Questionnaires
- Research Tools
- Teacher Behavior
- Teacher Competencies
- Teacher Effectiveness
- Teacher Evaluation
- Teacher Improvement
- Teaching Methods
- Videotape recordings

In PsycInfo, the descriptors used were:
- Teacher Attitudes
- Teacher Characteristics
- Teacher Effectiveness Evaluation
- Teacher Student Interaction
- Teachers
- Teaching Methods
- Teaching

The following limiters were specified:
- Published between 2002-2008
- Published in English
- Peer-reviewed journal article
- Tagged as Elementary Education or Secondary Education in ERIC; tagged as School Age (6-12) or Adolescence (13-17) in PsycINFO
- In-service teachers (in ERIC)

We limited our search to 2002-2008 because we were interested in reviewing recent research, and this is the period for which the databases marked peer-reviewed literature. Additionally, our primary interest is in the measurement of K-12 in-service teaching. Thus, although some of the instruments and evidentiary sources we review are prevalent in the literature on pre-service and student teaching experiences, such as video observation and teaching portfolios, we did not include these studies unless they measured full-time teachers. The ERIC search yielded 1575 unique articles, and the PsycInfo search yielded an additional 1665 unique articles.

In order to narrow these results, abstracts were reviewed and coded according to the following additional criteria:
- *Empirical*\(^1\): Study refers to specific collected data upon which the author is basing his/her conclusions.

\(^1\) Note that although this paper only reviews empirical articles, our database includes several non-empirical, theoretical articles to help us construct a conceptual understanding of teacher effectiveness and ways to measure it.
• **Location:** Study was conducted in an English speaking country with an educational culture similar to our own (includes the United States, Canada, Great Britain, Ireland, Australia, or New Zealand).

• **Population:** Research addresses the K-12 student population, and measures in-service teachers.

• **On Topic:** The article addresses the topic of measuring instructional quality, focused on methods used to measure teaching.

• **Direct Measure:** The methods used to measure teaching are direct sources of evidence (e.g., observations, videos, instructional artifacts, portfolios, instructional collections), meaning that what is being measured occurred or was produced inside a classroom.

If the abstract contained insufficient information to make a determination on any of these criteria, the article was coded as needing further review. Articles meeting all five criteria and those needing further review (277 total) were added to the database and proceeded to the next stage.

Next, the methods sections of all articles in the database were examined to (1) verify that all previously mentioned criteria were indeed met, and (2) to determine whether the article contained documentation of an instrument that directly measured teaching practice. We call this category “Replicable Protocol,” to indicate whether a study’s methods consisted of some specific protocol with a defined or referenced coding scheme. This excludes studies which present only ethnographic or qualitative/descriptive accounts of teaching. The resulting number of studies meeting all above criteria and including a Replicable Protocol was 157. Finally, the articles considered in this review are those which included sufficient information about the methods to inform our conversation about the validity and reliability of evidentiary sources of instructional quality.

By beginning our search with only published research reports and branching out to unpublished reports mainly through the review of reference lists and expert suggestions, we acknowledge that our findings have a publication bias (Kennedy, 2007) and are, to some extent, backward looking. Where possible we have incorporated recent work into the review that is not yet published but does have a bearing on our findings. These studies tend to be incorporated on more general trends rather than specific findings, because they have not yet been written about formally and therefore have not had the benefit of clarification and completeness that tends to come from writing up a more formal account of one’s research.

**Observation Protocols**

There is a wide range of observation protocols, all of which rely on evidence collected when a rater observes a teacher teaching. The protocols vary with respect to grade level, subject area(s), the degree to which the protocol is used in a form that matches the form its developers created, and the level of inference required by the rater. Each of these variations has important implications for assessing the validity argument of this evidentiary source.

Because this manuscript is evolving as additional instruments become incorporated, the findings below only pertain to the eight instruments reviewed for this draft. These instruments include Danielson’s Framework for Teaching (FFT), the elementary and secondary versions of the
Classroom Assessment Scoring System (CLASS), the Mathematical Quality of Instruction (MQI), the TEX-IN3, the Local Systemic Classroom Observation Protocol (LSCOP), the English Language Learner Classroom Observation Instrument (ELLCOI), the Mathematics Enhancing Classroom Observation Recording System (MECORS), and the Reformed Teaching Observation Protocol (RTOP).

Scoring
The strongest parts of the validity argument for scoring concern the care with which the instruments were developed and vetted with experts. The evidence also suggests it is possible to train raters to reliably score when compared with one another. There is less support for claims that the scoring rules have been applied accurately and without bias. Finally, support for the application of scoring models to data is thin and deserves further attention.

1.1 The scoring rule is appropriate.

Almost all of the rubrics employed were developed after a careful review of the literature (Danielson, 1996; Gersten, Baker, Haager, & Graves, 2005; La Paro, Pianta, & Stuhlman, 2004) and in many cases, vetting with external audiences such as teachers, administrators, and experts in measurement or subject area teaching and learning (Hoffman, Sailors, Duffy, & Beretvas, 2004). In fewer cases, protocol development also included alignment with national standards (Danielson, 1996; Piburn & Sawada, 2000). These standards ranged from general standards for all subject areas such as the Interstate New Teacher Assessment and Support Committee (INTASC) core standards to subject-specific standards such as the National Council of Teachers of Mathematics’ Principles and Standards for School Mathematics. Among both the subject-specific and general protocols, some were aligned with standards, others were the result of extensive literature searches, and some were a combination of both.

The rubrics or scoring rules that are being used in these instruments have an articulated rationale. Instruments that make low inference judgments (e.g., MECORS, Muijs & Reynolds, 2003; CIERA Classroom Observation Scheme, Taylor, Pearson, Peterson, & Rodriguez, 2003) used scoring rules that were reasonable with the inferences being made. For example, in MECORS there is a list of statements about the classroom: “The teacher starts a lesson on time,” or, “The teacher corrects behaviour constructively”. These statements are marked on a seven-point scale of agreement for each statement. In the case of instruments which require higher level inferences there were a range of rubric points, from four to seven. In the FFT, there is a four point scale – unsatisfactory, basic, proficient, and distinguished – and each scale point is described for each dimension of the instrument. In contrast, all the CLASS instruments use a seven point scale, with written descriptors at the 2, 4, and 6 score points. The articles we reviewed did not provide information on why specific instruments used particular scales or if they changed scales over the development process.

Only a handful of studies report on the internal reliability of scales or on the refinement of those scales based on factor analyses or other statistical analyses (Blunk, 2007; La Paro et al., 2004). In studies that did report internal consistency (alpha), scales were generally reliable at the 0.6 to 0.8 level. Internal reliability and factor analyses may have been conducted on many of the instruments reviewed but it was not a part of the documentation we found. This is an important
omission, in particular, for the instruments that were developed only out of a review of the literature. While constructs may be theoretically distinct in the literature, it is important to understand if they are being measured well in a specific protocol.

1.2 The scoring rule is applied accurately and consistently.

Interrater reliability is usually reported but the level of depth is reported unevenly. There is a good deal of information documenting what percentage of the time raters agreed with one another during training. For example, interrater reliability (usually reported as Cohen’s Kappa) on lower inference instruments ranged from .81 to .96. Higher inference instruments reported using the general approach of making sure raters met a .80 interrater agreement prior to being certified to code for the actual study.

The practice of double scoring observations to ensure consistency in rating over the course of a study was reported much less often. For those studies that report double scoring to investigate interrater reliability over the time of the study and as compared to master coders, ratings were typically checked at the beginning and end of a study and/or a master coder was employed throughout the study to detect drift and correct it with prompt training (Hoffman et al., 2004). Details about what was done to account for any rater drift (retraining, statistically modeling the drift, etc.), is rarely described. In the cases where it was described, instrument developers reported dealing with interrater agreement problems by retraining raters until they reached the desired levels of agreement. We did not review any articles that modeled the drift which may have occurred.

