Key Considerations When Measuring Teacher Effectiveness:

A Framework for Validating Teachers' Professional Practices

Carole Gallagher, Stanley Rabinowitz, and Pamela Yeagley

Assessment and Accountability
Comprehensive Center

AACC • A WestEd and CRESST partnership

AACC: Assessment and Accountability Comprehensive Center: A WestEd and CRESST partnership.
aacompcenter.org

The work reported herein was supported by WestEd, grant number 4956 s05-093, as administered by the U.S. Department of Education. The findings and opinions expressed herein are those of the author(s) and do not necessarily reflect the positions or policies of AACC, WestEd, or the U.S. Department of Education.

To cite from this report, please use the following as your APA reference:
Gallagher, C., Rabinowitz, S., and Yeagley, P. (2011). *Key considerations when measuring teacher effectiveness: A framework for validating teachers' professional practices* (AACC Report). San Francisco and Los Angeles, CA: Assessment and Accountability Comprehensive Center.

The authors thank the following for reviewing this policy brief and providing feedback and recommendations:
Ron Dietel (CRESST); and for editorial and design support: Tamara Lau and Karisa Peer (CRESST).

Key Considerations When Measuring Teacher Effectiveness: A Framework for Validating Teachers' Professional Practices

Carole Gallagher, Stanley Rabinowitz, and Pamela Yeagley

Researchers recommend that policymakers use data from multiple sources when making decisions that have high-stakes consequences (Herman, Baker, & Linn, 2004; Linn, 2007; Stone & Lane, 2003). For this reason, a fair but rigorous teacher-effectiveness rating process relies on evidence collected from different sources (Goe, Bell, & Little, 2008; Center for Educator Compensation Reform [CECR], 2009; Domaleski & Hill, 2010; Economic Policy Institute [EPI], 2010; Little, 2009; Mathers, Oliva, & Laine, 2008; National Comprehensive Center for Teacher Quality [NCCTQ], 2010a; Steele, Hamilton, & Stecher, 2010). Yet policymakers must take into account that (a) certain types of information are more trustworthy than others for the purposes of measuring teacher effectiveness and (b) the availability of technically sound data varies across content areas, grade ranges, states, districts, and schools.

Currently, a key source of data for measuring teacher effectiveness is statewide achievement testing. Because state tests are associated with stringent technical adequacy expectations in terms of validity, reliability, and fairness and are administered in a standardized fashion, they are considered a trustworthy source of information (i.e., they provide "Level 1" data) about the effectiveness of teachers' instructional practices (Linn, 2008; Toch & Rothman, 2008). Results from statewide testing can be particularly useful when statistical analyses of growth (e.g., value-added modeling) are used to determine a teacher's unique contribution to student learning during one grade or course (Braun, 2005; Goldhaber & Hansen, 2010; Hanushek & Rivkin, 2010; Harris, 2009; Kane & Staiger, 2008). Yet, to date, these measures are typically available only for a subset of teachers: those who teach English language arts (ELA), mathematics, or science at certain grades. Other sources of information—those that provide "Level 2" data—may be available for teachers in all grades and content areas but their trustworthiness for measuring teacher effectiveness may be uncertain and/or their use may require allocation of additional resources to ensure that the data can be collected systematically in all classrooms. A third category of data sources—those that provide "Level 3" data—yields richly descriptive information about instructional practices that may be available for all teachers in all grades and content areas, but does not bring the level of technical adequacy necessary for judgments about teacher effectiveness; data from these sources may be more appropriately used as *supplements* to Level 1 or Level 2 data. Because valid determinations of teacher effectiveness (or school- or classroom-level accountability) should focus on student learning and other valued outcomes, decision-makers will want to access all available Level 1 data and then consider how best to combine these data strategically with other types of information from Levels 2 and 3. Doing so holds real promise as a means for fairly validating the professional practices of all teachers.

This report is intended to highlight the range of data sources that can be tapped to validate teacher effectiveness. Section I describes broad considerations to support identification of those sources of information most appropriate for a specific purpose or context. Section II highlights the strengths and limitations of different types of information about teacher effectiveness, beginning with sources of Level 1 data and proceeding through typical sources of data at Levels 2 and 3. Section III offers a set of final recommendations about effective data use when measuring teacher effectiveness. State and local decision-makers are encouraged to consider all of the data options presented—and weigh possible tradeoffs associated with their use—when determining which *combination* of sources is most likely to yield the information that best meets their needs.

I. General Considerations Related to Evidence About Teacher Effectiveness

Different types of information can be useful for measuring teacher effectiveness, depending on the specific *purpose* and *context* for their use and the degree to which trustworthy data are readily *available*.

Purpose of Data Collection. Data related to teacher effectiveness may be collected for different purposes. Certain combinations or sets of information are more appropriate than others, depending on how and by whom the data will be used and what is at stake for teachers, students, and schools. It is important to note that in many cases, the reason(s) for collecting data about teacher effectiveness may change over time. Guiding questions to support informed decision-making about purpose-driven data use are provided below.

- ***High Stakes:*** Will these data be used for accountability purposes at the state, local, and/or classroom level?
- ***Medium Stakes:*** Will these data be used for decision-making related to hiring, promotion, or tenure?
- ***Low Stakes:*** Will these data be used formatively to support individual teacher growth or to inform decision-making about professional development for all staff?

Context for Data Collection. Decision-makers must consider the unique context for data collection, including review of factors related to culture, history, and policy. As the context for data collection will change over time, data collection strategies must be revisited at regularly scheduled intervals. Guiding questions to support informed decision-making about context-appropriate data use are provided below.

- How has the state or district defined teacher effectiveness? Given this definition, what types of evidence will best inform decision-making?
- Will existing or emerging national, state, and/or local policies enable or constrain this work?
- What state or local resources are available to support this work? Will existing resources support use of multiple data sources during decision-making about teacher effectiveness?
- Who will select the sources of evidence from which conclusions about teacher effectiveness will be drawn?
- What is the timeline for implementation of formal data collection procedures?
- How will consequences (positive and negative) of using these data sources for this purpose be monitored?

Accessibility of Trustworthy Data Sources. As previously stated, the availability of technically sound data varies across content areas, grade ranges, states, districts, and schools. Guiding questions are provided below to support informed decision-making about accessibility of data with different levels of technical adequacy expectations

- ***Level 1 Data:*** Are technically sound student-level assessment data readily available (either currently or in future plans) to serve as the centerpiece measure of teacher effectiveness?
- ***Level 2 Data:*** Are other assessment data (e.g., from aligned interim measures administered at the district level) available that can be used as direct measures of student learning? To what extent are these data available uniformly across the state? Can these data be collected systematically in all schools for all teachers?
- ***Level 3 Data:*** Are more readily available sources of information (e.g., observations, surveys, preservice academic history) available that can be used to supplement the more technically rigorous data from Level 1 or 2? What evidence suggests that these sources of information are sufficiently trustworthy to supplement Level 1 or 2 assessment data?

States are encouraged to think of their teacher evaluation systems in a dynamic fashion. Over time, given sufficient research and professional development, some Level 2 or Level 3 indicators may become more trustworthy and hence may shift to a higher level (i.e., up to Level 1 or Level 2, respectively). States are encouraged to use the lack of Level 1 data in some content areas to undertake research studies designed to demonstrate the adequacy of a broader range of indicators to measure teacher effectiveness. The benefits of this research can expand into grades for which Level 1 data currently are available, meeting the strong recommendation that multiple data sources should be used for any high-stakes accountability decisions.

II. Sources of Evidence About Teacher Effectiveness

To support informed decision-making about data options, the following sections provide detailed information about the following:

- ***Level 1 Data Sources:*** Student-level data was collected via standardized annual statewide tests of achievement, end-of-grade or end-of-course assessments, or customized pre-post measures of achievement.

- ***Level 2 Data Sources:*** Student-level data was collected from vendor-, district-, or school-developed measures of achievement (non-annual) or performance measures; test-based and non-test-based aggregate data (e.g., graduation rate); or data about teachers that are collected by a qualified external agency.
- ***Level 3 Data Sources:*** Teacher-level data was collected through formal or informal classroom observation; surveys or interviews with parents, students, or teachers; portfolio review; teacher-level performance assessment or performance checklists; or peer review.

