AMERICAN
PSYCHOLOGICAL
ASSOCIATION

ASSESSING AND EVALUATING **TEACHER PREPARATION PROGRAMS**

APA TASK FORCE REPORT

ASSESSING AND EVALUATING **TEACHER PREPARATION PROGRAMS**

APA TASK FORCE REPORT

Task Force Members
Mary M. Brabeck, PhD
Carol Anne Dwyer, PhD
Kurt F. Geisinger, PhD
Ronald W. Marx, PhD
George H. Noell, PhD
Robert C. Pianta, PhD
Frank C. Worrell, PhD (Chair)

Staff
Rena F. Subotnik, PhD

The American Psychological Association wishes to acknowledge the support of APA's Board of Educational Affairs (BEA) and the Council for the Accreditation of Educator Preparation (CAEP) in developing this report. This report was received by the American Psychological Association's Council of Representatives on February 21, 2014.

Assessing and Evaluating Teacher Preparation Programs

Available online at: www.
Printed copies available from:
Center for Psychology in Schools and Education
Education Directorate
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242
Phone: 202.336.5923
TDD/TTY: 202.336.6123
Email: rsubotnik@apa.org

Suggested bibliographic reference:
Worrell, F. Brabeck, M., Dwyer, C., Geisinger, K., Marx, R., Noell, G., and Pianta R. (2014). *Assessing and evaluating teacher preparation programs.* Washington, DC: American Psychological Association.

Copyright© 2014 by the American Psychological Association. This material may be reproduced in whole or in part without fees or permission provided that acknowledgment is given to the American Psychological Association. This material may not be reprinted, translated, or distributed electronically without prior permission in writing from the publisher. For permission, contact APA, Rights and Permissions, 750 First Street, NE, Washington, DC 20002-4242.

APA reports synthesize current psychological knowledge in a given area and may offer recommendations for future action. They do not constitute APA policy or commit APA to the activities described therein.

Cover photo used under Creative Commons license. Originally posted on Flickr by U.S. Department of Education.

CONTENTS

Abstract	01
Executive Summary	03
Assessing and Evaluating Teacher Preparation Programs	05
Using Student Learning Outcome Data to Assess Teacher Education Programs	13
Using Standardized Observations to Evaluate Teacher Education Programs	19
Using Surveys to Evaluate Teacher Education Programs	23
Cross-Cutting Themes in This Report	27
Recommendations	29
Task Force Member Biographies	33
References	37
Appendix	40

ABSTRACT

Effective teaching has long been an issue of national concern, but in recent years focus on the effectiveness of programs to produce high-quality teachers has sharpened. Long-standing achievement gaps persist despite large-scale legislative changes at the federal and state levels, and American students continue to show poorer performance on international tests compared to peers in other developed nations. These and other factors have resulted in the creation of new accreditation standards for teacher education programs. These standards, developed by the Council for the Accreditation of Education Programs (CAEP), require teacher education programs to demonstrate their graduates are capable of having strong positive effects on student learning.

The data and methods required to evaluate the effectiveness of teacher education programs ought to be informed by well-established scientific methods that have evolved in the science of psychology, which at its core addresses the measurement of behavior. Recent work highlights the potential utility of three methods for assessing teacher education program effectiveness: (1) value-added assessments of student achievement, (2) standardized observation protocols, and (3) surveys of teacher performance. These methodologies can be used by institutions to demonstrate that the teacher candidates who complete their programs are well prepared to support student learning. In this light, we discuss the evaluation of teacher education programs using these three methodologies, highlight the utility and limitations of each of these methodologies for evaluating teacher education programs, and provide a set of recommendations for their optimal use by teacher education programs and other stakeholders in teacher preparation, including states and professional associations.

EXECUTIVE SUMMARY

Effective teaching has always been important, and, in recent years, the effectiveness of programs to produce high-quality teachers has become an issue of national concern. One fortunate outcome of this renewed focus on teacher education programs is the attention being paid to the creation of valid and efficient tools to assess that teaching force and teacher preparation. Recent scholarship has highlighted three methods—value-added models of student achievement, standardized observation protocols, and surveys of performance—that can be used by teacher education programs to demonstrate that the candidates who complete their programs are well prepared to support student learning. **The desire for evidence of program impact arises primarily from the acknowledged ethical and professional responsibility of teacher education programs to assure the public that they are preparing effective teachers for U.S. schools.** This report assumes the kinds of data and methods required to evaluate the effectiveness of teacher education programs ought to be informed by well-established scientific methods that have evolved in the science of psychology, which at its core addresses the measurement of behavior.

GUIDING PRINCIPLES OF THE REPORT

- **PreK–12 student learning is the central element of effective teaching and should be an ongoing part of teacher preparation, with implications for quality control, program improvement, and program fidelity-assurance.**
- **Validity is the most important characteristic of any assessment and is the foundation for judging technical quality.** Validity is a comprehensive concept, encompassing other critical concepts such as reliability, intended and unintended consequences of the assessment, and fairness. Irrelevant variation introduced by differences in assessment directions, observer training and biases, assessment locale, and a host of other factors will degrade the validity of the assessment system and the quality of decisions made on the basis of the data. Using multiple sources of data will result in better quality data for making valid inferences.
- Alignment of all of the elements of a program improvement effort is essential to determining what data to use, how good the data are, and what should and could be done with the data. Such alignment requires collaboration among teacher preparation programs, districts, and states. The design of explicit feedback loops from the data into program improvement activities is an important requirement of a good assessment process.
- Pursuit of some of the recommendations in this report would need to be phased in, because they involve considerable change for some programs, states, jurisdictions, and accrediting bodies. Professional associations, states, and

accrediting bodies can aid in the transitions by providing training for institutions and individuals that will permit programs to acquire the capacity to make the needed changes in a timely manner.

- **Faculty and administrators, state policymakers, and accrediting bodies must all make decisions about the merits of programs. These decisions should be made with the best evidence that can be obtained now, rather than the evidence we might like to have had, or that might be available in the future. Thus, we argue that we should not let the perfect be the enemy of the good. Decisions about program effectiveness need to be made using the most trustworthy data and methods currently available.**

RECOMMENDATIONS

Some of these recommendations can be implemented in the short term, whereas others will require a longer time frame to bring to full fruition. Teacher preparation programs can begin immediately to partner with schools, districts, and state education departments to develop plans for implementing these recommendations, leading to the best use of data for program improvement and accountability.

1. The Council for the Accreditation of Educator Preparation (CAEP) and local, state, and federal governments should require that teacher preparation programs have strong affirmative, empirical evidence of the positive impact of their graduates on preK–12 student learning.

2. States should work with teacher preparation program providers to design systems of data collection that include information collected at the stages of selection, progression, program completion, and postcompletion.

3. States and teacher preparation programs should track candidates' involvement in various preparation experiences and identify models of various program elements or candidate attributes that predict a positive contribution to preK–12 student learning.

4. States should work with teacher preparation programs to develop valid measures of student learning outcomes for all school subjects and grades to assess student learning outcomes similar to those currently available in mathematics, language arts, and science.

5. Teacher preparation programs, universities, not-for-profit organizations, school districts, states, and the federal government should dedicate appropriate resources for data collection and analysis.

6. Institutions and programs that prepare teachers should identify and retain staff with sufficient technical skills, time, and resources to conduct data analyses. They should partner with states and districts in this endeavor.

7. Institutions and programs that prepare teachers should commit to a system of continuous improvement based on examination of data about their programs.

8. Institutions that prepare teachers should train program faculty and supervising teachers in the use of well-validated observation systems and develop a system for regular "reliability" checks so that the observations continue to be conducted with a high degree of fidelity.

9. Federal agencies, state departments of education, research organizations, and teacher accreditation bodies should identify, develop, and validate student surveys that predict student achievement.

10. States, program faculty, and CAEP should continue to develop and validate developmental benchmarks and multiple metrics to be used by teacher preparation programs for graduation decisions to ensure that graduates are proficient teachers who make substantial impacts on student learning.

11. Teacher preparation faculty should develop curricula that prepare teacher candidates in the use of data such as student achievement scores, surveys, and observations so candidates can continue to self-assess, and faculty can assess the progress of their students.

12. CAEP and the states should report annually to the public any adverse impact of implementation of assessments on the teaching force or preK–12 learning.

13. States and CAEP should develop a time frame for implementing the recommendations made here.

ASSESSING AND EVALUATING TEACHER PREPARATION PROGRAMS

Effective teaching has always been important, and in recent years, this issue has become one of national concern. The increased focus on effective teaching is attributable to a variety of factors including (a) long-standing achievement gaps that persist despite large-scale legislative changes at the federal and state levels, (b) the poorer performance that American students continue to show on international tests compared to their peers in several other developed nations; and (c) the need to manage spending by governments at the national, state, and local levels. All of these factors have shined a spotlight on the nation's schools, the quality of the teachers in those institutions, and the effectiveness of the preparation that teachers receive in colleges and universities. The focus on teacher education is also being fueled by competition and comparison with alternative certification programs and the new standards proposed by the Council for the Accreditation of Educator Preparation (CAEP), which require programs to demonstrate that their candidates are capable of having strong positive effects on student learning.

One fortunate outcome of these trends is the attention being paid to the critical importance of teachers and the need for valid, effective, and efficient tools to assess the teaching workforce. Recent work has highlighted three methods—value-added assessments of student achievement, standardized observation protocols, and surveys of teacher performance—that are showing promising results in assessing teacher effectiveness. These methodologies can be used by teacher education programs to demonstrate that the teacher candidates who complete their programs are well prepared to support student learning while introducing these teacher candidates to the experiences that will continue to play an important role in their careers, assuming that future studies continue to yield findings similar to preliminary results.

A free and appropriate education for all students is one of the guiding principles of American public education, and there is a growing recognition that effective education has significant benefits for individuals and for the society. In addition to its many benefits to the individual, effective education confers public advantages such increasing society's productivity, tax revenues, and public health, as well as reducing costs of social services. Increasingly, as our society changes, we set higher standards for our students and thus recognize the need to require that teachers engage in practices that promote student learning. The effective education of our children and youth is thus premised on a cadre of effective teachers:

> In the present case, the story is about the visibility of teaching and learning; it is the power of passionate, accomplished teachers who focus on students' cognitive engagement with the content of what it is they are teaching. It is about teachers who focus their skills in developing a way of thinking,

reasoning, and emphasizing problem-solving and strategies in their teaching about the content that they wish students to learn. It is about teachers enabling students to do more than what teachers do unto them. (Hattie, 2009, pp. 237–238)

There is ample evidence that effective teachers are the most important in-school contributors to student learning in classrooms (Glazerman, Loeb, Goldhaber, Staiger, Raudenbusch, & Whitehurst, 2010; Harris, 2012; Hattie, 2009; MET Project, 2012b; Weisberg, Sexton, Mulhern, & Keeling, 2009). Furthermore, the majority of teachers are prepared in teacher education programs in the nation's colleges and universities. Given the importance of teachers and teaching, a focus on assessment and evaluation of teachers' performance, both for purposes of improvement and for accountability, should be no surprise. Teacher preparation programs need to demonstrate with evidence that teacher education makes a difference in preK–12 student learning. The need for evidence of teacher impact arises from the ethical and professional responsibility of teacher education programs to assure the public that they are preparing effective teachers for U.S. schools.

This report was prepared in the context of calls from many quarters for teacher education programs to show they prepare candidates who are ready to teach in ways that demonstrably impact preK–12 student learning in a positive way. Psychology's grounding in the measurement of behavior led the American Psychological Association's Board of Educational Affairs to support the development of this report's contribution to the policy arena. States play an important role in the process of ensuring a supply of well-prepared teachers: States certify and license teachers; they also must approve teacher preparation programs. Many states have partnered with major teacher education accreditation bodies in the approval process. There are, however, differences across states in how these responsibilities are carried out. Some states require program accreditation for state approval and some do not; some states require only state approval; and some states require both program accreditation and state approval. To date, however, **neither the state teacher approval process nor the accreditation process has required specific data demonstrating that candidates are effective** or that programs prepare teachers ready to teach all children to high levels.