1.3 The scoring is bias free.

For observational data it is critical that selection and training procedures for raters do not introduce significant sources of bias. On the selection of raters, there was very little reporting. Occasionally a report would provide some details about who the raters were – graduate students, faculty, district administrators. Because very few of the observation instruments were used for high stakes purposes and there are no studies comparing raters with different backgrounds, we can only hypothesize about the impact of each of these groups of raters. It is possible that graduate students may not have seen enough instruction or faculty may hold such strong beliefs about instruction that these two groups may be systematically biased. It is also possible that administrators (e.g., principals or central office personnel) may have prior knowledge that systematically biases the rater’s observations. It may also be that each of these respective groups introduces different face validity concerns for high stakes assessments. Because we have very little information on raters and even less information about how rater background influences scores on these instruments or other sources of bias, we must look especially carefully at how raters were trained.

Almost all of the reports document the process by which the raters were trained in detail. However, there were very thin descriptions on issues critical to understanding how bias was handled. For example, we know little about how bias is addressed in the training or what counts as evidence in support of a particular score point for each dimension in the protocol. Interrater reliability tells us how well the training worked. But because empirical evidence on interrater
reliability over the length of the study is spotty and double-coding procedures are only reported in a handful of studies, even with good descriptions of rater training it is unclear how well that training served the study in the longer term. The articles that do report double-coding procedures and interrater reliability across time suggest there is often a need to retrain raters and in some cases, remove raters who cannot apply the scoring rubrics as the developers intended. Overall, the evidence for scoring being bias free is weak.

1.4 The data fit the scoring model.

It was rare to find any discussion of the scoring model or the fit of the model to the data. Most often, straightforward scoring models were used. Scores were summed or averaged across observation cycles and there was no analysis of the appropriateness of that approach. In the case of instruments that used more complex scoring models (e.g., an IRT model), there was modest support that the model fit that data.

Generalization

Though there are a couple of instruments for which there are some relevant analyses, the validity argument for the relationship between the sample of lessons and the universe of lessons is exceptionally weak. The field does not have any general rules of thumb regarding how many lessons one must observe on these protocols in order to get a representative sample of lessons. Nor do we know how long we must observe, what time of day we should observe or what times of year we should observe. Further, we do not know how many years a teacher must be observed before we have a stable estimate of her practice. Nor do we know how different subject areas within teachers may vary. So if you are observing a 4th grade teacher and wish to use observations to support a decision to give her merit pay, should you observe one lesson in each of the four subject areas (English/language arts, social studies, science, and mathematics)? Will observations in just English/language arts and mathematics give you a representative view of her overall teaching practice? What about for an elementary music teacher? Do you visit one lesson at each grade level? The research literature offers very few answers to these basic questions.

2.1 The sample adequately represents the quality of all lessons, or artifacts, or collections of artifacts.

One way to establish the parameters for an adequate sample is to conduct a generalizability study. Our review did not uncover a single generalizability study, but we did find one report that considered sources of variation seriously (Pianta, Hamre, Haynes, Mintz, & La Paro, 2007). In the technical appendix of the CLASS manual, the developers analyze evidence from a large number of studies across the United States. These studies draw on data from use of the CLASS-Pre-K and K-5 instruments.

We did not find any studies that evaluated how many observations one needs in order to have a stable year-to-year estimate of teaching practice. However, there is some evidence that depending on the instrument and the length of a scoring session, an instrument might need between four and six lessons in order to get reliable scores for a teacher (Blunk, 2007). For instruments that use cycles, rather than lessons as the unit of analysis, CLASS developers found that four cycles (20’ observing, 10’ coding) resulted in a representative sampling of the
classroom (Pianta, et al., 2007). They found that correlations between the total score and different cycle scores (i.e., first cycle score, second cycle score, etc.) varied across dimensions of CLASS within acceptable limits.

The most extensive work on variation in scores by time of day has been done on the CLASS instrument (Pianta, Hamre et al., 2007). CLASS developers found that there are slight differences in the mean scores across the school day. For example, they found that emotional support scores on the Pre-K version of the instrument decreased linearly across the school day. They did not find the same time of day relationship on the K-5 version of the instrument, though there was a significant negative linear trend in scores on the classroom organization domain across the school day.

The K-5 and Pre-K CLASS documentation also considered how scores varied across days of the week and times during the school year. In both cases they found small significant differences across days of the week and over the school year. Comparisons between fall and spring scores revealed moderate stability in most domains. Instructional dimensions appeared to vary the most in these spring/fall comparisons, with some reaching statistically significant levels. Scores in all three domains – emotional support, instructional support, and classroom organization – decreased at the end of the school year. There were also small decreases in mean scores on the emotional support and classroom organization domains around the December holidays.

We also did not find a single study that looked carefully at the stability of scores across subject areas, within teachers. This is particularly important if evidence from observation protocols is to be used for teachers who teach in self-contained classrooms where they are responsible for multiple subject areas in which they may have differential strength. An overall description of teaching performance for these teachers would need to account for any variation that stems from subject matter.

2.2 The sample is sufficiently large that random error is controlled for.

We did not find any information about how much random error is reasonable to expect in observation scores. In general, it is not common practice to report confidence intervals around total or domain level scores.

Extrapolation

The extrapolation inference is one of the most critical in understanding the strength of the validity argument when an instrument is going to be used for high stakes decisions. The extrapolation considers the relationship between the scores one might expect on all the lessons a teacher might teach and the level of teaching quality. In other words, in reviewing extrapolation evidence, we ask ourselves how sure we are that the observed scores (with whatever error and uncertainty they contain) tell us something about the level of performance in the much larger teaching quality domain. The evidence in this area is mixed. Scores have been related to standardized student achievement outcomes but not much else. The validity argument would be stronger if scores on the instruments were related to broader notions of high quality teaching. This might include notions of professional practice and student outcomes such as motivation and future course taking patterns.
3.1 The score on all lessons, artifacts, or collections of artifacts is related to the teaching quality an individual is able to enact.

The most common convergent validity evidence used to assess the extrapolation from the observation score to the wider domain of teaching quality is a comparison of observation-based scores to students’ standardized achievement test gains or gains on other student learning measures. The logic for this common comparison suggests that high quality teaching helps students learn academic material. Thus, instruments that attempt to measure teaching practices inside the classroom should be related to gains in student achievement. Scores on a number of instruments (e.g., CLASS-Pre-K, CLASS-K-5, FFT, English Language Learner Classroom Observation Instrument (ELLCOI), Reformed Teaching Observation Protocol (RTOP), CIERA’s Classroom Observation Scheme) have been correlated with or shown to be predictive of student achievement gains on standardized tests (Archibald, 2007; Baker, Gersten, Haager, & Dingle, 2006; Gallagher, 2004; Hoffman et al., 2004; Holtzapple, 2003; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004; Sawada et al., 2002; Schacter & Thum, 2004; Taylor et al., 2003). Where there are statistically significant relationships, they range from modest to moderate and vary by subject area. For example, Holtzapple (2003) found correlations (Kendall’s tau) between composite scores summed across the four FFT domains and students’ gains on a state standardized test, which ranged from 0.28 to 0.37, depending on subject tested. These correlations were statistically significant at conventional levels. Other work on the FFT found similar relationships but these relationships varied by grade level and subject area (Gallagher, 2004; Kimball et al., 2004; Milanowski, 2004). It is important to note that there was wide variation in rater training, raters’ relationship with the teacher (peer, supervisor, etc.), the degree of adherence to Danielson’s recommendations for use, the use of the scores, the student measures, and the number of observations conducted for each teacher. This variation may be partially responsible for the range of findings.