Specific information about the usefulness of data at each level is provided in the following sections. For each level, a table is presented that describes the different sources of information, highlights guidelines for use, and provides references for additional information.

Level 1 Data Sources

Level 1 data can be appropriately used for high-, medium-, and low-stakes purposes when guidelines for use are heeded. Accessing Level 1 sources is especially important when the data will be used for high-stakes purposes such as accountability. This is because Level 1 measures are required to meet high standards for technical adequacy (valid, reliable, and fair) and are administered in a standardized fashion. Results from Level 1 data sources can be particularly useful when statistical analyses of growth (e.g., value-added modeling) are used to determine a teacher's unique contribution to student learning during one grade or course.

Currently, Level 1 data from statewide testing can be collected only for a small subset of teachers: those who teach English language arts (ELA), math, or science in grades 3–8 or who teach specific courses in high school (e.g., biology). For this reason, two additional options that meet the technical adequacy expectations necessary to support use for accountability purposes are presented in the following table, but allocation of additional resources may be required to allow for their development. Finally, while Level 1 data are expected to serve as the cornerstone for decision-making about teacher-level accountability, a fair and comprehensive evaluation also will take into account information from Levels 2 and 3 (e.g., principal's observations and/or findings from an external agency) to ensure that the full range of effectiveness indicators has been considered.

Table 1. Level 1 Data Sources

Level 1 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Annual Statewide Tests of Achievement (Status) (n) (e)	• Results from statewide tests that are associated with high technical adequacy expectations • Can capitalize on existing data that can be used as proxy for indicator of instructional effectiveness • Moderate correlations between student achievement and other measures of teacher effectiveness	• Not all subjects and grades have mandated tests • Tests measure only a portion of the curriculum • Tests must have technical adequacy evidence to support use for this purpose • Research suggests that teacher behavior is not the only factor influencing student learning • Researchers recommend using for high-stakes decision-making only in conjunction with other sources of evidence of teacher effectiveness	Accomplished California Teachers [ACT] (2010) Battelle for Kids (2009) CPRE (2006) Goe et al. (2008) Gordon, Kane, and Staiger (2006) Hinchey (2010) New Teacher Project (2010) REL Midwest (2007, 2008) Steele et al. (2010) Steiner (2009) Toch and Rothman (2008) Hillsborough County, FL—STAR program
Growth Via Gain Score Approach (e)	• Describes difference in one student's scores from prior year to current year (year-to-year change) • Provides a simple measure of teacher effect	• Requires capacity to link students to teachers and track students longitudinally • Does not take into consideration a student's starting point	Goe et al. (2008) NCCTQ (2010 a-e) Steele et al. (2010) Delaware Growth Model Texas Growth Index

Level 1 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• Best to use multiple years of data and, when possible, average teachers' estimated impact over multiple years • May be based on student learning objectives (SLOs) using teacher-developed instruments as well as large-scale assessments	• Gives no information on what aspects of teacher's practice were effective • Assumes that tests are aligned to instruction and that instruction is aligned to standards	
Growth Via Value-Added Modeling (VAM) or Other Type of Statistical Modeling (e)	• Provides a summary score of the contribution of a teacher to growth in student achievement; when most students in a particular classroom perform better than predicted on a standardized test, the teacher is credited with being effective • Focuses directly on student learning • Can take into account prior achievement/initial status, student characteristics (e.g., gender, race/ethnicity, free/reduced lunch status), and teacher characteristics (e.g., years of experience) • Can reveal variation among teachers in their contributions to student learning • Useful for identifying teachers who likely need professional development and/or schools that may need specific assistance • May provide evidence about which teacher characteristics and qualifications matter most for optimal student learning; useful for identifying teachers who can serve as a resource to colleagues • Especially in math, teachers' past record of value-added is among strongest predictors of students' achievement gains in other classes and across years • Teachers with high value-added on state tests also had students who were among the highest performers on measures of deeper conceptual understanding	• Requires capacity to (a) link students to teachers, (b) track students longitudinally, and (c) conduct sophisticated data analyses • Most reliable when using multi-year models and incorporating other measures in decision-making • Generally need vertical alignment of tests across grades • Susceptible to known sources of bias, depending on model used and quality of data • Findings mixed in terms of correlation between VAM estimate and other performance indicators in content areas other than math • Challenging to parse teacher effect from student and school effects as students are not randomly assigned to schools or classrooms and teachers are not randomly assigned to schools • Assumes that a teacher's effectiveness is the same for each student •Averages test scores across all students in a classroom despite wide variability in the ways in which teachers may contribute to score gains • Lack of agreement on how methodological issues (and model assumptions) affect the validity of interpretations; some are adamant about the need for vertical scales (Baker et al., 2010) while others argue against such scales (Martineau, 2006)	Baker et al. (2010) BMGF (2010a & 2010b) Braun (2005) CPRE (2006) Goe et al. (2008) Goldhaber (2010) Goldhaber and Hansen (2010) Hanushek and Rivkin (2010) Harris (2009) Hill (2009) Jacob, Lefgren, and Sims (2009) Koretz (2008) Lefgren and Sims (2010) Martineau (2006) Mathers et al. (2008) McCaffrey, Lockwood, Koretz, and Hamilton (2003) NCCTQ (2010 a-e) Rothstein (2009) Schochet and Chiang (2010) Thomas (2010) Dallas Value-Added Assessment System District of Columbia Growth Model Florida Growth Model Georgia Growth Model Hawaii Growth Model Houston Value-Added Assessment System Louisiana Value-Added Teacher Preparation Program Minneapolis Value-Added Model New York Growth Model Persistence Model Rhode Island Growth Model SAS Education Value-Added Assessment System

Level 1 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• Interpretations can be criterion- or norm-referenced • Cost efficient and non-intrusive; relies on existing data • VAM estimates most reliable when based on multiple years of data	• Need to use in conjunction with other types of information that indicate ways in which teachers might improve • Greatest risk is misclassification; however, consequences for students also must be taken into consideration (i.e., risk may be worthwhile if it increases likelihood that students will be exposed to effective teachers) • Unclear to what extent one teacher's effect persists over time; teacher-induced learning in particular has low persistence • Harris (2009) recommends normalizing test scores to a mean of zero and a standard deviation of one and assuming that tests are locally scaled (to narrow the range in which one point is considered equivalent) • Rothstein (2009) recommends keeping teacher comparisons to those whose students started at the same achievement and grade levels	School Performance Framework Tennessee Teacher Evaluation System Washington, DC, IMPACT
End-of-Grade or End-of-Course Assessments (n) (e)	• Using a student-level longitudinal tracking system, scores from standards-based, content-specific summative assessments administered at the end of grades 1–6 are compared with students' scores from the previous grade. Score gains are calculated for each student receiving instruction from any one teacher. Annual mean gains for each teacher are estimated via analytic models (e.g., value-added modeling) that take into account students' unique starting points (and other covariates, if desired) • Mean gains for each teacher can be compared to the expected annual gain for that grade and content area (criterion-referenced model) or to the gains for that teacher's peers (norm-referenced model)	• May not be cost-effective if only used for classroom accountability purposes • Measures may not have been fully validated for high-stakes purposes • Requires expert judgment in setting the criteria or standard for performance against which teachers will be evaluated (i.e., the expected annual gain for each grade and content area) and/or in determining how norm-referenced scores will be used (e.g., a ranking system)	CECR (2009) Hillsborough County, FL—STAR program

Level 1 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• This method has the advantage of capitalizing on existing measures in core content areas while requiring development of new end-of-grade or end-of-course measures only where none is available • In grades 1–8, the end-of-grade assessments (EOGs) are intended to measure student learning in relation to the Common Core State Standards (CCSS) in ELA and math and to the state's standards in science, social studies, visual and performing arts, and physical education. • In grades 7–8, EOGs measure student learning in relation to the CCSS Literacy Standards in Social Studies/History, Science, and Technical Subjects and the state's standards in foreign language. • In grades 9–12, course-specific exams (end-of-course exams, or EOCs) are used to assess gains in those content areas with pre-existing measures (e.g., Algebra I, Biology, English 10). For those courses for which no measure currently exists, EOCs would need to be developed. Content-appropriate performance assessments would be developed to assess course-specific gains in visual and performing arts and physical education. • Ensures that teachers in all content areas and grades are included in teacher effectiveness analyses; intended to allow for evaluation of teachers in the arts, physical education, or foreign languages; in grades K–2; and in self-contained classrooms • Data also may be used for other purposes (e.g., to meet graduation requirements or inform promotion/retention decision-making)		