Recently, many states have increased their standards for program approval, and the Council for the Accreditation of Educator Preparation (CAEP) has endorsed a more rigorous program accreditation process. As of 2016, CAEP will be the only educator preparation accreditation body. The 2013 CAEP standards require demonstration of program effectiveness by documenting that teacher education programs prepare teacher candidates ready to teach in ways that effectively promote preK–12 student learning.*

The Council of Chief State School Officers (CCSSO) is also committed to more demanding oversight of preparation programs in their states (see *Our Responsibility, Our Promise: Transforming Educator Preparation and Entry Into the Profession,* CCSSO Task Force on Educator Preparation and Entry Into the Profession, 2012). The public, including state and federal departments of education, parents, and principals, are demanding that teachers in our nation's schools are teaching effectively and that teacher education programs are successful in meeting this national need (see *Preparing and Advancing Teachers and School Leaders: A New Approach for Federal Policy,* Almy, Tooley, & Hall, 2013).

Teacher education program faculty and administrators must make concrete decisions about (a) whom to admit, (b) how to assess their teacher candidates' progress toward becoming effective teachers, and (c) whom to recommend for state licensure as teachers.

This report reviews the assessment strategies of teacher preparation programs considered by accreditation bodies and state departments of education, some of which are currently being used by teacher education programs. Approval of programs by states or accreditors may, of course, include additional requirements that are beyond the scope of this report. For example, many states require that programs report the retention of candidates in teaching and the numbers of teachers in understaffed areas (e.g., STEM, special education) or in hard-to-staff schools.

Specifically, this report discusses evaluation of teacher education programs through the use of three sources of data: (a) results of preK–12 student academic growth in academic learning as assessed by standardized tests (typically using what is called value-added assessment); (b) teacher performance as measured by observation instruments; and (c) surveys of teacher education program

* The Council for the Accreditation of Educator Preparation (http://www.caepsite.org) has restructured accreditation of teacher preparation programs with a stronger focus on increased selectivity of candidates, data-driven continuous improvement of programs, and preK–12-student learning outcomes.

completers, those responsible for hiring and supervising teachers, and the students taught by the graduates. These three types of data-collection methods are discussed in light of standards for technical quality (e.g. validity, reliability, and fairness) in three types of decisions:

1. Decisions about the *progress* of candidates in the teacher education program
2. Decisions about *recommending* candidates for licensure
3. Decisions about the *effects of teacher education program graduates on students' achievement* after completion of the teacher education program

Regarding decisions about selection (i.e., whom to admit to teacher education), we recognize that many programs collect data on candidates to inform their decisions about accepting a candidate. These data may include interviews, performance in preselection courses on teaching as a profession, grades in classes in their first year or two in college, standardized test results (e.g., GRE and SAT), and reports and references from previous instructors and supervisors in applied experiences such as tutoring. Moreover, there are some early stage selection or screening tools in development (e.g., Jamil, Hamre, Pianta, & Sabol, 2012) that have shown preliminary evidence of validity for predicting candidates' competence in classroom interactions. Some assessments of candidates' beliefs about teaching (e.g., persistence, views of child and adolescent learning) have also demonstrated early evidence of validity. There are few if any systematic uses of such instruments in teacher preparation, however, and very little, if any, validity data that predict competence in the classroom or are useful for making selection decisions. For these reasons, although we believe selection is a very promising area for research on measurement development and on the prediction of future competence in teaching, the current evidence does not support using the methods of data collection discussed in this report for purposes of selection. It is important, however, to acknowledge at the outset that whether sufficient tools are available today or not, programs will have to continue to make admissions decisions now, and they should do this with the best tools currently available.

No single methodology is perfect, and the types of data and methods discussed in this report all have assessment and psychometric limitations that must be taken into account when making decisions about programs or individuals. This is not to say these limitations mean the instruments should not be used: Important decisions that could benefit from improved data are currently being made and will continue to be made. The use of multiple measures generally assures the ability to make stronger inferences, but different programs with different foci may also need to tailor their assessments accordingly, making perfect standardization of measurement an important but elusive goal.

Despite this concern, decisions about program effectiveness need to be made consistently and fairly. **Using the most trustworthy data and methods currently available at any given decision point is the optimal way to proceed.** Using evidence derived from data that have been assessed scientifically and technically to make decisions is best practice, and enhancing the technical quality of data employed in decision making will improve decision making.

TIMELINESS OF THIS REPORT

Currently, there is far too little evidence regarding what aspects of teacher preparation lead to positive preK–12 learning outcomes in the classrooms of teacher education program completers. We believe that, over time, using the measures of behavior called for in this report will enhance the ability of programs to produce effective graduates. Historically, tools for evaluating the quality of graduating teacher education candidates were not widely available. However, we now have available several methods of data collection with good technical qualities that enable us to make better evidence-based decisions. We do not intend to imply there are no other methods than those discussed here; rather, we are reporting on approaches that are widely used and have been shown to be related to effective teaching.

There are also some infrastructure and policy developments that make this work timely. With Race to the Top (RTTT) funding, a number of states have developed systems to tie data on teacher education graduates to data on K-12 student learning; other states are doing so even without RTTT funding. Many states are now in the process of implementing Common Core Standards to ensure preK–12 students graduate from high school ready for college and careers. States are implementing plans to assess the attainment of higher order learning, and most states are in the

process of implementing new standards for evaluating teacher education programs in light of these new standards. All these developments demand reliable and valid data to inform decisions (see *Trends in Teacher Evaluation: How States Are Measuring Teacher Performance,* Center for Public Education, 2013).

New statistical methods have been developed to achieve the goals of understanding the contributions of teachers to student learning. Value-added assessment (VAA) has become a widely used method for evaluating effective teachers. In typical uses of VAA, each student's year-end scores are statistically adjusted to take into account different starting points and other potentially important factors such as student disabilities. When evaluations of teachers include consideration of the gains their students show on standardized tests, decisions are generally more effective and informed than decisions based on nonstandardized observations conducted by individuals who typically have not been trained in scientific principles of assessment observation (Harris, 2012; Glazerman et al., 2010; Goldhaber, 2010; MET Project, 2012b; Weisberg, Sexton, Mulhern, & Keeling, 2009). Despite this supportive evidence, some scholars have expressed concerns regarding the adequacy of value-added assessments (e.g., Baker et al., 2010), and these concerns are discussed in the following VAA-specific section. In this report, we consider the conditions under which VAA can be used to validly, reliably, and fairly evaluate teacher preparation programs, recognizing that VAA can yield more precise estimates for cohorts of teacher preparation graduates than for individual teachers because of advantages of aggregation (see for example, Goldhaber & Liddle, 2012).

In addition, new observation measures of teachers have been developed that are of high technical quality when the proper conditions are present. The use of well-established and rigorously validated observation measures is critical for several reasons. First, research on these measures indicates they can meet acceptable professional standards. **Measures that meet professional standards yield data that may be used to make better decisions; measures that do not meet such standards are likely to be misleading for fair and accurate decision making.** In this report, we examine how observation measures can be used to provide formative feedback to individual teacher candidates and useful information in the aggregate about educator preparation programs and the teaching effectiveness of their cohorts of graduates.

Finally, most teacher preparation programs conduct surveys pertaining to their graduates. Some institutional accreditation standards require such surveys, and some states are moving toward making these data a requirement for program approval. In addition, both principals who hire teachers and the preK–12 students in teachers' classrooms are in a position to provide feedback on teachers' performance. A brief overview of the survey research is provided in this report and recommendations for ensuring data of high technical quality from surveys that assess the effectiveness of a program's graduates. (A list of criteria to inform instrument selection appears in an appendix to this report.)

This report begins with a general overview of standards for technical quality then applies these standards to a discussion of the use of standardized achievement scores in assessing teacher preparation programs, observations of teaching, and surveys, as well as evaluation decisions programs make about candidates throughout the course of a teacher's development.

GENERAL OVERVIEW OF STANDARDS FOR TECHNICAL QUALITY

The effective use of data for program improvement assumes that the data are part of an integrated system that is valid, reliable, and fair. Alignment of all the elements of a program improvement effort is essential to determining what data to use, how good the data are, and what should and could be done with them. For this reason, the design of explicit feedback loops from the data into program-improvement activities is an important requirement of a good assessment process.

TECHNICAL QUALITY STANDARDS AND VALIDITY

The *Standards for Educational and Psychological Testing* (*Standards*) (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999) represents the consensus of the measurement field regarding the technical quality of assessment and is the product of a longstanding partnership between educators and psychologists on the importance of using assessment appropriately in the service of assessing learning outcomes.

***Standards* is clear that validity is the most important characteristic of any assessment and is the foundation for judging technical quality.**

Validity is a comprehensive concept, not a single mathematical value. It encompasses other critical concepts, such as reliability (the degree to which the data are replicable and accurate) and the intended and unintended consequences of the assessment activities under consideration. Very importantly, validity implies fairness. An assessment cannot be considered valid overall if it is only valid for a particular segment of those affected by it. It follows from this view of technical quality that a comprehensive, explicit, and detailed assessment design is needed in order to produce assessments of high technical quality and, ultimately, to improve decisions about educational programs for teacher candidates (Faxon-Mills, Hamilton, Rudnick, and Stecher, 2013).

DESIGNING THE EVALUATION CYCLE

This report focuses on assessment for program improvement and for high-stakes decisions such as program accreditation. **Formative evaluations should provide diagnostic information that helps produce successful summative evaluations.** To do so, formative assessments must be predictive of and highly related to the information used in summative evaluations. That is the basic rationale for providing information on the uses of assessment data for program faculty, relevant policymakers, accrediting bodies, and state regulators to guide their decision making. This information pertains to the components of preparation programs that contribute to teacher effectiveness, distinguishing these components from those that do not and focusing attention on increasing the impact and intensity of the components that are, in demonstrable fact, productive preparation experiences.

Determining the goals of the program-improvement process as a whole is thus a necessary first step. Deliberations on this point include several questions:

- What are the basic elements of the program that might be improved?
- What data are available or can be made available to evaluate each of these elements?
- What are the strengths and weaknesses of each type of data to be collected? Although every form of assessment has its strengths, each is also vulnerable to specific threats to its validity and utility. How can these strengths be maximized and weaknesses minimized over the data collection and analysis process as a whole?
- How can the data be analyzed and interpreted to yield useful information for decision making?
- Finally, can these data be used to inform the design of program improvement strategies?

Ultimately, program improvement is a necessary but not sufficient condition for a positive program evaluation outcome. CAEP or state approval implies that programs prepare teachers who are effective educators. Changes of the type just discussed cannot all be put in place immediately. Some need to be phased in because they involve considerable change for programs, states, jurisdictions, and accrediting bodies. CAEP and state education agencies will need to determine the time frame. However, **we do encourage those who judge educator preparation programs to implement these changes in a timely manner; sometimes even if that change may not be entirely comfortable for all participants.** Professional associations, states, and accrediting bodies can also aid in transitions by providing training for institutions and individuals involved in or impacted by the process. Groups such as the American Association of Colleges of Teacher Education (AACTE), the Council of Chief State School Officers (CCSSO), the Council for the Accreditation of Educator Preparation (CAEP), the American Educational Research Association (AERA), the National Council on Measurement in Education (NCME), and the American Psychological Association (APA) should all consider the kinds of training they believe most appropriate and offer such sessions in a manner that will permit programs to acquire the capacity to make the changes needed.

ALIGNMENT

A complete *assessment* design or assessment program starts with clear statements of what is to be measured and why; what data are to be collected and how are they to be analyzed; how are decisions to be made on the basis of the data; and how will the intended and unintended consequences of the assessment activities be evaluated. An important facet of the assessment design is the explicit alignment of the evaluation's overall goals with what is

actually measured and how inferences are drawn from the data and actions taken.