In addition to the evidence that associates observation scores with student test scores, there are a handful of studies that document some evidence of a relationship between scores on observation protocols and scores on knowledge tests which assess the knowledge required for teaching. The logic undergirding this relationship suggests that high quality teaching requires knowledge of teaching. Thus, we would expect there to be a direct relationship between teacher knowledge and scores on an observation protocol which can measure that knowledge in action. In support of this assertion, Hill and colleagues (2007; 2008) recently found strong relationships between teachers’ scores on various dimensions of MQI and their overall scores on the Learning Mathematics for Teaching project’s assessment of mathematical knowledge for teaching (LMT-MKT), a multiple-choice assessment of mathematical knowledge for teaching (Hill, Schilling, & Ball, 2004). There is also some evidence that observation protocols are correlated with teachers’ beliefs, with at least one study finding that teachers’ beliefs mediated changes in teachers observation scores when those teachers were participating in a curriculum reform intervention (MacIsaac, Sawada, & Falconer, 2001).

3.2 There are not systematic errors that undermine the extrapolation to teaching quality.
The most important threat to the validity of the inference from scores to the larger teaching quality domain rests with the relationship between what the scores measure and the broadly defined domain of teaching quality. As we have defined teaching quality, there are both professional responsibilities outside of the classroom and inside of the classroom which define teaching quality. For example, we would expect that high quality teaching would include working productively with parents, supporting colleagues, and contributing to the school community. We would also expect that high quality teaching would engender the habits of mind, social skills, and academic ability that would prepare a child for success in life. This is a tall order for any one instrument or even a set of instruments.

Our review suggests that observational evidence, as instantiated in the instruments reviewed here, covers a narrow slice of the teaching quality domain. There is strong support for that slice, but nonetheless, the slice is narrow.

It is important to note, for example, that most of the evidence for the strength of the extrapolation from observation-based scores to the larger construct of teaching quality is focused on standardized student achievement measures. With a handful of notable exceptions at the younger grade levels (Foorman et al., 2006; Hamre & Pianta, 2005; Perry, Donohue, & Weinstein, 2007), very little research considers the relationship between observation-based scores and other aspects of teaching quality we believe are critical for both in-school and out-of-school success. Generally speaking, we do not know what the relationship is between observation-based scores and the student attributes that ultimately bear on academic achievement, acquiring a job, and participating in a democracy. There is significant research suggesting quality teaching can influence attributes such as motivation and self-efficacy (Blackwell, 2007), however there is very little documentation of this relationship for the instruments we reviewed, especially beyond the early school grades. The exceptions to this generalization consider the relationship between observation-based scores and children’s self-perceived academic competence, inter- and intra-personal behavior, student engagement, and teacher-student conflict. Noticeably absent from the studies reviewed was the relationship between observation-based scores and broader student attributes which might help one contribute to a democracy or become gainfully employed (e.g., honesty, executive process skills, and inter-personal skills).

It is also important to note that the different instruments attempt to account for different aspects of the overall teaching quality construct. Some instruments are designed to measure teaching quality across all subject areas. Others specifically target a smaller part of the teaching quality construct – mathematics teaching quality or science teaching quality. These types of differences highlight a trade-off in the amount of evidence that must be amassed. General observation protocols must document how the instrument holds up across grade levels and subject areas. Subject specific protocols may only have to deal with grade level differences. That subject specificity comes however, with the reality that districts may find it very difficult to use different instruments for different groups of teachers.

**Implication**

There is moderate support for the relationship between teaching quality and teaching performance. There have been adequate development efforts that build on established knowledge about teaching quality. Further, there is modest evidence that judgments about
teaching performance are sensitive to the impacts of learning, which theory suggests they should be. The validity argument is weakest in its specification of the long-term stability of scores and what that stability may mean about teaching performance judgments.

4.1 The implications associated with teaching performance are appropriate.

Generally support for the appropriateness of the implications associated with a particular teaching performance is established during the development stages of the instrument. In the scoring part of this section, we reviewed this support. In general, the content of the protocols have been vetted with multiple stakeholders such as teachers, administrators, and researchers. The protocols reflect a community’s knowledge about what practices and approaches to teaching reflect strong teaching performances. Unlike early observation protocols from the 1970’s, many of the instruments reviewed here deliberately do not privilege particular teaching practices. For example, the MQI does not presume group work in mathematics is superior to whole class instruction. The FFT does not assign high scores for a particular classroom management approach. That said, each instrument embodies a different conception of good teaching and it is these details with which various constituencies may disagree.

For example, the FFT takes a constructivist approach. The highest levels of performance require that students play an active role in all aspects of the classroom, including making meaning, contributing to classroom processes, and supporting discussion and questioning techniques. Various constituencies may reasonably disagree about the importance of this specific approach. In another example, MQI focuses almost exclusively on mathematics in the classroom. The instrument does not have any dimensions that rate teachers on their ability to enact smooth transitions or provide emotional support to students. Again, various constituencies may take issue with such a conception of teaching.

Our review suggests that the instruments all have reasonable political constituencies and theoretical documentation that support the inferences about teaching performance. However, the potential for using these protocols for high-stakes decisions requires that they be vetted appropriately with the teachers, administrators, and school communities with whom they will be used. Strong disagreements about the approach a particular instrument takes will seriously undermine the validity argument for the protocol.

4.2 The properties of the observed scores on the lessons, artifacts, and collections of artifacts support the implications associated with the judgments of teaching performance.

We did not find a single study that looked carefully at the stability of observation scores over multiple years. This is a very important gap in the literature. If evidence from observations is to be used to make high-stakes decisions, we must be sure that the scores are stable over time. Currently there is very little support for this claim. It is reasonable to presume that after the initial learning curve that has been documented at the beginning of a teacher’s career (roughly 3 to 5 years), a teacher’s practice may remain stable and thus, we would expect observational evidence to also be stable. But this has yet to be documented in the literature.
Further, we found no studies that compare scores across different groups of teachers. It is possible, though perhaps not probable, that these instruments function differently for teachers at different points on the developmental continuum (e.g., preservice, beginning, and in-service).

We did find some evidence that certain instruments are sensitive to changes in teaching practice. For example, on RTOP, teachers participating in professional development seminars aimed at improving the reformed nature of their teaching showed positive changes in their scores (Roehrig & Kruse, 2005).

**Summary Trends on Observation-based Evidence Sources**

There are a number of general trends we found in our review of these observation protocols. First, the protocols tend to fall into two broad categories – protocols that can be used across subject areas, and those intended to be used in specific areas. We found subject specific protocols in mathematics, science, and English/language arts. We did not find any subject specific instruments for social studies classrooms (e.g., social studies, history, government, geography, etc.). We view this as an unfortunate gap in the literature.

Second, there were more protocols for use at the elementary grades than the secondary ones. Many of the subject-specific protocols were validated in K-3, K-5, or K-8 classrooms, making it unclear whether or how the protocols might function differently as the curriculum shifts and students get older. The development and validation of strong secondary observation instruments across the full range of subject areas and grade levels would be a welcome addition to the literature.

Finally, video-taped observations were often treated as interchangeable with in-person observations. Protocols like the RTOP used both video and in-person observations to develop the instrument. CLASS used both video and in-person observations to train raters and establish interrater reliability. Though the two modes were frequently used interchangeably, we did not find systematic evidence to support this treatment. It is possible for example, that sound quality differences make video-taped lessons systematically different on some instruments and not others, or on some dimensions within the same instrument. It is also possible that the two modes vary but the variations are small enough that they do not produce meaningful differences in scores. Systematic investigation into the nature of the relationship between in-person and video scoring would provide support for the common practice of treating them interchangeably.