Level 1 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Newly Developed Pre-Post Measures of Achievement (n) (e)	• State- or district-developed, instructionally sensitive measure administered at the beginning (pre) and end (post) of school year or course • Gains in performance for each student receiving instruction from any one teacher are calculated • The mean gain for that teacher can then be compared to expected annual gain for that grade and content area (criterion-referenced model) or to the mean gain for that teacher's peers (norm-referenced model); alternatively, mean gain for each teacher can be estimated via analytic models (e.g., value-added modeling) that take into account students' starting points (and other covariates, if desired) • Instructionally sensitive tests are administered at the beginning and end of the year to allow for cleaner estimation of growth • Uses measures specifically designed to assess the contribution of the teacher in the current year • Can be customized to fit specific standards, age groups, and content areas • Grades K–5: A comprehensive measure may be used to assess pre-post gains in ELA, math, science, and social studies. Age- and content-appropriate performance assessments administered pre- and post-instruction are used to assess gains in visual and performing arts and physical education. • Grades 6–8: Content-specific measures are used to assess pre-post gains in ELA, math, science, social studies, foreign language, and technology (and other electives, as needed). Age- and content-appropriate performance assessments administered pre- and post-instruction used to assess gains in visual and performing arts and physical education.	• Requires (a) the development of measures that are designed to assess a teacher's instructional impact in each content area and grade and/or course in all schools across the state; (b) expert judgment in setting the criteria or standard for performance against which teachers will be evaluated, i.e., the expected annual gain for each grade and content area; and/or (c) expert judgment in determining how norm-referenced scores will be used (e.g., a ranking system) • May be costly to develop measures if only used for classroom accountability purposes • Strong diagnostic/pre-post measures are challenging to develop	CECR (2009) EPI (2010) Goe (2008) Hillsborough County, FL—STAR program

Level 1 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• Grades 9–12: Course-specific exams (EOCs) are used to assess gains in those content areas with preexisting measures (e.g., Algebra I, Biology, English 10). For those courses for which no measure currently exists, tests that may be administered in pre-post fashion would need to be developed. Content-appropriate performance assessments would be developed and administered in pre-post fashion to assess course-specific gains in visual and performing arts and physical education. • Fully inclusive of all teachers, including those in the arts, physical education, or foreign languages and those who teach in grades K–2 or in self-contained classrooms • Data may be used for other purposes (e.g., to diagnose students' strengths and limitations prior to and following instruction) • Only research-supported measure for attributing growth in learning to teacher effect (causal relationship)		

(n) = appropriate for new teachers (0 - 2 years of teaching experience); (e) = appropriate for experienced teachers (3+ years of experience)

Level 2 Data Sources

Level 2 data sources can be used effectively for low- and medium-stakes purposes. A number of Level 2 measures meet technical adequacy expectations and are administered in a standardized way and/or are collected systematically, either by the district or by a qualified external agency. These data include test-based and non-test-based aggregate data (e.g., graduation rate); vendor-, district-, or school-developed assessments not administered on an annual basis (e.g., interim assessments); student-level data from locally administered performance assessments; and teacher-level data collected by a qualified external agency.

When Level 1 data are not available and resources do not allow for new development, Level 2 data may be used in conjunction with key Level 3 sources (e.g., classroom observation) for high-stakes purposes. In the following table, a number of Level 2 options are presented that can be used to support informed decision-making about teacher effectiveness, but that are best used in conjunction with Level 1 data (if available). In all cases, use of multiple types of information (e.g., assessment data and findings from an external agency) helps ensure that the full range of effectiveness indicators has been considered when validating teachers' professional practices.

Table 2. Level 2 Data Sources

Level 2 Data Source: Assessment Data	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Test-Based Aggregate Performance Data (n) (e)	• Teachers in all grades and content areas can be included (school wide averaging) • May be focused on status or growth	• Effects cannot be attributed to teacher alone	NCCTQ (2010 a-e)
Non-Test-Based Aggregate School-, District-, or Department-Level Data (Collective Performance) (n) (e)	• Attendance, dropout, or graduation rate; group or school wide gain • Trustworthy data tapped from existing source • Can include teachers in non-core content areas (arts, PE, or foreign languages) as well as those in early elementary grades (K–2) and in self-contained classrooms	• Effects cannot be attributed to teacher effect alone	CECR (2009) Denner, Salzman, and Bangert (2001) Instructional Quality Assessment Intellectual Demand Assignment Protocol [IDAP] Mathers et al. (2008) NCCTQ (2010a & 2010b)
Other Vendor-Developed or Locally Developed Measures of Achievement (n) (e)	• Interim assessments aligned to state standards • Measures of foundational skills, frequently used in early elementary grades • Some may also be intended for diagnostic use • May include alternate assess ments or ELP tests	• Lack of research to support validity of use for purpose of teacher evaluation • Items must be administered and scored in ways that promote high levels of consistency across students and over time	Accomplished California Teachers [ACT], 2010 CECR (2009) Delaware Performance Appraisal System (DPAS II) DIBELS SAT University of Virginia—CLASS
Student-Level Performance Assessments (n) (e)	• Especially useful for teachers in content areas such as visual and performing arts and physical education • Can measure cognitively demanding content • Expectations for performance must be clear • Exemplars can support differentiation across performance at different levels	• Need to confirm alignment with state standards and local curriculum • Need to ensure technical adequacy (fair, reliable, valid) for purpose intended	BMGF (2010a & 2010b) Balanced Assessment in Mathematics (BAM) SAT 9 Reading Open Ended Test

Level 2 Data Source: Evaluation by External Agency	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Formal Classroom Observation (live or recorded) (n) (e)	• Most direct way to examine instructional practices • Moderately linked to student achievement • Evaluators are generally well trained and familiar with the protocol and rating system	• Can be costly to do, especially if done frequently or for longer durations • Little information to support use for high-stakes teacher evaluation	Baker et al. (2010) BMGF (2010a & 2010b) Danielson (1996, 2007) Goe and Croft (2009) Hanushek and Rivkin (2010) Harris (2009) Hill (2009)

Level 2 Data Source: Evaluation by External Agency	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• Useful for moderately high-stakes purposes or for high-stakes purposes if used in conjunction with Level 1 data • Can be adapted for use at any grade level or content area • Can capture degree of fidelity to standards-based instruction • Useful for feedback and coaching to develop greater teacher effectiveness • When conducted by a qualified agency, this data source holds real promise as a trustworthy and useful source of information	• Can be hit-or-miss; may or may not catch the teacher during a particularly strong teaching event • Interrater reliability (across teachers and over time) is a concern • Evaluation criteria must be transparent and valid for this purpose • Hill (2009) suggests that consultants may have better understanding of statistical complexities of value-added modeling	Kimball and Milanowski (2009) Mathers et al. (2008) Mathematical Quality of Instruction (MQI) Classroom Assessment Scoring System (CLASS) Protocol for Language Arts Teaching Observation (PLATO) Quality Science Teaching Instrument (QST) Reformed Teaching Observation Protocol (RTOP) Teachscape TEX-IN3 Observation System
Interview with Teacher or Performance Assessment (n) (e)	• Can capture change in teaching practice • Can communicate program-specific philosophies or goals • Can be captured via video clips • Exemplars can support differentiation across levels of performance • External evaluators generally are more experienced in applying rubric than are school staff	• Little information to support use for high-stakes teacher evaluation • Need to balance data collection needs with burden on teachers • Scoring criteria must be validated by master teachers • Performance standards must be widely communicated to practitioners and supported by master teachers	AACT (2010) CPRE (2006) Danielson (1996, 2007) Darling-Hammond (2010) Goe et al. (2008) Koppich, Asher, and Kerchner (2002) Instructional Quality Assessment Instructional Demand Assignment Protocol National Board for Professional Teaching Standards Performance Appraisal Review for Teachers Performance Assessment for California Teachers (PACT) SCOOP Notebooks Teacher Performance Assessment (TPA)
Teacher-Submitted Portfolio or Work Samples (n) (e)	• May include lesson plans, student work, daily reflections • Intended to capture artifacts from effective teaching events • Can be used at any grade level in any content area • External evaluators generally are more experienced in applying rubric than are school staff • May reveal strengths and limitations in teachers' practice	• Scoring rubrics must be validated by master teachers • Time consuming to develop and challenging to standardize • Inconclusive results about usefulness for measuring teacher effectiveness; findings vary widely depending on system used	Darling-Hammond (2010) Denner et al. (2001, 2003) Beginning Educator Support & Training [BEST], (CT) Instructional Quality Assessment (CRESST) Intellectual Demand Alignment Protocol (CCSR) Performance Assessment for California Teachers (PACT) Renaissance Teacher Work Sample Teacher Work Sample Methodology of Western Oregon University