The following set of questions may be useful in understanding what the *Standards for Educational and Psychological Testing* implies about determining the technical quality of tests or other forms of data collection for purposes such as program improvement:

- How much of what you want to measure is actually being measured?
- How much of what you did not intend to measure is actually being measured?
- What are the intended and unintended consequences of the assessment?
- What evidence do you have to support your answers to the above questions?

Attention to alignment is critical to technical quality. Well-developed formal methodologies are available for planning and maximizing alignment and for ensuring alignment has actually happened as planned. These include logical and data analytical methods and attention given to the important practical and ethical questions of individual and institutional participation in the assessment design and implementation process. (See, for example, Porter, Polikoff, Zeidner, & Smithson, 2008; Porter & Smithson, 2001; Porter & Smithson, 2004.)

JUDGMENT AND ITERATIVE IMPROVEMENT IN TECHNICAL QUALITY

The technical quality of assessments cannot be determined by a few discrete methodologies summarized into simple numerical values. As the *Standards* clearly indicates, validity is a comprehensive human judgment that improves in accuracy over time and at multiple points in the educational cycle. Validity judgments take into account empirical and theoretical rationales for the inferences we draw from data.

Informed judgment also plays a key role in decision making when standards or cut-points must be established, such as when one decides what constitutes passing or failing or when labels that imply ordering or ranking are applied, such as "average," "satisfactory," or "excellent." A standard is not a quality inherent in an instrument itself, but is rather a judgment made on the basis of data and experience, taking into account the costs and risks of the different types of errors inherent in data-based judgments. For example, errors in decisions about the readiness of program completers to begin independent practice have different consequences for the potential new teachers than for their students. It is thus critical to consider the costs of different types of error and how to reduce error over time, while recognizing that human decision systems inevitably make errors. Again, numerous well-thought-out methodologies are available for planning and implementing standard-setting activities (Cizek, 2001; Cizek & Bunch, 2007).

Importantly, this point calls attention to a fundamental assumption and key infrastructure elements in the collection and use of data to inform program progress and approval of programs by either states or CAEP. The assumption is that **data-driven program improvement is a joint commitment of teacher preparation faculty, staff, and program leadership, along with deans, provosts, and presidents, and that cyclic feedback loops can and should be designed and used to focus change and improvement.**

From an infrastructure standpoint, this iterative, judgment-based, cyclic use of data implies the need for competent staff dedicated to this aim and capable of designing, collecting, analyzing, and reporting data for practical use in decision making. This assumption has both resource and personnel implications. It assumes, for example, that supervising teachers and "host" schools have effective student teaching supports in place and that clinical placements are aligned with the goals of the teacher preparation programs.

DEVELOPMENTAL PATHWAYS AND TEACHER PREPARATION

As shown in Table 1, the teacher preparation process can be seen as a developmental pathway that includes selection, progression, completion, and postgraduate/program evaluation. At each of these stages, it is necessary to consider which data-collection methodologies are appropriate and which types of data are to be obtained. For each type of data, the following are required:

- Standards for technical quality, including validity, reliability, and fairness
- An appropriate infrastructure for data collection and analysis
- Appropriate implementation plans, including cost mitigation

TABLE 1

Teacher Preparation Programs as Developmental Pathways

Selection	Progression	Completion Go/No Go/Recycle	Post Grad/ Program Evaluation
Methods for determining quality and diversity of candidates (*not discussed in this report*)	Observations during clinical experiences. Focus on feedback for individual as well as for program	Observations	Observations
		Student surveys aligned with supervising teacher survey	Student, employer, candidate surveys aligned with performance criteria
		Student learning outcomes Knowledge and understanding of content	Student learning outcomes

USING STUDENT LEARNING OUTCOME DATA TO ASSESS TEACHER EDUCATION PROGRAMS

Data regarding teacher candidate learning outcomes collected and analyzed throughout the program can serve as invaluable quality control, program improvement, and program fidelity-assurance measures. They describe the extent to which the program has hit its own internal benchmarks. However, these data do not address **what has emerged as the preeminent concern of consumers and policymakers: the effectiveness of educators in leading their students to high and increasing levels of achievement** (CCSSO Task Force on Educator Preparation and Entry Into the Profession, 2012; U.S. Department of Education, 2011a). It is not difficult to imagine a program that hits its own internal candidate benchmarks but prepares teachers whose impact on student achievement is unacceptably poor. This outcome could emerge due to such diverse factors as poor overall design of the program, critical content gaps in the program, or simply a mismatch between the contexts in which graduates teach and the design of their preparation program. Although the ability of programs to obtain meaningful student learning outcome data during preparation will be limited, it is possible to obtain some indicators that can and should contribute to formative decisions about teacher candidates as they progress through the program.

Initiating the process of collecting, evaluating, and making decisions using student learning data during teacher candidate training has a number of critical advantages for programs, candidates, consumers, and students (*Driven by Data,* Bambrick-Santoya, 2010). First, this data-based process is an overt expression of the core value that teaching is about producing measurable results for preK–12 students. Second, it provides a platform for coaching prospective teachers through data collection, evaluation, and decision making. This sort of progress monitoring and decision-making framework has been demonstrated to have substantial benefits for students (Fuchs & Fuchs, 1986; Ysseldyke & Tardrew, 2007). Third, it provides program faculty with candidate effectiveness data at a point at which the information is still actionable for current candidates. Fourth, it provides learning outcome data when observation and interview data about candidate practices are still readily available.

These sets of data provide a unique opportunity to examine questions about why results occur as they do. Finally, collecting, evaluating, and making decisions about student learning outcome data while candidates are progressing through the program should be viewed as more a design, management, and organizational commitment issue than as a cost, data availability, and capacity issue. Evaluating student learning is a critical element of effective teaching and, therefore, should be an ongoing part of preparation. **The key challenge should not be obtaining the data, but rather devising systems to capture data efficiently and systematically,**

creating standards for evaluating those data, and developing ways for using information from the data to inform faculty about the curriculum and the program in a thorough and timely manner.

A critical challenge in devising a measurement process to assess teacher candidates' contributions to student learning outcomes will be the diversity of the outcomes that educators are attempting to produce, such as in the arts, in social skills, or in the learning that takes place in kindergarten. The body of research identifying critical learning outcomes across the full range of educators' work is currently inadequate. The varied issues that have emerged in areas in which there is credible research supporting important metrics (e.g., Keller-Margulis, Shapiro, & Hintze, 2008; Wayman, Wallace, Wiley, Ticha', & Espin, 2007) should engender some caution in adopting metrics in untested areas. Additionally, newly prepared teachers will most often be distributed over many schools, districts, and hundreds of square miles after completing their preparation programs. The practical challenges in collecting such a diversity of measures across so many contexts argues for the need for teacher preparation entities and teacher preparation accreditation entities to partner with states engaged in systematic, statewide, teacher evaluation efforts. These state efforts may include student learning outcome data, value-added analyses, standardized observations, and supervisor surveys in some cases. These same data may in some cases be able to form the backbone of the assessment of teacher preparation when collected for new teachers and assessment results are shared in a coordinated and appropriate manner. The integration of state-mandated assessments of teaching and assessment of teacher preparation is an issue that cuts across all assessment and subject matter domains and that requires collaboration with many potential partners and alignment with state longitudinal data systems.

Waiting decades for the accumulation of research to identify key measures for every instructional domain is not a viable option. Teacher candidates are working with preK–12 students now, and faculty members have to make decisions about teacher candidate effectiveness and readiness to progress in the present. Faculty and administrators, state policymakers, and accrediting bodies must all make decisions about the merits of programs. These decisions will have to be made with the best evidence that can be obtained now, rather than the evidence we might like to have had or that likely will be available in the future.

STAGES OF TEACHER PREPARATION AND ASSESSING STUDENT LEARNING OUTCOMES

Progression

As teacher candidates progress through their preparation, programs typically engage in continuous assessment of the degree to which individual teacher candidates have achieved critical learning outcomes. The areas assessed usually include core professional values, content knowledge, pedagogical knowledge, and the ability to implement quality instruction to meet state and district standards. Assessment of candidates' achievement of these important milestones describes the success of the preparation program in achieving the institutional goal of producing a cadre of well-prepared novice teachers. Evaluation of these data will provide the program with evidence of the extent to which it has met its immediate proximal objectives of preparing effective teachers.

In some instances (e.g., teaching reading decoding skills to students in the early grades), well-developed practical progress-monitoring measures are available and are ready to be deployed. Fortunately, the research bases that undergird these measures have also yielded extensively documented criterion- and/or norm-referenced standards for acceptable performance or growth. Under these circumstances, selecting, implementing, and evaluating student progress measures for students taught by teacher candidates can be a relatively straightforward enterprise. Additionally, these types of measures lend themselves relatively directly to examining aggregate and program results. For most other subject areas (e.g., high school biology, instrumental band, special education for severely disabled students), similarly well-developed and technically adequate measures are not yet available. In these contexts, the only currently viable method is to devise explicit learning targets that are directly tied to immediate instructional goals and that can be directly and practically measured. If this process includes collaborative work involving candidates, supervising teachers, and faculty supervisors, it can also serve as a valuable teaching function. Additionally, this collaboration should serve to balance the practicality, relevance, technical adequacy, and the level of aspiration of the goals. Each member of the team will bring a different and valued perspective and type of expertise.

A critical unresolved challenge for this type of contemporaneously developed student learning assessment will be the establishment of standards for candidate performance and the aggregation of dissimilar data for program appraisal and continuous improvement. For example, aggregation of data based on differing scales and units presents unique challenges. In these circumstances, the best that can be achieved regarding standards for candidate evaluation may be reasonable standards based on faculty judgment, local norms, grade-level expectations, and typical growth patterns gleaned from published reports. Creating an index that aggregates the mean number of words read correctly per minute, parts of the cell correctly identified, and common denominators identified is meaningless. For purposes of program assessment, results will have to be converted to a common metric such as effect size or goal attainment scaling (e.g., Cohen, 1988; 1992; Kiresuk & Sherman, 1968).

It is also important to recognize that all participants in the assessment of student learning are interested parties in the evaluation process. It is reasonable to assume that all teacher preparation programs, school districts, and states will have a desire for positive outcomes, so that jointly creating substantive standards and creating and delivering training for faculty and staff in implementing the system will be particularly critical. Periodic peer review may also prove helpful in sustaining the quality of standards implementation.

Program completion

At the end of a teacher candidate's preparation program, the faculty needs to review an array of data to determine whether the candidate is ready to be recommended for licensure or whether additional preparation is needed prior to that endorsement. Strong affirmative evidence that candidates are able to facilitate and enhance student learning is clearly a critical prerequisite for teacher preparation programs to recommend candidates for completion and licensure. It is far less clear, however, how programs can obtain meaningful data about this or by what standards they should judge those data.

Two salient problems arise in assessing teacher candidates' teaching efficacy at the conclusion of their preparation program. The first critical challenge is identifying appropriate measures and performance standards across the array of topics, subjects, and grade levels at which candidates may teach. The second is how to separate the efficacy of the candidate from other intertwined factors, such as the efficacy of the supervising teacher and/or co-teaching partners.

Student teaching and internship can present unique challenges related to their duration, heterogeneity across settings, and distance from the preparation program. The key challenges that differentiate this stage of assessment from the progression stage are clarifying with the cooperating school the content the student teachers will be accountably responsible for teaching, and how to measure their students' progress with regard to this content. If these two major tasks can be accomplished, then the problem in many ways is reduced to that described under student learning outcomes in the Progression section of this report.