**Artifacts**

A second evidentiary source of teaching quality is artifacts. Artifacts such as in-class assignments, completed student work, and teacher feedback provide external audiences with evidence of both the nature of the conversation about subject matter and students’ learning of that subject matter. These protocols can gauge the teacher’s “assessment literacy” which is thought to provide a window into much larger aspects of teaching and learning (Stiggins, 1991). While observations tend to focus mostly on the teacher actions and student involvement visible in a lesson, instructional artifacts go one step deeper and allow for an investigation of the relationships between teacher involvement in student learning and student engagement with
academic content. Artifact protocols allow for investigations of important aspects of instruction, including the construction of knowledge, disciplined inquiry, the value of the subject matter beyond school, clarity of learning goals, and the quality of grading criteria (Matsumura, Garnier, Pascal, & Valdés, 2002; Newmann, Marks, & Gamoran, 1996).

This review hones in on two development efforts: one by Newmann and his colleagues for the Consortium on Chicago School Research, the Intellectual Demand of Assignments Protocol (IDAP; Newmann, Bryk, & Nagaoka, 2001; Newmann, Lopez, & Bryk, 1998); and various iterations of the Instructional Quality of Assessment (IQA) both within and outside of the National Center for Research on Evaluation, Standards, & Student Testing (CRESST) by Matsumura and her colleagues (Junker et al., 2006; Matsumura, Garnier et al., 2002). In both of these protocols, teachers select assignments of varying level (i.e., typical or challenging) and provide basic contextual information about the assignment in a cover sheet or during a brief interview. The protocols also assess the student work produced by these assignments.

**Scoring**

1.1 The scoring rule is appropriate.

The rubrics for evaluating the intellectual demand of teacher assignments were developed by researchers who based their selection of criteria on their own previously published work. Much of this previously published work had already been through the peer review process and was a part of the larger literature on teaching quality. In addition to the rubrics being based on existing scholarship, subject matter experts assisted in constructing sections of the protocols, such as the IDAP’s standards for student work in writing and math (Newmann et al., 1998). Separate rubrics for each grade and subject were used (Junker et al., 2006; Newmann et al., 2001). Rubrics for evaluating student work were based on state academic standards or district-provided tools for academic assessment.

In both protocols 3- or 4-point scales are used. The IDAP consists of sample scoring rubrics for separate sections of the protocol. Criteria for assigning a point value are provided, based on the format and nature of the task the assignment entails. For example, fill-in-the-blank or multiple choice questions are scored lower than assignments that call for generalization and support of key concepts. Moderate to high internal reliability (measured by Cronbach’s alpha) was reported (Matsumura, Garnier et al., 2002; Newmann et al., 1998). In one IQA validation study, scales were checked for redundancy and consolidated (Aschbacher, 1999).

1.2 The scoring rule is applied accurately and consistently.

A hierarchical-linear analysis was calculated to evaluate rater reliability for the IDAP. These ratings were high (Newmann et al., 1998) and scores were adjusted for potential rater bias (Newmann et al., 2001). Raters had moderate agreement, as shown by percentages, but there were no reports of statistical significance testing for rater agreement to show whether the level of agreement was above chance. Scores on the IDAP protocol were adjusted for relative severity of scorers and standards across three years (Newmann et al., 2001).
Rater drift was not addressed in great detail in the IQA studies. Cronbach’s alpha was calculated to measure consistency of ratings at the teacher level and indicated high consistency within each assignment for each dimension. Inter-rater reliability for the IQA was moderate (Aschbacher, 1999; Matsumura, Garnier et al., 2002). Cohen’s kappa's were calculated for level of agreement between all raters in each group on each dimension and it was found that the level of agreement was not always above chance (Matsumura, Garnier et al., 2002), perhaps due to differences between novice and experienced raters. Double scoring was used either for the entire set of artifacts or for 20% of the sample (Matsumura, Patthey-Chavez, Valdes, & Garnier, 2002).

Problems were encountered in the Aschbacher (1999) study of the IQA, in which inter-rater reliability could not be established between bilingual raters in rating elementary school essays written in Spanish. Consequently, they were excluded from the final analysis. The researchers concluded that the rubrics needed tighter definitions with corresponding anchor papers in addition to more extensive rater training with specialization in grade level and/or type of assignment.

1.3 The scoring is bias free.

Although studies did not provide elaborate detail on training methods, in several cases, minimal variance occurred between raters, which suggested that training adequately prevented large discrepancies in raters’ judgments (Aschbacher, 1999). In other instances, there seemed to be raters that continued to be too stringent or too lenient, even after training. In the largest study involving IDAP, a Rasch analysis was executed to adjust for rater bias (Newmann et al., 1998).

Raters for IQA and IDAP varied by study. Most often they were teachers (or former teachers) or researchers. For example, many of the IQA studies used CRESST researchers as raters, or current/former classroom teachers (Matsumura, Garnier et al., 2002). In one IDAP study, teachers from local non-participating schools were trained to score the quality of teacher assignments and student performance over the course of three days. It is unclear whether this introduces systematic biases. To reduce any systematic biases that may result from having the same rater(s) score a particular set of assignments from one school or classroom type, both IDAP and IQA randomly assigned raters to assignments and each assignment was double scored.

1.4 The data fit the scoring model.

There was moderate evidence that the scoring models used fit the data. For example, on the IQA, researchers found that the three teacher assignment scoring dimensions were correlated with each other as predicted by theory. Further, the combination of all three dimensions was most significantly associated with student achievement, suggesting that the scoring scheme was operating as intended (Matsumura, Garnier et al., 2002).

Generalization

2.1 The sample adequately represents the quality of all lessons, or artifacts, or collections of artifacts.
We located a few generalizability studies that analyzed how many assignments and raters are required for a reliable estimate of artifact quality (Aschbacher, 1999; Matsumura, Garnier et al., 2002). Aschbacher (1999) found that two raters and six assignments provided sufficient reliability for the IQA. If all dimensions of the ratings are used, then only three artifacts are necessary. In a subsequent study, researchers found that two raters and three assignments provided a stable estimate of assignment quality at the secondary, but not elementary level (Matsumura, Garnier et al., 2002).

While we do know something about how many assignments and pieces of student work are necessary for a stable estimate of assignment quality, much remains unknown. For example, how do assignments that are a part of these protocols compare to all the assignments a teacher might assign? These protocols require the assignments and student work to be written. This allows them to be scored using the scoring systems described earlier. This necessity limits the types of assignments that can be considered. For example, an assignment that requires students to give a speech would not be eligible for inclusion in these protocols. PowerPoint presentations, science experiments, and poster displays are all assignment types that would not be eligible for inclusion. We know very little about how the exclusion of these types of assignments and students work bias these estimates.

2.2 The sample is sufficiently large that random error is controlled for.

Through multiple regression analyses, research designs included control for rater and score variation, unequal test forms, socioeconomic variance, and prior achievement (Aschbacher, 1999; Newmann et al., 2001). Random error in scoring these artifacts has not been widely reported by researchers. It is not common practice to report confidence intervals around total or domain level scores either.

Extrapolation

3.1 The score on all lessons, artifacts, or collections of artifacts is related to the teaching quality an individual is able to enact.