Level 2 Data Source: Evaluation by External Agency	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
National Board Certification Application or Certification Renewal Materials (e)	• Evaluators are well trained and familiar with the protocol and rating system • Teacher must adopt practices that have been shown to be effective • Teachers report that going through the process improved their teaching; can lead to improved classroom management, design and delivery of content, subject matter knowledge, and evaluation of student learning • Standardized and well documented • Significant accomplishment for more experienced teachers	• Certification renewal materials can vary by state and are not always reviewed for quality • Board certification status is not strongly linked to other predictors of teacher effectiveness • Lack of conclusive evidence that certification is an effective indicator of teacher quality • Unclear whether process itself leads to improvement in practice or whether only the most effective teachers opt to complete the process • Requires long-term commit ment to accomplish this goal • While Board certified teachers are expected to act as mentors to colleagues, colleagues' effectiveness is not increased solely by this mentorship	Allen, Snyder, and Morley (2009) Cantrell, Fullerton, Kane, and Staiger (2008) Cavalluzzo (2004) Darling-Hammond (2010) Hakel, Koenig, and Elliot (2008) Harris and Sass (2007, 2009) Kane, Rockoff, and Staiger (2006) Mathers et al. (2008) Sanders, Ashton, and Wright (2005) Vandervoort et al. (2004)
Tests of Content Knowledge and/or Understanding of Pedagogy (e)	• Generic as well as content-specific knowledge and skills	• Instruments must be validated as appropriate for this purpose	Donovan and Bransford (2005) NCTM, NCTE University of Michigan's Learning Mathematics for Teaching

(n) = appropriate for new teachers (0 - 2 years of teaching experience); (e) = appropriate for experienced teachers (3+ years of experience)

Level 3 Data Sources

Richly descriptive Level 3 data include formal and informal classroom observations; surveys and interviews with students, teachers, or parents; teacher-level performance assessments or checklists; portfolio reviews; peer evaluations; focus groups; documentation of pursuit of advanced academic or leadership opportunities; and review of pre-service credentials (e.g., GPA in content major, score on credentialing exam). These data may focus on specific teacher behaviors, attitudes, credentials, or qualifications. Level 3 sources help inform decision-making when the data are collected for low- and medium-stakes purposes (e.g., hiring or promotion decisions, determination of professional development needs) or to supplement Level 1 or Level 2 data for high-stakes purposes. Relevant information about teachers' practices can be collected about all teachers, regardless of content area or grade level assignment.

However, Level 3 data can be time consuming to collect, and challenging to use effectively and reliably as rating instruments; also, many are associated with known sources of bias (e.g., self-report). For this reason, in the following table, a number of Level 3 options are presented that can be used in conjunction with data from Level 1 (if available) or Level 2, depending on the unique purpose for data collection and context. As previously stated, use of multiple types of information (e.g., assessment data, findings from an external agency, and classroom observation) helps ensure that the full range of effectiveness indicators has been considered.

Table 3. Level 3 Data Sources

Level 3 Data Source: School Administrator Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Formal Classroom Observation by School Administrator (n) (e)	• Most direct way to get a glimpse of teaching practice • High face validity and teacher buy-in • Moderately linked to student achievement • Useful as formative tool for coaching teacher performance • Captures information on teachers' instructional practice • Can be adapted for use at any grade level or content area • Can be standardized via use of protocol or rubric • Can be useful for moderately high-stakes purposes when used in conjunction with Level 1 data	• Costly to do frequently or for longer durations, but least useful with one-shot approach and may introduce interrater reliability issues if multiple observers are used • Observer may not have sufficient subject matter expertise to make informed judgment • Little information about reliability and validity for teacher evaluation purposes • Requires proper training to apply rubrics appropriately and make judgments about whether students are learning • Most useful in identifying teachers who produce the largest and smallest achievement gains in students (or, more broadly, the strongest and weakest teachers) • Teachers should be informed about criteria used to evaluate instructional methods, classroom management strategies, etc.	CDE (2010) Danielson (1996, 2007) Goe et al. (2008) Goe and Croft (2009) Jacob and Lefgren (2008) Junker et al. (2006) Mathers et al. (2008) NCCTQ (2010 a-e) Observations REL Midwest (2007, 2008) Weems and Rogers (2010) Classroom Assessment Scoring System (CLASS) Denver ProComp Instructional Quality Assessment (IQA) Protocol for Language Arts Teaching (PLATO) Quality Compensation (Q Comp)—State of Minnesota Quality Science Teaching Instrument (QST) Reformed Teaching Observation Protocol TEX-IN3 Observation System UTeach Observation Protocol (UTOP) Washington, DC, IMPACT
Interview with Teacher or Traditional Performance Review (one-time discussion) (n) (e)	• Can tap teachers' intentions, goals, thought processes, perspective, knowledge, and beliefs • Can help bring to surface teachers' underlying philosophies and attitudes • Can be structured via standards for performance and scoring rubric to add reliability and rigor • Convenient and cost effective • Useful for targeting professional development programs toward key needs	• Focuses more on teacher characteristics than on instructional effectiveness • Little information on validity and reliability for purposes of teacher evaluation • Content varies for each protocol, and focus of each may be quite different • Requires sufficient expertise, training, and capacity to conduct effectively • Not effective as incentive for improvement	Calabrese, Sherwood, Fast, and Womack (2004) Darling-Hammond (2010) Goe et al. (2008) Hanushek and Rivkin (2010) Harris (2009)
Teacher-Submitted Portfolio or Work Samples (n) (e)	• Intended to capture artifacts from teaching events • Encourages teacher self-reflection and growth • Promotes active teacher participation in evaluation	• Training in using scoring rubric is critical • Time consuming to develop and review and challenging to standardize across schools, years, content areas, grades	Denner et al. (2001) Fleak, Romine, and Gilchrist (2003) Goe et al. (2008) EDUCATE Alabama

Level 3 Data Source: School Administrator Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• Can show alignment between instruction and standards and long-term information about teaching practices • Can be used at any grade level in any content area • May focus on broad aspects of teaching seen as important for specific contexts or content areas	• Teacher selection criteria may be biased; materials included may not be fully representative of teachers' practice • Use for high-stakes decision-making has not been validated	Little, Goe, and Bell (2009) Mathematical Knowledge for Teaching Instrument Mathers et al. (2008) NCCTQ (2010 a-e) New Mexico Professional Development Dossier Sanders et al. (2005) Stronge (2007) Instructional Quality Assessment (CRESST) Memphis Teacher Effectiveness Initiative National Board for Professional Teaching Standards (NBPTS) Teaching and Learning International Survey Teaching as Leadership Tennessee Comprehensive Assessment System Vermont, Connecticut, Washington, and Wisconsin teacher portfolio assessments Web-Based Teaching Log
Teacher Performance Assessment (n) (e)	• Best format for evaluating teacher in the act of delivering instruction and interacting with students • Provides evidence of classroom practices • Performance standards can add reliability and rigor if linked theoretically and practically to quality instruction • Performance standards are most effective when linked to closing student achievement gaps in that school/district/state • Exemplars can support differentiation among levels of teaching efficacy • A continuum of instruments exist that make performance assessments suitable for both novice and experienced teachers • Well-developed tasks can be used to describe the full range of performance (from novice to expert) in each domain or in specific activities	• Need to balance data collection needs with burden on teachers • Training in using scoring rubric is critical • Standards for performance must be transparent to teachers • Scoring protocols and rubrics should align with professional standards • Most useful in identifying teachers who produce the largest and smallest achievement gains in students (or, more broadly, the strongest and weakest teachers) • Must specify clear criteria for desired behaviors	Center for Collaborative Education (2010) CPRE (2006) Danielson (1996, 2007) Darling-Hammond (2010) Heneman, Kimball, and Milanowski (2006) Jacob and Lefgren (2008) Sluijsmans and Prins (2006) Performance Assessment for California Teachers (PACT) TAP: The System for Teacher and Student Advancement (National Institute for Excellence in Teaching [NIET]) Teacher Performance Assessment (TPA) Washington, DC, IMPACT