There is one special circumstance for the assessment of teacher candidates' contribution to student learning that is of special note. Some alternative route programs require that teachers serve as the teacher of record for the final year of their program (U.S. Department of Education, 2011b). In instances in which alternative route candidates serve as the teacher of record for a school year and also teach in grades and subjects where necessary data are available, it may be possible to use value-added results as part of the array of data used to evaluate candidates' readiness to enter the profession as certified independent practitioners. Since value-added data will be most broadly relevant to postgraduate assessment and program evaluation, the discussion of the issues surrounding value-added results is considered in a following section.

Postgraduate assessment and program evaluation

Assessment of new teachers' impact on student learning is arguably the most critically needed type of data to engage in a cycle of evaluation and continuous improvement for teacher preparation programs. It is also, unfortunately, the most difficult data to obtain. A few often-cited issues are that program graduates are likely to disperse over a large geographic area, including states other than where they were prepared; they will teach in many schools; they will teach many subjects; and they will teach at many grade levels. The diversity of what is taught creates a tremendous challenge for devising appropriate measures. The diversity of locations creates a daunting challenge for the collection of data in those circumstances for which reasonable measures can be identified. The financial demand of collecting new student learning measures across this diverse and

dispersed array of settings currently appears to be logistically and financially prohibitive. In addition, many, perhaps most, teachers are responsible for content or grade levels not covered by their state's standards assessments.

Finally, even in circumstances where measures can be obtained, the heterogeneity of the classes served creates enormous challenges for the interpretation of data. Endpoint-only analyses are clearly inadequate because the heterogeneity of students' starting points will result in incorrect conclusions regarding new graduates' efficacy (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Even pretest, posttest assessments are fraught with methodological challenges. Although they may provide a more accurate assessment, they do not completely sidestep the issue of individual differences and classroom heterogeneity. For example, it is probably not reasonable to expect similar science gains for students with intellectual disabilities and for typically developing students who share the same starting points. Research has suggested students with intellectual disabilities typically progress more slowly even after accounting for prior achievement (Noell, Porter, Patt, & Dahir, 2008). Similarly, a class that tests uniformly near the ceiling on an assessment will not be able to increase their scores to the degree a class scoring near the mean on the pretest can.

In most circumstances, given the complexity of the task and the logistical challenges, assessing the impact of new teachers on student learning will be beyond the resources of teacher preparation programs working in isolation. The viable solutions appear to be primarily dependent on partnerships with state education agencies or, in some cases, with large school districts, as well as other teacher preparation programs in the state. These entities may already have data that can be leveraged to examine the impact of new teachers on student learning, or they may have the capacity to obtain relevant data. In many states, the state education agency will be the only entity that has data available across all schools and districts within that state. In some cases, graduates of specific preparation programs may disproportionately serve a single or a few large urban school districts (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009), creating the possibility of capturing data from a substantial number of graduates by creating a partnership with that district or a small number of districts in that region.

Recognizing the necessity of collaboration between teacher preparation entities and administrative educational entities that have a broad geographic reach, such as state education agencies, raises the question of what data are available that would reflect student learning outcomes. In instances in which relevant data are available, they will typically be either teacher evaluation data or student standardized test data. Teacher evaluation data will have limited utility in assessing new teachers' impact on student learning when they consist of entirely nonstandardized evaluations by principals or supervisors. The extent to which nonstandardized evaluations are related to student learning outcomes is largely unknown and when attempts have been made to detect associations, they have not been found. Preliminary evidence suggesting large differences in student learning gains across classes (Glazerman et al., 2010; Goldhaber, 2010), and minimal differences in nonstandardized evaluations of teacher effectiveness strongly suggest typical ratings by principals and supervisors are likely to lack utility (Weisberg et al., 2009).

A second type of teacher evaluation measure that appears to have potential for assessing student learning gains is one for student learning goals, targets, or objectives. Student learning objectives (SLOs; New York State Education Department, 2012) will be used here as an omnibus term for these measures. SLOs are being used in various jurisdictions across the United States to assess student learning as part of teacher evaluation. The general structure of the process appears to be somewhat consistent across implementations. The SLO process typically begins with teachers selecting a small number of target measures for their class for the year. The teachers then collect a baseline measure in the fall and set a target for student achievement for the spring. An assessment is then administered in the spring and actual student progress is compared to the goal set by the teachers. The results are then judged against some standard, typically a manager appraisal, and that information feeds into teacher evaluation in ways that differ from one jurisdiction to another. The extent to which teachers have autonomy in the selection of measures and goals varies by content area, grade level, and jurisdiction. Limits may be placed on teacher decisions by such means as requirements for administrator input or by the requirement to use specific measures in some grades and subjects (e.g., New York State Education Department, 2012). Processes that

provide stronger guidance as to what measures and goals are acceptable will provide a basis for data analyses that could support the validity, reliability, fairness, and utility of these measures and thus their appropriate use in aggregate assessments of outcomes.

The critical problem in using SLOs to assess teacher preparation outcomes at present is that the process is just beginning in most large jurisdictions where it has been adopted. The psychometric properties of the data emerging from these measures and principals' evaluation of this information are not yet known. In applications where the procedures are vaguely defined, their properties may be unknowable. Additionally, the practical reality that, in at least some instances, teachers alone select the measures, and they themselves conduct the measurements upon which they will be evaluated in a high-stakes employment decision raises concerns about data integrity. Finally, the process of setting standards to judge the data, taking into account differences in class composition, is in a nascent stage of development. Attempting to leverage these data to provide feedback to teacher preparation programs might be desirable as part of an exploratory process in which the teacher preparation programs are part of field trials to create stronger standards for design, training, implementation, and evaluation.

The SLO process has tremendous intuitive appeal due in large part to its close connection to teaching and its adaptability to diverse classroom contexts. It is potentially applicable to all teachers. In addition, the process of establishing and measuring progress toward goals for student achievement may be beneficial in its own right (Fuchs & Fuchs, 1986; Karpicke & Roediger, 2008; Ysseldyke & Tardrew, 2007). The critical and immediate challenge is committing the resources necessary to refine the process to make it valid and more amenable to aggregation for decision making in contexts such as educator preparation program evaluation and school assessment. Given the enormous resources that will be committed to implementing these measures, this challenge is a critical priority.

The final type of data likely to be available is standardized student test data. As noted in an earlier section of this report, these data, when they meet reasonable data-quality standards, can potentially be leveraged to create value-added assessments (VAA) of teacher preparation programs. VAA have both the advantage and the disadvantage of having a relatively extensive literature base providing empirical evidence on the process' vulnerabilities and strengths. The appraisal of this same evidence when considering the use of value-added data for evaluating an individual teacher's work has resulted in stark divisions among respected scholars (e.g., Baker et al., 2010; Glazerman et al., 2010; Goldhaber, 2010; Institute of Education Sciences, 2012). Although the scholarly disagreements have many dimensions, some of which are touched upon here briefly, one of the core criticisms is that the differences observed between individual teachers are not sufficiently reliable to use them as an element of high-stakes evaluation decisions for those teachers (Baker et al., 2010). In contrast, other scholars have argued the level of reliability is better than that observed for scores on instruments used in other employment contexts, and that the information value for VAA greatly exceeds that of current teacher evaluation practices (Glazerman et al., 2010; Goldhaber, 2010). **Recent evidence suggests vaa results may be more reliable than the observation alternatives frequently recommended (Harris, 2012). In the context of teacher preparation programs, as contrasted with individual uses of vaa results, reliability can be improved by leveraging the many observations available from program graduates across schools and years.** In essence, the data at hand are used to determine the amount of data needed to obtain the desired level of reliability prior to reporting results.

A second technical concern that has been raised is that the variables and statistical controls used in VAA are insufficient to overcome the non-random sorting of students into teachers' classes. This is an "omitted variable" problem that can potentially be tested for in a number of ways in an analysis of teacher preparation program results (Gansle, Noell, & Burns, 2012; Noell et al., 2008). A variety of tests could be devised to examine heterogeneity of new teachers' effectiveness across programs within the same or similar contexts to examine the degree to which placement drives program results. Again in contrast to individual uses of VAA results, the volume of data necessary for high-quality data for decision making will typically be available when examining aggregate units such as teacher preparation programs. Similarly, there are decisions to be made about what to do about students instructed by multiple teachers. Current research supports the full attribution of a single student to

each relevant teacher (Hock & Isenberg, 2012).

A third critical concern is the degree to which integrating VAA into teacher assessment will inappropriately narrow the curriculum to what is tested (Baker et al., 2010). An assessment of teacher preparation programs using VAA as one of its elements will be limited to a minority of new teachers in most jurisdictions. Due to the absence of standardized testing in areas in which there are likely to be few teachers (e.g., foreign languages or physical education) and at some grade levels (e.g., kindergarten), many teachers will be excluded from this form of evaluation. It is important to note that with the expanding use of standardized end-of-course examinations at the high school level in many states, the potential coverage of VAA is expanding. The critical challenge in choosing to use VAA as an element of teacher preparation program assessment is weighing the advantage of knowing something about the instructional effectiveness of new teachers where data are available (e.g., for English, reading, mathematics, science, and social studies) versus the potential inequities that stem from not having equivalent data for other grades or subjects. Reasoned systems can be devised that make use of what can be known now, while remaining appropriately conscious of and cautious about what remains unknown.

In order for standardized test data and educational administrative data to be used in the value-added assessment of teacher preparation, some data-quality assurance thresholds will need to be overcome. First, the assessments must be psychometrically sound, reasonably strongly related to one another across years, and aligned to the instructional expectations that teachers will be targeting. The linkage of tests across years will be a particularly challenging issue when transitions to new assessments are being made, as will happen soon in many states with the adoption of tests aligned with the Common Core State Standards. The ability of analysts to provide technically sound value-added results across these transition points is not an issue that can be judged in advance. It will have to be resolved based on data that compare results across years based on actual administration of the tests. Second, sufficient, sound data linking students to teachers and teachers to preparation programs have to be available. Third, beyond students' achievement histories, data describing other critical information about students anticipated to influence results are needed (e.g., special education disability diagnoses, English language learner status, attendance, or giftedness status). Fourth, the available links between students across years must be sufficiently complete so the analysis is not undermined by large-scale and/or selective attrition due to unmatched records.

Assuming the data requirements have been met, an additional issue that analysts and policymakers will have to wrestle with is the standard for the counterfactual (Rubin, 2005). In essence this asks, "What teacher do you assume a student is most likely to have been taught by if they had not been taught by a new teacher from that preparation program?" Viable choices include the average teacher in the state, the average teacher in the district, the average new teacher, or an uncertified teacher. Each counterfactual creates a different standard for comparison, with some being more or less rigorous (Noell & Burns, 2006).

Although the analytic and data management work necessary to implement a VAA of teacher preparation is considerable, it is important to recognize it is likely to be a small expenditure compared to what is already being spent on student assessment systems. Despite significant challenges, initiatives in several states (e.g., Texas, Tennessee, Florida, Arkansas, and Washington) have succeeded in making progress in the use of VAAs. Further, in those contexts already deploying VAA to assess teachers' work, adding VAA to assessing teacher preparation represents a relatively modest amount of additional work.

USING STANDARDIZED OBSERVATIONS TO EVALUATE TEACHER EDUCATION PROGRAMS

Observation of teachers' interactions and classroom processes helps identify effective practices and can be a valuable tool in building capacity for teaching and learning (Allen, Pianta, Gregory, Mikami, & Lun, 2011; MET Project, 2010; Pianta, Mashburn, Downer, Hamre, & Justice, 2008). It is evident from the work done on training with observation protocols that large scale (e.g., national) implementation of observation assessment of teacher performance is possible. For example, a combination of live and web-based training protocols can sustain the training of thousands of observers to acceptable levels of understanding and agreement. When standard protocols for training and observation are used, there is evidence that observation scores capture features of teachers' behavior that are consistent from day to day, and across times of year or the content area the teacher is teaching (e.g., MET Project, 2010). Thus, observation protocols can indeed meet the technical standards for measurement presented earlier.