Like observation ratings, researchers working with artifact sources found evidence to support the relationships between scores on artifacts and other aspects of teaching quality. For example, ratings on both protocols correlated highly with higher ratings on students classroom work in mathematics and writing. Protocol ratings correlated moderately with students’ gains on standardized achievement test scores in reading, math, and writing (Matsumura, Garnier et al., 2002; Matsumura & Pascal, 2003; Newmann et al., 2001; Newmann et al., 1998). For example, the quality of secondary teacher assignments predicted 19% of the variance in adjusted reading achievement outcome scores for students (Matsumura, Garnier et al., 2002).

In order to make the inference from the artifact quality to the quality of teaching, we need evidence that there was a relationship with broader subsets of teaching quality. For at least the IQA, there is evidence that the quality of classroom assignments was associated with the quality of observed instruction (Matsumura & Pascal, 2003). Though this is supportive evidence, additional research that documents the relationship between artifacts and other markers of teaching quality is needed.
Further, it is not always clear what sources teachers utilize in creating instructional materials. They may have originated on the Internet, been provided by other teachers, or taken from the textbook. Therefore, teachers may be limited or misrepresenting their individual teaching performance in what they offer as evidence of teaching quality. Newmann et al. (2001) report no significant correlations between the quality of teacher’s assignments and the socioeconomic composition of the classroom or the level of student’s prior achievement. They conclude: "Variations in assignment quality were apparently more a function of teachers’ dispositions and individual choices, than of any of the characteristics commonly used to describe students' capacity of students to engage challenging academic work" (p. 25). Concentrated professional activities outside of the classroom may contribute to enhancement of teacher choices in selecting instructional materials. This issue must be better understood before artifacts are used as a high-stakes measure of teaching quality.

One important consideration that bears on the extrapolation from artifact scores to levels of teaching quality concerns the role of teachers’ expectations. Aschbacher (1999) postulates that these protocols serve as an attempt to record the relationships between teacher expectations and student achievement. As described earlier, student learning was positively correlated with demanding tasks, but it is unclear if this is the true cause or if teacher expectations outside these artifacts is the major factor. The interpretive argument presumes that higher scores on artifact protocols reflect the quality inherent in those artifacts and higher artifact quality is positively related to teaching quality. If however, artifact quality also detects teacher expectations, and it is those expectations that support students’ learning, not the cognitive demands of the assignments themselves, we would need to know this. Such a scenario would require further empirical studies and refinement of the validity argument for artifacts.

A final consideration of the relationship between artifact scores and teaching quality concerns the students for whom the protocols can be used. Specifically, there has not been much development of instructional artifact protocols for special education students, who were excluded from the sample in one study (Matsumura, Garnier et al., 2002). Although bilingual raters were utilized in several studies, reliability was not adequately established for assessing student work written in Spanish. Teaching special education students well is certainly an important part of teaching quality. As such, the instruments would benefit from development that would allow them to be used for that population of students.

3.2 There are not systematic errors that undermine the extrapolation to teaching quality.

There is some suggestion that the extrapolation from artifact quality to teaching quality may have certain problems. Specifically, there is some evidence that teachers viewed teaching quality somewhat differently than their state standards, upon which the artifact scoring rules and the state achievement tests are based. For example, during interviews in one study, teachers reported that they had different expectations for student work than those included in the state learning goals for these assignments (Aschbacher, 1999). Therefore, rubrics assessing the clarity of learning goals and grading criteria may be compromised by implicit or unstated teacher expectations. This may introduce systematic error into the extrapolation inference.
Implication

4.1 The implications associated with teaching performance are appropriate.

If ratings of instructional artifact quality continue to be validated against student standardized test scores, these protocols will result with major implications for school accountability or teachers in programs that are piloting merit pay scales or other incentives.

Few longitudinal studies have been conducted to analyze the stability of teacher scores over time. Newmann et al. (2001) controlled for cross-year drift in assignment scoring as they found large differences in ratings from year to year. Inconsistent scoring across years further compounds the current problem of knowing whether teachers’ scores are stable over time. The lack of findings in this area greatly limits inferences of low, middle, or high teaching performance made from artifact scores. Although there are no studies that show an increase in IQA or IDAP scores after professional development intervention, Quint, Akey, Rappaport, & Willner (2007) found significant relationships between teachers’ frequency of professional development, their value placed towards it, and their ratings on the IQA.

4.2 The properties of the observed scores on the lessons, artifacts, and collections of artifacts support the implications associated with the judgments of teaching performance.

We have mentioned that some studies have used a wider set of artifacts than only assignments and student work. The reliance on a wider set of artifacts will increase support for the inference from artifact score to the domain of teaching quality. There are, however, some general features of these protocols that constrain the types of inferences we should make from scores on artifacts. Artifact protocols have been developed mainly for the areas of math and writing rather than other content areas such as history, art, or science. The nature of high-quality assignments in these other subject areas may significantly differ. In addition, researchers will need to continue to use student outcome measures other than standardized test scores for subject areas that are not as widely tested as math and English-language arts.

Trait underrepresentation is the most serious threat to the validity of using instructional artifacts to make inferences about the teacher quality domain. Teachers choose which assignments to provide for rating in these protocols. This could result in teachers selecting or designing their student assignments to be scored well on these protocols, with the consequence of artifact quality becoming less representative of teaching performance. Researchers should further investigate the impact of assignment format on ratings of artifact quality. The directions for types of assignments teachers are to include should be refined and less subjective. Collected artifacts should represent various kinds of tasks students complete. Lastly, artifact protocols need to be tailored to fit assignments prepared for student populations, such as special education students or English language learners.

Artifact protocols provide a rich source of evidence in making inferences about the quality of teaching performance. The field contains more research on observation protocols, some of which consider artifacts on a small scale, yet artifact protocols offer additional perspectives on student learning that observation data cannot provide. However, artifact protocols are limited in their ability to provide holistic information about teacher performance. The IQA and IDAP are
touchstones for future development efforts. These instruments have exhibited sufficient internal reliability and have been correlated moderately with student academic achievement. Although more is to be learned about the impact of artifacts, these protocols allow for a close view of classroom practices that significantly influence student learning and achievement, including non-academic outcomes, such as their participation in a democracy or success in acquiring a job.

Instructional Collections

Instructional collections consist of protocols which assess teaching practice using a range of artifacts, materials, reflections, and observations that were created both for regular classroom instruction and for the assessment itself. Evidence in this section comes from research on the Instructional Quality Assessment (IQA) toolkit considered as a whole (Junker et al., 2006; Matsumura et al., 2006); the Scoop Notebook (Borko et al., 2007; Borko, Stecher, Alonzo, Moncure, & McClam, 2005). National Board Certification (NBC) studies pertaining specifically to the portfolio component (e.g., Cantrell, Fullerton, Kane, & Staiger, 2008; Hakel, Koenig, & Elliott, 2008; M.Y. Szpara & Wylie, 2007) portfolio assessments for beginning teachers based on the INTASC standards, as explored by Moss and colleagues (2004); and district-level portfolio-based evaluations as described by Tucker, Stronge, Gareis, and Beers (2003) and Attinello, Lare, and Waters (2006).

The IQA toolkit consists of four main components: (1) classroom observation, (2) student interviews, (3) teacher interviews, and (4) teacher-generated assignments, with corresponding examples of student work (Junker et al., 2006). The Scoop Notebook is comprised of a large array of classroom materials, including those generated prior to class (handouts, scoring rubrics), during class (writing on the board, student work), and after class (student homework and projects). Also included are pictures of classroom layout, equipment, and board work that cannot be included in the notebook, and teacher responses to reflective questions about each lesson. The NBC portfolio requires video documentation of instruction and teacher-student interaction, samples of student work, and detailed reflections and analysis of all materials included. Three of these entries must be classroom-based, while a fourth should reflect the teacher’s work outside the classroom with families, colleagues, and the community (2008 Guide to National Board Certification, 2008). Other portfolios discussed in this section include similar collections of materials, such as 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.