Level 3 Data Source: School Administrator Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Performance Checklist (n) (e)	• Describes behaviors of interest that can be observed in different settings at different times • Easily administered and inexpensive • Can be standardized • Minimal training necessary if exemplars are used • Formal training and frequent calibrations increase consistency of ratings	• Provides no indication of level of quality of checked items • Prone to issues with reliability • May need multiple evaluators to monitor events over course of year • May "hit or miss" (observer must be at the right place at the right time)	Denner et al. (2001) Mathers et al. (2008) IDAP IQA PACT
Years of Experience (Tenure) (e)	• Based on assumption that years of experience are indicator of quality; hence, tenure is proxy for quality in some evaluation systems • Can be used with teachers in all content areas and at all grade levels	• Despite extensive study, has not been linked conclusively to improved teaching practices or increases in student achievement	Braun (2005) Goe (2007) McCaffrey et al. (2008) NCCTQ (2010 a-e) Sanders et al. (2005)

Level 3 Data Source: Student Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Student Survey (n) (e)	• Inexpensive and easily administered • Provides students' perception of value of interactions with teacher • Can be useful for connecting teacher with students • Shown to be more strongly correlated with student achievement than administrator- or self-reported teacher effectiveness ratings • Feedback can be used formatively by teacher	• Students only qualified to rate on certain areas of effective teaching • Little information on validity and reliability for teacher evaluation purposes	BMGF (2010a & 2010b) Ferguson (2008) Goe et al. (2008) McQueen (2001) NCCTQ (2010 a-e) Peterson et al. (2001) Wilkerson, Manatt, Rogers, and Maughan (2001) BMGF (2010a & 2010b) Little, Goe, and Bell (2009) Davis School District—Utah Experience Sampling Method Memphis Teacher Effectiveness Initiative Quality Assessment Notebook School Performance Framework Teacher Behavior Inventory Teaching as Leadership Tripod Project Surveys

Level 3 Data Source: Student Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Student Interview (n) (e)	• Can be conducted by school officials • Provides insight into students' perceptions of teachers' strengths and limitations, and whether these perceptions are consistent across different groups of students • Can capture affective and attitudinal elements • Student ratings are more strongly correlated with student achievement than administrator- or self-reported teacher effectiveness ratings • Feedback can be used formatively by teacher • Encourages student buy-in (face validity)	• Little information on validity and reliability for teacher evaluation purposes • Students lack knowledge about the full context of teaching, and ratings may be susceptible to bias	BMGF (2010a & 2010b) Little, Goe, and Bell (2009)

Level 3 Data Source: Peer Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Informal Classroom Observation by Teaching Peer (n) (e)	• Observer is someone with similar content knowledge, instructional background, and expertise; may be from same or different school • Useful as formative assessment for coaching teacher performance; captures important information about teachers' instructional practices • Can be adapted for use at any grade level or content area • Can be standardized via use of protocol or rubric	• Time consuming • Peer must have deep knowledge to provide suggestions for improvement • May pull teachers from instructional responsibilities	Danielson (1996, 2007) Mathers et al. (2008) Sawchuk (2009) Cincinnati Public Schools Hillsborough County, FL—STAR program National Board for Professional Teaching Standards (NBPTS) Peer Assistance and Review (PAR)—Toledo, OH Teacher Evaluation System (TES)
Professional Support Group (n) (e)	• Invites sharing of lesson plans and effective instructional strategies within a sustainable learning community • Can encourage learning or deepening of knowledge of pedagogy • Inexpensive and easy to implement • Designed to remain in place for extended periods of time	• Generally not associated with high-stakes use (e.g., for accountability purposes)	Gordon, Kane, and Staiger (2006) Monroe-Baillargeon and Shema (2010) Mullen and Hutinger (2008) Norman, Golian, and Hooker (2005) Sawchuk (2009)

Level 3 Data Source: Peer Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• May be valuable for learning new skills and renewing commitment to the profession • Supports growth of novice teachers • Feedback can be used formatively by teacher		

Level 3 Data Source: Parent Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Survey (n) (e)	• Inexpensive and easy to administer online to parents in remote locations • Questions can be targeted to particular grade or content area • Feedback can be used formatively by teacher • Promotes parent buy-in	• Little information to support use as part of high-stakes teacher evaluation • May present a burden to some parents • Important to inform parents about how data will be used	Gordon, Kane, and Staiger (2006) Koppich et al. (2002) McQueen (2001) Peterson, Wahlquist, Brown, and Mukhopadhyay (2003)
Focus Group Discussion (n) (e)	• May highlight teacher's emerging strengths or ongoing challenges (e.g., classroom management) • Feedback can be used formatively by teacher	• Little information to support use as part of high-stakes teacher evaluation	Gordon, Kane, and Staiger (2006)

Level 3 Data Source: Teacher Self-Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Survey (n) (e)	• Questions may focus on pedagogy, instructional materials, technology, or content • Can tap teachers' intentions, values, thought processes, perspectives, knowledge, attitudes, beliefs, and professional ethics • Convenient and cost effective	• Self-reported data are associated with known limitations • Concerns that teacher reports do not closely correspond to researcher or administrator comments or reports	Ball and Rowan (2004) Thornton (2006) New Teacher Center Study of Instructional Improvement Surveys of Enacted Curriculum
Viewing of Video Recording of Teaching Event (n) (e)	• Requires reflection that may encourage teacher growth and may reveal teacher characteristics as well as practices • Promotes teacher participation in the evaluation process	• Time consuming and challenging to create high-quality recording • Little information to support use as part of high stakes teacher evaluation	Goe et al. (2008) Kennedy (2008) Mathers et al. (2008) Surveys of Enacted Curriculum Teaching Log

Level 3 Data Source: Teacher Self-Evaluation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Accessing Professional Development or Leadership Opportunities (n) (e)	• Can focus on broad and overarching aspects of teaching or on specific content matter • Can be targeted to meet needs of specific age group or school needs	• Findings inconclusive about actual impact on teacher effectiveness	BMGF (2010a & 2010b) Harris and Sass (2007) Heneman et al. (2006) Toch & Rothman (2008) Continuum of Teaching Practice (CA) Delaware Performance Appraisal System (DPAS II) Denver ProComp Learning Math for Teaching Project

Level 3 Data Source: Pre-Service Preparation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
Teacher Preparation Program Attended	• Quality of preparation program viewed by employers as potential indicator of quality of teachers trained in that program • May be most useful when district focuses on characteristics of key feeder programs (e.g., state or regional colleges with teacher preparation programs that frequently send potential candidates for employment) • Recruitment of new teachers may be targeted toward candidates that attended programs with strong NCATE evaluations	• Preparation programs should provide evidence that they endorse performance expectations embraced by potential employers • Preparation programs should provide evidence that students are actively engaged in coursework and activities that foster development of widely valued knowledge and skills • Not a consistent predictor of teacher effectiveness	Darling-Hammond, Newton, and Wei (2010) Gimbert, Cristol, and Sene (2007) Glazerman, Mayer, and Decker (2006) Goe and Stickler (2008) Heneman et al. (2006) Noell (2005) REL Midwest (2007) Stanford Teacher Education Program Teach for America Tennessee Teacher Effectiveness Studies (TN Teacher Quality Reform Initiative) Transition to Teaching
College Courses Taken and GPA	• Often considered as key part of package of teacher's "qualifications" during hiring decision-making • Recruitment of new teachers may be targeted toward candidates who have elected coursework and activities that lead to depth of understanding in content area and in pedagogy • Serves as an indication of depth and range of content knowledge and used as a proxy for direct measure of teacher knowledge	• Programs should provide evidence that their coursework is sufficiently rigorous to prepare candidates for effective teaching at the elementary or secondary levels • Findings inconclusive about actual impact on teacher effectiveness	Center for Collaborative Education (2010) Goe & Stickler (2008) Toch & Rothman (2008)