The pattern of results from classroom observations is quite consistent across grades, studies, and data-collection protocols: relatively low levels of teachers' support for student cognition and deeper understanding of content (feedback, focus on conceptual understanding, rich conversational discourse) and relatively well-developed classroom-management skills. A growing body of research documents that systematic observations in classroom settings can identify components of teacher–student interactions that contribute to students' social and academic development (e.g., MET Project, 2010; Mashburn et al., 2008; Pianta et al., 2005). Evidence links observed features of teachers' interactions with student learning, social development, motivation, and engagement. Moreover, observations of teacher behavior can be used to drive professional development demonstrated to improve those behaviors and student outcomes (Allen et al., 2011; Kane, Taylor, Tyler, & Wooten, 2010; Pianta, Mashburn, et al., 2008).

KEY CONSIDERATIONS WHEN USING OBSERVATIONS TO ASSESS SETTINGS

The use of standardized observations, if conducted validly, reliably, and fairly to measure those classroom interactions that impact student learning, is a direct and effective mechanism for focusing on teachers' behaviors. Such observations have the potential to illuminate links between certain inputs (e.g., resources for teachers) and desired outcomes (e.g., optimized student learning). The advantage of using valid tools standardized and clearly related to student outcomes is that with these kinds of observations, users can know they are making comparisons on an even playing field when noting strengths and challenges across classrooms, and they can know the behaviors they are observing are directly related to student growth and development (MET Project, 2010; Pianta & Hamre, 2009).

It should be clearly understood that the use of standardized observation tools is in no way at odds with giving personalized feedback to teachers; rather, it allows for the provision of highly targeted individualized feedback in clearly defined areas consistent across all teachers. These tools in combination provide a strong background for interpretation of scores. Using well-developed standardized tools is preferable in most circumstances to a highly customized approach in which every teacher preparation program, classroom, school, or district develops a tool on its own for which there is little comparative data and incomplete or absent validity evidence. Three sets of questions that follow should be asked to ensure valid, reliable, and fair use of observation measures.

Is the observation instrument well standardized in terms of its administration procedures?

Does it offer clear directions for conducting observations and assigning scores? *Standardization* refers to the rules and procedures for collecting observations to ensure consistency and quality control. These procedures include the qualifications of observers, length of the observation, and other practical and logistical features. It is important to select an observation system that provides clear, standardized instructions for use, both in terms of how to set up and conduct observations and how to assign scores. Without standardized directions to follow, different people are likely to use different methods, severely limiting the potential for agreement between observers when making ratings, and thus seriously limiting the validity of inferences that can be drawn from the data.

We recommend that three main components of standardization should be considered when evaluating an observation instrument: (a) the training protocol, (b) the policies and practical procedures for carrying out the observations, and (c) scoring directions.

With regard to the training protocol, are there specific directions for learning to use the instrument? Is there a comprehensive training manual or user's guide? Are there videos or transcripts with "gold standard" scores available that allow for scoring practice? Are there guidelines for the level of training to be completed before using the tool (i.e., do all observers need to observe in a certain number of classrooms and demonstrate an acceptable level of consistency of judgments with their colleagues and the given assessment standards)?

It is critical to specify in detail *practical matters* such as the length of observations, the start and stop times of observations (are there predetermined times, times connected with start and end times of lessons/activities, or some other mechanism for determining when to begin and end?), directions for the time of day or specific activities to observe, whether observations are announced or unannounced, and other related issues. Many of these practical details can be profitably considered with reference to research and underlying educational policy implications, such as the desired degree of teacher autonomy and participation in the assessment development process (e.g., should the teacher choose the lesson to be observed?) and the degree to which the assessments are intended to foster careful lesson planning (e.g., should the observation include discussion of why a particular lesson was selected by the teacher for observation?).

With regard to scoring, are users conducting scoring during the observation itself or after the observation? Is there a predefined interval between the observation and scoring it? How are scores assigned? Is there a rubric that guides users in matching what they observe with specific scores or categories of scores (i.e., high, moderate, low)? Are there examples of the kinds of practices that would correspond to different scores?

Does the observation instrument include reliability information and training criteria?

Reliability of observer judgments is a key consideration in selecting an observation assessment tool. Reliability is a property of any measurement that refers to the degree of error or bias in the scores obtained. It addresses the extent to which a tool measures those qualities consistently across a wide range of considerations that could affect a score (e.g., different raters, length of the observation period, variability across lessons, rater training). In observation assessments of classrooms, a reliable tool is one that produces the same score for the same observed behaviors, regardless of features of the classroom that lie outside of the scope of the tool, and regardless of who is completing the ratings. It should be noted that an observation system, like any other assessment, can be reliable without being valid, but that it cannot be valid without being reliable.

Assuming an important assessment goal is to detect consistent and stable patterns of teachers' behaviors across situations in the classroom, the measures need to be demonstrably consistent across time. It is advantageous

if observation tools provide information on their test-retest reliability/stability and the extent to which ratings on the tool are consistent across different periods of time (e.g., within a day, across days, across weeks) and, of course, across observers.

Is there evidence for an association between observation data and desired student outcomes?

We must know that our assessment tools are directly and meaningfully related to criteria of interest before using them either for program improvement or for accountability. If an observation tool is well aligned with the questions to be answered about classroom practice and meets technical quality standards, it is possible there may not yet be evidence available on the relation of these observations to the particular outcomes to be evaluated (e.g., student learning). In these instances, it is possible to use the observation in a preliminary way and to evaluate whether it is, in fact, associated with the specific outcomes of interest. For example, a teacher preparation program, district, or organization could conduct a pilot test with a subgroup of teachers and students to determine whether scores assigned using the observation tool are associated with the students' achievement as indicated by, for example, standardized test scores. However, since it will likely require two or more years to gather sufficient usable data, it might be easier for a teacher education program to choose an instrument for which there is already validity evidence and concentrate on training raters.

The importance of selecting an observation system that includes rigorous evidence of validity with regard to student outcomes cannot be overstated. It may be difficult to find instruments that have been thoroughly validated, but this is essential for making observation methodology a useful part of teacher preparation program improvement and evaluation, assuming the end goal of such efforts is to increase the extent to which preparation program enrollment and experience lead to improved student learning. If the teacher behaviors evaluated in an observation are known to be linked with desired student outcomes, teachers will be more willing to reflect on these behaviors and "buy into" observation-based feedback. Teacher educators then can feel confident establishing observation-based goals and mechanisms for meeting those standards, and educational systems, teachers, and students will all benefit (Allen et al., 2011; MET Project, 2010; Pianta & Hamre, 2009).

THE QUALITY OF CURRENTLY AVAILABLE OBSERVATION INSTRUMENTS

The vast majority of protocols for observing teacher performance in present use, whether in teacher preparation or for practicing teachers in the field, lack evidence of reliability and validity. Most are "home-grown" assessments derived from focus groups or consensus. If they are "off the shelf," then the evidence for psychometric properties may well be lacking. In short, the "market" for selection and use of observation protocols lacks the very contingencies that would drive selection of appropriate instruments or the use of them in ways likely to produce results that are fair, valid, or useful for evaluation or improvement.

An important review of teacher observation assessment instruments, the Measures of Effective Teaching (MET) Study (MET Project, 2012b) found two observation instruments that provide technically acceptable descriptions of teacher behaviors that can be applied across all content areas and grades: Classroom Assessment Scoring System (CLASS) and Framework for Teaching (FFT). There are other measures (e.g., edTPA™) being used in classrooms as a way to assess teaching. In the appendix, we provide a list of criteria that can guide the evaluation and selection of observation measures for use in teacher education programs.

STAGES OF TEACHER PREPARATION AND OBSERVATION MEASURES

Progression

It seems logical that progression through a teacher preparation program would be marked by regular assessments of candidates' competency in the classroom, particularly in domains of interactions with students that could be assessed by observation. Such periodic assessments would allow programs to do several things, including:

- build a data-driven approach to program design and improvement
- track the growth of competence for individual candidates
- track group performance year after year and conduct tests of program innovations and elements
- provide accreditation agencies information pertinent to program quality
- help align training experiences to outcomes

- build program coherence around a common language and lens for practice
- assign teacher candidates early and preventively to appropriate training experiences

As just one example of the importance of standardized observation assessments of teacher–student interactions, consider the fact that well over 95% of the nation's teacher candidates are observed during their teaching placements, ostensibly to gauge their skill level with regard to competencies deemed desirable or even necessary by their higher education program or state licensure systems (LaParo, Scott-Little, Ajimofor, Sumrall, Kintner-Duffy, Pianta, Burchinal, Hamre, Downer, & Howes, 2013). Then consider that in fewer than 15% of these observations is there evidence for the reliability of the instrument, much less evidence of validity for the inferences being drawn (LaParo et. al., 2013). It behooves us to continue to refine the technical quality of observation measures and use the best measures in the development and assessment of teacher education programs. As noted earlier in this report, these qualities are clearly stated in the *Standards for Educational and Psychological Testing* and are based on a comprehensive view of validity that encompasses both reliability and fairness (AERA, APA, & NCME, 1999).

Programs might use standardized observations as a means to track students' progress across all practicum-related experiences as students move from more structured and fairly simple experiences (e.g., tutoring groups) to full responsibility for the classroom, such as in student teaching. Using observation tools as a means of tracking performance in the field should then link back to relevant didactic experiences, such as courses on classroom management or pedagogical methods. The use of standardized observations as a measure of progression should always be performed by raters (e.g., faculty, clinical supervisors, or peers) who have been trained to acceptable levels of proficiency in the use of the instruments.

Program completion

With observation measures of high technical quality, teacher candidates may in part be recommended for certification and licensure upon demonstration of effective practice. Moreover, developing and graduating teacher candidates who show high performance on valid measures may be used to evaluate programs. Valid, reliable, and fair standardized observations of teacher practice in the classroom, performed by observers trained to acceptable levels of inter-rater agreement, can be a robust and highly relevant marker of a candidate's competence to teach. Such a marker can be used as a gateway to the profession, a sign of the program's success in producing capable teachers, and a source of useful feedback for program faculty. When the observation used at program completion is the same as that used for earlier experiences (progression), the program's curriculum gains valuable coherence of practice. Moreover, when used at program completion under conditions in which some students will not reach the prescribed "bar" for completion, it may be possible to then target specific remediation and support activities to enhance the candidate's likelihood of passing the next time.

Postgraduate assessment and program evaluation

Currently, principals and supervisors engage in widespread use of observations in classrooms. Given the focus on the evaluation and accountability of teacher preparation programs in state and federal policy, reliable observations of graduates' competence in the classroom could be a particularly important tool in offering data for evaluating teacher education programs; offering the use of these observations in schools meets the technical standards described above. If valid and if aligned to program standards (and perhaps even to earlier assessments of candidates in teacher preparation), such *postgraduate follow-up assessments* could be powerful tools for program evaluation and improvement.

USING SURVEYS TO EVALUATE TEACHER EDUCATION PROGRAMS

Several types of surveys have been used in assessing teacher performance over the years. These include (a) surveys of teachers about their satisfaction with their training and their perceived competence in job performance, (b) surveys of employers (e.g., principals and school district personnel) asking about the performance of teachers from the various institutions that serve as teacher providers, and (c) surveys of students of graduates asking about their teachers' performance and behavior. Each of these types provides different information that can be used in assisting teacher education programs in their quest for data on effectiveness and ongoing improvement.

TEACHER, PRINCIPAL, AND OTHER TEACHER SUPERVISOR SURVEYS

Surveys of graduates are a common way for teacher education programs to assess the success of their programs (Darling-Hammond, Eiler, & Marcus, 2002). Some surveys are created and administered by individual programs; some are administered by researchers interested in examining individual programs or groups of programs; and still others are administered by states, federal agencies, and national teacher education organizations.