Thus, instructional collections can consist of a wide range of materials, mediums, and sources of evidence to capture a comprehensive and holistic view of teaching practice. In theory, instructional collection protocols can be adapted to any subject matter or grade level; however, scoring standards often vary by subject or grade. For example, IQA rubrics have been created specifically for math and reading comprehension, the Scoop Notebook has rubrics for math and science, and NBC consists of over 25 different certificates for various subjects and grades.

Scoring

1.1 The scoring rule is appropriate.
Documentation of the scoring rubric’s development is one way to determine the appropriateness of the scoring rule. All articles provided some account of the criteria on which scoring rules were based, and development generally included thorough reviews of the research and involvement of experts in the field. Sometimes rubrics were developed out of long-standing research programs that examined measures of teaching (Borko et al., 2005; Junker et al., 2006). The NBC portfolio is also known for carefully developing content standards in each of their specialty areas, using committees of expert teachers and specialists in education, child development, and other relevant areas, and soliciting public feedback on standards once they are developed. The other studies examine portfolio assessments as they have been implemented in a state or district, and include less information about how the scoring rubrics were developed but usually do explain how good teaching was conceived and defined (Attinello et al., 2006; Moss et al., 2004; Tucker et al., 2003).

Scales and rubrics were generally created to be consistent with other prominent standards in education. For example, the Scoop Notebook is meant to measure reform-oriented teaching practices as outlined by the National Science Education Standards and the Principles and Standards for School Mathematics (Borko et al., 2005), the IQA standards are based on the Principles of Learning (Resnick & Hall, 2001; cited in Junker et al., 2006), and many portfolio assessments are aligned with the INTASC standards, which are meant to be coherent with National Board standards (Moss et al., 2004). This consistency with other consensus documents suggests the scoring rules are aligned with the aspects of teaching quality most important to a range of constituents.

The Scoop Notebook, IQA toolkit, and NBC use numerical scales that can be analyzed statistically to ensure consistency, while the other portfolio assessments were scored more qualitatively and holistically. IQA researchers examined the internal consistency of scales using intra-class correlations and Cronbach’s alpha, finding excellent consistency for math (0.92) and lower but still adequate consistency for reading comprehension (0.76; Matsumura et al., 2006). Scoop Notebook researchers conducted factor analysis, revealing a dominant factor underlying their 10 dimensions of reform teaching. The variance accounted for by this dominant factor was moderate and higher for math (56 - 57%) than for science (42 - 49%; Borko et al., 2007). Hakel et al. (2008) report that for the NBC portfolio scoring, the Board conducts regular assessments of internal reliability, and finds that results are consistent with expectations for large-scale portfolio assessments but “low relative to generally accepted standards for high-stakes assessments” (p. 101-102).

1.2 The scoring rule is applied accurately and consistently.

The IQA and Scoop Notebook scales were designed to be low-inference, so that raters without special expertise could be reliably trained to use them. These instruments are still being refined to accomplish this (e.g., Junker et al., 2006). In contrast, the NBC and other portfolio protocols are high inference, requiring scorers with specific content knowledge and experience (e.g., 2008 Guide to National Board Certification, 2008; Tucker et al., 2003).
Interrater agreement is addressed in all the protocols using a quantitative rubric. One IQA study reported interrater agreement ranging from 81 - 86% on the observation rubrics and 71 - 76% on the assignment rubrics, an improvement from previous pilots due to enhanced rater training and use of raters with more content expertise (Matsumura et al., 2006). The Scoop Notebook field tests found adequate to high levels of interrater reliability; agreement within 1 point (on a 5-point scale) generally exceeded 75% and was similar across math and science notebooks (Borko et al., 2007). The NBC process includes double-scoring procedures and assessments of interrater agreement each year to ensure reliability of portfolio scores. A federal evaluation presented by Hakel et al. (2008) reports that NBC reliability analyses consistently indicate reliable, accurate, and fair scoring. Interrater reliability was generally not reported for the portfolios using more qualitative scoring, and when it was, the description was minimal (e.g., Attinello et al., 2006). This appears problematic, given that reliability for qualitative scoring tends to be more difficult to establish.

There is far less information about rater consistency over time. The IQA studies looked at this, and determined that “naïve” raters (graduate students in education with no prior knowledge of the assessment) continued to improve in consistency over the course of the study, indicating that the training period needed to be longer and include more opportunities for “live” scoring (Junker et al., 2006). NBPTS is also known for paying careful attention to rater training and calibration (Hakel et al., 2008). However, measures of rater consistency over time and how rater drift and recalibration are handled were not described by the articles reviewed.

1.3 The scoring is bias free.

Most of these studies describe some aspect of rater selection and training. Raters tended to be members of the research team, graduate students, or teachers familiar with the content areas being evaluated. Portfolio raters were usually teachers or administrators with subject matter expertise. NBC portfolio assessors are typically accomplished teachers with content expertise, who often have obtained NBC themselves (2008 Guide to National Board Certification, 2008). Little is said about whether raters from these groups might score in systematically different ways than raters from other groups.

Most of the articles describe how raters were trained, but since many are exploratory, they have not yet established what training procedures are needed to achieve adequate rater accuracy. Raters of the Scoop Notebook are required to write a justification for each score they give, which may serve to improve accuracy and detect systematic biases (Borko et al., 2005). However, if these protocols are to be used for high-stakes assessment, it is important to establish whether rater bias exists, to understand rater training issues, and to determine the amount and type of training necessary to ensure accuracy.

The exception is NBC, which includes an extensive training program that addresses bias directly. Assessors are required to participate in bias reduction training, so that they can identify and monitor potential sources of error such as personal bias, writing bias, and societal bias. A study by Szpara & Wylie (2005) determines that bias awareness exercises included in the training are
effective in reducing scorer bias. In addition, Hakel et al. (2008) find that the Board is extremely diligent in addressing issues of fairness, particularly as they pertain to performance differences between genders and racial/ethnic groups.

1.4 The data fit the scoring model.

Explicit descriptions of scoring models were not presented in these articles. Complex models such as item response theory (IRT) were not described. For instance, scores on the IQA and Scoop Notebook tended to be simple ratings that were aggregated to create composite scores. Differential weighting of dimensions was not generally discussed, although these studies did sometimes report examining reliability and consistency separately by dimension or subscale (Borko et al., 2007; Junker et al., 2006). NBC does weight the various components of the portfolio differently (Cantrell et al., 2008), however we did not find documentation of what those weights are and how they were determined.

Generalization

2.1 The sample adequately represents the quality of all lessons, or artifacts, or collections of artifacts.

Reports on the IQA toolkit do a good job of documenting generalizability studies to determine the number of assignments and observations needed to yield a stable estimate of instructional quality. Pilot studies indicated that two observations were enough to yield a stable estimate of teaching performance, with dependability coefficients of 0.80 for reading and 0.86 for math. However, this was only the case if teachers complied fully with data collection requirements; otherwise, an extra observation may be needed. Collecting four assignments produced stable estimates of quality for both content areas, with dependability coefficients around 0.8, and decision studies indicated that collecting three assignments may also be sufficient. These analyses allowed IQA researchers to determine, for instance, that requesting two “challenging” and two recent assignments produced greater variability and lower dependability coefficients than requesting four assignments that were all “challenging.”