Level 3 Data Source: Pre-Service Preparation	Description and Potential Uses for Data	Guidelines for Use of Data	Exemplars and/or References
	• Can indicate degree to which pre-service teacher was actively engaged in coursework and developed valued knowledge and skills in relation to peers		
Score on Qualifying Exam for Certification	• Teachers' subject area certification is one of the teacher qualifications most consistently and strongly associated with improved student achievement, especially in math at the secondary level	• User should seek evidence to ensure test score is valid for this purpose • Scant research on impact on teacher effectiveness in content areas other than math • Rigor and technical adequacy can vary widely across exams	Cavalluzzo (2004) Hanushek, Kain, O'Brien, and Rivkin (2005) Kane, Rockoff & Staiger (2006) Toch & Rothman (2008)
Pre-Service Performance Assessment	• Conducted during field experience (e.g., student teaching) • Best format for evaluating teacher in the act of delivering instruction and interacting with students • Standards for performance must be transparent to teacher candidates and grounded in theory and research • Standards for performance should add reliability and rigor to performance evaluation • Exemplars can support differentiation among levels of teaching efficacy		Center for Collaborative Education (2010) CPRE (2006) Danielson (1996, 2007) Darling-Hammond (2010) National Board for Professional Teaching Standards Performance Assessment for California Teachers (PACT) Teacher Performance Assessment (TPA)
Tests of Content Knowledge and/or Understanding of Pedagogy (Pre-Service)	• Direct measure of teacher knowledge in content area • Generic as well as content-specific knowledge and skills	• Depending on test design, may or may not capture deep understanding of content • User must seek evidence to ensure that test is valid for this purpose	Aaronson, Barrow, and Sanders (2003) Darling-Hammond (2010) Donovan and Bransford (2005) Goe and Stickler (2008)
	• Math pedagogical knowledge was strongest teacher-level predictor of student achievement; research has shown that completion of an undergraduate or graduate major in math was associated with higher student achievement at the secondary level	• Content knowledge without understanding of pedagogy can reduce teacher effectiveness	Harris and Sass (2007) Hill, Rowan, and Ball (2005) Toch and Rothman (2008) Learning Mathematics for Teaching—Michigan Marshall (2009) NCTM, NCTE Praxis

(n) = appropriate for new teachers (0 - 2 years of teaching experience); (e) = appropriate for experienced teachers (3+ years of experience)

II. Summary and Recommendations for Measuring Teacher Effectiveness

As described in the preceding sections, a comprehensive teacher effectiveness rating process relies on evidence collected from multiple sources (i.e., a comprehensive or "hybrid" approach; for example, see Baker et al., 2010; BMGF, 2010a; or Hanushek & Rivkin, 2010). Because the stakes associated with evaluating a teacher's professional practices can be high (e.g., when used for accountability purposes), use of trustworthy data is critical. Level 1 assessment data, particularly when statistical analyses can be incorporated to estimate a teacher's unique impact during one grade or course, offer the specific psychometric characteristics necessary for high-stakes decision-making. Because technically sound Level 1 measures currently are not available for all teachers, state and local decision-makers will want to consider strategies for supplementing existing data with different types of information from Level 2 and/or Level 3. In this model, data from Levels 2 and 3 may be weighted differently than data from Level 1 in determining a comprehensive effectiveness rating. In all cases, it is important to consider the strengths and limitations of each data source highlighted in the preceding tables and to heed the cautions for use.

For the purpose of validating teacher effectiveness, particular combinations of data may be stronger than others in certain scenarios. Four scenarios are provided in this section, each describing a unique context for examining teacher effectiveness and strategies for combining different types of information to develop a defensible system for validating teachers' professional practices.

Scenario 1: The data are to be used for accountability purposes for all teachers across a state (high stakes); the context provides for ample resources for data collection across multiple sources; and standardized test scores (Level 1 data) are available at the student level in core content areas only. In this scenario, the centerpiece data sources are newly developed comprehensive pre-post measures administered in grades K–6, content-specific pre-post measures administered in grades 7 and 8, and course-specific pre-post measures administered in high school. Measures would be developed to ensure that teachers are evaluated consistently across all grades and content areas. The measures developed for teachers in performance-focused content areas, such as visual and performing arts and physical education, would be performance-based. While these measures focus only on student achievement or performance outcomes, well-developed pre-post measures have the Level 1 characteristics necessary to allow defensible judgments about a teacher's instructional effectiveness to be made for accountability purposes. In addition, data from these measures can be used diagnostically by teachers to target instruction to meet students' individual needs. Even when Level 1 data are readily available, states should look to supplement the information obtained from these measures with a broader range of non-assessment indicators that can be shown to be sufficiently trustworthy for high-stakes purposes.

Scenario 2: The data are to be used for accountability purposes for all teachers across a state (high stakes); the context provides for limited resources for data collection; and standardized test scores (Level 1 data) are available at the student level, only in core content areas at key grades. In this scenario, the strongest combination of data includes (a) end-of-year statewide testing data for teachers in ELA, math, and science at grades 3–8; (b) newly developed content-specific end-of-year assessments and/or performance tasks in social studies, foreign languages, visual and performing arts, physical education, and other electives in grades 3–8; (c) newly developed comprehensive end-of-grade measures and/or performance tasks for grades K–2; and (d) end-of-course test scores for all high school teachers, using existing data when available and newly developed measures in other content areas (e.g., foreign languages) or electives. Using a student-level longitudinal tracking system, scores from summative assessments administered at the end of one grade (grade 1–high school) are compared with that student's scores from the previous grade. Annual mean gains for each teacher are estimated via analytic models (e.g., value-added modeling) that take into account students' unique starting points. Mean gains for each teacher can be compared to the expected annual gain for that grade and content area (criterion-referenced model) or to the gains for that teacher's peers (norm-referenced model). These student-level achievement scores should be used in conjunction with other sources of information from Level 3 (e.g., principal's classroom observation) or from an external agency.

Scenario 3: The data are to be used for decision-making about continued employment (or standard evaluation) for a second-year high school social studies teacher (medium stakes), the context provides for sufficient resources for data collection from multiple sources, and some standardized testing data (Level 1 or 2 data) are available. In this scenario, the strongest combination of data includes scores from end-of-course assessments in American History (Level 1 data), scores from interim measures administered to students in all non-EOC social studies courses (Level 2 data), and formal principal observation (Level 3 data).

Scenario 4: The data are to be used for formative purposes (low stakes) to promote science teachers' professional growth; the context provides for limited resources for data collection; standardized test scores (Level 1 data) are available only at grades 5 and 7; interim test scores (Level 2 data) are accessible for teachers in grades K–4, 6, and 8; and end-of-course scores are available for teachers of high school biology. In this scenario, the strongest combination of data includes all available data from Levels 1 and 2, peer observation in all grades and high school courses, and student survey data from all grades and courses.