These surveys typically assess teachers' ratings of the program they attended and their estimation of how prepared they think they were for their teaching role when they graduated, and subsequently, especially at the time the survey is administered. Other questions focus on the teaching behaviors and practices teachers are currently engaged in, often in relation to professional standards. Such surveys can be either general or tailored to specific subject matter domains and grade levels. Similar types of surveys are sometimes administered to principals and other teacher supervisors.

Surveys of teachers have several advantages and disadvantages. Although these surveys can yield data on a large number of teachers at a relatively low cost and allow for comparisons across programs and cohorts, they can also suffer from the same shortcomings associated with nonstandardized observations of teachers. Many of the instruments are developed locally, on an ad hoc basis (Ellett & Garland, 1987; Loup, Garland, Ellet, & Rugutt, 1996), raising concerns about the validity of the inferences that can be made, the potential bias in self-report assessments, and the relationship of the ratings to actual student achievement (Kimball & Milanowski, 2009; Medley & Coker, 1987). Low response rates can also affect the representativeness, and therefore the validity, of survey responses.

More recent work has yielded encouraging conclusions. For example, Jacob and Lefgren (2008) found that principals' ratings of teachers' ability to raise mathematics and reading scores distinguished between the best and the worst teachers (in terms of producing achievement gains), although they were not useful in distinguishing teachers in the middle of the distribution. These ratings also correlated

significantly with value-added scores in reading (r = .29) and mathematics (r = .32). In another study, Jacob and Walsh (2011) reported that principals give higher ratings to teachers with higher student achievement, stronger education credentials, more experience, and fewer absences (see also Jacob, 2010).

STUDENT SURVEYS

Although student surveys have been used for many years (Aubrecht, Hanna, & Hoyt, 1986), they are most frequently employed in colleges and universities, and there is a robust literature on their effectiveness at the college level (Benton & Cashin, 2012). Originally, surveys of college students included questions about the general effectiveness of the instructor alongside questions relating to instructor characteristics such as clarity, enthusiasm, and organization. In more recent research in this area, the emphasis has been on teacher characteristics that are observable and can be operationalized in behavioral terms. **Rating observable teaching behaviors requires less judgment on the part of the students completing the survey, and thus is more likely to produce consistent results than are global questions that require students to make inferences about teacher performance.**

Although not in widespread use in preK–12 schools currently, there are several studies of middle school and high school students rating teachers, including a recent, ongoing large-scale project with participants from the fourth to ninth grades (MET Project, 2010). At this time, the current consensus is that scores on surveys completed by children in the primary grades (preK to Grade 3) are not reliable enough to be used to inform decision making. However, surveys with language targeted to the correct developmental level can yield reliable and valid scores with students in the upper elementary, middle, and high school grades (MET Project, 2010; Worrell & Kuterbach, 2001).

Over the years, student ratings of teachers have generated considerable debate. Opponents of student surveys of college teachers have contended that the ratings are influenced by course difficulty, workload, and grading leniency (Greenwald & Gillmore, 1997). Criticisms have also been leveled against the use of surveys in preK–12 classrooms, with a particular focus on students' lack of knowledge about teachers' content knowledge, curriculum requirements, and professional development activities (Goe, Bell, & Little, 2008). Another central concern is the fact that scores from student surveys on teachers have not been validated for use in summative decisions.

These criticisms notwithstanding, **student surveys of teacher effectiveness have considerable support in the empirical literature.** Scores of constructs based on observable behaviors are internally consistent and stable (Benton & Cashin, 2012; Burniske & Meibaum, 2012; Worrell & Kuterbach, 2001), are related to achievement outcomes in both college and K–12 students (Benton & Cashin, 2012; Burniske & Meibaum, 2012; MET Project, 2012a, 2012b), are more highly correlated with student achievement than are teacher self-ratings and ratings by principals (Wilkerson, Manatt, Rogers, & Maughan, 2000), and distinguish between more- and less-effective teachers identified using other metrics (Benton & Cashin, 2012; MET Project, 2010, 2012a, 2012b). Moreover, student surveys can be particularly useful in formative evaluation contexts because the scores can isolate areas in which teachers need to improve. For example, surveys of students from ethnic minority backgrounds may be particularly important, not only because these students are often on the lower end of the achievement gap, but also because these students may be particularly susceptible to teachers' perceptions (Jussim & Harber, 2005). Finally, **surveys of students are useful in distinguishing between teachers who hold high expectations for all of the students in their classrooms and teachers who do not (Weinstein, Gregory, & Strambler, 2004), a factor related to student achievement in all classrooms, but especially in classrooms serving low-income students and ethnic minority students.**

Thus, despite the criticisms, survey instruments continue to be in widespread use in higher education and are becoming more popular in preK–12 settings. That being said, researchers agree that student surveys should not be used in isolation and that data should be collected at multiple time points and from multiple classes (e.g., Peterson, 2004). Finally, there is a practical concern about using student surveys of student teachers if the district the teacher is placed in does not use these instruments and therefore has no established routine for collecting student survey data.

STAGES OF TEACHER PREPARATION: USING SURVEY DATA

As noted above, validity is not inherent in instruments themselves, but is based on evidence indicating that the

inferences drawn for a particular purpose make sense. Here, we consider the utility of and evidence in support of surveys in the context of the three decision points beyond Selection listed in Table 1.

Progression

With regard to examining teacher candidates progressing through a program, surveys can be potentially useful in providing data from supervisors on candidates' growth in and mastery of particular skill sets. Surveys of the candidates themselves, their students, and the master teachers in whose classrooms the candidates are teaching are also useful. The utility of surveys in formative evaluation is dependent on several factors, including (a) the availability of instruments yielding reliable scores that have been validated for the purpose of assessing student learning; (b) appropriate normative standards or benchmarks, even if local to the program, based on a sufficient number of teacher candidates' responses; and (c) an appropriate blueprint linking candidate scores to the program standards such that the aspects of performance that need to be remediated are clearly evident. Ideally, programs should use surveys of students, teacher candidates, and supervising teachers developed from the same program standards and blueprint the program uses as a framework. The formal observations of the candidates should be highly congruent with the constructs being assessed on the surveys.

Thus, teacher education programs interested in using surveys of their candidates need to begin by identifying a valid survey with evidence of high internal consistency and predictive validity. Alternatively, they can design their own survey specific to the program's training standards and context. In either case, the scores will need to be examined for reliability and utility in that context and validated and calibrated for the purpose of providing formative (and summative) feedback for teacher candidates, including deciding on cut scores to determine the need for remediation or additional practice, taking another look at a candidate, or coming to a conclusion about mastery of the skills being assessed. These options require a faculty member or consultant with expertise in measurement to work with the program as it develops the standards and processes that will inform the use of the scores. Although programs with large numbers of students may be able to gather sufficient data in a few semesters, smaller programs may need to gather data for several years before they have enough data to assess the validity of the inferences they wish to draw. There should be periodic, ongoing examinations of scores over several years to ensure items and constructs continue to work as intended.

Program completion

The program completion decision is essentially a binary one, with the program faculty deciding if they can recommend a teacher candidate for certification as a teacher or not. As such, it is typically a decision with much higher stakes than the decisions about instruction and remediation made continuously while the teacher candidate is progressing through the program. In the program completion decision stage, as in the case in any high-stakes decision, the use of surveys alone is not recommended. By the time teacher candidates are being considered for certification, they should have obtained survey ratings at or beyond the minimally acceptable level of performance determined by the program. Even with surveys that have very reliable and well-validated scores, this decision should be based on multiple performance indicators, which may include formal standardized observations by the program, feedback from the supervising teacher, and student ratings. At the program completion stage, surveys may be most useful in helping identify areas for remediation in cases where teacher candidates fall short across the multiple indicators that are being considered in the certification decision.

Postgraduate assessment and program evaluation

Given their established utility with in-service teachers (Jacob & Walsh, 2011; MET, 2010, 2012a, 2012b), surveys can be very useful as a program evaluation tool with former teacher candidates within a year of graduation and several years after graduation. As noted above, graduates can provide useful feedback about how prepared they felt by many key aspects of their teacher preparation program for their role, now that they are actually in the field. Surveys of principals and students can also complement the surveys of the graduates themselves to create a multi-informant perspective on the variables that the program expects and is assessing in its graduates. Student surveys also have a role in postgraduation assessments, as they can provide data on teachers' perceived effectiveness. **In the absence of student achievement data, student surveys may take on additional**

significance, as they are more highly correlated with student achievement than are surveys completed by other raters (Wilkerson et al., 2000).

Additionally, data at this stage can also be compared to the data collected when members of the cohort were teacher candidates and to data of the current cohort of teacher candidates. These data can then be used to identify trends related to factors such as years of experience, district demographics, and other factors that may be specific to the program (e.g., programs that are preparing teachers to work in urban districts). Comparative analyses of survey data (e.g., candidates and graduates, students, principals, and self-surveys) can also be used to revisit program standards with regard to establishing minimal levels of mastery in specific domains and to identify competencies that need to be enhanced or even added to the preparation program.

CROSS-CUTTING THEMES IN THIS REPORT

1 **Centrality of student learning. This report proceeds from the guiding principle that evaluating student learning is a critical element of effective teaching and therefore should be an ongoing part of preparation.** Data regarding the learning outcomes of the students of teacher candidates can serve as invaluable quality control, program improvement, and program fidelity-assurance measures. It is important to recognize that the preeminent concern of the general public and policymakers is the effectiveness of educators in leading their students to high and increasing levels of achievement.

2 **Identifying good teaching. This report proceeds from the principle that good and less-good teaching exists and that distinguishing more-effective and less-effective practice, validly, reliably, and fairly, although difficult, is possible.** A comprehensive understanding of teaching and its place in society and in individuals' lives involves many sets of values and perspectives. Of necessity, only a few of these many perspectives are fully addressed in this report. Given the appropriate care, attention, and resources, however, we believe we have demonstrated that teaching skills can be analyzed and improved in order to benefit students and society.

3 **Validity as a basic framework.** Validity is the most important characteristic of any assessment and is the foundation for judging technical quality. Validity is a comprehensive concept, not a single mathematical value, and it involves human judgment. It encompasses other critical concepts, such as reliability and the intended and unintended consequences of the assessment. Validity also implies fairness. An assessment cannot be considered valid overall if it is only valid for a particular segment of those affected by it.

4 **Validity and multiple methods.** Because no single measure or methodology is sufficient in itself, it follows that using multiple sources of data will result in better quality data for making decisions. In creating an assessment system, it is useful to consider explicitly how much of what one intends to measure is actually being measured and how much of what one does not intend to measure is actually being measured.

5 **Validity and stakeholder participation.** A complete assessment system starts with clear statements of what is to be measured and why; what data are to be collected and analyzed; how decisions are to be made; and how the intended and unintended consequences of the assessment activities will be addressed. It is essential that relevant stakeholders be involved from the beginning. Jointly creating substantive standards and training for current and future faculty and staff in implementing the system will be particularly critical. Periodic peer review may also help sustain the quality of the assessment system's implementation.

6 **The data-decision-implementation loop.** The effective use of data for program improvement assumes the data are

part of an integrated system that is valid, reliable, and fair. Alignment of all the elements of a program improvement effort is essential to determining what data to use, how good the data are, and what should and could be done with the data. For this reason, the design of explicit feedback loops from the data into program improvement activities is an important requirement of a good assessment process.

7 **Standardization of implementation.** Valid and useful assessment systems that create a strong basis for inferences and decisions all rely on the premise that the data they use are logical, precise, and accurate. This implies a "level playing field" we have called standardization. This standardization may take different forms with different types of assessments, but the underlying principle is the same. Irrelevant variation introduced by differences in assessment directions, observer training and biases, assessment locale, and a host of other factors will degrade the validity of the assessment system and the quality of decisions made on the basis of the data.