Moss and colleagues (2004) attempt to investigate the generalizability of portfolio assessments by comparing scores on two portfolios created by the same teacher for different lessons at different points in the year. However, they found that teachers included far less detail in the second portfolio, leading researchers instead to examine whether the evidence in the second portfolio would cause scorers to reconsider their evaluation of teaching performance as determined by the first portfolio. Findings indicated that the second portfolio generally supported their impressions based on the first portfolio with respect to the scoring criteria. They also used observation, interviews, and contextual information to determine how representative the portfolio was of typical teaching performance, and found that teaching practice varied by contextual factors such as student ability level, cultural background, and subject matter taught. This lead the authors to caution that teaching performance as measured by portfolios may not be generalizable to the teacher’s performance in other classes.

Other articles do not provide this much detail on generalizability, but some describe evaluating the contents of collections to determine whether all the requested materials provided useful
information for scoring, and refining protocols accordingly (Borko et al., 2005; Tucker et al., 2003).

2.2 The sample is sufficiently large that random error is controlled for.

Most of these studies are conducted with very small sample sizes, and some protocols have not been developed fully enough to determine how scores operate on a larger scale. We did not find any information about how much random error is reasonable to expect in scores on instructional collections. In general, it is not common practice to report confidence intervals around total scores.

Extrapolation

3.1 The score on all lessons, artifacts, or collections of artifacts is related to the teaching quality an individual is able to enact.

The studies we examined present evidence of convergent validity with observational measures, and to a lesser extent, student achievement measures. Ratings on the Scoop Notebook were found to correlate with observational ratings, with average correlations ranging from 0.71 - 0.85. Authors concluded that this represents a “considerable degree of convergent validity between broad judgments of the extent of reform-oriented practice in individual classrooms made on the basis of notebooks and observations” (Borko et al., 2007, p. 45).

Evidence of convergent validity with other measures of teaching practice and with student outcome measures is less common for instructional collections. One exception is the IQA pilot work, which correlates observation and assignment ratings with student achievement on the math and reading comprehension subscales of the SAT-10. Results indicated that IQA observation and assignment ratings on both math and reading comprehension rubrics have the potential to predict related scores on the SAT-10. However, the results are only suggestive because the sample size was too small to conduct multilevel modeling (Matsumura et al., 2006).

Many studies have also correlated NBC status with student achievement, often finding that students of NBC teachers exhibit higher achievement than students of non-NBC teachers. But overall results are mixed, and studies are plagued by issues of selection bias (Hakel et al., 2008). Most of these studies consider NBC as a whole, thus we do not consider them in our analysis since they do not reveal whether NBC portfolio scores specifically relate to student achievement. One study by Cantrell and colleagues (2008) does attempt to break this out, examining whether subscores from the different components of the NBC assessment can predict student achievement on standardized tests in math, writing, and language arts. They describe complex relationships between subscores and achievement scores, finding for instance that video scores from the portfolios provide more predictive information than teacher commentary on student work.

Portfolio studies are more likely to provide self-report data indicating that teachers and administrators find portfolios to be an accurate and comprehensive portrayal of teaching performance, and a useful exercise for improving practice. However, there is little evidence that directly measures how portfolios relate to changes in teaching practice or to student outcomes. For example, Tucker et al. (2003) examined whether portfolios were perceived as useful for
professional growth, finding mixed results. They note that while teachers indicated feeling the portfolios helped them improve their practice, little impact on teaching practices was actually observed. Thus, work that more directly explores these relationships is needed.

Support for the relationship between scores on an instructional collection and overall teaching quality can also come from divergent validity evidence. Both the IQA and Scoop Notebook document that the protocols are sensitive to variations in teaching when that should be the case. For instance, the Scoop Notebook revealed expected differences between teachers in California and Colorado, where Colorado was known to have a more reform-oriented curriculum than California (Borko et al., 2007). The IQA pilot examines score differences between two districts, one of which was more familiar than the other with the Principles of Learning on which the rubrics are based, and they found the expected significant differences (Junker et al., 2006). This evidence suggests these instruments are able to detect differences that are theorized to be important aspects of the instructional collection protocols.

3.2 There are not systematic errors that undermine the extrapolation to teaching quality.

One critical issue with any measure of teaching quality is the degree to which it overlaps with the larger teaching quality construct. Attinello et al. (2006) and Tucker et al. (2003) provide evidence that teachers and administrators tend to perceive portfolios as a well-rounded and generally accurate portrayal of instruction, though this is evidence based on self-report. The developers of the Scoop Notebook took care to analyze how well the various notebook components functioned in practice. They refined their artifact collection procedure after systematically examining the completeness and usability of artifacts included in the notebooks (Borko et al., 2005). Similarly, Tucker and colleagues (2003) examined the extent to which portfolio materials were able to document all the aspects of teaching they were trying to rate. They note that the portfolios exhibited high construct validity and provided a useful augmentation to evaluation because they included information not easily observed by administrators, such as documentation of professional activities and assessment practices. Thus, evidence is encouraging that instructional collections can document a wide range of teaching practices that reflect the larger domain of teaching quality.

**Implication**

4.1 The implications associated with teaching performance are appropriate.

As mentioned in the scoring section, the rubrics and standards for these instructional collections were carefully developed based on widely accepted views of good teaching, and are congruent with the standards of similar instruments. Several studies indicate that these protocols have high face validity, suggesting that the collections are seen as appropriate measures of teaching performance. Instructional collections are generally considered holistic and complete portraits of practice, and are often applicable to multiple grades and subject areas. NBC candidates tend to report that the portfolio process was beneficial and contributed to improvements in their practice (Hakel et al., 2008). However, caveats are certainly noted. For example, Attinello et al. (2006) found quantitative support for the accuracy of portfolios, but qualitative reports revealed concerns that portfolios do not reflect all aspects of practice, and that a portfolio’s presentation may not accurately portray what actually occurs in a classroom. Teachers in the Scoop
Notebook study cautioned that artifact collections could not adequately reflect classroom interactions such as the quality of student-teacher discourse (Borko et al., 2007).

How these protocols are used is extremely important in determining whether the implications associated with their scores are appropriate. Because instructional collections have not yet moved far past the stages of establishing consistent scoring procedures and testing convergent validity with other measures of teaching quality and student outcomes, it is generally recommended that they are used formatively or summatively in low-stakes situations, and not as a sole or primary component of high-stakes decision making.

The IQA and Scoop Notebook were developed with both summative and formative uses in mind, however they both need more large-scale, external testing before being used for high-stakes purposes. Portfolios, such as those based on the INTASC standards, are used by some states to make decisions about beginning teachers (Moss et al., 2004), and these are the studies for which we found the least validity evidence to support such a use. Portfolio assessments have been used primarily with preservice teachers, thus this evidence may be prevalent in the preservice literature, but it is disconcerting that we did not find much research on the validity of portfolios for making decisions about in-service teachers. Protocols that have been validated for use with preservice teachers are not necessarily appropriate to use with in-service teachers. NBC status is also used to make consequential decisions about teacher salaries and bonuses, and while there is evidence that the standards, reliability, and consistency of the NBC portfolio are reasonable, issues of convergent validity particularly with student achievement are still being debated (Hakel et al., 2008).

4.2 The properties of the observed scores on the lessons, artifacts, and collections of artifacts support the implications associated with the judgments of teaching performance.

Many studies attempt to sample diverse groups of teachers and students to ensure that scores do not reflect systematic sources of error. Most of these studies tried to account for factors such as student demographic characteristics (e.g., race/ethnicity, socioeconomic status, gender), school location (urban/suburban/rural), student ability, teacher experience, grade level, subject matter, curriculum, and teaching style when conducting their analyses. However, sample sizes were small, studies were often exploratory, and no studies looked at all these factors at once. Thus it is hard to determine whether any of these scales capture construct-irrelevant variance. We also found no evidence on whether scores remain stable over multiple years and throughout different points in a teacher’s career trajectory.