Table 4. Assessment Plan for Scenario 2: Use for State Test Data, Supplemented with Newly Developed End-of-Year Assessments

	Grades K - 2	Grades 3 - 6	Grades 7 - 8	High School
English Language Arts	Newly developed comprehensive end-of-grade test	Existing state test	Existing state test	Grade 9 EOC Grade 10 EOC Grade 11 EOC Grade 12 EOC
Mathematics	Newly developed comprehensive end-of-grade test	Existing state test	Existing state test	Algebra I Algebra II Geometry Calculus
Science	Newly developed comprehensive end-of-grade test	Existing state test	Existing state test	Biology Chemistry Physics Earth/Space Science
Social Studies	Newly developed comprehensive end-of-grade test	Newly developed end-of-grade test	Newly developed end-of-grade test	Geography History Government/Civics Economics
Foreign Languages	NA	Newly developed end-of-grade tests	Newly developed end-of-grade tests	Course-specific EOCs
Visual/Performing Arts	Performance Tasks	Performance tasks	Performance tasks	Performance tasks
Physical Education	Performance Tasks	Performance tasks	Performance tasks	Performance tasks
Other Electives	NA	NA	Newly developed end-of-grade tests	Course-specific EOCs

Glossary

Artifacts	Teacher-developed instructional materials such as model lesson plans, assignments, student work samples, audio or video recordings of classroom performance, notes from students or parents, teacher reflections or journals, results from assessments, and/or special awards or recognitions. Artifacts frequently are collected in a portfolio.
Checklists	List of target actions for teachers, with spaces for marking when the action was performed and for recording comments. Target actions may range from activities such as engaging students during instruction to participation in IEP meeting decision-making.
Classroom Observation	Review of teacher performance during course of instruction, generally supported through use of a protocol and/or checklist. Evaluation of observation data is enabled by professional judgment and application of rubrics developed by master educators.
Effective Teacher	"A teacher whose students achieve acceptable rates of student growth. A method for determining if a teacher is effective must include multiple measures, and effectiveness must be evaluated, in significant part, on the basis of student growth. Supplemental measures may include, for example, multiple observation-based assessments of teacher performance." (Race to the Top Application for Initial Funding, 2010)
Highly Effective Teacher	"A teacher whose students achieve high rates of student growth. A method of determining if a teacher is highly effective must include multiple measures, provided that teacher effectiveness is evaluated, in significant part, on the basis of student growth. Supplemental measures may include, for example, multiple observation-based assessments of teacher performance or evidence of leadership roles that increase the effectiveness of other teachers." (Race to the Top Application for Initial Funding, 2010)
In-Service Information	Data that are collected while a teacher is actively employed in the teaching profession. This includes the type of artifacts described above; feedback from administrators, peers, students, parents, and licensing entities; and results from annual reviews, performance assessments, and other measures.
Interview	Review of performance during formal discussion with employer or other stakeholder, generally supported through adherence to a script or protocol. Interviews are useful for soliciting information unique to the interviewee such as attitudes.
Non-tested Grades and Subjects	The grades and subjects that currently are not required to be tested annually under the ESEA.
Performance Task	Review of performance during completion of a specific task, generally supported through adherence to a protocol or use of a checklist. Performance tasks have a broad range of applications and vary in scale, ranging from planning and implementing a unit of instruction to completing complex projects over the course of many months. Scoring of performance tasks is enabled by professional judgment and application of rubrics developed by master educators.
Portfolio	A portfolio is a collection of teacher-developed artifacts compiled by teachers to exhibit evidence of their teaching practices, school activities, and student progress. Portfolios generally include exemplary artifacts selected by the teacher. Scoring of portfolios is enabled by professional judgment and application of rubrics developed by master educators.
Pre- and Post-Tests	Locally developed or customized tests of achievement that measure the content of a grade-level curriculum or course. The same (or nearly same) test is administered at the beginning of a unit or course of instruction (usually the beginning of the year or semester) and again at the end of a unit or course of instruction (usually the end of a year or semester). The purpose of pre-post testing is to gather finely grained information about what individual students know and can do in relation to a particular unit of instruction by comparing preexisting understanding (pretest) to post-instruction understanding (posttest).

Pre-service Information	Data that are collected during each teacher's formal training period. This includes academic coursework related to a content major as well as performance in courses on pedagogy, curriculum, classroom management, and educational leadership. These data also may include results from certification tests or other assessments.
Student Work Samples	Samples of student-completed work that usually are focused on a specific teaching event or instructional unit. Samples may include multiple pieces from the same student to show development through a unit, work that demonstrates several levels of achievement within a unit, or a class set of work for a specific unit. Scoring of student work samples is enabled by professional judgment and application of rubrics developed by master educators.
Survey	Selected-response or open-ended questions about teacher performance that may be used to elicit information from a variety of stakeholders (e.g., students, parents).Surveys can be administered online or by paper and pencil.
Teacher Effect	A teacher's contribution to a valued student learning outcome relative to the average for that school, district, or state. In most models, the focus is on the academic growth for all students exposed to a particular teacher during the course of instruction, as measured by a standardized test score or other trustworthy measure.

References

American Association of Colleges for Teacher Education (AACTE). (2010). *Teacher performance assessment consortium.* Retrieved from http://aacte.org/index.php?/Programs/Teacher-Performance-Assessment-Consortium-TPAC/teacher-performance-assessment-consortium.html

Aaronson, D., Barrow, L., & Sanders, W. (2003). *Teachers and student achievement in the Chicago public high schools.* Chicago,IL: Federal Reserve Bank of Chicago.

Accomplished California Teachers (ACT). (2010). *A quality teacher in every classroom: Creating a teacher evaluation system that works for California.* Stanford, CA: National Board Resource Center, Stanford University.

Allen, P., Snyder, C., & Morley, J. (2009). The wheel has already been invented: Improving teacher effectiveness through National Board for Professional Teaching Standards. *Catalyst for Change, 36*(1).

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., & Linn, R. (2010). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper No. 278). Washington, DC: The Economic Policy Institute.

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

Battelle for Kids. (2009). *The importance of accurately linking instruction to students to determine teacher effectiveness.* Columbus, OH: Author.

Bill & Melinda Gates Foundation (BMGF). (2010a). *Working with teachers to develop fair and reliable measures of effective teaching.* Seattle, WA: Author.

Bill & Melinda Gates Foundation (BMGF). (2010b). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project.* Seattle, WA: Author.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service.

Calabrese, R., Sherwood, K., Fast, J., & Womack, C. (2004). Teachers' and principals' perceptions of the teacher evaluation conference: An examination of Model 1 theories-in-use. *The International Journal of Educational Management, 18*(2), 116–117.

California Department of Education. (2010). *Teacher and principal evaluation systems.* Retrieved from http://www.cde.ca.gov/nclb/sr/tq/tpevalsys.asp?

Cantrell, S., Fullerton, J., Kane, T., & Staiger, D. (2008). ***National Board Certification and teacher effectiveness: Evidence from a random assignment experiment*** (National Bureau of Economic Research Working Paper No. 14608). Cambridge, MA: National Bureau of Economic Research.

Cavalluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* Alexandria, VA: CAN Corporation.

Center for Collaborative Education. (2010). *Including performance assessments in accountability systems: A review of scale-up efforts.* Boston, MA: Author.

Center for Educator Compensation Reform (CECR). (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades.* Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.

Consortium for Policy Research in Education (CPRE). (2006). *Standards-based teacher evaluation as a foundation for knowledge-and-skill-based pay* (CPRE Policy Brief). Philadelphia, PA: Author.

Danielson, C. (1996). *A framework for teaching.* Alexandria, VA: ASCD.

Danielson, C. (2007). *Enhaning professional practice: A framework for teaching.* Alexandria, VA: ASCD.

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching.* Washington, DC: Center for American Progress.

Darling-Hammond, L., Newton, X., & Wei, R. (2010). Evaluating teacher education outcomes: A study of the Stanford Teacher Education Programme. *Journal of Education for Teaching: International research and pedagogy, 36*(4), 369–388.

Denner, P., Norman, A., Salzman, S., & Pankratz, R. (2003). *Connecting teaching performance to student achievement: A generalizability and validity study of the Renaissance Teacher Work Sample Assessment.* Paper presented at the annual meeting of the Association of Teacher Educators, Jacksonville, FL.

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

Domaleski, C., & Hill, R. (2010). *Considerations for using assessment data to inform determinations of teacher effectiveness.* Nashua, NH: Center for Assessment.

Donovan, M., & Bransford, J. (2005). *How students learn*. Washington, DC: National Academies Press.

Economic Policy Institute (EPI). (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper). Washington, DC: Author.