8 **Training for participants.** Thorough and effective training in analyzing and using data for decision making will be necessary to create a valid, fair, and useful assessment system. It is unlikely that some of the recommendations in this report can be put in place quickly. Rather, they will need to be phased in because they involve considerable change for some programs, states, jurisdictions, and accrediting bodies. Professional associations, states, and accrediting bodies should aid in the transitions by providing training for institutions and individuals. Groups such as CCSSO, CAEP, APA, AACTE, AERA, and NCME should all consider providing the training they believe most appropriate and that will permit programs to acquire the capacity to make the needed changes in a timely manner.

9 **Perfect vs. good. Important decisions that could benefit from improved data are being made every day and will continue to be made whether or not high-quality data are available.** Faculty and administrators, state policymakers, and accrediting bodies must all make decisions about the merits of programs. These decisions will have to be made with the best evidence that can be obtained now, rather than the evidence we might like to have had or that likely will be available in the future. This presents the classic challenge of not letting the perfect be the enemy of the good. **Decisions about program effectiveness need to be made as consistently and fairly as possible, using the most trustworthy data and methods currently available to determine candidate effectiveness and readiness to progress now.**

RECOMMENDATIONS

We recognize these recommendations are ambitious and in some cases are associated with substantial costs in terms of financial and human resources and time commitment. In some cases these recommendations will require a cultural change in teacher preparation. We also note several gaps in the current literature call for additional research. As noted above, we also recognize the more fundamental point that there are many ways to view teaching for different purposes, and that, of necessity, this report does not fully address them all. There will always remain some aspects of teaching that may not be evaluated, but this should not deter us from addressing those that can and should be addressed. Also, as district, state, and federal agencies set the accountability agendas of educator preparation programs, these recommendations may help the programs demonstrate the enhancement of preK–12 student learning through better teacher preparation.

Making teacher education optimally effective will require collaboration among teacher preparation and school personnel; private and government funders; professional organizations; policymakers in districts and states; and local, state, and federal agencies. Teacher preparation programs that partner with schools, districts, and states can benefit from more consistent data collection while also gaining perspective on what districts and states need institutions to do.

Clearly, some of these recommendations can be implemented in the short term, while others will require a longer time frame to bring to full fruition. Programs can begin immediately to partner with schools, districts, and state education departments to develop plans for implementing these recommendations, which are most likely to lead to the best use of data for program improvement and accountability.

1 CAEP and local, state, and federal governments should require that teacher preparation programs have strong, affirmative, empirical evidence of the positive impact of their graduates on student learning.

2 States should work with teacher preparation program providers to design systems of data collection that include information collected at the stages of selection, progression, program completion, and postgraduation, including relevant indicators of performance at each stage. These systems of data collection should include instruments with the best available technical features. These systems should aim to provide longitudinal, prospective information on multiple constructs across the various outcome/performance assessments described in this report.

3 States and teacher preparation programs should track candidates' involvement in various preparation experiences and identify models of various program elements or candidate attributes that predict a positive contribution to preK–12 student learning. Federal and foundation funding sources

should provide resources to accomplish this critical empirical work.

4 States should work with teacher preparation programs to develop valid measures of student learning outcomes for all school subjects and grades to assess student learning outcomes similar to those currently available in mathematics, language arts, and science. When available, validated student learning objectives will enable teacher preparation programs to assess all their program graduates' performance relative to their impacts on the students they teach. Federal agencies and foundations should provide funding for the development of these assessments.

5 Teacher preparation programs, universities, not-for-profit organizations, school districts, states, and the federal government should dedicate appropriate resources for data collection and analysis. They must assign resources (time, infrastructure, technical capacity, funding) for faculty and/or professional staff to collect pupil and teacher data of high integrity and to regularly analyze and use these data for program improvement.

6 Institutions and programs that prepare teachers should identify and retain staff for data analysis with sufficient technical skills, time, and resources to conduct such analyses. In domains that require external data systems, such as preK–12 student achievement, institutions should partner with states and districts on data access and analysis.

7 Institutions and programs that prepare teachers should commit to a system of continuous improvement based on examination of data about their programs. They should allocate meeting time to discuss the results of data collection and analysis so various program members are informed of and able to reflect upon the results of these analyses. CAEP and the states should require that programs use the results of their data analyses annually to improve programs. CAEP and the states should also require programs to document how they have considered the data and used the data to inform changes made in the program; programs should assess those changes to see if they are effective.

8 Institutions that prepare teachers should train program faculty and supervising teachers in the use of well-validated observation systems and develop a system for regular "reliability" checks so the observations continue to be conducted with a high degree of fidelity. Programs should implement these observation tools (and associated training supports) at appropriate points in the teacher preparation program pathway and use the data for purposes of feedback at the candidate and program levels, as well as for state and CAEP accountability.

9 Federal agencies (e.g., NSF, IES, NIMH), state departments of education, research organizations (e.g. APA, AERA, NCME), and CAEP should identify and develop student surveys that predict student achievement. They should collect baseline data on teacher preparation candidates to develop a large enough sample to conduct psychometric analyses leading to benchmarks for suboptimal performance, adequate performance, and mastery. These surveys can and should provide feedback at the program and individual candidate levels on features of performance such as skills in motivating students, classroom management, and instructional competence. Research organizations and federal agencies should provide funding to calibrate student surveys with teacher self-reports to create appropriate benchmarks for both instruments.

10 States, program faculty, and CAEP should continue to develop and validate developmental benchmarks and multiple metrics to be used by teacher preparation programs for graduation decisions to ensure graduates are proficient teachers who substantially and positively influence student learning.

11 Teacher preparation faculty should develop curricula that prepare teacher candidates in the use of data such as student achievement scores, surveys, and observations so that candidates can continue to self-assess, and faculty can assess the progress of their students.

12 CAEP and the states should report to the public, on a regular basis, any adverse impact of implementation of assessments on the teaching force or preK–12 learning.

13 The states and CAEP should develop a time frame for implementing the recommendations made here. In general, these changes should be phased in in a manner that permits programs to make the necessary changes, but to do so as quickly as programmatically possible.

This report assumes that the kinds of procedures, data, and methods required to evaluate the effectiveness of teacher education programs ought to be informed by well-established scientific methods that have evolved in the science of psychology, which at its core addresses the measurement of behavior. In this light, as with all high-stakes decisions, we strongly recommend programs use these methods in combination rather than relying on any single method. We also encourage teacher education programs, in partnership with school districts and states, to invest time and resources in the development of systems that allow them to state affirmatively and with confidence that candidates completing their programs are making substantive contributions as new teachers to the learning outcomes of all of the students that they teach.

TASK FORCE MEMBER BIOGRAPHIES

FRANK WORRELL (CHAIR)

Frank C. Worrell, PhD, is a Professor in the Graduate School of Education at the University of California, Berkeley, where he serves as director of the School Psychology Program, faculty director of the Academic Talent Development Program, and Faculty Director of the California College Preparatory Academy. His areas of expertise include academic talent development, at-risk youth, scale development, teacher effectiveness, and the translation of research findings into school-based practice. Dr. Worrell is Coeditor of *Review of Educational Research* for the 2012–2014 term and a member of the editorial boards of several journals in psychology and education. He is a Fellow in four divisions of the American Psychological Association, a Fellow of the Association for Psychological Science, and an elected member in the Society for the Study of School Psychology. In 2011, Dr. Worrell received the Chancellor's Award for Advancing Institutional Excellence from UC Berkeley, and he was a recipient of the 2013 Distinguished Scholar Award from the National Association for Gifted Children.

MARY BRABECK

Mary M. Brabeck, PhD, is Gale and Ira Drukier Dean and Professor of Applied Psychology in the Steinhardt School of Culture, Education, and Human Development at New York University. Dr. Brabeck has published more than 100 papers, books, and chapters, and her research interests include intellectual and ethical development, values and conceptions of the moral self, professional and feminist ethics, and interprofessional collaboration through schools. She is a Fellow of APA (Divisions 7, 35, and 52) and AERA and serves as chair of the Board of Directors of the Council for the Accreditation of Teacher Preparation. She is an elected member of the Board of the New York Academy of Science, and her honors include a Doctor of Humane Letters from St. Joseph's University, Outstanding Alumni Achievement Award from the University of Minnesota, Alumni Award from the University of Minnesota School of Education and Human Development, Distinguished Alumni Award from St. Cloud University, APA Distinguished Leader of Women in Psychology Award, APA Corann Okorodudu Distinguished International Women's Advocacy Award, and the Kuhmerker Award from the Association for Moral Education.

CAROL DWYER

Carol Dwyer, PhD, has been concerned with assessment and equity as they relate to teaching and learning in both higher education and K–12. Her work includes both policy analysis and technical aspects of assessment theory, development, interpretation, and use. She has published extensively in the field of assessments' validity, with an emphasis on using construct validity theory to promote test quality, fairness, and appropriate use. She has also written about fairness and gender issues in communication skills, mathematical reasoning, grading practices, educational competitions, and other non-test achievement indicators. Dwyer was also the Principal Investigator for Research and Dissemination for the federally funded National Comprehensive Center for Teacher Quality, whose mission is to strengthen the quality of teaching, especially in high-poverty, low-performing, and hard-to-staff schools. Dwyer's interests include research and development of new forms of assessment. She was the architect and overall leader of The Praxis Series™ of national teacher licensing tests.

KURT F. GEISINGER

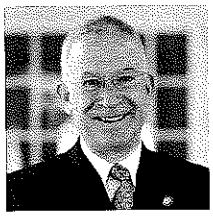
Kurt F. Geisinger, PhD, is Director of the Buros Center on Testing and Meierhenry Distinguished University Professor at the University of Nebraska. He served two terms as Council Representative for the Division of Measurement, Evaluation, and Statistics in the American Psychological Association. He also represented the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education on the ISO's International Test Standards committee. Within the American Psychological Association, he was elected to the American Psychological Association's Board of Directors for 2011–2013 and was nominated for its presidency in 2013. He was elected to the International Test Commission's Council/Board 2010–2012 and became its Treasurer in 2012, a position he holds. He served on the CAEP *Commission on Standards and Performance Reporting*, and currently chairs the CAEP Research Committee and serves on the CAEP Data Task Force. His primary interests lie in validity theory, admissions testing, proper test use, achievement testing, test use with individuals with disabilities and language minorities, and the adaptation of tests from one language and culture to another.

RONALD MARX

Ronald Marx, PhD, is Professor of Educational Psychology and Dean of Education at the University of Arizona, where he holds the Lindsay/Alexander Chair in Education. His previous appointments were at Simon Fraser University and the University of Michigan, where he served as the chair of the Educational Studies Program, codirector of the Center for Highly Interactive Computing in Education (winner of a Computerworld-Smithsonian Laureate for innovation in educational technology), and the codirector of the Center for Learning Technologies in Urban Schools (winner of the Council of Great Cities' Schools Urban Impact Award). In British Columbia, he conducted policy research that led to substantial reform of the province's schools. He has worked with computer scientists, science educators, scientists, and educational psychologists to enhance science education and develop teacher professional development models. His recent work with psychologists, early childhood educators, and public health researchers focuses on early education.

GEORGE NOELL

George Noell, PhD, is a Professor of Psychology at Louisiana State University and was previously Executive Director for Strategic Research and Analysis at the Louisiana Department of Education. At the Department of Education, his work focused on developing analytic systems that could support time-sensitive and long-term policymaking. Dr. Noell's research has focused on improving the quality and implementation of treatment plans for children in need of psychological services. Additionally, he has worked with partners to develop the Louisiana value-added assessment of teacher preparation, which is a statewide assessment linking student achievement to teacher preparation. Dr. Noell has worked with numerous state and national organizations to increase the use of quantitative data to inform policy decisions around programs for children and youth.