The notable exception is NBC, which focuses heavily on fairness and has been evaluated to detect sources of construct-irrelevant variance. One example is a study by Szpara and Wylie (2007) which investigates whether differences in writing style between African American and European American NBPTS applicants affect scorer’s judgments about their teaching performance. NBC training specifies that assessors should judge applicants on the basis of the teaching evidence presented, not on the form of the presentation, and assessors are trained to identify and monitor writing biases. Findings indicated that while writing differences were present between the two groups, they did not affect scoring. Likewise, Hakel et al. (2008) report that although systematic differences exist between applicants from different backgrounds (e.g.,
African American candidates show lower average passing rates than European Americans), research suggests that these differences do not result from biases in the assessment.

**Summary Trends on Instructional Collections**

Instructional collection rubrics tend to be well-developed and consistent with prevailing conceptions of high quality teaching. They are considered comprehensive and holistic representations of teaching performance, and generally enjoy high levels of face validity with stakeholders. However, validity evidence is much weaker in some of the other areas. For instance, scoring procedures are not always consistent and free of bias, evaluations of performance may not generalize beyond specific occasions, and implications associated with scores may not always be appropriate.

This review serves to reveal major gaps in the literature and the research field which need special attention. In the case of instructional collections, accounts of rater agreement and training procedures were usually present, but descriptions were thin and rarely discussed how potential rater bias was addressed. Evidence on consistency of scoring over time was scant and there were no descriptions of how to handle rater drift and recalibration. There were no explicit discussions of scoring models and how the data fit those models. There were some attempts to establish generalizability, but no studies fully examined score stability across lessons, content areas, grade levels, time of day, time of year, etc. No studies reported the amount of random error associated with scores. There was limited evidence on how scores might vary across different teachers in different contexts, and no studies discussed the stability of scores over several years and at different points in a teacher’s career.

While it is admittedly difficult to examine all of these factors, this work highlights particular areas in need of increased attention, such as issues of rater consistency, training, and bias. It also emphasizes that evidence should be documented and made available when it exists. The IQA and Scoop Notebook reports do a good job of this. In the case of a prominent, large-scale assessment like NBC, it is assumed that adequate measures have been taken to ensure the validity and reliability of protocols, however technical documentation of that work is often not made explicit or readily accessible. There should be more demand for access to this type of documentation. There should also be more external evaluations of these protocols, as several of them have only been tested by their developers.

**Discussion**

This review suggests that there have been some important validation efforts in each of the evidentiary areas examined. There is however, a lot of work left to do.

For each of the evidentiary sources, there are areas of significant weakness in the validity argument. These areas would benefit from immediate attention. In the case of observation instruments, we must know much more about how many times to observe, what the length of the observations should be, and when those observations should happen over the day and over the school year. Further, for teachers that teach multiple subjects (e.g., elementary teachers), we need to know more about what subjects should be observed and how those scores should be
combined. In the case of artifacts, one area of weakness concerns the relationship between scores on the instrument and levels of teaching quality. There are very few protocols in use and they can only be used in classrooms that teach mathematics and English/language arts. Scores on these instruments have not been linked to broader notions of teaching quality. And the instruments look narrowly at assignments and student work. In order to strengthen the validity argument for artifacts, additional work will need to document how these protocols relate to larger areas of the teaching quality domain. Finally, in the case of instructional collections, there is work to be done on how best to fairly score the instruments, train raters to reliable levels, and investigate how scores operate over time, content areas, and grade levels.

Across all three evidentiary sources, there is very little work that considers how scores on the reviewed instruments are related to important non-multiple choice academic outcomes and non-academic outcomes. There is, however, an emerging literature that links these evidentiary sources to scores on standardized multiple choice achievement tests. In all three categories – observations, artifacts, and instructional collections – there is modest to moderate evidence that the scores on the instruments are related to standardized student achievement measures. There is of course, variation by instrument and by study. We take the evidence on standardized achievement scores to suggest that these instruments may be able to detect teaching acts that contribute to student success.

One additional area for development concerns the use of instruments by non-developers. With a handful of notable exceptions, many of the instruments examined have only ever been written about by their authors. While this is reasonable and appropriate at the beginning of a development effort, in order to develop a robust validity argument, it is important for the field to systematically investigate the full range of an instrument’s interpretive argument.

Conclusion

The current policy environment presses districts to measure teaching practice in valid ways. This review suggests that relying on one evidentiary source to make claims regarding teaching performance is unwise given the current state of the field. There was no instrument in any evidentiary source that had a sufficiently strong validity argument that it could be used alone for high-stakes teaching decisions. Further, each evidentiary source struggles with particular measurement issues that would need to be addressed if an instrument that makes use of that type of evidence was going to be used in high-stakes situations. For example, guidance regarding the number, length, and timing of observations must be clearer for a particular observation protocol before that protocol is used for high stakes decisions. Finally, as with any form of assessment, specific evidentiary sources are neither good nor bad. There are certain observation protocols with stronger validity arguments; there are others with weaker ones. The final decision to include a particular protocol as part of a high-stakes decision will have to rest on the validity argument for that protocol.
## Appendix

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Citation</th>
<th>Grade level</th>
<th>Subject area</th>
<th>Constructs measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danielson’s Framework for Teaching (FFT)</td>
<td>Danielson (1996)</td>
<td>K-12</td>
<td>General</td>
<td>Planning and preparation, classroom environment, instruction, and professional responsibilities</td>
</tr>
<tr>
<td>Classroom Assessment Scoring System (CLASS)</td>
<td>Pianta, La Paro, &amp; Hamre (2007)</td>
<td>K-5</td>
<td>General</td>
<td>Emotional support, classroom organization, and instructional support</td>
</tr>
<tr>
<td>Classroom Assessment Scoring System -Secondary (CLASS-S)</td>
<td>Pianta, Hamre, Haynes, Mintz, &amp; La Paro (2007)</td>
<td>6-12</td>
<td>General</td>
<td>Emotional support, classroom organization, and instructional support</td>
</tr>
<tr>
<td>TEX-IN3</td>
<td>Hoffman, Sailors, Duffy, &amp; Beretvas (2004)</td>
<td>K-3</td>
<td>English/language arts</td>
<td>Text environment, text use, and text understanding</td>
</tr>
<tr>
<td>Mathematics Enhancing Classroom Observation Recording System (MECORS)</td>
<td>Muijs &amp; Reynolds (2003)</td>
<td>K-5</td>
<td>Mathematics</td>
<td>Classroom management, time on task, individual and group practice, interactive teaching, active teaching, instructional practices, classroom climate</td>
</tr>
<tr>
<td>Title</td>
<td>Authors/References</td>
<td>Grades</td>
<td>Content Areas</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td>--------</td>
<td>---------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Intellectual Demand of Assignments Protocol (IDAP)</td>
<td>Newmann, Lopez, &amp; Bryk (1998)</td>
<td>3-8</td>
<td>English/language arts, Mathematics</td>
<td>Design and grading of student tasks</td>
</tr>
<tr>
<td>Instructional Quality Assessment (IQA)</td>
<td>Aschbacher (1999)</td>
<td>2-8</td>
<td>English/language arts, Mathematics</td>
<td>Design and grading of student tasks</td>
</tr>
<tr>
<td>National Board Certification portfolio</td>
<td>Hakel, Koenig, &amp; Elliott (2008)</td>
<td>PK-12</td>
<td>General (25 specialty areas)</td>
<td>Commitment to students and their learning; knowledge of the subjects taught and how to teach those subjects to students; responsibility for managing and monitoring student learning; systematic thinking about practice and learning from experience; being members of learning communities.</td>
</tr>
</tbody>
</table>