Ferguson, R. (2008). The Tripod Project framework. *The Tripod Project.* Retrieved from http://www.tripodproject.org/uploads/file/The%20Tripod%20Project%20Framework%281%29.pdf

Fleak, S., Romine, J., & Gilchrist, N. (2003). Portfolio peer review: A tool for program change. *Journal of Education for Business, 78*(3), 139.

Gimbert, B., Cristol, D., & Sene, A. (2007). The impact of teacher preparation on student achievement in algebra in a "hard-to-staff" urban PreK–12–university partnership. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 18*(3), 245–272.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management, 25*(1), 75–96.

Goe, L. (2008). *Key issue: Using value-added models to identify and support highly effective teachers.* Washington, DC: National Comprehensive Center for Teacher Quality.

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

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://tqcenter.learningpt.org/publications/RestoPractice_EvaluatingTeacherEffectiveness.pdf

Goe, L., & Stickler, L. (2008). *Research and policy brief: Teacher quality and student achievement: Making the most of recent research.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2010). *When the stakes are high, can we rely on value-added? Exploring the use of value-added models to inform teacher workforce decisions*. Washington, DC: Center for American Progress.

Goldhaber, D., & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions.* Washington, DC: CALDER Urban Institute.

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institute.

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

Hanushek, E., & Rivkin, S. (2010). *Using value-added measures of teacher quality* (CALDER Brief #9). Washington, DC: Urban Institute.

Hanushek, E., Kain, J., O'Brien, D., & Rivkin, S. (2005). *The market for teacher quality.* Cambridge, MA: National Bureau of Economic Research.

Harris, D. (2009). Would accountability based on teacher value added be smart policy? An evaluation of the statistical properties and policy alternatives. E*ducation Finance and Policy, 4,* 319–350.

Harris, D., & Sass, T. (2007). *Teacher training, teacher quality, and student achievement* (Working Paper No. 3). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Harris, D., & Sass, T. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management, 28*(1), 55–80.

Heneman, H., Kimball, S., & Milanowski, A. (2006). *The Teacher Sense of Efficacy Scale: Validation evidence and behavioral prediction.* Madison, WI: Wisconsin Center for Education Research.

Herman, J., Baker, E., & Linn, R. (2004). *Accountability systems in support of student learning: Moving to the next generation.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Hill, H. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28,* 700–709.

Hill, H., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 317–406.

Hinchey, P. (2010). *Getting teacher assessment right: What policymakers can learn from research.* Boulder, CO: National Education Policy Center.

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

Jacob, B., Lefgren, L., & Sims, D. (2009). *The persistence of teacher-induced learning.* Unpublished manuscript, Brigham Young University, Provo, UT.

Junker, B., Weisberg, Y., Matsumura, L., Crosson, A., Wolf, M., Levison, A., & Resnick, L. (2006). *Overview of the Instructional Quality Assessment* (CSE Technical Report No. 671). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Kane, T., Rockoff, J., & Staiger, D. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City.* New York, NY: Columbia University.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

Kennedy, M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 61-66.

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

Koppich, J., Asher, C., & Kerchner, C. (2002). *Developing careers, building a profession: The Rochester Career in Teaching Plan.* Kutztown, PA: National Commission on Teaching and America's Future.

Koretz, D. (2008). A measured approach: Maximizing the promise, and minimizing the pitfalls, of value added models. *American Educator, 39*, 18–27.

Lefgren, L., & Sims, D. (2010). *Using subject test scores efficiently to predict teacher value-added.* Unpublished manuscript, Brigham Young University, Provo, UT.

Linn, R. (2007). *Educational accountability systems* (CSE Technical Report No. 687). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. (2008), Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40*, 699–711.

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

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality.

Marshall, J. (2009). School quality and learning gains in rural Guatemala. *Economics of Education Review, 28*(2), 207–216.

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

Mathers, C., Oliva, M., & Laine, S. (2008). *Improving instruction through effective teaching evaluation: Options for states and districts.* Washington, DC: National Comprehensive Center for Teacher Quality.

McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND.

McQueen, C. (2001). Teaching to win. *Kappa Delta Pi Record, 38*(1), 12–15.

Monroe-Baillargeon, A., & Shema, A. L. (2010). Time to Talk: An urban school's use of literature circles to create a professional learning community. *Education and Urban Society, 42*(6), 651–673.

Mullen, C., & Hutinger, J. (2008). The principal's role in fostering collaborative learning communities through faculty study group development. *Theory Into Practice, 47*(4), 276–285.

National Comprehensive Center for Teacher Quality (NCCTQ). (2010a). *Conducting a cost analysis for educational policies: Teacher effectiveness.* Washington, DC: Author.

National Comprehensive Center for Teacher Quality (NCCTQ). (2010b). *Guide to teacher evaluation products.* Retrieved from http://www3.learningpt.org/tqsource/GEP

National Comprehensive Center for Teacher Quality (NCCTQ). (2010c). *Research and policy update: Special edition highlights resources on teacher evaluation.* Washington, DC: Author.

National Comprehensive Center for Teacher Quality (NCCTQ). (2010d). *Research-to-practice brief: Challenges in evaluating special education teachers and English language learner specialists.* Washington, DC: Author.

National Comprehensive Center for Teacher Quality (NCCTQ). (2010e). *Research-to-practice brief: Methods of evaluating teacher evaluation.* Washington, DC: Author.

New Teacher Project. (2010). *Teacher evaluation 2.0*. Brooklyn, NY: Author.

Noell, G. H. (2005). *Technical report of assessing teacher preparation program effectiveness—A pilot examination of value added approaches.* Baton Rouge, LA: Louisiana Board of Regents.

Norman, P. J., Golian, K., & Hooker, H. (2005). Professional development schools and critical friends groups: Supporting student, novice and teacher learning. *The New Educator, 1*(4), 273–286.

Peterson, K., Wahlquist, C., Brown, J., & Mukhopadhyay, S. (2003). Parent surveys for teacher evaluation. *Journal of Personnel Evaluation in Education, 17*(4), 317–330.

Peterson, K. D., Wahlquist, C., Bone, K., Thompson, J., & Chatterton, K. (2001). Using more data sources to evaluate teachers. *Educational Leadership, 58*(5), 40-44.

Regional Education Laboratory (REL) Midwest. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest region.* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Regional Education Laboratory (REL) Midwest. (2008). *State policies on teacher evaluation practices in the Midwest region.* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Regional Educational Laboratory (REL) West. (2009). *Measuring teacher effectiveness: Implications for California's Race to the Top.* San Francisco, CA: REL West at WestEd.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4,* 537–571.

Sanders, W., Ashton, J., & Wright, S. (2005). *Final report: Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress*. Arlington, VA: National Board for Professional Teaching Standards.

Sawchuk, S. (2009). Judging their peers: An old concept that calls for teachers to assess their own is gaining traction as evaluation comes under the spotlight. *Education Week, 29*(12), 20–23.

Schochet, P., & Chiang, H. (2010). *Error rates in measuring teacher and school performance based on student test score gains.* NCEE 2010-4004. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies In Educational Evaluation, 32*(1), 6–22.

Steele, J., Hamilton, L., & Stecher, B. (2010). *Incorporating student performance measures into teacher evaluation systems.* Santa Monica, CA: RAND.

Steiner, L. (2009). *Determining processes that build sustainable teacher accountability systems.* Washington, DC: National Comprehensive Center for Teacher Quality.

Stone, C., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1–26.

Stronge, J. (2007). *Qualities of effective teachers (2nd ed.)*. Alexandria, VA: Association for Supervision and Curriculum Development.

Thomas, B. (2010). *Teacher evaluation literature review.* Washington, DC: Council of Chief State School Officers.

Thornton, H. (2006). Dispositions in action: Do dispositions make a difference in practice? *Teacher Education Quarterly, 33*(2), 53–68.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

Vandevoort, L., Amrein-Beardsley, A., & Berliner, D. (2004). National board certified teachers and their students' achievement. *Education Policy Analysis Archives*. Retrieved from http://www.nbpts.org/UserFiles/File/National_Board_Certified_Teachers_and_Their_Students_Achievement_Vandevoort.pdf

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

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