ROBERT PIANTA

Robert Pianta, PhD, joined the Curry School of Education in 1986 in the Clinical and School Psychology Program and was appointed dean in May 2007. The Novartis U.S. Foundation Professor of Education and a Professor of Psychology, he serves as director of the National Center for Research in Early Childhood Education and is the Founding Director of the Center for Advanced Study of Teaching and Learning at the University of Virginia. Dean Pianta's recent work focuses on the assessment of teacher quality, teacher–child interaction, and child improvement, using standardized observational assessment and video feedback. He has also extended his work into design and delivery of professional development using web-based formats and interactive video. Pianta is the Senior Author and Developer of the Classroom Assessment Scoring System (CLASS), a method for assessing teacher/classroom quality, being used in many district-, state-, and national-level applications. His assessments of teacher effectiveness are the national standard for Head Start classrooms and are included in the Gates Foundation's Measures of Effective Teaching study.

RENA F. SUBOTNIK (STAFF)

Rena F. Subotnik, PhD, is Director of the Center for Psychology in the Schools and Education (CPSE) at the American Psychological Association. The mission of CPSE is to generate public awareness, advocacy, clinical applications, and cutting-edge research to enhance educational and developmental opportunities for students at all levels of schooling, with a special focus on: (a) preK–12 teacher preparation programs, teaching skills, and professional development; (b) assisting educators and other school personnel in grounding both teaching and learning in current psychological science; (c) identifying attributes for success in school; and (d) advocating for and building a presence of psychology in the national and international education agenda, particularly in the preparation of teachers. Dr. Subotnik was a 1997 AAAS Congressional Fellow in Child Policy where she worked on teacher preparation program improvement and accountability for Senator Jeff Bingaman (D-NM). She previously served as professor of educational psychology at Hunter College, where she also coordinated the secondary teacher education program.

The task force wishes to acknowledge the invaluable contributions to this work on the part of Geesoo Maie Lee, Program Officer in the APA Center for Psychology in Schools and Education. The task force also greatly appreciates the advice and consultation provided by Jennifer Smulson, APA senior legislative and federal affairs officer.

REFERENCES

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333,* 1034–1037. doi:10.1126/science.1207998

Almy, S., Tooley, M., & Hall, D. (2013). *Preparing and advancing teachers and school leaders: A new approach for federal policy.* Retrieved from http://www.edtrust.org/sites/edtrust.org/files/publications/files/Preparing_and_Advancing_0.pdf

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Aubrecht, J. D., Hanna, G. S., & Hoyt, D. P. (1986). A comparison of high school student ratings of teaching effectiveness with teacher self-ratings: Factor analytic and multitrait- multimethod analyses. *Educational and Psychological Measurement, 46,* 223–231. doi:10.1177/0013164486461026

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L.,...Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers.* Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/page/-/pdf/bp278.pdf

Bambrick-Santoya, P. (2010). *Driven by data: A practical guide to improve instruction.* San Francisco, CA: John Wiley & Sons.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature.* Manhattan, KS: The IDEA Center.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Education Evaluation and Policy Analysis, 31,* 416–440. doi:10.3102/0162373709353129

Burniske, J., & Meibaum, D. (2012). *The use of student perceptual data as a measure of teacher effectiveness.* Austin, TX: The Texas Comprehensive Center at SEDL.

CCSSO Task Force on Educator Preparation and Entry Into the Profession. (2012). *Our responsibility, our promise: Transforming educator preparation and entry into the profession.* Washington DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/2012/Our%20Responsibility%20Our%20Promise_2012.pdf

Center for Public Education. (2013). *Trends in teacher evaluation: How states are measuring teacher performance.* Washington, DC: Author. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

Cizek, G. J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Corcoran, T., Mosher, F.A., & Rogat, A. (2009). *Learning progressions in science.* (CPRE Research Report #RR-63). New York, NY: Teachers College, Columbia University.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Princeton, NJ: Educational Testing Service.

Darling-Hammond, L., Eiler, M., & Marcus, A. (2002). Perceptions of preparation: Using survey data to assess teacher education outcomes. *Issues in Teacher Education, 11,* 65–84.

Ellett, C. D., & Garland, J. S. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up the 'state-of-the-art' systems? *Journal of Personnel Evaluation in Education, 1,* 69–92. doi:10.1007/BF00143280

Faxon-Mills, S., Hamilton, L.S., Rudnick, M., & Stecher, B.M. (2013). *New assessments, better instruction? Designing assessment systems to promote instructional improvement.* Santa Monica, CA: RAND.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53,* 199–208.

Gansle, K. A., Noell, G. H., & Burns, J. M. (2012). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education, 63,* 304–317. doi:10.1177/0022487112439894

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbusch, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Washington, DC: Brown Center on Education Policy at Brookings.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* 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. Retrieved from http://www.americanprogress.org/wp-content/uploads/issues/2010/12/pdf/vam.pdf

Goldhaber, D., & Liddle, S. (2012). *The gateway to the profession: Assessing teacher preparation programs based on student achievement* (Working Paper 65). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743-751. doi:10.1037/0022-0663.89.4.743

Harris, D. N. (2012). *How do value-added indicators compare to other measures of teacher effectiveness?* New York, NY: Carnegie Foundation. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Harris.pdf

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

Hock, H., & Isenberg, E. (2012). *Methods for accounting for co-teaching in value-added models.* (Report No. 7482). Washington, DC: Mathematica Policy Research.

Institute of Education Sciences. (2012). *Learning from recent advances in measuring teacher effectiveness.* Washington, DC: U.S. Department of Education.

Jacob, B. A. (2010). *Do principals fire the worst teachers?* (NBER Working Paper 15715). Ann Arbor, MI: University of Michigan and the National Bureau of Economic Research. Retrieved from http://closup.umich.edu/files/closup-wp-20-baj-principals-fire-worst.pdf

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26,* 101–136. doi:10.1086/522974

Jacob, B. A., & Walsh, E. (2011). What's in a rating? *Economics of Education Review, 30,* 434–448. doi:10.1016/j.econedurev.2010.12.009

Jamil, F., Hamre, B., Pianta, R., & Sabol, T. (2012). *Assessing teachers' skills in detecting and identifying effective interactions in classrooms.* Manuscript submitted for publication.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9,* 131–155. doi:10.1207/s15327957pspr0902_3

Kane, T. J., Taylor, E.S., Tyler, J.H., & Wooten, A.L (2010). *Identifying effective classroom practices using student achievement data.* Working Paper 15803. Cambridge, MA: National Bureau of Economic Research.

Karpicke, J.D., & Roediger III, H.L. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term accuracy of curriculum based measures in reading and mathematics. *School Psychology Review, 37,* 374–390.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45,* 34-70. doi:10.1177/0013161X08327549

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal, 4,* 443–453. doi:10.1007/BF01530764

LaParo, K., Scott-Little, C., Ajimofor, A., Sumrall, T., Kintner-Duffy, V., Pianta, R., Burchinal, M., Hamre, B., Downer, J. & Howes, C. (2013). Student teaching feedback and evaluation: Results from a seven-state survey. Manuscript under review, *Early Childhood Research Quarterly.*

Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest districts. *Journal of Personnel Evaluation in Education, 10,* 203–226. doi:10.1007/BF00124986

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D.,....Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79,* 732–749. doi:10.1111/j.1467-8624.2008.01154.x

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

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

MET Project. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project.* Seattle, WA: Bill and Melinda Gates Foundation.

MET Project. (2012a). *Asking students about teaching: Student perception surveys and their implementation.* Seattle, WA: Bill and Melinda Gates Foundation.

MET Project. (2012b). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill and Melinda Gates Foundation.

New York State Education Department. (2012). *Student learning objectives: Teacher overview.* Albany, NY: Author. Retrieved from http://www.nassauboces.org/cms/lib5/NY18000988/Centricity/ModuleInstance/1734/slo- teacher-overview.pdf

Noell, G. H., & Burns, J. L. (2006). Value added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57,* 37–50.

Noell, G. H., Porter, B. A., Patt, R. M., & Dahir, A. (2008). *Value added assessment of teacher preparation in Louisiana: 2004-2007.* Baton Rouge, LA: Louisiana Board of Regents. Retrieved from http://www.regents.state.la.us/Academic/TE/Value%20Added.htm

Peterson, K. (2004). Research on school teacher evaluation. *NASSP Bulletin, 88,* 60-79. doi:10.1177/019263650408863906

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109–119. doi:10.3102/0013189X09332374

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS.)* Baltimore: Paul H. Brookes.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre- kindergarten classrooms. *Early Childhood Research Quarterly, 23,* 431–451. doi:10.1016/j.ecresq.2008.02.001

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9,* 144–159. doi:10.1207/s1532480xads0903_2

Porter, A.C., Polikoff, M.S., Zeidner, T., & Smithson J. (2008). The quality of content analyses of state student achievement tests and content standards. *Educational Measurement: Issues and Practice, 27*(4), 2–14.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states—One hundredth yearbook of the National Society for the Study of Education, Part II* (pp. 60–80). Chicago: University of Chicago Press.

Porter, A. C., & Smithson, J. L. (2004). From policy to practice: The evolution of one approach to describing and using curriculum data. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability, NSSE Yearbook 103:2.* Chicago, IL: The National Society for the Study of Education.

Rubin, D. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100,* 322–332. doi:10.1198/016214504000001880

U.S. Department of Education. (2011a). *Our future, our teachers: The Obama administration's plan for teacher education reform and improvement.* Retrieved from http://www.ed.gov/sites/default/files/our-future-our-teachers.pdf

U.S. Department of Education. (2011b). *Preparing and credentialing the nations' teachers: The Secretary's eighth report on teacher quality based on data provided for 2008, 2009, and 2010.* Retrieved from http://title2.ed.gov/TitleIIReport11.pdf

Wayman, M. M., Wallace, T., Wiley, H., I., Ticha', R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41,* 85–120. doi:10.1177/00224669070410020401

Weinstein, R. S., Gregory, A., & Strambler, M. J. (2004). Intractable self-fulfilling prophecies: Fifty years after Brown v. Board of Education. *American Psychologist, 59,* doi:10.1037/0003-066X.59.6.511

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Washington, DC: The New Teacher Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360° feedback for teacher evaluation. *Journal of Personnel Evaluation in Education, 14,* 179–192. doi:10.1023/A:1008158904681

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

Ysseldyke, J., & Tardrew, S. (2007). Use of a progress monitoring system to enable teachers to differentiate mathematics instruction. *Journal of Applied School Psychology, 24,* 1–28. doi:10.1300/J370v24n01_01

APPENDIX:

Criteria for determining if an observation instrument should be used*

1. There has been empirical research on the instrument. (*Mandatory*)
2. There is evidence of inter-rater agreement among trained raters using the instrument. (*Mandatory*)
3. There is training available for users of the instrument. (*Mandatory*)
4. There is at least some preliminary research demonstrating the validity of the instrument that is "more than" just reliability information. Such information could include the following:
   - a. Correlations or other demonstrated relationships with student learning outcomes. (*Preferred*)
   - b. Correlations or other demonstrated relationships with other observation instruments, which have hopefully been validated themselves. (*Optional, but preferred*)
   - c. Strong evidence of theoretical underpinnings that have been at least in part shown to have been met (*Mandatory*).
   - d. Evidence of content validity such that individuals not related to the development of the instrument have found the content to represent the domain that is intended to be measured and have agreed that the domain is an appropriate one. (*Optional*)
5. The instrument has been used with diverse candidates in a manner that achieves fair results regardless of teacher, observer, or other group status. (*Optional, but highly preferred*)
6. The instrument has been used with positive results in a diverse grouping of schools. (*Optional, but highly preferred*)
7. The instrument has been used and/or cited in the teacher or teacher education research literature, with strong preference for inclusion in the peer-reviewed literature. (*Optional, but highly preferred*)

---

* Adapted from Spies, R. A., Carlson, J. F., & Geisinger, K. F. (2010). Introduction. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (ix-xviii). Lincoln, NE: Buros Institute of Mental Measurements.